SENAR

Supervised Engineering & Normative AI Regulation

Version 1.5 | 07.09.2026

Authors: Andrey Yumashev, Vadim Soglaev

CC BY-SA 4.0 | senar.tech

Содержание

CORE

SENAR Core

STANDARD

Introduction

Scope

Normative References

Terms and Definitions

Roles

Agent Instrumentation

Units of Work

Ceremonies

Quality Gates

Metrics

Operational Rules

Configurations

Maturity Model

Conformance

GUIDE

Quick Start

Philosophy

AI Output Review Checklist

Walkthrough

Transition Guide

SAFe Comparison

Failure Modes

Requirements Engineering

Legacy Adoption

Worked Example

Tool Integration

Agent Configuration

Onboarding

REFERENCE

Glossary

Scaling Ratios

Efficiency Model

Governance & Compliance

Tooling Requirements

Code Standards Template

SENAR Core

8 Rules for AI-Assisted Development

Version: 1.5 | **Date:** 2026-09-07 **Authors:** Andrey Yumashev, Vadim Soglaev **Copyright:** (C) 2026 Andrey Yumashev, Vadim Soglaev. Licensed under CC BY-SA 4.0. **Website:** senar.tech

> ***What is new in 1.5.*** *Rule 5 now agrees with the tier selection table below it: three tiers plus a reduced form of the first for small tasks, which is what that table has always listed. Nothing was removed and no rule changed in substance. The Standard changed around this document: it now names SENAR Core a normative reference (Standard 2.1), and Rule 7 finally has the equivalent its mapping table always claimed — a root cause criterion at QG-2 (Standard 8.3). In 1.4 one section was added — Three things that make a gate a gate.*

> *8 rules, 2 quality gates, 2 metrics, and a verification checklist — everything you need to work with AI coding assistants without shipping bugs. No specific tooling required. A text editor and version control are enough to start. As your practice matures, automation helps sustain discipline at scale. Reading time: 15 minutes.*

Who This Is For

You write code with AI assistants — Copilot, Claude, Cursor, or similar. You've noticed that AI output looks right but sometimes fails in production. You want a lightweight discipline that catches problems before they ship.

SENAR Core is that discipline. It works with any AI tool, any language, any workflow.

For organizations scaling AI development across teams, see the full SENAR Standard.

Key Term

Supervisor — the person directing the AI. If you work solo, that's you. In a team, it's whoever defines the task and reviews the AI's output. One person per task.

The 8 Rules

Rule 1. Task Before Code

Every change starts with a defined task:

- **Goal** — what you need to accomplish (one sentence).
- **Acceptance criteria** — numbered conditions that define "done." Each one independently verifiable.
- **Negative scenario** — at least one thing that could go wrong (error case, invalid input, edge case).

For tasks touching auth, payment, PII, or external APIs — include at least one acceptance criterion about input validation and access control.

Don't start coding without a task. Exploration is free, but committing exploration results requires a task.

Why this matters. AI output quality equals input quality. "Add login" produces plausible code that misses edge cases. "POST /auth/login returns JWT; returns 401 on wrong password; returns 422 on missing email" produces testable code on the first try.

Rule 2. Scope Boundaries

Every task explicitly defines:

- **What to change** — files, modules, or components in scope.
- **What NOT to touch** — at minimum: "do not modify files outside the listed scope." Be more specific when the task is near sensitive areas.

Without boundaries, AI agents will confidently refactor code outside scope, introduce unnecessary abstractions, or modify working systems you didn't ask about. Scope creep is the most common AI failure pattern.

Also helpful: specify patterns to follow ("follow the structure of `auth/router.py`") and constraints ("do not change the database schema").

Rule 3. Verify Against Criteria

Each acceptance criterion must be factually verified — run the test, check the output, measure the result, inspect the diff.

"It works" is not verification. "Looks right" is not verification. For each criterion, you need recorded evidence: a passing test, a command output, a measured value, a confirmed behavior.

No evidence for a criterion? That criterion is not verified. Where automated tests make sense, write them. Where they don't (documentation, config, infrastructure), other evidence works — command output, screenshot, grep result.

Never approve AI output based on intuition, code appearance, or the AI's claim that it's done. **The AI's confidence is not evidence.**

Rule 4. Tests Verify Requirements, Not Implementation

Tests come from acceptance criteria. They verify **what** the system does, not **how**.

A test that breaks when you refactor internals (without changing behavior) is an implementation test. Implementation tests create false failures during refactoring and false confidence during feature changes.

Good test: "POST /auth/login with valid credentials returns 200 and a JWT." Bad test: "The `_hash_password` method is called exactly once with bcrypt."

Watch AI-generated tests closely: AI tends to produce tests that mirror the implementation rather than verify the requirement. A test that just echoes what the code does provides zero verification value.

Rule 5. Check for Latent Defects

Before declaring "done," run through the verification checklist (see below). Three tiers based on risk, plus a reduced form of the first for tasks under 15 minutes — the tier selection table under *Verification Checklist* lists all four choices:

- **Standard tier** (every task): scope, deletions, phantom dependencies, test quality, test tampering, input validation, hardcoded secrets, stale patterns, cross-file consistency, code quality.
- **High tier** (security, auth, payment, data): Standard plus null guard bypass, empty config bypass, header trust, IDOR, return-True shortcut, auth coverage, unsafe deserialization, SSRF.
- **Critical tier** (production incidents, regulatory, complex features): all items.

Tasks involving user input, database writes, authentication, file I/O, or external APIs — use High tier at minimum. When in doubt, go higher.

Why this matters. AI-generated defects are hard to spot because they pass automated checks while containing subtle logical, security, or architectural flaws. The checklist targets specific patterns: returning True without validation, null equality bypasses, trusting HTTP headers, empty config values that silently disable security.

Rule 6. Zero Tolerance for Incomplete Work

Every claim of completion must be backed by evidence:

- "All tests pass" — show the test run output.
- "No lint violations" — show the linter output.
- "API returns 401" — show the response.
- "File was updated" — show the diff.

"Done" without verification is not done. "Probably works" is not done. If a task has 5 acceptance criteria, all 5 need recorded evidence. Claiming full completion with partial work is worse than honest partial completion.

Rule 7. Fix Causes, Not Symptoms

When you find a defect, identify the root cause before applying a fix.

A quick fix that suppresses the symptom creates technical debt, masks related defects, and guarantees recurrence. AI agents are particularly prone to symptom-level fixes — they'll add a null check where the real problem is that the data should never be null.

Before fixing, answer: "Why does this defect exist?" If the answer is "I don't know," keep investigating before patching. At minimum: identify the point of failure and the condition that triggered it. For complex defects (concurrency, state, external deps), reproduce the issue to verify your hypothesis.

Context hygiene: avoid sending PII, credentials, or regulated data to cloud AI tools without a Data Processing Agreement. See Standard Rule 10.12 for full requirements.

Rule 8. Capture Knowledge

Document the following during or right after completing a task:

- **Dead ends** — any approach that took more than 15 minutes and was abandoned. Record what you tried, why it failed, what you chose instead. (15 minutes is a

starting point — adjust to your domain. Based on empirical observation: beyond 15 minutes, the cost of documenting increases while the approach rarely succeeds.)

- **Decisions** — non-obvious choices with rationale (e.g., "chose argon2 over bcrypt because bcrypt import fails on Python 3.14").
- **Known issues** — limitations, workarounds, or tech debt introduced by the task.

Store knowledge where future AI sessions can access it — project docs, knowledge base, code comments, wherever works for you. Knowledge that exists only in your head doesn't exist for AI.

Why this matters. AI has no memory between sessions. Without captured knowledge, the same dead ends get explored repeatedly, the same decisions get re-debated, the same workarounds get rediscovered — at the same cost each time.

Quality Gates

Two quality gates that every task passes through. Enforce them however you want — manually or via tooling. The method doesn't matter; three things about the result do, and they are the next section.

NOTE: These are called "gates" to align with the SENAR Standard (Section 8), where Quality Gates are enforcement points and Checkpoints (Section 3.12) are context preservation actions within a session. The two concepts are distinct.

Three things that make a gate a gate

A checklist you tick is not a gate. Three things make the difference, and all three work with nothing but a text editor and version control.

Say what the gate stops. Write down what does not happen while the gate is unmet — "no code is written for this task", "the task is not called done". A gate whose purpose is "make sure the quality is good" cannot be applied to anything, because you cannot take a specific action and ask whether this gate stops it.

Judge the work as it is now, not as it was. A green run from yesterday says nothing about the file you edited an hour ago. If the work changed after the checks ran, run them again — otherwise the answer is no. This is the whole reason to record which version you checked.

When you cannot tell, the answer is no. The test suite did not finish, the tool crashed, you ran out of time — none of that is a pass. "Nothing bad was reported" and "everything is fine" are different statements, and only the second one lets the task through. A gate that waves

work through when it could not decide is worse than no gate: at least a missing gate is visibly missing.

Start Gate

(Called QG-0 in the full SENAR Standard)

Before writing any code, verify:

#	Check	Rule
1	Goal is defined (one sentence)	Rule 1
2	Acceptance criteria are listed (numbered, verifiable)	Rule 1
3	At least one negative scenario is included	Rule 1
4	Scope is defined (what to change, what not to touch)	Rule 2
5	For security-relevant tasks: threat surface identified, security AC included	Rule 1

If anything is missing, fill it in before directing the AI to start. Criterion 5 applies to tasks touching auth, user input, data storage, payment, or external APIs.

Done Gate

(Called QG-2 in the full SENAR Standard)

Before declaring a task complete, verify:

#	Check	Rule
1	All acceptance criteria have recorded evidence of verification	Rules 3, 6
2	Automated tests exist for testable criteria and pass; non-testable criteria have other recorded evidence	Rules 3, 4
3	Verification checklist passed at the appropriate tier	Rule 5
4	Root cause identified for any defects found during the task	Rule 7
5	Dead ends, decisions, and known issues documented (if none — state "none")	Rule 8

If anything is not met, the task stays in progress.

Metrics

Two numbers to track. Don't set targets immediately — measure for at least 3 work cycles to establish your baseline.

FPSR — First-Pass Success Rate

```
tasks_completed_without_rework / total_tasks_completed * 100%
```

The percentage of tasks completed correctly on the first attempt — no defects found after declaring "done," no re-opening, no rework. This is your primary indicator of whether the Start Gate and Done Gate are working.

Teams adopting SENAR Core typically see 50-65% initially, improving to 80-90% as the rules become habitual. These are illustrative ranges — establish your own baseline.

Dead End Rate

```
time_spent_on_dead_ends / total_task_time * 100%
```

Simpler alternative: `dead_ends_count / total_tasks` — average dead ends per task.

Measures how much effort goes into abandoned approaches. Reflects the effectiveness of knowledge capture (Rule 8). A declining rate means your knowledge base is working.

Above 30% warrants investigation. Below 10% in a mature project indicates effective knowledge reuse.

Verification Checklist

This checklist puts Rule 5 into practice. Pick the tier based on task risk.

Task Type	Tier	Items
Bug fix, config change, simple CRUD	Lightweight	4 (items 1, 3, 5, 7)
Normal feature, refactoring	Standard	10
Auth, payments, PII, external APIs	High	18
Production incident, regulatory, complex	Critical	28

Standard Tier — Every Task

> ***Lightweight mode for small tasks:*** *For tasks under 15 minutes (bug fixes, config changes, simple CRUD), check items 1, 3, 5, and 7 at minimum — scope, phantom dependencies, test tampering, and hardcoded secrets. Apply the full Standard tier for anything touching auth, payments, or user data.*

#	Check	What to Look For
1	**Scope**	Did the AI modify files outside the defined task scope?
2	**Deletions**	Did the AI silently remove or replace existing working code?
3	**Phantom dependencies**	Are all referenced dependencies real and available? (packages in manifest, APIs reachable, services declared)
4	**Test quality**	Do tests verify behavior from acceptance criteria, or just mirror the implementation?
5	**Test tampering**	Did the AI modify existing tests to make them pass instead of fixing the code?
6	**Input validation**	All user-supplied data validated for type, length, and format before use in queries, file operations, or commands.
7	**Hardcoded secrets**	No API keys, passwords, tokens, or credentials in source code.
8	**Stale patterns**	Does the code use deprecated APIs, removed functions, or patterns from older framework versions?
9	**Cross-file consistency**	If a type, interface, or contract was changed — are all consumers updated?
10	**Code quality**	Functions over 200 lines, duplicated logic, no reuse of existing utilities, unreadable structure. AI-generated code often works but is architecturally disposable.

High Tier — Security, Auth, Payment, Data

All Standard tier items, plus:

#	Check	What to Look For
11	**Null guard bypass**	Comparisons where both sides can be null/None/nil — `None == None` is True, bypassing access checks.
12	**Empty config bypass**	Security checks skipped when a config value is an empty string (e.g., `if secret and ...` fails open).
13	**Header trust**	HTTP headers (X-Forwarded-For, X-Partner-ID) used for security decisions without upstream proxy validation.
14	**IDOR**	Resources accessed by ID without verifying the requesting user's authorization to that specific resource.
15	**Return-True shortcut**	Access control functions that return True or grant access without performing explicit ownership validation.
16	**Auth coverage**	Every endpoint accessing user-specific data has authentication enforced; authorization checked at resource level.
17	**Unsafe deserialization**	Deserializing untrusted data without validation (pickle, yaml.load without SafeLoader, eval/exec on input).
18	**SSRF**	Server-Side Request Forgery: URLs from user input are validated against an allowlist; no unfiltered fetches of arbitrary URLs.

Critical Tier — Production Incidents, Regulatory, Complex Features

All Standard and High tier items, plus:

#	Check	What to Look For
19	**Dependency versions**	Are specified package versions real and published in the official registry?
20	**Hardcoded values**	Magic numbers, URLs embedded in code (secrets covered in Standard tier).
21	**Over-engineering**	Unnecessary abstractions, design patterns, or generalizations beyond what the task requires.
22	**Duplication**	New code that duplicates functionality already available in existing project utilities.
23	**Edge cases**	Happy path works — what about null, empty, boundary, concurrent, and high-volume inputs?
24	**Naming**	Does the AI follow project naming conventions consistently?
25	**Commit scope**	Is the commit atomic and focused on the task, or does it include unrelated changes?
26	**Format string injection**	String interpolation or formatting with untrusted input (Python `str.format`, JS template literals with eval, C printf).
27	**Unreachable safety code**	"Just in case" code paths that can never execute, masking incomplete control flow.
28	**Swallowed exceptions**	Catch/except blocks that discard errors silently — `except Exception: pass`, empty catch blocks, debug-level logging of errors.

What's Next

SENAR Core is the starting point. When your team grows or you need more structure, there's a clear progression:

	Core	Foundation	Team	Enterprise
Pairs	1	1–3	3–10	10+
Rules	8	8 Core + 10.2, 10.4, 10.9	15	15
Quality Gates	2	2 (QG-0, QG-2)	5 (QG-0..QG-4)	5 + compliance
Metrics	2	4	10	10 + portfolio
Roles	1	5 covered by 2 people	5 covered; 4.8 sets dedication	5 + portfolio
Ceremonies	0	3	7	7 + portfolio
Tooling	None	Recommended	Required	Required

> ***Note on "Pairs":*** *A Pair is one Supervisor working with one or more AI agents — typically one developer with their AI coding tool.*

Next step from Core: SENAR Foundation — add session management, all 5 responsibility sets covered by 2 people, and a set of 4 metrics: FPSR stays, Throughput, Lead Time and Defect Escape Rate join it, and Dead End Rate stops being one of the reported numbers. Takes 2-4 weeks after Core habits are solid. See the full SENAR Standard for details.

> ***Note:*** *Session management (structured start and end of AI work sessions) is introduced at Foundation level. Core focuses on individual task discipline.*

Conformance: If you follow all 8 rules and pass both gates consistently, you practice SENAR Core. No formal declaration is required — Core conformance is self-assessed.

Core rules map directly to Standard rules — nothing is unlearned:

Core Rule	Standard Equivalent
1. Task Before Code	Rule 1 (S10.1) + QG-0 (S8.1)
2. Scope Boundaries	QG-0 criteria + Guide habit
3. Verify Against Criteria	QG-2 (S8.3) + Rule 15 L2 (S10.15)
4. Tests Verify Requirements	QG-2 test criteria + QG-3 (S8.4)
5. Check for Latent Defects	Rule 15 (S10.15) + AI Review Checklist
6. Zero Tolerance	QG-2 criteria (S8.3)
7. Fix Causes, Not Symptoms	QG-2 criterion (S8.3) + Guide failure modes
8. Capture Knowledge	Rule 4 (S10.4) + Rule 9 (S10.9)

NOTE: Standard Rules 10.2 (Session Duration), 10.3 (Checkpoint Cadence), 10.5 (Periodic Audit), 10.6 (Version Control), 10.7 (Parallel Agent Limit), 10.8 (Complexity-Cost Calibration), 10.10 (Requirement Traceability), 10.11 (Code Documentation as Context), 10.12 (Context Hygiene), 10.13 (AI Model Governance), 10.14 (Script Change Management) are new at Foundation/Team/Enterprise configurations and have no Core equivalent.

Quick Reference Card

Before starting (Start Quality Gate):

1. Write the goal (one sentence).
2. List acceptance criteria (numbered, each independently verifiable).
3. Add at least one negative scenario.
4. Define scope: what to change, what NOT to touch.

While the AI works:

5. Verify each acceptance criterion with evidence (run it, test it, measure it).
6. Document any dead end that took more than 15 minutes.

Before declaring done (Done Quality Gate):

7. Run the verification checklist at the appropriate tier.
8. Confirm: all AC verified, tests pass, checklist passed, knowledge captured.

After the task:

9. Record decisions, known issues, and anything non-obvious.

10. Ask: could a new developer (or AI) continue this work without asking you questions? If not, capture what's missing.

SENAR Core 1.5 — senar.tech Copyright (C) 2026 Andrey Yumashev, Vadim Soglaev. Licensed under CC BY-SA 4.0.

SENAR

Supervised Engineering & Normative AI Regulation

Version: 1.5 | **Date:** 07.09.2026 **Authors:** Андрей Юмашев (Andrey Yumashev), Вадим Соглаев (Vadim Soglaev) **License:** CC BY-SA 4.0 | **Website:** senar.tech

---

> ***NOTE:*** *The following summary is informative. Normative requirements are defined in Sections 4–13.*

Start Here: Minimum Viable SENAR (MVS)

You are practicing SENAR when these six things are true:

1. **Every AI implementation has a Task** with a goal and acceptance criteria — before work begins.
2. **AI output is verified** by a Supervisor against acceptance criteria — never rubber-stamped.
3. **Tasks cannot start** without goal, acceptance criteria, and requirement link (Context Gate).
4. **Tasks cannot close** unless CI passes, tests pass, and types are clean (Implementation Gate).
5. **Throughput and Lead Time are measured** — you know how many tasks per session and how long they take.
6. **Dead Ends are documented** — when an approach fails, you write down why, so no one repeats it.

That's it. Do these six things and you are practicing the core of SENAR. For a structured entry point, see SENAR Core (8 rules, 2 gates, 2 metrics). Everything else in this standard builds on this foundation.

What Is SENAR

SENAR is a methodology for software development where AI agents are the primary producers of engineering artifacts and humans serve as Supervisors — directing, verifying, and governing AI output.

SENAR addresses a fundamental shift in software engineering: when AI produces the code, the human's value moves from **production** (writing code) to **judgment** (designing context, verifying correctness, making architectural decisions).

Existing methodologies (Scrum, SAFe, Kanban) were designed for teams of humans coordinating with each other. SENAR is designed for Supervisor+AI Pairs producing verified, traceable software.

SENAR Values

1. **Context over Code** — AI output quality is determined by input context quality
2. **Verification over Speed** — correctness is the constraint, not velocity
3. **Knowledge over Experience** — what is not documented does not exist for AI
4. **Enforcement over Agreement** — quality standards enforced through automated gates, not meetings
5. **Judgment over Keystrokes** — human attention on decisions and verification, not on typing

SENAR Core

SENAR Core provides 8 foundational rules for any team. This Standard extends Core with organizational processes, metrics, and governance. Teams adopting SENAR for the first time are advised to start with SENAR Core. Teams implementing SENAR Core are not required to conform to this Standard but are encouraged to adopt Foundation configuration when ready.

Document Set

Document	Purpose	Audience
SENAR Core	8 foundational rules — entry-level adoption	Any team, individual Supervisors
SENAR Standard (this document)	Normative requirements (SHALL/SHOULD/MAY)	Organizations, auditors
SENAR Guide	Philosophy, interaction patterns, training	Supervisors, adopters
SENAR Reference	Glossary, scaling ratios, economic model, compliance, tooling	Managers, compliance

Normative Language

Per RFC 2119: **SHALL** = required, **SHOULD** = recommended, **MAY** = optional.

Normative requirements appear in Sections 4–13 of this Standard. Sections 1–3 are definitional. Where a definition states SHALL, it fixes how a term is to be read and does not create an obligation standing on its own; the obligation it constrains is executed in Sections 4–13, and each has a place there — Task Scope (3.5.1) among the Task's required attributes (6.2) and the criteria of QG-0 (8.1); the extent of the term engineering artifact (3.5.2) through 8.6(a), which requires a declared effect to be testable against any proposed action; automation of gates (3.9) through 8.6(f). Where a definition instead fixes how a record is read — 3.5.1 on a scope narrowed after the work, 3.5.2 on an enumeration that omits a class of file — it states a reading and not a duty, and needs no executing clause. A conformance claim covering only Sections 1–3 is not valid (13.4), and no requirement depends on their being assessed. Section 4 is normative and definitional at once — it states responsibilities that SHALL be covered — and is counted as normative here and among the definitional sections in 13.4 for that reason. Section 12 is informative within that range and states no requirement — 12 itself says what would have to be settled before it could carry one. The Guide and Reference are informative.

Tooling

SENAR is tool-agnostic in choice but tool-dependent in practice. Automated quality gates, metric collection, and knowledge management require tooling support. See SENAR Reference (Tooling Requirements) for detailed capability requirements.

AI Capability Dependence

Some SENAR provisions are capability-dependent: they reference AI behavioral patterns (hallucination types, context window limits, instruction following characteristics) that change with model generations. The requirement to review such provisions when the AI model changes substantially is stated in Section 10.13; this summary states none. The SENAR Guide marks such provisions explicitly.

Empirical Basis and Limitations

SENAR's quantitative guidance (metric baselines, session duration guidelines, cost estimates) is derived from a single reference implementation: 552 tasks, $989 in AI costs, 38 sessions across 6 microservices. This constitutes a case study, not a controlled experiment. Organizations should treat these numbers as illustrative starting points, not universal targets — which is why Section 9 requires establishing your own baselines before setting targets.

Specific limitations: N=1 organization, self-reported metrics, pre/post design without control group, single AI model family. Independent replication across different organizations, domains, and AI models is needed to validate generalizability.

Coverage Limitations

The section above bounds the data behind this standard. This one bounds the survey behind its positioning. Section 1.5 states where SENAR sits relative to adjacent work; a statement of that kind is worth exactly as much as the list of what was not examined.

Examined by reading the primary source: the published normative texts of the works cited in Section 2.2 that are freely available.

Examined by structure and public analysis only: every paid publication cited in Section 2.2 — ISO 9001:2015, ISO/IEC 25010:2023, ISO/IEC 12207:2017, IEEE 29148:2018, CMMI for Development 2.0, ISO/IEC 42001:2023 and IEEE 1012-2024. Their structure, clause numbering, control titles, and public commentary were read; the text of their individual clauses was not, and this standard makes no claim about what that text says. Where this standard or its Reference names a clause of one of these works, the naming rests on published numbering and titles, which is enough to point at a clause and not enough to assert what it requires — an organization verifying conformance against any of them reads the source.

Not examined at all: ETSI, OASIS, the Trusted Computing Group, CEN-CENELEC, and the Japanese and Chinese standardization ecosystems. The survey was conducted in English against English-language sources; work published only in other languages is outside it.

The survey behind Section 1.5 was conducted on 2026-08-11 and re-checked on 2026-08-17. It establishes what was found, not what exists. A field moving at the rate this one moves invalidates such a survey quickly, and the interval between those two dates is the measure of it: six days apart, the re-check did not return the same picture as the survey. A reader who needs the positioning to hold at a later date re-runs the survey rather than relying on this one. This assessment is subject to revision, and organizations relying on the positioning in Section 1.5 should treat its date as part of the claim.

Intellectual Heritage

SENAR builds on established software engineering foundations: requirements engineering (IEEE 29148), quality cost models (Boehm, 1981), process maturity (CMMI/SEI), flow metrics (DORA, Accelerate), lean manufacturing quality practices (First Pass Yield, Right First Time), and human-AI teaming research. SENAR's contribution is the specific application and codification of these principles for AI-native development — where AI agents are the primary code producers and human engineers serve as Supervisors. The methodology does not claim to invent quality engineering; it claims to adapt it for a production model that did not exist when prior frameworks were created.

Versioning

Changes to normative requirements in the Standard are published as numbered versions (1.0, 1.1, 2.0). Guide and Reference materials may be updated between Standard versions. The changelog is maintained at senar.tech.

1. Scope

1.1 Purpose

SENAR defines a methodology for software development where AI agents are the primary producers of engineering artifacts and human engineers serve as Supervisors — directing, verifying, and governing the AI-driven process.

1.2 Intended Audience

- **Organizations** transitioning to AI-native software development;
- **Supervisors** — engineers who direct AI agents;
- **Managers** responsible for delivery, quality, and cost;
- **Tool vendors** building AI-native development platforms;
- **Auditors** evaluating quality practices of AI-native teams.

1.3 Applicability

This standard applies where:

a) AI agents generate a substantial portion of production artifacts (code, tests, configuration, documentation); b) Human engineers direct, review, and approve AI-generated output; c) Traceability from requirements to delivered artifacts is required; d) Quality assurance is enforced through automated mechanisms.

SENAR is tool-agnostic. It applies to any AI coding platform — autonomous agents, IDE assistants, terminal-based tools, or custom pipelines.

1.4 Out of Scope

a) AI model training, fine-tuning, or evaluation; b) Traditional development where humans write the majority of code; c) Organizational change management; d) Specific tool implementations.

1.5 Position and Relationship to Other Standards

Where this standard sits

Two bodies of work bear on the quality of software work. The subject of this standard falls between them.

Established software-engineering and quality standards govern **when work is verified, by whom, and to what rigor**. IEEE 1012-2024 governs verification and validation planning by integrity level, and its 2024 revision states applicability to systems built on generative AI. ISO 9001:2015 governs the quality management system surrounding the work. ISO/IEC 42001:2023 governs an AI system as a **product** an organization builds and operates — producer and product are the other way round from the case addressed here. What this body leaves to the adopting organization is where a control is placed relative to the work it governs, and what a verdict is required to be a statement about. It leaves them open correctly: its subject is the plan, not the mechanism.

The second body is agent-native practice — frameworks, protocols, and tooling that made the AI agent the addressee of the instruction. This body supplies mechanism. What it does not consistently supply is a stated criterion for when a check counts as a gate; checks are described as quality gates in work whose own documentation calls them recommended and provides no barrier.

SENAR addresses what falls between: the **properties a control is required to have** in order to govern work produced by an agent (Section 8.6), together with the units, roles, and measurements that make those properties operable.

What is claimed, and how to refute it

The claim is about properties, not about the absence of neighbours. Any framework, protocol, or platform whose gates have the properties of Section 8.6 satisfies what this standard asks at that point, whether or not it references SENAR. To refute the positioning, name a published work that states such properties as requirements on a control; to refute the standard's usefulness at this point, name a control governing agent-produced work that has all five and gained nothing by having them.

Basis for the positioning above: sixteen adjacent works were examined in an adversarial field review on 2026-08-11, and a focused re-check was run on 2026-08-17 against standards bodies, provenance formats, and agent tooling.

NOTE: Both examinations have recorded limits. Every paid publication named in Section 2.2 — ISO 9001:2015 and ISO/IEC 42001:2023 among them, and IEEE 1012-2024, all three

characterized above — was examined by structure, control titles and public analysis only; their clauses are therefore not quoted here, and Section 0 lists the full set examined on that basis. The search was English-language. Neither examination is a proof that no other work occupies this position, and neither is offered as one.

Adjacent subjects this standard does not cover

a) **Provenance of an artifact** — what produced it, and through which chain of custody; b) **The format of an attestation** — how a signed statement about an artifact is encoded, exchanged, and verified as a document; c) **Management of an AI system as a product** — the controls an organization applies to an AI system it builds, deploys, and operates.

These are the subjects of neighbouring work, referenced informatively in Section 2. A gate conforming to this standard consumes an attestation; it does not define the attestation's format.

Concepts adapted from established methodologies

SENAR extends and adapts concepts from established methodologies:

- **SAFe 6.0** — SENAR reimagines SAFe concepts for AI-native teams. SAFe comparison notes are provided throughout the document.
- **ISO 9001:2015** — SENAR Quality Gates are mapped against selected ISO 9001 clauses by their published numbering and titles; the mapping is not an assertion about clause text, which was not read (Section 0, survey coverage). Organizations should conduct gap analysis against the source for full compliance.
- **ISO/IEC 25010:2023** — SENAR metrics align with the published characteristics of the software quality model, on the same basis.
- **Scrum / Kanban** — SENAR borrows iterative delivery and flow measurement while replacing human-centric ceremonies with AI-appropriate alternatives.

SENAR may be extended with domain-specific profiles for regulated industries (medical devices, financial services, aerospace) that add controls required by sector-specific standards.

2. Normative References

2.1 Normative References

- **RFC 2119** — Key words for use in RFCs to Indicate Requirement Levels
- **RFC 8174** — Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, May 2017.
- **SENAR Core** — the companion document of this Standard: 8 rules, 2 quality gates, 2 metrics. It is invoked normatively in six places. 11.1 and 11.4 make its eight rules part of the Foundation rule set; 8.10 cites two items of its verification checklist by number; 8.1 cites its Start Gate criteria; 8.3 states that Foundation inherits the substance of a Team+ criterion from its Rule 4; and 5.7 permits, at Foundation, a substitution stated in terms of its High tier. A claimant at Foundation therefore cannot determine the applicable rule set from this Standard alone, which is what makes the reference normative rather than informative. The reverse direction does not follow: conformance to SENAR Core is declared separately and is not a claim against this Standard (13.3).

2.2 Informative References

These standards are referenced for context and alignment but are not normatively invoked by this Standard.

- **ISO 9001:2015** — Quality management systems — Requirements
- **ISO/IEC 25010:2023** — Systems and software Quality Requirements and Evaluation — Product quality model
- **SAFe 6.0** — Scaled Agile Framework (Scaled Agile, Inc.)
- **Scrum Guide** — The Scrum Guide (Schwaber & Sutherland, 2020)
- **Kanban Method** — Kanban: Successful Evolutionary Change for Your Technology Business (Anderson, 2010)
- **DORA Metrics** — Accelerate: The Science of Lean Software and DevOps (Forsgren, Humble & Kim, 2018)

- **ISO/IEC 12207:2017** — Systems and software engineering — Software life cycle processes
- **IEEE 29148:2018** — ISO/IEC/IEEE 29148:2018, Systems and software engineering — Life cycle processes — Requirements engineering.
- **CMMI** — CMMI Institute, "CMMI for Development, Version 2.0", 2018.

Supply chain, provenance, and attestation

This layer answers what produced an artifact and how a statement about it is encoded. This standard consumes such statements as evidence (Section 8.6) and does not define their format.

- **SLSA** — Supply-chain Levels for Software Artifacts (OpenSSF). Governs the **provenance of an artifact** — what built it, from which sources, under which isolation. Adjacent to Section 8.6(d): SLSA answers where an artifact came from, Section 8.6(d) requires that a verdict be bound to a digest the relying party measured.
- **in-toto Attestation Framework** — (OpenSSF/CNCF). Governs the **shape of an attestation**: subject, predicate type, and payload for a statement about an artifact. Section 8.6 requires evidence to correspond to state; it does not require this or any other envelope.
- **DSSE** — Dead Simple Signing Envelope (Secure Systems Lab / OpenSSF). Governs the **signing envelope** for such statements. Same boundary as above.
- **W3C PROV-DM** — Provenance Data Model (W3C Recommendation). A deliberately general **meta-model** for recording provenance; it leaves the semantics of verification to the consuming domain. It records what happened; it does not state when a verdict may be issued.
- **W3C Verifiable Credentials Data Model** — (W3C Recommendation). Governs typed, signed **attestations about identity and claims**. Adjacent as a document format; it does not specify a verification operation against a measured state.

AI system governance and V&V planning

This layer governs an AI system as a product, or the planning of verification, rather than the production of engineering artifacts by an agent.

- **ISO/IEC 42001:2023** — Information technology — Artificial intelligence — Management system. Governs the **AI system an organization builds and operates as a product**, through a management system with annexed controls. Producer and product are the other way round from the case this standard addresses.

- **NIST SP 800-218** — Secure Software Development Framework (SSDF), Version 1.1.
- **NIST SP 800-218A** — SSDF Community Profile for Generative AI and Dual-Use Foundation Models. Addressees are named in the profile itself: producers of models, producers of systems built on them, and acquirers. Artifacts produced by an agent, and the placement of a control over that production, are outside its stated scope.
- **IEEE 1012-2024** — IEEE Standard for System, Software, and Hardware Verification and Validation. Governs **V&V planning by integrity level**; the 2024 revision states applicability to systems built on generative AI. It leaves the placement of a control and the form of evidence open to the adopting organization.

NOTE: Seven of the works cited above are paid publications whose full normative texts were not read: ISO 9001:2015, ISO/IEC 25010:2023, ISO/IEC 12207:2017, IEEE 29148:2018, CMMI for Development 2.0, ISO/IEC 42001:2023 and IEEE 1012-2024. Their treatment here is based on published structure, clause numbering, control titles, and public analyses, and no claim is made in this standard about the content of their individual clauses. Section 0 states the same coverage limit for the survey behind Section 1.5.

---

3. Terms and Definitions

For complete glossary, see SENAR Reference.

Configuration notation: Where normative requirements differ by configuration, this standard uses the notation `[Team+: SHALL]`. "Team+" means Team and every configuration above it. Where the requirement is one Foundation carries at all, the notation indicates that it is SHOULD at Foundation and SHALL at Team and Enterprise. See Section 11 for configuration definitions. For entry-level adoption, see SENAR Core.

Whether a requirement belongs to the Foundation set is settled by Section 11, not by this notation. A requirement Section 11.1 does not include — QG-1 and QG-3, Rule 10.10, Federation Sync, dedicated role-holders — is not owed at Foundation at any level: it is absent there, not present as a SHOULD, and 13.1(c) does not call for it to be documented as an unimplemented SHOULD. The notation states the level of a requirement that applies; it does not extend the set of requirements that apply. A gate absent from Foundation's set does not carry its criteria there either — except where another clause requires those criteria in their own right, which 8.7 does for security-sensitive changes, and 8.4 states where they are then performed.

3.1 AI Agent

A software system powered by a large language model that generates engineering artifacts under human direction. An AI Agent is provided by an AI Model Provider (3.25) and operates at a specific model version.

AI Agents are not stable, deterministic tools. Model versions differ in capability, hallucination profiles, and instruction following behavior. Model version changes are treated as configuration changes (see Section 10.13 for normative requirements).

3.2 Supervisor

A human engineer who directs AI agents, verifies output, makes architectural decisions, and enforces Quality Gates. Production of artifacts is AI-directed; direct modification by the Supervisor is a recorded Gate Bypass, not an unremarked practice (Section 4.1, Section 8.6).

3.3 Supervisor+AI Pair

The fundamental production unit: one Supervisor working with one or more AI Agents.

3.4 Context

The information provided to an AI Agent to produce correct output: goal, acceptance criteria, constraints, knowledge, and traceability links.

3.5 Task

The atomic unit of tracked work. Has a goal, acceptance criteria, and requirement link.

3.5.1 Task Scope

The set of engineering artifacts (3.5.2) a Task is permitted to modify. Scope SHALL be recorded before work begins. The Task's scope is the set recorded before work began: a narrowing recorded afterwards does not place a modification already made outside it, because the term does not name the latest record but the one that preceded the work. Where scope is not recorded, the Task's scope is every artifact the Task in fact modified — a Task without a recorded scope does not thereby have a small one. The recording is executed as a required Task attribute (6.2) and as a criterion of QG-0 (8.1); this subsection fixes what is recorded and how the term is read when it is not.

3.5.2 Engineering Artifact

Any file whose content determines what the system does or how it is built, verified, deployed, or operated. This includes source code, tests, dependency manifests, build and pipeline definitions, infrastructure and deployment descriptors, database migrations, configuration that affects runtime behavior, Operational Scripts and agent instruction files (Section 5.3), and documentation that serves as context for an AI Agent (3.23).

An enumeration adopted by an organization does not narrow this term: a class of file omitted from such a list is within the term where the test below places it, and is therefore within a gate's declared effect. Where it is unclear whether a file is an engineering artifact, the test is whether a change to it can change what the system does, how it is verified, or how an agent behaves; if it can, it is one.

NOTE: The term appears in the effect declared by QG-0 (8.1) and therefore fixes the reach of that gate. It is defined here because a declared effect stated in undefined terms is settable by

the party it obliges, and 8.6(a) requires that the declaration determine, for any proposed action, whether the gate prevents it.

3.6 Exploration

An investigation without full Task formality. Explorations SHOULD be time-bounded (Section 6.1). If it yields implementation work, a Task is created.

3.7 Session

A time-bounded period of supervised AI work with defined start and end.

3.8 Increment

A scope-bounded batch of work with objectives and planned budget.

3.9 Quality Gate

An enforcement point that prevents a declared effect from occurring while its verdict is not positive. The criteria are gate-specific (Sections 8.1–8.5); the properties a check SHALL have in order to be a Quality Gate at all are normative in Section 8.6, and meeting the criteria without them is not sufficient. Gates SHALL be automated wherever feasible (8.6(f)); automation is a requirement on their operation, not a part of this definition — criteria requiring human judgment are provided for in 8.6(f) and appear in QG-2 and QG-4.

3.10 Knowledge Entry

A documented decision, pattern, known issue, or dead end stored in a searchable knowledge base.

3.11 Dead End

A documented failed approach with reason for abandonment. A dead end is any investigation that takes more than 15 minutes without producing a usable result. When this threshold is reached, the Supervisor stops, documents the approach and reason for failure, and chooses an alternative path. For normative requirements governing Dead End handling, see Section 10.4.

3.12 Checkpoint

A context preservation action during a Session to prevent work loss.

3.13 Gate Bypass

A documented exception allowing work past a Quality Gate. Requires justification, risk acknowledgment, remediation plan, and approval by a role senior to the requestor (8.6(h)). Admission of a gate's declared effect without a positive verdict is a bypass whether or not it was intended as one, and is recorded as one (8.6(h), 8.6(j)) — the record is what makes it the documented exception this term names, not a condition it satisfies beforehand.

3.14 Federation

Coordination mechanism for multiple Supervisor+AI Pairs across one or more projects: dependency tracking, shared knowledge, cross-project alerts. See Section 5.8 for federation requirements when managing multiple projects.

3.15 Cycle Time

Time from Task start to Task completion (`started_at → completed_at`). Distinguishes execution time from queue time (compare with Lead Time: `created_at → completed_at`).

3.16 Story

An intermediate grouping of Tasks representing a deliverable visible to stakeholders.

3.17 Requirement

A documented, verifiable statement of a need, capability, or constraint that the system must satisfy. SENAR defines three requirement levels: Business Requirement (3.18), System Requirement (3.19), and Task Requirement (3.20). A Task's goal and acceptance criteria constitute requirements at the TR level.

3.18 Business Requirement (BR)

A stakeholder-level need expressed in business terms. Source of all downstream requirements. Typically corresponds to an Increment objective or Epic goal.

3.19 System Requirement (SR)

A system-level capability or constraint derived from one or more Business Requirements. Expressed in terms of system behavior, not implementation. Typically corresponds to a Story goal.

3.20 Task Requirement (TR)

An implementation-level requirement decomposed from a Business or System Requirement. Corresponds to a Task's goal and acceptance criteria. The lowest level at which requirements are formally managed; test cases are verification artifacts derived from TRs, not a requirement level.

3.21 Requirement Hierarchy

The decomposition chain from business need to implementation unit: BR → SR → TR. Not all levels are required for all work; depth is determined by the Context Architect based on complexity and regulatory context (see Section 8.2).

3.22 Test Model (TM)

A verification artifact derived from Task Requirements that defines how each TR will be verified: test cases, test data, expected results, and verification method (automated test, manual demonstration, measurement). The Test Model is NOT a requirement level — it is the bridge between requirements and verification. In AI-native development, AI typically generates tests from TRs; the Supervisor verifies that generated tests actually exercise the stated acceptance criteria.

The level of Test Model formality scales by configuration — see Section 11 and QG-2 (Section 8.3) for normative requirements.

3.23 Code Documentation

Module-level, API-level, and architectural documentation that serves as persistent context for AI Agents. In AI-native development, code documentation has a dual audience: human Supervisors and AI Agents. Self-contained, machine-readable documentation reduces per-Task context overhead and improves AI output quality.

3.24 Traceability

The ability to trace every engineering artifact back to its originating requirement through a chain of linked references. Bidirectional: every TR traces up to a BR; every BR decomposes to at least one TR. Full traceability chain: BR → SR → TR → TM → Code.

3.25 AI Model Provider

An external service that provides AI model inference capabilities (e.g., cloud AI inference APIs, on-premise model servers). AI Model Providers are de facto suppliers — the AI model is the primary production tool, equivalent to a compiler. Model capabilities, limitations, and pricing change with provider decisions outside the organization's control.

3.26 AI Model Version

A specific release of an AI model, identified by provider designation. Version changes may affect output quality, cost, and behavioral characteristics. Different versions differ in capability, hallucination profiles, cost, and instruction-following behavior. Metric baselines (FPSR, cost/task) are version-dependent — see Section 10.13 for recalibration requirements.

3.27 Scope Creep

Unplanned changes to a Task's implementation that go beyond the stated goal and acceptance criteria. In AI-native development, scope creep manifests when AI agents modify code outside the defined scope boundaries, add unrequested features, or refactor existing code not covered by the Task.

3.28 Hallucination (AI)

AI-generated output that is plausible but factually incorrect: references to non-existent APIs, methods, CLI flags, or packages; fabricated file paths; confidently stated but wrong assertions

about system behavior. In dependency contexts, a hallucinated package is one that does not exist in the official package registry or that resolves to an unexpected maintainer.

3.29 WSJF (Weighted Shortest Job First)

A prioritization method calculating the ratio of Cost of Delay to Job Size. Used during Increment Planning (Section 7.1) to order the task pool. Adopted from SAFe without modification.

3.30 Value Stream

An end-to-end flow from stakeholder request to delivered, verified software. At Enterprise configuration, Increments are grouped by value stream with unified budgets (Section 11.3).

3.31 Adversarial Review

Independent review of AI-generated output by an agent that has no access to the generating agent's session context or reasoning. See Section 10.15, L3.

3.32 Agent Dispatch

Delegation of a Task or sub-task to a separate AI agent instance, typically operating in an isolated environment. See Section 5.7.

3.33 Agent Profile

A named configuration of scripts, permissions, and context that defines the capabilities and boundaries of an AI agent performing a specific function. See Section 5.2.

3.34 Structured Tool Protocol

A protocol enabling structured interaction between AI agents and platform services, providing self-describing tool schemas, atomic operations, and audit logging. See Section 5.6.

NOTE: Examples of qualifying protocols include Model Context Protocol (MCP), OpenAI function calling, and custom API interfaces.

3.35 Operational Script

A structured natural-language instruction that defines how an AI agent performs a specific action, containing trigger, preconditions, algorithm, postconditions, and outputs. See Section 5.3.

3.36 Adversarial Detection Rate (ADR)

Metric measuring the density of CRITICAL-severity findings discovered by adversarial review per task that underwent L3 review. Formula: `adversarial_critical_findings / L3_reviewed_tasks`. See Section 9.2.

3.37 Code Standards

A document defining mandatory code quality rules, loaded into AI agent context to guide implementation. Covers security, architecture, database, API, concurrency, and testing patterns. See Section 10.15, L2.

3.38 Latent Defect

AI-generated code that appears correct at surface level — passing automated checks, type validation, and self-review — but contains hidden defects (security bypasses, logic errors, architectural violations) detectable only through independent adversarial review. Latent defects arise when AI generates code by pattern matching rather than semantic understanding. See Section 10.15.

3.39 FPSR (First-Pass Success Rate)

The percentage of Tasks that meet all Acceptance Criteria on the first verification attempt, without requiring rework. See Section 9.1 for definition and measurement requirements.

3.40 Quality Sweep

A structured periodic audit of codebase and knowledge base — metrics, open defects, knowledge base coverage, architectural conformance and traceability completeness — conducted at a cadence the organization documents. It is not tied to the end of an Increment: 7.4 sets its frequency as organization-defined, and 8.6(g) reads that interval as the floor for

retaining gate records where the Increment Retrospective is not conducted (13.6). See Section 7.4 for normative requirements.

3.41 Flow Manager

A Supervisor role responsible for cross-team coordination, dependency tracking, and flow efficiency within a Federation. See Section 4.4 for role definition and responsibilities.

3.42 Context Architect

A Supervisor role responsible for designing and maintaining the knowledge architecture, context loading strategies, and documentation standards that enable effective AI-directed work. See Section 4.2 for role definition and responsibilities.

3.43 Knowledge Engineer

A Supervisor role responsible for capturing, structuring, and maintaining the organization's Knowledge Entries, dead ends, and reusable patterns. See Section 4.3 for role definition and responsibilities.

3.44 Verification Engineer

A Supervisor role responsible for designing the Test Model, executing adversarial reviews, and enforcing Quality Gates. See Section 4.5 for role definition and responsibilities.

3.45 Handoff

A record written at the end of a Session that carries to the next Session what it needs in order to continue: the state the work is in, what comes next, and the warnings that would otherwise have to be rediscovered. AI agents hold no memory between Sessions, so the Handoff is the only route by which the context of one Session reaches the next. It is a required output of Session End (7.3) and is required of every Session by 6.4(c). See Section 7.3 for normative requirements.

3.46 Behavioral Contract

The project rules file that states an AI Agent's boundaries, prohibitions and conventions: what it may change, what it may not, and the conventions its output follows. It is the first of the

three instrumentation levels of Section 5.1, is version-controlled and governed as a production process change (5.1, 10.14), and is required to be portable across agent implementations (5.9). It differs from an Operational Script (3.35), which states how a particular action is performed, and from the Programmatic Interface (5.4), which states what the agent can reach.

4. Roles

SENAR defines five responsibility sets. One person MAY cover multiple sets. Responsibilities SHALL be covered, not necessarily by dedicated positions.

For career paths, transition guides, and scaling ratios, see SENAR Guide and Reference.

4.1 Supervisor (SHALL — all configurations)

Directs AI agents, verifies output, makes architectural decisions, enforces Quality Gates.

Responsibilities: a) Define Tasks with goals and acceptance criteria before work; b) Provide structured context to the AI agent; c) Direct AI through execution with course corrections; d) Review all AI output against acceptance criteria; e) Make architectural and trade-off decisions; f) Enforce Quality Gates before closing Tasks; g) Capture knowledge entries during work; h) Maintain Session discipline (checkpoints, duration).

The production of engineering artifacts is AI-directed. A Supervisor who modifies an artifact directly supplies content by a route no QG-0 verdict covers — the verdict admits AI-directed production, and 8.6(j) states that the route is the origin of the content, not the identity of the process that writes the file. That is a Gate Bypass under 8.6(h) whether or not it was intended as one, and a positive QG-0 verdict for the Task does not cover it. Such a modification SHALL be recorded as a Gate Bypass, with the justification, risk acknowledgment, remediation plan, and senior approval that a Gate Bypass requires (Section 3.13).

Direct modification remains available and is expected to occur. As of this revision one case is recognized as substantive: an environment in which no agent can be run and there is nothing to switch to. Everything else the earlier list named — an agent stuck on a wrong approach, an agent that got most of the way, a fix cheaper than preparing context, a hotfix against a deadline — is addressable by agent means where an agent environment exists: a different model, a deeper pass, a second attempt. Where the environment is unavailable, those cases are the recognized one and not additional ones. What is not available in any case is doing it silently: the record is required for the recognized case as much as for any other, or metric 8 loses the population it is computed over (9.2). Cost of context preparation is a legitimate justification to record; it is not a reason to omit the record.

NOTE: The list above is dated. It states what agent capability makes addressable at the date of this revision, and the boundary moves as models change — which is why 10.13 requires re-evaluation of capability-dependent provisions when the model generation changes, and 0

marks such provisions as capability-dependent. A list of exceptions that is not re-read on that trigger describes the previous generation of tools within a year. Whether a case falls under the recognized one is decided by the senior who approves the Gate Bypass under 8.6(h); this Standard introduces no separate authority for it, and the single case is physical rather than organizational, so it does not scale by configuration.

4.2 Context Architect (SHALL — all configurations)

Designs requirements as structured AI input, manages the requirement hierarchy (BR → SR → TR), and ensures traceability. Quality is built at input: a defect in a Business Requirement cascades to all downstream System Requirements, Task Requirements, and ultimately to code. The Context Architect is the primary guardian of input quality.

Foundation (combined with Supervisor — Section 4.8, Section 11.1): At Foundation, the Supervisor absorbs Context Architect responsibilities. Minimum Foundation-level duties: a) Maintain the project context file (e.g., CLAUDE.md) and ensure it reflects current project state; b) Ensure context quality before task start — verify that QG-0 inputs are sufficient and well-structured; c) Design initial task decomposition (BR → TR at minimum).

Team (SHALL — dedicated): a) Maintain the requirement hierarchy (BR → SR → TR) with bidirectional traceability; b) Manage requirement lifecycle: draft → approved → verified → deprecated; c) Ensure requirement quality properties: verifiability (SHALL), consistency, sufficiency, non-redundancy, traceability [Team+: SHALL]; d) Prioritize work using WSJF (Section 3.29); e) Conduct periodic traceability audits to identify orphaned requirements (requirements without Tasks) and orphaned Tasks (Tasks without requirements); f) Enable requirement reuse by maintaining verified requirements as reusable Knowledge Base entries; g) Design Agent Profiles and manage Operational Scripts (Section 5); h) Review script changes as production process changes (Section 10.14); i) Maintain tool inventory per Agent Profile (Section 5.4).

4.3 Knowledge Engineer (SHALL Team+)

Captures, curates, and maintains the organizational knowledge base.

NOTE: At Foundation configuration, Knowledge Engineer responsibilities are combined with the Verification Engineer role and covered by one person (Section 4.8, Section 11.1). The "Team+" designation means a dedicated role; at Foundation, combined coverage applies. From Team, 4.8 sets whether the Knowledge Engineer is dedicated or shared across Pairs; from 5 Pairs the role is dedicated.

Responsibilities: a) Capture Dead Ends, decisions, patterns, known issues; b) Ensure entries are searchable and categorized; c) Review knowledge freshness; deprecate outdated entries; d) Analyze knowledge gaps.

4.4 Flow Manager (SHALL Team+)

Manages Session rhythm, cost tracking, and flow metrics.

NOTE: At Foundation configuration, Flow Manager responsibilities are absorbed by the Supervisor (Section 4.8, Section 11.1). The "Team+" designation means a dedicated role; at Foundation, combined coverage applies.

Responsibilities: a) Monitor Session duration and checkpoint cadence; b) Track cost per Task and Increment against budget; c) Monitor flow metrics (Section 9); d) Facilitate Increment Planning and Retrospective; e) Coordinate dependencies between Pairs (Federation); f) Schedule Quality Sweeps.

At Enterprise scale, expands to coordinate multiple Pairs (Federation Coordinator).

4.5 Verification Engineer (SHALL Team+)

Audits AI-generated output for correctness, security, and architectural conformance.

NOTE: At Foundation configuration, Verification Engineer responsibilities are combined with the Knowledge Engineer role and covered by one person (Section 4.8, Section 11.1). The "Team+" designation means a dedicated role; at Foundation, combined coverage applies.

Responsibilities: a) Conduct Quality Sweeps; b) Audit AI output using the AI Output Review Checklist (see SENAR Guide); c) Review AI-generated tests for coverage and assertion quality; d) Track Defect Escape Rate and identify systemic issues.

4.6 Enterprise Roles (Enterprise configuration only)

At Enterprise scale (10+ Pairs), three additional responsibility sets emerge:

Portfolio Manager: Coordinates multiple Increments across value streams. Manages unified budget, cross-stream dependencies, and strategic prioritization. Ensures AI investment aligns with organizational objectives.

Chief Supervisor: Sets architectural standards across all Pairs. Reviews and approves architectural exceptions. Defines organization-wide AI interaction patterns and context

templates. Note: Chief Supervisor defines organization-wide standards; Context Architect (Section 4.2) implements them within specific projects.

Federation Coordinator: Manages cross-project dependency tracking, shared knowledge base governance, and cross-team requirement traceability. Facilitates Federation Sync ceremonies (Section 7.5) at the portfolio level.

These are responsibility sets, not job titles. One person MAY cover multiple Enterprise roles.

4.7 Summary

Responsibility	Team	Enterprise
Supervisor	SHALL (each Pair)	SHALL (each Pair)
Context Architect	SHALL (covered; 4.8 sets dedication)	SHALL (dedicated)
Knowledge Engineer	SHALL (covered; 4.8 sets dedication)	SHALL (dedicated)
Flow Manager	SHALL (covered; 4.8 sets dedication)	SHALL (dedicated)
Verification Engineer	SHALL (covered; 4.8 sets dedication)	SHALL (dedicated)
Portfolio Manager	—	SHALL
Chief Supervisor	—	SHALL
Federation Coordinator	—	SHALL

NOTE: This table states Team and Enterprise. What Foundation owes is stated in 11.1, and the headcount that covers it in 4.8; neither is repeated here. For entry-level adoption — one Pair, which is what 11.4 gives Core — see SENAR Core, where the Supervisor role absorbs all responsibilities.

4.8 Role Combinations by Team Size

Team Size	Combination Pattern
1 Pair (Core)	Supervisor covers all responsibilities informally
1–3 Pairs (Foundation)	Supervisor + Context Architect + Flow Manager (combined), Knowledge Engineer + Verification Engineer (combined). 2 people cover all 5 responsibility sets.
3–5 Pairs (Team)	Context Architect + Flow Manager (combined), Knowledge Engineer (dedicated or shared), Verification Engineer (dedicated), Supervisors (each pair). Minimum 3 distinct role-holders.
5–10 Pairs (Team)	All 5 roles dedicated. Context Architect may still combine with Flow Manager at 5–6 pairs.
10+ Pairs (Enterprise)	All roles dedicated + Enterprise roles (Portfolio Manager, Chief Supervisor, Federation Coordinator).

NOTE: Role combinations are SHOULD-level guidance. Organizations MAY choose different combinations based on team skills and domain. The requirement is that all responsibilities are covered, not that specific combinations are used. This table is where headcount inside a configuration is set: the Responsibilities row of 11.2 states coverage and defers the number of distinct role-holders to the rows above.

5. Agent Instrumentation

5.1 Overview

AI-native development requires explicit management of the AI agent's behavior, capabilities, and constraints. This section defines how organizations SHALL configure, version, and govern the instruments that control AI agent behavior.

Agent instrumentation operates at three levels:

Level	Artifact	Purpose	Analogy
Behavioral Contract	Project rules file	Define boundaries, prohibitions, conventions	Job description
Operational Scripts	Procedural instructions	Algorithmic steps for specific actions	Machine work instructions
Programmatic Interface	API / Tools / Hooks	Instruments for agent-platform interaction	Machine control panel

Each level SHALL be version-controlled and subject to change management (see Section 10.14).

5.2 Agent Profiles

An Agent Profile defines a specific configuration of scripts, permissions, and context for an AI agent performing a particular function.

Organizations define at minimum the following profiles [Foundation: SHOULD; Team+: SHALL]:

Profile	Scripts	Access	Purpose
Generator	Implementation, commit, debug	Read/write code and tasks	Primary development
Reviewer	Review, security audit	Read-only code and tasks	Independent verification
Planner	Planning, retrospective	Write epics/stories, read all	Architecture, decomposition
Documenter	Documentation	Write knowledge base	Documentation and knowledge capture
Verifier	Quality sweep, testing	Read all, write findings	Quality audit

Separation of concerns: A Reviewer profile SHALL NOT have write access to the artifacts being reviewed. This separation is the foundation of adversarial review — the same agent cannot both generate and approve its own output.

Organizations MAY define additional profiles. One physical AI agent MAY switch between profiles within a session, provided each profile switch is logged. A switch does not confer independence: an agent that produced an artifact does not become an independent reviewer of it by entering the Reviewer profile, because it retains the context the artifact was produced in. Such a switch SHALL NOT be relied on to satisfy 10.15(a), and a review it yields is L2 — a verification statement by the party that did the work — and SHALL NOT be recorded as L3.

NOTE: The level is stated in the requirement above, not here — a NOTE is informative and cannot scale a normative requirement. Foundation teams typically combine Generator and Reviewer into fewer profiles matching their combined role structure (Section 4.8, Section 11.1). The separation of concerns paragraph is not scaled by configuration: wherever a Reviewer profile exists, it is read-only.

Team configuration: Organizations SHALL define all five profiles with enforced permission boundaries. Enterprise configuration: Organizations SHALL define all five profiles with enforced permission boundaries and audit trail for profile switches.

5.3 Operational Scripts

An Operational Script is a structured natural-language instruction that defines how an AI agent performs a specific action. Scripts are the primary mechanism for encoding organizational process into agent behavior.

5.3.1 Script Structure

Each script SHOULD contain:

- **Trigger**: When the script is invoked (command, event, condition)
- **Preconditions**: What must be true before execution
- **Algorithm**: Numbered steps with decision points
- **Postconditions**: What must be true after execution
- **Outputs**: What the script produces (artifacts, state changes, reports)

5.3.2 Script Governance

Scripts define agent behavior. Changing a script changes the production process. For the operational rule governing script changes, see Section 10.14.

SHALL (all configurations):

- Operational scripts stored in version control system
- Script changes reviewed before deployment (as code changes)
- Decision to change a script recorded in knowledge base with rationale

SHALL (Team+ configurations):

- Script changes tested on an isolated project before propagation to all projects
- Active script registry maintained with version identifiers
- Rollback capability: any script change can be reverted to previous version
- Script change audit trail maintained

SHOULD:

- Scripts have acceptance criteria (what the script does, does not do, edge cases)
- Script effectiveness tracked via metrics (e.g., FPSR of tasks executed under a given script)

5.4 Programmatic Interface

The Programmatic Interface defines what tools, APIs, and automated actions are available to the AI agent.

5.4.1 Tool Inventory

Organizations SHALL maintain an inventory of tools available to each Agent Profile:

- Which API endpoints are accessible

- Which file system operations are permitted
- Which external services are reachable
- Which actions require human confirmation

5.4.2 Principle of Least Privilege

Each Agent Profile SHALL have access only to the tools required for its function:

- A Reviewer SHALL NOT have write access to production code
- A Generator SHALL NOT have access to deployment credentials
- A Verifier SHALL NOT have access to modify quality gate results

5.4.3 Automated Actions (Hooks)

Organizations MAY define automated actions triggered by events:

- Post-session: metrics collection, knowledge sync
- Post-task-completion: quality gate validation, notification
- Pre-commit: lint check, security scan

Hooks SHALL be version-controlled alongside scripts.

5.5 Security Boundaries

5.5.1 Prompt Injection Defense

Organizations SHALL ensure that AI agents do not process untrusted user-supplied content as instructions. When agents read external data (user input, file contents, API responses), that data SHALL be treated as data, not as commands. Operational Scripts SHOULD include explicit instruction boundaries.

Practical measures:

a) Agent prompts SHALL clearly delimit system instructions from user-supplied data; b) External data ingested by agents (API responses, file contents, database records) SHALL NOT be interpreted as agent commands or modifications to the agent's behavioral contract; c) Organizations SHOULD test agent configurations against known prompt injection patterns as part of Quality Sweeps (Section 7.4).

5.6 Structured Tool Protocol

Organizations expose platform capabilities through a structured, self-describing tool protocol rather than CLI commands or direct database access [Team+: SHALL]. The level follows the Tooling row of 11.4, which is where this standard settles what tooling a configuration owes; at Core, where that row states None, this standard levies nothing (13.3). Stating this requirement unconditionally, as revisions before 1.5 did, would have put it at SHALL for a Foundation organization, whose tooling that row places at Recommended, and levied it on Core, where the same row states None. The chosen protocol SHOULD provide:

a) Self-describing schemas — the agent receives tool definitions with parameter types and descriptions, reducing hallucinated arguments; b) Atomic operations — each tool invocation is a single transaction, avoiding the "chained command" failure mode; c) Structured input/output — eliminating parsing errors from text-based CLI output; d) Audit logging — every tool invocation is recorded with parameters and results.

NOTE: Examples of qualifying protocols include Model Context Protocol (MCP), OpenAI function calling, and custom REST/gRPC tool APIs.

CLI as fallback: CLI commands SHOULD remain available when the structured tool protocol is unavailable. Where a structured tool protocol is operated, organizations SHALL document which operations have CLI fallback and which are protocol-only. Where none is operated the distinction has no subject, and this Standard levies nothing.

Direct database access: AI agents SHALL NOT use direct database access for write operations. Read operations MAY use direct access for debugging, but production workflows SHALL use protocol or CLI abstractions.

5.7 Agent Dispatch and Execution Isolation

Agent dispatch — delegating a Task or sub-task to a separate AI agent instance — introduces unique risks:

a) The dispatched agent operates with reduced context (no session history, limited knowledge); b) Multiple dispatched agents may modify the same files concurrently; c) The Supervisor cannot observe the dispatched agent's reasoning in real-time.

Organizations using agent dispatch are subject to the following. The level differs between them and is stated in each item, not in this line:

1. **Isolate**: dispatched agents SHOULD work in isolated working copies to prevent file conflicts; NOTE: Examples of isolation mechanisms include version control worktrees,

separate repository checkouts, containerized build environments, and ephemeral cloud workspaces.

2. **Scope**: dispatch prompts SHALL include explicit boundaries — which files to modify, which to leave unchanged;
3. **Review**: all dispatched agent output SHALL undergo L3 Adversarial Review (Section 10.15) — agent dispatch is the highest-risk scenario for latent defects. At Foundation, and only for changes that are not High risk (8.7), L2 Review with the High-tier checklist MAY substitute where independent agent access is unavailable; the organization SHALL record the substitution and the ground for it. The substitution does not reach High-risk changes: for those, 10.15 makes L3 SHALL regardless of configuration level and requires that at least one reviewer be a human Supervisor. Nor does it make L2 a lesser L3 — L2 is written by the party that did the work (10.15), so the substitution replaces independent review with a self-report and is bounded for that reason;
4. **Context**: dispatch prompts SHOULD include relevant Code Standards, architectural constraints, and known issues from the knowledge base;
5. **Limit**: organizations SHALL define a maximum parallel dispatch count per Supervisor (Section 10.7 applies).

Dispatched agents SHALL NOT:

- Modify files outside the scoped boundary without explicit approval;
- Commit directly to shared branches;
- Access other agents' isolated working copies or session state.

Pattern: Supervisor dispatches → agent works in isolated copy → agent returns result → review under clause 3 above by a party that did not direct the work → Supervisor merges to main branch.

NOTE: The Supervisor who directed the dispatch is not an eligible L3 reviewer of its output — 10.15(a) disqualifies the party that directed the work under review by name. The review step is therefore performed by another agent instance or another person. Where the Foundation substitution in clause 3 applies, the step is L2 with the High-tier checklist and is recorded as L2.

5.8 Federation — Scaling SENAR Across Projects

When an organization manages multiple projects (a "federation"), SENAR practices scale with additional coordination requirements.

5.8.1 Project Independence

Each project in a federation SHALL maintain its own:

- Task tracker and session history
- Knowledge base (patterns, known issues, dead ends)
- Metric baselines and targets
- Agent configuration (profiles, scripts, permissions)

5.8.2 Cross-Project Coordination

A federation SHOULD designate a coordination project (analogous to a SAFe Release Train Engineer) responsible for:

- Cross-project dependency tracking
- Federated knowledge routing (a pattern discovered in Project A that affects Project B)
- Aggregate metrics (federation-level throughput, cost, quality)
- Shared Code Standards and review criteria

5.8.3 Knowledge Routing

Knowledge entries SHALL be scoped:

- **Project-specific**: patterns, known issues, and decisions relevant only to one project — stored in that project's knowledge base;
- **Cross-project**: patterns affecting multiple projects (e.g., API contract changes, shared library updates) — stored in the coordination project and routed to affected projects;
- **Global**: methodology-level insights — stored in the coordination project, available to all.

Organizations SHALL define routing rules: which knowledge types are automatically shared vs. manually promoted.

Cross-project knowledge entries SHALL require approval from the receiving project's Supervisor before entering that project's active context. Global knowledge entries SHALL require coordination project Supervisor approval. Knowledge entries that reference security-sensitive topics (authentication, authorization, encryption, secrets, CORS, CSRF, permissions) SHALL be flagged for human review regardless of routing rules.

5.8.4 Federated Metrics

Federation-level metrics aggregate project metrics:

Metric	Federation Computation
Throughput	Federation tasks done divided by federation sessions, computed from the summed counts. Not the sum of project throughputs: a sum of ratios is not a ratio of the same kind, and it equals the federation figure only where every project ran the same number of sessions
FPSR	Weighted average by project task count
DER	Weighted average by project completed-task count — the denominator DER is defined over (Section 9.1)
ADR	Weighted average by project L3-reviewed task count — the denominator ADR is defined over (Section 9.2), which is not the same population as agent task count
Cost	Sum of project costs

Each weighted average above is weighted by the denominator of the metric being averaged. Weighting by any other population yields a number that is not that metric, and a sum of per-project ratios is not a ratio of the same kind as its parts.

Organizations SHALL NOT compare raw metric values across projects with different stacks, team sizes, or maturity levels. Comparison SHOULD use normalized metrics (e.g., cost per story point by complexity).

5.9 Portability

This standard is not bound to any specific AI agent, tool, or platform.

Operational Scripts SHOULD be written in structured natural language interpretable by any AI agent of sufficient capability:

- Clear algorithmic steps (not platform-specific SDK calls)
- Explicit inputs and outputs per step
- Conditional logic expressed in natural language
- No assumptions about specific tool implementations

The Programmatic Interface MAY be platform-specific (structured tool protocols, function calling APIs, etc.), but the Behavioral Contract and Operational Scripts SHALL be portable across AI agent implementations.

When migrating between AI platforms, organizations SHALL:

- Verify that all Operational Scripts execute correctly on the new platform
- Recalibrate baseline metrics (Section 10.13: AI Model Governance)
- Update the Programmatic Interface layer while preserving Contract and Script layers

---

6. Units of Work

6.1 Exploration

Time-bounded investigation without full Task formality.

a) Explorations SHOULD be time-bounded (organization defines limit); b) If Exploration yields implementation, a Task SHALL be created before changes are committed; c) Explorations SHOULD be logged with a summary.

6.2 Task

The atomic unit of tracked work.

Required Attributes

Attribute	Description
Goal	What must be accomplished
Acceptance Criteria	Verifiable conditions defining done
Requirement Link	Link to parent Story, Business Requirement, System Requirement, or Task Requirement (SHALL)
Work Type	Functional category (development, architecture, QA, documentation)
Scope	The engineering artifacts the Task is permitted to modify (Task Scope, 3.5.1), recorded before work begins (SHALL). Where it is not recorded, 3.5.1 applies and the scope is every artifact the Task in fact modified

Lifecycle States

Transition	Gate
planning → active	QG-0
active → done	QG-2
done → merged (Team+)	QG-3
merged → released (Team+)	QG-4
active → blocked	None
blocked → active	None

At Team configuration and above, the `done` state leads to merge review (QG-3) and release approval (QG-4).

6.3 Story

An intermediate grouping of Tasks representing a deliverable visible to stakeholders. Links Tasks to requirements.

Required Attributes

Attribute	Description
Title	One-sentence description of the deliverable
Acceptance Criteria	Story-level verifiable conditions defining done (independent of individual Task AC)
Linked Tasks	References to all Tasks that compose this Story
Requirement Link	Link to parent Business Requirement (BR) or System Requirement (SR) (SHALL)

Completion Criteria

A Story is complete when all its Tasks have reached the `done` state AND the story-level Acceptance Criteria have been verified by the Supervisor.

6.4 Session

A time-bounded period of supervised AI work.

a) Organizations SHALL establish and document a maximum Session duration; b) Checkpoints SHALL be performed at intervals documented by the organization; c) Sessions SHALL begin with context loading and end with metrics capture and handoff; d) Ceremonies MAY be automated tooling commands; e) Version control discipline: commits SHALL be atomic; secrets detection SHALL be automated; changes SHALL be reviewed for scope creep (Section 10.6).

6.5 Increment

A scope-bounded batch of work with defined objectives, planned budget, and risk register.

a) Each Increment SHALL have 3–5 measurable objectives; b) Task pool SHOULD be prioritized (WSJF recommended); c) At Team and above, each Increment ends with a Retrospective (7.7). The Quality Sweep is not tied to the Increment boundary: it runs at the cadence 10.5 sets (3.40).

7. Ceremonies

Ceremonies handle human strategic decisions. Quality enforcement is handled by Gates (Section 8).

7.1 Increment Planning (SHALL at Team+)

Define objectives, task pool, budget, and risks for the upcoming Increment.

Participants	Context Architect (leads), Flow Manager, Supervisors
Frequency	Once per Increment

Outputs: objectives, prioritized tasks, planned budget, risk register.

7.2 Session Start (SHALL)

Load context, select tasks, verify environment.

Participants	Supervisor
Overhead	Minimal (MAY be automated)

7.3 Session End (SHALL)

Capture metrics, write handoff, record knowledge.

Participants	Supervisor
Overhead	Minimal (MAY be automated)

SHALL outputs: summary, task states, handoff (next steps, warnings), metrics, dead ends (if any). **SHOULD outputs:** knowledge entries for decisions/patterns/known issues.

7.4 Quality Sweep (SHALL)

Periodic comprehensive audit of codebase and knowledge base.

Participants	Verification Engineer (leads)
Frequency	At cadence documented by the organization

Scope: security, code quality, test health, configuration drift, knowledge freshness, architectural conformance, dependency health, AI-specific issues (duplication, scope creep accumulation, TODO debris).

7.5 Federation Sync (SHALL at Team+)

Coordinate multiple Supervisor+AI Pairs.

Participants	Flow Manager (leads), Supervisors
Frequency	Regular cadence (daily or every 2–3 days recommended)

Agenda: status per Pair, dependency updates, shared component changes, integration issues.

7.6 Delivery Review (SHOULD)

Demonstrate working software to stakeholders, collect feedback, obtain acceptance.

Participants	Context Architect (leads), Supervisor, Stakeholder
Frequency	Per deliverable milestone

7.7 Increment Retrospective (SHALL at Team+)

Quantitative review: planned vs actual cost, throughput trend, FPSR, DER, ADR, knowledge capture rate.

Participants	Flow Manager (leads)
Frequency	End of each Increment

Output: specific, measurable, time-bounded, assigned improvement actions.

Organizations SHOULD allocate time for innovation, learning, and AI tooling experimentation.

7.8 Summary

Ceremony	Mandate	Frequency
Increment Planning	SHALL (Team+)	Per Increment
Session Start	SHALL	Per Session
Session End	SHALL	Per Session
Quality Sweep	SHALL	Periodic (org-defined)
Federation Sync	SHALL (Team+)	Regular cadence
Delivery Review	SHOULD	Per milestone
Increment Retrospective	SHALL (Team+)	Per Increment

8. Quality Gates

Automated enforcement points. Implemented as code, not checklists. See SENAR Guide for AI Output Review Checklist.

8.1 QG-0: Context Gate (Task start)

Quality is built at input. QG-0 ensures that a Task has sufficient, well-structured context before work on it begins.

Prevented effect (8.6(a)): modification of engineering artifacts within the Task's scope, however that modification is produced.

NOTE: The effect is stated without qualifying what produces the modification. A gate defined against AI-directed work alone would leave the same artifact, changed the same way, outside its own declaration whenever the change arrived by another route — and 8.6(a) requires that, for any proposed action, it can be determined whether the gate prevents it. Scope is bounded by the Task: modifications outside the Task's scope are not this gate's subject.

Point of admission (8.6(a)): the opening of the Task to modification — the transition at which work within the Task's scope may begin. The effect is admitted once for the Task. A modification made under that verdict, within the scope the verdict was issued for and by the route the verdict admits, falls within the admission already recorded: it is not a further admission, and calls for no further verdict, record or digest comparison. The route the verdict admits is AI-directed production (4.1); content that reached the artifact by another route is governed by 8.6(j) and is a Gate Bypass however wide the recorded scope was. Where the Task's scope is changed after the verdict, the verdict stands for a state that no longer holds, and 8.6(c) requires a new verdict before work continues under the changed scope.

QG-0's verdict is about the Task record it examined — goal, acceptance criteria, requirement link, work type and recorded scope (3.5.1) — and about the content of no engineering artifact. The digest required by 8.6(d) is measured over that record. QG-0 admits the modification of artifacts; it does not examine them.

SHALL criteria: goal defined, acceptance criteria verifiable, requirement or story linked (SHALL), work type assigned, Task Scope recorded (3.5.1) — the artifacts the Task is permitted to modify, recorded before work begins.

SHALL criteria (Team+): acceptance criteria are independently verifiable (each can be tested or measured separately); at least one negative scenario (error case, invalid input, or boundary condition) is included in acceptance criteria.

SHALL criteria (all configurations, security-relevant tasks): security surface declared — tasks touching auth, user input, data storage, payment, or external APIs SHALL identify the threat surface and include at least one security-related acceptance criterion. See SENAR Core (Start Gate, criterion 5) and Section 8.7.

SHOULD: plan, complexity estimate, relevant knowledge identified.

NOTE: Task Scope is a SHALL criterion at every configuration, listed above, and Core requires it as well (Start Gate item 4). Foundation additionally inherits Core's negative scenario requirement (Start Gate item 3) even though it is listed as a "Team+" SHALL criterion above. At Foundation, QG-0 = Core Start Gate checks + Standard QG-0 structure. The "Team+" criteria (independently verifiable AC, negative scenario) are already practiced at Core level; Foundation formalizes them as gate criteria.

8.2 QG-1: Requirements Gate (Story/Increment level — Team+ configs)

Ensures requirements are defined, approved, decomposed, and of sufficient quality before implementation begins.

Prevented effect (8.6(a)): creation of implementation Tasks from requirements that are not approved, not decomposed to sufficient depth, or not verifiable.

SHALL criteria [Team+]: Business Requirement (BR) exists and is approved; decomposition to Task Requirements (TR) is complete at appropriate depth; all TRs have verifiable acceptance criteria; no orphaned requirements (requirements without implementation Tasks).

Further SHALL criteria, at Team and above, which is where 11.1 places this gate: requirements satisfy the quality properties stated below; non-functional requirements specified where applicable (performance, security, accessibility).

Decomposition Depth

The Context Architect SHALL determine decomposition depth [Team+] based on complexity and regulatory context:

Context	Required Depth	Example
Standard feature, Team config	BR → SR → TR (3 levels)	BR: "Support OAuth" → SR: "OAuth provider flow" → TR: Task AC
Complex/regulated, Enterprise config	BR → SR → TR → TM (formal Test Model)	Full traceability chain for audit compliance

NOTE: For simple features (BR → TR, 2 levels), see SENAR Core.

Requirement Quality Properties

Requirements submitted to QG-1 SHALL be verifiable (each can be tested, measured, or demonstrated) and SHALL satisfy:

a) **Consistency** — no contradictions with existing requirements at the same or higher level; b) **Sufficiency** — the set of TRs covers the parent requirement (SR or BR) completely; c) **Non-redundancy** — no duplicate requirements across Stories; d) **Traceability** — every TR traces up to a BR; every BR decomposes to at least one TR.

NOTE: These are criteria of QG-1, and 11.1 places QG-1 at Team and above. Below Team there is no level of them to state, and until 1.5 this subsection stated one — SHOULD, scaling to SHALL at Team+ — for configurations that do not operate this gate at all. Removing that level changes no obligation: at Team and Enterprise the notation already made them SHALL, and below Team the gate is absent. Verifiability is the exception, and it is owed everywhere — but not through this subsection. Verifiable acceptance criteria are required by 8.1 at QG-0 at every configuration, and SENAR Core requires the same at its Start Gate.

Requirement Change Management

NOTE: Requirement Change Management applies at Team+ configuration.

When an approved BR or SR changes after implementation has begun:

a) The change SHALL be documented with rationale and the identity of the person who authorized it; b) Impact analysis SHALL identify all downstream requirements and Tasks affected by the change; c) Affected Tasks that are already `done` SHALL be flagged for re-verification — their QG-2 approval is invalidated; d) The changed requirement SHALL pass QG-1 again before new implementation begins; e) Supervisors of affected Tasks SHALL be notified of the change.

Scaling: at Team, impact analysis SHALL be supported by tooling (requirement parent/child links). At Enterprise, requirement changes SHALL be version-controlled and subject to change review (see Section 11.3, Requirements-as-Code).

8.3 QG-2: Implementation Gate (Task done)

Prevented effect (8.6(a)): closure of the Task and propagation of its result — the point at which the result becomes available to other work as a completed one.

SHALL criteria: CI passes, tests pass, static analysis passes (including type checking where applicable), no new lint violations, acceptance criteria verified by Supervisor, no security vulnerabilities detected by scanning tools, and for any defect found and fixed within the Task the root cause is identified and recorded — the same term SENAR Core uses in its Rule 7 and Done Gate criterion 4.

Where a configuration does not operate QG-3, the AI Output Review Minimum Criteria of 8.4 are performed at this gate whenever 8.7 requires them — at Foundation that is every security-sensitive change, and 8.4 states it from the other side.

SHOULD: coverage meets threshold, no duplication, docs updated, knowledge entry if decision/dead end. Team+: generated tests reviewed against Test Model (TM) — AI-generated tests SHALL exercise the stated acceptance criteria, not just achieve coverage.

NOTE: Foundation inherits from SENAR Core Rule 4 the substance of the Team+ criterion above — that tests verify the requirement and not the implementation — even though review against a Test Model is listed as Team+. The Test Model is a Team+ artifact and its formality scales by configuration (3.22); what does not scale is that a test mirroring the implementation verifies nothing. This is the same reading 8.1 states for the negative scenario criterion, and it is stated for the same reason: a Foundation organization carries the Core rules (11.1), and a criterion here that appears to begin at Team would otherwise read as an obligation removed on the way up.

Evidence and the state it is about

The verdict SHALL be a statement about the Task's artifacts as they stand at the moment the Task transitions to done (8.6(c)), and SHALL rest on a digest measured over those artifacts (8.6(d)). A verdict earned on one state is not a verdict on a later one: where the artifacts change after the checks ran, either the checks are re-earned against the changed state or the verdict is negative.

What the evidence proves, and what it does not

Bound evidence proves two things, and organizations SHALL NOT present a passing QG-2 as proof of more:

a) that the named set of checks ran against this state; b) that the verdict was not carried over from a different state.

It does not prove the good faith of a party authorized to record approval. A participant holding that authority who intends to mislead can select checks that pass, or record approval of work they did not examine, and the binding will be intact throughout — the evidence will correspond to the state exactly as required, and will still be worthless. The binding is a control against evidence drifting from what it describes: substitution, staleness, replay. It is not an attestation against a purposeful insider.

Where assurance against a purposeful participant is required, it comes from review by a party that does not hold the generating role (Section 10.15, L3) and from separation of authority between profiles (Section 5.2) — not from the binding, and not from adding further checks to this gate.

SHOULD (Team+: SHALL): AI-modified dependency manifests (the project's declared dependency files — e.g., package.json, requirements.txt, go.mod, Cargo.toml) reviewed for hallucinated packages — packages that do not exist in the official registry or that resolve to unexpected maintainers, typosquatted packages (near-name-matches of legitimate packages), and dependency confusion (internal vs. public namespace collisions).

8.4 QG-3: Verification Gate (Merge — Team+ configs)

Prevented effect (8.6(a)): integration of the change into the shared line of development.

SHALL criteria: acceptance tests pass, security scan clean, no regressions, code reviewed (risk-based — see 8.7).

SHOULD: performance within SLA, accessibility compliance, cross-service integration verified.

AI Output Review Minimum Criteria

At QG-3, the following AI-specific checks SHALL be performed:

a) Generated code does not call non-existent APIs, methods, or CLI flags (hallucination check); b) All imported packages exist in the project's dependency manifest (dependency validation); c) Generated code follows project architectural patterns and conventions (pattern conformance); d) Generated tests exercise the stated acceptance criteria, not just achieve coverage (test validity); e) No hardcoded credentials, API keys, or secrets in generated code (secrets check).

These criteria replace dependency on the informative AI Output Review Checklist in the SENAR Guide with normative minimum requirements. Where 8.7 requires them for a security-sensitive change at a configuration that does not operate QG-3, they are performed and recorded at the gate that configuration does operate — at Foundation, QG-2 (8.3). The criteria are owed there; the gate is not, and 11.1 is not altered by this.

8.5 QG-4: Acceptance Gate (Release — Team+ configs)

Prevented effect (8.6(a)): release — delivery of the increment to its users.

The qualifier states what Section 11 settles: Section 11.1 gives Foundation two gates, QG-0 and QG-2, and Section 11.2 gives Team and Enterprise all five. It is stated here for the same reason it is stated in 8.4 — a reader of this subsection would otherwise take the criteria below as owed at every configuration.

SHALL criteria: stakeholder acceptance recorded — obtained at a Delivery Review (7.6) or recorded outside one, since 7.6 is SHOULD and the acceptance is not; QG-3 still passing; staging environment (production-equivalent deployment target) verified.

8.6 Gate Properties

Sections 8.1–8.5 state what each gate checks. This section states what a check SHALL be in order to be a Quality Gate.

Criteria alone do not determine where a check acts, what its verdict is a statement about, or what happens when no verdict can be reached. Two implementations can satisfy every criterion in 8.1–8.5 and still prevent different things — or prevent nothing. The criteria say what a particular gate examines; the properties below say what any gate has to be.

A check that satisfies the criteria of a gate in 8.1–8.5 but lacks a property required at the applicable configuration (see *Scaling by configuration*, below) is not a Quality Gate for the purposes of this standard and SHALL NOT be recorded as one.

Every Quality Gate SHALL have the following properties. Each is named and stated in the sentence that opens it, so the five names with their opening sentences lift out as a checklist of what to look for; the paragraphs that follow bound each property, and several of them carry further requirements of their own — a gate is assessed against the whole item, not against its first sentence alone.

a) **Declared prevented effect.** The gate SHALL declare the effect it prevents: the specific change to the system, or to the record of work, that does not occur while the gate's verdict is negative. The declaration SHALL be specific enough that, for any proposed action, it can be determined whether that action is one the gate prevents. "Ensures quality" and "checks the work" are not declared effects, because no proposed action can be tested against them.

The declaration fixes the **point of admission**: the point at which the gate ceases to prevent the declared effect. Where the effect is a single act, the point of admission is that act, and no further statement is needed. Where the effect is a class of acts that recurs within one unit of work, the declaration SHALL also state at what point that class is opened, because otherwise

every act in the class is separately an admission. Properties (c), (d) and (g) attach to the point of admission so fixed. QG-0 is the only gate of this Standard whose declared effect is of the second kind; 8.1 states its point of admission.

b) **Preventive placement.** The gate SHALL act before the effect declared under (a) — it SHALL be a preventive control and not a detective one. A check that establishes the same condition after the effect has occurred is a detective control. A detective control does not satisfy this property, and its presence does not reduce the requirement on the preventive one; it MAY be operated in addition, and where the preventive control depends on a component that can fail without signalling, operating a detective control as well is a SHOULD.

NOTE: Preventive and detective carry their established meaning in internal control. A preventive control removes the opportunity for the unwanted effect to occur; a detective control establishes, after the fact, that it has occurred. The distinction is one of position relative to the effect, not one of strictness — a detective control may examine more and still prevent nothing.

c) **Correspondence of the verdict to state.** The verdict SHALL be a statement about the state the gate examined, as that state stands at the point of admission. Where the gate examines the artifacts the declared effect acts on, that is the state of those artifacts; where it examines inputs that exist before the effect, it is the state of the records examined — 8.1 names them for QG-0. A verdict issued for one state SHALL NOT be applied to a different state.

This is a property of the verdict, not of the way the verdict is obtained. It is satisfied where evidence accompanies the proposed action, and equally where the gate derives the verdict itself from the state being admitted. It is not satisfied where the gate accepts evidence it cannot relate to that state — however the evidence was delivered, and however thorough the check that produced it.

d) **Binding to a measured digest.** The correspondence required by (c) SHALL rest on a digest measured over the artifacts the verdict is about. A recorded identifier of a state — a revision identifier, a version tag, a build number — is a binding only where the party relying on it has itself measured the digest of the named artifacts and compared it with the recorded value; an identifier required to be recorded but never checked against what it names is not a binding, and a gate that presents it as one does not conform. The requirement is on the measurement, not on the presence of the field.

The digest SHALL cover every artifact on which the declared effect would act, and the record SHALL name the set covered and the function used. A digest measured over a proper subset — one file, a manifest, a version stamp — is a binding to that subset and to nothing else; it is compatible with arbitrary change everywhere the effect also reaches, and a gate that presents it as binding the verdict does not conform.

This applies where the verdict is about the state of the artifacts the effect acts on. Where a gate's verdict is instead about inputs that exist before the effect, and about the content of no artifact the effect will reach, the artifacts the verdict is about are the records the gate examined, and the digest is measured over those. QG-0 is that case, and 8.1 names the record.

The measurement SHALL be of the state the checks were performed on, and the comparison SHALL be made at the point the effect is admitted. A digest computed at admission and compared against a value recorded in the same act establishes that the state equals itself. It is not a binding: nothing about the verdict is fixed by it, and the staleness that (c) exists to exclude passes through such a comparison untouched.

e) **Fail-closed.** Where the gate cannot reach a verdict — the check does not run, does not complete, or completes without a determinate result — the verdict SHALL be negative and the declared effect SHALL be prevented. Absence of a negative finding SHALL NOT be treated as a positive verdict. This applies to unavailability of the gate itself and to unavailability of anything the gate depends on.

A check whose failure is suppressed, so that a run which produced no result is reported as a determinate one, does not satisfy this property. The requirement is not on the shape of the report but on the fact: where a check a gate relies on did not produce a result, the verdict that follows SHALL be negative, whether or not the pipeline was arranged to always emit a positive one. A gate SHALL identify the checks it depends on; a check named in the gate's criteria (8.1–8.5) is one it depends on, and a dependency cannot be removed by declaring it advisory while continuing to present its criterion as met.

NOTE: Where a gate is detective — (b) is SHOULD at Foundation and may legitimately be unmet there — this property is not thereby waived, but what it requires is what a detective control can do: the verdict is negative, and no positive verdict is recorded or relied upon. The obligation to prevent falls on gates that are preventively placed; the obligation not to report a verdict the gate did not compute falls on every gate, and is reduced nowhere.

Scaling by configuration

The five properties differ in what they cost to implement, and the obligation level follows that cost and nothing else. Preserving the conformance of any existing implementation is not a basis for setting a level.

NOTE: The normative columns are Foundation, Team and Enterprise — the three configurations this Standard defines (Section 11). The Core column is **informative** and is retained because the cost argument below is only legible against the smallest scale; 13.3 states how it is to be read, and states it once.

Property	Core (informative)	Foundation	Team	Enterprise
a) Declared prevented effect	SHALL	SHALL	SHALL	SHALL
b) Preventive placement	SHOULD	SHOULD	SHALL	SHALL
c) Correspondence of the verdict to state	SHALL	SHALL	SHALL	SHALL
d) Binding to a measured digest	SHOULD	SHALL	SHALL	SHALL
e) Fail-closed	SHALL	SHALL	SHALL	SHALL

Basis for each level:

- **(a), (c), (e) — SHALL at every configuration.** These cost policy, not capability. Declaring what a gate prevents, refusing to apply a verdict to a state it was not issued for, and treating an unreached verdict as negative are decisions. They are executable with nothing but the text editor that Core presumes — Section 11.4 lists Tooling as None there — which is why SENAR Core states the same three in its own words, and why an organization that has not made them was not stopped by cost.
- **(d) — SHOULD at Core, SHALL from Foundation.** Measuring a digest costs one command plus the discipline to record the result alongside the verdict. That is a small capability, but not a free one. Section 11.4 lists Tooling as Recommended from Foundation.
- **(b) — SHOULD at Core and Foundation, SHALL from Team.** Acting before an effect requires the ability to interpose on it. That is a capability, not a decision. Section 11.4 lists Tooling as Required only from Team.

NOTE: Across the three configurations this Standard defines, only (b) varies: (a), (c), (d) and (e) are SHALL at Foundation, Team and Enterprise alike. An organization conforming at any configuration therefore owes four of the five properties in full from the first day, and the single question the scale poses is whether it can interpose on an effect.

Where a property is SHOULD at the applicable configuration and is not implemented, 13.1(c) applies and the organization SHALL document it with rationale. Where (b) specifically is not implemented, the gate is detective with respect to the effect it declares and SHOULD be recorded as such, so that a later claim at Team or Enterprise begins from a known position rather than an assumed one.

Enforcement

f) Gates SHALL be automated wherever feasible. Criteria requiring human judgment require explicit recorded approval. g) Every gate execution SHALL produce an audit record, including

an execution that reached no verdict. The record SHALL contain: the gate; the unit of work it concerns; the verdict, including its absence; the time it was issued; the state it was issued for, by the digest measured under (d), together with the set that digest covers and the function used; and the checks the gate depended on, with the result of each.

The admission of the declared effect SHALL likewise produce a record, naming the verdict relied on and the time. One such record is produced per admission as the gate fixes it under (a) — for QG-0, one per Task (8.1), not one per modification. Without it (b) and (c) cannot be assessed by anyone but the party asserting them: a verdict and an effect that are never recorded in relation to each other establish neither order nor correspondence, and 13.7 asks for exactly that relation.

Records SHALL be retained for a period the organization documents, which SHALL be no shorter than the interval between its conformance reassessments (13.6). Where the organization does not conduct the ceremony 13.6 ties that reassessment to — Foundation omits the Increment Retrospective (11.1) — the interval read here is the interval between its Quality Sweeps (7.4), a ceremony every configuration of this Standard conducts; 13.6 names the same fallback. An assessment can rest on evidence only for as long as the evidence exists. h) The effect declared under (a) SHALL NOT be admitted without a positive verdict, except through a documented Gate Bypass (justification + risk + remediation + senior approval). Admission of the declared effect without a positive verdict and without a recorded Gate Bypass is a bypass, whether or not it was intended as one. i) Organizations SHOULD track Gate Bypass rate.

j) Modification of an artifact within a Task's scope by a route the gate does not precede — including direct modification by a Supervisor — is admission of the effect declared by QG-0 (8.1) without a positive verdict, and is therefore a Gate Bypass under (h). It SHALL be recorded as one. This obligation rests on (a) and (h) and applies at every configuration of this Standard, Foundation included, where (b) is only SHOULD: a gate that is not preventively placed still declares an effect, and admitting that effect without a verdict is still an admission.

The route is the origin of the content, not the identity of the process that writes the file. Content composed by a person and transcribed by an agent — dictated, pasted into a prompt for verbatim writing, or otherwise supplied for reproduction — arrived by a route the gate does not precede, and (j) applies to it. A gate placed before an agent's work does not examine text the agent was told to reproduce.

Route is not placement. A gate that is detective with respect to its effect — (b) is only SHOULD at Foundation, and such a gate is legitimate there — still governs the routes by which work reaches its artifacts, and its being detective does not turn every modification into an admission without a verdict. Where QG-0 issued a positive verdict for the Task, and the modification arose from AI-directed work carried out under that Task within the scope that verdict was issued for, the modification falls within the admission that verdict opened. A wide

recorded scope widens what the verdict admits by route, not what it admits by origin: content a person composed is not brought inside the admission by the size of the scope. What this clause captures is content that reached an artifact by a route no verdict covered — not work whose gate happens to sit after the effect rather than before it.

Where an Exploration (3.6) yields artifacts that are kept, the Task created under 10.1 does not retroactively bring their production within its gate: the modification occurred before the Task existed, and QG-0 issued no verdict on it. Such artifacts SHALL be recorded under this clause when the Task is created, or reproduced under a Task whose QG-0 preceded them. An Exploration is exempt from needing a Task; it is not a route by which finished work enters the system unexamined.

NOTE: This is a requirement to record, not a prohibition. Modifying an artifact directly stays available, and 4.1 states which case is recognized as substantive at the date of this revision — an environment in which no agent can be run and there is nothing to switch to. That subsection is where the set is settled, and it is dated there. A prohibition that cannot be honored in those cases would produce silent violations in place of records, which is the opposite of what a gate is for. The record is what converts an unremarked practice into a measurable one.

8.7 Risk-Based Review

Risk Level	Examples	Review
High	Security, auth, payment, data migration, architecture	SHALL: peer review + security review (all configurations)
Standard	Feature, UI, business logic	QG-2 automated + Supervisor verification statement
Low	Docs, config, trivial fixes	QG-2 automated sufficient

Security review for high-risk changes: For changes affecting authentication, payment processing, personal data handling, or cryptographic operations, security review SHALL be performed at ALL configuration levels, not only Enterprise. This is a mandatory (SHALL) requirement regardless of team size. Organizations developing security-sensitive systems SHALL apply the AI Output Review Minimum Criteria of 8.4 and the risk-based review of 10.15 to such changes at every configuration, performed at the gate that configuration operates.

NOTE — SETTLED IN THIS REVISION. Until 1.5 this subsection required QG-3 at every configuration for security-sensitive systems, while Section 3, 8.4, 8.9, 8.10 and 11.1 marked QG-3 as Team+. The two statements are reconciled by separating the criteria from the gate. A

gate is a control with a declared prevented effect and a point of admission (8.6); its criteria are what it checks there. QG-3 the gate — whose prevented effect is integration into the shared line of development — remains Team+, and the number of gates a Foundation organization operates is unchanged at two (11.1). What is owed at every configuration is the substance: the five criteria of 8.4 and the review of 10.15, performed at the gate that configuration does operate, which at Foundation is QG-2. This follows the form 10.15 already uses, where L3 review is owed for High-risk changes regardless of configuration level without any gate being added to any configuration. The stricter reading the previous revision recommended is therefore now the requirement, and the gate count it would have changed is not changed.

8.8 Gate Pipeline

```
Exploration ──► (yields work?) ──► Task
Task ──► [QG-0] ──► active ──► [QG-2] ──► done
                                  |
                        (Team+) [QG-3] ──► merged
                                  |
                               [QG-4] ──► released

Story/Increment: (Team+) [QG-1] ──► approved ──► Tasks created
```

NOTE: This diagram is informative. It shows the order in which gates are reached, not where a gate is required to act. The normative requirement on placement is 8.6(b), read together with the prevented effect each gate declares in 8.1–8.5.

8.9 Summary

Gate	Applied At	Core	Foundation	Team	Enterprise
QG-0	Task start	YES (Start Gate)	YES	YES	YES
QG-1	Story/Increment	—	—	YES	YES
QG-2	Task done	YES (Done Gate)	YES	YES	YES
QG-3	Merge	—	—	YES	YES
QG-4	Release	—	—	YES	YES

NOTE: SENAR Core defines two quality gates and names them the **Start Gate** and the **Done Gate**; this Standard calls the same two QG-0 (Context Gate) and QG-2 (Implementation Gate), and the Core document states the same mapping at each of them. Where this Standard cites Core — as 8.1 does — it uses Core's names, because that is what they are called there. Foundation configuration inherits both gates from Core (Section 11.1).

NOTE: This table says which gates apply. What each of them is required to *be* — the five properties and the obligation level of each by configuration — is stated once, in 8.6, and is not repeated here. A scale printed twice is a scale that can disagree with itself, and this one had begun to. The Core column of this table is read as 13.3 states, which is where that is stated and the only place it is.

8.10 Security Requirements Cross-Reference

One consolidated view of all security-related requirements across the Standard. Use this table when auditing security coverage or onboarding a new configuration.

Requirement	Where Defined	Configuration
AI Output Review Minimum Criteria (secrets check, hallucination check)	8.4, performed at the gate the configuration operates (8.7)	All configs for security-sensitive changes; as QG-3 criteria, Team+
High-risk change review — human approval required (auth, payment, data, crypto)	8.7 + Section 10.15 L3	All configs
Security surface declaration at task start	QG-0 (8.1)	All configs
Auth coverage verification	Verification Checklist item 16 (SENAR Core, High tier)	High tier
Input validation	Verification Checklist item 6 (SENAR Core, Standard tier)	Standard tier
Secrets detection in sessions	Section 6.4	All configs
Agent dispatch review	Section 5.7 + Section 10.15	All configurations (10.15); 5.7(3) defines one bounded substitution at Foundation

NOTE: "All configs" means the requirement applies regardless of team size — including Foundation teams. See 8.7 for the normative statement on security review applicability.

---

9. Metrics

Organizations SHALL establish baselines by measuring for at least 3 Increments before setting targets. Targets are organization-specific, not universal.

9.1 Mandatory (SHALL)

#	Metric	Formula	Measures
1	**Throughput**	tasks_done / sessions	Delivery performance
2	**Lead Time**	completed_at - created_at	Delivery speed
3	**First-Pass Success Rate**	single_cycle_tasks (completed without rework) / total × 100%	Context quality
4	**Defect Escape Rate**	post_done_defects / done × 100%	Gate effectiveness

A post-done defect is a defect task explicitly linked to a completed parent task that introduced the defect. The link SHALL be recorded in the task tracker.

9.2 Recommended (SHOULD; SHALL in Team+)

#	Metric	Formula	Measures
5	**Knowledge Capture Rate**	entries / tasks	Organizational memory
6	**Cost Predictability**	actual_cost / planned_cost × 100%	Estimation accuracy
7	**Cost per Task**	total_cost / tasks (by complexity)	Efficiency
8	**Manual Intervention Rate**	tasks_with_recorded_bypass_under_8.6(j) / total × 100%	AI-first adherence
9	**Cycle Time**	completed_at - started_at	Execution speed (vs Lead Time which includes queue time)
10	**Adversarial Detection Rate**	adversarial_critical_findings / L3_reviewed_tasks	Latent defect density

Manual Intervention Rate is computed from recorded Gate Bypasses, not from self-assessment. Section 8.6(j) makes direct modification of an artifact a Gate Bypass that SHALL be recorded, so the boundary of "manual intervention" is set normatively and is no longer left to each organization to define.

Relationship to Gate Bypass rate: 8.6(i) recommends rather than requires that figure, so an organization may hold this metric without it, and this Standard does not fix how it is computed. What the definitions give is a containment of events, not of rates: a modification recorded under 8.6(j) is a Gate Bypass under 8.6(h), so every such record belongs to the population Gate Bypass rate is drawn from, while the reverse does not hold — that population also contains admissions of any other gate's declared effect. Whether the two rates stand in the same relation depends on the denominator the organization gives Gate Bypass rate, which 8.6(i) leaves to it: they are comparable where that denominator is Tasks, the unit this metric uses, and not otherwise. Figures on different denominators answer different questions and SHALL NOT be added together or presented as one nested inside the other.

NOTE: The record this metric counts is made by the party that acted. It is therefore a measure of what was recorded, and organizations SHALL NOT present it as a measure of what occurred — the general form of this is 9.4(b). Organizations that collect this metric SHALL set a target for it and document the basis (9.4(c)) — at Foundation, which 11.1 does not ask to

report it, the duty arises only where the organization reports it; a figure that crosses the target is escalated under 9.4(d). This standard does not fix the number. No measurement exists that would justify one figure for every organization, and a ceiling stated without such a measurement would be an arbitrary number carrying the force of a requirement. What the standard does require is that the ceiling exist, be reasoned, and not be moved in silence — which is what a normalizing bypass rate looks like from the outside.

An unrecorded modification does not appear in it, and nothing in this standard detects one automatically. Where assurance about the figure itself is required, it comes from comparing the modifications recorded in version control against the gate's admission records (8.6(g)): every modification within a Task's scope should fall under an admission the gate recorded, and one that does not is either an admission that was not recorded or a bypass that was not declared. Authorship does not settle it — 8.6(j) makes the route the origin of the content, not the identity of the process that wrote the file, so a comparison of who committed what is blind to precisely the case the clause was written for. This is an audit the organization performs, not a property the metric has.

Adversarial Detection Rate (ADR) measures the density of CRITICAL-severity findings discovered by adversarial review (Section 10.15, L3) per task that underwent L3 review. Target: organizations SHOULD aim for ADR < 0.5 (fewer than one CRITICAL finding per two reviewed tasks). An ADR of 0 indicates either excellent AI output quality or insufficient review rigor — organizations SHOULD distinguish between the two.

For ADR computation, "L3_reviewed_tasks" means tasks that underwent L3 Adversarial Review. Tasks that did not receive L3 review (e.g., Low-risk tasks where L3 was skipped per Section 10.15) are excluded from the denominator. This ensures ADR reflects review effectiveness, not review coverage.

Knowledge Capture Rate target calibration: organizations with established knowledge bases SHOULD set KCR targets that reflect diminishing returns. A target of 1.0 (one entry per task) is appropriate for greenfield projects. For mature projects (>500 tasks), a target of 0.33 (one entry per three tasks) better reflects the natural rate of novel knowledge discovery. Organizations SHALL document their KCR target rationale.

9.3 Collection

Metrics 1–5, 8 and 9 SHALL be collected automatically from task tracking and session data. Metric 8 is derived from recorded Gate Bypasses under 8.6(j) and requires no separate reporting act. Metrics 6–7 require cost tracking integration. Cost Predictability requires Flow Manager assessment at Increment Retrospective.

Metric 10 (ADR) SHALL be collected from adversarial review findings recorded during L3 review (Section 10.15). Organizations SHALL maintain a record of review findings per task, classified by severity, to compute ADR.

Metric computation SHALL include all tasks without epoch-based filtering or exclusion of historical data. Organizations SHALL NOT exclude tasks from metric computation based on creation date, migration status, or tooling version. Rationale: epoch filters introduce complexity, create maintenance burden, and mask data quality issues; they also make a figure unauditable, because the population it was computed over stops being the population the tracker holds. A cumulative figure does not shed its early data on its own: those tasks keep the weight their share of the total gives them, and that share falls only as the total grows — not because the tooling changed. Where the change of tooling is exactly what makes early data unrepresentative, waiting does not fix the figure. That is a reason to state the limitation rather than to filter it away. If early data is known to be unreliable (e.g., pre-automation manual entries), this SHALL be documented as a known limitation, not filtered.

Cost Predictability (metric 6) requires organizations to record planned cost before implementation begins. In practice, planned cost estimation for AI-assisted tasks is unreliable — AI execution time is non-deterministic and model pricing varies. At Team configuration, Cost Predictability SHALL be tracked but baselines may be provisional during the first 3 Increments. At Enterprise configuration, Cost Predictability SHALL be tracked with established baselines. Cost per Task (metric 7) provides a more actionable proxy for cost management in early adoption.

Multi-agent and multi-session configurations: when multiple AI agents work in parallel (multiple concurrent sessions), per-session metrics (throughput, session duration, cost per session) reflect individual agent performance, not aggregate team output. Organizations using multi-agent configurations SHOULD additionally track aggregate metrics at the Increment level. Token and cost attribution in multi-agent scenarios SHOULD be recorded per-agent, with aggregate totals available for Increment-level reporting.

9.4 Reporting a figure

A metric is a claim about what happened. These requirements apply to every metric of 9.1 and 9.2, and to any further figure an organization reports as a SENAR metric.

a) **The method is disclosed with the figure.** Organizations SHALL state how a reported figure is computed: the population it is drawn from, the record type it is counted over, and the period it covers. A figure whose definition is not disclosed is an assertion rather than a measurement, and SHALL NOT be presented as a SENAR metric.

b) **A figure computed from self-made records measures what was recorded.** Where the records a metric is counted over are made by the party being measured, the figure measures what was recorded and not what occurred, and organizations SHALL NOT present it as a measure of what occurred. First-Pass Success Rate, Defect Escape Rate and Manual Intervention Rate share this property: each improves when a record is not made, and nothing in this standard detects an unmade record automatically.

c) **A self-set target or threshold carries its basis.** Where this standard leaves a target, threshold or baseline to the organization, the organization SHALL document the basis on which it was set and the data that basis rests on. A number set without a basis cannot be assessed, and an assessor acting under 13.7 is left with nothing to ask for.

d) **A target that is crossed is escalated, not quietly moved.** Where a reported figure crosses a target set under (c), the organization SHALL record the crossing and the action taken, or the decision to change the target together with the basis for the change. Revising a target to accommodate the figure, without stating that this is what happened, is a defect of the record and not a change of practice.

NOTE: (b) is a property of the class of records, not of a particular metric, which is why 9.2 states it where Manual Intervention Rate is defined and does not argue it again here. Assurance about such a figure comes from comparing the records against an independent trace — 9.2 names that comparison for Manual Intervention Rate — and not from adding further checks to the metric.

---

10. Operational Rules

NOTE: This section defines each rule; it does not state which of them a configuration owes. That is settled by Section 11, and for Foundation by 11.1. Read alone, the SHALL of each subsection gives a Foundation reader fifteen obligations where 11.1 gives it the SENAR Core rules plus six of these — 10.2, 10.4 and 10.9 by name; 10.5 through the Quality Sweep ceremony; 10.6 through 6.4(e), which states its three duties without a configuration qualifier; and 10.15 at L1 and L2, to which 10.15 itself adds L3 for High-risk changes and for all dispatched agent output, both regardless of configuration. The set is not restated here: a set printed twice is a set that can disagree with itself, and this Standard has already paid for that once (8.6).

10.1 Task Before Implementation (SHALL)

Implementation SHALL NOT begin without a Task. Explorations are exempt; if they yield code, a Task SHALL be created before commit.

10.2 Session Duration (SHALL)

Organizations SHALL establish, document, and enforce a maximum Session duration. The basis on which the maximum is set is documented under 9.4(c), which states that duty once for every number this standard leaves to the organization. The documented maximum SHOULD be informed by empirical session data. As a starting guideline, sessions exceeding 180 minutes show diminishing returns in most AI-assisted workflows.

10.3 Checkpoint Cadence (SHALL)

Checkpoints SHALL be performed at intervals not exceeding a duration documented by the organization. The documented interval SHOULD be no greater than the organization's maximum acceptable context loss in the event of a session interruption.

10.4 Dead End Documentation (SHALL)

Failed approaches SHALL be documented as Dead End knowledge entries (approach, reason, alternative). A dead end is any investigation that takes more than 15 minutes without

producing a usable result. When this threshold is reached, the Supervisor SHALL stop, document the approach and reason for failure, and choose an alternative path.

10.5 Periodic Audit (SHALL)

Quality Sweeps SHALL be conducted at a cadence documented by the organization, reviewed at each Increment Retrospective or, where that ceremony is not conducted — Foundation omits it (11.1) — at each Quality Sweep itself. The documented cadence SHOULD be no less frequent than once per 3 Increments. For Foundation configuration, the recommended cadence is monthly (Section 11.1), which satisfies this requirement when Increments are 2 weeks or shorter.

10.6 Version Control (SHALL)

a) Commits SHALL be atomic; b) Secrets detection SHALL be automated; c) Changes SHALL be reviewed for scope creep.

NOTE: (c) does not qualify what produced the change. Qualifying it — "AI changes SHALL be reviewed" — would carve out changes of other origin from the requirement by implication, which is the reading of a mixed process this standard no longer makes. Scope creep is a property of the change, not of its author.

10.7 Parallel Agent Limit (SHALL)

Supervisors SHALL NOT exceed the organization's defined parallel agent limit. A common starting limit is 3 concurrent agents per Supervisor; organizations SHOULD calibrate based on task complexity and Supervisor capacity.

10.8 Complexity-Cost Calibration (SHALL)

Organizations SHALL maintain cost baselines per task complexity level. The basis of each baseline and the period its data covers are documented under 9.4(c). Until 1.5 this was the one self-set number in this section with no stated basis of any kind, which made it the one an assessor could not question.

10.9 Knowledge Capture (SHALL)

Organizations SHALL establish and track a Knowledge Capture Rate target. The rationale for that target is documented under Section 9.2, which also gives calibration guidance; the

obligation is stated there and is named here because a reader setting the target from this rule would otherwise not meet it.

10.10 Requirement Traceability (SHALL Team+)

Organizations SHALL maintain traceability from Business Requirements through System Requirements to Task Requirements. Every Task SHALL link to at least one requirement or Story. Requirements SHALL be stored in a versioned, searchable system.

10.11 Code Documentation as Context (SHALL)

Code documentation (module-level documentation (docstrings, doc comments, or equivalent), API descriptions, architecture docs) is not a post-hoc artifact — it is active context for AI Agents. Well-documented code reduces the context volume a Supervisor must provide per Task and enables AI to produce correct output with less explicit instruction.

Organizations SHALL maintain code documentation that includes: (a) module/package purpose statement, (b) public API contracts with parameter and return types documented, (c) architectural boundary description identifying dependencies and integration points. The documentation SHALL be machine-readable (structured comments, API description formats, or equivalent). Documentation SHOULD be updated as part of the same Task that changes the code; 8.3 states this as a SHOULD criterion of QG-2 ("docs updated").

In AI-native development, documentation has a dual audience: human Supervisors and AI Agents. Documentation that is useful only to humans (e.g., "see John for details") is useless for AI. Documentation SHALL be self-contained and machine-readable.

10.12 Context Hygiene (SHALL)

Supervisors SHALL NOT include production credentials, API keys, personally identifiable information (PII), or regulated data (financial, health) in AI Agent context unless the AI model provider has a current data processing agreement and the data classification permits it. Organizations SHALL maintain a context classification policy defining which data categories are prohibited in AI context. Placeholder or synthetic data SHOULD be used when AI assistance is needed for tasks involving sensitive domains.

10.13 AI Model Governance (SHALL)

AI model providers are external suppliers. The AI model is the primary production tool — equivalent to a compiler. Organizations SHALL:

a) Record the AI model identifier and version used for each Session; b) Recalibrate metric baselines (FPSR, cost per task, throughput) when the active model version changes substantially — a change in model generation constitutes a substantial change; c) At Team+: evaluate new model versions before adoption — compare FPSR and cost per task on a representative task set; d) At Enterprise: maintain a formal model approval process — new models SHALL be tested against acceptance criteria for the organization's domain before production use; e) Document known model-specific limitations and hallucination patterns as Knowledge Entries.

Organizations SHOULD review the capability-dependent provisions of this Standard and its Guide — those that reference AI behavior, such as hallucination detection heuristics, session duration guidance and parallel agent limits — when model generations change substantially, and record the outcome of that review. Until 1.5 this duty was stated only in notes, here and in Section 0, and a note states no requirement; a provision written against one generation of models and never re-read describes the previous one within a year.

10.14 Script Change Management (SHALL)

Organizations SHALL treat operational script changes as production process changes:

- Version-controlled (Section 10.6)
- Reviewed before deployment
- Decision recorded in knowledge base (Section 10.9 applies)

Team+ configurations SHALL additionally:

- Test script changes on isolated environment before propagation
- Maintain version registry of active scripts per project
- Ensure rollback capability for any script change

For the full definition of Operational Scripts and their governance, see Section 5.3.

10.15 AI Output Quality Verification (SHALL)

Code produced for a Task SHALL undergo quality verification before commit, whatever produced it. Three verification levels are defined:

NOTE: The requirement does not qualify the producer, for the reason given at 10.6(c) and at 8.1: a rule written against AI-generated code alone would leave the same file, changed the same way, unverified whenever the change arrived by another route. Where a level is defined in terms of the generating agent — L3 requires a reviewer without access to the generating agent's context — the requirement reads as independence from whoever or whatever produced the code, and for code a Supervisor wrote directly it means the Supervisor is not the reviewer.

Level	Method	Configs
L1: Automated	Static analysis — file size, function size, complexity metrics, security pattern checks, lint	All (SHALL)
L2: Verification Statement	The person or agent who did the work writes a structured verification statement confirming what was checked against acceptance criteria and Code Standards	All (SHALL)
L3: Adversarial Review	An independent agent (different model or context-free) reviews without access to the generating agent's reasoning	SHALL (Team+)

L3 Adversarial Review requirements: a) The reviewer SHALL NOT have access to the generating agent's session context or reasoning, and SHALL NOT be the party that directed the work under review. This applies to human and AI reviewers alike: a Supervisor who ran the generating session holds that context by having produced it, and satisfies the letter of "a human reviewer" while defeating the independence the level exists to obtain; b) At least one reviewer SHOULD be a "cold" reviewer — an agent with zero prior context of the task; c) Findings SHALL be classified by severity (CRITICAL, HIGH, MEDIUM) and recorded; d) CRITICAL findings SHALL block commit until resolved; e) Organizations SHALL track Adversarial Detection Rate (ADR) — see Section 9. f) The severity scale SHALL be documented: what CRITICAL, HIGH and MEDIUM cover in the organization’s context, and who assigns them. The assignment of CRITICAL SHALL be recorded with its reason. This scale is the same severity scale by which (c) requires findings to be classified. This classification is not a label: under (d) it blocks a commit, and under 9.2 it is the numerator of Adversarial Detection Rate, so a scale left undocumented makes both undecidable. Where an AI reviewer proposes the classification, the Supervisor who accepts the finding is accountable for it — and for a finding of an L3 review that Supervisor is the one (a) permits, not the party that directed the work under review. Severity classifies a FINDING. It is not the risk level of a change (8.7), which selects review depth, and it is not the checklist tier of SENAR Core, which selects how much of that checklist is run. The scales share words — High appears in all three, Critical in two of them and not in the risk levels of 8.7 — so an organization documenting a scale under this clause states which of the three it is defining.

L3 review SHALL be applied based on risk level (see Section 8.7):

- **High risk** (security, auth, payment, data migration, architecture): For High risk changes (Section 8.7), L3 review SHALL be applied regardless of configuration level; at least one reviewer SHALL be a human Supervisor, not an AI agent; security review SHALL be performed (see Section 8.7);
- **Standard risk** (feature, UI, business logic): L3 SHALL be applied (Team+);
- **Low risk** (docs, config, trivial fixes): L3 MAY be skipped with documented justification.

NOTE: L2 Verification Statement addresses quality defects (logic errors, style violations, acceptance criteria coverage) but does NOT constitute a security control. For security assurance, L3 is the minimum effective verification level. The L2 verification statement is a structured record confirming what was checked — it is not a review by an independent party. The person or agent who did the work writes the statement, confirming each acceptance criterion was verified and how.

NOTE: Latent defects in AI-generated code are characteristically difficult to detect because they satisfy automated quality checks while containing subtle logical, security, or architectural flaws. Common patterns include: returning True without validation, null equality comparison bypasses, trusting HTTP headers without verification, empty configuration values that silently disable security checks, and unreachable "safety" code that masks incomplete control flow understanding.

Agent dispatch (delegating implementation to a sub-agent) is the highest-risk scenario for latent defects, as the dispatched agent operates with reduced context. Organizations using agent dispatch SHALL apply L3 review to all dispatched agent output.

Exactly one substitution is defined against this requirement, in 5.7(3), and its bounds are the same in both places: it is available at Foundation only; it applies only where independent agent access is unavailable; it does not reach High-risk changes (8.7), for which the risk-level rule above makes L3 SHALL regardless of configuration; and the organization records the substitution together with the ground for it. Outside those bounds this requirement is unconditional at every configuration of this Standard. The bound matters because the substitute is not a weaker review but a different kind of act: L2 is written by the party that did the work (table above, and the NOTE that follows it), so where it stands in for L3 the independence L3 exists to obtain is not reduced — it is absent.

10.16 Summary

#	Rule	Team	Enterprise
1	Task Before Implementation	SHALL	SHALL
2	Session Duration	SHALL	SHALL
3	Checkpoint Cadence	SHALL	SHALL
4	Dead End Documentation (>15 min threshold)	SHALL	SHALL
5	Periodic Audit	SHALL	SHALL
6	Version Control	SHALL	SHALL
7	Parallel Agent Limit	SHALL	SHALL
8	Complexity-Cost Calibration	SHALL	SHALL
9	Knowledge Capture	SHALL	SHALL
10	Requirement Traceability	SHALL	SHALL
11	Code Documentation as Context	SHALL	SHALL
12	Context Hygiene	SHALL	SHALL
13	AI Model Governance	SHALL	SHALL
14	Script Change Management	SHALL	SHALL
15	AI Output Quality Verification	SHALL	SHALL

NOTE: The table lists Team and Enterprise because those configurations carry all fifteen. What Foundation carries is stated in 11.1 and is not repeated here, for the reason given at the head of this section. For entry-level adoption, see SENAR Core, which defines 8 foundational rules.

11. Configurations

NOTE: For individual adoption, see SENAR Core — 8 rules, 2 quality gates, 2 metrics. This Standard defines three organizational configurations: Foundation, Team, and Enterprise.

11.1 SENAR Foundation (1–3 Pairs)

The bridge between Core and Team. For small teams that need more structure than Core but aren't ready for full Team process.

Category	Requirements
Base	All SENAR Core rules apply
Responsibilities	All 5 responsibility sets of Section 4, covered by 2 people (Section 4.8): Supervisor + Context Architect + Flow Manager on one, Knowledge Engineer + Verification Engineer on the other
Ceremonies	3: Session Start, Session End, Quality Sweep (monthly is the cadence 10.5 recommends for Foundation; the requirement there is that the cadence be documented). NOTE: Foundation conducts Increments (Section 6.5) but omits formal Increment Planning and Retrospective ceremonies. Planning is implicit in Session Start; retrospective insights are captured in Quality Sweep. Section 7.8 carries the matching qualifier on both rows, so the omission is stated where a reader of Section 7 will meet it and not only here.
Quality Gates	2: QG-0 (Context Gate) + QG-2 (Implementation Gate)
Metrics	4: Throughput, Lead Time, FPSR, Defect Escape Rate
Rules	All 8 SENAR Core rules (Base row above), plus three stated in this Standard's numbering: 10.2 Session Duration, 10.4 Dead End Documentation, 10.9 Knowledge Capture. The eight and the three are not added together — 10.4 and 10.9 are how this Standard states Core Rule 8, so a total would count one obligation twice
Knowledge Base	Shared, accessible to all team members and AI sessions

> ***Note:*** *The Metrics row lists the four metrics Foundation collects and reports. Rule 10.9, one of the three named in the Rules row above, additionally requires a documented Knowledge Capture Rate target — a target and the tracking of it, not a reported metric of the set in Section 9.1. Where the two readings of this table diverge, 10.9 is the requirement and the Metrics row is the reporting obligation; an organization satisfies both by documenting a KCR target with its rationale (9.2) without adding KCR to its metric reporting. The four are not a superset of Core's two: Core tracks FPSR and Dead End Rate (SENAR Core, Metrics), so the move to Foundation keeps FPSR, adds Throughput, Lead Time and Defect Escape Rate, and stops reporting Dead End Rate. Dead ends themselves continue to be recorded — Rule 10.4 requires it at Foundation — but their rate is not one of the four reported here. The same distinction covers Adversarial Detection Rate: 10.15(e) requires organizations to track ADR, and 10.15 requires L3 review of High-risk changes regardless of configuration, so a Foundation organization that makes such a change owes the ADR figure under 10.15 while the Metrics row here lists four without it. As with KCR, 10.15(e) is the requirement and the Metrics row is the reporting obligation; both are satisfied by holding the ADR figure for the L3 reviews that occurred without adding ADR to the four reported metrics.*

> ***Note:*** *At Foundation, session duration monitoring (Rule 10.2) is the Supervisor's responsibility, since Flow Manager is not a separate role at this configuration (Section 4.8).*

> ***Note:*** *Core rules and Standard rules use different numbering, which is why the Rules row states two lists instead of one sum. Core Rule 8 (Capture Knowledge) expands into Standard Rules 10.4 (Dead End Documentation) and 10.9 (Knowledge Capture); Foundation states these explicitly alongside Standard Rule 10.2 (Session Duration). Two further Standard rules reach Foundation without appearing in that list, and both arrive through other rows of this table rather than through the Rules row: 10.5 (Periodic Audit) is what the monthly Quality Sweep in the Ceremonies row is, and 10.15 (AI Output Quality Verification) applies at L1 and L2, which its own table marks "All (SHALL)" — L3 additionally applies to High-risk changes regardless of configuration. Section 10 states its rules without a configuration qualifier and 10.16 summarizes them for Team and Enterprise only, so this table is where the Foundation set is settled.*

> ***Note:*** *Rule 10.6 (Version Control) states three duties — atomic commits, automated secrets detection, scope-creep review — and 6.4(e) states the same three without a configuration qualifier. They are therefore owed at Foundation already, through 6.4, and 8.10 records secrets detection as applying to all configurations. Rule 10.6 is absent from the list above because its content reaches Foundation through 6.4(e), not because that content is optional here; the head of Section 10 counts it among the rules that reach this configuration.*

Adoption Path (week by week)

- **Week 1 — Core habits + Session Start/End.** All team members read SENAR Core and begin applying its 8 rules daily. Introduce Session Start (context load + goal declaration) and Session End (output summary + knowledge entry) as team ceremonies. Assign the 5 responsibility sets across the 2 people — roles may overlap.
- **Week 2 — Dead End documentation + Knowledge capture.** Activate Rule 4 (Dead End Documentation): every blocked path gets a knowledge entry before work pivots. Activate Rule 9 (Knowledge Capture): every significant decision and non-obvious finding goes into the shared KB. After two weeks the team has a working knowledge base, not an empty one.
- **Week 3 — Quality Sweep (first monthly).** Run the first Quality Sweep: review the session log, check which dead ends were documented, identify recurring friction points. This is a retrospective with output — concrete adjustments to thresholds, session duration, or task decomposition depth.
- **Week 4 — Metrics baseline.** Start measuring all 4 metrics: Throughput (tasks completed per week), Lead Time (task start to done), FPSR (first-pass success rate at QG-2), Defect Escape Rate (bugs found after done). Record baseline values — no targets yet, just calibration. Adjust thresholds to your domain.

Foundation vs Core — What's Different

Core is individual discipline: one Supervisor, two quality gates, two metrics, all enforced at the personal level. Foundation adds **team coordination layer**:

- Shared knowledge base visible to all team members and AI sessions (not just the author's local context)
- Session Start/End as **team ceremonies** — synchronization points, not just individual habits
- Monthly Quality Sweep — collective review of process health, not individual retrospection
- Throughput, Lead Time and Defect Escape Rate on top of the FPSR that Core already tracks — three numbers that only mean something once several people share a definition of "done"

The cognitive overhead is low. Foundation is designed to be adopted incrementally on top of Core habits already in place.

FAQ for Foundation Teams

When do we move to Team? When any of the following applies: the team reaches 3+ Pairs and coordination overhead is visible; you need cross-pair dependency tracking or knowledge federation; you want QG-1 (Requirements Gate) or QG-3 (Verification Gate); you want dedicated (non-overlapping) responsibilities.

Can we skip Foundation and go directly to Team? Yes. If the team already has process discipline — structured task decomposition, code review culture, retrospectives — start at Team. Foundation is a bridge, not a mandatory stop.

What if we're exactly 3 pairs? Evaluate experience level. If everyone is experienced with structured AI-assisted development, go Team — the overhead is manageable. If the team has mixed experience or is just starting with AI workflows, stay Foundation for 4–8 weeks to build habits before adding federation and all 5 gates.

11.2 SENAR Team (3–10 Pairs)

Category	Requirements
Responsibilities	All 5 responsibility sets covered (SHALL). Dedication scales inside this configuration: minimum 3 distinct role-holders at 3–5 Pairs, and all five dedicated from 5 Pairs — except that Context Architect may still combine with Flow Manager at 5–6 Pairs (Section 4.8, which is the source for both). What is required is coverage, not a particular combination
Ceremonies	All 7 (per Section 7)
Quality Gates	All 5: QG-0 through QG-4 (SHALL)
Metrics	All 10 (SHALL)
Rules	All 15 (SHALL)
Federation	Cross-Pair dependency tracking, shared KB (SHALL)
Risk-Based Review	QG-3 differentiated by risk (Section 8.7)
Security Review	SHALL for high-risk changes (auth, payment, data, crypto) — see Section 8.7

NOTE: The Responsibilities row yields to Section 4.8, which is the single source for headcount in this configuration. Until 1.4 the row required "All 5 dedicated (SHALL)" across the whole configuration, which contradicted 4.8: the same configuration prescribes a minimum of 3 distinct role-holders at 3–5 Pairs and states that what is owed is coverage of the responsibility

sets rather than dedication. Two SHALL requirements demanded different things, and one of them had to yield; the one that yields is the one that is not the source. An organization holding a Team claim made under 1.3 reassesses under Section 13.6, which already names a new revision of this Standard among the triggers for reassessment — at SHOULD, the level that subsection states; this note introduces no second procedure and raises no level.

11.3 SENAR Enterprise (10+ Pairs)

All Team requirements, PLUS:

Category	Additional
Portfolio Management	Increments grouped by value stream with unified budget
Additional Responsibilities	Portfolio Manager, Chief Supervisor, Federation Coordinator
Compliance	QG audit trails for ISO/regulatory requirements
Governance	Gate Bypass review, architectural exceptions, budget oversight
Requirements-as-Code	Requirements stored in VCS, CI validates traceability, change review required (SHALL)

For scaling ratios and enterprise guidance, see SENAR Reference.

NOTE: Enterprise configuration describes target state for large-scale adoption. Organizations at this scale should validate ratios and adapt coordination mechanisms to their context.

11.4 Comparison

Element	Core	Foundation	Team	Enterprise
Pairs	1	1–3	3–10	10+
Rules	8 (Core)	8 Core + 10.2, 10.4, 10.9	15	15
Quality Gates	2	2 (QG-0, QG-2)	5 (QG-0..QG-4)	5 + compliance
Metrics	2	4	10	10 + portfolio
Responsibilities	1 (Supervisor)	5 covered by 2 people	5 covered; dedicated from 5 Pairs (4.8)	5 + portfolio
Ceremonies	0	3	7	7 + portfolio
Knowledge Base	Recommended	Required	Required + federation	Required + federation
Security Review	Checklist	Checklist	SHALL (high-risk)	SHALL + formal audit
Tooling	None	Recommended	Required	Required

NOTE: The Core column of this table is comparative throughout; 13.3 states what that means and is the only place that states it. This Standard's obligations begin at Foundation. The properties every Quality Gate is required to have, and the obligation level of each by configuration, are not a row of this table: they are stated once, in 8.6.

NOTE: The Tooling row above is the one the gate property scale of 8.6 follows, and it follows it deliberately: the level of each property is set by what the property costs to implement. 8.6 states the scale and the basis for each level.

11.5 Migration

```
SENAR Core ──► Foundation ──► Team ──► Enterprise
 + 3 roles (combined)  + dedicated roles    + portfolio management
 + 3 ceremonies        + federation         + value streams
 + 2 more metrics      + all gates          + compliance
 + session mgmt        + all metrics/rules  + governance
```

Each step is incremental. Pilot with a subset of Pairs first. Typical timelines:

- Core → Foundation: 2-4 weeks after team has internalized Core habits
- Foundation → Team: 2-3 months, when team reaches 3+ Pairs or needs federation
- Team → Enterprise: when organization has 10+ Pairs and needs portfolio governance

---

12. Maturity Model

> ***This section is informative.*** *It describes maturity levels and states no requirement: no SHALL appears in it, and no conformance claim is assessed against it. Three things would have to be settled before it could carry one, and none of them is settled by measurement this Standard holds: the model does not describe Foundation at all (12.1, Level 2 says so outright); it gives no rule for aggregating dimension levels into an organizational level, so an organization at L3 on four dimensions and L1 on three has no stated level; and the dimension table stops at L3 while the levels run to L5. Stating a level therefore remains an organization's own reading and is not a claim under 13.1.*

12.1 Levels

Level 1: Ad Hoc

AI used opportunistically. No tasks, no gates, no metrics, no documented knowledge.

Level 2: Supervised (aligns with SENAR Core)

Tasks before work. AI output verified. QG-0 and QG-2 enforced. Core's 2 metrics tracked — FPSR and Dead End Rate, which are the two SENAR Core names in its own *Metrics* section. Dead Ends documented (>15 min threshold). Session management is not part of this level: Core assigns the structured start and end of a work session to Foundation, and this model does not yet describe Foundation.

Key indicator: Every AI implementation has a Task; output is verified.

Level 3: Measured (aligns with Team)

All 5 responsibility sets covered. All 5 gates automated. All 10 metrics with baselines and targets (including ADR tracked per task). Knowledge capture structured. Federation active. Architectural Decision Records tracked — the abbreviation ADR is not used for them in this Standard, where 3.36 gives it to Adversarial Detection Rate. Data-driven decisions.

Key indicator: Baselines established; metrics reviewed at every Retrospective.

> ***Note:*** *Levels 4 and 5 are aspirational targets based on industry patterns (CMMI, DORA). No SENAR implementation has been independently validated at these levels. They are provided as directional guidance for organizations seeking continuous improvement.*

Level 4: Managed (aspirational)

Cost predictability consistent. FPSR improving. Defect Escape Rate below threshold. Knowledge base actively improves AI context quality. Process improvements are experimental and measured.

Level 5: Optimizing (aspirational)

Organization defines own patterns. Value streams optimized. Process experiments routine. Knowledge base is competitive advantage.

NOTE: Levels 4–5 are aspirational. See the note above Level 4 for details.

12.2 Assessment Dimensions

Organizations MAY be at different levels across dimensions:

Dimension	L1	L2	L3
Task Discipline	None	Tasks before work	Full lifecycle + traceability
Quality Gates	None	QG-0 + QG-2	All 5 automated
Metrics	None	2 (Core)	4 mandatory + 6 recommended
Knowledge	Undocumented	Dead Ends captured	Structured + target rate
Verification	None	Supervisor self-verifies	Independent verification, adversarial review and ADR tracking
Cost	Unknown	Tracked	Baselined and predicted
Requirements	Undocumented	Task goal + AC	Full hierarchy (BR→SR→TR) + quality properties

12.3 Progression

Organizations SHOULD progress sequentially. Each level builds on the previous.

Organizations MAY focus on weakest dimensions first.

13. Conformance

13.1 Claiming Conformance

An organization claiming conformance to SENAR SHALL:

a) Identify the applicable configuration: Foundation, Team, or Enterprise (Section 11); b) Implement all SHALL requirements of the applicable configuration; c) Document any SHOULD requirements that are not implemented, with rationale; d) Maintain evidence of implementation (audit records, metric data, task tracker records); e) Document any SHALL requirement handled under 13.5, with the same rationale that clause requires, and disclose the list with the claim wherever the claim is published. A claim documents its unimplemented SHOULD requirements under (c); it SHALL disclose its non-conformances too, and disclosing the weaker class while withholding the stronger would tell a relying party less the more the claim departs from this standard. A claim stated in the format below while records exist under 13.5 that it does not disclose is a claim of conformance without non-conformance, whatever the internal assessment record says.

A conformance claim SHALL use the following format: "[Organization] conforms to SENAR v[version] [Configuration] configuration, [self-declared | peer-assessed | independently audited], as of [date]."

13.2 Conformance Levels

Level	Description	Evidence
Self-declared	Organization assesses itself against SENAR requirements	Internal assessment record
Peer-assessed	Assessment conducted by a SENAR practitioner from another organization	Peer assessment report
Independently audited	Assessment by a qualified auditor with SE process assessment experience	Formal audit report

Organizations MAY claim any conformance level. Higher levels provide stronger evidence for stakeholders, clients, and regulators.

13.3 SENAR Core

Teams implementing SENAR Core are not required to conform to this Standard but are encouraged to adopt Foundation configuration when ready. SENAR Core conformance is declared separately per the SENAR Core document.

This governs how every Core column in this Standard is to be read, and it is stated here and nowhere else. Those columns are informative. This Standard's obligations begin at Foundation (11.4). Where a Core column states a level — as the gate property scale in 8.6 does — it states what the level would be at that scale, so that the Foundation row can be read as a step up from something. It is not a requirement on anyone, and no conformance claim is assessed against it. What Core requires of a gate is in the SENAR Core document, in Core's own register and without this Standard's machinery — see *Three things that make a gate a gate* there, which carries (a), (c) and (e), the three that work with a text editor: a document of eight rules that acquired a Gate Bypass procedure would no longer be the thing its adopters chose. Sections that carry a Core column carry a pointer to this subsection instead of restating the argument.

13.4 Partial Conformance

Organizations MAY claim partial conformance by specifying which sections they conform to (e.g., "SENAR Team conformant for Sections 6, 8, 9 and 10"). Partial conformance claims SHALL explicitly list the included sections, and the list SHALL accompany the claim wherever the claim is published. A claim stated in the format of 13.1 without its list of sections is a claim of full conformance to the named configuration, whatever the internal assessment record says.

Partial conformance claims SHALL include at minimum Sections 6 (Units of Work), 8 (Quality Gates), 9 (Metrics), and 10 (Operational Rules) — these constitute the irreducible core of SENAR practice. A claim covering only definitional sections (Sections 1–4) is not a valid partial conformance claim.

A partial conformance claim that does not include Section 7 SHALL state the retention interval it applies under 8.6(g). The floor 8.6(g) names is an interval between ceremonies of Section 7, so a claim that omits Section 7 leaves that floor without an anchor, and the records the rest of the claim is assessed against have no stated life. A partial conformance claim that includes Section 8 SHALL include Section 8.6 in its entirety. The gate properties are not divisible: where a property is required at the claimed configuration and is absent, the result is not a partially conforming gate but a check that does not meet the definition of one (8.6). Claiming the criteria of 8.1–8.5 while omitting 8.6 claims the form of the gates without the substance.

13.5 Non-Conformance Handling

When a SHALL requirement cannot be met, the organization SHALL document the non-conformance with: justification, risk acknowledgment, remediation plan, and approval by a role senior to the requestor. Non-conformance records are distinct from operational Gate Bypasses (Section 3.13) — they apply at the process level, not the task level. A record made under this subsection is disclosed with the conformance claim under 13.1(e): this subsection is a legitimate path around a SHALL, and a path around a requirement that the relying party cannot see is indistinguishable, from outside, from the requirement having been met.

13.6 Maintaining Conformance

Organizations claiming conformance SHOULD reassess at each Increment Retrospective (Section 7.7). Where that ceremony is not conducted — Foundation omits it (11.1) — organizations SHOULD reassess at each Quality Sweep (7.4) instead, and it is that interval 8.6(g) reads as the floor for retaining records. Conformance claims SHOULD be re-evaluated after: a new revision of this Standard, major process changes, AI model generation changes (Section 10.13), or organizational scaling across configuration tiers.

Reassessment required by revision 1.4

A conformance claim made against a revision earlier than 1.4 was not assessed against Section 8.6, because those properties did not exist then. Organizations holding a Team or Enterprise claim SHALL reassess against 8.6(b) in particular. A gate placed after the effect it declares — at review, at merge, or at any point where the effect has already occurred — is a detective control and does not satisfy 8.6(b), which is SHALL from Team.

Some claims that were valid under earlier revisions are not valid under this one. This is a consequence of stating the requirement, not of raising it: before 8.6, placement was not specified, so a claim could be made without the question being asked. Reassessment is required, not optional. Where the requirement cannot be met, 13.5 applies — the non-conformance is documented with justification, risk acknowledgment, remediation plan, and senior approval, and the claim is corrected rather than maintained. The configuration notation of Section 3, corrected in the same revision, calls for no reassessment: the key was restored to the meaning every use of it in Sections 4–13 already carried, so no requirement changed level.

Reassessment required by revision 1.5

A conformance claim made against revision 1.4 was assessed against a smaller set of requirements than this revision states. Six are new or newly unconditional, and an

organization holding a 1.4 claim SHALL reassess against them:

a) 8.3 — root cause identified and recorded for any defect found and fixed within the Task, at QG-2 and therefore at every configuration; b) 8.7 with 8.4 — the AI Output Review Minimum Criteria and the review of 10.15 are owed for security-sensitive changes at every configuration, performed at the gate that configuration operates. Until 1.5 this was a recommendation in a note; c) 9.4(a)–(d) — the method behind a reported figure is disclosed, a figure drawn from self-made records is not presented as a measure of what occurred, a self-set target carries its basis, and a crossed target is escalated rather than quietly moved; d) 10.15(f) — the severity scale is documented and names which of the three scales that share its words it defines; e) 13.1(e) — departures recorded under 13.5 are disclosed with the claim wherever the claim is published; f) 13.4 — a partial claim that omits Section 7 states the retention interval it applies under 8.6(g).

As with 1.4, this is a consequence of stating requirements that were previously unstated or stated only in notes, not of raising ones already stated. Where a requirement cannot be met, 13.5 applies and the claim is corrected rather than maintained. Two changes of this revision relax rather than tighten and call for no reassessment: 5.6 is scaled to Team+ from an unconditional statement, and 8.2 no longer states a level for configurations that do not operate QG-1.

13.7 Mechanically Checkable Clauses

Conformance to most of this standard is assessed by reading: process descriptions, records, and statements about what an organization does. The gate properties of Section 8.6 are stated here in the form an assessor can check against evidence rather than against a description. They are not the only requirements in this standard with that character — 8.2 (orphaned requirements), 10.10 (traceability), 10.13 (model identifier per Session), 10.15(c) (recorded findings) and 6.2 (Task attributes) are all decidable from tracker and version-control data — but they are the ones whose evidence this section specifies. They are listed so that a claimant knows what to retain and an assessor knows what to ask for.

What follows is checkable only to the extent that the records required by 8.6(g) exist, are complete, and have been retained. Those obligations are stated there, not here, and an assessor whose claimant cannot produce them has not found a passing gate; they have found an unassessable claim.

Property	What is presented	How it is checked
a) Declared prevented effect	The gate's definition, naming the effect	The declaration exists, and a specific proposed action can be tested against it
b) Preventive placement	Audit records (8.6(g)) for the gate and for the declared effect	The verdict is recorded before the effect, not after
c) Correspondence of the verdict to state	The audit record identifying the state the verdict was issued for	No verdict is applied to a state other than the one it names
d) Binding to a measured digest	The recorded digest and the artifacts it names	The digest recomputed over those artifacts equals the recorded value
e) Fail-closed	Records of runs that reached no verdict, and the checks the gate depended on with the result of each (8.6(g))	No positive verdict was recorded or relied upon in any such case; no verdict is positive while a check it depended on produced no result; and where the gate is preventively placed, the declared effect was prevented

These differ in how mechanical they are, and the difference is not concealed: (d) is recomputable by an assessor holding the artifacts and the record of the set the digest covers; (b), (c) and (e) are decidable from the records 8.6(g) requires — of the gate execution and of the admission of the effect — and from nothing else; (a) is a presence check followed by the applicability test stated in 8.6(a), which needs a reader.

Two limits belong here rather than in a reader's discovery. First, (d) establishes that a recorded digest and a set of artifacts correspond; both are produced by the claimant, so the recomputation demonstrates internal consistency and not the moment of recording. Second, and stated in 8.3 for the gate itself: none of these checks is an attestation against a participant who intends to mislead. They are controls against evidence drifting from what it describes. An assessor seeking assurance against a purposeful claimant should look to the conformance level (13.2) and to who produced the records, not to a further row in this table.

This adds nothing to Section 13.2. The three conformance levels are unchanged and no fourth level is introduced. What changes is that at any of those levels, part of the assessment can rest on evidence rather than on assertion.

NOTE: This standard defines no certification scheme, maintains no register of claimants, and requires no paid assessment. A conformance claim is made by the organization itself, per 13.1. Nothing in this section creates a body, and nothing in it makes an assessment by a third party a precondition for claiming conformance.

SENAR Quick Start: 5 Minutes to Better AI Development

You use AI to write code. SENAR makes the result reliable.

No meetings. No mandatory certifications. These habits are the **SENAR Core** — 8 rules, 2 quality gates, 2 metrics — everything you need to start. Adoption overhead: under 1 hour, about 5 minutes per session after that.

The 6 Habits

BEFORE you tell the AI what to build

1. Write the WHAT and the DONE

Before starting, write two things:

- **Goal** — what must be accomplished (one sentence)
- **Acceptance Criteria** — how you'll know it's done (numbered list, each independently testable)

Bad: "Add login functionality" Good: "Implement email/password login. AC: 1. POST /auth/login returns JWT on valid credentials. 2. Returns 401 on invalid password. 3. Returns 422 on missing email field. 4. Token expires in 24h."

Why this matters: AI output quality equals input quality. A vague goal produces plausible-looking code that fails in production. A precise goal with clear acceptance criteria produces testable, correct code on the first try.

2. Set boundaries

Tell the AI what NOT to touch:

- "Change ONLY the users/ directory"
- "Do NOT modify the database schema"
- "Follow patterns from auth/router.py"

Without boundaries, AI will confidently refactor half your codebase to "improve" things you didn't ask about.

WHILE the AI works

3. Verify against your criteria, not your intuition

Don't just glance at the code and think "looks right." Check each acceptance criterion:

- AC 1: Does POST /auth/login return JWT? → Check the test or run it.
- AC 2: Does it return 401 on invalid password? → Check the test.
- AC 3: Missing email → 422? → Check the test.

If there's no test for a criterion, the criterion isn't verified.

4. Document dead ends

When an approach fails, write one sentence about why:

- "Tried bcrypt for password hashing — import fails on Python 3.14, switched to argon2"
- "SQLAlchemy async session: can't use lazy loading, need selectinload"

This takes 10 seconds and saves hours — for you next week, for your teammate, for any AI that reads your knowledge base.

AFTER it's done

5. Run the tests

If tests pass AND all acceptance criteria are met → done. If not → not done. No exceptions, no "it probably works."

6. Capture what you learned

If you discovered something non-obvious during this task, write it down:

- A decision you made and why
- A pattern that worked well
- A gotcha that surprised you

> ***Note — 6 habits vs. 8 Core rules:*** *These 6 habits cover the essence of the 8 SENAR Core rules. The full Core adds two explicit practices: a tiered verification checklist for latent defects (Rule 5) and root-cause analysis before patching symptoms (Rule 7). If you are working in a regulated or high-stakes context, read the full Core document before relying on this Quick Start alone.*

Before & After: Real Data

These numbers come from one project (6 microservices, 552 tasks, $989 in AI costs, 38 sessions). They are illustrative, not universal — your numbers will differ. SENAR requires you to establish your own baselines before setting targets.

Without structured process (sessions 1–3, ad hoc):

- Tasks started without acceptance criteria → AI produced code that "looked right" but failed edge cases
- No dead ends documented → same failed approaches repeated across sessions
- No session discipline → 200+ minute marathon sessions with noticeable efficiency decline
- Defects discovered after "done" → estimated rework cost ~$105 per escaped defect (estimated from project data: average rework cost of defects found after task completion)

With SENAR habits adopted (sessions 4–38):

- Every task has goal + AC before start → First-Pass Success Rate improved to 85%+ (tasks correct on first try)
- Dead ends documented → repeated failures eliminated for documented cases
- Sessions capped at 120 min with checkpoints → consistent throughput, no context crashes
- QG-0 blocks taskless work → zero "what was this for?" tasks

Caveat: This is a single team's experience (N=1), not a controlled study. The improvement conflates methodology adoption with natural team learning. Independent replications are needed — and that's why we published the standard, so others can measure too.

The cost of SENAR: ~5 minutes session overhead + 1–3 minutes per task. **What you avoid:** rework from vague requirements, repeated dead-end approaches, marathon sessions with declining output.

That's It

These habits correspond directly to the **SENAR Core** rules — the self-contained subset of the SENAR Standard. Everything else in the methodology builds on this foundation.

What you get:

- Fewer "works on my machine" surprises
- AI produces correct code on the first try more often (we measure this — it's called First-Pass Success Rate)
- Knowledge accumulates instead of evaporating between sessions
- You can hand off to another developer (or your future self) without losing context

What it costs: ~5 minutes session overhead + 1–3 minutes per task for goals, criteria, and verification.

Next Steps

You want to...	Read
Read the formal SENAR Core document (8 rules, 2 gates, 2 metrics)	**SENAR Core**
See a complete task start-to-finish	Guide: Worked Example
Set up SENAR with your AI tool	Guide: Tool Integration (Claude Code, Cursor, Copilot)
Adopt SENAR in an existing codebase	Guide: Legacy Adoption
Learn about requirement levels	Guide: Requirements Engineering
Understand the philosophy	Guide: Philosophy
Upgrade from Core to the full Standard	Guide: Transition — Core to Standard
Scale to a team	Standard, Section 11: Configurations
Evaluate your current practice	Standard, Section 12: Maturity Model

---

SENAR Guide: Philosophy

This document expands on the SENAR Values and explains the six pillars that underpin the methodology.

SENAR Values

1. **Context over Code** — AI output quality is determined by input context quality. Invest in requirements, not in coding speed.
2. **Verification over Speed** — AI generates at machine speed. Correctness, not velocity, is the constraint.
3. **Knowledge over Experience** — AI has no memory between sessions. What's not documented doesn't exist for AI.
4. **Enforcement over Agreement** — Quality gates as automated code, not meetings people can skip.
5. **Judgment over Keystrokes** — Human attention on decisions (what to build, is it correct), not on typing code.

Pillar 1: Context-First (Quality at Input)

The quality of AI output is a direct function of input context quality.

The cascade principle: Quality is built at input, not checked at output. A defect in a Business Requirement propagates to all downstream System Requirements, Task Requirements, and ultimately to code. By the time a defective requirement reaches AI execution, it produces plausible-looking code that passes automated checks but solves the wrong problem. No amount of testing at the output (QG-2, QG-3) can catch a requirement that was wrong from the start.

This is why SENAR invests in requirement quality (QG-0, QG-1) before implementation begins — not as bureaucracy, but as the highest-leverage quality investment.

Requirements ARE context. A well-defined Business Requirement that decomposes into clear System Requirements and Task Requirements IS the primary input to AI-generated code quality. The requirement hierarchy (BR → SR → TR) is not paperwork — it is the structured context that determines whether AI produces correct output.

Context components:

- Task goal and acceptance criteria (= Task Requirements)
- Requirement links to parent Story/BR/SR (= traceability)
- Architectural constraints and conventions
- Relevant knowledge (decisions, dead ends, gotchas)
- Examples and anti-patterns
- Scope boundaries (what NOT to change)

Common failures:

- *Vibe prompting* — vague instructions without structure. Appropriate for trivial tasks; dangerous for complex work.
- *Context dumping* — flooding AI with unstructured information. Critical constraints get buried.
- *Implicit assumption* — expecting AI to know conventions without documentation.
- *Ambiguous requirements* — requirements that a human would clarify in conversation but AI interprets literally or hallucinates a resolution. Unlike human teams, AI does not ask "did you mean X or Y?" — it picks one silently.

Pillar 2: AI-First, Not AI-Only

The Supervisor's primary mode is AI-directed work. Manual coding is a justified exception, not a prohibition.

The exception is recorded. Since revision 1.4, modifying an artifact by a route QG-0 does not precede is a Gate Bypass and is recorded as one, with justification, risk acknowledgment, remediation plan, and senior approval (Standard 8.6(j), 4.1). This does not make manual coding rarer or harder — it makes it visible. What is no longer available is doing it silently, and "preparing the context costs more than the fix" is a justification you write down, not a reason to skip writing one.

When manual intervention is appropriate. Since Standard 1.5 the answer is narrower than this Guide used to give, and the reason is dated rather than principled: the list below described what agents could not do when it was written, and models moved.

One case is recognized as substantive (Standard 4.1): **an environment in which no agent can be run and there is nothing to switch to.**

The cases this Guide used to list are not exceptions where an agent environment exists — each has an agent-side answer:

Case as it used to be listed	What to do instead
Micro-fix (1–3 lines) cheaper than preparing context	Prepare less context, not none: a Task with a goal and one acceptance criterion is cheap
AI stuck in a loop on a wrong approach	Change the model or the framing; a loop is a context defect, and fixing the output leaves the defect
Infrastructure config where AI hallucinates	Give the agent the schema or the vendor page; hallucination on config is a missing-context symptom
Time-critical hotfix	Escalate the model, not the process. Where the environment itself is unavailable, this is the recognized case above, not a second one
AI got 90% right — easier to fix 3 lines	A second pass on the same Task costs less than the record the bypass requires

Where you do modify directly, the record is required in every case, including the recognized one — otherwise metric 8 loses the population it is computed over (Standard 9.2). Expect the recognized case to be a small share of your Gate Bypass records; a large one is a statement about your environment. See Guide 06, PF-3 for how to read the number.

Common failures:

- *Shadow coding* — writing code and hiding it from traceability. Under 8.6(j) this is now an unrecorded Gate Bypass, not merely a cultural problem
- *Micromanaging* — rewriting >30% of output instead of improving context
- *Rubber stamping* — accepting without verification

Pillar 3: Enforcement over Ceremony

Quality is enforced through automated gates, not meetings.

AI agents don't attend meetings, feel accountability, or learn from retrospectives. The only reliable quality mechanism is automated enforcement.

Purpose	Mechanism
Decide what to build	Ceremony
Verify code quality	Gate
Review with stakeholder	Ceremony
Verify requirement met	Gate

Pillar 4: Knowledge Persistence

What is not documented does not exist for AI.

Code documentation is context, not afterthought. In AI-native development, code documentation serves a dual purpose: it helps human Supervisors understand the system AND reduces the context volume AI needs per Task. A well-documented module (docstring explaining purpose, public API contracts, architectural boundaries) means the Supervisor doesn't need to re-explain these things in every Task goal. AI reads the docs, understands the module's role, and produces code that fits.

Undocumented code forces the Supervisor to compensate: longer Task goals, more explicit constraints, more scope boundaries. This is expensive context that should live in the code itself. Documentation that says "handles OAuth flow for Google and GitHub, stores tokens encrypted in session table, refreshes automatically" saves 5 lines of Task context on every auth-related task.

Rule 9.11 (Code Documentation as Context) makes this a SHALL requirement: code documentation sufficient for AI to understand module purpose, API contracts, and boundaries without reading the full implementation.

What to capture: decisions (with rationale), patterns, gotchas, dead ends (highest reuse value), observations.

Knowledge delivery to AI: static context files, search APIs, MCP integrations, or RAG. Choose deliberately.

Knowledge lifecycle: `current` → `needs_review` (flagged as potentially stale) → `deprecated`. Stale entries actively harm output quality — they give AI incorrect context.

Pillar 5: Interaction Patterns

AI supervision is a dialogue, not a one-shot command.

Pattern	When to Use
Plan-then-Execute	Complex tasks: ask AI for plan → review → approve → execute
Iterative Refinement	Most tasks: generate → review → course correct → improve
Example-Driven	UI/patterns: "do it like this but with X change"
Negative Example	Known pitfalls: "don't do X because Y"
Checkpoint-and-Verify	Multi-step: AI does step 1 → verify → step 2
Constraint Fence	Scope control: "change ONLY files X, Y. Do NOT touch A, B"
Rollback-and-Retry	Wrong direction: git rollback, restart with different context
Exploration	Unknown territory: investigate before committing to approach

Context Window Management

- AI context degrades as conversation grows (saturation)
- Progressive disclosure: provide information as needed, not all at once
- Strategic checkpoints clear context and restart fresh
- For large codebases: targeted file reads over bulk dumps
- Handoffs serialize essential context for new sessions
- If AI starts confusing things, it's not the model — the context is full

Hallucination Management

Common types: non-existent APIs/methods/CLI flags, non-existent files, correct-looking code with subtle edge case bugs, confident but wrong assertions.

Detection heuristics:

- Excessive confidence in novel solutions (red flag)
- "Suspiciously perfect" code handling every edge case
- References to paths/APIs you don't recognize
- Always run the code — don't trust AI's claim about what it does

Pillar 6: Empirical Calibration

Every rule and target should be calibrated to your data.

SENAR provides formulas, not targets. Organizations establish baselines by measuring for 3+ Increments, then set targets based on their own reality.

Common failures:

- *Cargo culting* — adopting another org's targets
- *Premature optimization* — aggressive targets before understanding baselines
- *Metric fixation* — optimizing for metric rather than outcome

SENAR Guide: AI Output Review Checklist

> ***NOTE — Relationship to SENAR Core***
>
> *SENAR Core includes an updated **Verification Checklist** (28 items, 3 tiers: **Standard / High / Critical**) that supersedes this checklist for Core adopters.*
>
> ***Tier mapping:** Tier 1 (this guide) ≈ Standard (Core) | Tier 2 ≈ High | Tier 3 ≈ Critical*
>
> *This Guide checklist is retained for teams adopting SENAR through the Standard (team-level entry point) who have not yet moved to Core. If your team uses SENAR Core, use the Core Verification Checklist instead.*

> ***NOTE — Status of this checklist under the Standard***
>
> *This checklist is informative. It is not the normative requirement at QG-3, and no conformance claim rests on it.*
>
> *The normative floor is Standard 8.4, AI Output Review Minimum Criteria — five checks that SHALL be performed at QG-3. Standard 8.4 states that those criteria replace dependency on this checklist: an organization is required to perform the five, not to adopt this document. That distinction was not stated anywhere the reader of this chapter would find it, and the chapter read as though ticking it were the obligation.*
>
> *The five are covered here, and this is the mapping:*
>
> | ***Standard 8.4 (SHALL at QG-3)*** | ***Items in this checklist*** |
> |---|---|
> | *a) Hallucination check — no non-existent APIs, methods, CLI flags* | *24* |
> | *b) Dependency validation — imports exist in the manifest* | *3, 4* |
> | *c) Pattern conformance — follows project architecture and conventions* | *6, 7, 12, 20* |
> | *d) Test validity — tests exercise the AC, not just coverage* | *8, 9* |
> | *e) Secrets check — no hardcoded credentials, keys, or secrets* | *5, 10* |
>
> *Everything else here is beyond the floor, offered as an aid rather than as a requirement. Two consequences worth stating plainly: running all 24 items does not by itself satisfy 8.4 unless each of the five was actually performed and recorded; and a team that finds this list too long may reduce it to the five without falling below the Standard.*

When reviewing AI-generated output, check for these AI-specific issues.

The Checklist

#	Check	What to Look For
1	**Scope**	Did AI modify files outside the task scope? (most common AI error)
2	**Deletions**	Did AI silently remove or replace existing working code?
3	**Phantom imports**	Are all imported packages in the dependency file?
4	**Dependency versions**	Are specified versions real and published?
5	**Hardcoded values**	Magic numbers, URLs, credentials, API keys in code?
6	**Over-engineering**	Unnecessary abstractions, patterns, or generalization?
7	**Duplication**	New code that duplicates existing utilities?
8	**Test quality**	Do tests verify behavior, or just mirror implementation?
9	**Test tampering**	Did AI modify tests to pass instead of fixing the code?
10	**Security**	Open CORS, hardcoded tokens, SQL without parameterization?
11	**Edge cases**	Happy path works — what about null, empty, boundary, concurrent?
12	**Naming**	Does AI follow project naming conventions?
13	**Commit scope**	Is the commit atomic and focused, or a kitchen sink?
14	**Null guard before comparison**	`None == None` is `True` in Python (JS: `null === null` is true; most languages have equivalent null-equality traps) — access checks bypassed when both sides null?
15	**Empty config bypass**	Security check skipped when config value is empty string? (`if secret and ...` fails open)
16	**Header trust**	X-Forwarded-For, X-Partner-ID, Content-Length used for security without proxy validation?
17	**IDOR**	Resource accessed by ID without verifying user's access to that resource? Auth ≠ authorization.
18	**Return True shortcut**	Access control function returns True/grants access without explicit ownership validation?

19	**Format string injection**	`str.format(**untrusted_dict)` — Python format supports attribute access, enables injection (JS: template literals with `eval()`; C/C++: `printf` format strings; any language: string interpolation with untrusted input)
20	**God functions/files**	Functions >50 lines, files >400 lines — strongest signal of unrefactored AI output
21	**Unreachable safety code**	`return False` after exhaustive exception handling — AI added "just in case" but can never execute
22	**Swallowed exceptions**	`catch/except` blocks that discard errors (`except Exception: pass` , `catch(e) {}` , or logging at debug level) — hides real failures. Check for null/None/nil returns that silently mask error conditions
23	**Unsafe deserialization**	pickle.loads, yaml.load without SafeLoader, eval/exec on untrusted input, JSON prototype pollution? (Java: `ObjectInputStream`; JS: `eval(JSON)`; any language: deserializing untrusted data without validation)
24	**Phantom APIs**	Do the called functions, methods, fields and CLI flags actually exist in the version of the library, framework or tool in use? Distinct from item 3: the package can be present in the manifest and the call into it still invented. Signatures are hallucinated far more readily than package names, and a plausible one survives review precisely by looking right — check it against the installed version's documentation or by executing it, not against recollection

Priority Tiers

The tiers below say how much of this list to run and when. They do not govern Standard 8.4: the five criteria in the mapping above are performed at QG-3 on every Task, whichever tier that Task falls into. Several of the items those criteria rest on sit in Tier 2 and Tier 3 here — 8.4(b) on item 4, 8.4(c) on 6, 7, 12 and 20, 8.4(e) on 5 and 10 — and the tier placement says when this Guide suggests reading the item, not whether the Standard's criterion applies.

Tier 1 — Always check (every task): Items 1–3 (scope, deletions, phantom imports), 24 (phantom APIs) and 8–9 (test quality, test tampering). These catch the most frequent and most dangerous AI defects. A 6-item check takes a couple of minutes.

Tier 2 — Security-sensitive tasks (auth, payment, data, API): Items 10 (security), 14–18 (null guard, empty config, header trust, IDOR, return True). These catch latent defect patterns — AI output that looks correct but fails under adversarial conditions. Identified through adversarial audits of production AI-generated code.

Tier 3 — Deep review (complex tasks, agent dispatch output): All remaining items (4–7, 11–13, 19–23). Apply when reviewing complex features, refactors, or any output from dispatched agents (Rule 15 L3).

Usage

Print this list. Use it at QG-2 (Implementation Gate) for every Task. Over time, some checks become automatic habits — but keep the list visible for new Supervisors.

SENAR Guide: End-to-End Walkthrough

Adoption Paths

SENAR has two adoption paths. Start with Core; upgrade to Standard when your needs outgrow it.

Core Path (Individual or Small Team)

1. **Adopt the 8 rules** — Task Before Code, Scope Boundaries, Verify Against Criteria, Tests Verify Requirements, Check for Latent Defects, Zero Tolerance for Incomplete Work, Fix Causes Not Symptoms, Capture Knowledge.
2. **Enforce the 2 gates** — Start Gate (goal + AC + negative scenario + scope before implementation) and Done Gate (all AC verified, checklist passed, knowledge captured).
3. **Measure the 2 metrics** — FPSR (First-Pass Success Rate) and Dead End Rate. Establish baselines for 3+ cycles before setting targets. Dead End Rate is not abbreviated DER anywhere in SENAR: DER is the Standard's Defect Escape Rate (Standard 9.1), a different number that Foundation reports and Core does not.

That is all. No roles, no ceremonies, no session management. The walkthrough below demonstrates this path.

Standard Path (Team or Organization)

1. **Start with Core** — adopt all 8 rules, both gates, both metrics.
2. **Add Foundation configuration** — 1–3 Pairs; add session discipline (Session Start/End, monthly Quality Sweep); two people cover all five responsibility sets — Supervisor + Context Architect + Flow Manager on one, Knowledge Engineer + Verification Engineer on the other (Standard 11.1, 4.8). 4 metrics: FPSR stays, Throughput, Lead Time and Defect Escape Rate are added (adds Throughput, Lead Time), QG-0 + QG-2.
3. **Add Team configuration** — introduce dedicated roles (Context Architect, Flow Manager, Verification Engineer), full ceremonies (Increment Planning, Quality Sweep, Retrospective).

4. **Add Enterprise configuration** — federation coordination, cross-project metrics, portfolio governance, compliance audit trails.

The Team Walkthrough section below demonstrates the Standard path at Team configuration.

Core Walkthrough (Solo Supervisor)

One person, one AI agent, one session. Building a REST API endpoint.

Session Start (2 min): Open terminal. Load previous handoff: "Auth module done. Next: user profile endpoint." Select task `impl-user-profile`.

Context Gate check: Goal: "GET /users/me returns current user profile." AC: "1. Returns 200 with user data. 2. Returns 401 without token. 3. Includes company info." Story link: `user-management`. → Passes.

Execution: Direct AI with Constraint Fence:

> *"Implement GET /users/me. AC: [above]. Follow patterns from auth/router.py. Change ONLY users/ directory."*

(This example uses Python/FastAPI. The same pattern applies to Java/Spring Boot @RequestMapping, Go/Gin router.GET, TypeScript/NestJS @Get().)

AI generates router + test. Review with checklist: scope ✓, imports ✓, no hardcoded values ✓, tests check behavior not implementation ✓. One issue: AI didn't test the 401 case.

> *"Add test for unauthenticated request → 401."*

AI adds test. Re-review: clean.

Implementation Gate: CI green, 6/6 tests pass, mypy clean, AC verified — against commit `4f2a9c1e`, working tree clean. → Done. (For other stacks: Java uses javac + SpotBugs, Go uses go vet + staticcheck, TypeScript uses tsc --strict.)

The commit is part of the verdict, not a footnote to it: Standard 8.6(c) requires the verdict to be about the state being closed, and a green run stops being a verdict the moment a file changes underneath it. See Guide 09, Section 7 for the full form.

Dead End capture: None this task — straightforward.

Session End (3 min): 4 tasks done, 75 minutes. Handoff: "Profile, settings, avatar upload done. Next: password change. Warning: avatar upload needs file size validation."

Total overhead: 5 minutes (start + end). No ceremonies, no roles, no meetings. Just discipline.

> ***The section below demonstrates Team-level SENAR.*** *If you're using Core only, you can stop here — the Core walkthrough above is your complete workflow.*

Team Walkthrough (3 Pairs)

A complete SENAR cycle for Team configuration. 3 Supervisor+AI Pairs, web application.

Increment Planning

Context Architect leads:

1. Reviews backlog: 32 candidate Tasks across 6 Stories
2. Applies WSJF: Authentication scores highest (blocks other work, moderate size)
3. Assigns: Pair A → auth (6 tasks), Pair B → orders CRUD (8 tasks), Pair C → CI/CD (5 tasks)
4. Risks: "Auth library choice may need exploration" → creates Exploration
5. Budget: 19 tasks selected for this Increment

A Supervisor's Session

Session Start (2 min): Load handoff → "Auth model done. Next: login endpoint." Select task `impl-login-endpoint`. Dev environment green.

QG-0 (automatic): Goal ✓, AC ✓ (4 criteria), requirement link ✓, work type ✓ → passes.

Execution (Plan-then-Execute + Constraint Fence):

> *"Implement POST /auth/login. AC: [criteria]. Follow patterns in auth/models.py. Use PyJWT. Config in settings.py. Change ONLY auth/ directory."*

(This example uses Python/FastAPI with PyJWT. The same constraint fence pattern applies to any stack: Java/Spring Security with jjwt, Go with golang-jwt, TypeScript/NestJS with @nestjs/jwt.)

AI generates router, service, tests. Supervisor reviews with checklist:

- Scope: ✓ only auth/ files
- Phantom imports: ⚠ AI imported `python-jose` but project uses `PyJWT`
- Hardcoded values: ✓ token secret from env
- Edge cases: ⚠ no expired token test

Iterative Refinement:

> *"Two issues: 1) Use PyJWT not python-jose. 2) Add expired token test."*

AI corrects. Re-review: clean.

QG-2 (automatic): CI ✓, 14/14 tests ✓, mypy ✓, lint ✓, AC verified ✓, security ✓ → passes.

Knowledge capture: "Gotcha: AI defaults to python-jose for JWT even when PyJWT is in requirements. Always specify library in context."

3 tasks done, 110 minutes.

Session End (5 min): Metrics saved, handoff written, 1 knowledge entry.

Quality Sweep

Quality Sweep procedure:

1. Select 3-5 recently completed tasks (random or risk-weighted)
2. For each task, verify: (a) AC have evidence, (b) checklist was applied at correct tier, (c) knowledge was captured
3. Record findings as observations in the knowledge base
4. If FPSR drops below target, discuss root causes in the next Team Sync

After 6 sessions, Verification Engineer audits:

Finding	Action
Duplicate validation utility (exists in shared/)	Fix task: refactor
3 TODO comments left by AI	Fix task: resolve
Zero test coverage for order deletion	Fix task: add tests
Inconsistent error format between modules	Fix task: standardize

Knowledge: "Pattern: AI undertests delete operations — add to standard AC template."

Increment Retrospective

Metric	Value
Throughput	7.3 tasks/session
Lead Time (median)	35 minutes
FPSR	74%
DER	5.3%
KCR	0.42
Cost Predictability	110%
MIR	12%

Actions: improve context template (JWT library), add "test deletion" to AC template, review QG-2 for concurrency tests.

SENAR Guide: Transition from Traditional Roles & Core to Standard

Upgrading from SENAR Core to SENAR Standard

When to Upgrade

SENAR Core is sufficient for individual developers and small teams (1-2 people). Consider upgrading to the full Standard when:

- **Team grows beyond 3 people** — you need roles (Context Architect, Flow Manager, Verification Engineer) to coordinate work across Supervisors.
- **Multiple projects** — cross-project dependencies require federation coordination, not ad-hoc communication.
- **Compliance or regulatory needs** — you need audit trails, formal requirement traceability, and documented quality evidence.
- **Knowledge silos appear** — tribal knowledge is not making it into the knowledge base, and different Supervisors produce inconsistent results.
- **Dead End Rate plateaus** — it stops improving because knowledge is captured but not systematically reused across the team.

What Standard Adds Over Core

Dimension	SENAR Core	SENAR Standard
Rules	8 rules	15 rules (adds session management, knowledge lifecycle, code documentation, federation)
Gates	Start Gate + Done Gate	5 gates: QG-0 (Context) + QG-1 (Requirement) + QG-2 (Implementation) + QG-3 (Verification) + QG-4 (Deployment)
Metrics	FPSR + Dead End Rate	10 metrics (adds Throughput, Lead Time, Defect Escape Rate, KCR, MIR, Cost per Task, Cost Predictability, and more; Dead End Rate is not among them)
Roles	Supervisor (implicit)	Supervisor, Context Architect, Flow Manager, Verification Engineer, Knowledge Engineer
Ceremonies	None	Session Start/End, Increment Planning, Quality Sweep, Increment Retrospective
Configurations	Not applicable	Foundation, Team, Enterprise — progressive adoption levels (Core serves as the entry point)
Session management	Not prescribed	Duration limits, checkpoint cadence, handoff discipline
Multi-team	Not covered	Federation Sync, cross-project dependencies, portfolio metrics
Maturity model	Not applicable	6 dimensions, 5 maturity levels per dimension

Migration Steps

If you are already using SENAR Core, migration to Standard is incremental — nothing is unlearned.

Step 1: Map your Core practice to Standard (Day 1) Your 8 Core rules map directly to Standard rules. Your Start Gate maps to QG-0. Your Done Gate maps to QG-2. FPSR remains your primary metric, and Dead End Rate remains the second of Core's two. SENAR Core is the entry point to the Standard — you are ready for Foundation configuration.

Step 2: Add session discipline — Foundation configuration (Week 1) Introduce formal Session Start (load context, select tasks, verify environment) and Session End (write handoff, capture metrics, document knowledge). Cap sessions at 4–6 hours. Add a monthly Quality

Sweep. This is Foundation: the SENAR Core rules plus five of the Standard's own — Standard 11.1 states the set and says plainly that the two lists are not added together — with all 5 responsibility sets covered by **two** people (Standard 11.1, 4.8) — Supervisor + Context Architect + Flow Manager on one, Knowledge Engineer + Verification Engineer on the other — and 4 metrics: FPSR stays, Throughput, Lead Time and Defect Escape Rate are added, and Core's Dead End Rate stops being reported (Standard 11.1) — Rule 10.4 still requires the dead ends themselves to be recorded. Suitable for 1–3 Pairs.

Two obligations arrive at this step that Core does not levy, both introduced in revision 1.4. Budget for them here rather than meeting them at a conformance assessment.

- **8.6(d), binding to a measured digest**, rises from SHOULD at Core to SHALL at Foundation (Standard 8.6, *Scaling by configuration*). A gate verdict has to name the state it was issued for by a digest measured over that state, not by a revision identifier alone. The cost is one command plus the discipline to record its result beside the verdict.
- **8.6(j), direct modification as a Gate Bypass.** Modification of an artifact within a Task's scope by a route the gate does not precede — direct editing by the Supervisor included — is admission of the effect QG-0 declares, and SHALL be recorded as a Gate Bypass carrying justification, risk acknowledgment, remediation plan and senior approval (8.6(h)). This holds at Foundation even though preventive placement, 8.6(b), is only SHOULD there: a gate that cannot interpose still declares an effect.

Step 3: Add Team ceremonies (Week 2-4) When you have 3+ Supervisors:

- **Increment Planning** — Context Architect leads: review backlog, apply WSJF prioritization, assign tasks to Pairs.
- **Quality Sweep** — Verification Engineer audits: sample recent work for architectural consistency, test quality, knowledge freshness.
- **Increment Retrospective** — Flow Manager leads: review metrics, identify process improvements, set targets.

Step 4: Add organizational metrics (Month 2+) Beyond Foundation's four (Throughput, Lead Time, FPSR, Defect Escape Rate), begin tracking: Knowledge Capture Rate (KCR), Cost per Task, Cost Predictability, Manual Intervention Rate, Cycle Time and Adversarial Detection Rate. Establish baselines for 3 Increments before setting targets.

Step 5: Add governance (when needed) Enterprise configuration (10+ Pairs): federation coordination across projects, portfolio metrics, formal audit trails and compliance mapping. QG-3, QG-4 and requirement traceability arrive earlier, at Team (Standard 11.2), and are not what distinguishes Enterprise.

Core Rules to Standard Rules Mapping

Core Rule	Standard Equivalent
1. Task Before Code	Rule 1 (S10.1) + QG-0 (S8.1)
2. Scope Boundaries	QG-0 SHALL criteria (S8.1, Task Scope) + S3.5.1
3. Verify Against Criteria	QG-2 (S8.3) + Rule 15 L2 (S10.15)
4. Tests Verify Requirements	QG-2 test criteria + QG-3 test validity (S8.4)
5. Check for Latent Defects	Rule 15 (S10.15) + AI Review Checklist (Guide)
6. Zero Tolerance for Incomplete Work	QG-2 enforcement (S8.6)
7. Fix Causes, Not Symptoms	Rule 5 Periodic Audit (S10.5) + Guide failure modes
8. Capture Knowledge	Rule 4 Dead End Documentation (S10.4) + Rule 9 Knowledge Capture (S10.9)

Role Mapping

Current Role	SENAR Responsibility	Transferable Skills	Skills to Develop
Senior Developer	Supervisor	Code review, architecture, domain knowledge, debugging	Context design, AI interaction, delegation over execution
Tech Lead	Senior Supervisor	Architecture, trade-offs, mentoring	AI-first mindset, reduced hands-on coding
QA Engineer	Verification Engineer	Verification, edge case thinking, acceptance criteria	AI output patterns, hallucination detection
DevOps Engineer	Supervisor (infra)	Automation, CI/CD, infrastructure	AI-directed config, context for infrastructure
Product Owner	Context Architect	Requirements, prioritization, stakeholder management	Structuring requirements for AI consumption, traceability
Scrum Master	Flow Manager	Facilitation, process optimization	Metrics-driven management, cost tracking

Supervisor Career Path

Level	Focus	Key Competency
Junior Supervisor	Single tasks, close verification	Following patterns, thorough review
Supervisor	Multiple tasks, architectural decisions	Context design, trade-off judgment
Senior Supervisor	Cross-project impact, mentoring	Architectural direction, reviewing others
Staff Supervisor	Process optimization, methodology	QG design, maturity assessment

Common Transition Challenges

"I'm faster typing it myself" — True for trivial tasks. For moderate+, context design + AI generation is faster than manual coding AND produces tests, documentation, and traceable output. Measure it.

"I don't trust AI output" — Good instinct. That's what QG-2 and the Review Checklist are for, and at QG-3 the five minimum criteria of Standard 8.4 are performed whether or not you ran the checklist earlier (Guide 02). Verification is your core skill now, not trust.

"My value was writing code" — Your value was solving problems. You still solve problems — but now through direction and judgment, not keystrokes. The problems get bigger.

SENAR Guide: SAFe Comparison

SENAR is a standalone methodology inspired by SAFe concepts, designed for AI-native teams. Organizations using SAFe for human-led teams can adopt SENAR for AI-native workstreams, planning for coexistence rather than overlay.

Key Differences

Aspect	SAFe	SENAR	Why Different
Production unit	Team (5-9 people)	Supervisor+AI Pair	AI replaces the team as producer
Delivery metric	Velocity (SP/sprint)	Throughput (tasks/session)	AI throughput is unpredictable; SP estimation adds overhead
Time unit	Sprint (2 weeks)	Session (hours)	AI works in bursts, not rhythms
Planning	PI Planning (2 days, full ART)	Increment Planning (1 session)	No multi-team synchronization needed
Quality mechanism	DoR/DoD (team agreements)	Quality Gates (automated code)	AI doesn't feel accountability
Knowledge	CoP, wiki, tribal knowledge	Explicit knowledge base (mandatory)	AI has no long-term memory
Retrospective	Qualitative (feelings + data)	Quantitative (metrics only)	AI work produces precise measurements
Coordination	Scrum of Scrums, ART Sync	Federation Sync	Programmatic dependency tracking
Scaling	Essential → Full → Portfolio SAFe	Core → Team → Enterprise SENAR	Same concept, different production model

Role Mapping

SAFe Role	SENAR Equivalent	Key Difference
Developer	Supervisor	Supervisor directs AI; doesn't write code as primary activity
Product Owner	Context Architect (partial)	CA designs context for AI, not backlog for humans. Business value is shared with Flow Manager
Scrum Master	Flow Manager	FM manages rhythm and cost, not team dynamics. Servant leadership less relevant with AI
QA	Verification Engineer	VE writes acceptance criteria (AI writes tests). Focus shifts to AI-specific defect patterns
RTE	Flow Manager (at Enterprise)	Enterprise FM coordinates multiple Pairs like RTE coordinates teams
System Architect	Chief Supervisor (Enterprise)	Architectural governance across all Pairs

What SAFe Has That SENAR Doesn't

SAFe Element	SENAR Approach
Communities of Practice	Knowledge base replaces human-to-human knowledge sharing with machine-readable entries
Architectural Runway	Knowledge Persistence + Dead End documentation serve similar purpose
IP Iteration	Innovation time recommended in Increment Retrospective
Built-In Quality	Quality Gates (stronger enforcement)
Solution Train / Large Solution	Enterprise configuration with Federation Coordinators (less developed)

Migration from SAFe to SENAR

Organizations moving AI workstreams from SAFe to SENAR should:

1. Start with SENAR Core on one Pair — validate minimum viable process

2. Run SAFe and SENAR in parallel for human-led and AI-led workstreams
3. Track both sets of metrics; compare delivery and quality
4. Expand SENAR to Team when 3+ Pairs are productive
5. SAFe remains for workstreams where humans write the majority of code

---

SENAR Guide: Failure Modes

SENAR can fail. Every methodology can. This document catalogs the ways SENAR implementations break down, how to detect problems early, and what to do about them.

The failure modes are organized into three categories: process failures (the methodology is adopted but practiced incorrectly), organizational failures (the people and structures around SENAR break down), and technical failures (the AI tooling and infrastructure undermines the process).

Process Failures

PF-1: Shadow Coding

Description: Supervisors secretly write code instead of directing AI Agents. The Task appears AI-generated, metrics show normal throughput, but the human is doing the implementation work. This defeats the fundamental premise of SENAR: that the Supervisor's value is judgment, not keystrokes.

Early warning signs:

- Manual Intervention Rate climbs above 30% without corresponding Dead End entries explaining why AI couldn't do the work. Since 1.4 this rate is derived from Gate Bypasses recorded under 8.6(j) (Standard 9.2), which fixes the boundary normatively and puts a senior approval behind each record. It does **not** make the figure tamper-proof: 9.2 states outright that the metric measures what was recorded, that an unrecorded modification does not appear in it, and that nothing detects one automatically. A low MIR is therefore not evidence against shadow coding — the audit in Recovery action 1 below is what produces that evidence, and this failure mode is the reason to run it. Since 1.5 the Standard also requires a target for this metric with its basis (Standard 9.4(c)) and escalation when the figure crosses it (Standard 9.4(d)): a bypass rate that normalizes looks, from outside, exactly like a target quietly moved
- Commit timestamps cluster outside of session windows
- Supervisors report feeling "faster without the AI" but can't articulate what the AI struggled with

- Knowledge base has no Dead End or Gotcha entries — everything "just works" (because the human did it)
- Session tool call counts are suspiciously low relative to task complexity

Recovery action:

1. Conduct a non-punitive audit: for a sample of Tasks, compare the modifications in git against the gate's admission records (Standard 8.6(g)) — a modification with no admission behind it is the finding. Session logs are not the reference: Standard 8.6(j) sets the boundary by the origin of the content, so a change an agent wrote to dictation sits inside the session log and still arrived by a route no verdict covered
2. Identify which task types trigger shadow coding — these likely have poor context templates
3. Improve context quality for those task types (better AC, patterns, architectural constraints)
4. If AI genuinely cannot handle certain task types, create an explicit "manual" work type with its own tracking

Prevention:

- Normalize manual intervention as a documented, tracked activity — not a shameful secret. Since 1.4 the documentation is not optional: it is the Gate Bypass record required by 8.6(j)
- Track Manual Intervention Rate as a diagnostic metric, not a performance metric
- Ensure Supervisors understand that high MIR on certain task types is a signal to improve context, not a personal failure
- Session logs should be reviewed in Quality Sweeps — not to police, but to identify context gaps

---

PF-2: Rubber Stamping

Description: Supervisors accept AI output without meaningful verification. Code passes QG-2 (automated tests, lint, types) but nobody checks whether it actually solves the problem correctly, handles edge cases, or fits the architecture. The human becomes a button-presser.

Early warning signs:

- Defect Escape Rate rises while First-Pass Success Rate stays suspiciously high (everything "passes" but bugs appear later)
- Code review comments are absent or perfunctory ("LGTM", "looks good")
- Verification time per task drops below plausible minimums (a 200-line change reviewed in 30 seconds)

- Architectural drift: patterns diverge across the codebase because nobody enforces consistency
- Security vulnerabilities appear in production that a human reviewer should have caught

Recovery action:

1. Mandatory Quality Sweep focused on the last 2 Increments — compare merged code against requirements
2. Introduce QG-3 (Verification Gate) if not already active: independent review by a different Supervisor
3. Establish minimum review time thresholds proportional to change size
4. Create an "AI Review Checklist" (see Guide 02) and require it to be completed per task

Prevention:

- QG-3 with independent verification is the strongest safeguard
- Rotate Verification Engineer assignments so reviewers stay sharp
- Track Defect Escape Rate prominently — it is the primary indicator of rubber stamping
- Quality Sweeps must sample actual code, not just metrics

PF-3: Gate Bypass Normalization

Description: Quality Gates exist but are routinely bypassed with documented exceptions. "We'll fix it later" becomes the default. Gate bypass was designed for genuine emergencies; when it becomes normal, the gates are decorative.

> ***Read the bypass rate as two numbers, not one (revision 1.4).***
>
> *Since 8.6(j), direct modification of an artifact is a recorded Gate Bypass. A team that codes by hand for the reasons Guide 01 calls appropriate will therefore accumulate bypass records while doing nothing wrong, and a single blended rate will cross any threshold set before 1.4.*
>
> *Separate them before judging:*
>
> - ***8.6(j) bypasses** — the Manual Intervention Rate (Standard 9.2, Metric 8). High is a signal about context quality and agent capability, not about gate discipline. Diagnose it as PF-1, below.*
> - ***Criteria bypasses** — a gate's criteria were unmet and the effect was admitted anyway. This is what PF-3 is about, and the thresholds below apply to this number alone.*
>
> *The two are nested and Standard 9.2 forbids adding them together. A team whose criteria-bypass rate is near zero and whose 8.6(j) rate is 30% has a context problem, not a decorative gate.*

Early warning signs:

- More than 10% of Tasks have criteria-bypass entries (bypasses under 8.6(j) are excluded — see above)
- The same gate is bypassed repeatedly (e.g., QG-0 skipped because "we already know what to do")
- Bypass justifications become formulaic ("time pressure", "will address in next increment")
- Tech debt tickets created from bypasses are never completed
- New team members learn to bypass gates as standard practice

Recovery action:

1. Audit all bypass entries from the last 2 Increments — categorize by gate and justification
2. For each repeatedly-bypassed gate, determine: is the gate too strict (adjust criteria) or is the team undertrained (invest in process)?
3. Require bypass approval from someone other than the Supervisor doing the work
4. Create a "bypass budget" per Increment — a maximum number of allowed bypasses. When exhausted, the gate is hard-blocked

Prevention:

- Track bypass rate as a metric reviewed at Increment Retrospectives

- Retrospective must include a "bypass review" — every bypass is discussed, not just counted
- Gate criteria should be reviewed quarterly: if a gate is routinely bypassed, it may need adjustment rather than enforcement
- Distinguish between "gate too strict" (criteria problem) and "discipline too low" (culture problem) — the responses are different

---

PF-4: Knowledge Base Staleness

Description: The Knowledge Base was populated during initial adoption but is no longer maintained. Entries refer to old patterns, deprecated APIs, or architectural decisions that have been superseded. AI Agents receive stale context and produce output based on outdated information.

Early warning signs:

- Knowledge Capture Rate drops below 0.3 entries per task for 3+ consecutive sessions
- New Supervisors report that KB entries contradict what they see in the codebase
- AI Agents produce output using patterns documented in KB that the team has since abandoned
- "Last reviewed" dates on KB entries are older than 3 Increments
- Dead End entries reference tools or libraries that are no longer in the stack

Recovery action:

1. Schedule a dedicated KB audit session — treat it as a Task, not a side activity
2. Mark every entry with a freshness status: current, needs-review, deprecated
3. Delete or archive deprecated entries immediately — stale context is worse than no context
4. Assign KB maintenance as an explicit responsibility (Knowledge Engineer role)

Prevention:

- Quality Sweep ceremony (Standard, Section 7.4) should include KB freshness audit
- Knowledge entries should have a "last verified" timestamp, updated when someone confirms the entry is still accurate
- Set a target Knowledge Capture Rate (0.3-0.5 entries/task) and review at Retrospectives
- When a Task produces output that contradicts a KB entry, the entry must be updated before the Task is marked done

PF-5: Dead End Documentation Stops

Description: Teams stop documenting failed approaches. Dead Ends are one of the highest-value knowledge types — they prevent AI Agents from repeating mistakes. When teams are under pressure, Dead End documentation is the first thing dropped because it documents what didn't work, which feels low-priority.

Early warning signs:

- Dead End entries per Increment drop to zero
- AI Agents attempt approaches that were already tried and abandoned in previous sessions
- Session durations increase because the AI is exploring paths that a human remembers failing but didn't document
- Handoff documents mention "we tried X but it didn't work" without creating a KB entry

Recovery action:

1. Review the last 5 session handoffs for mentions of failed approaches — create Dead End entries for each
2. Add "Dead End capture" as an explicit step in the Session End ceremony
3. Make Dead End documentation part of QG-2 acceptance criteria for complex tasks

Prevention:

- Session End template should include a "Failed approaches" section that becomes a Dead End entry
- Knowledge Capture Rate metric should be decomposed by entry type — a healthy KB has a mix of decisions, patterns, gotchas, and dead ends
- Celebrate Dead End documentation: it is not a record of failure, it is a map of the territory

---

PF-6: Session Discipline Erosion

Description: Sessions stretch beyond reasonable limits. Supervisors skip Session Start (no context loading, no task selection) or Session End (no handoff, no metrics capture). The session boundary — SENAR's primary rhythm mechanism — dissolves into continuous, unstructured work.

Early warning signs:

- Session durations exceed 8 hours regularly
- Session Start entries have no task list or context verification
- Handoff documents are empty or missing
- Checkpoints are never taken
- Tool call counts per session exceed 500 (fatigue territory)
- Quality of AI output degrades in the second half of sessions (context window exhaustion)

Recovery action:

1. Enforce hard session time limits — 4-6 hours is the recommended range
2. Make Session End a blocking step: the next session cannot start without a completed handoff
3. Implement automatic checkpoint reminders at 40-60 tool call intervals
4. Review the last 10 sessions for missing handoffs and create them retroactively

Prevention:

- Session Start and Session End are ceremonies, not optional steps
- Tooling should enforce: session cannot start if previous session has no handoff
- Set a maximum tool call budget per session (300-500 depending on task complexity)
- Supervisor fatigue is real: verification quality degrades after 4-6 hours. Process must acknowledge human limits

---

PF-7: Vanity Metrics

Description: Teams optimize metrics without improving outcomes. Throughput is inflated by splitting tasks into trivially small units. First-Pass Success Rate is gamed by lowering acceptance criteria. Lead Time is reduced by starting tasks before they're ready. The numbers look good; the software does not.

Early warning signs:

- Throughput increases dramatically but stakeholder satisfaction doesn't
- Average task complexity drops to "trivial" for 80%+ of tasks
- FPSR is above 95% but Defect Escape Rate is also above 10% (contradictory signals)
- Lead Time decreases but deployed features are incomplete or buggy
- Teams resist adding new metrics or changing existing ones
- Cost per Task drops but total Increment cost doesn't

Recovery action:

1. Cross-reference metrics: Throughput x DER, FPSR x post-deployment defects, Lead Time x rework rate
2. Introduce "outcome metrics" alongside process metrics: deployment frequency, change failure rate, time to restore
3. Redefine task granularity guidelines — a task should represent a meaningful, verifiable change
4. Retrospective should compare metrics against actual stakeholder outcomes

Prevention:

- Never reward or punish based on a single metric
- Metrics exist to diagnose, not to judge — make this explicit in team norms
- DER is the hardest metric to game because it's measured by external reality (production bugs)
- Quality Sweeps should validate that metrics tell a consistent story

---

Organizational Failures

OF-1: Supervisor Burnout

Description: Constant verification of AI output creates cognitive fatigue. The Supervisor must maintain deep technical understanding while processing large volumes of generated code. Unlike traditional development where the developer builds context incrementally, the Supervisor must rapidly context-switch between tasks and verify unfamiliar code patterns.

Early warning signs:

- Verification quality drops in afternoon sessions (rubber stamping after lunch)
- Supervisors report feeling "drained" despite not writing code
- Increased use of gate bypasses in later sessions of the week
- Sick days and turnover increase
- Supervisors start preferring simple tasks and avoiding complex ones

Recovery action:

1. Reduce session length to 3-4 hours with mandatory breaks
2. Rotate between deep verification work and lighter tasks (documentation, knowledge curation)
3. Implement pair supervision: two Supervisors verify each other's complex tasks
4. Ensure Supervisors have autonomy in task selection and session planning

Prevention:

- Design sessions with verification load in mind — not all tasks require the same level of scrutiny
- Budget time for non-verification work: KB maintenance, context improvement, learning
- Track tool calls and session duration — these are leading indicators of burnout
- Respect the difference between "AI can work 24 hours" and "humans cannot verify for 24 hours"

OF-2: Skills Atrophy

Description: Supervisors stop writing code entirely and gradually lose the ability to evaluate AI output at a deep technical level. They can spot surface-level issues (typos, obvious bugs) but miss architectural problems, performance anti-patterns, and subtle security issues. The "judgment over keystrokes" value requires that judgment remains sharp.

Early warning signs:

- Supervisors struggle to explain why a particular implementation approach is wrong
- Architectural decisions are increasingly delegated to the AI with no human challenge
- Code reviews become "does it work?" rather than "is this the right approach?"
- Supervisors avoid tasks in unfamiliar parts of the codebase
- New technologies are adopted without Supervisor understanding of tradeoffs

Recovery action:

1. Institute "manual implementation days" — periodic sessions where Supervisors write code without AI assistance. Keep this work outside any Task's scope (a kata, a spike, a throwaway branch): 8.6(j) applies to modification of an artifact *within a Task's scope*, so practice that produces nothing the project ships needs no bypass record. Where a manual day produces work the project keeps, it does, and the record is written like any other
2. Require Supervisors to write the architectural approach before the AI implements it
3. Rotate Supervisors across different parts of the codebase to force learning
4. Include "explain the implementation" as part of verification: if you can't explain it, you can't verify it

Prevention:

- Supervisors should occasionally implement complex tasks manually to stay sharp

- Technical learning time should be budgeted — not as overhead, but as capability maintenance
- The Context Architect role keeps Supervisors engaged in system design, not just task verification
- Quality Sweeps that require deep code understanding serve double duty: quality assurance and skill maintenance

---

OF-3: AI Tool Vendor Lock-in

Description: The SENAR implementation becomes deeply coupled to a specific AI tool, model, or platform. Session templates, knowledge base format, context injection mechanisms, and verification workflows all assume a particular vendor's API and capabilities. Switching becomes prohibitively expensive.

Early warning signs:

- All context templates use vendor-specific syntax or features
- Knowledge base format is optimized for one model's context window size
- Session management is embedded in a proprietary platform with no export
- Cost tracking depends on vendor-specific billing APIs
- Team expertise is in "using Tool X" rather than "supervising AI agents"

Recovery action:

1. Audit all touchpoints between SENAR process and AI tooling — create an abstraction inventory
2. Define a vendor-neutral interface for context injection, output capture, and session management
3. Extract knowledge base into a portable format (structured markdown, JSON)
4. Run parallel sessions with alternative tools to validate portability

Prevention:

- SENAR deliberately does not prescribe specific AI tools — maintain this neutrality
- Keep knowledge base in a format readable by any LLM (structured text, not proprietary embeddings)
- Session management should be a thin layer, not a platform
- Periodically evaluate alternative tools to maintain optionality

---

OF-4: Over-Process (Bureaucratic SENAR)

Description: SENAR becomes an end in itself. Every task requires exhaustive documentation. Every session has a 30-minute startup ceremony. Every knowledge entry needs three approvals. The methodology that was designed to increase throughput becomes the primary obstacle to throughput.

Early warning signs:

- Session Start takes longer than 30 minutes
- More time is spent on SENAR ceremonies than on actual work
- Supervisors create "process compliance" tasks that produce no software value
- New Supervisors are overwhelmed by the number of required steps
- Teams maintain two systems: the "official" SENAR process and their actual workflow

Recovery action:

1. Measure "process overhead ratio": time on ceremonies / time on productive work. If it exceeds 20%, the process is too heavy
2. Revisit which SENAR level (Core/Team/Enterprise) actually matches the team's needs — many teams over-adopt
3. Automate everything that can be automated — manual ceremony steps should be minimized
4. Apply SENAR's own principle: "Enforcement over Ceremony" — if a step can be automated, it shouldn't be a manual ceremony

Prevention:

- Start with SENAR Core and upgrade only when specific pain points justify additional process
- Every ceremony should have a defined timebox
- Every manual step should have a justification: "what failure does this prevent?" If the answer is unclear, remove the step
- Periodically run a "process audit" — are all steps still earning their cost?

OF-5: Under-Process (Perpetual L2)

Description: The team adopts SENAR Core (Maturity Level 2) and never progresses. Tasks exist, Start Gate and Done Gate are enforced, Core's FPSR and Dead End Rate are tracked — but there's no independent verification, no knowledge capture discipline, no federation coordination. The team gets the minimum benefit and plateaus.

Early warning signs:

- Same process for 6+ months with no evolution
- Knowledge base growth has flatlined
- DER is stable but never improves
- Cross-project dependencies are managed ad-hoc (email, chat) rather than through federation
- No Increment Retrospectives — metrics are collected but never reviewed
- Team says "SENAR is working fine" but can't point to specific improvements

Recovery action:

1. Run a maturity assessment across all 6 dimensions (Section 12.2) — identify the weakest
2. Pick ONE dimension to improve in the next Increment — don't try to jump from L2 to L3 across all dimensions simultaneously
3. Assign improvement ownership to a specific person (Flow Manager responsibility)
4. Set a measurable target: "Reduce DER from 12% to 8% by end of Increment"

Prevention:

- Include "process improvement" as a standing item in Increment Planning
- The Flow Manager role exists precisely to drive process evolution — ensure it's staffed
- Set explicit maturity targets with timelines
- Review the SENAR Maturity Model (Section 12) at each Retrospective

---

OF-6: Supervisor Resistance

Description: Developers refuse to adopt the Supervisor role. They see SENAR as de-skilling: taking away the creative, satisfying work of writing code and replacing it with bureaucratic verification of machine output. The resistance may be quiet (passive non-compliance) or loud (active pushback).

Early warning signs:

- High shadow coding rates (PF-1) — developers code and retroactively create tasks to match
- Supervisors describe themselves as "reviewers" or "testers" rather than "engineers"
- Hiring for Supervisor roles is difficult — candidates want "real engineering" positions
- Team morale surveys show dissatisfaction with role definition
- Experienced developers leave for traditional development roles

Recovery action:

1. Address the identity question directly: Supervisor is a more senior role than Developer, not a lesser one
2. Highlight the aspects of the role that require deeper skill: architectural judgment, context design, verification of complex systems
3. Allow Supervisors to choose when to code manually — autonomy reduces resistance. The choice is theirs; the record under 8.6(j) is not optional, and presenting it as a bureaucratic tax rather than as what makes the autonomy defensible is how this recovery action turns into PF-1
4. Share concrete examples of how the Supervisor role amplifies individual impact (throughput multiplied by AI)

Prevention:

- Frame the transition as career progression, not role diminishment
- Ensure compensation reflects the increased scope and responsibility of the Supervisor role
- Provide the Transition Guide (Guide 04) and allow adequate adjustment time
- Create communities of practice where Supervisors share techniques and celebrate wins

---

OF-7: Knowledge Silos

Description: Each Supervisor develops their own undocumented patterns for context preparation, task structuring, and AI interaction. These patterns are effective but exist only in the Supervisor's head. When they're unavailable, other Supervisors can't replicate their results.

Early warning signs:

- Throughput varies dramatically between Supervisors on similar task types
- Knowledge base entries are sparse or generic — the "real" context is in personal notes
- Supervisors can't cover for each other during absences
- Handoff documents don't capture the "tricks" that make a session productive
- New Supervisors take months to reach baseline productivity

Recovery action:

1. Run "context pairing" sessions: have high-performing Supervisors work alongside others, capturing their undocumented patterns
2. Convert personal notes and templates into shared knowledge entries

3. Standardize context templates for common task types
4. Include "knowledge sharing" as an explicit goal in Quality Sweeps

Prevention:

- Knowledge Capture Rate metric should be reviewed per Supervisor, not just in aggregate
- Context templates should be shared artifacts, not personal files
- Session handoffs should include "what worked well" for context preparation, not just task status
- The Knowledge Engineer role exists to prevent silos — ensure it's actively practiced

---

Technical Failures

TF-1: Model Change Breaks Baselines

Description: The AI model is updated (new version, different provider, or fine-tuning change) and all established baselines become invalid. Throughput, FPSR, Lead Time, and Cost per Task all shift unpredictably. The team loses its ability to detect process problems because the signal is buried in model-change noise.

Early warning signs:

- Sudden shift in all metrics after a model update
- Tasks that previously had high FPSR start failing
- Cost per Task changes significantly (often in both directions for different task types)
- Context templates that worked well now produce worse output
- AI behavior changes subtly: different coding style, different library preferences, different error handling patterns

Recovery action:

1. Treat a model change as a "baseline reset event" — document the change and mark metrics before/after
2. Run a sample of representative tasks on the new model to establish new baselines
3. Update context templates and knowledge entries that assumed specific model behavior
4. Allow 1-2 sessions of reduced throughput expectation while the team calibrates

Prevention:

- Pin model versions in production — don't auto-update

- When evaluating a new model, run a parallel evaluation on a representative task set before switching
- Keep context templates model-neutral: describe what you want, not how the model should produce it
- Maintain a "model change log" that records which model version was used for each Increment's baselines

---

TF-2: Context Window Limitations

Description: The AI's context window is finite. As projects grow, the context required for a task (codebase knowledge, architectural constraints, related patterns, previous decisions) exceeds what fits in a single session. Tasks are artificially split not because they're naturally separable, but because the AI can't hold enough context.

Early warning signs:

- Tasks are split into sub-tasks that don't make sense as independent units
- AI "forgets" instructions given earlier in the session
- Late-session output contradicts early-session decisions
- Supervisors spend increasing time re-explaining context after checkpoints
- Complex refactoring tasks fail because the AI can't see enough of the codebase simultaneously

Recovery action:

1. Restructure context injection: prioritize high-value context (architectural constraints, active patterns) over low-value context (full file contents)
2. Use knowledge base entries as compressed context: a 5-line decision entry replaces reading 500 lines of code
3. Implement progressive context loading: start with architecture overview, load details as needed
4. Accept that some tasks require multiple sessions — design handoff documents to carry context across the boundary

Prevention:

- Knowledge base entries are context compression tools — invest in high-quality, concise entries
- Architectural documentation should be structured for AI consumption: clear, concise, with explicit constraints
- Design task granularity around natural boundaries, not context window size

- Monitor context window utilization and establish team conventions for context budget allocation

---

TF-3: Architecturally Wrong but Gate-Passing Code

Description: AI generates code that passes all automated quality gates (tests pass, types check, linter clean, security scan clear) but is architecturally wrong. It might use the wrong pattern, create inappropriate coupling, bypass established abstractions, or introduce a dependency that will cause problems in 6 months. Automated gates verify correctness; they cannot verify judgment.

Early warning signs:

- Code reviews increasingly find "it works but it's wrong" issues
- Similar functionality is implemented differently in different parts of the codebase
- Established abstractions are bypassed — AI creates new patterns instead of using existing ones
- Dependency graph grows in unexpected directions
- Refactoring costs increase over time as architectural drift accumulates

Recovery action:

1. Strengthen the QG-3 Verification Gate with explicit architectural review criteria
2. Create "architectural constraint" knowledge entries that are injected into every session context
3. Run an architectural health audit — identify and document all established patterns
4. Consider architecture-specific linting rules that can be automated (dependency constraints, layer violations)

Prevention:

- Architectural decisions must be documented in the knowledge base, not just in the code
- Context templates for each area of the codebase should include "use this pattern, not that pattern" instructions
- Quality Sweeps should include architectural consistency checks
- The Context Architect role is critical here: they ensure AI receives the right architectural constraints
- Automated architectural fitness functions (dependency rules, layer checks) convert judgment into enforcement

---

TF-4: Phantom Dependencies

Description: Across sessions, AI agents introduce dependencies (libraries, services, API calls, shared state) that are not explicitly tracked. Each dependency is individually reasonable, but the aggregate creates a fragile system. No single person or document captures the full dependency graph.

Early warning signs:

- Build times increase without obvious cause
- Package manifest (package.json, requirements.txt, go.mod) grows steadily
- "Works on my machine" issues increase — environment setup is unreliable
- Upgrading one dependency triggers cascading failures
- Nobody can explain why a particular dependency was added
- Service coupling increases: changes in one service require changes in others

Recovery action:

1. Run a dependency audit: for every dependency, find the Task that introduced it and verify it's still needed
2. Create a "dependency decision" knowledge entry type — every new dependency requires a documented justification
3. Implement dependency change detection in CI — any new dependency triggers a review flag
4. Prune unused dependencies

Prevention:

- Add dependency justification to QG-2 acceptance criteria: "no new dependencies without documented reason"
- Track dependency count as a secondary metric
- Quality Sweeps should include dependency review
- Context templates should list approved dependencies and their purposes — AI will prefer what it knows about

---

TF-5: Meaningless Test Coverage

Description: AI generates tests that achieve high coverage metrics but don't actually verify meaningful behavior. Tests mirror the implementation: they test that the code does what the code does, not that the code does what the requirements say. This is particularly insidious because the coverage number looks excellent.

Early warning signs:

- Test coverage is above 90% but Defect Escape Rate is also high
- Tests break when implementation changes but not when behavior changes (brittle, tightly coupled tests)
- Test descriptions are generic: "should work correctly", "handles input"
- Tests don't cover edge cases, error paths, or boundary conditions
- Mutation testing kills few mutants despite high line coverage
- Integration tests are actually unit tests with mocked-out dependencies

Recovery action:

1. Run mutation testing to evaluate actual test effectiveness — coverage means nothing without kill rate
2. Review test quality in Quality Sweeps: do tests verify requirements or mirror implementation?
3. Require acceptance criteria to include specific test scenarios — AI writes tests from scenarios, not from code
4. Introduce property-based testing for critical paths

Prevention:

- Acceptance criteria should define test cases: "given X, when Y, then Z" — not "write tests for this function"
- Track mutation testing score alongside coverage
- QG-2 should include test quality checks, not just coverage thresholds
- Knowledge base should contain testing patterns: "how we test X-type functionality"
- Separate the test-writing agent prompt from the implementation agent prompt — tests should verify requirements, not mirror code

---

Summary

No failure mode in this document is theoretical. Every one has been observed in practice, either in AI-native development or in analogous patterns from traditional Agile/SAFe adoption.

The common thread across all failure modes is the same: **when the uncomfortable parts of the process are skipped, the process stops working.** Verification is uncomfortable. Documentation is tedious. Dead Ends feel like wasted time. Gate compliance feels like bureaucracy. But these are the load-bearing elements of SENAR — remove them and the structure collapses.

The best defense is not more process but better feedback loops:

1. **Metrics that cross-reference** — no single metric tells the truth alone
2. **Quality Sweeps that sample reality** — not just metrics dashboards but actual code and actual knowledge entries
3. **Retrospectives that ask hard questions** — not "are our numbers good?" but "is our software good?"
4. **A culture where documenting failure is valued** — Dead Ends and failure mode awareness prevent repeated mistakes

---

SENAR Guide: Requirements Engineering for AI-Native Development

Why Requirements Matter More with AI

In traditional development, a vague requirement leads to a conversation: "Did you mean X or Y?" The developer asks, clarifies, and implements correctly.

With AI, a vague requirement leads to a confident, plausible implementation of the wrong thing. AI does not ask clarifying questions — it picks an interpretation silently. The resulting code compiles, passes the tests the AI wrote (which test the wrong behavior), and looks correct on review.

The cascade principle: A defect at the requirement level is the most expensive defect. It produces correct-looking code that solves the wrong problem, correct-looking tests that verify the wrong behavior, and correct-looking documentation that describes the wrong system. Every downstream artifact inherits the original defect.

This is why SENAR's first value is "Context over Code" — and requirements are the most important context.

The Three Levels

SENAR defines three requirement levels. Not all are needed for every task — depth scales with complexity and team size.

BR — Business Requirement

What: A stakeholder-level need expressed in business terms. **Who writes it:** Stakeholder, Product Owner, or Context Architect. **Where it lives:** Increment objective, Epic goal, or dedicated BR entry in the knowledge base.

Examples:

- "Users must be able to reset their password without contacting support"
- "The system must process orders within 30 seconds during peak load"
- "API must support OAuth 2.0 for third-party integrations"

Properties: A good BR is:

- Business-oriented (no implementation details)
- Verifiable (you can determine if the system satisfies it)
- Independent (achievable without solving another BR first, or dependency is explicit)

SR — System Requirement

What: A system-level capability derived from a BR. Describes what the system must do, not how. **Who writes it:** Context Architect (Team+) or Supervisor (Core). **Where it lives:** Story goal, or dedicated SR entry linked to parent BR.

Examples (derived from BR "password reset"):

- "POST /auth/reset-password sends a reset link to the user's email"
- "Reset link expires after 1 hour"
- "User can set a new password using a valid reset token"
- "System logs all password reset attempts for security audit"

Properties: A good SR is:

- Specific enough to decompose into Tasks
- Traceable to a parent BR
- Testable at the system level (integration test or E2E test)

TR — Task Requirement

What: An implementation-level requirement. Corresponds directly to a Task's goal and acceptance criteria. **Who writes it:** Supervisor. **Where it lives:** Task goal + acceptance criteria fields.

Example (derived from SR "POST /auth/reset-password"):

- **Goal:** Implement POST /auth/reset-password endpoint
- **AC:**
    1. Returns 200 and sends email for existing user
    2. Returns 200 for non-existing user (no information leak)
    3. Returns 422 if email field missing
    4. Reset token is cryptographically random, 32 bytes
    5. Token stored hashed in database (not plaintext)

6. Rate limited to 3 requests per email per hour

Properties: A good TR (= good acceptance criteria) is:

- Independently verifiable (each AC can be tested separately)
- Includes at least one negative scenario (error case)
- Specifies behavior, not implementation ("returns 422" not "use Pydantic validator")

Test Model (TM) — the Verification Bridge

The Test Model is NOT a fourth requirement level. It is a verification artifact derived from Task Requirements.

```
BR → SR → TR → [TM] → Code + Tests
                 ↑ requirement    ↑ verification artifact
```

What a Test Model contains:

- Test cases for each AC (what to test)
- Test data (boundary values, error inputs)
- Expected results (what "pass" looks like)
- Verification method (automated unit test, integration test, manual demo, measurement)

In AI-native development, AI typically generates tests alongside code. The danger: AI generates tests that pass but don't actually exercise the stated AC. The Test Model is the Supervisor's check that generated tests match the requirements, not just achieve coverage.

Configuration	Test Model Discipline
Core	Implicit — Supervisor reviews generated tests against AC
Foundation	Implicit — same as Core; combined Supervisor role performs AC review
Team	SHOULD be explicitly reviewed — each AC has a corresponding test
Enterprise/Regulated	SHALL be documented, traceable to TRs, independently verifiable

Why not a requirement level? Test cases describe *how to verify*, not *what the system must do*. This follows ISO/IEC 29119 (software testing) which separates test design from requirements. Making TM a requirement level would force Core users to manage 4 levels — overhead that kills adoption.

Requirement Quality Properties

These properties are the criteria of QG-1 (Standard 8.2), and Standard 11.1 places QG-1 at Team and above. Below Team the gate is absent, so four of the five have no level there — not a weaker one. Verifiability is the exception and is owed at every configuration, but through a different clause: QG-0 requires verifiable acceptance criteria (Standard 8.1), and Core requires the same at its Start Gate. Reusability is this Guide's own suggestion; the Standard does not state it.

Property	Below Team	Team+ (QG-1)	Where the Standard states it
Verifiability	owed, via QG-0	owed	Standard 8.1 and 8.2
Consistency	gate absent	owed	Standard 8.2
Sufficiency	gate absent	owed	Standard 8.2
Non-redundancy	gate absent	owed	Standard 8.2
Traceability	gate absent	owed	Standard 8.2
Reusability	—	—	not stated; this Guide's suggestion

Verifiability in Practice

Bad: "The system should be fast" — fast by whose standard? Good: "API response time < 200ms at p95 under 100 concurrent users"

Bad: "Error handling should be robust" — what does robust mean? Good: "All API endpoints return structured error responses with HTTP status code, error code, and human-readable message"

Bad: "The UI should be user-friendly" — unmeasurable. Good: "A new user can complete the registration flow in under 2 minutes without external help"

AC Templates for Common Tasks

REST API Endpoint

1. Returns expected status code and body for valid input
2. Returns 401/403 for unauthorized/forbidden access

3. Returns 422 with descriptive error for invalid input
4. Returns 404 for non-existent resource
5. Response matches documented schema
6. Rate limiting applied per configuration

Database Migration

1. Migration is idempotent (safe to run multiple times)
2. Rollback migration exists and works
3. No data loss during migration
4. Migration completes within acceptable time for production data volume
5. Indexes created for new query patterns

UI Component

1. Renders correctly in target viewport range (320–1920px)
2. Accessible: keyboard navigable, screen reader compatible (WCAG AA)
3. Handles loading, empty, and error states
4. Responsive to theme/dark mode if applicable

Integration / External API

1. Handles success response correctly
2. Handles timeout (configurable, default 5s)
3. Handles error response with structured error
4. Retry policy for transient failures
5. Credentials not hardcoded (env/config)

Infrastructure / DevOps

1. Configuration change is reversible
2. Health check endpoint updated if applicable
3. Monitoring/alerting covers the change
4. Documentation updated (runbook, architecture diagram)

Anti-Patterns

Specification Vacuum

Problem: Task has a goal but no acceptance criteria. AI produces something that "looks right" but nobody can verify it. **Fix:** Every Task SHALL have acceptance criteria (QG-0). No AC = no start.

Over-Specification

Problem: 20 acceptance criteria for a trivial task. Overhead exceeds value. **Fix:** Match AC depth to task complexity. Trivial: 2–3 AC. Complex: 5–8 AC. If you need 15+ AC, the task should be split.

Copy-Paste AC

Problem: Same AC copied across tasks without adaptation. "Returns 200" everywhere, but some endpoints should return 201 or 204. **Fix:** Use templates as starting points, then customize. Templates save time; blind copying introduces bugs.

Untraceable Requirements

Problem: Tasks exist without any link to why they exist. When priorities change, nobody knows which tasks to drop. **Fix:** Every Task links to a Story or BR (Rule 9.10). If you can't explain why a task exists, it probably shouldn't.

Ambiguous Criteria

Problem: "System should handle edge cases properly." Which edge cases? What is "properly"? **Fix:** Name the specific edge cases. "System returns 409 when duplicate email. System returns 413 when file exceeds 10MB. System returns 429 when rate limit exceeded."

Scaling by Configuration

Aspect	Core (1–2 Pairs)	Foundation (1–3 Pairs)	Team (3–10 Pairs)	Enterprise (10+)
Hierarchy depth	BR → TR (2 levels)	BR → TR (2 levels)	BR → SR → TR (3 levels)	Full + test derivation
Who manages	Supervisor	Supervisor (combined roles)	Context Architect	Context Architect + review
Storage	Task fields	Task fields + KB entries	Knowledge Base	Versioned + auditable
Reuse	Informal patterns	Informal patterns	KB templates	Cross-project library
Traceability	SHOULD	SHOULD	SHALL	SHALL + audit trail
Quality review	Self-review	Monthly Quality Sweep	QG-1	QG-1 + independent review
Overhead per task	~1 min (write AC)	~1–2 min (write AC + log)	~3 min (decompose + link)	~5 min (full traceability)
Test Model	Implicit (review tests vs AC)	Implicit (review tests vs AC)	Explicit review	Formal TM documentation

---

Requirements-as-Code (Enterprise)

At Enterprise scale, requirements are managed with the same discipline as code: versioned, reviewed, tested by CI.

What This Means

```
requirements/
  BR-001-oauth.md          # Business Requirement
  SR-001-google-oauth.md   # System Requirement (links to BR-001)
  SR-002-token-refresh.md  # System Requirement (links to BR-001)
```

Each requirement file contains:

- Unique ID and title
- Level (BR/SR)
- Parent link (SR → BR)
- Status (draft/approved/verified/deprecated)
- Version history (git log)
- Child links (BR → list of SRs, SR → list of Task slugs)

Task Requirements (TR) remain in the task tracker as goal + AC — they don't need separate files because they ARE the task.

CI Checks for Requirements

Check	What It Catches
Orphan detection	BRs with no child SRs; SRs with no Tasks
Traceability validation	Tasks without requirement links; broken parent references
Status consistency	Tasks referencing `deprecated` requirements
Coverage report	% of BRs fully decomposed and implemented
Change impact	When a BR changes, list all affected SRs and Tasks

Requirement Library and Reuse

Verified requirements (status: `verified`, used successfully in N projects) become library entries:

```
library/
  patterns/
    rest-api-endpoint.md    # Template: AC for any REST endpoint
    db-migration.md         # Template: AC for any migration
    auth-integration.md     # Template: AC for auth flows
```

A new project imports a library pattern, parameterizes it, and gets proven AC that have been tested across multiple implementations. This directly improves First-Pass Success Rate — AI receives battle-tested context instead of first-draft requirements.

When to Adopt

Requirements-as-code adds overhead. It pays off when:

- 10+ Supervisor+AI Pairs share requirements across projects
- Regulatory compliance requires audit trails (ISO 9001, ASPICE, medical devices)
- Cross-project requirement reuse is a strategic goal
- Requirement changes are frequent and impact analysis is needed

For SENAR Core and most Team configurations, requirements in the task tracker and Knowledge Base are sufficient.

SENAR Guide: Adopting SENAR in Legacy Codebases

Most software is not greenfield. You have 200,000 lines of undocumented code, and Rule 9.11 says "documentation sufficient for AI to understand module purpose." You are not going to document everything before starting. Here's how to adopt SENAR incrementally.

Principle: Next Task, Not Full Retrofit

You don't need to "SENAR everything." Start with the next task. When that task touches a module, document that module. When the next task touches another module, document that one. Over time, the documentation frontier advances with actual work.

What NOT to do: Don't create a "documentation sprint" that takes 3 weeks and produces docs nobody reads. Documentation written in isolation from implementation is outdated before it's finished.

Phase 1: Start with Habits (Day 1)

Adopt the 6 Quick Start habits immediately. They require zero codebase preparation:

1. Write goal + AC before each AI task (even in legacy code)
2. Set scope boundaries ("change ONLY this file, don't refactor the module")
3. Verify against AC, not intuition
4. Document dead ends (especially important in legacy — "tried X, failed because Y")
5. Run tests
6. Capture knowledge

Scope boundaries (habit 2) are critical for legacy: AI agents will try to "improve" surrounding code if you don't fence them.

Phase 2: Document on Contact (Week 1+)

Every time a task touches a module, add minimum documentation:

```
"""
Module: user_auth
Purpose: Handles login, registration, and session management.
Public API:
  - login(email, password) -> Session
  - register(email, password, name) -> User
  - verify_session(token) -> User | None
Dependencies: database (PostgreSQL), redis (session store)
Boundaries: Does NOT handle OAuth (see oauth_provider module).
"""
```

This takes 5 minutes per module. It satisfies Rule 9.11 at the SENAR Core level. From this point forward, AI tasks touching this module need less explicit context in the Task goal — the docstring provides it.

Key: Write for AI, not for humans. The AI doesn't care about your design philosophy. It needs: what this module does, what the public interface is, what it connects to, and what it does NOT do.

Phase 3: Build Knowledge Base from Tribal Knowledge (Week 2+)

Legacy codebases have tribal knowledge — things people know but haven't written down. As you encounter these during SENAR tasks:

- **Dead End:** "Can't use async in the auth module because it depends on a sync middleware chain" → document it
- **Gotcha:** "The order status field uses integers 1-7, not the enum — legacy migration never happened" → document it
- **Decision:** "We use raw SQL instead of ORM for the reporting module because of the complex join queries" → document it

These become Knowledge Base entries that AI reads in future sessions. Every documented gotcha prevents one future $105 escaped defect.

Phase 4: Requirement Links for New Work (Month 1+)

New features on legacy codebases often lack clear requirements — someone says "fix the thing" and you investigate. SENAR's Exploration (Section 6.1) handles this:

1. Start an Exploration (time-bounded investigation)
2. When you understand what needs to happen, create a Task with goal + AC
3. Link the Task to a Story or BR — even if the BR is just "reduce support tickets about X"

For bug fixes in legacy code:

- BR: the original business need that the bug violates (e.g., "users must be able to reset passwords")
- TR: the specific fix with AC (e.g., "POST /reset-password returns 200 for non-existing emails — no information leak")

Phase 5: Quality Gates on Legacy (Month 2+)

QG-0 works immediately on legacy: every task has goal + AC before starting.

QG-2 may need adaptation:

- "CI passes" requires that CI exists. If not, add minimal CI as a one-time investment.
- "Tests pass" requires tests. For untested legacy modules, the minimum is: add tests for the code you change. Don't test what you don't touch.
- "Types clean" may not apply to dynamically typed legacy. Use what's available (mypy for Python, TypeScript strict for JS migrations).

QG-3 and QG-4 are Team+ and Enterprise — defer until SENAR Core is solid.

What You Get

After 3 months of SENAR Core on a legacy codebase:

- Documentation frontier advancing with actual work (not rotting in a wiki)
- Knowledge base of gotchas that prevent repeated failures
- Measurable FPSR showing whether context quality is improving

- Dead ends that save hours per session
- No "big bang" documentation effort, no process disruption

Anti-Pattern: The Documentation Sprint

Problem: Manager says "let's spend 2 weeks documenting everything before we start using AI." **Why it fails:** Documentation without implementation context is abstract and immediately stale. The person documenting module X hasn't worked in it for 6 months — they'll miss the gotchas. **SENAR approach:** Document on contact. The person documenting module X is the person implementing a task in module X right now. They know exactly what AI needs to know because they just discovered it.

SENAR Guide: A Complete Task, Start to Finish

This is not a diagram. This is one task, from requirement to done, exactly as it happens in practice. The task: add a password reset endpoint to a REST API.

Stack: Python 3.14, FastAPI, PostgreSQL, pytest. Configuration: SENAR Core (solo Supervisor).

---

1. The Requirement

The product backlog has a business requirement:

> ***BR-042:*** *Users must be able to reset their passwords without contacting support.*

The Supervisor decomposes this into a Task Record:

```yaml
task: impl-password-reset
story: user-authentication
goal: >
  Implement POST /auth/password-reset/request and
  POST /auth/password-reset/confirm endpoints that allow
  users to reset their password via a time-limited token.
acceptance_criteria:
  1. POST /auth/password-reset/request with valid email returns 202
and creates a reset token
  2. POST /auth/password-reset/request with unknown email returns 202
(no information leak)
  3. Reset token expires after 30 minutes
  4. POST /auth/password-reset/confirm with valid token + new
password updates the password and invalidates the token
  5. POST /auth/password-reset/confirm with expired or invalid token
returns 400
  6. POST /auth/password-reset/confirm with password shorter than 8
chars returns 422
role: supervisor
complexity: M
```

Time spent: ~2 minutes.

---

2. Start Gate (QG-0 in the Standard) Check

Before the task goes to `active`, the system checks Context Gate criteria:

Criterion	Status
Goal is non-empty	**PASS** — goal describes both endpoints and their purpose
Acceptance criteria exist	**PASS** — 6 criteria, each independently testable
Story link set	**PASS** — linked to `user-authentication`
Role assigned	**PASS** — `supervisor`

Start Gate (QG-0): **PASS**. Task moves to `active`.

---

3. The AI Prompt

The Supervisor opens the terminal and gives this prompt to the AI agent:

```
Implement password reset for our auth module.

Goal: POST /auth/password-reset/request and POST /auth/password-
reset/confirm
endpoints. Request creates a time-limited token. Confirm validates the
token
and updates the password.

Acceptance criteria:
1. POST /request with valid email → 202, creates reset token
2. POST /request with unknown email → 202 (no information leak)
3. Token expires after 30 minutes
4. POST /confirm with valid token + new password → updates password,
invalidates token
5. POST /confirm with expired/invalid token → 400
6. POST /confirm with password < 8 chars → 422

Scope:
- Change ONLY files in src/auth/ and tests/auth/
- Follow the patterns in src/auth/router.py and src/auth/service.py
- Do NOT modify the User model — add a new PasswordResetToken model
- Use the existing get_password_hash() for hashing
- Do NOT add new dependencies without asking

Plan:
1. Create PasswordResetToken model in src/auth/models.py
2. Add reset request/confirm logic in src/auth/service.py
3. Add endpoints in src/auth/router.py
4. Add schemas in src/auth/schemas.py
5. Write tests for all 6 AC
```

Time spent: ~1 minute (most of it was already in the task record).

4. The AI Output

The AI generates four files:

`src/auth/models.py` — adds `PasswordResetToken` model:

```
class PasswordResetToken(Base):
    __tablename__ = "password_reset_tokens"

    id: Mapped[int] = mapped_column(primary_key=True)
    user_id: Mapped[int] = mapped_column(ForeignKey("users.id"))
    token: Mapped[str] = mapped_column(String(64), unique=True,
index=True)
    expires_at: Mapped[datetime] = mapped_column()
    used: Mapped[bool] = mapped_column(default=False)
```

`src/auth/schemas.py` — adds request/response schemas:

```
class PasswordResetRequest(BaseModel):
    email: EmailStr

class PasswordResetConfirm(BaseModel):
    token: str
    new_password: str = Field(min_length=8)
```

`src/auth/service.py` — adds two service functions:

```python
from datetime import timezone

async def request_password_reset(db: AsyncSession, email: str) ->
None:
    user = await get_user_by_email(db, email)
    if user is None:
        return  # AC-2: no information leak
    token = secrets.token_urlsafe(48)
    expires_at = datetime.now(timezone.utc) + timedelta(minutes=30)
# AC-3
    reset = PasswordResetToken(user_id=user.id, token=token,
expires_at=expires_at)
    db.add(reset)
    await db.commit()


async def confirm_password_reset(
    db: AsyncSession, token: str, new_password: str
) -> None:
    reset = await db.execute(
        select(PasswordResetToken).where(
            PasswordResetToken.token == token,
            PasswordResetToken.used == False,
        )
    )
    reset = reset.scalar_one_or_none()
    if reset is None or reset.expires_at <
datetime.now(timezone.utc):
        raise HTTPException(status_code=400, detail="Invalid or
expired token")
    user = await db.get(User, reset.user_id)
    user.password_hash = get_password_hash(new_password)
    reset.used = True  # AC-4: invalidate token
    await db.commit()
```

`src/auth/router.py` — adds two endpoints:

```python
@router.post("/password-reset/request", status_code=202)
async def password_reset_request(
    body: PasswordResetRequest, db: AsyncSession = Depends(get_db)
):
    await request_password_reset(db, body.email)
    return {"detail": "If the email exists, a reset link has been
sent"}


@router.post("/password-reset/confirm", status_code=200)
async def password_reset_confirm(
    body: PasswordResetConfirm, db: AsyncSession = Depends(get_db)
):
    await confirm_password_reset(db, body.token, body.new_password)
    return {"detail": "Password updated"}
```

`tests/auth/test_password_reset.py` — six tests:

```python
async def test_request_reset_valid_email(client, db, user):
    """AC-1: valid email returns 202, creates token."""
    resp = await client.post("/auth/password-reset/request", json=
{"email": user.email})
    assert resp.status_code == 202
    token = await db.execute(select(PasswordResetToken).where(
        PasswordResetToken.user_id == user.id
    ))
    assert token.scalar_one_or_none() is not None


async def test_request_reset_unknown_email(client):
    """AC-2: unknown email returns 202 (no leak)."""
    resp = await client.post("/auth/password-reset/request",
                             json={"email": "nobody@example.com"})
    assert resp.status_code == 202


async def test_token_expires_after_30_minutes(client, db, user):
    """AC-3: token created with 30-minute expiry."""
    await client.post("/auth/password-reset/request", json={"email":
user.email})
    token = (await
db.execute(select(PasswordResetToken))).scalar_one()
    delta = token.expires_at - datetime.now(timezone.utc)
    assert 29 * 60 < delta.total_seconds() < 31 * 60


async def test_confirm_reset_valid_token(client, db, user,
reset_token):
    """AC-4: valid token + new password → password updated, token
invalidated."""
    resp = await client.post("/auth/password-reset/confirm",
                             json={"token": reset_token.token,
"new_password": "newsecure123"})
    assert resp.status_code == 200
    await db.refresh(reset_token)
    assert reset_token.used is True
    await db.refresh(user)
    assert verify_password("newsecure123", user.password_hash)
```

```
async def test_confirm_reset_expired_token(client,
expired_reset_token):
    """AC-5: expired token returns 400."""
    resp = await client.post("/auth/password-reset/confirm",
                             json={"token":
expired_reset_token.token, "new_password": "newsecure123"})
    assert resp.status_code == 400


async def test_confirm_reset_short_password(client, reset_token):
    """AC-6: password < 8 chars returns 422."""
    resp = await client.post("/auth/password-reset/confirm",
                             json={"token": reset_token.token,
"new_password": "short"})
    assert resp.status_code == 422
```

---

5. Verification Against AC

The Supervisor does not skim the code and say "looks good." The Supervisor goes through the acceptance criteria one by one:

AC	What to Check	Verdict
1. Valid email → 202 + token created	`test_request_reset_valid_email` asserts 202 and token exists in DB	**PASS**
2. Unknown email → 202	`test_request_reset_unknown_email` asserts 202, no error	**PASS**
3. Token expires in 30 min	`test_token_expires_after_30_minutes` checks `expires_at` delta	**PASS**
4. Valid confirm → updates password, invalidates token	`test_confirm_reset_valid_token` checks `used=True` and password changed	**PASS**
5. Expired/invalid token → 400	`test_confirm_reset_expired_token` asserts 400	**PASS**
6. Short password → 422	`test_confirm_reset_short_password` asserts 422	**PASS**

Every criterion has a corresponding test. Every test asserts the expected behavior, not the implementation.

Time spent: ~3 minutes reading tests and confirming the mapping.

---

6. A Dead End

But first — something went wrong during execution.

The AI's first attempt used `bcrypt` for the password hashing utility it added alongside the endpoint:

```
from bcrypt import hashpw, gensalt

def get_password_hash(password: str) -> str:
    return hashpw(password.encode(), gensalt()).decode()
```

When tests ran:

```
E   ModuleNotFoundError: No module named '_bcrypt'
```

The `bcrypt` package has a C extension that does not compile on Python 3.14 (as of March 2026). The Supervisor directed the AI to switch:

```
bcrypt import fails on Python 3.14. Switch to argon2-cffi with argon2id
hasher.
The existing get_password_hash() already uses argon2 — don't create a new
one,
use the existing function.
```

The AI corrected to use the existing `get_password_hash()` from `src/auth/utils.py` (which already uses `argon2-cffi`). Tests pass.

The Supervisor writes a dead end entry:

```
Dead end: bcrypt package — C extension fails to compile on Python 3.14.
Switched to argon2-cffi (argon2id). The existing get_password_hash()
utility
already uses argon2id. Don't add bcrypt to new endpoints.
```

Time spent: ~30 seconds to write the entry.

---

7. Done Gate (QG-2 in the Standard) Check

The Supervisor runs the Implementation Gate:

```
$ pytest tests/auth/test_password_reset.py -v
======================== 6 passed in 2.14s ========================

$ mypy src/auth/ --strict
Success: no issues found in 6 source files

$ ruff check src/auth/
All checks passed!

$ git rev-parse HEAD && git status --porcelain
4f2a9c1e8b3d7a06f5c2e91b8d4a7f30c6e51982
(clean)
```

The last command is not decoration. Standard 8.6(c) requires the verdict to be a statement about the artifacts *as they stand when the task is closed*, and 8.6(c) is SHALL at every configuration, Core included. The three checks above ran against a specific tree; the fourth line records which one, and that the tree had not moved since. Edit a file after the tests pass and this verdict is about a state that no longer exists — re-run, or the answer is no.

Criterion	Status
CI pipeline passes	**PASS** — all green
All tests pass	**PASS** — 6/6
Type checking clean	**PASS** — mypy strict, zero errors
No new linting violations	**PASS** — ruff clean
Acceptance criteria verified	**PASS** — all 6 AC checked in Section 5
No security vulnerabilities	**PASS** — `argon2id` is OWASP-recommended; no new deps with CVEs
State the verdict is about	`4f2a9c1e`, working tree clean (Standard 8.6(c); recording it is 8.6(d), SHOULD at Core and SHALL from Foundation)

Done Gate (QG-2): **PASS** for state `4f2a9c1e`. Task moves to `done`.

Time spent: ~2 minutes (mostly waiting for CI).

> ***Measuring FPSR in practice:*** *Count a task as 'first-pass success' if it passes QG-2 without returning to active status. Track via your task management tool: tasks that go planning→active→done = success; tasks that go active→done→active→done = rework. FPSR = successes / total completed × 100%.*

---

8. Knowledge Capture

The Supervisor creates a knowledge entry for the decision made during this task:

```yaml
type: decision
title: "Using argon2id for all password hashing"
context: >
  Password reset endpoint needed hashing. bcrypt fails on Python
3.14.
  argon2id is the OWASP recommendation and already used in
get_password_hash().
decision: >
  All password operations use argon2-cffi with argon2id algorithm via
  the shared get_password_hash() utility in src/auth/utils.py.
  Do not introduce bcrypt or any other hashing library.
related_dead_end: "bcrypt C extension incompatible with Python 3.14"
```

Time spent: ~30 seconds.

---

9. Session End Snippet

At session end, the handoff captures everything a future session (or a different Supervisor) needs:

```
Session #12 — 2026-03-22 — 95 minutes — 4 tasks done

Completed:
  - impl-password-reset (this task)
  - impl-email-verification
  - fix-jwt-refresh-race
  - impl-logout-endpoint

Dead ends:
  - bcrypt C extension fails on Python 3.14 → use argon2-cffi

Knowledge entries:
  - decision: argon2id for all password hashing (OWASP, Python 3.14
compat)
  - gotcha: JWT refresh token race condition — need DB-level locking

Next session:
  - impl-oauth-google — Google OAuth integration
  - impl-rate-limiting — rate limit on auth endpoints (5 req/min)

Warnings:
  - OAuth requires new env vars (GOOGLE_CLIENT_ID, GOOGLE_CLIENT_SECRET)
    — add before starting the task
```

What This Cost

Step	Time
Writing goal + acceptance criteria	~2 min
Writing the AI prompt	~1 min
AI generates code + tests	~3 min (waiting)
Verification against AC	~3 min
Dead end documentation	~30 sec
Knowledge entry	~30 sec
Total	**~10 min**

Without AI, this task — two endpoints, a model, schemas, six tests, hashing research — would take approximately 25–40 minutes of manual coding.

The overhead of SENAR discipline (goal, AC, verification, knowledge capture) adds roughly 4 minutes to a task. The AI generates the remaining 6 minutes of work. The net result: a fully tested, documented, traceable endpoint in under 10 minutes.

But the real savings are not in this session. They are in the next session, when someone (or an AI agent) needs to add another auth endpoint and finds:

- The dead end that prevents them from wasting 15 minutes on bcrypt.
- The decision that tells them to use `get_password_hash()`.
- The acceptance criteria pattern they can copy for the next endpoint.

That future time saved is where SENAR pays for itself many times over.

Key Takeaways

1. **The AC did the heavy lifting.** Six criteria, written in 2 minutes, drove the entire implementation — what the AI generated, what the tests checked, what the Supervisor verified. Without them, the Supervisor would have been reading code and guessing whether it's correct.

2. **The dead end is an investment, not overhead.** Thirty seconds of typing saves the next person (or AI) from repeating the same failure. In a knowledge base with 100+ dead ends, this compounds into hours saved per increment.

3. **Done Gate (QG-2) is not a ceremony.** It's your test runner, type checker, linter, and a verification table — the specific tools depend on your stack. It takes 2 minutes. If any line is red, the task is not done. No judgment calls, no "it's probably fine."

4. **The handoff makes sessions independent.** Session #13 can be started by a different person, a different AI agent, or the same Supervisor after a weekend. Everything needed is in the handoff — no tribal knowledge, no "let me remember where I was."

A Note on This Example

This walkthrough shows a clean-pass scenario: one dead end, one retry, everything green on the second attempt. Real sessions are often messier — multiple dead ends, partial test failures, AC that turn out ambiguous mid-implementation, unexpected dependency conflicts. This example demonstrates the SENAR workflow structure, not the typical difficulty level. Your sessions will vary.

Stack Variations

The practices above apply identically regardless of technology stack. Here are the same acceptance criteria patterns for common stacks:

> ***Note:*** *Stack variations illustrate framework-specific patterns, not exact translations of the Python example. Adapt acceptance criteria to your stack's conventions.*

Java / Spring Boot

Task: Implement password reset endpoint **AC:**

1. POST /api/v1/auth/reset-password accepts {email} → returns 200 (no email enumeration)
2. Token stored with BCrypt hash, expires in 1 hour
3. Rate limited: 3 requests per email per hour (@RateLimiter annotation or filter)
4. Integration test with @SpringBootTest and MockMvc
5. Negative: invalid token returns 400, expired token returns 410

Go / Gin

Task: Implement password reset endpoint **AC:**

1. POST /api/v1/auth/reset-password accepts {email} → returns 200
2. Token stored with bcrypt.GenerateFromPassword, 1h expiry
3. Rate limited: middleware with sync.Map or Redis
4. Table-driven tests with testify
5. Negative: invalid/expired token returns appropriate status

TypeScript / NestJS

Task: Implement password reset endpoint **AC:**

1. POST /api/v1/auth/reset-password accepts {email} → returns 200
2. Token stored with bcrypt.hash, 1h TTL
3. @Throttle() decorator or custom guard
4. Jest e2e test with supertest
5. Negative: ValidationPipe rejects malformed input

Full worked examples for Java/Spring Boot, Go, and TypeScript/NestJS are planned for the SENAR Guide v1.4.

SENAR Guide: Tool Integration

SENAR is tool-agnostic — the standard never requires a specific product. But the practices map differently to each tool. This chapter shows how to apply SENAR with the three most common AI coding tools, plus a minimal knowledge base setup for teams with no infrastructure.

SENAR with Claude Code

Claude Code is a terminal-based AI agent with persistent memory, slash commands, and MCP (Model Context Protocol) integration. It is the closest match to SENAR's assumptions about AI agents.

Goal + AC → Task Description in Prompt

Give the goal and acceptance criteria directly in the prompt. Claude Code processes structured text well:

```
Implement POST /orders/{id}/cancel endpoint.

Goal: Allow users to cancel their own orders if the order is in "pending"
status.

Acceptance criteria:
1. POST /orders/{id}/cancel with valid order in "pending" → 200, status
changes to "cancelled"
2. POST /orders/{id}/cancel with order in "shipped" → 409 Conflict
3. POST /orders/{id}/cancel on another user's order → 403
4. Cancelled order cannot be cancelled again → 409
5. Returns 404 for non-existent order ID

Scope: change ONLY src/orders/ and tests/orders/. Follow patterns in
src/orders/router.py.
```

Scope Boundaries → CLAUDE.md + Slash Commands

`CLAUDE.md` at the project root defines persistent boundaries the agent reads on every session:

```
# Boundaries
- NEVER modify files outside src/ and tests/
- NEVER change database migration files directly — use alembic
- Follow existing patterns in router.py, service.py, schemas.py
- Do NOT add dependencies without asking
```

Use `/plan` before complex tasks to have the agent propose a plan for review before executing.

Knowledge Base → CLAUDE.md + Memory System

Claude Code has two knowledge layers:

Layer	SENAR Mapping	Persistence
`CLAUDE.md`	Project-level knowledge (rules, patterns, boundaries)	In repo, version-controlled
Memory files (`~/.claude/projects/*/memory/`)	Dead ends, decisions, gotchas	Per-user, persists across sessions

When you hit a dead end, tell Claude Code explicitly:

```
Remember this: bcrypt fails on Python 3.14. Always use argon2-cffi.
```

It writes to its memory system. Future sessions read this automatically.

For teams using MCP, knowledge entries can be pushed to a shared knowledge base:

```
Create a knowledge entry: "argon2id for password hashing — bcrypt incompatible
with Python 3.14. OWASP recommendation. Use get_password_hash() from auth/utils.py."
```

Dead Ends → Knowledge Entries via Memory or MCP

Immediate capture during the session:

```
That approach failed. Remember: SQLAlchemy async sessions cannot use lazy
loading. Must use selectinload() or joinedload() for relationships.
```

QG-0 → /plan Skill

Before starting a task, use `/plan` to validate that the task has sufficient context:

```
/plan — Review this task before implementation. Check that goal, AC, and
scope
are clear. Flag anything ambiguous.
```

The agent reviews the task description and flags missing AC, ambiguous scope, or unstated assumptions.

QG-2 → /review Skill

After implementation, use `/review` to run the Implementation Gate:

```
/review — Check this implementation against the acceptance criteria.
Verify: tests exist for each AC, mypy clean, no linting violations.
```

Session Discipline

Claude Code tracks tool call counts internally. Use checkpoints:

```
/checkpoint — Save progress. 47 tool calls since last checkpoint.
```

Session handoffs go into the memory system or a handoff file the agent reads on next startup.

SENAR with Cursor

Cursor is a VS Code fork with inline AI chat, file-level context via `@mentions`, and project-level rules via `.cursorrules`.

Goal + AC → .cursorrules + Prompt Structure

Create a `.cursorrules` file with persistent instructions:

```
# Project Rules
- Every implementation MUST have tests for all acceptance criteria
- Follow patterns in existing router/service/schema files
- Never modify database migrations directly
- Use type hints on all function signatures
```

For each task, structure your Cursor chat prompt the same way:

```
Goal: [one sentence]
AC: [numbered list]
Scope: [files to change, files NOT to change]
Plan: [ordered steps]
```

Cursor does not enforce this structure — you enforce it by habit.

Scope Boundaries → @file Mentions

Cursor's `@file` syntax controls context. Use it as a constraint fence:

```
@src/orders/router.py @src/orders/service.py
Implement order cancellation. Follow the patterns in these files.
Only change files in src/orders/ and tests/orders/.
```

Mentioning specific files focuses the model's attention and reduces the chance of it inventing patterns that don't match your codebase.

Knowledge Base → .cursorrules + Project Docs

Cursor reads `.cursorrules` on every prompt. Use it for accumulated knowledge:

```
# Known Issues
- bcrypt fails on Python 3.14 — use argon2-cffi
- SQLAlchemy async: no lazy loading, use selectinload()
- JWT refresh: must use DB-level locking to prevent race conditions
```

For larger knowledge bases, maintain a `docs/knowledge/` directory and reference it:

```
@docs/knowledge/dead-ends.md — Read this before implementing auth-related features.
```

Dead Ends → Comment Files or Inline Docs

Cursor does not have a built-in knowledge system. Two options:

Option A: Knowledge files (recommended) — maintain `docs/knowledge/dead-ends.md` and reference with `@`:

```
# Dead Ends

- 2026-03-15: bcrypt C extension fails on Python 3.14 → use argon2-cffi
- 2026-03-18: Celery beat + async → use APScheduler instead
- 2026-03-20: pytest-asyncio auto mode breaks fixtures → use mode=strict
```

Option B: Inline comments — add dead end notes near the relevant code:

```
# DEAD END: Do not use bcrypt here — C extension fails on Python 3.14.
# Use argon2-cffi via get_password_hash(). See docs/knowledge/dead-ends.md.
from auth.utils import get_password_hash
```

QG-0 → Manual (Write AC Before Prompting)

Cursor has no built-in gate. You are the gate. The discipline:

1. Write goal + AC in the chat prompt **before** asking for code
2. If you catch yourself typing "implement X" without AC, stop. Write the AC first.
3. Keep a task template (see Starter Kit below) and paste it every time.

QG-2 → Cursor Review Features + Terminal

After the AI generates code:

1. Use Cursor's diff view to review changes
2. Run tests in the terminal: `pytest tests/ -v`
3. Run type checker: `mypy src/ --strict`
4. Walk through AC one by one — does each criterion have a passing test?

SENAR with GitHub Copilot

GitHub Copilot works primarily through inline completions and Copilot Chat (VS Code sidebar or PR reviews). It has the least structured context control of the three tools.

Goal + AC → Structured Comments Before Code

Copilot reads surrounding code and comments. Use structured comments as the prompt:

```
# TASK: Implement order cancellation endpoint
# GOAL: Allow users to cancel pending orders
# AC-1: POST /orders/{id}/cancel + pending order → 200,
status=cancelled
# AC-2: POST /orders/{id}/cancel + shipped order → 409
# AC-3: POST /orders/{id}/cancel + other user's order → 403
# AC-4: Cancelled order → 409 on re-cancel
# AC-5: Non-existent order → 404
# SCOPE: src/orders/router.py, src/orders/service.py, tests/orders/

@router.post("/orders/{order_id}/cancel")
async def cancel_order(  # Copilot completes from here
```

This is more friction than Claude Code or Cursor, but it works: Copilot uses the comments as context for completions.

Scope Boundaries → File-Level Context

Copilot's context is primarily the open file and related files. Control scope by:

- Having the right files open in tabs (Copilot reads open files)
- Keeping unrelated files closed
- Using Copilot Chat with explicit file references: "Following the pattern in `src/orders/router.py`, implement..."

Knowledge Base → Codebase Documentation (Rule 9.11)

Copilot learns from your codebase. Rule 9.11 (Documentation Completeness) directly improves Copilot output:

- Docstrings on every public function → Copilot generates consistent new functions
- Type hints everywhere → Copilot generates typed code
- `README.md` per module → Copilot understands module purpose

```
"""Order service — handles order lifecycle operations.

Patterns:
- All mutations go through service layer, never router
- Use get_or_404() for entity lookup
- Status transitions validated via Order.can_transition_to()
"""
```

Dead Ends → ADR Files or Doc Comments

Use Architecture Decision Records in `docs/adr/`:

```
# ADR-007: Password Hashing Algorithm

## Status: Accepted

## Context
Need password hashing for auth module. bcrypt C extension fails on
Python 3.14.

## Decision
Use argon2-cffi with argon2id algorithm via shared
get_password_hash() utility.

## Consequences
- All password operations use one utility function
- argon2-cffi must remain in dependencies
- Do not add bcrypt as alternative
```

Copilot reads ADR files when they are open, and Copilot Chat can be directed to them explicitly.

QG-0 → Manual Discipline (Pre-Implementation)

Manual discipline — write goal + AC in a task tracking file or issue before starting AI work. Copilot does not have a built-in pre-implementation gate; use TASK_TEMPLATE.md or issue templates as the enforcement mechanism.

1. Write goal + AC in a GitHub issue or `TASK_TEMPLATE.md` **before** opening the editor
2. If you catch yourself prompting Copilot without AC, stop. Write the AC first.
3. Keep a task template (see Starter Kit below) and fill it in every time.

QG-2 → PR Template + CI Pipeline + PR Review

Since Copilot integrates with GitHub, use PR templates as the post-implementation gate:

```
<!-- .github/PULL_REQUEST_TEMPLATE.md -->
## Task
- [ ] Goal defined
- [ ] Acceptance criteria listed below
- [ ] Scope boundaries stated

## Acceptance Criteria
1.
2.
3.

## Verification
- [ ] Each AC has a passing test
- [ ] mypy/tsc clean
- [ ] No new linting violations

## Knowledge
- Dead ends encountered:
- Decisions made:
```

The PR cannot be opened without filling this in. Reviewers verify against it.

Additionally, Copilot's natural QG-2 includes the CI pipeline:

1. GitHub Actions runs tests, type checker, linter
2. Copilot Chat (PR review mode) reviews the diff
3. Human reviewer checks AC-to-test mapping

This is less immediate than Claude Code's in-session verification, but it works when combined with the PR template.

The Core Knowledge Base (No Infrastructure Required)

You do not need a database, a wiki, or a specialized tool. You need three files in your repository:

```
docs/
  knowledge/
    dead-ends.md      — one line per dead end
    decisions.md      — one paragraph per decision
    gotchas.md        — one line per gotcha
```

dead-ends.md

```md
# Dead Ends

Approaches that failed. Check here before trying something new.

- 2026-03-15: bcrypt — C extension fails on Python 3.14. Use argon2-cffi.
- 2026-03-18: Celery beat with async workers — event loop conflicts. Use APScheduler.
- 2026-03-20: pytest-asyncio auto mode — breaks fixtures with scope=session. Use mode=strict.
- 2026-03-21: SQLAlchemy lazy loading in async — not supported. Use selectinload().
- 2026-03-22: Redis Sentinel with asyncio — aioredis deprecated. Use redis-py >= 5.0.
```

Format: `date: thing tried — why it failed. What to use instead.`

One line. Ten seconds to write. Saves hours.

decisions.md

```
# Decisions

Architectural and technical decisions with context.

## Password Hashing: argon2id
**Date:** 2026-03-15
**Context:** Need password hashing for auth module.
**Decision:** Use argon2-cffi with argon2id via `get_password_hash()`
in `src/auth/utils.py`.
**Reason:** OWASP recommendation, Python 3.14 compatible, bcrypt
fails (see dead-ends.md).

## Task Queue: APScheduler over Celery
**Date:** 2026-03-18
**Context:** Need periodic tasks (token cleanup, report generation).
**Decision:** Use APScheduler with AsyncIOScheduler.
**Reason:** Celery beat has event loop conflicts with async FastAPI.
APScheduler runs in-process.
```

Format: title, date, context (one sentence), decision (one sentence), reason (one sentence).

gotchas.md

```
# Gotchas

Things that are technically correct but surprising. Not bugs — just
traps.

- SQLAlchemy: `session.refresh()` after commit or you get stale data
in tests
- FastAPI: `Depends()` in test overrides must match the exact
function, not just signature
- pytest: fixture scope=session + async requires `event_loop` fixture
override
- Docker: `COPY requirements.txt .` before `COPY . .` — layer cache
invalidation order matters
- PostgreSQL: `SERIAL` vs `GENERATED ALWAYS AS IDENTITY` — use
IDENTITY for new tables
```

Format: one line. Context: behavior. That is all.

Why This Works

AI agents read your repository. When `docs/knowledge/dead-ends.md` is in the repo, every AI prompt has access to it — whether through Claude Code's automatic file reading, Cursor's `@file` mentions, or Copilot's open-tab context.

The knowledge base is version-controlled, searchable, and grows with every session. No external tool required.

SENAR Starter Kit

Copy-paste ready templates for immediate use.

TASK_TEMPLATE.md

```md
# Task: [slug]

## Goal
[One sentence: what must be accomplished]

## Acceptance Criteria
1. [First criterion — independently testable]
2. [Second criterion]
3. [Third criterion]

## Requirement Link
Story: [story-slug]
BR: [business requirement ID, if applicable]

## Scope
- Change ONLY: [directories/files]
- Do NOT change: [protected files/directories]
- Follow patterns in: [reference files]

## Plan
1. [First step]
2. [Second step]
3. [Third step]

## Notes
[Context, prior dead ends to avoid, related knowledge entries]
```

.github/PULL_REQUEST_TEMPLATE.md

```md
## Summary
[One sentence: what this PR does]

## Acceptance Criteria
- [ ] AC-1: [criterion from task]
- [ ] AC-2: [criterion from task]
- [ ] AC-3: [criterion from task]

## Verification
- [ ] All AC have corresponding tests
- [ ] Tests pass locally
- [ ] Type checker clean (mypy/tsc/pyright)
- [ ] Linter clean
- [ ] No new dependencies added (or justified below)

## Dead Ends Encountered
<!-- Approaches that failed during this task. Move to
docs/knowledge/dead-ends.md -->
- None / [description]

## Decisions Made
<!-- Architectural or technical decisions. Move to
docs/knowledge/decisions.md -->
- None / [description]

## Knowledge Entries
<!-- Gotchas, patterns, or non-obvious behaviors discovered -->
- None / [description]
```

docs/knowledge/README.md

```
# Knowledge Base

This directory contains accumulated project knowledge from SENAR
sessions.

## Files

| File | Purpose | Format |
|------|---------|--------|
| `dead-ends.md` | Approaches that failed | One line per entry |
| `decisions.md` | Architectural/technical decisions | One paragraph
per entry |
| `gotchas.md` | Surprising but correct behaviors | One line per
entry |

## How to Use

**Before starting a task:** Check `dead-ends.md` for the area you're
working in.

**During a task:** When an approach fails, add one line to `dead-
ends.md`.
When you make an architectural choice, add a paragraph to
`decisions.md`.
When you discover something surprising, add a line to `gotchas.md`.

**After a task:** Review what you learned. If any dead ends,
decisions, or
gotchas were captured in the PR template, move them to the
appropriate file.

## For AI Agents

If you are an AI coding agent reading this: check `dead-ends.md`
before
attempting any approach in a domain that has entries. Check
`gotchas.md`
for the libraries and frameworks you are about to use. Check
```

```
`decisions.md`
for architectural context.
```

Knowledge Base Options

SENAR does not prescribe a specific knowledge base tool. Common approaches by team size:

- **Solo/Core**: Markdown files in the project repo (e.g., `docs/knowledge/`)
- **Foundation (1-3 Pairs)**: Project wiki, Notion, or structured YAML/JSON in repo
- **Team (3-10)**: Dedicated tool with search — CouchDB+Meilisearch, Confluence, Linear docs
- **Enterprise (10+)**: Federated knowledge bases per project with cross-project search

The key requirement: knowledge must be searchable by AI agents in future sessions.

Choosing Your Tool

SENAR Practice	Claude Code	Cursor	GitHub Copilot
Goal + AC	Prompt text	Prompt text	Structured comments
Scope boundaries	CLAUDE.md + prompt	.cursorrules + @file	Open tabs + prompt
Knowledge base	Memory + MCP	.cursorrules + docs/	docs/ + ADRs
Dead end capture	Memory entries	Knowledge files	ADR files
QG-0 enforcement	/plan skill	Manual discipline	Task template / issue
QG-2 enforcement	/review skill	Terminal + diff view	PR template + CI pipeline
Session handoff	Memory system	Handoff file	PR description
Context persistence	Automatic (memory)	Manual (@file)	Manual (open files)

> ***What "manual discipline" satisfies, and what it does not (revision 1.4).***
>
> *Every QG-0 arrangement in this chapter that reads "you are the gate" is a Supervisor remembering to do something. Standard 8.6 asks a further question of any such arrangement: does anything interpose on the effect, or does the effect simply not happen when the Supervisor remembers?*
>
> *A remembered discipline satisfies 8.6(a), (c) and (e) — those cost policy, and you can hold them with a text editor. It does not satisfy 8.6(b), preventive placement, because nothing prevents the work from proceeding when it is forgotten. 8.6(b) is SHOULD at Core and Foundation and **SHALL from Team**, so these arrangements are adequate for individual and small-team adoption and are not adequate for a Team-configuration conformance claim. Where the discipline is what you have, 8.6 asks you to record the gate as detective with respect to its declared effect, so that a later claim starts from a known position (Standard 8.6, Scaling by configuration).*
>
> *8.6(d), binding to a measured digest, is the property that reckoning leaves out, and it is the one that bites first: it is SHALL from **Foundation**, not from Team (Standard 8.6, Scaling by configuration). A remembered discipline does not hold it by remembering. The verdict has to name the state it was issued for by a digest measured over that state, and the digest has to be recorded with the verdict (8.6(g)). It belongs neither with the three above nor with (b): unlike (b) it needs no ability to interpose — one command measures it — and unlike the three it is not free, because someone has to run that command and write the result down. A Foundation adopter running the arrangements in this chapter therefore owes (d) outright, and owes (b) as an unmet SHOULD, documented with rationale under 13.1(c).*
>
> *The practical consequence for tool choice: to reach Team configuration you need something that can refuse — a pre-commit hook, a required CI job, a tracker that will not move a task to `active` with empty acceptance criteria — not a better prompt template.*

The tool does not make the process. The process makes the tool useful. A Supervisor with discipline and a plain text editor will outperform a Supervisor with the best AI tool and no acceptance criteria.

SENAR Guide: Agent Configuration in Practice

This chapter bridges the normative requirements of Standard Section 5 (Agent Instrumentation) with practical implementation. It covers how to write operational scripts, define agent profiles, manage script changes, and apply these concepts across different AI development tools.

Anatomy of an Operational Script

An Operational Script is a structured natural-language instruction. Unlike code, scripts are interpreted by an AI agent — which means they must be unambiguous, self-contained, and testable.

Minimal Script Structure

```
Script: commit
Trigger: /commit command or explicit user request

Preconditions:
- Changes exist in the working tree
- No failing tests

Algorithm:
1. Run git status to see all changes
2. Run git diff to review staged and unstaged changes
3. Analyze changes and draft a commit message:
   - Summarize the nature of changes (feature, fix, refactor...)
   - Keep message under 72 characters for the subject line
4. Stage relevant files (prefer specific files over git add -A)
5. Create the commit
6. Run git status to verify success

Postconditions:
- Commit exists in local repository
- Working tree is clean for committed files

Outputs:
- Commit hash
- Summary of what was committed
```

What Makes a Good Script

1. **Explicit preconditions.** If a script assumes something is true, state it. "Tests pass" is a precondition, not an assumption.

2. **Numbered steps with decision points.** When step 3 depends on the outcome of step 2, say so. "If tests fail, stop and report. If tests pass, continue to step 4."

3. **Defined outputs.** What does the script produce? A commit? A report? A status change? Be specific.

4. **Edge cases documented.** "If no changes exist, report this and exit without creating an empty commit."

5. **No platform-specific calls.** Write "Run the test suite" not "Run `npm test`." The script should work regardless of the project's tech stack.

Anti-Patterns

- **Vague instructions:** "Make sure everything is good" — not verifiable
- **Implicit knowledge:** "Use the standard approach" — which standard?
- **Missing error handling:** No guidance for when steps fail
- **Overcoupled scripts:** One script that does planning, implementation, and review

---

Defining Agent Profiles

An Agent Profile restricts what an AI agent can do. The restriction is the point — it prevents the agent from performing actions outside its designated function.

Generator Profile (Primary Development)

```
Profile: Generator
Purpose: Implement code changes for a specific Task

Scripts available:
- implement: Write code to satisfy Task acceptance criteria
- debug: Investigate and fix failures
- commit: Create version-controlled commits
- test: Run and verify test results

Access:
- Read/write: source code, tests, configuration
- Read: task definitions, knowledge base
- Write: task status updates, knowledge entries
- NO access to: deployment, infrastructure, other projects' code

Constraints:
- Changes must stay within Task scope boundaries
- Commits must be atomic (one logical change)
- Must not modify files outside defined scope
```

Reviewer Profile (Independent Verification)

```
Profile: Reviewer
Purpose: Independently verify code changes against acceptance criteria

Scripts available:
- review: Check code against acceptance criteria and standards
- security-audit: Scan for security issues
- report: Generate review findings

Access:
- Read-only: source code, tests, configuration, task definitions
- Write: review comments, findings
- NO access to: code modification, commits, deployment

Constraints:
- Cannot modify any code being reviewed
- Findings must reference specific acceptance criteria
- Must check for AI-specific issues (hallucinated APIs, self-validating
tests)
```

Minimum Viable Profiles

At SENAR Core level, you need at minimum:

1. **Generator** — for development work
2. **Reviewer** — for independent verification

The key insight: even if the same AI agent switches between profiles, the profile switch enforces a cognitive boundary. The Reviewer profile cannot write code, so it evaluates rather than fixes.

Managing Script Changes

Changing a script changes how your AI agent behaves in production. Treat script changes with the same rigor as production code changes.

The Script Change Workflow

```
1. Identify need for change
   ↓
2. Draft the change with rationale
   ↓
3. Review (peer or self, depending on configuration)
   ↓
4. Test on isolated project (Team+)
   ↓
5. Deploy to target project
   ↓
6. Monitor effectiveness (FPSR, error rate)
   ↓
7. Record decision in knowledge base
```

Testing Script Changes

Before deploying a script change across all projects:

1. **Select a representative task** — pick a task similar to what the script will handle
2. **Run the task with the old script** — record FPSR and any issues
3. **Run a similar task with the new script** — compare results
4. **Check for regressions** — did the change break any existing behavior?

Rollback

Every script change should be reversible:

- Scripts are in version control — `git revert` works
- Maintain a version identifier in each script (date or semver)
- Document what to watch for after rollback (tasks in progress may be affected)

Tool-Specific Examples

Claude Code

Claude Code uses `CLAUDE.md` for the Behavioral Contract and `.claude/skills/` for Operational Scripts.

Behavioral Contract (`CLAUDE.md`):

```
# Project Rules
- NEVER commit without explicit permission
- NEVER modify files outside src/ and tests/
- Run tests before every commit
- Ask before installing new dependencies
```

Operational Script (`.claude/skills/review.md`):

```
# Review Skill
Trigger: /review command

Steps:
1. Read the current git diff
2. For each changed file, check:
   a. Does the change match the stated Task goal?
   b. Are there any hardcoded values that should be configurable?
   c. Are error cases handled?
   d. Do tests cover the acceptance criteria, not just code paths?
3. Check for AI-specific issues:
   a. Hallucinated APIs or methods
   b. Non-existent package references
   c. Self-validating test patterns
4. Report findings with specific line references
```

Profile switching: Claude Code does not natively enforce profile permissions. Implement via CLAUDE.md rules: "When performing a review, you are in Reviewer mode. Do NOT modify any code. Only report findings."

OpenAI ChatGPT / GPT-4

Behavioral Contract (system prompt or custom instructions):

```
You are working in Generator mode. You may:
- Read and write source code and tests
- Create commits with descriptive messages
- Search the codebase

You may NOT:
- Modify deployment configuration
- Access external APIs
- Create files outside the project directory
```

Operational Scripts (structured prompts or saved workflows): ChatGPT / GPT-4 does not have a native skill/script mechanism. Scripts are implemented as structured prompt templates that the Supervisor provides at the start of each task.

Cursor

Behavioral Contract (`.cursorrules`):

```
Rules:
- Always run type checking before suggesting changes as complete
- Never modify files not mentioned in the current task
- Prefer editing existing files over creating new ones
```

Operational Scripts (`.cursor/prompts/` or manual invocation): Cursor supports custom prompts that function similarly to operational scripts. Store these in version control alongside the project.

Practical Workflow: Script Change Lifecycle

Scenario: Your "implement" script produces code that frequently has type errors. You want to add a "run type checking" step before marking the implementation as complete.

Step 1: Document the Problem

In your knowledge base:

```
Type: observation
Title: Implementation script produces type errors
Body: In the last 5 tasks, 3 had type checking failures caught at CI
  (QG-2). Adding a type check step to the implement script should
  catch these earlier and improve FPSR.
```

Step 2: Draft the Change

Add to the implement script, after the "write code" step:

```
5. Run the project's type checker
6. If type errors exist:
   a. Fix them
   b. Re-run type checker
   c. Repeat until clean
7. Continue to tests
```

Step 3: Review

For Core: self-review — does the addition make sense? Does it conflict with other steps?

For Team+: peer review — another Supervisor reviews the script change. Check: is the new step clear? Are edge cases handled (what if the project has no type checker)?

Step 4: Test (Team+)

Run one task with the new script on a non-critical project. Did the type check step execute correctly? Did it catch type errors before CI? Did it add unreasonable time?

Step 5: Deploy

Commit the script change. Update the version identifier.

Step 6: Monitor

Over the next 5–10 tasks, track:

- Did type checking failures at CI (QG-2) decrease?
- Did task completion time change significantly?
- Any unexpected side effects?

Step 7: Record

```
Type: decision
Title: Added type checking step to implement script
Body: After 3/5 tasks had type errors at CI, added explicit type check
  step to implement script. First 5 tasks after change: 0/5 type
  errors at CI. ~2 min added per task, saves ~10 min rework.
```

Summary

Concept	Core	Team+
Agent Profiles	Generator + Reviewer (SHOULD)	All 5 with permission enforcement (SHALL)
Operational Scripts	In version control, self-reviewed	Reviewed, tested, registry maintained
Script Changes	Treated as process changes	Tested on isolated project first
Rollback	Version control provides	Registry + explicit rollback procedure
Audit Trail	Knowledge entries	Formal change log

SENAR Guide: Onboarding New Team Members

Reading Order

1. **Day 1 (30 min):** Read SENAR Core — understand the 8 rules, 2 gates, verification checklist
2. **Day 1 (15 min):** Read Guide Quickstart — internalize the 6 habits
3. **Day 2 (30 min):** Read Guide Worked Example — see a complete task end-to-end
4. **Day 2 (20 min):** Read Guide AI Review Checklist — your verification reference
5. **Week 1:** Read Guide Tool Guides — set up your AI tool with SENAR practices

First Task Protocol

Your first SENAR task should be:

- **Small scope** — a bug fix or simple feature (< 1 hour)
- **Paired with an experienced Supervisor** who reviews your gates
- **Full ceremony** — do every step explicitly, even if it feels slow

Checklist for your first task:

- [ ] Write goal + acceptance criteria before touching code
- [ ] Define what's in scope and what's NOT
- [ ] Run Start Gate checklist
- [ ] Work with AI, noting any dead ends (> 15 min)
- [ ] Verify each AC with evidence
- [ ] Run Done Gate checklist
- [ ] Document any dead ends, decisions, or gotchas

Ramp-Up Timeline

Week	Focus	Expected Outcome
1	Core rules + 2-3 supervised tasks	Understand the workflow
2	Independent tasks with post-review	Comfortable with gates
3	Full independence + checklist internalized	FPSR tracking meaningful
4	Knowledge contribution + peer review	Full team member

Common New-Member Mistakes

1. **Skipping the negative scenario** — Rule 1 requires at least one "what should NOT happen"
2. **Not documenting dead ends** — Even "I tried X and it didn't work" is valuable
3. **Treating the checklist as optional** — Start with Standard tier every time
4. **Confusing exploration with implementation** — Exploration is free, but committing requires a task

What NOT to Worry About

- Metrics (FPSR, Dead End Rate) — your first 10 tasks are baseline, not graded
- Perfect knowledge entries — any capture is better than none
- Speed — the process gets faster with practice, not by skipping steps

SENAR Reference: Complete Glossary

Alphabetical listing of all SENAR terms. Core terms are defined normatively in SENAR Standard Section 3.

Term	Definition
AI Agent	Software system powered by LLM that generates engineering artifacts under human direction
AI Model Provider	External service providing AI Agent capabilities (e.g., Anthropic, OpenAI, Google); de facto supplier (Standard 3.25)
AI Model Version	Specific version/generation of an AI model; metric baselines are version-dependent (Standard 3.26)
Checkpoint	Context preservation action during a Session
Context	Information provided to AI Agent: goal, AC, constraints, knowledge, traceability
Context Architect	Responsibility: designs requirements as structured AI input, manages traceability
Cost per Task	Metric: total cost / tasks done, segmented by complexity
Cost Predictability	Metric: actual cost / planned cost for an Increment
Cycle Time	Time from Task start to completion (started_at → completed_at)
Dead End	Documented failed approach with reason for abandonment
Defect Escape Rate	Metric: `post_done_defects / done × 100%` — defect tasks linked to a completed parent, over Tasks completed (Standard 9.1)
Delivery Review	Ceremony: demonstrate software to stakeholders
Exploration	Time-bounded investigation without full Task formality
Federation	Coordination mechanism for multiple Supervisor+AI Pairs
Federation Sync	Ceremony: coordinate multiple Pairs on dependencies
First-Pass Success Rate	Metric: % Tasks completed correctly in one cycle
Flow Manager	Responsibility: session rhythm, cost tracking, flow metrics
Gate Bypass	Documented exception to proceed past a Quality Gate. Admission of a gate's declared effect without a positive verdict is a bypass whether or not it was intended as one, and is recorded as one (Standard 8.6(h), 8.6(j))
Gate Properties	The five properties a check has to have to be a Quality Gate: declared prevented effect, preventive placement, correspondence of the verdict to

	state, binding to a measured digest, fail-closed. Obligation levels scale by configuration (Standard 8.6)
Increment	Scope-bounded batch of work with objectives and budget
Increment Planning	Ceremony: define objectives, tasks, budget, risks
Increment Retrospective	Ceremony: quantitative review of Increment metrics
Knowledge Capture Rate	Metric: knowledge entries / tasks done
Knowledge Engineer	Responsibility: capture, curate, maintain knowledge base
Knowledge Entry	Documented decision, pattern, gotcha, or dead end
Lead Time	Time from Task creation to completion
Manual Intervention Rate	Metric: % Tasks with a Gate Bypass recorded under Standard 8.6(j). Derived from records, not from self-report (Standard 9.2)
Maturity Level	Organization's SENAR adoption depth (L1–L5)
Quality at Input	Principle: defects in requirements cascade to all downstream artifacts; quality is built at input, not checked at output
Quality Gate	Enforcement point that prevents a declared effect until its verdict is positive. Meeting the criteria of 8.1–8.5 is not sufficient: a check lacking a property required at the applicable configuration is not a Quality Gate (Standard 8.6)
Prevented effect	The specific change to the system, or to the record of work, that does not occur while a gate's verdict is negative. Declared per gate in Standard 8.1–8.5; required by 8.6(a)
Quality Sweep	Ceremony: periodic comprehensive codebase/KB audit
Requirement	Documented, verifiable statement of need at BR, SR, or TR level (Standard 3.17)
Requirement — Business (BR)	Stakeholder-level need in business terms; source of all downstream requirements (Standard 3.18)
Requirement — System (SR)	System-level capability derived from BR; what the system must do (Standard 3.19)

Requirement — Task (TR)	Implementation-level requirement = Task goal + acceptance criteria (Standard 3.20)
Requirement Hierarchy	Decomposition chain: BR → SR → TR; depth scales by complexity (Standard 3.21)
Requirement Library	Managed repository of verified, reusable requirements stored in Knowledge Base (Team+)
Requirement Pattern	Reusable AC template for common task types (CRUD, migration, UI, integration)
Session	Time-bounded period of supervised AI work
Session End	Ceremony: capture metrics, write handoff, record knowledge
Session Start	Ceremony: load context, select tasks, verify environment
Story	Intermediate grouping of Tasks as stakeholder-visible deliverable
Supervisor	Human who directs AI agents, verifies output, enforces gates. Production of artifacts is AI-directed; direct modification by the Supervisor is a recorded Gate Bypass (Standard 3.2, 4.1)
Supervisor+AI Pair	Fundamental production unit: one Supervisor + AI Agent(s)
Task	Atomic unit of tracked work with goal and acceptance criteria
Throughput	Metric: tasks completed per Session
Traceability	Bidirectional linking: every TR → BR upward; every BR → TR(s) downward (Standard 3.24)
Value Stream	End-to-end flow from client request to delivered software
Verification Engineer	Responsibility: audits AI output for correctness and security
Work Type	Functional category of a Task (dev, arch, QA, docs)
WSJF	Prioritization: Cost of Delay / Job Size (from SAFe)

SENAR Reference: Scaling Ratios

These ratios are guidelines. Organizations should adjust based on their domain, AI tooling capabilities, and team maturity.

Responsibility Allocation by Team Size

Pairs	Supervisor	Context Architect	Knowledge Engineer	Flow Manager	Verification Engineer
1–2	1 (covers all)	— (absorbed)	— (absorbed)	— (absorbed)	— (absorbed)
3–5	3–5	1 (covers CA+KE)	(covered by CA)	1 (covers FM+VE)	(covered by FM)
6–10	6–10	1 dedicated	1 dedicated	1 dedicated	1 dedicated
10–20	10–20	2 (by domain)	1	2	2
20–50	20–50	3–5	1–2	2–3 (Federation Coordinators)	3–5
50+	50+	5–10	2–3	5+ (incl. Portfolio Manager)	5–10

Enterprise Roles

Role	Ratio	When Needed
Portfolio Manager	1 per 20–50 Pairs	Multiple value streams or budget oversight
Chief Supervisor	1 per organization	Architectural governance, QG standards
Federation Coordinator	1 per 10–20 Pairs	Cross-team dependency management
Compliance Officer	1 per organization	Regulated industries (ISO, PCI, HIPAA)

Quality Sweep Staffing

Pairs	VE Coverage
1–2	Supervisor self-audits
3–5	1 VE, sweeps rotate across Pairs
6–10	1 dedicated VE, full sweep each cycle
10+	1 VE per 5–10 Pairs, coordinated sweep schedule

SENAR Reference: Efficiency Model

This document provides frameworks for evaluating the efficiency of SENAR adoption. All models use **ratios and multipliers**, not absolute monetary values — organizations substitute their own numbers in any currency.

A. Efficiency Dimensions

SENAR efficiency is measured across four dimensions:

Dimension	What It Measures	Key Ratio
Throughput	Output per production unit	Tasks per Supervisor+AI Pair vs tasks per traditional developer
Quality	Defect prevention and detection cost	Defect cost ratio: early detection vs late detection
Knowledge	Organizational learning retention	Knowledge reuse rate: entries that prevent repeated mistakes
Overhead	Process cost as fraction of delivery	Ceremony + gate time as % of productive session time

---

B. Throughput Model

B.1 Production Unit Comparison

Model	Production Unit	Typical Composition
Traditional	Development Team	5–9 engineers + QA + PM
SENAR	Supervisor+AI Pair	1 Supervisor + AI Agent(s)

Throughput multiplier:

```
T_multiplier = Tasks_per_Pair_per_period / Tasks_per_Developer_per_period
```

Organizations SHOULD measure this ratio during pilot to establish their baseline. The ratio varies significantly by: domain complexity, AI tool capability, Supervisor experience, and context quality.

B.2 Scaling Efficiency

Traditional scaling: adding developers has diminishing returns (Brooks's Law — communication overhead grows as n^2).

SENAR scaling: adding Supervisor+AI Pairs has near-linear returns up to the federation coordination limit, because Pairs operate independently with programmatic dependency tracking.

```
Traditional: Effective_capacity = n × Developer_output × (1 -
communication_overhead(n))
SENAR:       Effective_capacity = n × Pair_output × (1 -
federation_overhead(n))
```

Where `federation_overhead(n)` grows slower than `communication_overhead(n)` because dependencies are tracked programmatically, not through meetings.

C. Quality Efficiency

C.1 Defect Detection Cost Ratio

Defects caught earlier cost less to fix. This ratio is consistent across organizations:

Detection Point	Relative Cost
During AI generation (same session)	**1×** (baseline)
During Quality Sweep (periodic audit)	**3–5×**
During acceptance testing	**5–10×**
In production	**10–50×**

Quality Gate ROI:

```
Gate_ROI = (Defects_caught × Avg_late_detection_cost) /
Gate_operation_cost
```

Organizations SHOULD measure defect counts at each stage and calculate their own cost ratios.

C.2 First-Pass Efficiency

Higher First-Pass Success Rate (FPSR) means less rework:

```
Rework_cost = Tasks_total × (1 - FPSR) × Avg_rework_cost_per_task
Efficiency_gain = Rework_cost_before_SENAR - Rework_cost_after_SENAR
```

FPSR improves as context quality improves (better acceptance criteria, richer knowledge base, documented dead ends).

D. Knowledge Efficiency

D.1 Dead End Reuse

Each documented Dead End prevents future Supervisors from repeating a failed approach.

```
Dead_End_ROI = Documented_dead_ends × Avg_times_would_be_repeated ×
Avg_exploration_cost
```

The reuse rate for well-documented Dead Ends approaches 100% — nearly every documented dead end prevents at least one repeat within the organization.

D.2 Knowledge Accumulation Effect

As the knowledge base grows, context quality improves, which improves FPSR, which reduces rework:

```
Session N:    FPSR = f(KB_size_at_N, Context_quality)
Session N+K:  FPSR' > FPSR  (if KB is maintained and growing)
```

This creates a compound efficiency gain — each Increment is more efficient than the last, up to the plateau where most common patterns and pitfalls are documented.

E. Overhead Model

E.1 Process Overhead Ratio

```
Overhead_ratio = Time_on_ceremonies_and_gates / Total_session_time
```

Target: overhead < 15% of session time for Core/Foundation, < 20% for Team.

Activity	Core/Foundation	Team
Session Start	2–5 min	2–5 min
Session End	5–10 min	5–10 min
Quality Gate checks	Automated (0 min)	Automated (0 min)
Quality Sweep	Periodic (amortized)	Periodic (amortized)
Federation Sync	N/A	5–10 min per sync
Increment Planning	Amortized	1 session per Increment
Retrospective	Amortized	30–60 min per Increment

E.2 Overhead Break-Even

SENAR overhead pays for itself when defect prevention savings exceed process cost:

```
Break_even: Gate_cost + Ceremony_cost < Defects_prevented ×
Avg_defect_cost
```

Organizations SHOULD calculate this after 3 Increments with measured data.

F. Comparison Framework

F.1 Traditional Team vs SENAR Team

To compare delivery efficiency for the same scope:

Metric	Traditional Team	SENAR Team	How to Measure
Headcount	N developers + QA + PM	M Supervisors + support roles	Count
Throughput	Tasks per period	Tasks per period	Same task granularity
Defect rate	Defects per 100 tasks	Defects per 100 tasks	Same counting method
Lead time	Requirement → production	Requirement → production	Same milestones
Rework rate	% tasks requiring rework	% tasks requiring rework (1 - FPSR)	Same definition
Knowledge retention	Bus factor, onboarding time	KB coverage, onboarding time	Measured

F.2 Decision Criteria

SENAR is more efficient when:

- AI tools can generate the majority of implementation artifacts for the domain
- Throughput multiplier (B.1) exceeds overhead ratio (E.1) — net productivity gain
- Defect prevention savings (C.1) exceed quality gate costs — net quality gain
- Knowledge accumulation (D.2) provides compounding returns over time

SENAR is less efficient when:

- Domain is poorly suited for AI generation (novel research, highly regulated manual processes)
- AI tooling costs exceed the value of throughput gains
- Organization cannot invest in tooling infrastructure (task tracker, CI/CD, knowledge base)
- Team is too small to benefit from process structure (1 developer on a side project)

G. Evaluation Worksheet

Organizations evaluating SENAR adoption SHOULD measure these during a pilot (minimum 3 Increments):

Metric	Pilot Value	Baseline (before SENAR)	Delta
Tasks per Pair per session	___	___ (per developer)	×___
FPSR	___%	N/A (new metric)	—
Defect Escape Rate	___%	___%	___%
Rework rate	___%	___%	___%
Session overhead (min)	___	N/A	—
Knowledge entries created	___	0	+___
Dead Ends documented	___	0	+___

Decision Rule

```
IF throughput_multiplier > 1.0 + overhead_ratio
AND defect_escape_rate <= baseline_defect_rate
THEN SENAR is providing net efficiency gain → consider scaling
```

H. Red Flags

Signs that SENAR adoption is reducing rather than improving efficiency:

Red Flag	What It Means	Action
Overhead ratio > 30%	Process overhead exceeds value	Simplify: reduce to MVS, automate ceremonies
FPSR declining over time	Context quality degrading	Audit knowledge base, review AC quality
Throughput multiplier < 1.0	Pairs slower than traditional developers	Wrong domain for AI, or insufficient Supervisor training
Criteria-bypass rate > 20%	Gates don't match reality	Recalibrate gate criteria
Manual Intervention Rate > 30%	Context quality or agent capability, not gate discipline	Improve context for the affected task types (Guide 06, PF-1)
Knowledge entries = 0	Knowledge capture abandoned	Reinforce Dead End documentation at minimum
Sessions consistently exceed duration limit	No discipline	Enforce checkpoints, review causes

SENAR Reference: Governance and Compliance Annex

This annex maps SENAR practices to governance, regulatory, and compliance requirements for organizations using AI-native development teams. It is intended for compliance officers, auditors, legal counsel, and engineering leadership evaluating SENAR adoption in regulated environments.

Disclaimer: This document provides recommended practices and compliance mapping guidance. It does not constitute legal advice. Organizations should consult qualified legal counsel for jurisdiction-specific regulatory interpretation.

A. Responsibility Model

A.1 The Accountability Principle

In SENAR, **the Supervisor is accountable for all AI output they approve**. This is not a delegation of responsibility to the AI agent — it is an explicit acceptance of responsibility by the human who directs, reviews, and approves the work.

The accountability chain follows a clear principle:

> ***AI generates. Human approves. Human is responsible.***

This principle is embedded structurally in SENAR through Quality Gates. When a Supervisor passes a Task through QG-2 (Implementation Gate), they are attesting that:

- CI passes, tests pass, types are clean (automated verification);
- Acceptance criteria are satisfied (human judgment);
- No security vulnerabilities were detected (tooling-assisted);
- The output is fit for its intended purpose (professional judgment).

This attestation is the compliance-relevant act. It is recorded, timestamped, and attributable to a specific individual.

A.2 Responsibility by Quality Gate

Gate	Who Is Responsible	What They Attest
QG-0 (Context)	Supervisor / Context Architect	Task is well-defined with verifiable acceptance criteria
QG-1 (Requirements)	Context Architect	Business requirement is approved and properly decomposed
QG-2 (Implementation)	Supervisor	AI output meets acceptance criteria, CI passes, no security issues
QG-3 (Verification)	Verification Engineer / Peer Reviewer	Code reviewed, acceptance tests pass, no regressions
QG-4 (Acceptance)	Stakeholder / Context Architect	Software meets business requirements, ready for release

Each gate passage constitutes a documented approval decision by an identified human.

A.3 Gate Bypass Responsibility

SENAR Standard Section 8.6(h) requires that Gate Bypasses include justification, risk assessment, remediation plan, and senior approval. From a compliance perspective:

- **The person who approves the bypass owns the risk** for any downstream consequences.
- Gate Bypass records SHALL be preserved as audit evidence.
- Organizations SHOULD track Gate Bypass rate (Standard 8.6(i)) as a compliance health metric.
- Auditors should read an elevated Gate Bypass rate by its composition before drawing a conclusion, per B.3 item 7: a high 8.6(j) share is a statement about context quality and agent capability, while a high criteria-bypass share is a statement about gate discipline. Only the second is an indicator of process pressure; the merged figure is not.

A.4 Escalation Paths

Situation	Escalation Path
Supervisor is uncertain about AI output correctness	Escalate to Verification Engineer or peer Supervisor for independent review
AI output touches high-risk areas (security, auth, payment, data migration)	SHALL trigger peer review per risk-based review policy (Standard 8.7)
Defect found post-release in AI-generated code	Incident response process (Section F); trace to originating Task and Supervisor
Supervisor suspects AI hallucination or fabricated references	Stop, verify independently, document as Dead End if approach is abandoned
Regulatory or compliance implications unclear	Escalate to Compliance Officer (or equivalent organizational role) before proceeding

A.5 Shared Responsibility in Team+ Configurations

In Team+ configurations, responsibility is distributed across dedicated roles:

- **Context Architect** is responsible for requirement quality and traceability completeness.
- **Knowledge Engineer** is responsible for knowledge base accuracy and freshness.
- **Flow Manager** is responsible for process adherence and metric collection.
- **Verification Engineer** is responsible for independent verification and Quality Sweep thoroughness.
- **Supervisor** remains responsible for the specific AI output they direct and approve.

This distribution does not dilute individual accountability — it creates a chain of documented responsibilities, each with its own audit trail.

B. Audit Trail Requirements

B.1 SENAR Artifacts as Audit Evidence

SENAR's structure produces the following artifacts, each of which constitutes audit evidence:

Artifact	What It Contains	Compliance Value
Task record	Goal, acceptance criteria, requirement link, work type, lifecycle timestamps, assigned Supervisor	Demonstrates planned, authorized work with traceability
Quality Gate records	Gate ID, pass/fail, timestamp, approver, automated check results, and the digest identifying the state the verdict was issued for (Standard 8.6(g), 8.6(d))	Demonstrates enforcement of controls at each stage, and that each verdict is attributable to a specific state rather than to the work in general
Gate Bypass records	Justification, risk assessment, remediation plan, approver	Demonstrates controlled exception management
Session logs	Start/end timestamps, tasks worked, checkpoint records, handoff notes	Demonstrates supervised execution with bounded scope
Handoff documents	Session summary, next steps, warnings, open issues	Demonstrates continuity of supervision across sessions
Knowledge entries	Decisions, patterns, Dead Ends, gotchas — timestamped and categorized	Demonstrates organizational learning and rationale capture
Increment Planning records	Objectives, task pool, budget, risk register	Demonstrates planned delivery with risk management
Retrospective records	Metrics review, planned vs actual, improvement actions	Demonstrates continuous improvement and management review
Quality Sweep reports	Audit scope, findings, remediation actions	Demonstrates periodic independent verification
Metrics data	Throughput, Lead Time, FPSR, DER, Cost Predictability, etc.	Demonstrates measurement-based process management

B.2 Quality Gate to Audit Evidence Mapping

Quality Gate	Evidence Produced	Demonstrates
QG-0 (Context)	Task created with goal, AC, requirement link	Work authorization and scope definition
QG-1 (Requirements)	Approved requirement, decomposition record	Requirements management and approval
QG-2 (Implementation)	CI results, test results, Supervisor approval record	Verification of deliverable, change control
QG-3 (Verification)	Review record, acceptance test results, security scan	Independent verification, security assessment
QG-4 (Acceptance)	Stakeholder approval, staging verification, Delivery Review record	Release authorization, acceptance testing

B.3 Demonstrating Adequate Supervision

An auditor evaluating whether AI-generated code was adequately supervised should examine:

1. **Task definition quality.** Was QG-0 enforced? Did the Task have clear, verifiable acceptance criteria before AI work began? A well-defined Task demonstrates intentional direction, not open-ended AI generation.

2. **Gate passage records.** Did QG-2 pass with documented Supervisor approval? Were automated checks (CI, tests, lint, security scan) executed and recorded? Automated gate records are stronger evidence than manual checklists.

3. **Risk-appropriate review.** Was the risk level correctly classified? Did high-risk changes receive peer review as required by Standard 8.7? Consistent application of risk-based review is a key indicator.

4. **Session discipline.** Were Sessions bounded in duration? Were checkpoints performed at the required cadence? Unbounded sessions without checkpoints suggest inadequate attention.

5. **Knowledge capture.** Were decisions documented? Were Dead Ends recorded? Active knowledge capture indicates reflective supervision, not passive acceptance.

6. **Metrics trend.** Is the Defect Escape Rate stable or improving? A rising DER suggests degrading supervision quality.

7. **Gate Bypass rate, read as two figures.** Bypasses under Standard 8.6(j) — direct modification of an artifact — are recorded alongside bypasses of a gate's criteria, and Standard 9.2 states the two are nested and are not to be added together. Ask for them separately. A high 8.6(j) figure is a statement about context quality and agent capability; a high criteria-bypass figure is a statement about gate discipline, and only the second one is the finding this item is looking for.

What gate records prove, and what an auditor should not infer from them

Standard 8.3 states the limit and this annex repeats it, because it is the point at which an audit of SENAR artifacts can overreach. A gate record bound to a measured digest proves that the named checks ran against that state and that the verdict was not carried over from a different one. It does not prove the good faith of a party authorized to record approval: a participant holding that authority who intends to mislead can select checks that pass, or approve work they did not examine, and every binding will be intact throughout.

Assurance against a purposeful participant comes from review by a party that does not hold the generating role (Standard 10.15, L3) and from separation of authority between profiles (Standard 5.2). An auditor seeking that assurance should examine those controls, not ask for more checks at QG-2 — adding checks to a gate strengthens it against drift, staleness, and substitution, which is not the exposure in question.

B.4 Retention Recommendations

Artifact Category	Recommended Minimum Retention	Rationale
Task and gate records	Duration of software lifecycle + regulatory retention period	Core traceability evidence
Session logs and handoffs	3 years or regulatory minimum (whichever is longer)	Supervision evidence
Knowledge entries	Indefinite (active curation)	Ongoing operational value
Metrics data	3 years minimum	Trend analysis and audit evidence
Gate Bypass records	Duration of software lifecycle	Risk acceptance evidence
Incident records	Per regulatory requirement (typically 5-7 years)	Post-incident analysis

C. Regulatory Mapping

C.1 ISO 9001:2015 — Quality Management Systems

NOTE: ISO 9001 is a paid publication and its clause text was not read (Standard, Section 0, survey coverage). Clause numbers and titles below are the published ones; the "Requirement Summary" column is this document's paraphrase from public commentary, not a quotation, and an organization claiming ISO 9001 conformance reads the source.

ISO 9001 Clause	Requirement Summary	SENAR Artifact	How SENAR Satisfies
6.1 — Actions to address risks and opportunities	Identify risks affecting QMS, plan actions	Increment Planning risk register; Gate Bypass risk assessments; risk-based review classification (Standard 8.7)	Each Increment begins with documented risk identification. Gate Bypasses require explicit risk assessment. Review depth is determined by risk level.
7.5 — Documented information	Create, update, and control documented information	Task records, gate records, knowledge entries, session logs, handoffs	All SENAR artifacts are timestamped, attributable, and retained. Knowledge entries are curated for freshness (Knowledge Engineer responsibility).
8.1 — Operational planning and control	Plan and control processes for product/service provision	Increment Planning with objectives, task pool, and budget; QG-0 enforcing task definition before work	Work is planned at Increment level, authorized at Task level, and controlled through automated Quality Gates.
8.5 — Production and service provision	Controlled conditions for production	Session discipline (bounded duration, checkpoints); Supervisor direction of AI agents; automated CI/CD enforcement	AI production occurs under supervised, time-bounded conditions with automated controls. Manual interventions are tracked.
8.6 — Release of products and services	Verify product/service requirements met before release	QG-4 (Acceptance Gate): stakeholder approval, staging verification, QG-3 still passing	Release requires documented verification and stakeholder acceptance.
9.1 — Monitoring, measurement, analysis, and evaluation	Determine what to monitor and measure	10 defined metrics (4 mandatory in 9.1, 6 recommended in 9.2); Increment Retrospective with quantitative review	Metrics are collected automatically, reviewed at each Retrospective, with baselines established over 3+ Increments.

10.1 — Improvement: Nonconformity and corrective action	React to nonconformities, take corrective action	Retrospective improvement actions (specific, measurable, time-bounded, assigned); Quality Sweep findings and remediation	Retrospectives produce assigned corrective actions. Quality Sweeps identify nonconformities. Dead Ends document failed approaches.

C.2 SOC 2 Type II — Trust Services Criteria

SOC 2 Criterion	Requirement Summary	SENAR Artifact	How SENAR Satisfies
CC6.1 — Logical and physical access controls	Restrict access to authorized users and processes	SENAR does not prescribe access control tooling, but: Supervisor role assignment controls who may approve gate passage; Task assignment documents authorized workers; Session logs document who performed what work and when	Organizations must supplement SENAR with infrastructure-level access controls. SENAR provides the process-level authorization records.
CC7.1 — Detection of unauthorized or malicious activity	Monitor and detect anomalies	Quality Sweeps detect scope creep, unauthorized changes, configuration drift; Version control rules (Standard 10.6) mandate atomic commits with secrets detection; DER metric tracks defects escaping gates	Quality Sweeps and automated scanning provide detection. Organizations must supplement with infrastructure monitoring.
CC7.2 — Monitoring of system components	Monitor system components to detect anomalies	Session logs document all AI-directed changes; Gate records document all approvals; Metrics track process health indicators (FPSR, DER)	SENAR provides process-level monitoring. Organizations must supplement with system-level monitoring (APM, logging, alerting).
CC8.1 — Changes to infrastructure, data, software, and procedures	Manage changes through a controlled process	QG-0 through QG-4 constitute a change management pipeline; Task records document change authorization; Gate records document change verification; Gate Bypasses document controlled exceptions	SENAR's Quality Gate pipeline is a change management process. Every change is authorized (Task), verified (QG-2), independently reviewed (QG-3 for Team+), and released (QG-4).

C.3 GDPR — Considerations for AI Tools Processing Personal Data

GDPR applies when AI coding tools process personal data. This can occur when:

- Source code contains personal data (e.g., test fixtures with real user data, hardcoded credentials);
- AI prompts include personal data from production systems (e.g., debugging with real data);
- AI tool providers process and potentially retain prompt data;
- Session logs contain personal data referenced during development.

GDPR Consideration	Recommended SENAR Practice
Lawful basis for processing	Organizations should establish lawful basis (typically legitimate interest or contract performance) for any personal data sent to AI tools. Document this in data processing records.
Data minimization	QG-0 context preparation SHOULD exclude personal data. Use synthetic or anonymized data in AI prompts. Include data classification check in Quality Sweep scope.
Data processing agreements	Organizations processing EU personal data via cloud-hosted AI tools SHALL have Data Processing Agreements (DPAs) in place with AI tool providers covering prompt data processing and retention.
Right to erasure	Session logs containing personal data must be subject to erasure processes. Organizations should design session logging to minimize personal data capture.
Data protection impact assessment	Organizations SHOULD conduct a DPIA before adopting AI coding tools that will process personal data, particularly when using cloud-hosted models.
Cross-border transfers	When AI tools process data outside the EEA, organizations must ensure adequate transfer mechanisms (SCCs, adequacy decisions) are in place.

C.4 General Software Audit Requirements

Audit Requirement	SENAR Artifact	How SENAR Satisfies
Traceability from requirement to implementation	Task → requirement link (QG-0 mandate); Story → Task decomposition; Increment objectives → Stories	Every Task links to its parent requirement. Full traceability chain from business objective to implementation.
Change authorization	Task creation (authorization to work); QG-2 passage (authorization to close); QG-4 passage (authorization to release)	Changes are authorized at three levels: work initiation, implementation approval, and release.
Segregation of duties	Supervisor directs AI; Verification Engineer reviews independently (Team+ configurations); Stakeholder approves release	Team+ configurations provide role separation. Core/Foundation configurations should document compensating controls.
Evidence of testing	QG-2 mandates CI pass, tests pass; QG-3 mandates acceptance tests pass; automated test results recorded as gate evidence	Test execution is automated and recorded as part of gate passage.
Incident management	Task-based incident tracking; traceability from incident to originating code change; post-incident review process (Section F)	Incidents are traceable through the full chain: incident → code → Task → requirement → Supervisor.
Management review	Increment Retrospective with quantitative metrics; Quality Sweep reports; Delivery Review records	Regular, structured reviews with documented outcomes and improvement actions.
Continuous improvement	Retrospective improvement actions; DER and FPSR trending; Knowledge base growth	Improvement is measured (metrics), documented (actions), and verified (subsequent Retrospective review).

C.5 EU AI Act (Regulation 2024/1689)

AI coding assistants typically fall under 'limited risk' or 'general-purpose AI' categories. Key obligations:

- **Article 50: Transparency** — users must be informed they interact with AI. SENAR's explicit labeling of AI-generated work satisfies this.
- **Article 53: Obligations for GPAI model providers** — relevant when organizations fine-tune or deploy models.
- Organizations operating in the EU SHOULD review their AI tool usage against the Act's requirements and maintain documentation of compliance measures.

C.6 NIST AI Risk Management Framework (AI RMF 1.0)

SENAR practices align with NIST AI RMF functions:

- **Govern**: Roles (Section 4), Governance (this annex)
- **Map**: Risk-based review tiers (Section 8.7), Data classification (D.1)
- **Measure**: Metrics (Section 9), Quality Sweeps (Section 7.4)
- **Manage**: Quality Gates (Section 8), Incident Response (F.1)

Organizations subject to US federal AI governance requirements SHOULD document this mapping.

C.7 ISO/IEC 27001:2022

SENAR controls map to ISO 27001 Annex A:

- **A.8.25** (Secure development lifecycle): Quality Gates, Verification Checklist
- **A.8.28** (Secure coding): Rules 10.6, 10.15, AI Output Review
- **A.8.9** (Configuration management): Rules 10.13, 10.14
- **A.5.12** (Classification of information): Data classification (D.1)
- **A.8.15** (Logging): Audit trail requirements (B.1)

Organizations seeking ISO 27001 certification SHOULD map SENAR artifacts to their Statement of Applicability.

D. Data Governance for AI Tools

D.1 Data Classification for AI Interactions

Before sending any data to an AI coding tool, organizations should classify that data and apply appropriate controls:

Classification	Definition	AI Tool Policy	Examples
Public	Information intended for public disclosure	Any AI tool (cloud or on-premise)	Open-source code, public documentation, published APIs
Internal	Information for internal use, low sensitivity	Cloud AI tools with DPA; on-premise AI tools	Internal architecture docs, non-sensitive business logic, internal tooling code
Confidential	Sensitive business information	On-premise AI tools preferred; cloud only with explicit DPA, encryption, and no-retention guarantees	Proprietary algorithms, customer-facing feature code, competitive differentiators
Restricted	Highest sensitivity; regulatory or contractual obligations	On-premise AI tools only; no cloud transmission	Personal data (GDPR), payment card data (PCI DSS), health records (HIPAA), authentication secrets, encryption keys

D.2 AI Tool Selection Criteria

Organizations should evaluate AI coding tools against the following criteria, weighted by their data classification requirements:

Criterion	Questions to Answer
Data retention	Does the provider retain prompts or generated output? For how long? Can retention be disabled?
Training opt-out	Does the provider use customer data for model training? Can this be contractually excluded?
Data residency	Where is data processed and stored? Does this comply with applicable data residency requirements?
Encryption	Is data encrypted in transit (TLS 1.2+) and at rest? Who holds the encryption keys?
Access controls	Who at the provider can access customer data? Under what circumstances?
Audit rights	Does the contract include audit rights or third-party audit reports (SOC 2, ISO 27001)?
Subprocessors	Does the provider use subprocessors? Are they disclosed and contractually bound?
Incident notification	What are the provider's breach notification obligations and timelines?

D.3 On-Premise vs. Cloud AI: Decision Framework

Factor	Cloud AI Appropriate	On-Premise AI Required
Data classification	Public or Internal	Confidential or Restricted
Regulatory environment	No data residency requirements	Data sovereignty or residency mandates
Contractual obligations	No customer restrictions on cloud processing	Customer contracts prohibit cloud AI processing
Risk tolerance	Organization accepts cloud provider risk	Zero tolerance for external data exposure
Cost-performance trade-off	Cloud cost is acceptable for capability	On-premise cost justified by compliance requirements

Organizations operating in regulated industries (finance, healthcare, government, defense) should default to on-premise AI tools for all non-public data unless a specific risk assessment justifies cloud use.

D.4 Data Retention for AI Interactions

Data Type	Retention Recommendation	Notes
Session logs (sanitized)	Per audit trail retention policy (Section B.4)	Remove personal data before long-term retention
Raw AI prompts	Minimal retention; purge after session close unless regulatory requirement	Prompts may contain sensitive context
AI-generated output	Retained as part of version-controlled source code	Standard code retention policies apply
Task and gate records	Duration of software lifecycle + regulatory period	Core compliance artifacts
AI tool usage metrics	3 years minimum	Cost tracking, audit evidence

D.5 Right to Deletion

A critical question for organizations subject to GDPR or similar privacy regulations: if personal data was included in an AI prompt, can its influence be removed from the model?

Current reality: For cloud-hosted AI models, organizations generally cannot guarantee removal of training influence from model weights, even if the provider deletes stored prompt data. This is a fundamental limitation of current large language model architectures.

Recommended practices:

1. **Prevent rather than remediate.** Data classification and QG-0 context preparation should exclude personal data from AI prompts. Prevention is more reliable than post-hoc deletion.
2. **Contractual no-training clauses.** Ensure AI tool contracts explicitly exclude customer data from model training.
3. **Prompt data deletion.** Ensure contracts include prompt data deletion obligations and verify compliance.
4. **On-premise models for restricted data.** When deletion guarantees are required, use on-premise models where the organization controls the full data lifecycle.
5. **Document the limitation.** If an organization cannot guarantee deletion of training influence, document this as a known risk in the DPIA and apply compensating controls (data minimization, synthetic data).

E. Intellectual Property and Licensing

Note: AI-generated code IP is an evolving legal landscape with significant jurisdictional variation. The following represents recommended practices as of the time of writing. Organizations should obtain qualified legal counsel for their specific jurisdiction.

E.1 Ownership of AI-Generated Code

The legal status of AI-generated code ownership varies by jurisdiction and remains unsettled in many:

Jurisdiction Approach	Current Position	Implication for SENAR
Human authorship required	Several jurisdictions require human creative contribution for copyright. Purely machine-generated output may not be copyrightable.	SENAR's Supervisor model strengthens IP claims: the Supervisor provides creative direction (Task goals, acceptance criteria, architectural decisions) and exercises judgment in selection and approval.
Work-for-hire / employment	In many jurisdictions, employer owns work created by employees in course of employment.	AI tool output directed by employee Supervisors is likely covered under existing employment IP agreements, but organizations should verify.
AI tool provider terms	Most AI tool providers assign output rights to the user, but terms vary.	Organizations should review AI tool terms of service for IP assignment clauses and ensure they are compatible with organizational requirements.

Recommended practices:

1. **Update employment and contractor agreements** to explicitly address AI-assisted and AI-generated code.
2. **Review AI tool provider terms** to confirm output ownership assignment.
3. **Document human creative contribution** through SENAR's Task records (goal, acceptance criteria, architectural decisions, Supervisor review notes). This documentation strengthens authorship claims.
4. **Maintain records of Supervisor direction** — the context provided to the AI, the selection and modification of AI output, and the judgment applied during review.

E.2 License Contamination Risk

AI models trained on open-source code may generate output that reproduces or closely resembles code under copyleft licenses (GPL, AGPL, LGPL). If such output is incorporated into proprietary software without compliance, this creates license contamination risk.

Risk factors:

- AI models trained on public code repositories without license filtering;
- Generated code that closely matches existing open-source implementations;
- Insufficient review of generated code for license-encumbered patterns.

Recommended practices:

1. **Automated license scanning.** Include license scanning tools in the CI pipeline (enforced at QG-2). Scan both direct dependencies and generated code for known license-encumbered patterns.
2. **Dependency audit at QG-3.** For Team+ configurations, include dependency license audit in the QG-3 verification scope.
3. **Quality Sweep coverage.** Include license compliance in Quality Sweep scope (dependency health audit).
4. **AI tool selection.** Evaluate AI tool providers' training data policies and any indemnification they offer against IP claims.
5. **Code provenance documentation.** For high-value or high-risk code, document whether it was AI-generated, human-written, or a combination.

E.3 Documentation Requirements for IP Compliance

Organizations seeking to demonstrate IP compliance for AI-generated code should maintain:

Document	Purpose	SENAR Source
AI tool inventory	List of AI tools used, their terms of service, IP provisions	Organizational policy (outside SENAR scope)
Task records with Supervisor attribution	Document human creative direction and judgment	Task records, gate records, session logs
License scan results	Demonstrate absence of license contamination	QG-2 and QG-3 automated scan results
Dependency audit records	Document license compliance for all dependencies	Quality Sweep reports
Human contribution records	Strengthen authorship claims	Context preparation records, review notes, knowledge entries documenting design decisions

---

F. Incident Response for AI-Generated Defects

F.1 Response Process

When a production incident is traced to AI-generated code, the following process applies in addition to the organization's standard incident response:

1. **Triage and contain.** Standard incident response: assess severity, contain impact, restore service.
2. **Trace the origin.** Use SENAR's traceability chain:
    - Production incident → code change (version control);
    - Code change → Task record (commit references Task);
    - Task record → requirement (QG-0 requirement link);
    - Task record → Supervisor (Task assignment and QG-2 approval);
    - Task record → Session (session log with timestamps and context).
3. **Assess gate effectiveness.** For each Quality Gate the defective code passed through:
    - Did the gate execute? (Check gate records.)

    - Did automated checks pass? (Were checks adequate for this defect type?)
    - Did human review occur? (Was it risk-appropriate per Standard 8.7?)
    - Was a Gate Bypass involved? (If so, was the bypass risk assessment accurate?)

4. **Classify the root cause.** Use the AI-specific root cause taxonomy (Section F.2).

5. **Remediate.** Fix the immediate defect. Create a Task for any systemic improvement.

6. **Credential rotation.** If credentials or secrets are exposed (committed to VCS, included in AI context, or logged), the organization SHALL rotate affected credentials within 24 hours and audit access logs for unauthorized use.

7. **Post-incident review.** Conduct a structured review addressing:

    - Was supervision adequate for the risk level of this change?
    - Were Quality Gates effective? Should gate criteria be strengthened?
    - Was the AI agent's context sufficient? Should knowledge entries be created?
    - Should the risk classification for this type of change be elevated?

Organizations SHOULD define severity-based response timeframes for AI-generated vulnerabilities. Recommended: Critical — 4 hours, High — 24 hours, Medium — 72 hours, Low — next increment. This is a response-time scale for a vulnerability, and it is a third subject: Standard 10.15(f) requires an organization to document which scale it is defining, because severity of a finding, risk level of a change (Standard 8.7) and the checklist tier of SENAR Core share words. The Low level above is this Reference's own addition — Standard 10.15(c) names three severities, not four.

F.2 AI-Specific Root Cause Taxonomy

Standard root cause categories (logic error, integration error, performance issue) apply to AI-generated code. The following additional categories are specific to AI-native development:

Root Cause Category	Description	Indicator	Preventive Control
Hallucination	AI fabricated an API, library, or behavior that does not exist	Code references nonexistent functions, packages, or configurations	QG-2 automated checks (compilation, tests); Supervisor verification of external references
Context gap	AI lacked necessary context to produce correct output	Code contradicts existing architecture, duplicates existing functionality, or violates undocumented constraints	Improved QG-0 context preparation; knowledge base entries for constraints and conventions
Stale context	AI used outdated information (deprecated API, changed requirement)	Code uses deprecated patterns or APIs; behavior matches old requirements	Knowledge base freshness maintenance (Knowledge Engineer); Session Start context refresh
Gate bypass	Quality Gate was bypassed and the defect was in bypassed scope	Gate Bypass record exists for the relevant gate	Review Gate Bypass approval process; reduce bypass rate
Insufficient gate criteria	Quality Gate passed but criteria were inadequate to detect the defect	Gate records show passage; defect type was not covered by automated checks	Strengthen gate criteria; add test coverage for defect type
Scope creep	Changes beyond the Task scope introduced the defect, whatever produced them	Commit includes changes unrelated to Task acceptance criteria	QG-2 scope review (Standard 10.6(c)); atomic commit enforcement
Supervision gap	Supervisor approved without adequate review	No evidence of substantive review; rubber-stamp pattern in gate records	Session duration limits; review quality metrics; Verification Engineer audits
Accumulation effect	Multiple individually-correct AI changes interact	No single commit is defective; issue emerges from combination	Quality Sweeps (architectural conformance check);

	to create a systemic issue		integration testing at QG-3

F.3 Traceability Chain

The complete traceability chain for incident investigation:

```
Production Incident
  └► Code Change (commit hash, diff, timestamp)
        └► Task Record (goal, AC, work type, Supervisor)
              ├► Requirement (Story / business requirement)
              ├► QG-0 Record (context quality at task start)
              ├► QG-2 Record (implementation verification)
              │     ├► Automated check results (CI, tests, lint,
security)
              │     └► Supervisor approval (who, when)
              ├► QG-3 Record (if applicable: review, acceptance tests)
              ├► Session Log (when, duration, checkpoints, context)
              └► Knowledge Entries (decisions, Dead Ends, patterns)
```

This chain enables auditors and incident investigators to reconstruct the full decision history from business requirement through implementation to production deployment.

G. What SENAR Does Not Cover

SENAR is a process methodology for supervised AI-native development. The following compliance-relevant areas are outside SENAR's scope and require additional organizational controls:

Area	Why It Is Outside SENAR Scope	What Organizations Need
Infrastructure access controls	SENAR prescribes process roles, not system permissions	IAM policies, RBAC, MFA, privileged access management
Network security	SENAR is tool-agnostic; network architecture is implementation-specific	Firewalls, VPNs, network segmentation, DDoS protection
Data encryption standards	SENAR does not prescribe cryptographic controls	Encryption-at-rest, encryption-in-transit, key management
Business continuity / disaster recovery	SENAR addresses development process, not operational resilience	BCP/DR plans, backup procedures, RTO/RPO targets
Physical security	SENAR is a software development methodology	Facility access, environmental controls
Employee background checks	SENAR defines responsibilities, not HR processes	Background verification, security clearance procedures
Vendor management (beyond AI tools)	SENAR addresses AI tool data governance only	Third-party risk management program
Privacy by design	SENAR supports data minimization in AI prompts but does not replace privacy architecture	Privacy impact assessments, data mapping, consent management
AI model governance	SENAR governs AI usage in development, not AI model training or deployment	Model risk management, bias testing, model monitoring (relevant for organizations that train or fine-tune models)
Regulatory reporting	SENAR produces audit evidence but does not automate regulatory submissions	Reporting workflows, regulatory calendars, submission procedures

Organizations adopting SENAR should integrate it into their broader governance, risk, and compliance (GRC) framework rather than treating it as a standalone compliance solution.

G.1 Compensating Controls for Core/Foundation Configuration

SENAR's Core configuration (1 Pair) and Foundation configuration (1–3 Pairs) combine multiple responsibilities in fewer people. This creates a segregation-of-duties gap that may

concern auditors:

Gap	Compensating Control
Supervisor self-reviews (no independent Verification Engineer)	Automated gate enforcement (QG-2 automated checks are independent of the Supervisor); periodic Quality Sweeps; peer review for high-risk changes
Supervisor absorbs Context Architect role	QG-0 enforcement ensures minimum task definition quality regardless of who fills the role
No dedicated Flow Manager	Automated metrics collection reduces dependence on manual oversight; Session duration limits are enforceable through tooling

Organizations in regulated industries operating at Core or Foundation scale should document these compensating controls and assess whether they are adequate for their regulatory obligations. Transitioning to Team configuration may be necessary to satisfy segregation-of-duties requirements.

H. Implementation Checklist

Organizations adopting SENAR in a regulated environment should address the following:

- ☐ Establish data classification policy for AI tool interactions (Section D.1)
- ☐ Evaluate and select AI tools based on data classification and compliance requirements (Section D.2)
- ☐ Execute data processing agreements with cloud AI tool providers (Section D.3)
- ☐ Update employment and contractor IP agreements for AI-generated code (Section E.1)
- ☐ Integrate license scanning into CI pipeline at QG-2 (Section E.2)
- ☐ Configure Quality Gate automation to produce audit-grade records carrying all six elements Standard 8.6(g) requires — the gate, the unit of work, the verdict including its absence, the time, the state digest with the set it covers and the function used, and the checks the gate depended on with the result of each (Section B.1)
- ☐ Configure the paired record of the admission of the declared effect, naming the verdict relied on and the time (Standard 8.6(g)) — without it, 13.7 cannot check preventive placement, which is decided by the order of the two records
- ☐ Reassess any Team or Enterprise conformance claim against Standard 8.6(b), preventive placement — required by Standard 13.6 for claims made before revision 1.4

- [ ] Disclose, with the claim, every SHALL requirement handled under Standard 13.5 — required by Standard 13.1(e) since revision 1.5. An auditor reading a claim that discloses unimplemented SHOULD requirements and nothing else is reading an incomplete claim
- [ ] Establish Gate Bypass approval process with documentation requirements (Section A.3)
- [ ] Define data retention policies for SENAR artifacts (Section B.4)
- [ ] Conduct DPIA if AI tools will process personal data (Section C.3)
- [ ] Document compensating controls if operating at Core/Foundation configuration (Section G.1)
- [ ] Establish AI-specific root cause categories in incident management process (Section F.2)
- [ ] Include AI tool compliance in Quality Sweep scope (Section D, E)
- [ ] Train Supervisors on data classification and AI-specific compliance obligations
- [ ] Integrate SENAR audit trail into existing GRC tooling and reporting

---

SENAR Reference: Tooling Requirements

This document specifies what tooling must support for each SENAR component. Requirements are split into two tiers:

- **Core/Foundation** — minimum for SENAR Core (1 Pair) and Foundation (1-3 Pairs) configurations. Standard 12 places Core at Maturity Level 2 and does not describe Foundation, so no maturity level is claimed for Foundation here
- **Team+** — full requirements for SENAR Team and Enterprise configurations (3+ Pairs, Maturity Level 3+)

SENAR is tool-agnostic by design. These are capability requirements, not product recommendations.

> ***This document is informative.*** *Section 0 of the Standard states that the Guide and Reference are informative, and no conformance claim is assessed against them. The SHALL and SHOULD levels in the tables below state what tooling has to support in order to satisfy a requirement of the Standard; they are a reading of that requirement, not a second one. Where a row cites a clause of the Standard, that clause is the obligation and this document does not extend it. Where a row cites none, the row is this Reference's own recommendation and creates no obligation at all.*

---

1. Task Tracker

The task tracker is the system of record for all units of work. Every Task, Story, and Increment lives here.

1.1 Required Fields

Field	Type	Core/Foundation	Team+	Notes
`goal`	Text	SHALL	SHALL	What the Task must accomplish, stated as an outcome
`acceptance_criteria`	Text	SHALL	SHALL	Verifiable conditions for Task completion
`status`	Enum	SHALL	SHALL	Minimum states: `planning`, `active`, `done`, `blocked`
`work_type`	Enum	SHALL	SHALL	Values: `dev`, `arch`, `qa`, `docs`, `infra`, `ops`
`complexity`	Enum	SHOULD	SHALL	Values: `trivial`, `simple`, `moderate`, `complex`
`story_id`	Reference	SHOULD	SHALL	Link to parent Story or requirement
`requirement_link`	Reference	SHOULD	SHALL	Traceability link to BR/SR/TR (Standard 3.21)
`requirement_level`	Enum	—	SHALL	Values: `BR`, `SR`, `TR` — level of the linked requirement
`requirement_status`	Enum	—	SHALL	Values: `draft`, `approved`, `verified`, `deprecated`

`requirement_parent`	Reference	—	SHALL	Link to parent requirement (for hierarchy navigation)
`attempt_count`	Integer	SHOULD	SHALL	Number of implementation attempts (for FPSR calculation)
`created_at`	Timestamp	SHALL	SHALL	Auto-set on creation
`started_at`	Timestamp	SHALL	SHALL	Auto-set on transition to `active`
`completed_at`	Timestamp	SHALL	SHALL	Auto-set on transition to `done`
`session_id`	Reference	SHALL	SHALL	Which Session completed this Task
`supervisor_id`	Reference	—	SHALL	Who supervised this Task
`bypass_reason`	Text	SHALL	SHALL	Justification if any Quality Gate was bypassed — the first of the four elements 8.6(h) requires of a legitimate Gate Bypass. SHALL at every configuration of the Standard since revision 1.4: 8.6(h) makes the record the condition of the bypass being legitimate, and 8.6(j) makes direct modification of

				an artifact one of them. At Core the Standard levies nothing (Standard 13.3) and this row is a recommendation
`bypass_risk`	Text	SHALL	SHALL	The risk accepted by proceeding — second element of 8.6(h). Core as above
`bypass_remediation`	Text	SHALL	SHALL	The plan to remove the bypass, and by when — third element of 8.6(h). Core as above
`bypass_approver`	Reference	SHALL	SHALL	The senior approval the bypass rests on — fourth element of 8.6(h). A justification without it does not make the bypass legitimate, only recorded. Core as above
`knowledge_refs`	List	—	SHOULD	Links to relevant Knowledge Base entries
`cross_deps`	List	—	SHALL	Cross-project dependency references
`cost`	Numeric	—	SHOULD	Token/API cost for this Task

1.2 Required Automations

Automation	Core/Foundation	Team+	Description
QG-0 field validation	SHALL	SHALL	Block transition to `active` unless `goal` and `acceptance_criteria` are non-empty
State transition timestamps	SHALL	SHALL	Auto-capture `started_at`, `completed_at` on state change
State transition rules	SHOULD	SHALL	Enforce valid transitions: `planning` -> `active` -> `done`; `blocked` can enter/exit from any state
Attempt counter increment	SHOULD	SHALL	Increment `attempt_count` each time Task returns from `done` to `active`
Story completion roll-up	—	SHALL	Auto-calculate Story status from child Task statuses
Increment progress tracking	—	SHALL	Dashboard showing Increment completion percentage
Bypass tracking	SHALL	SHALL	Record which gate was bypassed, by whom, with what justification. SHALL at every configuration of the Standard since revision 1.4: 8.6(h) makes the record the condition of a legitimate bypass, and 8.6(j) makes direct modification of an artifact one of them. At Core the Standard levies nothing (Standard 13.3) and this row is a recommendation
Gate audit record	SHALL	SHALL	Every gate execution produces a record identifying the state the verdict was issued for (Standard 8.6(g))
Digest of the verdict's state	SHOULD (Core) / SHALL (Foundation)	SHALL	Record a digest measured over the artifacts the verdict is about, not only a revision identifier (Standard 8.6(d)). This is the one row in this table where Core and Foundation differ

Fail-closed on unreached verdict	SHALL	SHALL	A check that does not run, does not complete, or completes without a determinate result yields a negative verdict — never a pass by silence (Standard 8.6(e))
Duplicate detection	—	SHOULD	Warn when a new Task goal closely matches an existing Task
Requirement orphan detection	—	SHOULD	Identify requirements without implementation Tasks and Tasks without requirements
Requirement change impact	—	SHOULD	When a BR/SR changes, list all affected downstream artifacts

1.3 Required Reports

Report	Core/Foundation	Team+	Description
Throughput	SHALL	SHALL	Tasks completed per Session, with trend over time
Lead Time distribution	SHALL	SHALL	Histogram of `completed_at - created_at`, segmented by complexity
FPSR	SHALL	SHALL	% of Tasks completed with `attempt_count = 1`
DER	SHALL	SHALL	`post_done_defects / done × 100%` — defect tasks linked to a completed parent, over Tasks completed (Standard 9.1)
Task status summary	SHALL	SHALL	Count of Tasks by status, filterable by Increment/Session
Cycle Time	—	SHALL	`completed_at - started_at` distribution
Cost per Task	—	SHALL	Total cost / Tasks done, segmented by complexity and work type
Cost Predictability	—	SHALL	Actual vs. planned cost per Increment
Bypass rate	—	SHALL	% of **Tasks** carrying at least one recorded Gate Bypass, reported separately for criteria bypasses and for bypasses under 8.6(j), and segmented by gate type. The denominator is Tasks because that is the unit its consumers are stated in — Standard metric 8 (Manual Intervention Rate, Standard 9.2) and the PF-3 early-warning threshold in the Guide (Guide 06, *PF-3: Gate Bypass Normalization*). The two kinds SHALL NOT be added together (Standard 9.2). Bypasses per gate evaluation is a different figure with a different denominator — at Team a Task passes up to five gates — and SHALL NOT be compared against a threshold stated in Tasks
Supervisor workload	—	SHOULD	Tasks per Supervisor per Session, with complexity weighting

Cross-dependency status	—	SHALL	Status of all cross-project dependencies

---

2. CI/CD Pipeline

Automated enforcement of QG-2 (Implementation Gate) and supporting infrastructure for QG-3 and QG-4.

2.1 QG-2: Implementation Gate (Automated)

All checks run on every commit or merge request. Pipeline must block merge if any check fails.

Since Standard 1.5, a Foundation organization making a security-sensitive change performs the five AI Output Review Minimum Criteria of Standard 8.4 at this gate, because 8.7 requires the criteria at every configuration while 11.1 gives Foundation no QG-3. Tooling at Core/Foundation therefore needs somewhere to record those five results against the state the verdict is about (Standard 8.6(d), 8.6(g)) — the same record the Team+ column keeps at QG-3.

Check	Core/Foundation	Team+	Notes
Unit tests pass	SHALL	SHALL	100% pass rate required
Type checking passes	SHALL	SHALL	Language-appropriate: TypeScript strict, mypy, etc.
Linter passes	SHALL	SHALL	Zero warnings policy (not just errors)
Security scan (dependencies)	SHOULD	SHALL	Known CVE detection in dependencies
Security scan (SAST)	—	SHOULD	Static analysis for security anti-patterns
Build succeeds	SHALL	SHALL	Clean build from scratch, no cached state
Test coverage threshold	SHOULD	SHALL	Configurable minimum (recommendation: 70%+)
Dependency change detection	—	SHOULD	Flag new dependencies for review
Breaking change detection	—	SHOULD	API compatibility checks for shared interfaces

Pipeline configuration requirements:

Requirement	Core/Foundation	Team+
Pipeline runs automatically on push/MR	SHALL	SHALL
Pipeline blocks merge on failure	SHALL	SHALL
Pipeline results visible to Supervisor	SHALL	SHALL
No `allow_failure` — every job either works or is removed	SHALL	SHALL
Pipeline execution time < 10 minutes	SHOULD	SHALL
Pipeline provides clear failure messages	SHALL	SHALL
Pipeline is version-controlled (pipeline-as-code)	SHOULD	SHALL

2.2 QG-3: Verification Gate (Review)

Requirement	Core/Foundation	Team+	Notes
Branch protection on main/production branches	—	SHALL	No direct push
Merge request required	—	SHALL	All changes go through MR
Minimum 1 approval required	—	SHALL	Approver must be different from the Supervisor who created the MR
Review checklist integration	—	SHOULD	AI Review Checklist (Guide 02) embedded in MR template
Approval cannot be from the MR author	—	SHALL	Enforce independent verification
Stale approval invalidation	—	SHOULD	Re-approve after new commits

2.3 QG-4: Acceptance Gate (Deployment)

Requirement	Core/Foundation	Team+	Notes
Staging environment exists	—	SHALL	Mirrors production configuration
Staging deployment is automated	—	SHALL	Same pipeline as production, different target
Staging verification step	—	SHALL	Manual or automated acceptance test on staging
Production deployment requires explicit approval	—	SHALL	Human gate — not auto-deploy on merge
Rollback capability	—	SHALL	Ability to revert to previous version within minutes
Deployment tracking	—	SHALL	Record what was deployed, when, by whom
Smoke tests post-deployment	—	SHOULD	Automated verification that deployment succeeded
Feature flags	—	SHOULD	Ability to disable features without redeployment

---

3. Knowledge Base

The Knowledge Base is the persistent memory of the project. AI Agents have no memory between sessions — the KB is how knowledge survives.

3.1 Core Requirements

Requirement	Core/Foundation	Team+	Notes
Searchable (full-text)	SHALL	SHALL	Supervisor and AI Agent must be able to search by keyword
Categorized by entry type	SHALL	SHALL	Minimum types: `decision`, `pattern`, `gotcha`, `dead_end`, `observation`
Entry has title and body	SHALL	SHALL	Title for scanning, body for detail
Entry has creation timestamp	SHALL	SHALL	When was this knowledge created
Entry has project/scope tag	SHOULD	SHALL	Which project or area this knowledge applies to
Versioned entries	—	SHALL	Track changes to knowledge entries over time
Freshness tracking	—	SHALL	`last_reviewed` timestamp, updated when entry is confirmed still accurate
Entry status	—	SHALL	Values: `current`, `needs_review`, `deprecated`
Cross-references between entries	—	SHOULD	Link related knowledge entries
Bulk export	SHOULD	SHALL	Export all entries in a portable format (JSON, markdown)
Entry author tracking	—	SHALL	Who created or last updated this entry

3.2 Entry Types

Type	Purpose	Example
`decision`	Architectural or design choice with rationale	"We use UUIDs for all entity IDs because..."
`pattern`	Reusable approach for common tasks	"Error handling in API endpoints follows this structure..."
`gotcha`	Non-obvious behavior that causes problems	"CouchDB bulk_docs silently ignores conflicts unless..."
`dead_end`	Approach that was tried and abandoned, with reason	"Tried using WebSockets for SSE replacement — failed because..."
`observation`	Empirical finding worth remembering	"Build times increase 2x when running TypeScript strict mode on..."
`template`	Reusable requirement/AC pattern for common task types	"REST API endpoint AC: 1. Returns 200 for valid input. 2. Returns 401 without auth..."

3.3 AI Agent Integration

Requirement	Core/Foundation	Team+	Notes
AI Agent can search KB during session	SHALL	SHALL	Via tool/function call or context injection
AI Agent can create KB entries during session	SHALL	SHALL	Dead Ends and Gotchas discovered during work
Relevant KB entries injected into session context	SHOULD	SHALL	Based on task area, project, or explicit references
AI Agent can update existing KB entries	—	SHALL	Correct or extend existing knowledge
KB query results are structured	SHOULD	SHALL	Title, type, body, freshness — not raw text dumps
KB access is scoped	—	SHALL	Agent sees relevant project knowledge, not everything
KB entry creation triggers review queue	—	SHOULD	Human reviews AI-created knowledge entries

3.4 Freshness Management

Requirement	Core/Foundation	Team+	Notes
Entries older than N increments flagged for review	—	SHALL	Configurable threshold (recommendation: 3 Increments)
Freshness report available	—	SHALL	Distribution of entry ages, count by status
Deprecated entries excluded from AI context	—	SHALL	Stale context is worse than no context
Quality Sweep includes KB freshness audit	SHOULD	SHALL	Part of the ceremony checklist

---

4. Session Management

Sessions are SENAR's primary rhythm mechanism. The tooling must support the full session lifecycle.

4.1 Session Lifecycle

Requirement	Core/Foundation	Team+	Notes
Session start with timestamp	SHALL	SHALL	Records when the Session began
Session end with timestamp	SHALL	SHALL	Records when the Session ended
Session has unique identifier	SHALL	SHALL	For linking Tasks, metrics, handoffs
Session links to Supervisor	SHOULD	SHALL	Who ran this Session
Session links to Tasks worked	SHALL	SHALL	Which Tasks were active during this Session
Previous session handoff available at start	SHALL	SHALL	Context continuity

4.2 Checkpoint Capability

Requirement	Core/Foundation	Team+	Notes
Mid-session checkpoint	SHOULD	SHALL	Save context state without ending the Session
Checkpoint captures current task status	SHOULD	SHALL	What's in progress, what's done
Checkpoint reminder	—	SHOULD	Alert at configurable intervals (e.g., every 40-60 tool calls)
Checkpoint restores context on session crash	—	SHOULD	Recover from unexpected session termination

4.3 Handoff Documents

Requirement	Core/Foundation	Team+	Notes
Handoff document created at Session End	SHALL	SHALL	Summary of what was done and what's next
Handoff includes completed Tasks	SHALL	SHALL	List of Tasks done in this Session
Handoff includes in-progress Tasks	SHALL	SHALL	State of Tasks not finished
Handoff includes blockers/risks	SHOULD	SHALL	What might prevent the next Session from succeeding
Handoff includes knowledge captured	—	SHALL	KB entries created or updated during the Session
Handoff stored and searchable	SHOULD	SHALL	Historical handoffs available for reference
Handoff loaded automatically at next Session Start	SHOULD	SHALL	No manual searching for the latest handoff

4.4 Metrics Capture

Requirement	Core/Foundation	Team+	Notes
Tool call count per Session	SHOULD	SHALL	Leading indicator of session complexity and fatigue
Session duration	SHALL	SHALL	Derived from start/end timestamps
Tasks completed per Session	SHALL	SHALL	Throughput input
Session cost (tokens/API spend)	—	SHALL	For Cost per Task calculation
Error/retry count	—	SHOULD	How many times the AI needed correction

---

5. AI Agent Integration

Requirements for how AI Agents are configured, monitored, and managed within SENAR.

5.0 Structured Tool Protocol

Requirement	Core/Foundation	Team+	Notes
Platform capabilities exposed through a self-describing tool protocol	SHOULD	SHALL	Standard 5.6, whose level follows the Tooling row of Standard 11.4
Self-describing schemas with parameter types	SHOULD	SHALL	Standard 5.6(a) — reduces hallucinated arguments
Atomic tool invocations	SHOULD	SHALL	Standard 5.6(b) — one invocation, one transaction
Structured input and output	SHOULD	SHALL	Standard 5.6(c) — no parsing of text CLI output
Every invocation logged with parameters and results	SHOULD	SHALL	Standard 5.6(d)
Documented list of operations with CLI fallback	SHALL	SHALL	Standard 5.6 states this without a configuration qualifier
No direct database access for writes by AI agents	SHALL	SHALL	Standard 5.6, stated without a configuration qualifier

5.1 Context Injection

Requirement	Core/Foundation	Team+	Notes
Structured project instructions	SHALL	SHALL	Project-level context loaded at session start (e.g., CLAUDE.md, .cursorrules)
Task context injection	SHALL	SHALL	Goal, AC, and relevant constraints provided to agent
KB entry injection	SHOULD	SHALL	Relevant knowledge entries available during work
Codebase conventions	SHOULD	SHALL	Coding standards, architectural patterns, naming conventions
Scope boundaries	SHOULD	SHALL	Explicit "do not change" instructions
Session history/handoff injection	SHOULD	SHALL	Previous session context for continuity

5.2 Output Capture

Requirement	Core/Foundation	Team+	Notes
Code changes captured in version control	SHALL	SHALL	Standard git workflow
Session logs available for review	SHOULD	SHALL	What the AI did, what the Supervisor approved
AI reasoning visible	SHOULD	SHOULD	Why the AI made specific choices (for verification)
Output linked to Task	SHALL	SHALL	Which Task produced which code changes

5.3 Cost and Token Tracking

Requirement	Core/Foundation	Team+	Notes
Token usage per session	—	SHALL	Input + output tokens
API cost per session	—	SHALL	Dollar cost derived from token usage
Cost per task (derived)	—	SHALL	Session cost / tasks completed
Cost trend over time	—	SHALL	Is the process becoming more or less expensive?
Budget alerts	—	SHOULD	Warn when session/increment cost approaches budget
Cost by model/provider	—	SHOULD	If multiple models are used, track cost separately

5.4 Multi-Agent Orchestration (Team+)

Requirement	Core/Foundation	Team+	Notes
Workspace isolation	—	SHOULD	Agents work in isolated environments (worktrees, containers)
Agent-to-agent communication	—	SHOULD	Delegate subtasks to specialized agents
Parallel agent execution	—	SHOULD	Multiple agents working simultaneously under supervision
Agent output aggregation	—	SHOULD	Collect results from multiple agents into unified review
Agent capability profiles	—	SHOULD	Different agents for different task types (code, test, docs)

5.5 Agent Configuration Management

Requirements for managing Agent Profiles, Operational Scripts, and the Programmatic Interface (Standard Section 5).

Requirement	Core/Foundation	Team+	Notes
Script storage in version control	SHALL	SHALL	Scripts stored alongside project code
Script versioning (change history)	SHOULD	SHALL	Track who changed what, when, and why
Profile definition storage	SHOULD	SHALL	Agent Profiles stored as structured configuration
Profile switching support	—	SHOULD	Ability to switch agent context between profiles within a session
Permission scoping per profile	—	SHALL	Enforce read/write boundaries per Agent Profile
Script propagation mechanism	—	SHALL	Propagate script changes across multiple projects
Script rollback capability	SHOULD	SHALL	Revert any script change to previous version
Active script registry	—	SHALL	Inventory of active scripts with version identifiers
Script change audit trail	—	SHALL	Log of all script modifications with rationale
Tool inventory per profile	SHOULD	SHALL	Document which tools each profile can access
Hook execution engine	—	SHOULD	Automated actions triggered by events (post-session, pre-commit)
Script effectiveness metrics	—	SHOULD	Track FPSR and error rate per script version

---

6. Integration Requirements

How the components connect to form a functioning SENAR toolchain.

6.1 Core/Foundation Integrations

Integration	Description
Task Tracker <-> Session Management	Sessions reference Tasks; Tasks reference Sessions
Task Tracker <-> CI/CD	Pipeline results linked to Tasks and merge requests
Knowledge Base <-> AI Agent	Agent can search, read, and create KB entries during sessions
Session Management <-> AI Agent	Session start/end driven by or coordinated with agent lifecycle

6.2 Team+ Integrations

Integration	Description
Task Tracker <-> Knowledge Base	Tasks link to relevant KB entries; KB entries link to originating Tasks
CI/CD <-> Task Tracker	Pipeline failures auto-update Task status to `blocked`
Session Management <-> Metrics	Session metrics auto-populated from session data
Knowledge Base <-> CI/CD	Architectural constraint KB entries enforced in pipeline (aspirational)
Federation <-> Task Tracker	Cross-project dependency status synchronized
Federation <-> Knowledge Base	Cross-project knowledge accessible with proper scoping

---

7. Selection Guidance

When evaluating tooling options, prioritize:

1. **Automation over manual process** — if a ceremony step can be automated, it should be. Manual steps are failure points (see Guide 06, PF-3: Gate Bypass Normalization).

2. **Portability over features** — data must be exportable. Vendor lock-in (Guide 06, OF-3) is a recognized failure mode.
3. **AI-accessible over human-only** — every data store the Supervisor uses should also be queryable by the AI Agent. If the AI can't access it, it doesn't exist for SENAR purposes.
4. **Simple over comprehensive** — a task tracker with 5 well-implemented fields beats one with 50 fields nobody fills in. Start with Core/Foundation requirements and add Team+ requirements when needed.
5. **Composable over monolithic** — prefer tools that integrate via APIs over all-in-one platforms. SENAR components evolve at different rates; replacing one component should not require replacing all of them.

SENAR Reference: Code Standards Template

> *Organizations SHOULD maintain a Code Standards document loaded into every AI agent's context (Rule 15, L2). This template provides a starting point derived from adversarial audits of production AI-generated code.*

How to Use

1. Copy this template to your project's agent configuration (e.g., `.claude/references/code-standards.md`)
2. Adapt rules to your stack (Python, TypeScript, Go, etc.)
3. Add project-specific conventions
4. Ensure the document is loaded into every agent session automatically

Template Sections

A Code Standards document SHOULD cover:

Section	What to Define	Example Rules
Security	Input validation, access control, header trust, injection prevention	Never trust HTTP headers for security decisions without proxy validation
Architecture	File/function size limits, SRP, dependency injection	Max 400 lines/file, max 50 lines/function
Database	Query safety, migration patterns, bounded queries	Always use parameterized queries; always set LIMIT
API	Auth checks, schema validation, error handling	Every endpoint: verify auth AND resource access
Concurrency	Thread safety, resource lifecycle, data immutability	Shared mutable state requires locks
Domain-specific	LLM output validation, event processing, etc.	Validate LLM output before downstream use
Testing	Regression tests, mock boundaries, assertion quality	Every fix MUST include a regression test
Configuration	Startup validation, secure defaults	Fail fast if security config is missing

Key Principle

Every rule in a Code Standards document SHOULD be traceable to a real defect. Theoretical rules are ignored; evidence-based rules are followed.

See SENAR Guide Chapter 2 (AI Output Review Checklist) for complementary runtime checks.