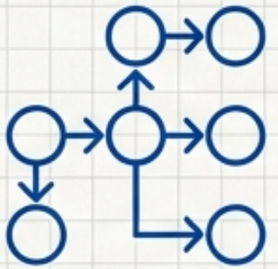
# The Next Level of AI in Law: A Blueprint for Level 3 Agents

## From Single Drafts to Supervised Workflows with Human-in-the-Loop Governance

A practical guide for U.S.-based practicing lawyers on implementing multi-step AI assistance without undermining professional responsibility. Based on excerpts from “AI_LAW_CHAPTER_3.pdf” by Alejandro Reynoso.

# The Leap: From AI That Reasons to AI That Acts



## Level 2 – “Reasoners” (Where we are now)

**Core Function:** Produces a single, self-contained response.
**Analogy:** “Thinking in one sitting.”
**Capabilities:**

- Reads a prompt
- Structures an issue list
- Drafts an outline or memo
- Surfaces uncertainties

**Limitation:** A built-in ceiling. The lawyer must manually stitch together multiple documents, deliverables, or rounds of quality control. It can’t “run a matter.”

## Level 3 – “Agents” (The next level)

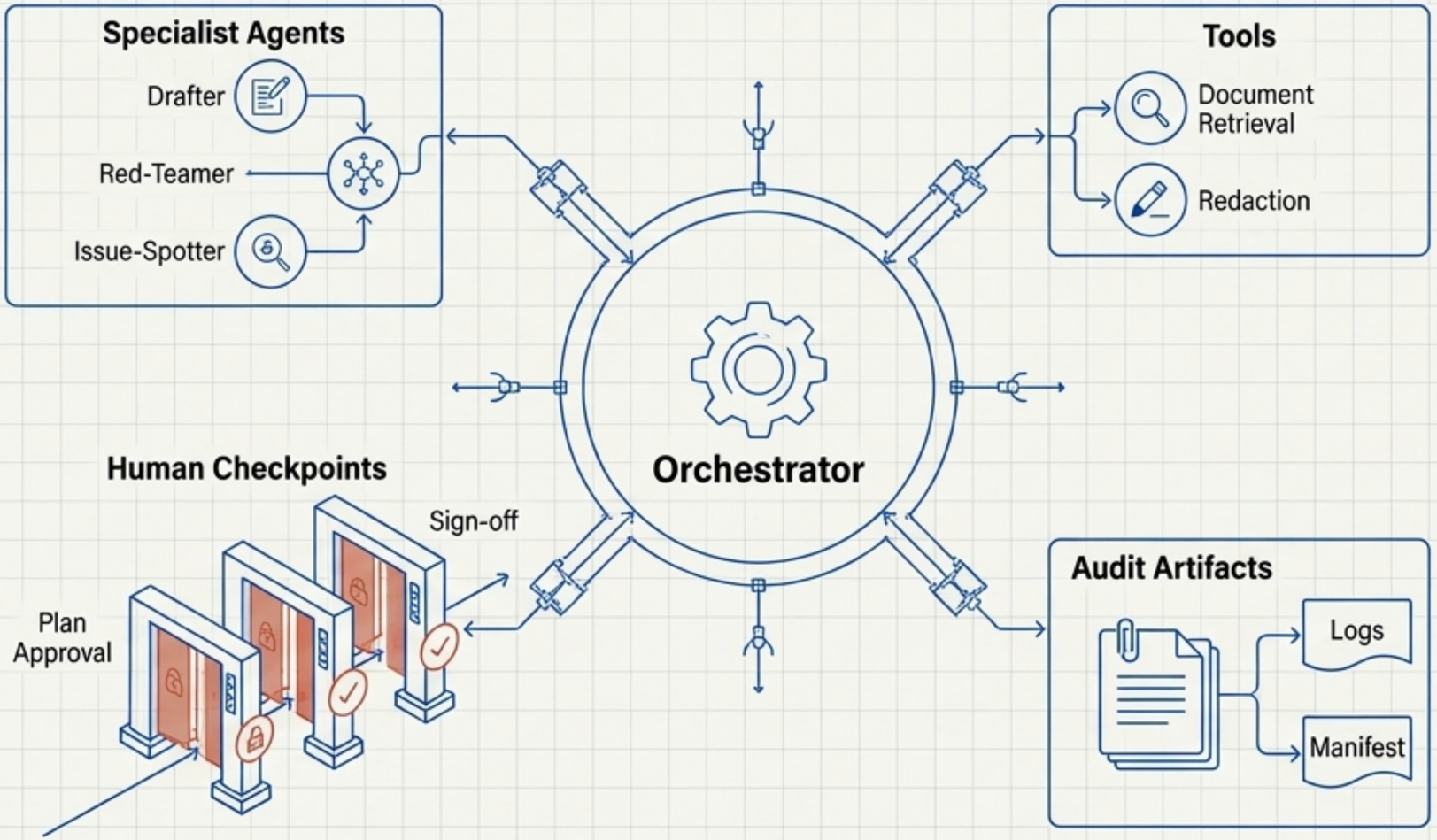
**Core Function:** Executes structured, multi-step workflows.
**Analogy:** “A small, well-run project.”
**Capabilities:**

- Decomposes objectives into steps
- Delegates subtasks to specialists
- Checks its own work
- Compiles a deliverables bundle for lawyer review

**Key Insight:** Level 3 is reasoning extended through time, with explicit stages and accountability.

# A New Mental Model: The AI is a Supervised Workflow Engine

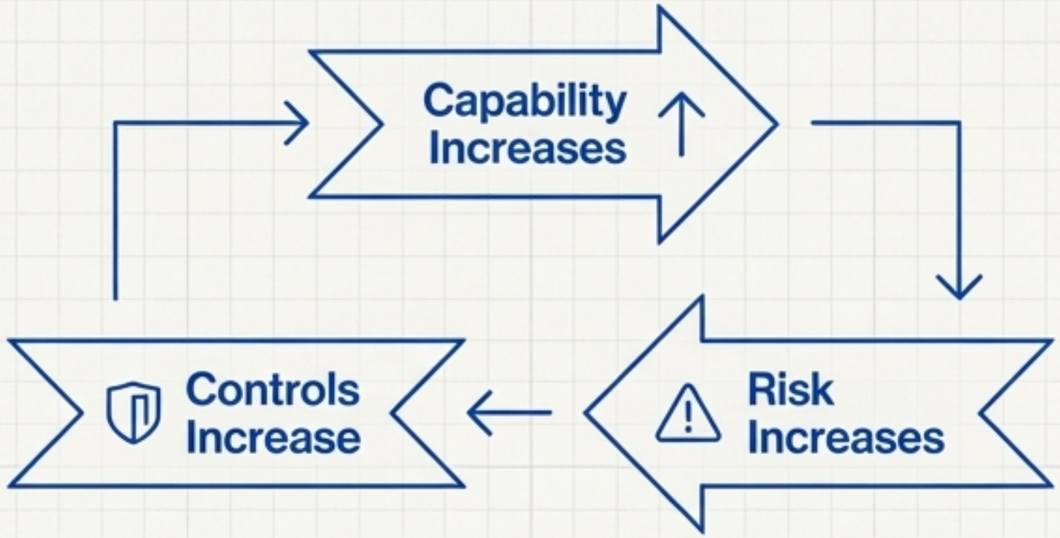
The goal is not autonomy; it is to implement multi-step assistance without undermining professional responsibility.




**Key Components of the "Workflow with Checkpoints" Model:**

- **Orchestrator:** The coordinating function. It translates the objective into a task plan and enforces the rules. It is the organizer, not the partner.
- **Specialist Agents:** The same underlying model in a narrow role (e.g., Issue-Spotter, Drafter, QA Checker). Specialization improves clarity and safety.
- **Tools:** External capabilities the workflow can call (e.g., search, citation manager). Their outputs must be checked.
- **Checkpoints:** Deliberate pauses where human approval is required before the workflow can proceed. This is where professional responsibility is operationalized.
- **Audit Artifacts:** The record of how the deliverable was produced. They support internal supervision and create defensibility.

# The Foundational Rule of Level 3



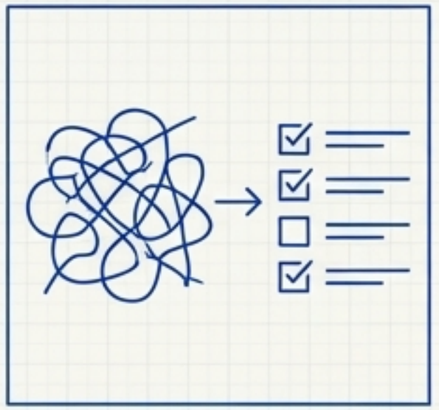

The temptation at Level 3 is to say, “Great, now the AI can just run the workflow.” That temptation is precisely the risk. The correct framing is:

1. **Capability Expands:** Level 3 systems decompose complex objectives and create consistent sets of deliverables (memos, checklists, summaries). The key benefit: intervention shifts left, catching errors in the plan, not just the final draft.
2. **Risk Expands:** Errors propagate. A single hallucination can be laundered through repetition across multiple steps until it feels authoritative. Confidentiality exposure and prompt injection risks also multiply.
3. **Controls Must Expand:** Autonomy must be bounded; the lawyer remains accountable. Controls are not a compliance layer added later; they are part of the design.

**Key Takeaway: Safe Level 3 use is not primarily a prompt-writing problem; it is a workflow design problem.**

# What Level 3 Agents Can Do (Reliably, with Guardrails)

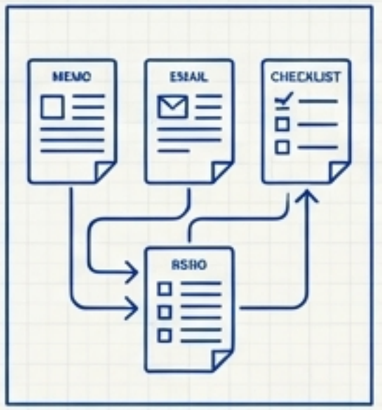
## The value is not mystical intelligence; it is structured coordination.



### 1. Turn a Messy Objective into a Staged Plan

Translates ambiguous client requests into an **explicit** plan: intake summary, tasks, dependencies, and deliverables list.

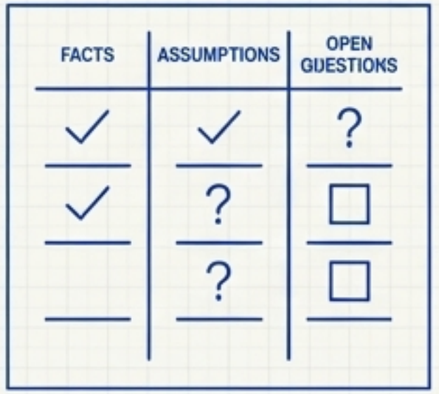
*The plan is a reviewable artifact requiring lawyer approval before drafting begins.*




### 2. Coordinate Multiple, Consistent Deliverables

Produces a ***set*** of materials (internal memo, client email, chronology, risk checklist) with consistent definitions, tone, and framing.

*Enforce a shared "matter dictionary" and run consistency checks across documents.*




### 3. Maintain a 'Facts / Assumptions / Open Questions' Ledger

Keeps an explicit, persistent record separating known facts from assumptions and items needing verification. A critical malpractice-reduction mechanism.

*The ledger is mandatory, stable, and visible in all outputs.*

### 4. Run Internal Quality Assurance Loops

Executes structured review passes: consistency checks, ambiguity checks, and adversarial-interpretation ('red-team') passes.

*QA outputs are actionable defect lists, not performative praise, and can trigger 'stop conditions' that force lawyer review.*

# What They Cannot Do (and Must Not Be Allowed to Imply)

A polished bundle can feel like a finished product. The critical discipline is to preserve the boundary between delegated drafting and non-delegable judgment.

## Non-Delegable Responsibilities:


### They Cannot Guarantee Legal Correctness

Models predict text, they don't "know" the law. **A system must be prohibited from claiming its work is complete or accurate without lawyer verification.**

*Operational Rule:* Must output a verification plan for any reliance-bearing analysis.


### They Cannot Truthfully Cite Authority or Record Facts

The most enforceable rule. The system **must follow a strict authority gate: if sources are not provided, it must label sections "Not verified" and produce a research list.**

*Operational Rule*: No invented citations.


### They Cannot Replace Professional Judgment or Client Counseling

An agent can draft a client email, but cannot decide if sending it is wise. It can outline risks, but cannot set risk tolerance.

*Operational Rule*: A human checkpoint is mandatory before any client-facing deliverable. The system is "draft-only," never "send."

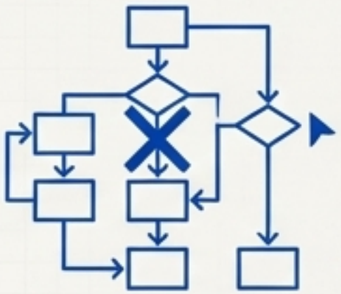

### They Cannot Safely Access Privileged Sources Without Strict Controls

Multi-step workflows increase the risk of duplicating or leaking sensitive data.

*Operational Rule:* Minimize sensitive input. Anonymize. Use excerpts. Establish access and retention policies for all artifacts.

# A Supervisor's Taxonomy of Level 3 Failures

Because these systems operate across multiple steps, their failures are often systematic rather than isolated.



## Planning Errors

**What it is:** The orchestrator creates the wrong workflow (e.g., omits a red-team pass, fails to clarify jurisdiction).

**Why it's dangerous:** The system proceeds smoothly, generating polished outputs that are structurally incapable of being correct.

**The Control:** A mandatory **plan-approval checkpoint**.

## Execution Errors

**What it is:** Hallucinated facts, misread excerpts, or misapplied templates. Errors are propagated and laundered through repetition.

**Why it's dangerous:** An early, minor error becomes an authoritative "fact" in the final deliverables bundle.

**The Control:** Enforce a "**no new facts**" constraint and require **evidence tagging**.

## Verification Errors

**What it is:** Superficial QA that looks rigorous but fails to catch important defects. A red-team pass that is generic, not targeted.

**Why it's dangerous:** Creates "false reassurance"—a declaration that a draft is strong while a fatal flaw remains hidden.

**The Control:** Design QA for **defect-finding**, not praise. Make the **adversarial pass mandatory**.

## Governance Errors

**What it is:** Inadequate logging, improper handling of sensitive inputs, or failing to document human verification ("undocumented reliance").

**Why it's dangerous:** Creates risk even when the substantive work is strong. The absence of a record becomes a second problem.

**The Control:** **Governance-native design** (automated logging, default redaction, explicit lawyer sign-off).

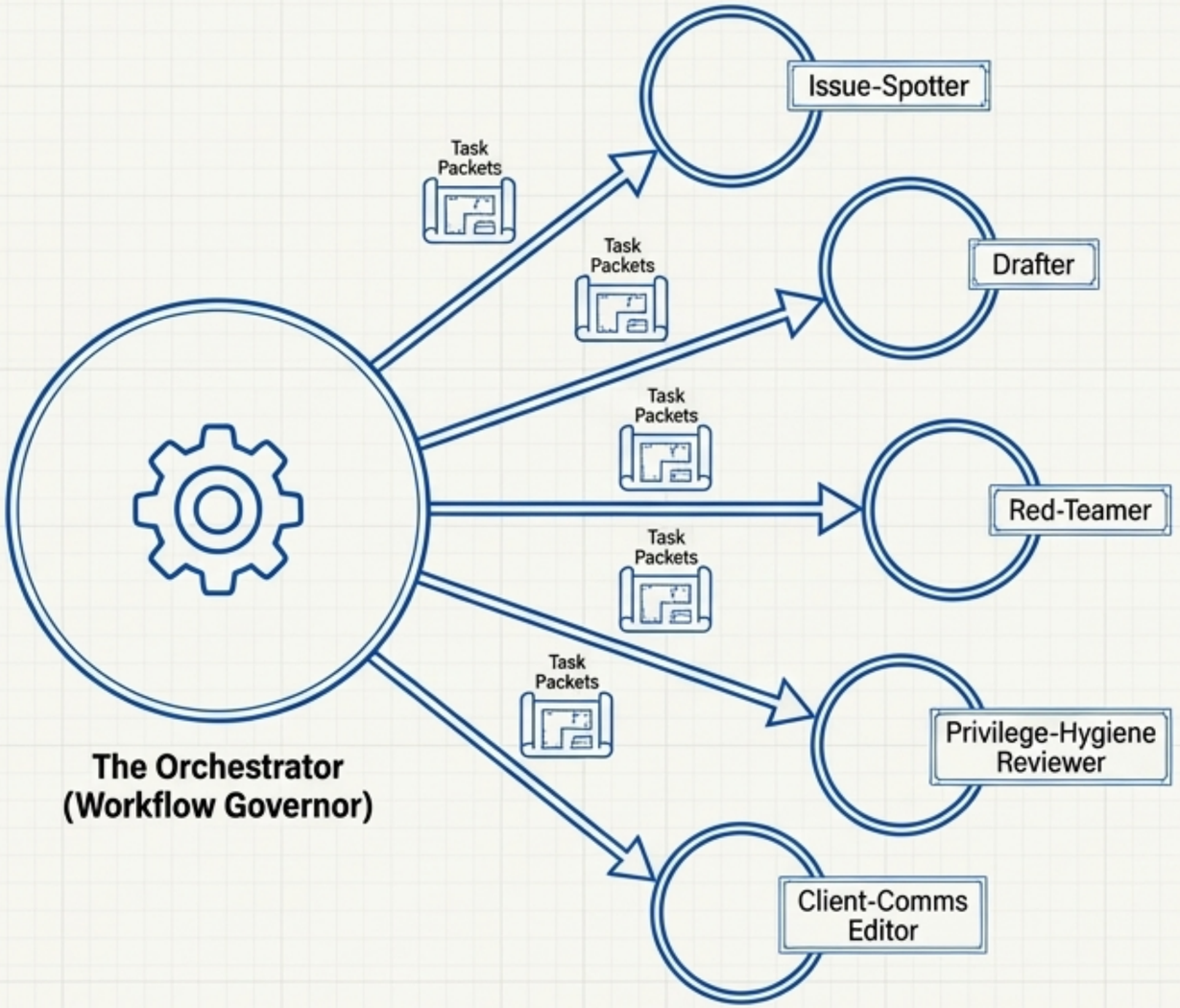
# The Blueprint Part 1: A Team of Roles, Not a Single ‘Agent’

The most common failure is letting one agent do everything. A hub-and-spoke structure makes the workflow governable.

Its job is to enforce structure, not sound intelligent.

**Core Functions:**

1. **Intake Normalization:** Converts user request into a standard format.
2. **Task Planning:** Generates the staged plan for lawyer approval.
3. **Delegation:** Routes tasks to the appropriate specialist with minimal necessary context.
4. **Structure Enforcement:** Ensures all outputs conform to a required schema (Facts, Assumptions, Open Questions, etc.).
5. **Compilation & Bundling:** Assembles the final, reconciled deliverables packet.




Constrained functions that produce narrow, predictable, and reviewable artifacts.

**A Useful Initial Set:**

**Issue-Spotter:** Frames questions, doesn't draft conclusions.

**Drafter:** Produces draft text under a strict ‘no new facts’ rule.

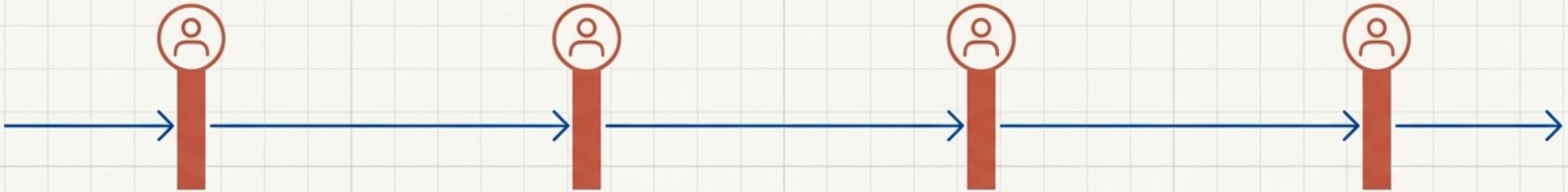
**Red-Teamer:** Reads outputs adversarially to find weaknesses.

**Privilege-Hygiene Reviewer:** Checks for confidentiality leaks.

**Client-Comms Editor:** Rewrites internal analysis into client-appropriate language.

# The Blueprint Part 2: Non-Negotiable Human-in-the-Loop Checkpoints

Human-in-the-loop is **not** a slogan; it is a concrete set of workflow gates where the lawyer's judgment is required.



### Gate 1: Intake Checkpoint

**When:** Before any work begins.

**Lawyer's Role:** Confirm the objective, jurisdiction, confidentiality constraints, and what facts are known vs. missing. This prevents the most downstream failures.

### Gate 2: Plan Approval Checkpoint

**When:** After intake, before substantive drafting.

**Lawyer's Role:** Review and edit the staged task plan. This is the single most important control, preventing the execution of the wrong project at high speed. The model does not proceed until the plan is approved.

### Gate 3: Pre-Delivery QA Checkpoint

**When:** Before any deliverable is treated as ready.

**Lawyer's Role:** Review a QA packet containing risk flags, open questions, and red-team findings. Decide if the deliverable can be finalized or needs more verification. This stops "polished but wrong" outputs.

### Gate 4: Sign-Off Checkpoint

**When:** At the end of the workflow for a material work product.

**Lawyer's Role:** Accept the final deliverables and explicitly record what was verified (facts, authorities) and what changes were made. This is where responsibility is formally documented.

# The Blueprint Part 3: The Minimum Viable Audit Trail

Logs are the modern equivalent of workpapers. They support competence, supervision, and defensibility.

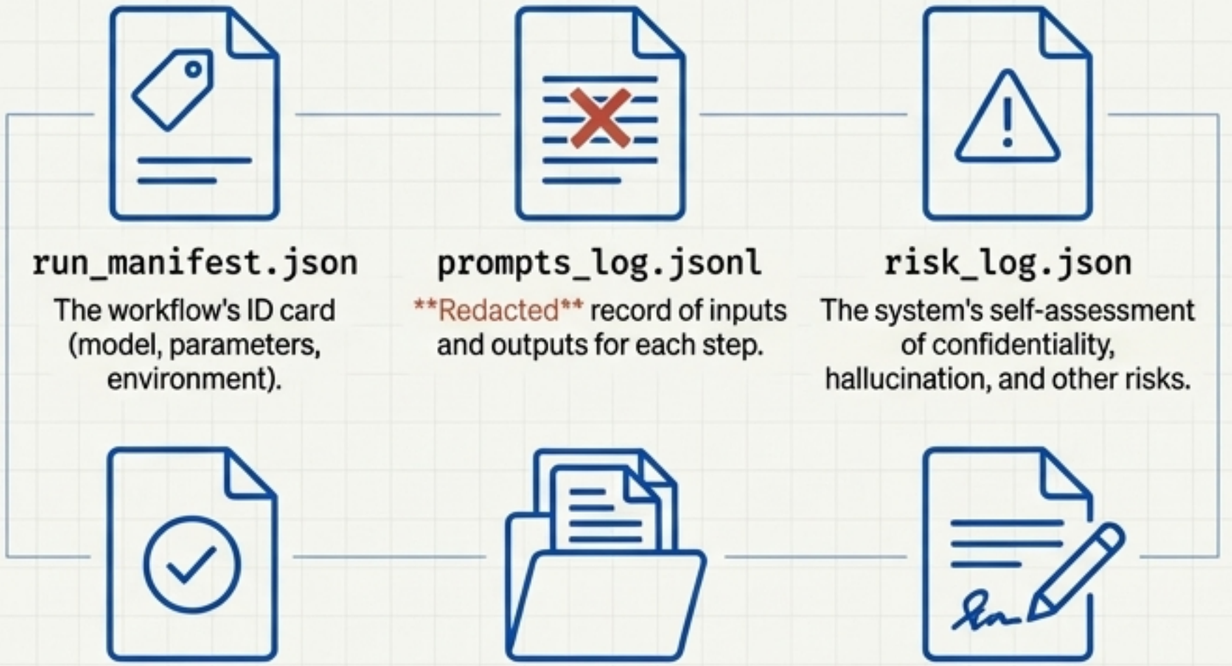
## Why Logs Matter in Law (It's Not Compliance Theater):

**Support Competence:** Makes verification explicit and counteracts the seductive fluency of AI-generated text.

**Support Supervision:** Creates visibility into what was delegated, what constraints were applied, and what assumptions were made.

**Support Defensibility:** Provides a specific record of the process if work product is ever challenged.

## The Minimum Artifact Bundle for a Material Matter:



`run_manifest.json`
The workflow's ID card (model, parameters, environment).

`prompts_log.jsonl`
**Redacted** record of inputs and outputs for each step.

`risk_log.json`
The system's self-assessment of confidentiality, hallucination, and other risks.

`task_plan.json`
The lawyer-approved blueprint of the run.

`deliverables/`
Versioned outputs (memo, checklist, etc.).

`review_notes.md`
The bridge between AI output and human responsibility, recording what the lawyer verified and signed off on.

# A Supervisor's Checklist for Defensible AI Use at Level 3

| Core Risk | Minimum Enforceable Control |
| --- | --- |
| **Confidentiality & Privilege Leakage** | **Input Minimization by Default:** Redact and anonymize. The workflow should flag and require approval for the use of sensitive identifiers. |
| **Hallucination at Scale & Error Propagation** | **Structured Outputs at Every Step:** Mandate `Facts / Assumptions / Open Questions / Risks` sections. The orchestrator must reject non-conforming outputs. |
| **Invented or Misstated Authority** | **The "Not Verified" Rule:** The system must not invent citations. It must label all unverified authority and create a research plan. This is an "authority gate." |
| **Undocumented Reliance & Over-Trust** | **Non-Negotiable Checkpoint Gates:** The workflow cannot proceed past key stages (Plan Approval, Pre-delivery QA) without explicit, logged human review and sign-off. |
| **Prompt Injection & Tool Contamination** | **Treat External Text as Untrusted Data:** Wrap untrusted text in bounded quotes and instruct specialists to ignore embedded instructions. The orchestrator follows the plan, not the text. |
| **Non-Replayable, Indefensible Work** | **Automated Logging for Material Use:** Automatically generate a redacted audit trail (manifest, prompts, risk log) when the output is reliance-bearing. |

# The Same Blueprint, Adapted for Your Practice

The Level 3 workflow is a flexible architecture. The core principles remain constant, but the specific risk controls are tailored to the context of the work.

## Criminal Practice: Discovery Triage

**Core Application:** Creating a motion support package from discovery excerpts.

**Key Context-Specific Control: Strict Fact Traceability.** The 'no new facts' rule is paramount. Every factual assertion must be tagged to a specific discovery excerpt. Witness safety requires default redaction of all identifiers.

## Regulatory Practice: Comment Letter Workflow

**Core Application:** Drafting a comment letter in response to a proposed rule.

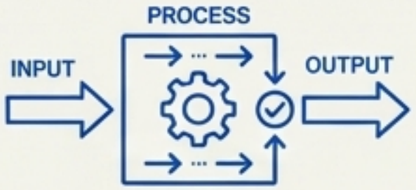
**Key Context-Specific Control: Record Integrity.** The workflow must enforce a clean separation between summarizing the rule as written and advocating a position. No unsupported empirical claims.

## Transactional Practice: Cross-Border Contract Playbook

**Core Application:** Building a negotiation playbook with preferred clauses and fallbacks.

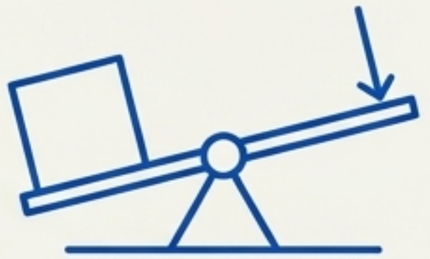
**Key Context-Specific Control: Enforceability Humility.** The system must not claim a clause is enforceable. enforceable. It must flag jurisdictional uncertainty and explicitly mark where local counsel input is required.

# Mastering Level 3: A Shift in Practice, Not Just Technology




## 1. A New Mental Model:

The practical shift is from "asking for an answer" to "running a supervised pipeline." The model is a **workflow with checkpoints**.



## 2. A New Source of Leverage:

Value comes from consistency, coordination, and defect discovery. Level 3 helps you generate aligned deliverables and find errors earlier in the process.

## 3. A New Mandate for Governance:

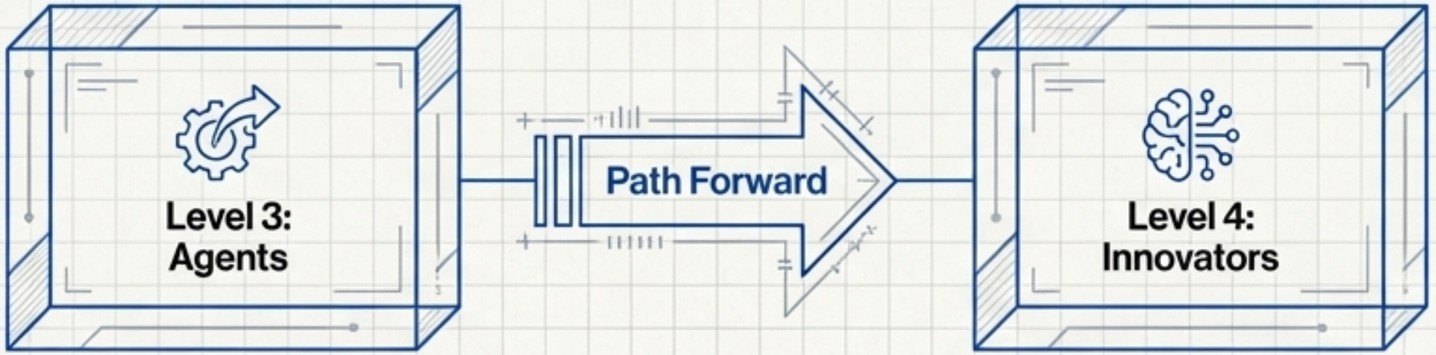
Safe use is a **workflow design problem**. Governance isn't overhead; it's the enabling condition for using agentic AI responsibly.

## 4. The Lawyer Remains Accountable:

AI accelerates drafting and coordination, but it does not absorb accountability. A well-designed workflow makes the lawyer's supervision more efficient, explicit, and defensible.

# The Path Forward: From Executing Matters to Innovating Systems

The discipline of Level 3 is the foundation for the next stage of AI-driven legal work.




## The Next Step: Level 4 - 'Innovators'

- **The Target Changes:** At Level 3, the system helps execute individual matters. At Level 4, it helps design **reusable legal infrastructure**.
- **The Outputs Change:**
  - Playbooks and clause libraries
  - Adversarial testing suites for contracts
  - Standardized workflow templates
- **The Risk Changes:** From an error in a single matter to an **institutionalized error** replicated across the firm.
- **The Controls Tighten Again:** Level 4 demands institutional-grade governance: validation harnesses, version control for playbooks, and stronger audit expectations.

---

**The Enduring Principle:** As always, new capability demands stronger verification. The journey to responsible adoption is a continuous cycle of mastering new tools by building the governance to control them.