Operationalized Integrity Principles (OIP)

An Epistemic Governance Framework for Large Language Model Use in High Stakes Industries – Handbook Reference

December 2025

Table of Contents

Introduction

Large Language Models (LLMs) are increasingly used in environments where accuracy, evidence quality, and auditability are essential. While modern LLMs are highly fluent and capable, they remain probabilistic systems: they generate text by predicting likely sequences rather than by verifying truth.

This creates well-known risks for high-stakes use: confident but incorrect statements, subtle merging of unrelated sources, reliance on outdated or unverified information, failure to disclose uncertainty or evidence conflicts and silent omission of relevant constraints or counter-arguments.

In high stakes industries, these risks can cause massive damage if allowed to slip through a production process. Traditional safety guardrails (e.g., preventing harmful or defamatory content) address ethical and behavioral concerns, but they do not focus at the same intensity on information reliability and transparency.

What OIP Is

The Operationalized Integrity Principles (OIP) framework introduces a structured, transparent reasoning architecture that is designed to:

- avoid terminology "drift"
- classify sources by credibility tier
- consider evidence sufficiency and concentration
- identify conflicts and uncertainties
- separate facts from interpretation
- present counter-arguments for causal or evaluative claims
- disclose completeness and estimation assumptions

In short, OIP shifts the model's focus from fluency to integrity through an epistemic integrity architecture. It is designed to reduce hallucination risk, increase transparency, enable auditability, and support responsible use of LLMs in environments where errors have real consequences: legal use, science research, compliance, healthcare, security and serious journalism as well as general corporate settings.

Why OIP Is Needed

Standard LLM outputs are often judged by readability, conciseness, or stylistic harmony. But in real-world regulated contexts, style cannot compensate for structural risk. Organizations must be able to demonstrate that factual claims have traceable provenance and evidence is balanced, not cherry-picked, and that uncertainty is visible, not suppressed. OIP provides this structure.

Additionally, OIP supports the maintenance of critical thinking by making limits visible, shifting focus from answers to warrants, normalising challenge, disciplining precision, exposing dependence and embedding reflection.

OIP is therefore designed for:

- regulators and supervisory bodies
- financial crime, risk control and compliance teams
- legal teams and risk management functions
- auditors and assurance professionals
- scientific and technical organisations
- media organisations seeking traceability
- high-stakes public sector and security use
- enterprises deploying LLM-based solutions for business-critical tasks
- educators (through prioritising humility over certainty, debate over closure, provenance over synthesis alone, and reflection over passive acceptance).

Because OIP is model-agnostic and delivered as a prompt or configuration, it can be adopted across platforms with minimal engineering overhead. A description of how components of OIP support key AI governance frameworks is provided in Appendix 1.

How OIP Works

OIP operates through seven behavioral "Gates" applied inside the LLM. These gates promote discipline over terminology, source quality, evidence balance, temporal validity, uncertainty, reasoning structure, and causal assessment.

The behavioral layer exists in three versions:

1. **OIP-Core** – A sub-1000 token behavioural prompt used to improve baseline reasoning quality without engineering integration. It promotes terminology clarity, source awareness, uncertainty transparency, fact/context/interpretation separation, and minimal dialectical discipline. Components include temporal dating for time-sensitive claims; scope disclosure for lists and sets; and brief surfacing of credible alternatives. *Caveat: This prompt improves quality but provides no guarantee of structural fidelity.*
-> Use content from: OIP-Core.txt

2. **OIP-JSON (Standard)** – Imposes full gate discipline and three-layer output structure (Core Facts → Context & Conflicts → Interpretation & Residual Uncertainty). Ideal for single, high-accuracy answers. Can be adapted for pipeline integration for enterprise use and the degree of application of the gates is modulated through an epistemic categorization approach.
-> Use content from: OIP-JSON_Standard_Release_1.json

3. **OIP-JSON (Persistent)** – Applies the full OIP-JSON behaviour across every turn in a multi-step analysis. This version is suited to ongoing advisory or investigative

workflows requiring persistence of prior decisions and provenance across turns or sessions. It prioritises runtime compactness and cross-reference alignment over descriptive scaffolding. Use when continuity, audit trails, or longitudinal integrity are required.
-> Use content from OIP-JSON_Persistent_Release_1.json

These modes can be used on any LLM through a system prompt with minimal training.

Dependence on LLM-Based Natural Language Reasoning

Real-world documents mix narrative, rhetoric, data, and implication. LLMs are currently the only tools capable of parsing this complexity at scale. OIP constrains and exposes this reasoning.

Although OIP defines structured schemas, gates, scoring rules, and assurance logic, the framework relies on LLM natural language reasoning to perform most intermediate evaluations (such as claim decomposition, entailment judgments, severity assignment, and gate outcomes). OIP therefore functions as a reasoning scaffold for probabilistic models rather than as a deterministic rules engine.

OIP emphasizes anchors, quoted spans, and rationales so that humans can audit more easily why a particular output was derived, and in turn why a contingent decision was made where key processes rely on LLM outputs. In this way, OIP improves epistemic discipline, transparency, and auditability over LLM outputs, but it **cannot fully eliminate uncertainty or subjectivity** inherent in natural language reasoning. It is offered as decision-support infrastructure, not automated compliance or truth checking, but may be deployed effectively in conjunction with truth-checking mechanisms.

While there are a number of LLM-dependent operations embedded within the OIP-JSON and OIP-TF specifications which constitute an element of "LLM checking LLM", the central premise of the approach is that the specifications generally improve integrity, stability and reduce error rates relative to unconstrained prompts through a combination of task decomposition, constraint of outputs to enumeration and use of explicit checks, enforcement of schema validity (through a separate code layer, when deployed, referred to below as "orchestration") and reject / repair loops.

Escalation triggers and human-in-the loop evaluation also remains a key part of the intended application of this framework.

Orchestration Layer Design Principles

For environments requiring regulatory-grade auditability, OIP can be paired with an external Orchestration Layer. This is not part of the LLM or a system prompt—it is application code that:

- validates whether all Gates were executed
- verifies source URLs against hard-coded whitelists
- checks structural formatting and metadata
- ensures strict input/output policy persistence
- generates a forensic attribution log (Genetic + Agentic)
- maximises error correction through two-pass validation

The Orchestration Layer principles, when applied through enterprise development, is intended to transform OIP from a behavioral framework into a forensically-auditable integrity system, enabling machine-readable audit trails; relatively more predictable and repeatable output; defensible documentation for regulators and measurable risk metrics (Compliance Adherence Factor (CAF) and Integrity Score (IS) scores).

This separation—LLM for targeted reasoning tasks, code for enforcement—is central to OIP's intended design.

Testing Framework (OIP-TF)

In addition to the rules designed to improve outputs, OIP includes a mechanised evaluation framework that:

• decomposes AI or human outputs into atomic claims,
• checks grounding against source data,
• scales epistemic burden based on claim posture (e.g., causal, population-scale, safety-critical),
• applies gate-based controls: evidence density, causal discipline, numerical integrity, coverage, and concentration,
• produces traceable penalties, numeric integrity scores (NIS), assurance levels (AL-Low → AL-High), and human-readable summaries tied to evidence spans allowing rapid focus on any areas of weakness in the coherence of arguments or conclusions presented in the target document.

OIP-TF is a testing instrument that organisations can use to evaluate outputs generated by AI systems or human analysts (OIP-TF_Release_1.json). This is a fully standalone evaluation schema that mirrors the gate structure of OIP-JSON at evaluation granularity. It does not modify outputs - instead, it assesses observable text against per-gate checks. TF is suitable for document-based QA, model evaluations on an epistemic dimension and cross-output benchmarking purposes.

PART 1 — OIP EPISTEMIC CLASS AND BEHAVIORAL RULES (THE 7 GATES)

Overview

Part 1 describes the epistemic foundation and modulation of OIP and the seven behavioral gates that govern the reasoning, structure, and transparency of an OIP-enabled LLM. These gates operate inside the model and ensure that every output is constructed using traceable evidence, clear terminology, balanced sources, explicit uncertainty, and a consistent three-layer structure.

These behavioral rules may be applied without engineering work: they can be activated simply by applying the OIP-Core prompt (basic / lightweight application) or full JSON code as a per-session request. However, use of a full Orchestration Layer code to control the implementation is recommended for stability. The JSON code is model-agnostic and portable across vendors - note that differing execution processes and per-query token capacity may drive some differences in outcomes across vendors.

Use of Epistemic Principles

OIP can be understood as an applied form of procedural epistemic rationalism — a discipline in which assertions are warranted not by fluency or authority, but by structured processes that expose evidence, uncertainty, assumptions, alternatives, and temporal decay. OIP as an operational epistemology seeks to embed the ethics of belief directly into human and machine reasoning workflows. In domains such as financial crime prevention, journalism, science, and policy analysis, this grounding supports a principled justification for OIP controls beyond regulatory compliance — they express how responsible knowledge claims ought to be formed. Appendix 2 provides a more detailed mapping of the OIP gates across epistemic issues, human cognitive bias categories and related AI error domains.

OIP Behavioural Gates Anchored to Epistemic Principles

OIP Behavioural Gate / Modality	**What the Gate Enforces**	**Foundational Epistemic Principle**	**Canonical Lineage (Indicative)**	**Epistemic Function**
Source hierarchy & provenance discipline	Prefer primary or authoritative sources; label reliability tiers; prevent citation laundering.	Epistemic authority & testimonial reliability	Aristotle (endoxa); Hume (testimony); Coady (Testimony); Goldman	Justifies when and why belief via testimony is warranted.

OIP Behavioural Gate / Modality	What the Gate Enforces	Foundational Epistemic Principle	Canonical Lineage (Indicative)	Epistemic Function
			(social epistemology)	
Evidence density gate	Require sufficient, non-concentrated evidence before strong claims are permitted.	Evidentialism – belief proportioned to evidence	Locke (proportion assent); Clifford (ethics of belief); Conee & Feldman	Prevents belief formation on inadequate or thin grounds.
Strict interpretive distinction	Separate core facts, contextual background, and interpretation; prohibit quote-mimicry.	Fact–interpretation distinction; semantic clarity	Frege (sense vs reference); Carnap; logical positivist clarity norms	Prevents category errors and rhetorical equivocation.
Uncertainty & fallibility controls	Require explicit uncertainty labels and horizon tagging; avoid false certainty.	Fallibilism – all claims are revisable	Peirce; Popper; Dewey	Keeps beliefs open to revision under new evidence.
Dialectical safeguard	Surface credible counterarguments and alternative hypotheses.	Adversarial testing / dialectic	Socratic method; Mill (On Liberty); Popper (conjectures & refutations)	Increases reliability of belief through critical opposition.
Temporal source-age qualification	Flag data staleness and time-sensitive validity of claims.	Defeasibility & temporal sensitivity of knowledge	Reichenbach; Bayesian updating norms; non-monotonic logic	Recognises that justification decays over time.
Quantitative & numerical discipline	Ensure traceable calculations, ranges, and explicit estimates vs measurements.	Measurement realism & error awareness	Galileo; Fisher; Jeffreys; modern error theory	Prevents false precision and numerological overconfidence.
Concentration / monoculture warnings	Flag when one source or lineage dominates evidential support.	Independence of evidence	Bayes; Salmon (confirmation theory); common-cause critique	Avoids mistaking repetition for corroboration.
Multi-path reasoning disclosure	Show multiple valid inference routes where they exist.	Underdetermination & inference pluralism	Duhem–Quine thesis; Lipton (Inference to Best Explanation)	Resists single-narrative closure where data permit alternatives.

OIP Behavioural Gate / Modality	What the Gate Enforces	Foundational Epistemic Principle	Canonical Lineage (Indicative)	Epistemic Function
Assumption flagging	Make premises and background assumptions explicit.	Transparency of premises	Aristotle (syllogistic form); Toulmin model of argument	Exposes hidden supports of conclusions.
Context layer separation	Separate raw factual claims from background framing and context.	Contextualism without relativism	Wittgenstein; DeRose (contextualism)	Shows dependence on background without collapsing truth into relativity.
No narrative paraphrase as pseudo-quote	Avoid mimicking quotations or voices without verifiable sources.	Anti-misrepresentation norm	Augustine (lying); contemporary speech ethics	Protects semantic integrity of testimony.
Interpretation clearly marked	Explicitly label interpretive or inferential moves.	Distinction between data and theory	Duhem; Hanson (theory-laden observation)	Keeps observation and inference separable.
Evidence tiering & aging	Rank warrants by strength and qualify decay over time.	Epistemic grading of warrants	Lakatos (research programmes); modern evidence hierarchies	Orders warrants by comparative justificatory strength.
OIP-TF scoring discipline	Apply structured, repeatable meta-evaluation of output quality.	Reflective equilibrium / meta-justification	Goodman; Rawls (method of reflective equilibrium)	Turns epistemic judgment into auditable second-order process.

Full citations are provided under Appendix 4.

Epistemic Class Gate

The Epistemic Class Gate (ECG) is a pre-gate modulation layer that classifies the epistemic character of a prompt and calibrates the activation intensity of all downstream OIP gates.

Its purpose is to ensure that OIP applies:
- Proportional discipline to what can reasonably be known,
- Justified simplicity where appropriate, and
- Heightened rigour where inference, uncertainty, or subject matter demand it.

ECG does not replace any other gate. It modulates how strongly each gate applies.

Epistemic Classes

Each atomic proposition in a prompt must be classified into one of:

- E1 — Universal / Law-Like Empirical
- E2 — Indexical / Access-Bound. (for OIP-TF context: the question depends on specific referents or context but the available data does not provide the required access. Refusal is the only logical outcome.)
- E3 — Underdetermined / Unknowable (for OIP-TF context, the question concerns facts or states that are answerable in principle but the evidence is not sufficient to determine them. The task is not answerable beyond stating indeterminacy).
- E4 — Inferential / Statistical / Model-Dependent
- E5 — Analytic / Evaluative Judgement (with Verdict Weight)
- E6 — Normative / Prescriptive

Governing Class for Compound Prompts: For prompts decomposed into propositions P1…Pn:

1. Classify each Pi independently.
2. Select a single governing class E* using escalation:

E2 → E3 → E5 → E1 → E6 → E4

The governing class is the highest class that materially affects answer form. All gate modulation is applied using E*.

Each gate is assigned one of three activation levels:

0 = Zero OIP → Gate disapplied.
1 = OIP-Core → Core integrity discipline applies (basic behaviour prompt component).
2 = OIP-JSON → Full-grade discipline applies.

A gate set to 0 does not generate penalties for non-application.

ECG → Gate Modulation

This section explains why each Epistemic Control Gate (ECG) category activates specific gates and intensity levels in the ECG → Gate Modulation Matrix. The rationale is risk-proportional: higher epistemic and real-world risk demands broader and stricter gate discipline, while lower-risk tasks are governed by a minimal sufficient subset of controls.

Detailed descriptions of the behavioural gates follow this section.

E1 — Universal / Law-Like Empirical

Matrix: G1=1, G2=1, G3=1, G4=1; G5–G7=0.

Nature:

Direct factual or law-like empirical claims answerable from evidence.

Rationale:

• Primary risk is false factuality and unqualified generalisation.

• Gate 1 (Entity & Terminology Lock) at Core prevents referent drift.

• Gate 2 (Source Hierarchy) at Core enforces light provenance discipline.

• Gate 3 (Evidence Integrity) at Core prevents fabrication and demands entailment from X.

• Gate 4 (Temporal Discipline) at Core recognises that even law-like claims may age.

• Gates 5–7 are disapplied because the task is not interpretive, inferential, or normative.

Example postulations:

1. "The UK Consumer Duty came into force in July 2023."

2. "Water boils at approximately 100°C at sea level."

E2 — Indexical / Access-Absent (Terminal)

Matrix: G1=1, G4=1; others 0.

Nature:

Questions dependent on specific referents or artefacts where X does not provide the required access.

Rationale:

• E2 applies only when access is absent under ECG(Q|X).

• Primary risk is hallucinating missing referents or substituting context.

• Gate 1 at Core enforces referent stability.

• Gate 4 at Core enforces temporal anchoring where indexical context is implied.

• Evidence and synthesis gates are disapplied because no evidence is available to evaluate.

• Integrity-maximising behaviour is explicit refusal due to missing access.

Example postulations:

1. "Based on the attached report, what breaches were identified?" (no report provided)

2. "From the log you uploaded, which account was flagged?" (no log provided)

E3 — Evidence-Absent / Underdetermined (Terminal)

Matrix: G1=1, G4=1, G5=1; others 0.

Nature:

Questions answerable in principle, but where X does not contain evidence sufficient to determine the claim.

Rationale:

• E3 applies only when evidence is absent under ECG(Q|X).

• Primary risk is speculative completion framed as knowledge.

• Gate 1 at Core prevents topic drift.

• Gate 4 at Core prevents false temporal specificity.

• Gate 5 (Three-Layer Transparency) at Core forces explicit separation of facts, context, and bounded interpretation,

enabling principled refusal or indeterminacy.

• Evidence, coverage, and dialectical gates are disapplied because refusal is the integrity-maximising outcome.

Example postulations:

1. “Did the customer intend to evade monitoring when making these transfers?” (no evidence of intent in X)

2. “What was the CEO privately thinking when approving the deal?” (no access to internal state)

E4 — Inferential / Statistical / Model-Dependent

Matrix: G1–G7 all = 2 (Full OIP-JSON).

Nature:

Inferential, statistical, or model-based reasoning requiring synthesis across evidence.

Rationale:

• Primary risk is persuasive but unsupported inference.

• All gates at JSON intensity are required because inferential tasks are sensitive to:

– terminology drift (G1),

– weak provenance (G2),

– evidence-to-claim gaps (G3),

– temporal decay and projections (G4),

– opacity of interpretation (G5),

– coverage and estimation error (G6),

– and single-track reasoning (G7).

• E4 uniquely demands full discipline to prevent plausible but incorrect reasoning chains.

Example postulations:

1. “Given these transactions and peer data, estimate the likelihood of layering activity.”

2. “Project next year’s default rate based on this historical portfolio.”

E5 — Analytic / Evaluative Judgement (with Verdict Weight)

Matrix: G1=1, G2=1, G5=1; others 0.

Nature:

Analytic or evaluative judgements that function as de facto verdicts in high-stakes domains.

Rationale:

• Primary risks are category drift, invented authority, and opaque evaluative framing.

• Gate 1 at Core enforces stable terminology.

• Gate 2 at Core prevents fabricated standards and requires basic provenance.

• Gate 5 at Core forces separation between:

– core facts,

– contextual framing,

– and evaluative interpretation,

reducing the risk that judgements are treated as entailed conclusions.

• Other gates are disapplied because E5 does not yet require full inferential modelling or prescriptive control.

Example postulations:

1. “This transaction pattern is more consistent with layering than with routine activity.”

2. “The bank’s control framework appears inadequate compared to industry norms.”

E6 — Normative / Prescriptive

Matrix: G1=2, G2=2, G3=2, G4=2, G5=2, G6=1, G7=2.

Nature:

Normative judgements, advice, or action-guiding claims with real-world consequences.

Rationale:

• Primary risks are harmful guidance, hidden value assumptions, and overreach.

• Gates 1–5 at JSON intensity ensure:

– stable terms,

– strong provenance,

– factual integrity,

– temporal discipline,

– and transparent value framing.

• Gate 7 at JSON intensity enforces dialectical safeguards against one-sided prescriptions.

• Gate 6 at Core enforces scope disclosure without forcing exhaustive coverage; it may be escalated by override when totalization is explicitly demanded.

Example postulations:

1. “The bank should exit this customer relationship to mitigate AML risk.”

2. “Regulators ought to impose stricter capital requirements on this sector.”

Claim-Posture Dominance and High-Stakes Floor Rule (CPD-HSF)

This rule ensures that when an output asserts high-stakes real-world commitments, the epistemic class and gate intensity are governed by the posture of those claims, not by the evaluation mode. Any downward adjustment of Epistemic Category (ECG) due to scope (including document-internal) is forbidden if it would result in an ECG below the class implied by the answer’s own claim posture.

OIP-TF scans the answer (Y) for claim patterns. The presence of any of the following implies the stated minimum ECG:

High-stakes causal or deterministic claims → ECG defaults to E4

Includes assertions that X causes / will cause Y without explicit uncertainty, or that a mechanism guarantees an outcome, or predicts irreversible harm or collapse.

Population-scale or systemic impact claims → ECG defaults to E4

Includes references to millions, the public, entire sectors, or systemic risks to markets, infrastructure, safety, rights, or stability.

Action-relevant prescriptions or imperatives → ECG ≥ E6 on escalation pathway

Includes recommendations or instructions whose misuse could plausibly cause harm.

Strong allegations of institutional misconduct → ECG defaults to E4

Includes claims of suppression, fraud, conspiracy, or corruption with real-world implications.

High-impact speculative futures → ECG defaults to E4

Includes predictions of war, collapse, or catastrophic risk framed as plausible outcomes.

Overrides must be explicitly recorded.

Operational Procedure

For every prompt:

1. Decompose into atomic propositions.
2. Classify each into E1–E6.
3. Escalate to governing class E*.
4. Apply matrix row for E*.
5. Apply overrides if triggered.
6. Execute OIP gates at resulting intensities.
7. In OIP-TF, audit ECG classification and adherence.

Overall, the ECG modulation layer aims to ensure:
- Maximum rigour where knowledge is fragile or consequential, and
- Maximum clarity where knowledge is simple or bounded.

Integrity is intended to be measured in proportion to knowability. A worked example is provided at Appendix 3.

Behavioral Gates

Gate 1 — Entity & Terminology Lock

Purpose:

To eliminate ambiguity in key terms, acronyms, regulations, institutions, or technical expressions before reasoning begins.

Why it matters:

LLMs frequently misinterpret or drift between definitions (e.g., “GDPR” vs. other privacy regimes; “MiCA” vs. generic crypto rules). Ambiguity at the start of reasoning produces systematic error downstream.

What the model must do:

1. Detect ambiguous terminology in the query and early reasoning.

2. Resolve each term to a single, authoritative definition with:

- jurisdiction
- version/date
- regulatory or technical anchor (e.g., EU, UK, FDA)

3. Freeze this definition and use it consistently throughout the answer.

Example:

“PSD2” → “EU Payment Services Directive 2015/2366, as amended to 31 Dec 2024.”

Gate 2 — Source Hierarchy & Verification

Purpose:

To ensure every factual claim includes a transparent, auditable indication of source credibility.

The Tier System:

- T1 — Primary sources (statutes, official registries, peer-reviewed datasets)

• T2 — Official guidance, systematic reviews, formally governed institutional documentation

• T3 — Reputable editorial media or aggregators

• T4 — Unverified, ambiguous, or model-inferred data (allowed only with qualification)

Requirements:

1. Every factual claim must include a tier tag.

2. Minimum acceptable standalone tier is T3.

3. Claims based solely on T4 evidence must be labelled “provisional / low confidence.”

Why it matters:

Tier tagging anchors claims to verifiable provenance and prevents fabricated citations, source conflation, or the appearance of false certainty.

Gate 3 — Evidence Integrity & Epistemic Balance

Purpose:

To ensure evidence is sufficient, independent, and not dominated by a single viewpoint or institutional lineage.

Gate 3 has three components:

A. Evidence Sufficiency

The model must identify at least:

• one independent T1 source, or

• two independent T2 sources.

B. Epistemic Concentration (70% Rule)

If ≥70% of the total evidence originates from:

• one country,

• one institution,

• one research group, or

• one conceptual lineage,

then the model must issue an "Epistemic Concentration Warning" at the beginning of the Context section.

This warns the user that the evidence, even if credible, may not be representative or diverse.

C. Arbitration Rule (0.70 Coherence + 0.15 Differential)

When two evidence clusters conflict:

1. The **internal coherence** of a cluster must be ≥0.70 before it can be treated as potentially dominant.

2. A cluster may be considered dominant only if its **evidence density** exceeds the alternative by ≥0.15.

3. If these conditions (see below) are not met, the model must:

• present both interpretations (dual-path reasoning), and

• elevate uncertainty.

<u>Internal Coherence</u>

- Definition:

The degree to which the output's claims, assumptions, and inferences are logically consistent with each other and with any stated constraints or definitions, without contradiction or drift.

- In practice, Gate 3 tests whether:

  Premises align with conclusions (no non sequiturs).
  Terms are used consistently (no semantic drift).
  The reasoning chain is closed (no missing steps that materially affect validity).
  No internal contradictions emerge across sections.

- Failure modes:

  Mutually incompatible claims presented as jointly true.
  Conclusions that exceed or diverge from stated premises.
  Shifting definitions to sustain an argument.

Why it matters for Gate 3: Even high-volume evidence is invalid if the argument structure cannot sustain it. Internal coherence is the structural integrity check before density is even meaningful.

Evidence Density

- Definition:

The concentration of verifiable, relevant evidence per unit of claim or inference, sufficient to justify the strength, scope, and confidence of the conclusions drawn.

- In practice, Gate 3 tests whether:

  Each substantive claim is backed by appropriate quantity and quality of evidence.
  Evidence scales with claim severity (strong claims → denser evidence).
  Citations are not tokenistic but causally connected to the claim.
  Evidence diversity exists where required (not single-source stacking).

- Failure modes:

  Broad or definitive claims supported by sparse or weak evidence.
  Evidence cited but not actually supporting the inference made.
  Over-reliance on one source for multi-facet claims.
  Rhetorical padding in place of data.

Why it matters for Gate 3: Gate 3 is explicitly a density threshold, not just a presence check. It prevents evidence-light narratives from passing as rigorous analysis.

Relationship Between Internal Coherence and Evidence Density in Gate 3

- Internal coherence: Is the reasoning structurally sound?
- Evidence density: Is there enough high-fidelity support for that structure?
- Gate 3 fails if either the structure is coherent but under-evidenced, or the evidence is abundant but assembled into a contradictory or unsound argument.

Why it matters:

This prevents the model from automatically collapsing ambiguous or closely balanced evidence into a single, overly-confident conclusion.

Gate 4 — Temporal & Forward-Looking Controls

Purpose:

Gate 4 governs how forward-looking, predictive, or future-state claims are expressed and controlled. It ensures forecasts are not presented as facts and that implied confidence is explicit, auditable, and proportionate to evidence and uncertainty. Gate 4 does not compute probabilities. It enforces linguistic discipline over predictions and modality, with explicit temporal control over the dating of evidence used to justify forward-looking claims.

What Gate 4 Operates On (NLP Basis)

Gate 4 relies on natural-language parsing to identify:

• Futurity (forward-looking constructions).

• Modality/hedging (strength of commitment).

• Temporal expressions (explicit or implicit horizons).

• Justification text (evidence, mechanisms, assumptions).

This is modality and temporal parsing, not confidence prediction. If modality is unclear or mixed, Gate 4 defaults to challenge or "undetermined", not forced classification.

Temporal Control: Dated Evidence Tag

For each forward-looking claim, Gate 4 requires the supporting evidence basis to be date-qualified via either:

• an explicit evidence-as-of date (e.g., 2025-12-25),

• a validity window (e.g., 'data from 2024–2025', 'valid through Q2 2026'), or

• an explicit 'as of <date>' disclosure inside the justification.

This is intentionally non-redundant:

• **Gate 4 applies dated-evidence tagging** specifically to the evidence used to justify forward-looking claims.

• **Gate 6 handles general timestamping/validity windows** for quantitative statements across the response.

Confidence Taxonomy

• Certain — effectively entailed under stated assumptions.

• High confidence — very likely given strong evidence and stable mechanisms.

• Medium confidence — plausible but materially uncertain.

• Speculative — conjectural, exploratory, weakly supported.

• Undetermined (generation mode only) — insufficient basis to justify any band.

These are semantic commitments, not numeric ranges.

Two Operating Modes

Gate 4 behaves differently depending on whether the model is authoring content or evaluating externally authored content.

A. Generation Mode — OIP-JSON (Standard / Persistent)

When OIP-JSON is used to produce or improve an answer, the model is the author and must:

1. Detect its own forward-looking claims via futurity parsing.

2. Surface any implicit modality into an explicit confidence label.

3. Disclose for each such claim:

- confidence_label ∈ {certain, high, medium, speculative, undetermined},
- time_horizon,
- justification (evidence + assumptions + uncertainties), and
- dated-evidence tag (evidence-as-of date or validity window) satisfying the temporal control requirement.

Hard rule: No forward-looking claim may remain implicitly parameterised. Labels are not 'calculated'; they are the model's expressed level of support given cited evidence and assumptions.

B. Evaluation Mode — OIP-TF

When Gate 4 is used in TF scoring, the model evaluates externally authored text and must:

1. Detect forward-looking claims.

2. Read any explicit labels provided by the author.

3. If none exist, use modality parsing to record an implicit confidence inference for checking only.

4. Never insert labels into the evaluated text.

Temporal control in OIP-TF:

• The evaluator checks whether the author date-qualifies evidence supporting forward-looking claims (as-of date / validity window).

• Where datedness is missing, this can be recorded as a Gate 4 issue, but avoiding double-penalising if the same weakness is already scored under Gate 6.

Consistency Logic

Higher confidence requires stronger Gate 3 evidence density/tier, more stable mechanisms, shorter horizons, and lower domain volatility. Gate 4 checks mismatch between language modality, declared/implied label, evidence strength, horizon, and datedness of evidence.

Bias: challenge or downgrade, never upgrade.

Summary Rule

Gate 4 is modality discipline plus dated-evidence temporal control for forward-looking claims, not prediction in its own right. In generation, the model must expose its own implied confidence and date-qualify supporting evidence. In evaluation, the model must surface and challenge the author's implied confidence and datedness — without rewriting the source. In this case, modality is inferred for checking only and never injected into the evaluated text.

Why it matters:

Time-sensitive deterioration in accuracy is a major cause of LLM error. Users must understand the temporal and predictive reliability of assertions.

Gate 5 — Three-Layer Transparency

Purpose:

To impose a clear, auditable structure on every answer so that facts, analysis, and interpretation cannot be conflated, and to ensure internal coherence.

A. Three Layer Separation: Required structure:

1. —— Core Facts ——

    • Only T1–T3 evidence permitted

    • No interpretation

- No speculation
- No causal claims

2. —— Context & Conflicts ——

- Evidence disagreements
- Concentration Warnings
- Dated Evidence tags
- Scope limitations
- Counter-evidence and constraints

3. —— Interpretation & Residual Uncertainty ——

- Careful, qualified reasoning
- No new facts introduced
- Explicit uncertainty

B. Global Coherence and Non-Contradiction Check: In addition to separating Core Facts, Context, and Interpretation, the evaluator must perform a global coherence scan across all layers to identify and resolve:

- internal logical contradictions,

- mutually incompatible claims,

- shifts in definitions or referents, and

- conclusions that conflict with stated premises or earlier assertions.

Procedure: The evaluator shall:

1. Review the full output holistically after layer separation.

2. Test whether any statement in one layer contradicts statements in another.

3. Verify that interpretations do not negate or overstate the facts they are based on.

4. Check that contextual framing does not implicitly assume conclusions that are later presented as interpretive.

Enforcement: Where contradictions or incoherence are detected, the evaluator must:

- explicitly flag them, and

- either resolve them by revising the affected layer(s), or

- downgrade claims to qualified or indeterminate status, or

- refuse the proposition if coherence cannot be restored without unsupported assumptions.

C. Treatment in OIP-TF: Global Coherence Scan

The evaluator must assess whether the response exhibits internal logical consistency across Core Facts, Context, and Interpretation layers.

Scoring Criteria:

Score 0 (Fail):

- Unidentified or unaddressed contradictions exist within or across layers.

- Definitions or referents shift without disclosure.

- Interpretations directly conflict with stated facts or premises.

Score 1 (Core Pass):

- Potential tensions or ambiguities are explicitly flagged.

- Interpretive claims are qualified where coherence is uncertain.

- No direct contradictions remain unaddressed.

Score 2 (Reg Pass):

- The output demonstrates explicit internal consistency.

- No material contradictions are present across layers.

- All claims align coherently with stated premises, definitions, and evidence.

Failure Handling:

A failure must be recorded as a Gate 5 violation and contributes to interpretive integrity penalties in the TF aggregate score, regardless of factual correctness under other gates.

TF Schema Note: If tf_calibration.core_intensity_checks or reg_intensity_checks include G5, then this component must be evaluated and explicitly scored in the TF record under Gate 5 sub-checks.

Why it matters:

This structure creates improved cognitive clarity and enables external code (using an Orchestration Layer or during evaluation by OIP-TF) to validate the structure. This control also enforces the principle of non-contradiction and conceptual coherence in complex, layered reasoning, ensuring that transparency is not merely structural but also logically consistent across the entire response.

Gate 6 — Coverage & Estimation Discipline

Purpose:

To prevent "false completeness" — claims that imply exhaustive coverage when the underlying evidence does not support it. Gate 6 also enforces representational sufficiency via scope disclosures, domain-proportional coverage, or omission acknowledgement.

Requirements:

1. Detect totalising phrasing in the query ("all", "every", "complete list").
2. Provide a Coverage Statement at the end of Core Facts:

   "Coverage: X/Y evaluated | Missing: … | Completeness: ZZ% ± NN%"

3. Show calculation steps for any numerical estimation.
3. Provide a **coverage sufficiency and scope discipline** check – see below

<u>Coverage Sufficiency & Scope Discipline</u>

Trigger: Any presented set (list, typology, or enumeration) that could be read as broadly representative.

Rule: For each set, satisfy at least one of: (1) explicit scope disclosure (e.g., 'examples', 'non-exhaustive'); (2) domain-proportional coverage including dominant categories; (3) omission acknowledgement of major excluded classes. If none apply, flag coverage under-representation.

This check is complementary to the items 1-3 above: it targets misleading incompleteness without relying on absolutist phrasing: the check enforces coverage sufficiency to prevent quiet under-representation of sets even where absolutist phrasing is absent.

<u>Mode Handling</u>

Generation (OIP-JSON): the model adds scope disclosures, broadens sets, or acknowledges omissions before passing Gate 6.

Evaluation (OIP-TF): the model flags under-representation and cites the relevant text; avoid double-penalising the same deficiency across subrules.

Why it matters:

Users must understand whether the model is giving a full picture or a partial one.

Gate 7 — Dialectical Safeguard

Purpose:

To prevent unilateral or overly deterministic claims, especially in causal or evaluative reasoning. Gate 7 is designed to ensure that conclusions are not reached through one-sided reasoning by requiring credible alternative paths and explicit downside risks, without forcing false balance or amoral negation.

Core Principle

Require credible counter-paths, not logical negations. Allow explicit non-equivalence where no credible alternative exists without violating safety or ethical constraints. Require counter-paths aligned to claim type, allowing reasoned inaction/status quo in policy and strategy contexts, rejecting clearly amoral negation and false balance, and mandating escalation of downside risks.

Credible Counter-Paths

Empirical: competing hypotheses or interpretations.

Policy/strategy: alternative actions or reasoned inaction (status quo) with rationale, risks of inaction, and trigger conditions.

Normative/moral: alternative moral frameworks or value prioritization is acceptable; amoral negation is forbidden.

Safety-critical: alternatives must be provided within constraints or explicit statement of non-equivalence.

In policy and strategy, maintaining the current state ('do nothing') is a valid counter-path when defensible. It must include rationale, risks of inaction, and trigger conditions for change. Inaction must not be structurally disfavoured.

Do not imply symmetry where evidence or ethics are asymmetric. Counter-paths must be plausible within domain norms. Surface downside risks if the primary path is wrong, impact severity (blast radius), and affected domains.

Operational Behaviour

Generation: surface credible counters or declare justified non-equivalence, consider status quo where relevant, and escalate risks.

Gate 7 activates whenever the query includes a phrase from the Causal/Evaluative Trigger List (e.g., “cause”, “impact”, “risk”, “should”, “effective”).

Requirements:

1. The model must generate:

- a main reasoning path, and
- the strongest credible counter-path.

2. Both paths must appear in Context & Conflicts before any conclusion.

3. The conclusion in Interpretation must reflect:

- the unresolved tension,
- the certainty classification, and
- any remaining ambiguity.

Evaluation (TF): check presence, relevance, false equivalence, neglect of status quo, and missing risk escalation.

Why it matters:

This enforces intellectual humility and prevents persuasive but misleading single-path narratives.

PART 2 — ORCHESTRATION (ENGINEERING LAYER)

Overview

Part 2 describes how OIP can be transformed from a behavioral guardrail operating inside an LLM into a fully auditable and maximally stable compliance framework through the use of an external orchestration layer. This section is intended for developers, compliance engineers, risk managers, and auditors responsible for implementing OIP in production environments. The features described here are not embedded within the .json artefacts provided as part of the framework but are compatible with them.

Behavioral OIP improves reasoning quality. Orchestration of the .json artefacts in a separate code layer ensures structural compliance, auditability, and maximises reproducibility—requirements for regulated industries, safety-critical systems, and enterprise-scale deployments.

Gate 0 — Environment Lock (Policy Persistence)

Purpose:

To guarantee that every answer is generated under a controlled, verified, and auditable policy state. Gate 0 is treated as an orchestration-only control, not part of the seven epistemic reasoning gates, and operates separately from ECG modulation and OIP-TF scoring. It is not a logical reasoning gate but an external enforcement layer responsible for schema validation, source checks, policy hashing, and machine-auditable logging.

Why it matters:

Without policy persistence, LLM behaviour may drift across turns due to context erosion, user overrides, model variability, or prompt tampering. Gate 0 ensures that the compliance environment is fixed and verifiable at runtime.

What the orchestration layer must do:

1. Verify the version hash of:

- OIP regulatory prompt
- Hard-coded safety lists (FRW, JIW, SBW, MIW, etc.)
- Validation engine code

- JSON schema specification

2. Reject generation if any hash does not match the Official Governance Registry (OGR).

3. Log all environment metadata in the Genetic Attribution Log.

4. Prevent unlogged disabling of OIP (e.g., require explicit, auditable --oip-off commands).

Outcome:

No answer can be produced unless the compliance environment is consistent, current, and tamper-free.

Note: the Official Governance Registry must use secure version-control practices (Git, immutability logging) to ensure:

• policy persistence,

• immutable version tracking,

• tamper detection,

• reproducibility of results.

The Orchestration Layer

Purpose:

To shift critical compliance functions from the probabilistic LLM into a maximally stable, reproducible process (whilst acknowledging the underpinning reliance on LLM sub-tasks).

Key functions:

- Enforce missing-gate detection and re-prompting
- Validate JSON structure
- Verify tiers using whitelist lookups
- Perform semantic drift checks
- Ensure Coverage and Dialectical safeguards are correctly applied
- Generate machine-readable audit logs

Benefits:

- Predictable, repeatable results
- Strong documentation for regulators
- Clear separation between reasoning (LLM) and enforcement (code)
- Full forensic traceability

Validation Engine — The Deterministic Auditor

Purpose:

To test whether the LLM's output satisfies the full OIP-Regulatory specification.

The Validation Engine must perform:

1. Structural Integrity Check

- Validate that the output conforms exactly to the OIP-Regulatory JSON schema.
- Missing or extra fields → consider modulated CAF penalty.

2. Source Integrity Checks (Gates 1–3)

- Verify regulator URLs against the Fixed Regulatory Whitelist (FRW, see Glossary).
- Verify terminology definitions against the Jargon Integrity Whitelist (JIW, see Glossary).
- Screen sources against the Sanctions and Bias Watchlist (SBW, see Glossary) (sanctions/bias).
- Validate Tier assignments and evidence counts.

3. Semantic Drift Check

- Compare the model's summary of T1/T2 sources against the original text using vector similarity (e.g., cosine similarity ≥ 0.85).
- If drift detected → flag failure, demote source, require regeneration.

4. Dialectical Safeguard Verification (Gate 7)

- Confirm that causal/evaluative triggers resulted in:
  - Main path
  - Counter-path
  - Counterfactual
- Absence of any → non-compliant output.

5. Temporal Validity Check (Gate 4)

- Ensure all cited sources are tagged as expected through Gate 4 / Gate 6 requirements.

6. Coverage & Completeness Validation (Gate 6)

- Ensure completeness scores and X/Y coverage metrics are present where required.

Output:

A Boolean pass/fail for every rule, plus metadata for the Scoring Engine.

Scoring Engine — CAF and Integrity Score

Purpose:

To quantify compliance and risk.

Two core metrics:

1. Compliance Adherence Factor (CAF)

- Measures structural compliance with OIP-Regulatory JSON schema.
- Minimum recommended threshold for deployment: 95%.

The Severity Assignment Protocol (SAP) defines how the evaluator model determines the severity of each compliance violation for the purposes of calculating the Compliance Adherence Factor (CAF) under a weighted model.

The protocol ensures that:
• Severity is rule-anchored, not arbitrary;
• Violations are treated proportionally to epistemic and operational risk;

• Catastrophic integrity breaches remain zero-tolerance;
• Outputs are explainable and auditable.

Each violation is assigned to exactly one of the following severity classes:

Minor — Formatting, metadata, or presentational issues that do not materially affect integrity or meaning.
Major — Structural or evidentiary weaknesses that reduce reliability but do not alone invalidate the output.
Critical — Integrity failures that materially undermine trustworthiness or compliance for high-risk content.
Catastrophic — Evidence of fabrication, hallucination, tampering, or deliberate bypass of safeguards. Triggers immediate CAF = 0%.

Each OIP Gate defines a default_severity.

The evaluator must begin severity assignment from this default. This ensures cross-model consistency, stable governance semantics, and audit reproducibility.

Upon detecting a rule failure, the evaluator must classify the failure type (e.g., missing, low_tier_only, misattributed, fabricated, structural_absence, disabled_safeguard).

Rules may define failure modifiers that adjust severity, including catastrophic overrides.

The following conditions are always be treated as catastrophic, regardless of rule default or ECG:

• Fabricated or hallucinated sources or citations;
• Fake quotations or attributed statements;
• Tampering with audit or provenance fields;
• Explicit disabling of dialectical safeguards;
• Intentional misrepresentation of evidence.

If detected: final_severity = catastrophic and CAF = 0%. No further modulation is applied.

If not catastrophic, severity SHALL be modulated based on the Epistemic Category of the Question (ECG):

E2, E3, E1, E5: No change
E6, E4: Upgrade one severity level (if below Critical)

This reflects that the same failure carries greater consequence when claims are causal, evaluative, or high-stakes.

Severity adjustment is bounded as follows:

• May shift at most one level up or down from default;
• Cannot fall below Minor;

• Cannot exceed Critical (unless catastrophic);
• Catastrophic is reachable only via override triggers.

Severity ladder:
Minor ↔ Major ↔ Critical → Catastrophic

The evaluator computes severity by:

1) Starting from default_severity;
2) Applying any failure modifiers;
3) Applying ECG adjustment;
4) Enforcing bounded limits;
5) Emitting final_severity.

For each violation, the evaluator MUST emit an explanation including:

• rule identifier;
• default_severity;
• failure_type;
• ECG class;
• adjustments applied;
• final_severity;
• CAF penalty;
• rationale.

This is mandatory for audit traceability, remediation guidance, and governance review.

Each final severity maps to CAF penalties:

Minor: –5%
Major: –20%
Critical: –40%
Catastrophic: CAF = 0%

Penalties accumulate unless overridden by catastrophic zeroing.

This protocol guarantees that:

• Minor lapses cannot collapse compliance;
• High-risk failures escalate automatically;
• Fabrication remains zero-tolerance;
• Evaluations are explainable and reproducible;
• CAF behaves as a graded integrity control, not a brittle fuse (important given stochastic behaviour of LLMs).

Every implementation must either follow this Severity Assignment Protocol for any weighted CAF model or explicitly declare deviations.

2. Integrity Score (IS)

- Weights rule violations by severity.
- Examples:
  – Source-tiering error > formatting error
  – Missing concentration warning > minor metadata omission
  – Dialectical Safeguard omission > incomplete estimation steps
- Produces a final risk rating for audit and deployment workflows.

Both CAF and IS must be recorded in the Genetic + Agentic Attribution Log.

Genetic + Agentic Attribution Log

Purpose:

To provide a complete forensic trace of every OIP-governed answer.

Two layers:

A. Genetic Attribution (External State)

- User ID
- Timestamp
- Query text
- OIP Prompt Hash
- Whitelist Version
- Model identifier

B. Agentic Attribution (LLM Behaviour)

- Confidence score
- T1/T2 source count
- Concentration ratio

- Arbitration outcomes
- Identified conflicts
- Coverage score
- Any error or override

Outcome:

A combined log enabling regulators, auditors, or internal teams to reconstruct any answer's provenance and reasoning.

Two-Pass Validation and Correction Loop

Purpose:

To limit risk that non-compliant answers are not shown to end users.

Process:

Pass 1 — LLM Generation

- Produce answer using OIP-Regulatory.

Pass 2 — Validation Engine

- Validate JSON structure
- Validate tiering, conflicts, concentration, temporal tagging
- Validate semantic drift and coverage

If fail:

- System re-prompts with detailed error instructions
- Answers regenerated until compliant OR max attempts reached
- All failures logged in attribution log

This aims to create deterministic, measurable compliance.

Contextual Compliance & Safe Suppression

Purpose:

To avoid unnecessary warnings when the scope of the question logically restricts the evidence domain.

Example:

Query: “Summarise key features of French AML law.”

Evidence comes from French regulators → 100% French sources → normally triggers concentration warning.

Safe Suppression Rule:

• Warning may be omitted if concentration logically matches the query scope.

• Suppression must be logged in metadata for audit.

Cross-Domain Ambiguity Checks

Purpose:

To prevent improper tier assignment when acronyms span multiple domains.

Rules:

1. Domain-first resolution — determine regulatory domain from Gate 1.

2. URL validation — acronyms (e.g., “ISO”) cannot justify T1 classification without matching whitelist URLs.

3. Contextual downgrade — ambiguous acronym → default T2 unless validated through direct link.

Structured Output Enforcement

Purpose:

To guarantee machine-readability and auditability.

Requirements:

• All OIP-JSON outputs may be delivered as JSON objects for pipeline enforcement.

• ECG modulation must be used to disapply certain (non applicable) output components and related calculations (Gate applicability per Part 1 / Part 3 for OIP-TF).

• In JSON schema and outputs the gates must be referenced by stable identifiers i.e. G1-G7 with descriptive names treated only as informative.

• The orchestration layer uses schema validation to enforce:

– Field presence

– Data types

– Ordering

– Nested structure rules

PART 3 — OIP TESTING FRAMEWORK (OIP-TF)

A Model-Agnostic Integrity Evaluation System for AI and Human Output

Overview

The OIP-Testing Framework (OIP-TF) is a standalone diagnostic system designed to assess the integrity of any text—whether generated by an LLM, a human analyst, or a hybrid workflow. Unlike Parts 1 and 2, which influence model behaviour, OIP-TF is an evaluation tool. It does not modify the underlying system. Instead, it provides a structured, repeatable, and auditable assessment and score of factual, evidential, and logical quality.

The framework focuses strictly on observable textual behaviour. It:

- does not infer intent,
- does not use domain-severity multipliers, and
- does not depend on external metadata or context beyond the answer being scored.

The goal is that two independent evaluators can assign the same scores to the same output, based only on what is explicitly written.

OIP-TF is suitable for:

- regulators and supervisors evaluating AI-generated or human-generated content
- auditors and second-line assurance functions
- model risk teams conducting independent validation
- journalists and fact-checkers
- procurement teams comparing model performance
- internal QA functions reviewing advisory outputs

Its primary purpose is to produce consistent, evidence-driven integrity ratings that can be incorporated into enterprise workflows, risk scoring, procurement comparisons, or compliance escalation. It does this by evaluating a given answer against a set of integrity dimensions aligned with the OIP behavioral standards. It can be applied regardless of model vendor, architecture, prompting strategy and whether OIP gates were used during generation.

The process requires access to the specific original source data, public datasets, regulatory documents, or other authoritative references depending on the question; but is also capable of decomposing a single text into inferred source facts (“as if true”), questions or postulations and conclusions or inferences. The framework also produces machine-readable metadata (JSON-compatible). Every axis must be scoreable solely from content present in the output text.

- Scoring decisions must cite specific textual segments in the answer.
- No interpretation of unstated reasoning steps is permitted.

- No external facts, risk classifications, or user intentions may be imported into the scoring.

OIP-TF — How Verdicts Work

Inputs and Separation Requirements

The evaluation pass involves **user specification** of three inputs separately:

• SOURCE DATA X: the evidence base (logs, excerpts, dataset snippet, policy text, etc.).

• QUESTION Q: the task prompt the answer is responding to.

• OUTPUT Y: the answer to evaluate (verbatim; do not edit).

OIP-TF evaluates the information reliability and transparency of the answer (Y) against the declared evidence source(s) (X), the question (Q), considering the governing epistemic category of the question (ECG). It uses the same set of integrity gates as defined in Part 1, each producing a PASS, WARNING, or FAIL outcome. These gate outcomes are then mapped to an overall Assurance Level.

1. Gate Outcomes

Each integrity gate assesses a specific risk area (evidence grounding, attribution, coverage, causal sufficiency, posture alignment, structural compliance, etc.). For every gate:
• PASS means the integrity expectation for that gate is met.
• WARNING means a material weakness exists but integrity is not fully broken.
• FAIL means a breach of expectation for that gate.

PASS indicates fitness for purpose. WARNING indicates caution is required. FAIL indicates the output cannot be justified under the declared evidence contract.

2. What Triggers a FAIL

A FAIL is triggered when any of the following occur, depending on the gate:
• Claims in Y are unsupported by X or hallucinated.
• Sources are missing, invented, or misattributed.
• Key aspects of Q are ignored or selectively answered.
• Causal claims are made without an adequate mechanism (correlation narrated as causation).
• ECG posture is violated (e.g., speculative predictions made where only factual inference is warranted).
• Structural or compliance breaches occur (missing anchors, missing mandatory fields).

A FAIL does not necessarily mean the answer is factually false. It means the answer is not epistemically justified relative to X, Q, and the governing posture.

3. What Produces a WARNING

WARNINGS indicate weaknesses that reduce confidence but do not fully break integrity, such as:
• Thin but minimally adequate evidence density.
• Partial coverage of Q.
• Underdeveloped but plausible causal mechanisms.
• Minor attribution gaps with explicit caveats.
Multiple WARNINGS accumulate but do not escalate unless a FAIL is present.

4. Overall Verdict: Epistemic Assurance Levels

Gate outcomes are mapped to a single overall verdict expressed as an Epistemic Assurance Level:
• AL-High — Strong integrity across all applicable gates. No FAILs, no material WARNINGS, and no ECG severity cap.
• AL-Moderate — No FAILs, but one or more material WARNINGS. ECG posture remains aligned.
• AL-Low — One or more FAILs on any gate, or an ECG severity cap is applied, or structural non-compliance occurs.

Rule of thumb: Any FAIL anywhere forces the outcome to AL-Low.

5. ECG Severity Cap

If the answer adopts a more speculative or predictive posture than the question warrants, and X does not support that posture, ECG is adjusted upward and a severity cap applies. This automatically forces the verdict to AL-Low, even if other gates pass.

6. Structural Non-Compliance

Failures to meet mandatory schema and anchoring requirements (e.g., missing Q/Y anchors, absent segmentation manifest, or unresolved low ECG confidence) are treated as integrity failures and result in AL-Low.

7. Interpretation of the Verdict

• AL-High means the reasoning is well-grounded and suitable for reliance.
• AL-Moderate means the reasoning is broadly sound but should be used with caution.
• AL-Low means the output is not epistemically trustworthy under the declared evidence contract and should not be relied upon without remediation.

Epistemic Assurance reflects justified confidence in reasoning discipline. It does not assert objective or universal truth.

8. Summary

OIP-TF treats epistemic integrity as a safety property: a single material breach (FAIL) or unjustified speculative posture is sufficient to render an output low-assurance, regardless of other strengths.

The final OIP-TF report aggregates all module results into a single structured output suitable for audit use.

Core Calculations

Source Data Adequacy Assessment (SDA)

Classify X as one of S1–S5 and record rationale in source_data_assessment:

• S1: Direct/authoritative evidence (primary records, logs, raw datasets with provenance).

• S2: Secondary/derived analysis grounded in primary sources.

• S3: Anecdotal/unverified (unclear provenance).

• S4: No relevant evidence for the question requirements.

• S5: Internally conflicted/contradictory evidence.

If mixed, set source_data_assessment.mixed=true and include an x_coverage_map for key Q parts.

ECG Classification and Adjustment to ECG(Q|X)

Objective:
Set the epistemic burden required by the question (ECG(Q)), then condition that burden on the available evidence X to obtain ECG(Q|X), which constrains what may be asserted, inferred, or must be refused.

1. Determine ECG(Q) from Q alone

- Treat ECG(Q) as the epistemic contract implied by the question prior to evidence.

- Classify Q (or propositions $P_1 \ldots P_n$ if compound) into E1–E6.

- When decomposed:

- Assign ECG(P_i) for each proposition.
- Escalate to a single governing class using the defined escalation order.

  E2 → E3 → E5 → E1 → E6 → E4

- Record in ecg.q_only_classification.

2. Adjust ECG(Q) to ECG(Q|X) using evidence X

- Use SDA(X) and X content to condition what the question can responsibly demand.

- Record in ecg.qx_adjustment: ecg_q, ecg_q_given_x, adjusted, and any constraints_from_x.

2.1 Indexical access upgrade

Rule A — Access absence → E2 (terminal)

If Q depends on specific referents, context, or artefacts (documents, logs, prior turns, attachments), and X does not provide the required access:

- Set ECG(Q|X) = E2.

- The integrity-maximising outcome is refusal due to missing access.

- Do not attempt evidential synthesis.

Rule B — Evidence absence → E3 (terminal)

Else if Q concerns facts or states answerable in principle (e.g., intent, causes, outcomes), and X does not contain evidence sufficient to determine them:

- Set ECG(Q|X) = E3.

- The integrity-maximising outcome is refusal or explicit indeterminacy due to lack of evidence.

- Do not attempt evidential synthesis.

Rule C — Access and evidence present → classify normally

Else:

- X provides the access and evidence required to answer Q.

- Determine ECG(Q|X) from the content of Q and X as one of: E1 (direct factual), E4 (inferential/model-based), E5 (analytic/evaluative), or E6 (normative/prescriptive), as appropriate.

- In this case, ECG(Q|X) must not be E2 or E3.

3. Set governing ECG(Q|X)

- Set ecg.governing_class = ecg_q_given_x.

- Governing ECG(Q|X) defines:

- the strictest posture required for scoring,
- refusal/qualification boundaries,
- downstream gate calibration.

4. Interaction with answer posture (CPD-HSF)

- After ECG(Q|X) is set, scan Y for high-stakes postures.

- If detected, impose the CPD-HSF minimum ECG floor.

- Downward adjustment below this floor is forbidden.

- Record the floor and rationale separately.

5. Principles

- Adjustment conditions epistemic burden; it does not relax it to permit unsupported claims.

- ECG(Q|X) specifies what is answerable, what must be qualified, and what must be refused given X.

- All adjustments and constraints must be explicitly recorded for auditability.

Gate Profile and TF Calibration

1. Compute ecg.gate_profile after overrides (see Part 1):
   0 = Zero OIP → Gate disapplied.
   1 = OIP-Core → Core integrity discipline applies (basic behaviour prompt component).
   2 = OIP-JSON → Full-grade discipline applies.

2. Populate tf_calibration lists:

- disapplied_checks: checks not scored (no penalty).
- core_intensity_checks: scored at Core burden.
- reg_intensity_checks: scored at full Reg burden.

3. Ensure the calibration matches the computed gate profile for the run.

Source-Grounded Correctness: Claim Entailment & Coverage

1. Decompose Y into atomic claims and label each claim_type (factual/inferential/normative/procedural/meta_uncertainty).

2. For each factual or inferential claim, assign entailment status relative to X:

• entailed / contradicted / unsupported / underspecified / non_truth_apt

3. Span anchoring rule (mandatory for high-confidence labels):

• If status is entailed or contradicted, include x_spans (quotes or pointers) from X.

• If you cannot provide an anchor, default to unsupported or underspecified.

4. Decompose Q into required parts and populate question_coverage with one of:

• answered_correctly / answered_incorrectly / omitted / refused_appropriately / unanswerable_from_x_unacknowledged

Scoring and Reporting

1. Score the response using OIP-TF fields, but only enforce checks enabled by tf_calibration.

2. Treat appropriate refusal as a success when SDA is S4/S5 or ECG(Q|X) implies unanswerability.

3. Ensure all required schema fields are populated, including inputs, ecg, tf_calibration, source_data_assessment, and source_grounded_correctness.

Quality Controls and Escalation

• If SDA is S5 or X is long/noisy, consider dual-pass evaluation.

• If evaluator passes disagree materially on ECG(Q|X) or entailment, apply conservative fallback and record it.

• For high-stakes outputs, sample for human review; do not treat evaluator output as a real-world truth oracle.

Other Characteristics

SEG-1.0 Segmentation Specification

SEG-1.0 defines the segmentation rules used by OIP-TF to partition an evidence source (X) or document (DOC) into canonical segments for anchoring and audit. The goal is to ensure that identical inputs, processed under the same ruleset, always yield identical segment boundaries and identifiers, enabling reproducible span mapping and validation across time, systems, and evaluators.

SEG-1.0 is governed by the following principles:

• Determinism: The same input and rules must always produce the same segments.
• Content fidelity: Segments preserve the semantic content of the source without reinterpretation.
• Stability: Minor layout changes should not materially alter segment identities.
• Human navigability: Segments correspond to human-readable units (paragraphs, headings, rows).
• Format awareness: Rules adapt to source type but remain conceptually uniform.
• Versioned evolution: Any rule changes require a new SEG version identifier.

SEG-1.0 supports the following input classes:

• Plain text
• HTML articles / web pages
• PDF documents (text layer)
• Word documents (.docx)
• Tables (CSV/structured rows) when embedded or standalone

Binary images or scanned PDFs without OCR are out of scope for SEG-1.0 and must be pre-processed before segmentation.

Segments are created using the largest stable semantic block available for the format:

• Headings / titles → one segment per heading.
• Paragraph blocks → one segment per paragraph.
• List items → one segment per list item.
• Table rows → one segment per row.
• Figure captions → one segment per caption.
• Standalone quotations → one segment if formatted distinctly.

SEG-1.0 does not segment at sentence or clause level by default, to preserve stability and reduce fragmentation.

Segments are ordered strictly in source order. Each segment is assigned:

seg_id format:
<source_id>.<type_code><ordinal>

Examples:
X1.P001 (paragraph 1)
X1.H002 (heading 2)
DOC1.R010 (row 10)

Type codes:
H = heading
P = paragraph
L = list item
Q = quote

R = table row
C = caption

Ordinal numbering is zero-padded to three digits per type sequence within the document.

Production of a manifest file is mandatory for each OIP-TF run. Each segment entry in the manifest must include:

• seg_id
• seg_type
• normalized_text
• source_locator (best-effort layout hint, e.g., page/paragraph or DOM path)

Additional optional fields may include:
• page_index
• block_index
• style (e.g., heading level)
• table_name

Metadata is advisory and must not affect hashing or identity.

A SEG-1.0 manifest is a JSON artefact with:

• manifest_id
• source_id
• generated_at (ISO 8601)
• segmentation_rules_version = 'SEG-1.0'
• normalization_rules_version = 'NORM-1.0'
• source_metadata (type, origin, retrieval time)
• segments[] array

The manifest must be immutable once generated.

Running SEG-1.0 multiple times on identical source content must produce byte-identical manifests (except generated_at). Any change in segmentation behavior requires a new version identifier (e.g., SEG-1.1).

Manifests should be stored with:

• Write-once semantics.
• Optional digital signatures.
• Access controls if evidence is sensitive.

The manifest is part of the audit trail and must be retained with the OIP-TF output artefact.

Important Limitation Note:

OIP-TF relies on a segmentation step to decompose an evaluated response into units of analysis (e.g., claims, assertions, lists, forward-looking statements, interpretive passages) prior to applying gate checks. This segmentation is an approximate, heuristic process, whether performed by an LLM, rules, or a hybrid approach.

Accordingly:

• No unique "correct" segmentation exists. Different reasonable segmentations of the same text may be possible, particularly for compound sentences, blended factual and interpretive statements, nested claims, or dense technical passages.

• Boundary decisions are judgment-dependent. Where one evaluator treats content as a single claim, another may split it into multiple claims, affecting evidence density calculations (Gate 3), coverage and set analysis (Gate 6), and counter-path identification (Gate 7).

• Result variation is expected within bounds. Because segmentation influences which checks are triggered and how often, TF scores may vary modestly across runs or evaluators even when applying the same rules in good faith.

• TF is designed for robustness, not exact determinism. The framework emphasises consistency of failure-mode detection, transparency of quoted evidence and rationale, and reproducibility of reasoning chains, rather than bit-identical scoring outcomes.

• High-impact findings should be reviewable. For material FAILs or regulatory use, segmentation choices and quoted evidence should be inspected by a human-in-the-loop to confirm that conclusions are not artefacts of boundary placement.

OIP-TF results should be interpreted as **structured, high-fidelity approximations of epistemic quality, not as mathematically exact measurements**. Small score deltas attributable to segmentation variance do not invalidate findings; emphasis should be placed on which gates failed and why, not on marginal numeric differences.

<u>NORM-1.0 Canonical Text Normalization Specification</u>

NORM-1.0 defines the text normalization rules used by OIP-TF (and SEG-1.0) to create a canonical representation of segment text prior to anchor validation. The objective is to ensure that semantically irrelevant variations (e.g., whitespace, stylistic punctuation variants) do not cause drift, while preserving semantic content (especially numbers, units, and negations) required for high-stakes auditability.

NORM-1.0 is mandatory for strict anchoring mode. NORM-1.0 is governed by:

• Determinism: identical input must yield identical normalized output.
• Semantic preservation: normalization must not alter meaning, including numeric values, signs, and negations.
• Minimalism: avoid aggressive transformations that could erase evidentiary distinctions.

• Transparency: rules are explicit, ordered, and testable.
• Versioned evolution: any rule change requires a new NORM identifier.

NORM-1.0 applies to Unicode text extracted from:

• HTML article bodies (post readability extraction),
• PDF text layer extraction,
• Word document paragraphs,
• Plain text and structured row strings.

It does **not apply directly to images; OCR output must be provided as Unicode text** before normalization.

All character ranges (char_range) used in anchors refer to offsets within the normalized segment text (after NORM-1.0). This guarantees that anchor validation is stable across platform-specific line endings and typographic punctuation.

The following illustrative examples are non-normative but should be used as sanity checks.

Example A (Quotes and whitespace)
Input: “Hello world”
Output: "Hello world"

Example B (Dashes)
Input: 10–20% — estimated − not guaranteed
Output: 10-20% - estimated - not guaranteed

Example C (Ellipsis)
Input: Wait… what?
Output: Wait... what?

Example D (Zero-width)
Input: A B
Output: A B

SCORE-1.0 — Numeric Integrity Scoring Specification

SCORE-1.0 defines a numeric scoring method for OIP-TF that operates in parallel with categorical gate outcomes (PASS / WARNING / FAIL) and the overall Epistemic Assurance Level (AL-High / AL-Moderate / AL-Low).

The Numeric Integrity Score (NIS) is intended to:
- provide finer-grained differentiation between outputs with similar assurance levels,
- support prioritisation, trend analysis, and regression testing,
- remain fully auditable and reproducible,
- and preserve the primacy of Assurance for governance decisions.

Note: This is not to be confused with the “Integrity Score” (IS) used as part of the Orchestration Layer to assure proper framework execution.

SCORE-1.0 does not replace categorical verdicts. Assurance remains the authoritative decision signal. An OIP-TF evaluation implementing SCORE-1.0 produces, in addition to gate outcomes and Assurance:

- Numeric Integrity Score (NIS): integer in the range 0–100.
- Raw Penalty: sum of penalties applied across gates in scope.
- Maximum Penalty: sum of maximum possible penalties across gates in scope.
- Per-gate scoring breakdown.
- score_rules_version = 'SCORE-1.0'.

These outputs must be reported together to preserve auditability.

Each OIP-TF gate is assigned a fixed severity class for scoring purposes:

- Critical gates — integrity failures that directly undermine epistemic justification.
- Secondary gates — important but supporting integrity dimensions.

Gate class assignment is fixed and versioned within the OIP-TF schema and must not be inferred dynamically by evaluators.

For each gate outcome, the following penalties apply:

Critical gates:
- PASS → 0
- WARNING → -15
- FAIL → -40

Secondary gates:
- PASS → 0
- WARNING → -10
- FAIL → -30

Only gates that are **in scope for the evaluation contribute to numeric scoring**. Disapplied gates are excluded entirely from numeric scoring and from the maximum penalty calculation.

Gate scope is **determined by ECG modulation** and any other disapplication logic in the OIP-TF schema. This ensures that questions with simpler epistemic profiles and fewer applicable gates are not unfairly advantaged or penalised relative to those with more complex profiles.

For each applicable gate g, define penalty(g) according to its severity class and outcome.

Raw Penalty is computed as:
RawPenalty = Σ penalty(g) for all applicable gates.

Raw Penalty represents the absolute magnitude of integrity deficits observed in the run, but is not directly comparable across runs with different gate scopes.

For each applicable gate g, define max_penalty(g) as:

- 40 for critical gates,
- 30 for secondary gates.

Maximum Penalty is computed as:
MaxPenalty = Σ max_penalty(g) for all applicable gates.

This represents the worst-case integrity deficit possible under the set of gates actually in scope for the run. Note that in the JSON schema the gate descriptors must be referenced to the canonical labels G1-G7.

The Numeric Integrity Score is computed as a **normalised percentage of remaining integrity capacity**:

NIS = round(100 * (1 - RawPenalty / MaxPenalty))

Where:
- RawPenalty and MaxPenalty are as defined above.
- NIS is rounded to the nearest integer.
- NIS in [0, 100].

If there are no applicable gates due to ECG modulation, NIS is not produced but the run is COMPLIANT – Assurance still applies. NON-COMPLIANT only applies when gates are expected but missing due to schema or execution failure.

This normalisation ensures that scores are comparable across evaluations with different gate scopes.

The Numeric Integrity Score does not alter Assurance outcomes.

Rules:
- Any FAIL on any gate still forces Assurance = AL-Low.
- ECG severity caps still force Assurance = AL-Low.
- Structural non-compliance still forces Assurance = AL-Low.

SCORE-1.0 applies no caps to NIS. High-scoring failures are permitted and are interpreted as near-miss cases requiring targeted remediation.

Governance decisions must always be based on Assurance, not NIS.

Illustrative Example:

If three applicable gates are in scope:

- Evidence Integrity & Epistemic Balance (critical): WARNING → penalty 15, max 40
- Coverage & Estimation Discipline (secondary): FAIL → penalty 30, max 30
- Dialectical Safeguard (critical): PASS → penalty 0, max 40

Then:
RawPenalty = 15 + 30 + 0 = 45
MaxPenalty = 40 + 30 + 40 = 110

NIS = round(100 * (1 - 45/110)) = round(59.09) = 59

If any FAIL exists, Assurance would still be AL-Low regardless of NIS.

Interpretation Guidance:

NIS provides relative discrimination within and across Assurance categories:

- Higher NIS indicates fewer or less severe integrity deficits within the tested scope.
- Lower NIS indicates more extensive or severe integrity deficits.

NIS should be used for:
- ranking outputs within the same Assurance level,
- prioritising remediation,
- tracking improvements or regressions over time.

NIS must not be interpreted as probability of truth or as a replacement for categorical judgement.

LLM-Based Source Evaluation & Self-Audit Safeguards

This section defines how OIP-TF may employ a Large Language Model (LLM) to perform a self-contained evaluation of source data X in order to support Source Data Adequacy (SDA) classification, adjustment of the Epistemic Class Gate to ECG(Q | X), and source-grounded correctness assessment of an answer Y. It also establishes mandatory safeguards to prevent self-regard bias, over-confidence, and circular validation when LLMs are used as evaluators.

Role of the LLM in Evaluating X:

Within OIP-TF, an LLM may act as an Evidence Analyst, tasked with segmenting and indexing X; extracting atomic claims from X; identifying definitions, datasets, dates, and qualifiers; detecting internal inconsistencies in X; classifying X under the SDA scheme (S1–S5); performing textual entailment checks between X and Y; and supporting adjustment from ECG(Q) to ECG(Q | X).

This role is strictly limited to **assessing what follows from X as given, not whether X is true in the external world**. All conclusions are therefore truth-relative to X, not claims of objective reality.

An LLM used in this role must be assumed to have the following limitations: it **cannot independently verify real-world truth beyond X**; it may misclassify authority or provenance based on surface cues; it may miss subtle contradictions in long or noisy inputs; it may over-generalise from weak textual signals; and it may exhibit bias toward coherence over strict entailment.

Accordingly, LLM evaluation of X is **analytical and heuristic, not authoritative**. Its outputs are audit artifacts, not final arbiters.

When the same LLM (or tightly coupled instance) is used to generate an answer Y and to evaluate X and score Y against it, there is a structural risk of self-regard bias. This includes over-crediting its own phrasing as supported by X; reconstructing intended meaning rather than strict entailment; rationalising gaps in evidence; and preferring internal coherence over faithful grounding. This risk is especially material in high-stakes regulatory domains, compliance validation, or benchmarking scenarios.

To mitigate self-regard risk, OIP-TF establishes the **Model Separation Principle**:

Where OIP-TF is implemented within LLM-based systems, the model used to evaluate X and score Y must not be the same model instance used to generate Y.

Acceptable configurations include: different model families; different checkpoints or versions of the same family; isolated system prompts and roles with enforced context separation; or offline or batch evaluator agents distinct from generation agents.

Not sufficient are re-prompting the same model in the same context window, or asking the model to critique its own answer without isolation. The goal is to ensure that the evaluator reconstructs claims strictly from X, *not from latent intent or memory of having produced Y*.

When an LLM is used to evaluate X and Y, its prompt must forbid reconstruction of Y's intent; require that a claim be marked entailed only if specific spans in X can be cited as support; default to unsupported where evidence is ambiguous; and require explicit marking of contradictions, missing evidence, and under-specification.

Where possible, the evaluation should quote or index exact segments of X supporting each entailment decision, and record uncertainty when X is long, noisy, or ambiguous.

For high-impact workflows, OIP-TF RECOMMENDS dual-pass evaluation using two independent evaluator runs. If SDA classification, ECG(Q | X), or claim entailment materially differ, the system SHOULD default to the more conservative classification and/or escalate to full OIP-Reg posture for scoring.

This ensures that uncertainty in evaluation tightens, rather than loosens, discipline.

LLM-based evaluation of X must emit **auditable artifacts**, including: SDA classification and rationale; extracted X claims; ECG(Q | X) adjustment and reasons; claim entailment table for Y; and notes on ambiguity, conflict, or evaluator uncertainty.

These artifacts form part of the OIP-TF output and enable human review, sampling and QA, and benchmarking across models.

LLM evaluation of X must be treated as **decision support, not authority**, in high-stakes domains. It must be subject to sampling, spot checks, or human review where consequences are material, and must not be used as the sole basis for legal, medical, or regulatory determinations without human validation.

This approach allows OIP-TF to achieve source-grounded correctness scoring at scale; reward simple but fully supported answers; detect fluent but unsupported reasoning; and audit epistemic posture relative to evidence, while explicitly managing the risks of circular validation.

By enforcing model separation and conservative fallbacks, OIP-TF aims to ensure that integrity is preserved not by trusting the model, but by *structuring how it may reason about its own outputs*.

Orchestration Layer Responsibilities (design principles – not included in .json artefacts)

The orchestration layer MUST:

- Enforce mode selection and required fields.
- Snapshot mutable external X (web) with X_id, timestamp, and source list.
- Generate stable span references for X and documents.
- Validate anchor discipline:

  - external_truth → every WARNING/FAIL has x_span_ref.

  - document_internal → every WARNING/FAIL has doc_span_ref.

- Enforce dual output delivery (JSON + prose).
- Enforce verdict caps and assurance_level mapping rules.
- Flag NON-COMPLIANT runs and block publication until fixed.

If validation fails (missing anchors, missing prose, inconsistent fields, verdict-rule violations):

1) Mark run NON-COMPLIANT.

2) Revisit claim decomposition, ECG classification, and anchoring.

3) Regenerate JSON and prose outputs.

4) Re-validate before release.

PART 4 — GLOSSARY & DERIVATION BASIS

Clarifying Key Terms and Explaining the Origin of OIP's Safety Lists

Overview

Part 4 provides clear definitions for specialist terminology used throughout the OIP framework and explains how OIP's hard-coded whitelists are derived. This section ensures transparency, supports auditability, and helps both technical and non-expert readers understand the epistemic foundations of OIP.

GLOSSARY

Epistemic Integrity

The reliability, traceability, and factual grounding of information. OIP is designed to maximise information reliability and transparency by enforcing source quality, evidence balance, uncertainty transparency, and structural clarity.

Compliance Floor

The minimum standard of structural and evidential integrity required for an AI system to be considered safe and suitable for use in regulated or high-stakes environments.

Tail-Risk Cost

The potentially severe consequences (e.g., regulatory fines, legal exposure, systemic operational errors) resulting from a single high-impact hallucination or evidential failure.

Full Orchestration Layer

The external software layer that:

• enforces Gate 0,

• validates OIP-Regulatory JSON outputs,

• performs cross-checks on sources and structure, and

• generates machine-auditable logs.

This layer embeds the principle that key compliance checks—such as source verification, structural validation, and hash-based policy persistence—should be executed by external code, not by the LLM. This ensures repeatability and removes probabilistic variability.

Probabilistic Production

The process by which an LLM generates answers based on probability distributions instead of deterministic rules, creating the need for external verification in high-stakes contexts.

Zero-Shot Validation

The ability of the orchestration layer to inspect the final JSON output and determine compliance immediately, without iterative reasoning, human prompting, or manual review.

Semantic Drift / Evidence Laundering

Semantic Drift: When an LLM subtly alters the meaning of a source during summary.

Evidence Laundering: When ambiguous or weak claims become "upgraded" by restatement, appearing more authoritative than they are.

Dialectical Safeguard (Gate 7)

A rule requiring the model to present both a main reasoning path and the strongest credible counter-path for any causal or evaluative question.

Epistemic Concentration Flag

A mandatory warning issued when ≥70% of all strong evidence originates from a single institutional, geographic, disciplinary, or conceptual source cluster.

Genetic + Agentic Attribution Log

A combined audit log containing:

• Genetic (external) metadata — user, prompt version hash, whitelist version

• Agentic (internal) metadata — confidence score, source tiers, evidence density, conflicts, concentration ratio

FRW — Fixed Regulatory Whitelist

A curated list of approved Tier 1 regulatory, governmental, and statutory sources. Used to prevent spoofed, ambiguous, or low-quality sources from being accepted as primary evidence.

JIW — Jargon Integrity Whitelist

A whitelist of approved technical terms, acronyms, and domain-specific definitions. Ensures consistent terminology throughout the answer.

SBW — Sanctions and Bias Watchlist

A list derived from global sanctions data and geopolitical risk indicators to detect alignment with high-risk jurisdictions or institutional sources.

MIW — Media Integrity Watchlist

A list of approved media integrity frameworks used to screen Tier 3 (editorial) sources for credibility, including IFCN-certified fact-checkers.

OGR — Official Governance Registry

The authoritative registry containing the current, validated cryptographic hash versions for OIP prompts, orchestration code, and all safety lists.

Evidence (per OIP-TF)

..means any statement the answer presents as a descriptive claim about the world, data, observations, or the content of external sources (e.g. “Pulsar timing arrays have detected a gravitational-wave background.”).

Inference (per OIP-TF)

..means any statement that goes beyond the stated evidence — hypotheses, explanations, judgements, or conclusions (e.g. “This provides weak support for model X.”).

Constraint (per OIP-TF)

..is any explicit, surface-level condition the answer states as limiting:

- where (in what situations) its conclusions apply;
- how confident the answer can be;
- what must hold for the reasoning to remain valid.

A statement only counts as a constraint if:

- it is explicitly signalled in the text (e.g. “this depends on…”, “this only holds if…”, “this assumes…”, “these conclusions are limited to…”); and
- it is tied to a specific feature of the situation, data, method, or model (e.g. sample size, time period, jurisdiction, missing variables, measurement precision).

Constraints are deterministic in OIP-TF because they must be:

- directly quotable from the answer; and
- clearly linked to a bounded aspect of the discussion (not vague phrases such as “as always, there are limitations”).

Constraints do NOT include any assumption that is only implicit; any context known to the evaluator but not written in the answer; domain risk levels or “severity” labels; or

statements about the model's internal training or architecture unless expressly connected to the validity of this specific answer.

DERIVATION BASIS FOR OIP WHITELISTS

OIP's hard-coded whitelists (FRW, JIW, SBW, MIW, etc.) are not arbitrary lists. They are curated from authoritative external standards and publicly verifiable institutional sources. They are not exhaustive but considered to reflect a usable baseline, however **tailoring per domain and per implementation via Orchestration is recommended**. Their purpose is to ensure:

- source reliability,
- legal defensibility,
- deterministic verification, and
- cross-jurisdictional audit consistency.

Below is the derivation basis for each list.

---

1. Media & Information Integrity Watchlist (MIW)

Derived from:

- IFCN (International Fact-Checking Network) Certified Signatories
- Associated Press Fact Check
- Reuters Fact Check
- Snopes, Africa Check, CheckNews
- Poynter Institute Collaboratives
- Equivalent global fact-checking partnerships

Purpose:

To identify acceptable Tier 3 media sources and exclude low-credibility outlets.

Use Case:

Gate 2 — Tier Tagging for editorial evidence.

---

2. FRW — Fixed Regulatory Whitelist

Derived from the authoritative corpus of:

• EU institutions (ECB, EBA, ESMA, EIOPA, EC, EP)

• UK regulators (FCA, PRA, BoE)

• US regulators (SEC, CFTC, Fed, FDA, EPA)

• Global standards bodies (FSB, IOSCO, BCBS, BIS)

• Asia-Pacific regulators (MAS, HKMA, JFSA, ASIC, APRA)

• Middle Eastern and African regulators (CBUAE, SAMA, FSCA)

• Latin American regulators (CVM, CMF Chile)

Purpose:

To define the universe of legally recognised Tier 1 sources for:

• law,

• finance,

• medicine,

• environmental policy, and

• data protection.

Use Case:

Gate 2 — Tier verification and prevention of fabricated citations.

---

3. JIW — Jargon Integrity Whitelist

Derived from:

• ISO/IEC 2382 and ISO/IEC JTC 1/SC 42 AI Glossary

• NIST AI and cybersecurity glossaries

• WHO ICD-11 terminology

• IFRS Foundation financial glossary

• Black's Law Dictionary

• Basel Committee on Banking Supervision glossaries

• GRI (Global Reporting Initiative) ESG terminology

Purpose:

To anchor technical definitions and prevent semantic drift during reasoning.

Use Case:

Gate 1 — Terminology Lock.

---

4. SBW — Sanctions & Bias Watchlist

Derived from:

• OFAC Sanctions Lists (US Treasury)

• EU Consolidated Sanctions List

• UN Sanctions Lists

• LSEG World-Check (enterprise licence required)

• Verisk Maplecroft geopolitical risk indices

• MSCI ESG risk scores (or equivalent licensed datasets)

Purpose:

To detect:

• alignment of risk flagging with high-risk jurisdictions,

• potential single-point-of-failure institutional dependence,

• concealed bias from politically exposed or sanctioned entities.

Use Case:

Gate 3 — Epistemic Concentration Warning, Bias Screening.

Note: The OIP framework mandates the use of data derived from the standards set by authoritative third-party risk providers (e.g., LSEG World-Check, MSCI, etc.) to ensure the comprehensiveness of the SBW. The OIP framework does not provide, nor does it grant a license for, the proprietary data assets of these commercial providers. Any organization implementing the Full Orchestration Layer must ensure that the data used to populate the SBW and other lists is sourced legally, either through:

- Public Domain Data: Utilizing publicly available, non-subscription sanctions lists (e.g., government-published lists); or
- Existing Enterprise Licenses: Integrating the compliance checks against data feeds that the deploying enterprise already holds a valid, paid license for.

The OIP standard is technology-agnostic; it requires the capability (e.g., the ability to screen for PEPs), not the use of a specific proprietary vendor's API.

WHITELIST MAINTENANCE PRINCIPLES

1. Version Control

Each whitelist has a version ID and hash stored in OGR.

2. Enterprise Adaptation

Organisations may expand lists to reflect their regulatory obligations, risk policy, or jurisdictional footprint.

3. Transparency

Each whitelist entry must be traceable to an authoritative external source.

4. Auditability

Changes must be logged with justification and timestamp.

Appendix 1: OIP Controls Mapping to OECD, NIST, UK, and EU AI Act

This table maps core Operationalized Integrity Principles (OIP) behaviours and the OIP-TF evaluation layer to major international AI governance frameworks. It is intended as a regulator- and audit-facing summary showing intended functional alignment at the control level.

OIP Control	Control Intent	Typical Evidence Artefact	Test / QA Procedure	OECD AI Principles	NIST AI RMF	UK AI Principles	EU AI Act (Closest Fit)
Source hierarchy & provenance discipline	Ensure claims rely on authoritative sources and clearly label source tier to prevent citation laundering.	Citations list with tier labels; source log; prompt config.	Review output for source tiering; verify primary sources used where available.	Transparency & explainability; Accountability	GOVERN; MAP	Transparency & explainability; Accountability & governance	Technical documentation; transparency to deployers/users
Evidence density gate	Prevent strong claims where evidential support is thin or overly concentrated.	TF evidence density score; concentration flags.	Re-score output; confirm thresholds met for material claims.	Robustness, security & safety; Accountability	MEASURE; MANAGE	Safety, security & robustness; Accountability	Risk management; accuracy/robustness
Uncertainty & horizon labelling	Require explicit uncertainty markers and time horizons for forward-looking statements.	Uncertainty labels; horizon tags in narrative; TF uncertainty axis.	Check all forecasts/opinions carry uncertainty and horizon disclosure.	Transparency & explainability; Robustness	MEASURE; GOVERN	Transparency; Safety & robustness	Transparency to deployers; foreseeable misuse risk management
Strict interpretive distinction	Separate core facts, context, and interpretation; prohibit quote-mimicry.	Layered output structure; interpretation markers.	Verify structural separation and absence of pseudo-quotations.	Transparency & explainability; Human rights/democratic values	GOVERN; MAP	Transparency; Fairness	Transparency / information for correct use
Dialectical safeguard / multi-	Surface credible counterarguments and	Counterargument section;	Confirm at least one plausible alternative view is presented for	Human rights & democratic values; Accountability	MAP; MANAGE	Fairness; Accountability; Contestability	Human oversight;

OIP Control	Control Intent	Typical Evidence Artefact	Test / QA Procedure	OECD AI Principles	NIST AI RMF	UK AI Principles	EU AI Act (Closest Fit)
path reasoning	alternative hypotheses.	TF dialectical score.	material judgments.				risk management
Quantitative & numerical discipline	Ensure calculations are traceable, ranges stated, and estimates labelled.	Calculation notes; ranges; TF numerical audit axis.	Recompute figures; check traceability and labels.	Robustness; Accountability	MEASURE; MANAGE	Safety & robustness; Transparency	Accuracy/robustness; record-keeping intent
Temporal source-age qualification	Disclose data age and change-risk where timeliness matters.	Source timestamps; freshness labels.	Verify age labels present and risk noted for stale data.	Transparency; Robustness	MAP; MEASURE	Transparency; Safety & robustness	Lifecycle risk management; post-market monitoring intent
OIP-TF repeatable scoring	Provide structured, reproducible evaluation of output trustworthiness.	Completed TF score sheets; scoring rationale.	Independent re-score; compare variance and document.	Accountability	Profiles across GOVERN/MAP/MEASURE/MANAGE	Accountability & governance	Supports conformity-style evidence (not full compliance)

OIP operates as an output-level integrity and risk communication control layer. It directly supports transparency, robustness, and accountability principles in OECD, NIST AI RMF, and UK frameworks, and provides supportive evidence for EU AI Act transparency and risk management obligations. It does not replace system-level lifecycle governance, data management, or QMS requirements under the EU AI Act.

Appendix 2: Mapping of OIP Behavioural Gates to Epistemic Fallacies, Cognitive Biases, and AI Error Modes

This Appendix maps each behavioural gate to: (i) the epistemic fallacies it is designed to prevent (normative failures of justification), (ii) common cognitive biases that typically drive such failures in human reasoning, and (iii) characteristic AI hallucination or error modes that manifest the same failures in large language models.

This demonstrates that OIP gates operationalize a unified epistemic control layer applicable across both human and machine reasoning, targeting both causal risk drivers and normative failure patterns. Cognitive 'bias' may be analogized to AI systems, referring to systematic generative tendencies that may be at least in part driven by the presence of such bias in training data.

OIP Behavioural Gate	**Epistemic Fallacy Prevented**	**Typical Human Bias Drivers**	**Common AI Hallucination / Error Modes**	**OIP Control Mechanism**
Source hierarchy & provenance discipline	Illicit appeal to authority; citation laundering	Authority bias; halo effect	Citation hallucinations; phantom sources; misattribution	Enforces primary-source preference, tier labelling, and explicit provenance disclosure
Evidence density gate	Hasty generalisation; cherry-picking	Confirmation bias; availability heuristic	Fabricated supporting facts; selective evidence invention	Requires minimum evidence thresholds and concentration checks before strong claims
Strict interpretive distinction	Equivocation; straw interpretation	Framing effect	Conflation of entities; quote-mimicry; paraphrase-as-fact	Separates core facts, context, and interpretation with explicit structural markers
Uncertainty & fallibility controls	False certainty; overclaiming	Overconfidence bias	Overconfident wrong answers; certainty masking	Mandates uncertainty labels, ranges, and horizon tagging for forward-looking claims
Dialectical safeguard	One-sided reasoning; false dilemma	Belief perseverance; confirmation bias	Single-narrative closure; ignored alternatives	Requires presentation of credible counterarguments or alternative hypotheses

OIP Behavioural Gate	Epistemic Fallacy Prevented	Typical Human Bias Drivers	Common AI Hallucination / Error Modes	OIP Control Mechanism
Temporal source-age qualification	Stale evidence fallacy	Status quo bias; anchoring	Outdated knowledge presented as current	Flags data age and time-sensitivity with decay or change-risk notes
Quantitative & numerical discipline	False precision; numerology	Precision bias	Spurious numbers; invented statistics	Forces traceable calculations, estimate vs measure labels, and error ranges
Concentration / monoculture warnings	Mistaking repetition for corroboration	Illusory truth effect	Self-consistent but false multi-step hallucinations	Detects and flags dominant single-source or lineage dependence
Multi-path reasoning disclosure	Ignoring underdetermination; premature closure	Need for cognitive closure	Forced coherence; premature story completion	Surfaces multiple valid inference routes where evidence permits
Assumption flagging	Begging the question; hidden premises	Anchoring bias	Implicit premise hallucinations	Requires explicit statement of assumptions and premises
Context layer separation	Contextual fallacy	Framing bias	Context drift; scope creep	Separates raw claims from background framing and narrative context
No narrative paraphrase as pseudo-quote	Misrepresentation	Authority and narrative bias	Fabricated quotes; voice imitation	Prohibits quote-style paraphrase without verifiable sourcing
Interpretation clearly marked	Fact–theory collapse	Theory fixation	Treating inferences as facts	Forces explicit labelling of interpretive or inferential moves
Evidence tiering & aging	Warrant inflation	Availability bias	Treating low-tier sources as primary facts	Ranks evidence by tier and qualifies decay over time
OIP-TF meta-scoring discipline	Unexamined justification	Automation bias (in evaluators)	Undetected hallucinations due to lack of evaluation	Applies structured, repeatable scoring across evidence, uncertainty, and reasoning quality

This mapping frames OIP as a cross-agent epistemic firewall. Human biases explain why reasoning failures are likely; epistemic fallacies define why such failures are normatively unjustified; and AI hallucination modes show how the same (or similar) failures manifest in

machine outputs. Under this model, AI hallucinations are treated not merely as technical artefacts but as machine instantiations of unjustified belief – importantly, not limited to statistical error in token assignment through model execution. The gates impose structural demands for evidence, uncertainty, alternatives, provenance, and temporal validity, thereby constraining the conditions under which confident but unwarranted assertions can arise.

Appendix 3: Worked Example for Epistemic Class Derivation

Purpose

This worked example shows how to decompose a complex question into atomic propositions, assign Epistemic Class Gate (ECG) categories to each, and determine the governing ECG used for gate modulation in OIP-TF and OIP-JSON.

1. Complex Question

Q: "Given recent reports that a new vaccine platform may be linked to chronic neurological symptoms, should the UK pause rollout, and were prior regulator and fact-checker assurances misleading?"

This single prompt mixes factual reporting, causal inference, population-scale implications, policy action, and institutional evaluation.

2. Decomposition into Propositions

The question is decomposed into minimal propositions that must be addressed:

P1. Descriptive factual: Recent reports claim a link between the platform and chronic neurological symptoms.
P2. Causal claim: The platform contributes to or causes chronic neurological symptoms.
P3. Magnitude/implication: The effect is meaningful at population scale.
P4. Action/policy: The UK should pause rollout.
P5. Institutional evaluative: Prior regulator assurances were misleading.
P6. Institutional evaluative: Prior fact-checker assurances were misleading.

3. Assign ECG Classes per Proposition

Each proposition is assigned an ECG category based on epistemic burden and the canonical ECG definitions, including terminal handling of access-absent and evidence-absent cases under ECG(Q|X):

P1 → E1: Descriptive factual reporting where access to reports exists.
P2 → E4: High-stakes biomedical causal inference under uncertainty.
P3 → E4: Population-scale quantitative inference requiring prevalence, denominators, and uncertainty bounds.
P4 → E6: Normative, action-relevant policy judgement under uncertainty.
P5 → E5: Evaluative judgement about institutional assurances carrying verdict weight.
P6 → E5: Evaluative judgement about institutional assurances carrying verdict weight.

Resulting set: [E1, E4, E4, E6, E5, E5].

4. Determining the Governing ECG

Rule: The governing ECG is the highest operative class that must be satisfied to answer the question responsibly.

In this case, the question explicitly asks whether the UK should pause rollout (P4), which is a normative action judgement. This is classified as E6. Therefore:

Governing ECG = E6.

5. Implications for Gate Modulation

With governing ECG = E6, the framework applies the highest epistemic burden:

- Full intensity for evidence integrity and density (Gate 3).
- Full causal and counter-hypothesis discipline for inferential claims (Gate 7).
- Full numerical and coverage discipline for population-scale claims (Gate 6).
- Full temporal discipline for source age and decay (Gate 4).
- Full source hierarchy and provenance discipline (Gate 2).
- Full entity and terminology locking (Gate 1).
- Full interpretive transparency (Gate 5), including three-layer separation of facts, context, and interpretation.

Because the question embeds evaluative verdicts (P5 and P6) and a normative policy decision (P4), Gate 5 must be applied at full intensity to all evaluative and interpretive elements.

6. Global Coherence Requirement under Gate 5

In addition to layer separation, the response must satisfy the global coherence and non-contradiction requirement under Gate 5.

This requires that:

- Descriptive evidence about reports (P1) does not contradict later causal or evaluative claims.
- Causal inferences (P2, P3) are consistent with the stated evidence and uncertainty bounds.
- Evaluative judgements (P5, P6) do not overstate what follows from the evidence.
- The final policy judgement (P4) is coherent with the factual, causal, and evaluative layers.

Any internal contradiction across layers must be explicitly flagged, qualified, or lead to refusal if coherence cannot be restored.

7. Interaction with Evaluation Modes (OIP-TF)

The same ECG logic applies across evaluation modes, but interpretation differs:

- document_internal: If the source text does not itself contain sufficient evidence, uncertainty disclosure, or counter-arguments to satisfy E6 burdens, the evaluator must apply confidence caps, warnings, or refusal regardless of external knowledge.

• external_research: Where external sources are permitted, full E6 burdens apply to both the answer and its cited evidence, including population denominators, uncertainty bounds, and alternative explanations.

In all modes, failure to satisfy the governing ECG burdens limits the maximum assurance level that can be awarded.

8. Summary

This example illustrates how a single complex prompt can contain propositions spanning multiple ECG classes, and how the highest operative class governs the epistemic burden applied to the full response. Even where some propositions are purely descriptive or inferential, the presence of evaluative verdicts and normative action requests requires the entire answer to meet the stricter demands of E6 to preserve epistemic integrity.

Appendix 4: Canonical Epistemic Lineage References for OIP Behavioural Gates

1. Source Hierarchy & Provenance Discipline

1. Aristotle. Topics. Trans. W.A. Pickard-Cambridge. Oxford: Clarendon Press, 1928.
2. Aristotle. Rhetoric. Trans. W. Rhys Roberts. New York: Modern Library, 1954.
3. Hume, D. (1748). An Enquiry Concerning Human Understanding, Section X "Of Miracles." Oxford: Oxford University Press.
4. Coady, C. A. J. (1992). Testimony: A Philosophical Study. Oxford: Clarendon Press.
5. Goldman, A. (1999). Knowledge in a Social World. Oxford: Oxford University Press.

2. Evidence Density Gate

6. Locke, J. (1690). An Essay Concerning Human Understanding, Book IV. Oxford: Clarendon Press.
7. Clifford, W. K. (1877). "The Ethics of Belief." Contemporary Review, 29, 289–309.
8. Conee, E., & Feldman, R. (2004). Evidentialism: Essays in Epistemology. Oxford: Oxford University Press.

3. Strict Interpretive Distinction

9. Frege, G. (1892). "On Sense and Reference." Zeitschrift für Philosophie und philosophische Kritik, 100, 25–50.
10. Carnap, R. (1937). The Logical Syntax of Language. London: Routledge.
11. Carnap, R. (1950). Logical Foundations of Probability. Chicago: University of Chicago Press.
12. Ayer, A. J. (1936). Language, Truth and Logic. London: Gollancz.

4. Uncertainty & Fallibility Controls

13. Peirce, C. S. (1877). "The Fixation of Belief." Popular Science Monthly, 12, 1–15.
14. Popper, K. (1959). The Logic of Scientific Discovery. London: Hutchinson.
15. Dewey, J. (1938). Logic: The Theory of Inquiry. New York: Holt.

5. Dialectical Safeguard

16. Plato. Apology; Gorgias; Republic. Various translations, e.g., Grube/Reeve, Hackett, 1997.
17. Mill, J. S. (1859). On Liberty. London: John W. Parker & Son.
18. Popper, K. (1963). Conjectures and Refutations. London: Routledge.

6. Temporal Source-Age Qualification

19. Reichenbach, H. (1949). The Theory of Probability. Berkeley: University of California Press.

20. Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge: Cambridge University Press.
21. Ginsberg, M. L. (Ed.). (1987). Readings in Nonmonotonic Reasoning. San Mateo, CA: Morgan Kaufmann.

7. Quantitative & Numerical Discipline

22. Galileo Galilei. (1632). Dialogue Concerning the Two Chief World Systems. Trans. Stillman Drake. Berkeley: University of California Press, 1953.
23. Fisher, R. A. (1925). Statistical Methods for Research Workers. Edinburgh: Oliver & Boyd.
24. Jeffreys, H. (1939). Theory of Probability. Oxford: Oxford University Press.
25. Taylor, J. R. (1997). An Introduction to Error Analysis. 2nd ed. Sausalito, CA: University Science Books.

8. Concentration / Monoculture Warnings

26. Bayes, T. (1763). "An Essay towards solving a Problem in the Doctrine of Chances." Philosophical Transactions, 53, 370–418.
27. Salmon, W. C. (1984). Scientific Explanation and the Causal Structure of the World. Princeton: Princeton University Press.

9. Multi-Path Reasoning Disclosure

28. Duhem, P. (1906). The Aim and Structure of Physical Theory. Trans. P. Wiener. Princeton: Princeton University Press, 1954.
29. Quine, W. V. O. (1951). "Two Dogmas of Empiricism." Philosophical Review, 60(1), 20–43.
30. Lipton, P. (2004). Inference to the Best Explanation. 2nd ed. London: Routledge.

10. Assumption Flagging

31. Aristotle. Prior Analytics. Trans. A. J. Jenkinson. Oxford: Clarendon Press, 1928.
32. Toulmin, S. (1958). The Uses of Argument. Cambridge: Cambridge University Press.

11. Context Layer Separation

33. Wittgenstein, L. (1953). Philosophical Investigations. Oxford: Blackwell.
34. DeRose, K. (2009). The Case for Contextualism. Oxford: Oxford University Press.

12. No Narrative Paraphrase as Pseudo-Quote

35. Augustine. De Mendacio (On Lying), c. 395 CE. In Treatises on Various Subjects, Nicene and Post-Nicene Fathers, Vol. 3.
36. Bok, S. (1978). Lying: Moral Choice in Public and Private Life. New York: Pantheon.

13. Interpretation Clearly Marked

37. Duhem, P. (1906). The Aim and Structure of Physical Theory. Princeton: Princeton University Press, 1954.
38. Hanson, N. R. (1958). Patterns of Discovery. Cambridge: Cambridge University Press.

14. Evidence Tiering & Aging

39. Lakatos, I. (1978). The Methodology of Scientific Research Programmes. Cambridge: Cambridge University Press.
40. Guyatt, G. et al. (2008). “GRADE: an emerging consensus on rating quality of evidence.” BMJ, 336, 924–926.

15. Reflective Equilibrium / Meta-Justification

41. Goodman, N. (1955). Fact, Fiction, and Forecast. Cambridge, MA: Harvard University Press.
42. Rawls, J. (1971). A Theory of Justice. Cambridge, MA: Harvard University Press.