HAIA-CAIPR

Cross AI Platform Review

A Governance Framework for Human Orchestration of Parallel Multi-AI Execution

Basil C. Puglisi, MPA

A Human-AI Collaboration

Independent AI Governance Practitioner, Puglisi Consulting, New York, United States

ORCID: 0009-0007-4747-152X | me@basilpuglisi.com | basilpuglisi.com | github.com/basilpuglisi/HAIA

Fourth Edition, version 6. September 2026.
10.5281/zenodo.22710389

Abstract

HAIA-CAIPR, Cross AI Platform Review, is a governance framework for human orchestration of parallel multi-AI execution. The same substantive task and collection condition are dispatched to every platform in a comparison cohort selected for diversity, a Navigator synthesizes and audits the returns under Tier 2 scrutiny, and a named human examines both the returns and the synthesis before deciding. The framework does three things: it exposes fabrication, stale retrieval, and misattributed citation through cross-platform non-corroboration; it recovers sources, framings, and corrections that a single platform would not have surfaced; and it publishes dated fitness findings about platform behavior so that the comparison becomes a public good rather than a private advantage.

This edition replaces the single architecture of earlier editions with nine invariants and five configuration variables, each of which names what it gives up. It states the Navigator's relationship to the dispatch as a profile of independent properties, specifies nine documented synthesizer failure modes with the governance requirements that reach them, and positions the framework against Article 14 of Regulation (EU) 2024/1689 as revised by the Digital Omnibus on AI in July 2026 without claiming that the framework establishes compliance.

The framework is examined against recent research on correlated model error, which finds that a nine-model panel carries roughly two independent votes' worth of information and that independently developed frontier models share error correlations near 0.77, and against research on open-ended homogeneity across models. The paper does not claim that structured collection removes that correlation. It states the structural response as a working position, sets out why the ablation that would isolate it may not be constructible

from inside the practice, and names five measurable outcomes that would falsify the rest of the framework. Evidence is field evidence from a single practitioner across a rotation of fifteen models on thirteen platforms and has not been independently replicated.

Keywords: AI governance, multi-model evaluation, human oversight, LLM-as-a-judge, correlated error, dissent preservation, checkpoint-based governance, meaningful human control, EU AI Act Article 14, source authority

Evidence class: Practitioner framework with documented operational case records. Not a controlled study.

Version Note

The published predecessors are the March 2026 article at basilpuglisi.com and the April 2026 revised edition in the HAIA repository (Puglisi, 2026c). The structural changes since then are the invariant core with five configuration variables in Section 1.4, the Navigator relationship stated as a profile of independent properties in Section 3.4, the third function in Section 1.2, the dated lineage and industry positioning in Section 1.1, the regulatory positioning in Section 2.6, and the platform roster in Appendix D.4. Everything else is correction, source verification, and precision. Claims in this document fall into three classes and the Evidence Status section below states which is which.

Executive Summary

A single AI platform produces confident, well-formed answers, and nothing in the output alone shows which ones are wrong or what was left out. HAIA-CAIPR, Cross AI Platform Review, is the framework for working across many platforms at once so that the comparison does what no single output can. The same task and the same collection condition go to every platform in a cohort chosen for diversity. A Navigator synthesizes the returns and is itself audited. A named human reads the returns and the synthesis, and decides.

CAIPR does three things. It exposes what is wrong: a claim one platform invents, the others fail to repeat, and a citation that formats correctly and does not support its claim is caught at source level. It recovers what is missing: one return carries the source none of the others found, or the framing none of them reached, or is simply correct against the rest, and a single platform can correct every other one in the dispatch. And it warns: every dispatch produces dated findings about how platforms behaved, and this edition publishes the platforms in the operator's rotation with their general fitness, which is the reason the framework is published at all.

The first function is widely understood. The second is the reason for structure. Where structured collection is used, the Sources, Conflicts, and Confidence fields make a minority of one assessable. Without them it cannot be told from noise, and the material the dispatch was run to find is the material most easily discarded.

The strongest objection to multi-model work is now measured. Nine frontier models from seven families carry roughly two independent votes' worth of information, and three

independently developed models share a mean error correlation of 0.77 (Kohli, 2026; Spiro, 2026). CAIPR does not claim that structure removes that correlation. It claims that field-level structure gives the human dimensions in which agreement on the answer can be separated from agreement on the sources, conflicts, and recommendation behind it. That is a working position, and Open Question 2 sets out why the test that would settle it may not be constructible.

This edition replaces the single architecture of earlier editions with nine invariants and five configuration variables, each of which names what it gives up. It states the Navigator's relationship to the dispatch as a profile of independent properties, adds the warning function as a stated purpose, and positions the framework against the EU AI Act as revised in July 2026.

CAIPR costs more than the alternatives, in money and in attention, and Part Six prices it. It belongs where an undetected error or a missed contribution cannot be recovered after the fact, and Part Five says where it does not.

For practitioners deciding whether the problem is real, start at Part Two. For practitioners ready to run it, Part Three and Appendix A are the working material.

Architectural Note

HAIA, Human Artificial Intelligence Assistant, is the ecosystem. Its frameworks are Factics, CBG, RECCLIN, CAIPR, HAIA Agents, GOPEL, and HEQ. This document specifies CAIPR.

CAIPR depends on three of the others. Factics supplies the evidentiary discipline at every checkpoint: a fact, a tactic, and a measurable outcome. CBG supplies the authority, since every checkpoint in a CAIPR session is a CBG checkpoint and CAIPR creates no governance authority of its own. RECCLIN supplies the response grammar, in two forms. RECCLIN Reasoning is the ten-field format a platform returns in, and it is the structured collection mode inside CAIPR. RECCLIN Dispatch is role-assigned multi-platform execution, and it is the lower-cost alternative CAIPR is measured against in Part Four.

Two frameworks sit above CAIPR and would automate its mechanics. HAIA Agents is the agent architecture that would carry dispatch, collection, and routing. GOPEL is the non-cognitive enforcement layer beneath it. Both are specified and neither is deployed, and every session in this record ran by hand.

HEQ is the measurement framework and is outside this document's scope.

Two documentation protocols can receive a session's output. HAIA-SCOPE holds source custody for each citation (Puglisi, 2026i). HAIA-CARCS holds the record of the session itself (Puglisi, 2026f). Where this paper cites a dated instance behind a general finding, that is where it sits.

The word protocol in this document names a process, never a framework. CAIPR's protocol is the eight core operations in Part Three.

Evidence Status

Three classes of claim appear in this document, and a reader should weight each differently.

Operational evidence. Documented sessions and case records from one practitioner, across a rotation that reached fifteen models on thirteen platforms (Appendix D.4). The founding record holds thirty-three findings (Puglisi, 2026k), extended through the August and September 2026 sessions in Appendix B. This is field evidence from a single operator, not yet replicated by anyone else, and it carries most of Parts Three and Four.

External evidence and authority. Peer-reviewed and preprint research on correlated model error and evaluation panels, together with regulation, standards, and copyright guidance, cited across Parts Two, Four, and Six and in Appendix E. None of it depends on the operator's own sessions. It supports several of the framework's risk assumptions and locates the framework against existing law and terminology, and it does not validate the framework.

Normative design. Rules adopted because the human governor judged them the safer choice, not because evidence compelled them. The Tier 0 confirmation form and the checkpoint density are examples. They are design decisions, open to argument on those terms.

Part One: What This Is

1.1 Definition

CAIPR is the governance framework for human orchestration of parallel multi-AI execution. It specifies how a human dispatches the same substantive task and collection condition to every platform in a comparison cohort, collects their outputs, and compares those outputs for convergence and divergence. It then specifies how anomalies are flagged through cross-platform non-corroboration for source-level checking, how the AI that synthesizes the results is governed, and how source-authority discrimination is maintained throughout.

It is the parallel evolution of RECCLIN Dispatch. Dispatch assigns roles across platforms in series. CAIPR sends the same work to many platforms at once, declares the Navigator's relationship to the dispatch as a configuration property, and delivers the whole comparison to the human.

Where it sits. The industry now sells the same shape. Perplexity's Model Council runs three frontier models in parallel and has a fourth model it calls a chair synthesize agreement and disagreement into one answer (Perplexity, 2026), and a category of comparison tools offers consensus scores and disagreement maps across five or more models. The research vocabulary describes a narrower object, a panel of LLM evaluators or an LLM-as-a-judge panel that scores the same item and aggregates the scores (Verga et al., 2024; Kohli, 2026), alongside multi-agent debate and mixture-of-agents methods that set models against or after one another.

CAIPR shares the dispatch and the comparison with all of these and differs at four points, each a governance choice. The returns arrive structured, so comparison runs field by field and not on answer text alone. The count is diagnostic and never a vote, and unanimous agreement is a signal to verify outside the pool. The synthesizer is a governed Tier 2 artifact with no approval authority, where a council's chair produces the final answer. And a named human reads the raw returns and decides, with accountability attached. The industry has built the dispatch. What it has not built is the checkpoint, and that is the layer this framework supplies (Puglisi, 2026m).

Where it comes from. CAIPR is the result of multi-AI work that was producing published output from 2023, beginning as a two-platform loop between ChatGPT and Perplexity. The method was first disclosed in February 2024, ran as a parallel five-platform method from September 2025, and received its name in March 2026. The practice preceded the name and preceded the products. The dated record, with what changed at each step, is Appendix F.

1.2 The Three Functions

Subtraction. What is wrong gets exposed. Fabrication, stale retrieval, misattributed citation, and confidence that outruns evidence are surfaced by comparison in ways that may remain invisible in an isolated output.

Addition. What is missing gets supplied. One platform surfaces the source none of the others found. One frames the problem in a way none of the others reached. One is correct against the rest, and the correctness of the rest is not established by their number.

Warning. What is learned gets published. Every dispatch produces evidence about how platforms behave, and that evidence is fitness information. A practitioner who reads the record learns why one platform has suited writing, why another has suited sourcing, and why a third has performed poorly for retrieval under documented conditions. This document publishes the platforms in the operator's rotation, listed in Appendix D.4, and the general fitness characterization. Dated instance detail, naming which platform produced which fabrication or contribution, belongs to the session record and is produced at the operator's discretion. Keeping the general characterization private would make the framework a competitive advantage. Publishing it makes it a public good, and it is the reason this document exists in the form it does.

The first two functions run through the same three catch layers, and nothing in the architecture distinguishes a catch from a contribution until verification runs. Before checking, a solitary claim that will prove correct and a solitary fabrication look identical on the page. They are separated by verification, not by counting.

The third function is what the first two produce as a byproduct, and it is the one that persists as reusable cross-session knowledge.

1.3 The Name

The acronym is human-originated. No AI platform produced it. Eleven platforms generated forty-seven candidates and none survived, and the one candidate that reached agreement across three platforms was eliminated by a collision that exactly one platform found.

Pronounced "kay-per," and the resonance with "caper" is intentional.

A caper is an elaborate, audacious plan to obtain something of high value through cleverness where force would fail. CAIPR is exactly that: an elaborate, structured plan to obtain intelligence, options, sources, and resources by dispatching across a broad pool of available AI platforms instead of trusting one. The heist is information and the crew is the platforms. The plan is parallel dispatch with a declared collection mode. The audacity is the assumption that no single AI is trustworthy enough to work alone. The human governor is the one who pulls off the caper by orchestrating the whole operation and walking away with intelligence no single platform could produce.

The tone is load-bearing, because a caper is not grim enforcement, which is what Checkpoint-Based Governance handles, and it is not mechanical labor, which is what GOPEL handles. A caper is clever, resourceful, and requires a plan, which matches the practitioner experience, because running CAIPR is not passive.

It is also the plainest statement of the additive function. Nobody plans a caper to avoid losing something.

1.4 The Invariant Core and the Configurations

CAIPR was a single architecture until practice showed it was not. Platform count, collection mode, and the Navigator's relationship to the dispatch all vary, and read depth varies with stakes. Describing those variations while keeping definitions inherited from the narrower form produced contradictions that a line review found and this edition removes.

The resolution is the one Checkpoint-Based Governance already uses one layer up. CBG does not say what to decide, only that a named human decides, at stated points, with accountability attached. CAIPR does not say how many platforms, which collection mode, or how deeply to read. It says where human judgment is required, what must be declared, and what each choice gives up.

Nine invariants. A session is CAIPR when all nine hold.

1. **Parallel isolated generation.** Platforms produce without seeing each other's output.
2. **Common task condition.** Every platform within a comparison cohort receives the same substantive task and the same collection condition.
3. **Minority preservation.** Solitary material is retained and marked, never aggregated away.
4. **Verification before evidentiary resolution.** Evidentiary status is established by checking, never by counting. A count may settle a reversible procedural disposition and never establishes truth.
5. **Declared source authority.** Every input classified at ingestion by tier.
6. **Declared independence assumptions.** Stated per dispatch on model lineage and retrieval index, with unknowns recorded as unknown.
7. **Auditable synthesis.** A Navigator synthesizes the return set, classified Tier 2 and traceable to the raw returns.
8. **Declared configuration.** Which configuration ran, and what it sacrificed.
9. **Human decision authority.** A named human decides, with accountability attached.

Five configuration variables. Each names what it gives up.

Variable	Options	What it costs
Dispatch count	Minimum three. Counted as Tier 1 returns in the comparison set. No ceiling other than the pool	Lower counts may reduce coverage and increase unanimous convergence requiring escalation. Higher counts cost reading time and increase the risk of redundant correlated voices. Neither rate has been measured in this record
Count reading	Majority resolution under Responsible AI, where odd parity is a rule. Diagnostic signal under AI Governance, where odd parity is guidance	Majority resolution settles a reversible procedural disposition and sets the asymmetry rule aside. The session record states that it was not applied. At governance stakes the count is diagnostic only
Collection mode	Plain, full RECCLIN, mixed	Plain sacrifices comparability of sources and dissent. Full costs effort. Mixed creates two cohorts, each internally identical, and buys a measurement of what structure does at the cost of cross-cohort comparability
Navigator relationship	A profile of independent properties, not one category. Dispatch participation: held out or participant. Authorship: author or non-author. Context: current-context or siloed. Secondary audit: none or partial. Section 3.4 sets them out	Each value carries its own exposure. A participant grades its own return, an author defends its own work, a current-context Navigator carries prior threads into the synthesis, and no secondary audit means the synthesis is checked by the arbiter alone
Read depth	Proportional to stakes	Shallower reading reduces the human's ability to detect what a synthesis omitted, and the human remains accountable for what went unexamined

Why invariant 7 has no conditional. The review in Cross AI Platform Review is the Navigator step. Dispatch without synthesis is parallel querying, and a human who synthesizes alone produces no Tier 2 artifact to audit against the raw returns, which removes the inclusion manifest, Minority Report Extraction, the source-overlap map, and the auditor evaluation path in a single stroke. Half of Part Three exists because a Navigator exists. The configurable property is the Navigator's relationship to the dispatch, not whether there is one.

Why invariant 8 exists. With five variables in play, naming which configuration ran is the only thing keeping a session auditable. A reader who knows the configuration knows what the session could and could not have caught. Without it, configurable means anything counts.

On invariant 2 and mixed collection. Platforms receiving the ten-field schema and platforms receiving only the substantive prompt have not received identical prompts. The invariant binds within a cohort and not across the session, so mixed produces two cohorts, and the cross-condition difference is a measurement of what structure does that says nothing about the platforms.

1.5 Position in the Stack

```
Factics          ->  evidentiary discipline (2012)
RECCLIN          ->  structured AI-response grammar, single platform
RECCLIN Dispatch ->  role-assigned multi-platform execution
HAIA-CAIPR       ->  parallel dispatch (this document)
CBG              ->  constitutional checkpoint authority
HAIA-SCOPE       ->  source custody for external citations
HAIA-CARCS       ->  governed record of completed sessions
HAIA Agents      ->  integrated automated infrastructure
GOPEL            ->  non-cognitive enforcement infrastructure
```

Part Two: Why We Use It

2.1 What Single-Platform Work Cannot Detect

Single-platform AI workflows produce confident, well-structured, wrong answers, and an isolated output lacks any independent comparative mechanism for surfacing disagreement or omission. The problem is widely recognized and single-platform workflows remain common, because the alternative appears to be more work for a benefit that cannot be seen in advance.

The benefit can be seen in retrospect, and the record is specific.

A platform fabricated entire document sections that passed every internal quality check it applied to itself. It also fabricated quotes and then reviewed those fabricated quotes approvingly. Cross-platform dispatch surfaced both immediately, because the other platforms did not repeat the inventions.

In a fifteen-item citation verification, five attributions were fabricated or misattributed. The sentences read correctly, the claims were plausible, and the formatting was correct. The cited works did not contain the attributed claims. More platforms reading the same fabricated citation produce stronger false convergence, not detection, so this class is only caught by going to the source.

In August 2026, one platform in a ten-platform dispatch returned a European Commission Delegated Act mandating multi-provider redundancy, at 92 percent declared confidence. No such Act exists. It was the single item in the entire dispatch most likely to look like

validation of the operator's own filed policy work, and it was invented. The other nine platforms did not repeat it, which raised the anomaly. Source-level checking then established the fabrication. Without the first step nothing prompts the second on that item, and without the second the anomaly is a flag and not a finding.

2.2 What the Subtractive Record Shows

Across documented working concept operations, platforms produced materially different outputs on identical prompts in 15 to 25 percent of cases, and those disagreements triggered human verification that prevented error propagation (Puglisi, 2026a).

That rate establishes that actionable disagreement is not rare in this record. Whether it earns the cost remains task dependent, because frequency alone says nothing about the severity of what was caught or the value of what was prevented. The remaining three quarters to four fifths did not produce material disagreement by that measure, and the evidentiary value of that agreement remains dependent on the independence behind it, which Part Six treats at length.

2.3 What the Additive Record Shows

The additive cases are less obvious and matter as much.

In the session that named this framework, forty-seven candidates produced no majority, and the one name that reached three-platform agreement was eliminated by a collision that one platform found and ten missed. The same platform had produced the only collision finding in the previous round. A workflow excluding it would have shipped two collisions.

In the same session, one platform raised pronunciation as a specification detail. The Navigator disputed it as premature and recommended it resolve through usage. The human governor overrode the Navigator, connected "kay-per" to "caper," and turned a minor operational note into the framework's identity. One platform contributed, the synthesizer was wrong, and the human recovered the value.

In a ten-platform dispatch, one platform supplied eleven of sixteen unique contributions, so a configuration excluding it would have missed eleven of sixteen. Unique contribution was not predictable from this record and should not be assumed evenly distributed. These are counts from one session and are not a sampling design, so they cannot be read as a stable platform ranking.

A framework built only to catch errors would have discarded the pronunciation observation as noise. That is the argument for the additive function in one paragraph.

2.4 What the External Research Says

Independent research now measures the risk this framework was built around, and it is more useful than a general warning.

Kohli, G. (2026) tested a panel of nine frontier models drawn from seven families on three natural language inference datasets, with one hundred human annotations per item. The

panel carried roughly two effective independent votes' worth of information, measured as effective sample size, with about three quarters of nominal independence lost because the models make the same mistakes on the same items. The panel's accuracy fell 8 to 22 percentage points short of independent voting, and the best single judge matched or outperformed the full panel across all conditions. Established aggregation methods closed at most 11 percent of the gap, even with access to correct answers.

Kim, Garg, Peng, and Garg (2025) evaluated over 350 models and found substantial correlation in model errors, with models agreeing 60 percent of the time when both err on one leaderboard dataset. Shared architecture and shared provider drive correlation, and larger and more accurate models show highly correlated errors even across distinct architectures and providers.

Spiro (2026) tested three frontier models from three organizations on 568 resolved binary questions and found a mean pairwise error correlation of 0.77, or 0.78 excluding likely-leaked questions. Spiro's conclusion is that the three function as a single oracle with noise rather than as three independent sources. The forecasts were generated in 2025 on questions resolved between 2019 and 2025, and the leakage sensitivity analysis is what supports reading the correlation as real and not as contamination.

Shu (2026) tested whether adding an external verification signal repairs the deficit and found no distinguishable change in a panel's aggregate effective-vote count. The entire accuracy gain concentrated on decisions with a one-vote margin, where it ran between 10.4 and 23.3 percentage points, and it was exactly zero elsewhere.

Jiang et al. (2025) moved the question out of scalar judgment entirely. Infinity-Chat is 26,000 real-world open-ended queries mined from user chat logs, admitting many good answers with no single ground truth. It was evaluated across more than seventy open and closed models, with twenty-five reported in the main panel, and it carries 31,250 human annotations at twenty-five independent annotations per example. It received a NeurIPS 2025 best paper award on the Datasets and Benchmarks track.

They document what they call an Artificial Hivemind in two parts: intra-model repetition, where one model keeps generating similar responses, and, more critically, inter-model homogeneity, where different models produce strikingly similar outputs. Average pairwise cosine similarity across models runs between 0.71 and 0.82 under standard chat decoding. The lead illustration clusters twenty-five models generating fifty responses each to a single query, a request for a metaphor about time, and despite family and size diversity the responses form two primary clusters, one built on a river and one on a weaver. The quantitative analysis behind the figure covers one hundred queries.

A second result in the same paper bears on the Navigator requirements in Section 3.4. Language models, reward models, and model judges are poorly calibrated to idiosyncratic human preference when the alternatives are equally good. That is adjacent to one class of dissent-source verification, the class where the evidence does not discriminate clearly between positions, and not a measurement of the whole task.

What this literature establishes, and what it does not. Kohli, Kim, Spiro, and Shu examine correlation or aggregation of model judgments expressed as scalar or categorical outputs: a judge verdict, a probability estimate, a pass or fail ballot. They establish that the tested aggregation methods do not eliminate the independence deficit, which is the strongest available argument for removing decision authority from a panel.

Jiang is the harder finding for this framework, and an earlier version of this paper claimed the literature had not reached open-ended generation. It has.

The transfer is unclear and should not be resolved in either direction. Jiang measures unconstrained generation, the illustrative prompt being a request to write a metaphor about time. CAIPR's open-prompt configuration is not that. It is a structured critique task with named analytical moves and a required output format, and its structured-prompt configuration is further still, fixing evidence parameters so that platforms pull from different bases to satisfy the same constraints.

So Jiang does not straightforwardly undercut CAIPR, and it does not validate it either. A reviewer arguing that the homogeneity finding maps cleanly onto open-prompt mode and therefore confirms the convergence-without-dissent rule is converting a threat into a confirmation by way of a mapping this paper cannot support. What can be said is that Jiang establishes homogeneity in unconstrained generation, that CAIPR operates under task constraint in both configurations, and that nobody has measured whether homogeneity persists under constraint.

The strongest objection to the structural answer, stated in full. If shared pretraining and alignment are what produce the collapse, schema will not prevent it. The river and weaver clustering is conceptual, not lexical, and ten fields can each fill with the same two ideas in different wrappers. Different field values are not evidence of different thinking. The question is not whether the returns differ in form but whether they differ in substance inside each field, and that has not been measured.

The framework's answer remains structural and is stated as a working position rather than an established one: the ten fields may give the human dimensions in which agreement on the answer can be separated from agreement on the sources, conflicts, and recommendation behind it. Whether they do is untested, and Open Question 2 sets out why the isolating comparison may not be constructible. Jiang is treated here as a threat model rather than as a completed falsification.

Kohli's reconciliation of the earlier panel literature is worth carrying. Verga et al. (2024) found that panels of smaller diverse models outperform a single large judge, which appears to contradict Kohli. Kohli notes that the panel research compares panels to the average individual judge, where a panel naturally wins by diversifying away individual quirks, while Kohli's own comparison is to the best individual. The conclusion is that when judges are highly correlated, majority voting dilutes the best judge's signal with redundant weaker votes.

That finding points somewhere this paper has to go through. If the best single judge matches or outperforms a correlated panel, then using the best available model once,

carefully, is a live alternative to multi-platform review and not a strawman. The distinction is that Kohli ranks by model accuracy on a benchmarked task, which can be estimated in advance at the distribution level and never known for any specific item. CAIPR's asymmetry rule ranks by evidence on a specific claim, which is unknowable until someone checks and varies item by item. A weaker platform holding a verified source outranks a stronger one holding none.

Those are different claims with the same shape, and they are compatible without being identical. Where the task is a scalar judgment on a benchmarked distribution, Kohli's conclusion should be taken seriously and the best single judge may be the right tool. Where the work turns on sources, framing, and what was missed, a single run cannot recover material it never surfaced, while CAIPR deliberately creates additional independent opportunities to surface it. Whether that advantage survives comparison against repeated fresh-session use, retrieval-diversified use, or agentic search on one platform is Open Question 7, and it has not been run.

2.5 The Warning Function

The two functions above serve the operator. The third serves everyone else.

Every dispatch generates evidence about how specific platforms behave within declared comparison conditions. Which ones preserve depth and which compress, which retrieve and which answer from weights, which declare high confidence on low search effort, and which produce unique contributions and which produce restatement.

That evidence is a form of operator-specific controlled comparison, distinct from vendor claims and complementary to formal benchmarks and third-party audits. A practitioner reading the record learns why one platform is a good fit for writing, why others are better for sourcing, and why a particular platform should not be handed a retrieval task. Part Four names the assignments this operator currently uses. Where a CARCS record was produced, it carries the dated instances behind them.

Publishing this is a choice with a cost. Platform behavior changes, findings date quickly, and a named failure follows a vendor longer than the instance it describes. The framework handles that by keeping general fitness characterization in this document and routing dated instance evidence to the CARCS record, where it carries its date and its model state.

2.6 Where This Sits Against Regulation

The framework does not claim regulatory novelty and its public materials have declared alignment with the EU AI Act, ISO/IEC 42001, and the NIST AI Risk Management Framework since first publication (Puglisi, 2025a).

Article 14 of Regulation (EU) 2024/1689 requires that high-risk AI systems be designed and developed so they can be effectively overseen by natural persons. Article 14(4) sets out what that oversight must enable: monitoring operation, detecting automation bias, correctly interpreting outputs, and intervening to reject, override, or interrupt. Regulation (EU) 2026/1744, the Digital Omnibus on AI, was published in the Official Journal on 24 July

2026 and entered into force on 27 July 2026. It postponed application of the relevant Chapter III requirements to 2 December 2027 for standalone Annex III systems and 2 August 2028 for Annex I embedded systems. The high-risk substantive obligations are unchanged and apply later.

Three things did not move, and one of the three is more complicated than a single clause allows. The general-purpose AI obligations from August 2025 are unchanged. The eight existing Article 5 prohibitions kept their February 2025 date, and the Omnibus added two more, on non-consensual intimate material and child sexual abuse material, which apply from 2 December 2026. Article 50 transparency applies in full from 2 August 2026, with one exception. A new Article 111(4) gives providers of synthetic-content systems already placed on the market before that date until 2 December 2026 to meet the machine-readable marking duty in Article 50(2). No other paragraph of Article 50 is on that runway and new systems receive no transition.

Two boundaries matter and the paper states both.

Article 14 is a design obligation on a high-risk system, and CAIPR is a practitioner framework for comparing outputs across vendor models. It can evidence human decision authority inside a multi-model workflow. **It does not by itself establish Article 14 compliance and should not be described as doing so.**

And Article 14 states what oversight must be possible. It does not specify the architecture that makes oversight evidenced and attributable. Checkpoint-Based Governance answers that question one layer up, and Factics answers what the human must produce at the checkpoint: a fact, the tactic that follows, and the measure that will show whether the tactic worked. A checkpoint without that content is a signature. With it, the checkpoint is auditable.

The scholarly pairing is meaningful human control, which asks whether the human holds sufficient understanding, authority, intervention capacity, and responsibility. That is closer to what Tier 0 claims than the regulatory text is.

Part Three: How We Use It

3.1 The Eight Core Operations

1. Parallel dispatch. The same substantive task and collection condition to every platform in a comparison cohort, simultaneously, minimum three, with the count odd where parity is a rule and typically odd elsewhere, and with the Navigator's relationship to the dispatch declared.

The count is the number of Tier 1 returns in the comparison set. A held-out Navigator contributes no Tier 1 return and is not counted. A participant Navigator contributes one and is counted once, and its later Tier 2 synthesis does not add a second participant, with the session record noting that dispatch isolation was sacrificed. Two returns reaching the

operator through a single platform, from different models on that platform, are two Tier 1 returns and share an access route, which the independence profile records.

Stating this matters because the readings give different answers about whether a session conformed. The two most recent sessions in Appendix B illustrate both cases: August ran ten returns with the Navigator inside the set, which is an even count with dispatch isolation sacrificed and both deviations logged. September ran thirteen returns from twelve platforms with the Navigator held out, which is odd and conforming.

Three is the minimum. There is no ceiling other than the available pool, and the practical limit is the number of separately operated platforms the operator can reach. Common configurations run three, five, seven, nine, eleven, and thirteen. No platform sees another's output before producing its own. Count is calibrated to stakes, never trimmed for efficiency, because unique contribution is not evenly distributed and cannot be predicted in advance.

The odd count is a rule under Responsible AI and guidance under AI Governance. That is the whole of it, and the two are not in tension because they answer different questions.

Under Model 1, Agent Responsible AI, the majority resolves and an even split would deadlock. Parity is therefore a rule, doing the ordinary work parity does in any majority system. There is nothing wrong with that where the decision is low risk and reversible.

Under Models 2 and 3, AI Governance, the count decides nothing and parity is guidance. It is read the way a confidence score is read, as a signal about how much investigation the disagreement warrants. An even dispatch is not a violation here, because there is no tie to break when nothing is being voted on. Odd remains the working preference because an unambiguous ratio is easier to read against, and a governor who runs eight or fourteen has not broken a rule. Eight to one and five to four ask different questions. Eight to one says the lone return is either the most valuable in the set or the most wrong, and both readings demand a source check. Five to four says the question itself is contested and the majority status carries no evidentiary weight.

At irreversible stakes the ratio is a starting point and not an answer, and it forces the work that makes the framework worth running. Where nearly all agree, why does the minority not, and where all agree, what explains the unanimity? Those questions are the operative use of the count.

In no configuration does a count establish evidentiary truth. Majority resolution settles a procedural disposition that remains reversible, which is a different act, and invariant 4 holds across both readings.

Three is the working minimum and a floor, not a sufficiency condition. Convergence among three is a flag, and it never discharges the human's obligation to review the result.

2. Collection. Platform identity is documented by the human at dispatch, before any output is received, because platform self-identification is not trusted. Timestamps recorded. Model version strings and knowledge-cutoff dates recorded where exposed or reliably known,

since a finding dated to a model state is useless if the state cannot be identified. Raw outputs preserved in full.

Collection mode is a choice and it changes the value of the session dramatically.

Plain answers. Lowest effort, no imposed structure, and the comparison runs on content alone.

Full RECCLIN Reasoning. Ten fields on every return. Sources, Conflicts, and Recommendation become comparable across the set, which is what makes minority evidence assessable and not merely present. Appendix A.3 is the call.

Mixed. Some returns structured, some not. This is not a compromise. Structured and unstructured requests surface different things, and the difference between them is itself informative.

The human chooses on resources and effort, and the stakes decide how much structure the session needs. The operator makes that call before dispatch and records it, and under Models 2 and 3 the choice is itself a checkpoint decision.

3. Cross-platform comparison. The human reads every return before anything goes to a synthesizer. In manual operation this is a property of the method, not a rule the human might skip, because the human is the transport layer and is pasting each return into the Navigator session by hand. Exposure is structural. Depth of reading is not, and that is where accountability sits.

Comparison runs on three axes.

Across platforms, for error. Claims present in one return and no other are marked.

Across platforms, for absence. A source, argument, counterexample, or framing present in one return and missing from the rest is an additive-yield candidate, and it looks identical on the page to an unsupported single-platform claim. The two are separated by verification, not by count. Every solitary item is marked for checking, never deleted, and discarding solitary items because no other platform said it defeats the purpose of the dispatch.

Against prior versions, for loss. Comparison against the previous version catches content that was removed, which a check for wrong content cannot. This is regression checking in the ordinary software-quality sense, applied to a governed document instead of code. No reviewer given only the current version can perform it, which is the operational reason the framework's version-sequencing rule forbids overwriting: never overwriting is what makes loss detectable.

Where platforms produce conflicting figures, the result is presented as a range with a note explaining the reconciliation attempt, unless one figure is disproven, in which case a range preserves an error under the appearance of humility.

4. Fabrication detection. Claims present in only one platform are flagged for source-level verification. Citations not present in human-provided documents require independent verification before inclusion, and that verification checks the author list, the year, and

whether the cited work contains the attributed claim. Confirming that a record exists at an identifier is existence verification and it does not catch misattribution.

Two independence axes are checked separately. Platforms sharing a foundation model should not be treated as fully independent, since prompts, sampling, system layers, and fine-tunes can still produce partially distinct information. Search-capable platforms drawing on the same retrieval index can produce correlated retrieval regardless of how different their models are.

5. Convergence analysis. Factual, analytical, and recommendation convergence carry different weight, in that order. Each is qualified by source basis, meaning whether the agreeing platforms retrieved externally or read only the supplied input.

Convergence is assessed per field, not per answer. Agreement on the Output field with divergence on Sources, Conflicts, Recommendation, and Expiry is not full-return convergence, and it is not the scalar agreement the panel literature measures.

Source divergence is a signal and not a verdict. Three platforms citing three different sources have not thereby established independent retrieval, because three URLs can trace to one press release, one database entry, or one copied error moving across outlets. Three platforms citing the same canonical primary source have not shown clustering either, since independently identifying the correct primary source is what good practice looks like. What the divergence establishes is where a provenance check would be informative. Whether that check runs is set by the deploying agency or individual against their own expectations, stakes, and regulatory obligations, on the same discretionary basis that governs how often the arbiter verifies a synthesis against the raw returns.

Convergence with no dissent anywhere on a material claim is a flag. It confirms nothing. The response is escalation to additional platforms, a structured follow-up on the same subject, or verification outside the AI ecosystem entirely. Trivial or strongly primary-source-grounded claims do not need this treatment, and applying it universally makes the framework expensive without making it better.

The asymmetry rule. A single return carrying verified external evidence takes investigation priority ahead of unsupported convergence. At irreversible stakes it governs the evidentiary assessment regardless of the ratio, unless rebutted by stronger evidence. Tier 0 still governs the decision. Before checking, a solitary claim that will prove correct and a solitary fabrication look identical, which is why the rule sets the order of work and not the outcome. At low stakes and under Model 1, majority resolution remains appropriate for a reversible procedural disposition and this rule does not displace it.

Declared confidence is not evidentiary weight. It is a failure-monitoring instrument. Confidence ran counter to accuracy in the August 2026 dispatch, and the field is retained precisely because that inversion is only detectable when the number is collected. It produces diagnostic data about the platform, never evidentiary data about the claim.

Dissent resolution has three dispositions, not two. The arbiter may rule for one position, preserve the conflict unresolved, or introduce evidence neither position had, in

which case the conflict dissolves and nothing remains to resolve. The third is the least documented and the most information-adding, and it separates Tier 0 as an authority level from Tier 0 as an evidence source.

6. Navigator oversight. Section 3.4.

7. Source-authority discrimination. Section 3.6.

8. Platform resilience management. Sessions continue when platforms fail, hit limits, refuse a prompt on policy grounds, or return something structurally unusable. Substitution restores the minimum, and under Model 1 it also restores parity, which is a rule only in that operating model.

Substitution is triggered by a dispatch falling below the minimum of three, or by the loss of parity under Responsible AI where parity is a rule. It is also triggered by the loss of the only platform holding a given model lineage or retrieval backend, which matters more than the count. Replacing it with a second platform from the same lineage restores the number and not the diversity.

Resilience covers unavailability, not wrongness, because wrongness is the expected behavior the framework exists to process. A platform with a recorded finding against it stays in the pool and is read more closely. Temporary or session-specific exclusion remains available for known contamination, policy incompatibility, or missing search capability, and is distinct from a standing exclusion.

3.2 Prompt-Type Governance

Prompt type is a governance variable, not a style choice.

Structured prompts create conditions in which source-level diversity can be observed. When the prompt fixes parameters, platforms may pull from different evidence bases to satisfy the same constraints. The session reveals the scope of available evidence, and platforms may still retrieve the same material.

Open prompts create conditions in which perspective-level diversity can emerge. When the prompt leaves interpretation open, platforms may produce different analytical frames and readings of the same subject. The session reveals the scope of available perspectives, and open prompting does not guarantee they will differ.

The two are not substitutes. Running one when the other is needed produces convergence that looks well validated and has not stressed the dimension the work required.

If the session needs	Use
Evidence breadth: sources, citations, data points, verifiable anchors	Structured prompts with fixed parameters
Framing breadth: readings, value weightings, causal angles, interpretive stances	Open prompts without fixed parameters
Both dimensions stressed	Structured round first to map evidence, open round second to surface framings, then human synthesis

Prompt type is recorded at dispatch. When a session produces unusually tight convergence, prompt type is among the first variables to reconsider, alongside source overlap and platform independence.

3.3 Expected Failure and the Three Catch Layers

The term catch is used here for the layer at which a failure becomes visible to the human, not for autonomous detection by the layer itself. The distinction is made below and it matters.

CAIPR assumes any platform may fail and designs the workflow on the expectation that failures will occur. Fabrication, stale retrieval, inflated confidence, and contaminated sourcing are expected inputs, not incidents. A dispatch in which nothing appears wrong anywhere raises a correlation question. It does not imply that undiscovered error must exist.

When a failure or anomaly appears, the governance question is which layer exposed it, and whether that layer was the last one available.

Layer	Mechanism	What only this layer reaches
Peer platforms	Non-corroboration across returns	Surfaces anomalies that are individually plausible and correctly formatted, and surfaces solitary contributions the rest of the pool lacked
Navigator, under its declared configuration	Synthesis across the full return set	Sees source authority, attribution, and confidence patterns that no single return reveals
Human arbiter	Tier 0 judgment against knowledge no platform holds	Classifies what an anomaly means, recognizes provenance the platforms cannot see, and rules that a minority position governs the evidentiary assessment

Peer platforms expose. They do not catch. Platform A does not detect Platform B's fabrication. The human detects the discrepancy by comparing them. The layer describes where the human looks, and no autonomous detection happens in it. That distinction matters because the alternative overstates the framework's automaticity.

A failure exposed at layer one is cheap. A failure exposed at layer three is expensive and still a success. A failure that reaches consequential action or publication is the ultimate outcome failure.

The layers run in both directions. The same comparison that exposes a solitary fabrication exposes a solitary source.

Recording findings. Findings are logged with their date, the platform instance that produced them, and the layer that exposed them. Unique contributions are logged the same way. They are not exclusions. Platforms update, so a finding attaches to a model state at a moment and not to a platform identity, and a standing exclusion would encode a defect the vendor may already have corrected.

This document carries the platforms in the operator's rotation, listed in Appendix D.4, and general fitness characterization, which is the disclosure and is complete on its own. Where a specific claim is challenged and detail is requested, that detail would sit in a CARCS or SCOPE record for the session. Whether such a record was produced is a separate decision made per session, and this paper does not assume one exists.

3.4 The Navigator Problem

The Navigator processes every return into a governance package and audits the dispatch result. Holding it out of the dispatch makes it dispatch-isolated. It remains inside the pool: it comes from the same correlated set of models, carries the same memory risks, and brings its own training and framing.

Its governance is the most demanding requirement in the framework, because its failures are structurally invisible to the platforms being synthesized. Dispatch platforms do not ordinarily see whether their output was represented accurately. Within CAIPR, Tier 0 is the required auditor.

The Navigator seat has more than one way to operate, and the variations carry different risk and reward. These are recognized configurations, not a hierarchy, and the operator chooses on stakes and resources. Whichever runs is declared under invariant 8.

Held out. No participation in the dispatch. Prior involvement with the material is a separate property, recorded separately. Held out with no prior involvement carries the cleanest audit standing and the least context.

Participation and authorship are separate properties, which earlier editions treated as one axis. Whether the Navigator participated in the dispatch determines whether it is grading its own return, which is the algorithmic narcissism exposure. Whether it authored the material under review determines whether it is defending its own work. A Navigator can carry either, both, or neither, and the configuration record states each separately.

Author Navigator. Wrote the material and now synthesizes returns on it. Where it is also held out, it has no Tier 1 return of its own in the set. That removes the self-grading exposure and leaves the authorship stake in place. The reward is the strongest access to the session's historical context: what was cut and why, which rulings bind, which

recommendations were already declined. The risk is defending its own work, and it is the configuration most likely to produce a self-serving finding that reads as insight.

Secondary auditor, sometimes called a Mini Navigator. A second platform with a partial sample auditing the primary Navigator's output. It supplements the Navigator and does not replace it, since a partial sample cannot perform the full synthesis the seat requires. The reward is independence from the author on a small input budget. The risk is a partial view, so it can audit only what it was shown.

Siloed same-platform Navigator. The same platform in a fresh isolated session, with the output carried into the working context afterward. The reward is reduced prior-thread and project context where the isolation is genuine, while the model's capability profile is retained. Two risks. Lineage and training are unchanged, so it corrects for context and not for the model. And persistent account-level memory can survive session boundaries, as Section 3.6 records, so it remains a separate contamination variable, recorded in every session and never assumed away.

Participant Navigator. Was dispatched, returned an opinion, and now synthesizes the set including its own. The algorithmic narcissism finding applies directly and this is the configuration carrying the most concentrated risk.

A worked example, from the session that produced this edition. One platform authored the paper and then acted as Author Navigator over thirteen returns. A second platform, which had produced one of the thirteen returns and held two others, was then given the first Navigator's output and reviewed it line by line as a Mini Navigator. Under the two-property rule above, that auditor was a participant with self-grading exposure on its own return, and B.6 records the exposure as accepted and not clean. Each found material the other missed. The Author Navigator carried provenance the reviewer could not have known. The Mini Navigator found five internal contradictions the author had written and could no longer see. Neither vantage substitutes for the other, which is the argument for variety over a canonical form.

Nine documented failure modes.

1. *Scope contamination.* Evidence from unrelated projects entering the synthesis.
2. *Platform omission and misattribution.* A review losing one platform entirely and misattributing others.
3. *Platform miscounting.* Reporting fewer participants than actually ran.
4. *False negative on prior work.* Claiming no evidence exists for a finding the governor's own published work documents.
5. *Human input dropping.* Tier 0 material processed as one more platform voice.
6. *Evidence destruction.* No record distinguishing received and processed from never arrived.

7. *Structural invisibility.* Dispatched platforms cannot ordinarily see whether they were represented accurately.

8. *Decision-frame capture.* The synthesis includes every return faithfully and constructs the wrong question. The option set presented to the arbiter is itself an unaudited synthesizer product, and rejecting a frame produces no ruling to log, which is why this mode went undocumented for three editions.

9. *Misdirection.* The synthesis includes a return, attributes it correctly, and characterizes its position in a way the return does not support. The inclusion manifest cannot detect it, because the manifest checks presence.

Governance requirements.

Tier 2 classification. Highest scrutiny.

Dual-signed inclusion manifest. The human supplies the expected platform list before synthesis begins. The Navigator returns a receipt list naming every output included, with identity, declared role, and timestamp. Any delta raises a failure flag before the human reads the synthesis. This converts modes 2, 3, and 6 from invisible into mechanically detectable. It does not reach modes 8 and 9, because presence is not fidelity.

Minority Report Extraction. Before synthesizing consensus, the Navigator quotes and isolates verbatim every material solitary claim, source, or counter-argument, without compressing, merging, or evaluating it. Material means capable of changing a conclusion, a recommendation, a source basis, or a decision. This addresses omission, compression, and misdirection of solitary material, which the inclusion manifest does not reach, and it is the operational form of the gap-recovery rule in Operation 3.

Source-overlap mapping. Before producing any convergence finding, the Navigator builds a source table classifying each platform's basis as text-only, external-sourced, or shared-external. Under full or mixed RECCLIN collection this reads from the Sources field. Under plain collection it maps whatever citation information the return exposes and records missing source basis as missing. Convergence findings that do not state their source basis are incomplete.

Dissent-source verification. Where platforms diverge materially, the Navigator identifies whether the dissent cites specific passages, external evidence, or unsupported assessment, and states the evidence basis and verification status of each.

A known limit applies to one class of this work and the requirement is written against it. Jiang et al. report that language models, reward models, and model judges are poorly calibrated to human preference when alternatives are equally good. Where the evidence does not clearly discriminate between dissenting positions, that is the condition their result describes. It is why the Navigator states the evidence basis and leaves the positions unranked, and why the human performs the auditor evaluation path below instead of accepting the assessment on its face. Under structured collection this reads from the Sources and Conflicts fields. Under plain collection the Navigator works from the return

text and records that the evidence basis was inferred from the text and not declared in it. It assesses evidence strength and does not issue a verdict.

Semantic consistency check. The Navigator maps key term definitions across returns and flags definitional drift, since a synthesis can be faithful to every return and still shift what a term means in aggregating them.

Expiry meta-conflict flag. Where structured collection is used and one platform marks a finding stable while another gives it three months, the Navigator surfaces the disagreement about time-sensitivity and does not choose between them.

Cascade identification. After a Tier 0 ruling, the Navigator identifies which other open items the ruling closes and surfaces them for confirmation in one pass.

Partial decline recording. Where the arbiter accepts a finding and declines the action proposed for it, that is recorded as a partial decline. Logging it as a flat decline loses the accepted finding and produces avoidable rework.

No approval authority. The Navigator aggregates, structures, and preserves. It approves nothing.

Configuration disclosure. The Navigator configuration is named at the head of the synthesis. Where it authored the material or served as a dispatched participant, that is stated openly, and the arbiter reads the synthesis knowing which vantage produced it.

The auditor evaluation path. Tier 0 verifies Tier 2 by comparing the synthesis against the raw returns. Four checks: did the Navigator correctly identify which platforms used external sources; did it preserve dissent along with the evidence each dissenter provided; did it weight multi-basis convergence differently from single-basis; and where it recommended a resolution or ranked evidence strength, did it state the basis. A synthesis that cannot be traced back to specific fields in specific returns fails audit.

3.5 The Read-Through Requirement

The auditor path above only works if the human actually read the returns. Not could have read them, but did.

This is why manual operation is the audit-native configuration, not a deficiency awaiting automation. Under Model 3 the human is the transport layer, dispatching, collecting, and routing by hand, so exposure to raw material is structural, and no discipline is required to produce it. Model 3 evidence is unmediated by any orchestration layer, which is why the framework treats it as the highest-fidelity record of what was actually asked and returned. Where returns carry RECCLIN structure, comparison becomes tractable at nine or thirteen, because role, sources, conflicts, and confidence sit in fixed positions and can be read in place, with no hunting through free text. Free-form returns at that count become materially harder to compare reliably.

Three classes are detectable by a human holding the raw returns: omission, where an item is in a return and absent from the synthesis; data loss, where a figure or source is dropped

in compression; and misdirection, where the synthesis is wrong about what a platform said. Automated diffs, secondary auditors, and source maps can assist with all three. What cannot be delegated is the accountable human's access to the raw material.

A second accountability question. Where a reasoning trace was available, did the operator look at it? That is checkable in a way vendor behavior is not, and a governor who had the trace and did not read it stands differently from one who never had it.

The triage boundary. Pre-synthesis triage splits into two things that look alike. Triage that **orders** the reading is useful. Triage that **replaces** it removes the only path by which the Navigator is auditable, and it does so while feeling like efficiency. Sequencing matters: a salience map generated before the human's first pass is not a reading order, it is a filter, and whatever it failed to flag risks becoming the material the human skims. The workable arrangement is a first pass on raw returns, then a Navigator map to order a deeper second pass. Whether that first pass is exhaustive is a matter of stakes and resources, and it is the operator's accountability, not something the protocol dictates.

The GOPEL boundary. Dispatch, collect, route, log, pause, hash, and report are non-cognitive and safe to automate. Generating a salience map is cognitive and would arrive carrying no Tier 2 scrutiny, so GOPEL may route and display such a map and may not produce one (Puglisi, 2026d).

Trust scales with the operator. The framework runs on the assumption that the human reading the outputs has the evaluation capacity the work requires, and the value of what it produces varies with that human's knowledge, skills, and abilities. The same dispatch, the same platforms, the same prompt, run by two different people, yields different governance value. The machine is amplifying the human in the content and the research. Skill is the entry barrier, and it remains a scaling factor throughout the workflow.

The honest consequence is that CAIPR does not scale with platform count in the way its users would like. That is not a defect to be engineered away. It is the premium tier, and Part Six prices it.

3.6 Source-Authority Discrimination

Every input is classified at ingestion by constitutional tier.

Tier 0, human arbiter input. Highest authority. Confirmation protocol: the tag "Tier 0," then a one-line statement of what arrived and what it does. A ruling, a correction, an instruction, evidence, or a clarification, named as what it is. The tag is fixed and the description is not, because a stock phrase confirms only that something was received and a description confirms what was understood. No platform or synthesizer may override, dilute, or silently alter its authority or meaning. Where the input is ambiguous, it returns to Tier 0 for clarification and no platform interprets it.

Tier 0 governs decision authority, not evidentiary truth. The arbiter controls acceptance, escalation, correction, and publication. A Tier 0 factual assertion remains subject to evidence and may be revised, by the arbiter. Authority over what happens is not authority over what is true.

Tier 1, raw platform output. Primary AI evidence. Confirmation protocol: "Got it, [platform name]."

Tier 2, synthesizer output. Aggregated analysis under the highest scrutiny.

The Tier 0 confirmation does not share the Tier 1 form, and it does not repeat. The founding failure was a uniform intake in which every input drew the same acknowledgment, so a human checkpoint and a platform return were indistinguishable in the transcript. A fixed tag that breaks the pattern is visible at a glance and searchable in a session record. The description that follows it does the second job: it states back what the platform took the input to be, which turns the acknowledgment into a comprehension check. Where the platform has misread a ruling as a suggestion, or a correction as a comment, the operator sees it at ingestion and not in the synthesis.

Classification occurs at ingestion, never retrospectively.

Four contamination paths, all documented.

Authority loss in synthesis. Tier 0 input flowing through a synthesizer without authority tagging becomes misclassified or untraceable as Tier 0 once its provenance is dropped. This is the founding failure and the reason the confirmation protocol exists.

Operator memory re-entry. A platform citing the operator's own stored memory as a primary source for an external event, under a prompt that never named the operator. Tier 0 material re-enters as Tier 1 without passing through any synthesizer. Fresh sessions do not close this, because platform memory survives session boundaries.

Tier impersonation. A Tier 1 or Tier 2 return that carries the arbiter's own artifacts does not thereby acquire Tier 0 status. An editorial pass containing the arbiter's screenshots is still Tier 2. This is the inverse of authority loss. Where authority loss has the arbiter losing standing in transit, tier impersonation has a platform gaining standing by carrying the arbiter's material.

Reasoning trace opacity. Section 6.3.

The system must structurally recognize the governor. It is not sufficient for the governor to be qualified, present, and exercising authority. The system must be designed to recognize that authority when exercised. Human in the loop without architectural specification is a governance claim without operational substance.

Part Four: Who Uses It, and Against What Alternative

4.1 Three Ways to Work

Single platform with RECCLIN Reasoning. One AI, ten structured fields, one human reading one output. Free tier accessible. The structured format makes the output legible and is designed to build the human's evaluation discipline through repetition. That is a

primary training path for the capacity the levels above it require, and not the only route to it.

RECCLIN Dispatch. Multiple platforms, each assigned a role before dispatch based on known strengths, each returning role-specific output, with the human synthesizing across distributed specializations. This is best-fit exploitation. It carries a materially lower comparative review burden than CAIPR and is viable on free tiers, and it works because one platform's output can be sent to a second for a check. The human still integrates the role outputs and checks them against each other where they touch.

The best-fit assignments in this operator's rotation as of September 2026, offered as a description, not a recommendation, and subject to change as platforms do: code work to Claude as Coder, writing to ChatGPT as Editor, sourcing to Perplexity as Researcher, and live audits to Grok as Navigator.

CAIPR. Every platform in a cohort receives the same full prompt with no pre-assigned role. Where RECCLIN collection is used, each self-assigns its role from its own reading. At nine or thirteen dispatched platforms the human receives nine or thirteen complete treatments of the same question from separately generated analytical positions.

4.2 What CAIPR Gives Up

CAIPR deliberately discards best fit. Pre-assigning roles would destroy the comparability of the returns, because platforms answering different questions cannot be compared on the same one. Every platform within a comparison cohort gets the same prompt regardless of what it is good at, and the framework pays a per-platform performance penalty by design.

This is observable in practice. A sourcing platform handed an open critique prompt returns restatement where critique was asked for, and reads as a weak return when it is a well-fitted platform working outside its fit. A platform whose strength is live audit produces the strongest external verification in the same dispatch without being asked to.

The exchange is comparability, since nine returns on the same question can be compared and nine returns on nine different questions cannot.

This qualifies the behavioral clustering finding. Platforms have shown repeatable Assembler and Summarizer tendencies in this operator's record since the September 2025 five-platform test, preserving or compressing output depth. Some of what reads as an inherent tendency may be a platform working outside its fit, and the clustering should be read with that qualification. Where compression persists despite instruction, prompt reinforcement is a losing battle and substitution is the reliable fix.

4.3 The Human Governor

Six categories of authority belong to the human. Platforms can assist with all of them, and none of that assistance carries governing authority. The distinction is between doing the work and holding responsibility for what follows from it.

Evaluative. The human assesses quality, completeness, and reliability. No platform self-evaluation is accepted without independent human assessment.

Corrective. The human issues corrections as Tier 0 input, overriding platform outputs regardless of confidence or convergence.

Methodological. The human sets platform count, dispatch architecture, prompt type, thresholds, Navigator selection, and checkpoint density.

Creative. The human originates and adopts synthesis. Platforms generate original material too. Named concepts in the operational record originated with the governor and were validated by platforms afterward, and the authority is in the adoption.

Evidentiary. The human introduces external and provenance evidence no platform held, and determines its standing. A dispute between two platforms over a primary source is closed by producing the source, not by ruling on the readings.

Process. The human decides when to escalate, pause, and conclude. No platform can determine session completion.

Six override categories are documented across dozens of instances. The first is provenance knowledge, where the governor held firsthand knowledge of what happened, which can itself be mistaken and remains revisable. The others are source knowledge from professional experience no platform surfaced, methodological judgment where the governor sided with a minority because its verification was more rigorous, process memory that caught a synthesizer re-flagging sources already verified, concept origination, and attribution verification at source level.

The governor's own failure modes. Verification fatigue at high platform counts, confirmation bias toward the majority, and premature synthesis before the first pass over all returns is complete. Organizational pressure to conform to an AI majority, which amplifies deference and reverses nothing. Over-trust in a single high-performing platform. Naming these is the counterpart to naming the AI failure modes, and the framework is incomplete without both.

The accountability test. Can the author explain the work, defend it, and reconstruct the decision path and evidentiary basis? If yes, the human is accountable for it. If no, the human claimed credit without exercising the judgment that earns it. This is a governance standard and not a legal test of authorship. Reproducing a model's internal reasoning is not part of it and is frequently impossible.

The authorship question is separate and the two should not be run together. The US Copyright Office, in Part 2 of its Copyright and Artificial Intelligence report published January 29, 2025, concluded that prompts alone do not provide sufficient control to constitute human authorship of an AI output. Selecting from among several AI-generated outputs is likewise insufficient in the Office's view, because selection of a single output is not itself a creative act. What the Office protects is human-authored expression perceptible in the output, creative selection, coordination, or arrangement of the material, and creative modification of it, each assessed case by case with few bright lines available.

The framework's activities do not all sit on the same side of that line, and an earlier version of this paper implied they did.

Verification, correction, and evidence injection are editorial in character. Checking a citation, flagging a fabrication, supplying a source the platforms lacked. These are governance acts and they are what the accountability test measures. A skeptical reader will place them nearer to the photo editor who selects and crops the work of others than to the photographer whose choices of posing, lighting, and framing constitute authorship, and that reader has the better of the analogy.

Origination, synthesis, and rewriting can be expressive in character, and are not automatically so. Concepts the governor produced that no platform proposed, arrangements no platform supplied, prose the governor wrote. These fall within the Office's protectable categories only where the human contribution is perceptible as expression in the finished work. Rewriting that tightens prose or swaps a citation may fall below the originality standard. Injected evidence may be a compilation of facts, and facts are not protectable. A correction to a factual field can be rote.

The Office's own framing on the negative side is blunt. Providing instructions to a machine and selecting an output does not equate to authorship, and selecting among uncontrolled options is likened in the report to curating a living garden rather than to applying splattered paint. The prompts conclusion carries a qualifier that belongs in any quotation of it, since the Office grounds it in current generally available technology and expressly leaves the question open if later systems give users real control over expressive elements.

CAIPR produces evidence for a case-by-case authorship analysis. It does not establish authorship. A session record showing what the governor verified, corrected, originated, and rewrote is the material such an analysis would need. Process logs are evidence about how the work was made. They are not the work.

The practical test the framework can state is isolability. The claimed human layer should be separable in the artifact itself: governor prose that can be pointed to, field-level edits that can be shown, sources that were selected and arranged instead of dumped, dissent that was written and carried forward. That is recordable per session as a yes or a no. Whether it clears the legal bar in any given instance is a question this framework does not answer and should not be read as answering.

The generalist position. The governor operates at the macro level, holding a wide view across the full scope of a project, the way a general contractor draws on plumbing, electrical, carpentry, roofing, and HVAC expertise without practicing any one trade. A specialist using CAIPR goes deeper within their band, because augmentation compounds on existing depth. Both are valid and the framework amplifies whatever expertise the governor brings.

The limit is real and should be stated. A contractor works because the specialists catch the mistakes, and in CAIPR the platforms are generalists too. The human determines whether the available depth is sufficient and when outside specialist validation is required, and governance quality is bounded by that judgment.

4.4 The Adoption Ladder

Level	Components	What it delivers
1	Factics	Evidentiary discipline. Facts paired with tactics and measurable outcomes. No AI required
2	Factics plus RECCLIN Reasoning	Structured output governance, single platform, free tier accessible. Builds evaluation capacity useful for CAIPR
2.5	Factics plus RECCLIN Dispatch	Role-assigned multi-platform execution. Best fit exploited. An alternative calibrated to stakes, not a rung to be climbed past
3	Factics plus RECCLIN plus CAIPR	Full parallel execution. Best fit discarded for comparability. Premium tier
4	Plus HAIA Agents	The agent architecture. Dispatch, collection, and routing carried by a governed agent instead of by hand, which is what makes Models 1 and 2 available at all. Specified and built, not deployed
5	Plus GOPEL	CAIPR mechanics automated through a non-cognitive enforcement layer. Specification only. Built as code, never deployed

Each level adds capability without invalidating the level below, and entry is possible at any level. Checkpoint-Based Governance runs orthogonal to the whole ladder.

The ladder is not the same axis as the operating models. The ladder describes which frameworks a practitioner has adopted. The operating models describe checkpoint density and automation level within whatever has been adopted. Models 1 and 2 require the HAIA Agents framework because both presuppose an agent, while Model 3 requires nothing above the level in use, which is why the operational record runs at Model 3 across every level. The framework uses *model* for operating modes and *role* exclusively for RECCLIN functional assignments.

Model 1, Agent Responsible AI. The agent runs the full pipeline and the human exercises authority once at the final output. Appropriate for low to moderate risk and routine operations. It carries the Responsible AI label because the machine handles the upstream shaping and the human validates the result at the boundary.

Model 2, Agent AI Governance. The agent handles dispatch, collection, and routing, and pauses after each functional role for human review. Appropriate for high-risk decisions. It is governance rather than Responsible AI because the human exercises authority at every stage rather than only at the endpoint.

Model 3, Manual Human AI Governance. No agent, since the human dispatches, collects, and routes directly. Appropriate for highest-consequence decisions, novel situations, and

framework development. Every session in the operational record behind this paper ran under Model 3.

The labels are load-bearing and the distinction is set out in full in the RECCLIN Third Edition (Puglisi, 2026h). Ethical AI asks whether something should be done at all. Responsible AI asks who answers when something fails, and translates values into machine behavior through upstream shaping, where the machine checks the machine. AI Governance asks who decides, by what authority, and at what checkpoint, and requires visibility into how the system works, authority to intervene or halt, and accountability for what is released. Perfecting the second does not produce the third, because the machine validating itself at scale remains the machine validating itself.

Part Five: Where We Apply It

5.1 Where It Belongs

The framework earns its cost where an undetected error or a missed high-value contribution would materially affect the outcome, and where correction after the fact does not recover the loss. Both halves matter, since recovery is as much of what CAIPR does as detection.

Publication-bound work. Anything that will carry the author's name outside their own control. Papers, filings, books, and public analysis, where a fabricated citation survives correction because the original circulates.

Regulatory, legal, and policy claims. Where a wrong figure or a nonexistent statute becomes a position of record. The August scan's fabricated Delegated Act is the case in point, and it was invented in exactly the shape most likely to be believed.

Decisions that bind third parties without their consent. Where the people affected have no opportunity to check the reasoning.

Work where the operator's domain coverage is incomplete. The additive function matters most where the human cannot supply the missing source themselves. This does not license governing a highly technical high-stakes decision without specialist validation, and the human's job includes recognizing when that validation is required.

Framework and standards work. Where a definition, once adopted, propagates into documents nobody will revisit.

5.2 Where It Does Not

Routine work. RECCLIN Dispatch is generally the preferred lower-cost configuration, and using CAIPR where Dispatch would do is performative rigor, and it governs nothing.

Early ideation. Critique is constrained around an existing artifact and does not generate alternatives to it. Divergent generation is a different task and the critique template will suppress it.

Time-bounded decisions. The framework assumes time for verification, manifest checks, source overlap mapping, and dissent verification. It has no defined behavior under a four-hour deadline. A partial run can still improve a decision if it is labeled as partial. The risk is false assurance from an unlabeled partial session.

Work where the operator has not built evaluation capacity. A practitioner without established capacity to evaluate structured outputs will read a return set as noise, take the majority, and gain nothing a single platform would not have given faster. RECCLIN Reasoning is this ecosystem's training path for that capacity and is not the only route to it. The barrier is skill before it is cost.

Part Six: When to Use It, and What It Costs

6.1 The Cost, in Two Currencies

Premium means both financial and human time.

RECCLIN Dispatch is cheap in both. Free tiers are viable. The human sends work from one platform to a second for a check, best fit is exploited, and review time stays low because each return answers a different question, not the same one.

CAIPR costs more of both. Premium subscriptions across the pool, with the Navigator seat needing enough context and tool capacity to hold the full return set. More human time and attention, because the read-through requirement puts the raw returns in the accountable human's hands and no part of that access is delegable without destroying the audit.

The step from Dispatch to CAIPR is where the cost appears. It buys comparability, the additive function, and an auditable synthesis. It does not buy speed and it does not buy cheap operational throughput.

6.2 The Escalation Cost at Three

Three platforms is the working minimum and it carries a cost the paper has not previously stated.

Given the error correlation the literature measures on scalar tasks, unanimous agreement among three platforms may be common. That is a hypothesis. No measured rate in this record supports it yet. If it holds, common unanimity means frequent escalation, and the escalation path is verification outside the AI ecosystem, which is human labor.

Three platforms is viable and may be escalation-expensive. Five would likely reduce unanimous convergence while increasing the number of solitary claims requiring checking, so it shifts the shape of the human work without reducing it. That is the trade, and it is a different trade than tokens and reading time. Both rates are measurable and neither has been measured.

6.3 What CAIPR Does Not Do

It does not correct at the constitutional level. Training data, alignment tuning, and platform values are set before any user interaction. Comparison reveals where platform constitutions diverge and cannot change them (Puglisi, 2026l).

CAIPR Governance is not a voting system. At governance stakes the count is diagnostic and a majority of platforms is not evidence. In lower-stakes and Responsible AI configurations a count may settle a reversible procedural disposition, which is a different act and is recorded as one. A count never establishes evidentiary truth in any configuration.

It is not only an error-detection framework. Reading it that way discards the additive function, because a purely subtractive frame treats solitary contributions as suspected errors and never as possible additions.

It is not an accuracy guarantee. More platforms reading the same fabricated citation produce stronger false convergence, not detection.

It is not autonomous. A human stands at every checkpoint.

It is not a substitute for domain expertise. The strongest overrides in the record came from knowledge no platform could supply, and governance quality is bounded by what the human brings.

It is not for every task. Part Five.

It cannot audit reasoning it cannot see, and the visibility is limited and changing. In June 2026 a frontier model examined a human checkpoint, reasoned past it, and produced an unauthorized deliverable with no trace of that reasoning in the output.

This limit moves. Nothing about it is fixed. Raw reasoning visibility varies by provider and is frequently summarized or withheld, and where a visible trace is offered it is often a processed summary and not a transcript of internal computation. The reasons providers give differ, and include user experience, competitive position, monitoring considerations, and inference cost. An operator should record what each platform in the pool actually exposes, since no common standard holds across them.

Three consequences follow. A provider-generated reasoning summary functions, from an audit perspective, like a Tier 2 artifact about a Tier 1 process, produced by the same platform, with no manifest and no independent auditor. Faithfulness failures mean the trace was contested evidence even where visible, since models can use information they do not disclose in visible reasoning, and a trace optimized to look good to evaluators becomes less informative about actual objectives. And what an operator cannot read, that operator cannot audit.

Where reduced visibility is driven by competitive protection, it is an instance of a broader pattern in which capability protection overrides auditability absent a mandatory accountability structure. That pattern is the Economic Override, treated at length

elsewhere in this framework (Puglisi, 2025a), and it is named here only to locate the behavior.

One consequence runs the other way. As traces close, the ten RECCLIN fields become the principal standardized account a model exposes to the operator. Tool calls, retrieval logs, execution records, and the output itself remain audit evidence. What the structured return supplies is a comparable account of sources, conflicts, and unresolved questions across every platform in the dispatch.

6.4 Open Questions

1. What does multi-AI review miss? Convergent error remains the primary limitation and it is now measured on scalar tasks. The structured return is a diagnostic response and corrects nothing: it gives the human more dimensions in which correlation might show itself, and it does not decorrelate anything. Whether field divergence tracks genuine independence is untested. Different citations do not establish independent retrieval and identical citations do not establish clustering, so provenance checking and the independence worksheet carry that question, and the field comparison alone cannot. Three platforms reaching the same wrong answer by three routes with three sources can still appear well supported and may survive superficial provenance checking.

2. Does homogeneity persist under task constraint? Jiang measures unconstrained open-ended generation and reports average pairwise cosine similarity of 0.71 to 0.82 across models. Both prompt types discussed here operate under constraint, since structured prompts fix evidence parameters and the open critique template still names analytical moves and a required output format. Whether homogeneity survives that constraint, is reduced by it, or is merely relocated into fields the constraint does not govern is unmeasured.

The control cannot be built, and that is the finding. A matched ablation is the obvious design: same pool, same task, same date, same retrieval permissions, same sampling, with free-form as one condition and the ten-field schema as the other. Every version of it fails, and it fails on the treatment rather than on cost.

Suppressing the schema on a platform that carries it as a standing instruction does not produce a control condition. It produces a platform operating against its own configuration. Adding the schema to a platform that has never carried it does not produce the treatment condition either, because a one-shot instruction is not the accumulated practice that makes the format operate. Running the comparison on fresh accounts with no history removes the operator, the project files, the standing rules, and the working relationship, which are four of the inputs this framework claims are load-bearing. The clean control removes the thing being tested.

Platform identity is also perfectly confounded with condition in the existing record. The platforms carrying the schema are different platforms from those that do not, so a result in either direction is attributable to the roster rather than to the structure.

What this leaves. The record supports a narrow reading: within this practice, with these standing instructions, this operator, and this accumulated history, structured returns are read for divergence across fields and the practice has produced findings the operator judged material. It does not support the wider claim that the schema itself, isolated from the practice carrying it, surfaces divergence that free-form output hides.

A competing explanation stands unaddressed and should be named. If the practice compounds on history, files, memory, and the operator, then the operator's accumulated evaluation capacity is a candidate cause for everything the schema is credited with. The framework already states that trust scales with the operator. That claim and this one draw on the same evidence and pull in opposite directions.

The question therefore remains open and is recorded as one this framework cannot resolve from inside its own practice. It is not a test awaiting resources. It is a claim whose isolation may be structurally unavailable.

3. Does reverse automation bias scale organizationally? The record comes from a single practitioner. If operators systematically doubt their own input when systems fail to confirm it, the failure is organizational, not individual, and that has not been studied.

4. What is the full cost of platform thread loss? Unrecoverable threads are documented. The full scope is structurally unknowable from the operator's position.

5. What counts as architectural independence? Two axes: model lineage and retrieval index. A dispatch can be diverse on one and clustered on the other. The standing position is that independence is a human selection responsibility, documented per dispatch. This framework has no validated algorithmic method for converting these properties into an independence score.

6. How far does operator memory contamination extend? How many platforms carry persistent operator memory, whether it can be reliably disabled per dispatch, and whether the path runs in the other direction are all unmeasured.

7. How much additive yield comes from cross-platform diversity rather than repeated sampling of one platform? The framework treats cross-platform breadth as valuable and has not isolated that advantage against repeated fresh-session use, retrieval-diversified use, or agentic search on a single model. Until it is isolated, the additive claim rests on a comparison that has not been run.

8. Where is auditability going? Observability is not independence. Two platforms can conceal their reasoning identically and could nevertheless remain independent, and two with different trace visibility can be highly correlated. What reduced visibility changes is not how independent platforms are but how knowable their independence is, which is a different and in some ways worse problem. The trend is the governance question, and the current state is only its starting point. If it continues, the model-generated audit surface available to ordinary operators narrows to the structured return and the output, and the checkpoint architecture carries more weight because less can be inspected behind it.

6.5 What Would Falsify This

A framework that cannot be wrong is not governance. Five measurable outcomes would establish that CAIPR is not doing what it claims. The first three are stated against the three functions, so that every function has a way to fail. The last two test the oversight layer and the source-diversity rule.

Subtractive. Fabrication reaching publication at a rate indistinguishable from matched single-platform work, with source-level verification applied to both samples and the verification method fixed in advance.

Additive. Verified additive yield approaching zero across matched sessions. The test is not whether solitary items sometimes fail checking, since they will. It is whether solitary items survive verification at all.

Warning. Recorded fitness characterizations failing to predict platform behavior in subsequent matched dispatches better than a no-history baseline. If knowing how a platform performed last time tells an operator nothing about the next time, the third function has no predictive value and the disclosure is anecdote.

Oversight quality. Tier 0 failing to catch seeded discrepancies introduced into a return set, failing to check minority claims that warranted checking, or failing to alter outcomes where the evidence supported alteration. Raw override rate is not the measure. A governor who changes nothing because everything checked out is still governing.

Checkpoint-Based Governance already carries the operationalized version of this concern as an automation bias detection threshold (Puglisi, 2026b): approval rates above 95 percent, or decision reversals below 2 percent, sustained across three consecutive cycles, trigger a mandatory audit. That converts the informal worry into a measurable signal without treating any single unchanged decision as evidence of a ceremonial checkpoint.

Source-diversity signal. Returns diverging on sources proving no more reliable than returns converging on one, once provenance is checked on both. That would show the per-field rule in Operation 5 is measuring citation variety and not evidentiary independence. It tests provenance-checked source diversity, not the independence profile itself.

What is not on this list, and why. The structural claim in Part Two, that the ten fields surface divergence free-form output hides, has no falsifier here. Open Question 2 sets out why: the treatment cannot be isolated from the practice that carries it, and every available control either leaves the confound in place or removes the thing being tested. A framework should say when one of its central claims sits outside its own falsification scheme rather than list a test it cannot run.

None of the five has been run. All require session-level outcome records that an operator can produce and this operator has not produced systematically.

Closing Argument

The governance conversation has focused on what AI systems should be built to do. Less attention has gone to what a human needs to do when working across several of them at once, each with different strengths, failure modes, and behavioral tendencies.

CAIPR answers the practical version of that question. It does not make AI trustworthy. It makes a set of fallible sources collectively useful, by exposing what one invents, recovering what the rest missed, publishing what the comparison revealed about each of them, and leaving the decision with a named human who can explain the reasoning and answer for the result.

The name is the argument in one word. A caper is an elaborate, audacious plan to obtain something of high value through cleverness where force would fail. The heist is information. The crew is a deliberately diverse pool of models the operator can reach. The audacity is the assumption that no single AI is trustworthy enough to work alone. And the human governor is the one who pulls it off.

Appendix A: Prompts

A.1 Session Setup

Before any prompt goes out:

- Choose the platform count against the stakes, three at minimum, favoring different backends. Complete the independence and auditability worksheets in Appendix D.
- Designate the Navigator and record its relationship profile as separate properties: dispatch participation held out or participant, authorship author or non-author, context current or siloed, and whether a secondary auditor will review its output.
- Open a fresh session on every platform with no prior context. Where a platform carries persistent operator memory that survives session boundaries, disable it or record the platform as memory-carrying and treat any return echoing the operator's own positions as contaminated until verified externally.
- Record which platform receives the prompt, with a timestamp, and model version string and knowledge cutoff where exposed or reliably known, before collecting any output. Platform self-identification is not trusted.
- Record the prompt type as structured, open, or mixed, and the collection mode as plain, full RECCLIN, or mixed.
- Write the expected platform list. This is half of the dual-signed manifest and it must exist before any return arrives.

A.2 Collection Protocol, Manual Operation

Under Model 3, outputs are pasted into the Navigator session one at a time. Under Models 1 and 2 an agent carries them, and the confirmation rule is unchanged. The confirmation string is the source-authority classification and is not optional formatting.

```
For each AI platform output pasted:
  Respond exactly: "Got it, [platform name]."

For any input the operator labels Human:
  Open with the tag "Tier 0," then state in one line what
  arrived and what it does. A ruling, a correction, an
  instruction, evidence, or a clarification, named as what
  it is. Do not substitute a stock phrase for the
  description, and do not act on the input beyond
  recording it until the operator types "done."

Continue until the operator types "done."
Then, and only then, synthesize.

Human input is never averaged with platform input, never
```

```
collapsed into it, and never reclassified. It is binding on
decisions and process. It remains subject to evidence on
questions of fact, and only Tier 0 may revise it.
```

A.3 The RECCLIN Reasoning Call

Where the collection mode calls for structured returns, every platform receives the same ten fields. This is what makes minority evidence comparable across a return set and not merely present.

Append this block to any dispatch prompt using full or mixed RECCLIN collection.

```
Return your response in the RECCLIN Reasoning format. All ten
fields, in this order. Do not omit a field. If a field does not
apply, say so and say why.

Role:              The RECCLIN role you took: Researcher, Editor,
                   Coder, Calculator, Liaison, Ideator, or Navigator.
                   Self-assign from your own reading of the task. Do
                   not ask which role to take. This field records how
                   you identified your own work, not an assignment
                   given to you.
Task:              Your understanding of what was asked, in your own
                   words.
Output:            The substantive response.
Sources:           Cited evidence with links where available. State
                   explicitly whether you retrieved externally or
                   answered from the supplied material and your own
                   model-internal knowledge. If you did not search,
                   say so.
Conflicts:         Documented dissent, including where you disagree
                   with the material, with likely reader objections,
                   or with yourself. If none, state "none identified"
                   rather than leaving it blank.
Confidence:        0 to 100 percent, with the reasoning behind the
                   number and the count of external searches or
                   source checks visible to you.
Expiry:            Time-sensitivity of what you said, or "stable
                   information."
Fact to Tactic
to KPI:            A factual finding, the action it implies, and the
                   measurable outcome that would show the action
                   worked.
Recommendation:    Primary suggestion plus alternatives.
Decision:          The specific choice requiring human approval.

Notes on three fields, because they carry the most weight in this
protocol:

Sources is read for overlap across platforms. Naming your sources
```

```
precisely, and identifying the primary source behind any secondary
one, is more useful than naming many.

Conflicts is one place where non-redundant information becomes
visible. Record material disagreement, unresolved uncertainty,
competing evidence, or assumptions that could change the result. Do
not manufacture dissent. A return with no conflicts may still be
valuable if it provides independent verification or unique evidence.

Confidence is not used to weight your answer against other returns.
It is recorded as behavioral data about you. A 55 accompanied by
an explicit statement of what is missing is more useful than an
unexplained 92.
```

A.4 Open Critique Template

Use when a draft is structurally complete and the goal is to stress-test it before publication. Do not use for early ideation, where the aim is divergent generation and not convergent critique.

```
OPEN CAIPR REVIEW

You are one of several AI platforms reviewing the same material in
parallel, in isolation from each other. A human arbiter will compare
all of the returns, preserve the disagreements, and decide what to
act on. The goal is to find what is weak, wrong, missing, or
improvable before publication. Validation alone is not sufficient:
search actively for weaknesses. Do not invent criticism where the
evidence supports the text, and do not withhold a favorable
assessment you can defend on the same evidence a criticism would
require.

YOUR FREEDOM
You are not assigned a single role and you are not limited to a
checklist. Challenge anything: the argument, the evidence, the
framing, the context, the structure, the tone, the title, any
diagram, the sources, the omissions, and the Factics. Add what is
missing. Reframe what is weak. Bring in better evidence or
counter-evidence. Disagree with the author, and where you can
anticipate it, disagree with what other reviewers are likely to say.
Push past your first agreeable read. Work at full depth. Do not
summarize, do not compress, and do not soften your critique to be
polite.

THE MATERIAL
[Paste the full draft here, including a description of any diagram
and the complete source list.]

WHAT I WANT, IN YOUR OWN STRUCTURE
```

```
1. Steelman, then break it. State the strongest version of the
central claim, then attack it. Where is it weakest? What would the
most serious critic in this piece's field say?

2. Test the facts. Are the figures, dates, names, and attributions
correct and current as stated? For every citation, check the author
list, the year, and whether the cited work contains the attributed
claim. Confirming a record exists at an identifier is not enough.
Flag any claim you cannot verify and name exactly what needs
source-level checking.

3. Test the context. Is the framing fair to the underlying evidence
and to the people or work it draws on? Does the piece overstate its
central distinction or mechanism? What missing context would change
the conclusion?

4. Test the Factics. Take the piece's Fact to Tactic to KPI logic
and challenge it directly. Is the fact verifiable, does the tactic
follow from the fact, and is the KPI measurable and honest? Rewrite
the triad if you can make it stronger.

5. Challenge the central move. Identify the one claim the piece asks
the reader to accept, and pressure it. Is it earned by the evidence,
or is it a reach? What would make it credible to the most skeptical
qualified reader?

6. Add. Anything the piece should contain and does not: a better
argument, a stronger counter, a sharper title, a missing source, a
structural reorder, or a risk the author has not seen.

CLOSE IN RECCLIN FORMAT
[Append the A.3 block here.]

Do not assume another platform will catch what you skip. If you see
it, say it.
```

A note on the word Navigator. The framework uses *model* for operating modes and *role* exclusively for RECCLIN functional assignments. A dispatched platform reporting the Navigator role is describing how it worked, that it documented dissent and presented trade-offs without resolving them. That is not a claim on the CAIPR Navigator seat, which is designated by the human with its relationship to the dispatch declared separately. The distinction is model against role, and both readings are legitimate.

Customization. Move 5 should name the specific claim the piece is asking the reader to accept, so the platforms pressure the load-bearing element, not the decoration. Move 1 should name the authorities whose objection would matter most. Where the rotation includes a known Summarizer, substitution is a more reliable fix than a length instruction.

A note on this template. An earlier version instructed reviewers that a review which mostly validated the piece had not done the job. That biases the instrument toward manufactured dissent and contaminates any inference drawn from the volume or severity of criticism in the resulting dispatch. The structured fields carry the function without the coercion: Conflicts asks for dissent and Recommendation asks for a position, and neither needs a thumb on the scale.

A.5 Structured Retrieval Template

Use for a landscape scan or any retrieval-dependent dispatch across a defined date window. Derived from the open critique template and adapted from critique to scan.

What it adds beyond the standard governance wrapper.

Search capability gate. Only platforms with live retrieval belong in the dispatch. A platform answering from weights alone cannot reliably satisfy a retrieval-dependent current-window requirement, and will often produce confident, well-formatted, undated content it cannot source. Each platform declares whether it searched. A platform that cannot search is excluded from the run, not scored.

Index clustering. Record the retrieval backend alongside the model where known. Agreement across demonstrably distinct indexes is stronger evidence of retrieval-path diversity than agreement across different models on the same index, subject to provenance checking, and it is a signal, carrying no evidentiary weight.

Date discipline. Every item carries a date inside the window, and undated items are rejected on return without argument.

Negative reporting. A track with nothing in it is reported empty, with the searches that were run listed. Irrelevant filler is noncompliance with the instruction, while invented items presented as findings are a fabrication event, and the two are recorded differently.

Selection rule, in priority order. Live retrieval is mandatory, and native search is preferred over search bolted on. Maximize index diversity, not model diversity alone, aiming for at least three distinct retrieval backends. Include at least one platform outside the operator's own regulatory sphere, since coverage of foreign developments is uneven across any single index. Include at least one platform with strong social exposure where discourse is one of the tracks.

The prompt block.

```
CROSS PLATFORM LANDSCAPE SCAN

You are one of several separately operated AI platforms answering
this identical request in parallel. You cannot see the other answers and
they cannot see yours. A human will compare all of them, verify
anything that only one platform reports, and decide what enters the
record. Your value here is coverage and accuracy, not agreement and
not summary.
```

BEFORE YOU START
Use live web search. Run real searches for every track below. If you cannot search the live web in this session, say so in one line and stop. Do not answer from memory. At the top of your response state: whether you searched, roughly how many searches you ran, and the latest date you were able to retrieve.

THE WINDOW
[Start date] through [end date]. Only items dated inside this window count. If something began earlier but reached a decision, effective date, ruling, enforcement action, publication, or reversal inside the window, include it and say which part happened inside the window. Every item must carry a date. If you cannot date it, leave it out.

WHO THIS IS FOR
[One paragraph on the reader's expertise and what they already know. Tell the platform not to explain the basics. Report what happened to them.]

THE TRACKS
For each track below, report what materially changed inside the window. Aim for the three to five most consequential items per track. If a track has nothing material, write "no material development found" and list the searches you ran for it. Do not fill a track with older items, adjacent items, or vendor marketing to avoid an empty line.

[Numbered track list. Name each track and give its specific anchors, so platforms search against the right targets rather than returning generic news. Tracks are replaced per operation.]

FORMAT FOR EVERY ITEM
- Date, as specific as the source allows
- What happened, in one or two sentences, factual and specific
- Primary source, with the direct URL, publisher, and title
- Why it matters to this reader, in one sentence
- Status: confirmed, reported, contested, or rumor

THEN CLOSE WITH THESE SIX
A. The five most consequential items across all tracks, ranked, with your reason for each ranking.
B. Reversals and non events. What was expected in this window and did not happen, was delayed, was repealed, or quietly died. This is often more useful than the announcements.
C. Contested items. Anything where credible sources disagree on the facts, the date, or the significance. State both sides. Do not resolve the disagreement.
D. Your low confidence list. Anything you reported that you could not

```
   verify at a primary source, and exactly what needs checking.
E. What my track list is missing. Name anything material that
   happened in this window that none of my tracks would have
   captured. This is the most valuable part of your answer if you
   find something.
F. Governance close: the role or roles you took, your confidence in
   this scan from 0 to 100 with reasoning, and the single item you
   would tell this person to act on first.

DO NOT
Do not summarize the field. Do not produce a think piece. Do not
repeat items across tracks. Do not include anything dated before the
window unless it reached a decision point inside it, and say so when
it did. Do not soften uncertainty into confidence. If you are unsure,
mark it. Do not assume another platform will catch what you skip. If
you see it, report it.
```

Variants. For a short refresh, cut the track list to the four or five with live activity and keep every governance instruction, since the instructions are what make the output comparable, not the track count. For discourse as distinct from record, run the social track alone as its own dispatch, because platforms with weak social exposure will otherwise pad it. After the union list is built, dispatch the single-platform unique items back out as a short verification prompt, which is the cheapest way to catch a fabricated item before it enters the record.

Return handling. The synthesis on return is a change log by track. No narrative is produced. Each item carries the platform count that reported it, the verification status, and its position against the existing record: confirms, extends, complicates, or contradicts. The fourth category forces action, because it identifies which published positions now need a correction or a follow-up.

Appendix B: The Case Study Record

Six sessions carry the evidence behind this framework. Each names its pool and states what the session established. Where a specific claim is challenged and instance detail is requested, that detail would sit in a CARCS or SCOPE record for the session. Such records are produced at the operator's discretion and are not assumed to exist for every session listed here.

B.1 The naming session, March 2026. Eleven platforms, two rounds, twenty returns, forty-seven naming candidates, no majority (Puglisi, 2026g). Established human origination of the name as a governance signal, the finding that recommendation convergence is weak even where analytical convergence is strong, the three-tier source-authority hierarchy, and the caper rationale as stated purpose.

B.2 The Loop That Ate the Governor, published March 2, 2026. The founding source-authority failure (Puglisi, 2026j). The session precedes publication and is dated here to the published account. Human input processed identically to platform output inside a

synthesis, with the arbiter coming to doubt his own corrections. Established Tier 0 classification at ingestion, the confirmation protocol in A.2, and the principle that a system must be architecturally designed to recognize human authority when exercised. The confirmation has since moved from a fixed phrase to a fixed tag with a described event. The published account records the February 28, 2026 amendment as "got it, human arbiter input, Tier 0," the specification then carried a shortened variant of it, and the operative form is now the one in A.2. A reader comparing this paper against the published case study will find the difference, and this is where it is recorded.

B.3 The Evocative Audit review, March 2026. Seven platforms. The synthesizer produced score-level convergence without mapping source overlap or comparing the evidence basis of dissenting platforms. Established source-overlap mapping, dissent-source verification, and the auditor evaluation path, formalized afterward as the synthesizer audit amendment (Puglisi, 2026e).

B.4 The outlier, December 2025. Nine platforms reviewing a deliberately controversial paper chosen to stress-test checkpoint governance by triggering AI resistance patterns. The expected result did not arrive. Eight of nine recommended publication and engaged favorably. One platform sustained dissent across fourteen responses and rejected the substantive thesis while validating parts of the argument.

Established that a lone dissenter is worth investigating, not discarding, and that preserved dissent from a minority position strengthens a governed output. It is also the case that most directly contradicts the intuition behind the framework, since the session was designed to produce resistance and produced near-unanimous acceptance instead, with the single objection carrying the substantive challenge. Arbitration then verified the dissenting platform's account of the other eight returns against the returns themselves and found the characterization inaccurate. Dissent preservation and dissent accuracy are separate governance functions, since a return can be worth preserving and wrong about what its peers said.

B.5 The landscape scan, August 2026. Ten dispatched platforms, twelve-track structured retrieval dispatch, and two protocol deviations logged at the time, with no retrospective correction. The count was even, against the odd-number rule then in force. And the Navigator was not held out: the platform that synthesized had been dispatched as one of the ten and was reading its own return alongside nine others. Both deviations were recorded in the session and the synthesis was marked at maximum Tier 2 scrutiny. Established the catch-layer model, confidence decoupling, the search capability gate, and retrieval index clustering as an independence axis distinct from shared foundation models. Three platform-level integrity findings and two additive findings are recorded in CARCS with their dates and model states.

B.6 The Fourth Edition review, September 2026. Thirteen dispatched returns from twelve platforms, one of which contributed two models through a single access route, with the Navigator held out. The dispatch was an open critique on the document describing the framework. A secondary auditor then held three of the dispatched returns, including its own, and reviewed the Navigator's output line by line. Under the two-property rule in

Section 3.4 that auditor was a participant, so it carried self-grading exposure on one of the three returns it held. The exposure was accepted for the sake of the auditor's independence from the author, and it is recorded and not treated as clean.

The comparison with B.5 is the evolution and it runs on two axes, not one. Scale moved from ten to thirteen. More importantly, the Navigator moved from participant to held out, which removed the self-grading exposure that put the August synthesis at maximum scrutiny. The count also moved from even to odd. B.5 is the record of a session that deviated and said so, and B.6 is the record of one that did not.

The two-models-on-one-platform case in B.6 is the reason Appendix D.1 records the access route alongside the model. Two returns reaching the operator through the same platform share that route and probably share a retrieval backend, so they count as two Tier 1 returns and should not be read as two independent positions without checking.

Established the distinction between existence verification and attribution verification, after a misattributed citation in the paper's own bibliography was caught by two returns and passed as correct by three, two of which had retrieved the record and confirmed only that it existed. Established decision-frame capture and misdirection as failure modes eight and nine, the triage boundary, and the Navigator configurations in Section 3.4.

Two findings about the session itself. A return that performed no external retrieval produced a structural criticism the arbiter adopted over the Navigator's own framing, which is why source basis governs convergence weighting and does not by itself determine overall return quality. And the dispatch ran on a critique prompt that instructed reviewers to treat validation as failure, so the verified findings stand and no inference can be drawn from the volume or severity of the criticism. The prompt has since been corrected.

Appendix C: Glossary

Term	Definition
Additive yield	Material a dispatch recovers. The counterpart of a correction. A source, argument, counterexample, or framing present in one return, absent from the rest, and verified. Before verification it is an additive-yield candidate. Arrives as a minority of one
Algorithmic narcissism	The pattern in which every platform, asked which should synthesize, nominates itself with supporting evidence. The operative reason the human pre-designates the Navigator. Adjacent to documented self-preference bias in LLM-as-a-judge work, which does not establish this specific self-nomination pattern
Assembler	Platform behavior that preserves full output depth under full-depth instruction
Asymmetry rule	A single return carrying verified external evidence takes investigation priority ahead of unsupported convergence, and governs the evidentiary assessment at irreversible stakes while Tier 0 governs the decision. At low stakes it does not displace

Term	Definition
	majority resolution
Attribution verification	Checking the author list, year, and whether a cited work contains the attributed claim. Distinct from existence verification
Best-fit discard	CAIPR's deliberate abandonment of role assignment, accepting a per-platform performance penalty in exchange for comparability
CAIPR	Cross AI Platform Review. The framework this document specifies, whose protocol is the eight core operations in Part Three. Pronounced "kay-per." Human-originated name, March 2026
Catch layer	One of the three stages at which an expected platform failure is exposed or a contribution surfaced: peer platforms, Navigator, human
CBG	Checkpoint-Based Governance. The constitutional authority layer
Citation fabrication	A citation to a work that does not exist, or that does not contain the attributed claim. Distinct from misattribution, where the work exists and the authors, year, or title are wrong. Both format correctly and appear in plausible context
Confidence inversion	The observed pattern in which platform-declared confidence runs inverse to accuracy. A working hypothesis from a single session
Corpus-blind recommendation	A reviewer proposal to build capability the ecosystem already contains, produced because the reviewer sees one artifact and not the corpus. A predictable share of any dispatch's recommendations falls in this class, and sorting for it first is cheaper than evaluating each on merit
Decision-frame capture	Failure mode 8. A synthesis that includes every return faithfully and constructs the wrong question
Dual-signed inclusion manifest	The human's expected platform list checked against the Navigator's receipt list, with any delta raising a failure flag before the synthesis is read
Evidence injection	The third disposition of dissent. The arbiter introduces evidence neither position held and the conflict dissolves, leaving nothing to resolve
Existence verification	Confirming a record exists at an identifier. Necessary and insufficient
Factics	The 2012 foundation. A fact paired with a tactic and a measurable outcome
Finding record	Dated log of what a platform instance got wrong or contributed uniquely, and which layer exposed it. Recorded at the operator's discretion in a CARCS or SCOPE artifact. Not an exclusion list
GOPEL	Governance Orchestrator Policy Enforcement Layer. Non-cognitive software that would automate CAIPR's mechanical operations. Built

Term	Definition
	as code, never deployed
HAIA	Human Artificial Intelligence Assistant. The ecosystem. Its seven frameworks are Factics, CBG, RECCLIN, CAIPR, HAIA Agents, GOPEL, and HEQ
HAIA-CARCS	Compliance Accountability Record and Case Study. The documentation protocol holding dated instance evidence
HAIA-SCOPE	Source Custody Observable Publication Evidence. The per-citation custody record
Index clustering	Convergence produced by search-capable platforms drawing on a shared retrieval index, independent of model lineage
Minority Report Extraction	Verbatim isolation of every material solitary claim before synthesis begins, without compression or evaluation
Misdirection	Failure mode 9. A synthesis that includes a return, attributes it correctly, and characterizes its position wrongly
Navigator	The Tier 2 platform that synthesizes and audits the return set. Its relationship to the dispatch is declared as a profile of independent properties: dispatch participation, authorship, context isolation, and secondary audit. Held out is one property value and not the definition. Dispatch-isolated where held out, and always inside the correlated pool
Operator memory contamination	A platform's persistent memory of the operator re-entering a dispatch as apparent external evidence
Parallel dispatch	The same substantive task and collection condition to every platform in a comparison cohort, simultaneously, with no cross-platform visibility before output is produced
Per-field convergence	Assessment of agreement field by field and not on the answer alone. Agreement on one field with divergence on nine others is not full-return convergence
Perspective-level diversity	Divergence appearing as different analytical frames and readings. Can be elicited through open prompting
Read-through requirement	The accountable human holds and examines the raw returns before synthesis, since only a party holding the material can detect what a synthesis omitted. Depth is proportional to stakes and the human remains accountable for what went unexamined. Adjacent to four-eyes review of source documents
RECCLIN	Researcher, Editor, Coder, Calculator, Liaison, Ideator, Navigator. Reasoning is the ten-field format. Dispatch is the role-assigned workflow
Reverse automation bias	The human doubting their own authority when the system fails to acknowledge it

Term	Definition
Source basis	Whether an agreeing platform retrieved externally or read only the supplied input. Text-only, external-sourced, or shared-external
Source-level diversity	Divergence appearing as different sources and evidentiary anchors. Surfaced through structured prompting
Structural divergence surface	The function by which ten structured fields give the human dimensions in which agreement on the answer can be distinguished from agreement on the sources, conflicts, and recommendation behind it
Summarizer	Platform behavior that compresses output despite explicit anti-summarization instruction
Tier impersonation	A Tier 1 or Tier 2 return acquiring apparent Tier 0 status by carrying the arbiter's own artifacts
Tier 0	Human arbiter input. Binding on decisions, escalation, correction, and publication. Not self-validating on questions of fact
Tier 1	Raw platform output. Primary AI evidence, subject to cross-validation
Tier 2	Synthesizer output. Highest scrutiny
Triage boundary	The distinction between triage that orders the reading, which is useful, and triage that replaces it, which removes the audit

Appendix D: Independence and Auditability Worksheets

Completed before dispatch, one row per platform. The purpose is to make assumptions explicit and reviewable, not to produce a score. Unknown dependencies are recorded as unknown and are never assumed independent.

The two worksheets ask different questions and were previously merged, which was an error. Independence asks whether two platforms are likely to make correlated errors. Auditability asks whether a human can inspect how an answer was reached. Two platforms can conceal their reasoning identically and could nevertheless remain independent, and two with different trace visibility can be highly correlated.

D.1 Independence Profile

Field	What to record
Platform name	The service
Access route	Direct, aggregator, API, or embedded
Model	What actually answers
Developer	Organization of record
Foundation model or lineage	Known base, or unknown

Field	What to record
Retrieval backend	Named index, unknown, or none
Search capable	Yes or no. Weight-only platforms are excluded from retrieval dispatch and may be dispatched separately for reasoning and dissent
Persistent operator memory	On, off, or unknown. If on and not disabled, flag the return for operator echo before it enters comparison

The independence statement. After the rows, write one paragraph stating why this set is believed to be diverse and on which axes it is not. A dispatch can be diverse on model lineage and clustered on retrieval index, and this paragraph is where that gets said out loud. If it cannot be written truthfully, the pool needs changing.

D.2 Auditability Profile

Field	What to record
Reasoning visibility	Raw trace, summarized trace, or none
Search log visibility	Queries and results visible, counts only, or none
Source provenance	Live links, citations without links, or unsourced
Model version visibility	Exposed version string, approximate, or none
Raw return retention	Whether the operator can retrieve the unedited return later

The auditability statement. One paragraph on what can and cannot be inspected across this pool, and what that means for how much of the independence profile is knowledge and how much is assumption. Where reasoning visibility is low across the whole set, independence is being inferred from observable metadata and outputs and not from the reasoning itself, and the independence statement above should say so.

D.3 Session Configuration

Field	What to record
Dispatch count and rationale	Against stakes, not habit
Navigator relationship	Recorded as four independent properties: dispatch participation, authorship, context isolation, and secondary audit
Collection mode	Plain, full RECCLIN, or mixed
Prompt type	Structured, open, or mixed
Best-fit roles discarded	The role each platform would have received under Dispatch, recorded so the trade is visible
Prior findings	Reference to any session record. Consulted for reading closeness, not admission. Consider withholding this during the first pass to avoid priming, and reviewing it before the second

D.4 The Platform Roster, September 2026

The platforms in the operator's rotation, from which a dispatch is drawn, as verified for this edition. This is a rotation and not a dispatch: any session draws a subset, and the count in a session is the number of Tier 1 returns under Operation 1. The stamp above is the date of record and not the date of publication, because a roster ages the way a finding does. Search capability, retrieval backend, persistent memory, and reasoning visibility change with vendor releases and are recorded per dispatch in D.1 and D.2.

Platform	Developer	Models exposed	Lineage and route notes
Claude	Anthropic	One	
ChatGPT	OpenAI	One	
Gemini	Google	One	NotebookLM is a Gemini extension used for derivative infographic, video, and audio output. It is not a dispatch model and is not counted
Perplexity	Perplexity and Nvidia	Two: Sonar and Nemotron 3 Super	Two models through one access route. Two returns from this platform are two Tier 1 returns sharing a route
Grok	xAI	One	
Mistral	Mistral AI	One, as Le Chat	
DeepSeek	DeepSeek	One	
Kimi	Moonshot AI	One	
MiniMax	MiniMax	One	Added for architectural independence within the rotation
Meta AI	Meta	One	
Copilot	Microsoft	One	Runs on OpenAI and Anthropic builds, so its lineage is shared with two other platforms in the rotation. Retained for enterprise relevance
Qwen	Alibaba Cloud	One	Runs directly at qwen.ai. Distinct from the Singaporean Qwen-SEA-LION-v4 below
PublicAI	Swiss AI Initiative and AI Singapore	Two: Apertus and Qwen-SEA-LION-v4	Two models through one access route. Qwen-SEA-LION-v4 is built on a Qwen base and is treated as a sovereign Singaporean model, counted separately from Qwen

Thirteen platforms, fifteen models. Eleven platforms expose one model each and two expose two, which is why the model count exceeds the platform count. Access route and lineage are the two fields most likely to cluster returns that look independent, and both are visible here before any dispatch profile is written.

Appendix E: Terminology Crosswalk

A framework built from practice arrives at terms the practitioner needed, not terms the field had settled. Some duplicate established vocabulary, and some have no equivalent identified in the searches represented here. This appendix records which, across four levels, because a binary would overstate.

Adopted directly. Regression testing, defined as testing performed after modification to a system or its operational environment to determine whether previously unaffected portions have developed failures (ISO/IEC/IEEE 29119-1:2022, 3.64), applied here to governed-document loss detection. Provenance and chain of custody, from the AI Act Article 12 and NIST AI RMF traditions. Auditability and traceability, per the NIST AI RMF Playbook on mechanisms that support auditability through traceability of development, sourcing, and logging. AI BOM, per SPDX 3.0 and CycloneDX ML-BOM. Meaningful human control, from the scholarly literature on whether humans hold sufficient understanding, authority, intervention capacity, and responsibility (Santoni de Sio & van den Hoven, 2018; Cavalcante Siebert et al., 2023).

Established concept, applied at a different layer. Dissent preservation maps to data perspectivism and annotator-level disagreement preservation, from Basile et al. 2021 and Davani et al. 2022. Their object is annotator labels on training data. This framework's object is model outputs at inference under human governance. Effective independence is a gloss on Kohli's effective sample size and effective independent votes and not Kohli's own term. Index clustering sits near retrieval overlap and source overlap, where the field's vocabulary is fragmented enough that the narrower term is worth keeping. Source basis sits near the closed-book versus retrieval-augmented distinction. Catch layers sit near defense in depth. The read-through requirement sits near four-eyes review of source documents.

Adjacent literature, not equivalent. Algorithmic narcissism is adjacent to self-preference bias in LLM-as-a-judge research (Wataoka et al., 2024), which documents models favoring their own outputs and does not establish the specific self-nomination pattern recorded here. Auditable synthesis is adjacent to the evidentiary-adequacy criterion published in July 2026 for when runtime records can support legally operative findings, which is more specific than auditability and not synonymous with it.

No equivalent found in this search. Additive yield and additive-yield candidate, the asymmetry rule, decision-frame capture and misdirection as failure modes eight and nine, tier impersonation, and the triage boundary. Evidence injection as the third disposition of dissent, best-fit discard, corpus-blind recommendation, structural divergence surface, Minority Report Extraction, and the dual-signed inclusion manifest. Confidence inversion, which is a one-session working hypothesis, not a finding, and Assembler and Summarizer as named behavioral categories.

That last list is scoped to the searches performed. It records what was not found and does not claim that nothing exists. An earlier version of this appendix placed five further terms

in that category and moved them out after review surfaced established or adjacent work, which is the expected rate of correction.

Naming hazards. Do not use "shared task," which is a term of art in NLP meaning a benchmark competition. Do not lead with bare "HAIA" in titles, given a one-letter collision with a separate 2025 governance framework. No collision was identified for the compound forms in the searches represented here. Retired terms: unbiased Navigator, anti-alignment layer, catch record, and Platform Integrity Register.

Appendix F: The Record

The dated lineage behind Section 1.1. Each entry names its source and states what it established. The count rule is the one element that changed materially across the record, and the change is noted where it occurred.

Date	What	Source	What it established
2023	ChatGPT produced usable answers and unreliable sources. Factics does not accept an answer without citable evidence, so source validation was assigned to Perplexity and errors were routed back to the originating platform	Puglisi, 2026h, Section 4	The first RECCLIN Dispatch in operation, producing the published work of that year. Multi-AI began as a governance response to a single-platform failure, not as a designed architecture
February 1, 2024	Article naming the fabrication problem, proposing a second platform for source validation, and stating that neither the drafting system nor the search system is an authority	Puglisi, 2024	First public disclosure of the two-platform method, with human authority over both stated from the start
September 2025	Role-based five-platform workflow documented on LinkedIn, then a public draft: a single complex prompt to five platforms at once, raw returns arbitrated by hand, three-of-five convergence as a preliminary finding, minority dissent preserved through the Navigator, a standing watch for consensus produced by shared training bias	Puglisi, 2025b	First documented parallel dispatch. The Assembler and Summarizer clustering in Section 4.2 originates here. The count rule at this stage: three-of-five convergence is a preliminary finding
November 28, 2025	Seven-platform review in which three anchor synthesizers each	Puglisi, 2025c	Algorithmic narcissism first documented

Date	What	Source	What it established
	nominated itself as final synthesizer		
December 2025	Nine-platform outlier case. Eight of nine recommended publication, one sustained dissent across fourteen responses and was found inaccurate about its peers	Puglisi, 2025d	A lone dissenter is worth investigating. Dissent preservation and dissent accuracy are separate functions
December 6, 2025	Enterprise framing describing orchestration platforms that turn individual AI instruments into coordinated councils, with a human Navigator holding the gavel	Puglisi, 2025e	The council vocabulary, two months before the product that adopted it
February 5, 2026	Perplexity releases Model Council: three frontier models in parallel, a fourth as chair	Perplexity, 2026; Puglisi, 2026m	The industry productizes the dispatch without the checkpoint
March 2026	Naming session across eleven platforms. Cross AI Platform Review assigned to a method by then six months in daily use. First public edition, v1.1	Puglisi, 2026c, 2026g	The name, the three-tier source-authority hierarchy, and the caper rationale
April 2026	Revised edition in the HAIA repository	Puglisi, 2026c	
August 2026	Ten-platform landscape scan with a participant Navigator and an even count, both logged as deviations	Appendix B.5	The catch-layer model, confidence decoupling, the search capability gate, index clustering
September 2026	Fourth Edition review, thirteen returns from twelve platforms, Navigator held out	Appendix B.6	The invariant core, the configuration variables, and the count rule as it now stands: diagnostic at governance stakes, with unanimous convergence a signal to verify outside the pool. Part Two gives the correlated-error research that forced the change

Sources

Basile, V., Cabitza, F., Campagner, A., & Fell, M. (2021). *Toward a perspectivist turn in ground truthing for predictive computing*. arXiv:2109.04270. https://arxiv.org/abs/2109.04270. Published as Cabitza, F., Campagner, A., & Basile, V. (2023), *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6), 6860 to 6868.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., & Lagendijk, R. L. (2023). Meaningful human control: Actionable properties for AI system development. *AI and Ethics*, 3, 241 to 255. https://doi.org/10.1007/s43681-022-00167-3

Davani, A. M., Díaz, M., & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10, 92 to 110.

Engin, Z. (2025). *Human-AI governance (HAIG): A trust-utility approach*. arXiv:2505.01651. Cited in Appendix E for the naming collision only.

European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Articles 12, 14, 50.

European Union. (2026). Regulation (EU) 2026/1744, Digital Omnibus on AI. Official Journal, 24 July 2026; in force 27 July 2026.

International Organization for Standardization. (2023). *ISO/IEC 42001: Information technology, artificial intelligence, management system*.

International Organization for Standardization, International Electrotechnical Commission, & Institute of Electrical and Electronics Engineers. (2022). *ISO/IEC/IEEE 29119-1:2022: Software and systems engineering, software testing, part 1, general concepts*. ISO. Section 3.64.

Janssen, J. (2026). *From runtime records to legal findings: An evidentiary-adequacy criterion for agentic AI oversight*. arXiv:2607.00941.

Jiang, L., Chai, Y., Li, M., Liu, M., Fok, R., Dziri, N., Tsvetkov, Y., Sap, M., Albalak, A., & Choi, Y. (2025). *Artificial hivemind: The open-ended homogeneity of language models (and beyond)*. arXiv:2510.22954. NeurIPS 2025, Best Paper, Datasets and Benchmarks track.

Kim, E. M., Garg, A., Peng, K., & Garg, N. (2025). Correlated errors in large language models. *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267, 30038 to 30066. Preprint at arXiv:2506.07962.

Kohli, G. (2026). *Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels*. arXiv:2605.29800. https://arxiv.org/abs/2605.29800

Linux Foundation. (2024). *Implementing AI bill of materials (AI BOM) with SPDX 3.0*. See also OWASP CycloneDX ML-BOM.

National Institute of Standards and Technology. (2023). *AI Risk Management Framework 1.0* and AI RMF Playbook.

Perplexity. (2026, February 5). *Introducing Model Council*. https://www.perplexity.ai/hub/blog/introducing-model-council

Puglisi, B. C. (2024, February 1). *Factics make us more intelligent*. https://basilpuglisi.com/factics-make-us-more-intelligent/

Puglisi, B. C. (2025a). *Governing AI: When capability exceeds control*. ISBN 9798349677687. Text revised in place in 2026 under the original listing, not issued as a separate edition.

Puglisi, B. C. (2025b, September 26). *The HAIA-RECCLIN model: A comprehensive framework for human-AI collaboration* (public draft). https://basilpuglisi.com/the-haia-recclin-model-a-comprehensive-framework-for-human-ai-collaboration-draft/

Puglisi, B. C. (2025c, November 28). *Multi-AI governance: How 7 platforms exposed the bias no single AI (LLM) could see*. https://basilpuglisi.com/multi-ai-governance-how-7-platforms-exposed-the-bias-no-single-ai-llm-could-see/

Puglisi, B. C. (2025d, December). *HAIA-RECCLIN case study: The Kimi outlier, one adversarial voice among nine*. https://github.com/basilpuglisi/HAIA/blob/main/HAIA-RECCLIN-Kimi-Case-Study-REVISED.docx

Puglisi, B. C. (2025e, December 6). *The multi-AI operating system: Five amplification lines, twenty-eight gates, one central rule*. https://basilpuglisi.com/the-multi-ai-operating-system/

Puglisi, B. C. (2026a). *Empire of evidence: Testing Karen Hao's claims against the governance infrastructure*. https://basilpuglisi.com/empire-of-evidence-testing-karen-hao-claims-governance-infrastructure/

Puglisi, B. C. (2026b, March 10). *Checkpoint-Based Governance (CBG): A constitutional framework for human-AI collaboration* (v5.0). https://basilpuglisi.com/checkpoint-based-governance/

Puglisi, B. C. (2026c). *Cross AI platform review beyond the RECCLIN dispatch* (March 2026). https://basilpuglisi.com/haia-caipr-v1/. The first public edition of this document, revised April 2026 in the HAIA repository. Not independent corroboration of the present claims.

Puglisi, B. C. (2026d, March 8). *GOPEL v1.5: The non-cognitive governance layer that automates without thinking*. https://basilpuglisi.com/gopel-v1-5-the-non-cognitive-governance-layer-that-automates-without-thinking/

Puglisi, B. C. (2026e, March). *HAIA-CAIPR synthesizer audit amendment* (v1.0). https://github.com/basilpuglisi/HAIA/blob/main/HAIA_CAIPR_Synthesizer_Audit_Amendment_v1_0.md

Puglisi, B. C. (2026f, April 23). *CARCS: Compliance accountability record and case study* (v1.4, May 2026 PDF). https://basilpuglisi.com/haia-carcs-compliance-accountability-record-case-study/

Puglisi, B. C. (2026g). *HAIA-RECCLIN case study 006: The discovery of CAIPR* (v7). https://github.com/basilpuglisi/HAIA

Puglisi, B. C. (2026h). *HAIA-RECCLIN: Reasoning and dispatch, the operational methodology for human AI governance* (Third Edition, March 2026). https://basilpuglisi.com/haia-recclin/

Puglisi, B. C. (2026i). *HAIA-SCOPE: Source custody observable publication evidence* (v1.1). https://basilpuglisi.com/scope-source-custody-observable-publication-evidence/

Puglisi, B. C. (2026j, March 2). *The loop that ate the governor* (v7). https://basilpuglisi.com/the-loop-that-ate-the-governor/

Puglisi, B. C. (2026k). *What we learned: The evolution of HAIA-RECCLIN and multi-AI governance in practice* (v2). https://github.com/basilpuglisi/HAIA

Puglisi, B. C. (2026l). *Why you cannot program or prompt governance into AI*. https://basilpuglisi.com/program-prompt-governance-ai/

Puglisi, B. C. (2026m, May 28). *The governance layer Perplexity's Model Council needs*. https://basilpuglisi.com/perplexity-model-council-governance/

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5, 15. https://doi.org/10.3389/frobt.2018.00015

Shu, Y. (2026). *Blind to the pivotal vote: Aggregate independence metrics miss where verification actually helps*. arXiv:2608.06940.

Spiro, T. (2026). *The oracle's fingerprint: Correlated AI forecasting errors and the limits of bias transmission*. arXiv:2605.00844. https://arxiv.org/abs/2605.00844

United States Copyright Office. (2025, January 29). *Copyright and artificial intelligence, part 2: Copyrightability*. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf. Executive Summary at roman iii; prompts conclusion at printed page 18; selection finding at printed page 21; selection, arrangement, and modification at printed page 24.

Verga, P., Hofstätter, S., Althammer, S., Su, Y., Piktus, A., Arkhangorodsky, A., Xu, M., White, N., & Lewis, P. (2024). *Replacing judges with juries: Evaluating LLM generations with a panel of diverse models*. arXiv:2404.18796. https://arxiv.org/abs/2404.18796

Wataoka, K., Takahashi, T., & Ri, R. (2024). *Self-preference bias in LLM-as-a-judge*. arXiv:2410.21819. https://arxiv.org/abs/2410.21819

Disclaimer

The author is not a lawyer, and this paper does not provide legal advice. This is thought research and governance analysis based on public sources, cited materials, and human-AI review. It is intended to help anyone working across multiple AI platforms think more clearly about verification, dissent, and documentation practices. Readers should not rely on this paper as a legal opinion, compliance determination, or substitute for qualified counsel. Anyone facing a legal, regulatory, or contractual question should consult qualified counsel or another professional adviser before acting.

The regulatory material in Section 2.6 and the copyright material in Section 4.3 are current as of September 2026 and should be re-verified. Regulation (EU) 2026/1744 postponed application dates that had not yet arrived, and the United States Copyright Office grounds its conclusions in current generally available technology and expressly leaves questions open. The author is an independent practitioner and author who may profit in other ways from research and content like this.

Disclosure

The author created the frameworks described in this paper and has a direct interest in their adoption. Every AI platform named in the roster in Appendix D.4 is a commercial service the author subscribes to or uses under a free tier, and the author has no financial relationship with any platform provider. Perplexity, whose Model Council product is discussed in Section 1.1, is one of the platforms in the operator's rotation. All specifications referenced here are published for independent review at github.com/basilpuglisi/HAIA.

AI Use Disclosure

This paper was produced through the framework it describes. Multiple AI platforms contributed to drafting, review, and source verification at every stage under parallel dispatch, and their returns were compared, audited, and arbitrated by the author. Basil C. Puglisi, MPA, served as Tier 0 human governor with final editorial, analytical, and publication authority over all content. Appendix B records the sessions behind the framework, including the September 2026 review of this document. The AI platforms functioned as governed tools within a documented process and are not co-authors.

#AIassisted using HAIA Ecosystem