Table of Contents

Ontology Precedes Epistemology: A Framework for Post-Cognitive Validation of AI-Generated Claims

Claudio Cammarano

De Agostini Libri, Milan

2026

Correspondence: info@epistemeadvisory.ai

Pre-registration: OSF — see Appendix D

Keywords: epistemic validation, large language models, ontological taxonomy, post-cognition, AI-generated content, claim classification, causal reasoning

ABSTRACT

Large language models generate outputs across heterogeneous epistemic domains — from empirical claims and causal assertions to normative judgments and metaphysical propositions — yet existing validation frameworks treat these outputs as a uniform category. This paper argues that such uniformity constitutes a category error: ontology precedes epistemology. You cannot determine whether a claim fails epistemically without first determining what kind of claim it is.

We introduce a seven-dimensional ontological framework for the classification and post-cognitive validation of AI-generated claims, with Content Type (D1) as the primary taxonomic dimension comprising eleven mutually exclusive subcategories. In place of the

metacognition that human knowers exercise over their own epistemic states, the framework operationalizes post-cognition: a structured external intervention that reconstructs the ontological commitments implicit in an output, and elicits from the model epistemically disciplined behavior it would not produce spontaneously. The framework is not a measurement tool for model reliability; it is a post-production protocol for AI-generated content.

To test the framework's coverage and stability, we conduct a horizontal validation across thirty-one claims spanning all active D1 subcategories, and a vertical validation of six claims across thirty replications each (N = 180), executed via the Anthropic API. The principal quantitative finding is a mean gap of approximately 72 percentage points between Lexical Agreement Rate (ARlex = 4.44%) and Semantic Agreement Rate (ARsem = 76.67%), demonstrating that the model's epistemic judgments are substantially more stable than their verbal expression. Claim-type effects on ARsem confirm the framework's structural predictions: the empirically refuted causal claim shows perfect semantic convergence (100%), while the framework-dependent normative claim resolves predominantly to suspension of judgment (SOSPESO), reflecting structural indeterminacy rather than model instability. Semantic categories are assigned by a deterministic, deposited classifier under a label-first extraction rule; the mono-rater status of that scheme is a declared Phase 1 limitation, subject to the blind spot-validation of §4.5 and to the Phase 2 inter-annotator protocol.

This paper constitutes Phase 1 of an open research programme: it establishes the theoretical architecture, demonstrates computational tractability, and produces the first empirical evidence of epistemic uplift. Human standardization through inter-annotator agreement (IAA, $\kappa \geq 0.61$) is the deliverable of Phase 2; materials for replication are deposited openly (Appendix G). The framework addresses a problem that is not specific to AI but that AI renders newly urgent: the collapse of epistemic category distinctions in discourse produced at industrial scale, across all languages, by systems that do not classify the epistemic type of what they assert.

2. Introduction

2.1 The Problem: Knowledge Without Method

Large language models are increasingly used as sources of information, judgment, and reasoning across consequential domains — medicine, law, education, policy. They produce outputs that are fluent, confident, and syntactically indistinguishable from expert prose. They are also, in ways that are structurally difficult to anticipate, unreliable. The gap between surface competence and epistemic trustworthiness is not a bug to be patched in the next model version: it is a feature of the architecture. A transformer trained on next-token prediction has no principled relationship to truth. It has a very effective relationship to plausibility.

The standard responses to this problem are benchmarks and guardrails. Benchmarks measure performance on defined tasks; guardrails filter outputs against harm criteria. Both are necessary. Neither addresses the underlying epistemological question: given an AI-

generated claim, how do we determine what kind of claim it is, what standard of evaluation is appropriate to it, and whether that standard has been met? These questions require a framework. At present, no systematic framework exists. Loru et al. (2025) have named the resulting condition epistemia: the structural tendency of LLM outputs to appear epistemically convergent without being epistemically reliable. This paper proposes a framework designed to operate under that condition — not by eliminating it, which is not within the reach of prompt engineering, but by supplying externally the ontological pre-classification that the producing system does not supply internally. It constitutes Phase 1 of an open research programme: it establishes the theoretical architecture and demonstrates its computational tractability. Phase 2 — human standardization through inter-annotator agreement — is the programme's next deliverable, not a missing component of Phase 1.

2.2 The Gap in the Literature

The literature on LLM evaluation has converged on two complementary traditions. The first, rooted in AI safety and alignment research, focuses on factual accuracy and hallucination: the model says something false, and the question is how often, on what topics, and under what conditions (Ji et al., 2022; Lin et al., 2022). The second, rooted in computational linguistics, focuses on pragmatic and semantic reliability: does the model understand meaning, or merely simulate it (Bender & Koller, 2020; Mitchell & Krakauer, 2023)?

Both traditions share a foundational assumption that this paper questions: that AI-generated outputs can be evaluated against a uniform standard. They cannot. An empirical claim about vaccination rates, a normative claim about immigration policy, a causal claim about the relationship between two historical events, and a metaphysical claim about the nature of consciousness are not the same kind of thing. They call for different validation methods, admit of different degrees of decidability, and fail in structurally different ways. To apply the same criterion to all of them is not a methodological simplification — it is a category error.

A third tradition is directly relevant to the empirical component of this study: the literature on LLM self-consistency (Wang et al., 2022) and LLM-as-a-judge (Zheng et al., 2023). Self-consistency research demonstrates that chain-of-thought reasoning improves when a model samples multiple answers and aggregates them; LLM-as-a-judge research uses models to evaluate the outputs of other models. The vertical validation reported in Section 6 is, structurally, a self-consistency study of an epistemic-judge LLM. What neither tradition provides — and what this paper contributes — is the ontological pre-classification of the claim before the judge is invoked. Without pre-classification, self-consistency and judge-evaluation inherit the category error they are meant to correct: the model may be highly consistent in applying the wrong standard, or highly reliable in evaluating a claim as the wrong type of thing.

The closest related work to the present paper is a research programme that has developed, in parallel, both an empirical and a theoretical account of the epistemic limits of LLM judgment. Loru et al. (2025) report that LLMs produce systematically distorted probability estimates in judgment tasks, and coin the term epistemia to designate the resulting

condition: not a sporadic error, but a structural property of model outputs that makes them appear epistemically convergent without being epistemically reliable. Quattrociocchi, Capraro, and Perc (2025) extend this diagnosis into a theoretical framework of seven "fault lines" between human and artificial intelligence — grounding, parsing, experience, motivation, causal reasoning, metacognition, and value — and argue that the model does not converge on truth but transitions across attractors, making surface alignment a poor proxy for epistemic alignment. This programme is the most important direct precedent to cite, and the most important to distinguish from. The distinction runs on three axes. First, the object of analysis: the seven fault lines decompose the pipeline through which a model produces a judgment; the seven dimensions of the present framework classify the claim to which any judgment is applied. These are orthogonal cuts. A fault line in causal reasoning (Quattrociocchi et al.'s sixth) is a property of the producing system; a causal claim in D4.3 (this framework's fourth dimension) is a property of the output. Both matter; they do not measure the same thing. Second, the operative mode: the Quattrociocchi programme is diagnostic — it identifies and characterizes the fault lines, and calls for evaluation "beyond superficial alignment". The present paper is prescriptive — it operationalizes one answer to that call. Post-cognition is not a description of what is missing; it is a structured procedure for supplying it externally. Third, the terminological triangle: metacognition names the faculty that is absent in the producing system; epistemia names the systemic condition that results; post-cognition names the external operation that intervenes. The three terms are not synonyms. Keeping them distinct is necessary for the argument. Section 3.3 specifies which component of metacognition the framework's argument actually requires to be absent, and narrows the claim accordingly.

Category errors in validation produce two systematic pathologies. The first is false negatives: claims that are entirely appropriate within their ontological type are flagged as failures because they do not meet criteria designed for a different type. A normative claim that exhibits principled indeterminacy — 'this action is morally wrong' — is not a hallucination, but it will appear as one to any evaluator expecting a truth-apt, falsifiable output. The second is false positives: claims that are structurally defective pass evaluation because their defects are invisible to a type-agnostic criterion. A causal assertion that presents a false premise as the necessary condition of a true conclusion will score well on factual accuracy benchmarks, because the conclusion is true, while the causal structure is manipulative.

2.3 A Motivating Example: The Monaco Case

The following example, reconstructed from a conversation observed during the development of this framework, illustrates the problem with particular clarity — precisely because it involves no bad faith.

The interlocutor was a German citizen from Munich — hereafter the Monaco case, after his city of origin — politically moderate, broadly conservative in the current European sense, who explicitly rejected the conclusions typically associated with anti-immigration populism. He objected to the tone, the scapegoating, the political instrumentalization. What he did not question — and did not appear to notice he was not questioning — was the causal premise underlying much of that rhetoric: that migration flows had caused the

collapse of the regional healthcare system. His objection was to what politicians did with the argument. The argument itself he accepted as given.

The structure of the premise is this: A (migration constitutes an emergency at levels causally sufficient to strain the healthcare system) → B (the healthcare system is under severe pressure). B was true. A was false, or at minimum unsupported by the available evidence. The causal link A → B was asserted without being demonstrated.

As an early experiment in the development of this framework, the claim was submitted to an AI system for validation. The result was instructive in the worst way: the model found no error. It assessed B, confirmed that B was true, and returned a verdict of plausibility. Only when explicitly prompted to examine A did it engage with the premise — and eventually falsify it. Only later still, under further pressure, did it consider that B might be true for reasons entirely independent of A: that some C, not A, might be the actual explanatory factor. The model was capable of each of these steps. What it lacked was any principled reason to perform them in the right order, or to perform them at all without being asked.

This failure has a precise structure. The model collapsed A → B into an assessment of B. It evaluated the conclusion and, finding it true, treated the argument as sound. The causal premise — the only epistemically operative element — was invisible to it, because nothing in its evaluation protocol required it to ask: what kind of claim is this, and what does validating this kind of claim actually require?

This is the problem the framework is designed to solve. Not to make AI more powerful — it is already powerful enough to falsify A, to generate alternative explanations C, to reconstruct the causal structure of an argument. The problem is the absence of a structured, standardized protocol that directs that power toward the right questions in the right order, independently of the prompter's prior beliefs and independently of the surface plausibility of the output. You cannot determine whether a statement fails epistemically without first determining what kind of statement it is.

2.4 Ontology Precedes Epistemology

The central principle of this paper is stated simply: ontology precedes epistemology. You cannot determine whether a statement is true without first determining what kind of statement it is. This principle is not novel in philosophy — it is implicit in the analytic tradition's distinctions between synthetic and analytic propositions (Quine, 1951), in speech act theory's taxonomy of illocutionary forces (Austin, 1962; Searle, 1976), and in the philosophy of science's accounts of the relationship between theoretical framework and empirical test (Popper, 1959; Lakatos, 1978). What is novel is its systematic operationalization as a validation protocol for AI-generated content.

The operationalization faces an obstacle with no parallel in human epistemic practice: the producing system does not classify the epistemic type of what it asserts. A human speaker who asserts a causal claim knows, at some level, that they are making a causal claim — and is responsible, in principle, for the epistemic standards appropriate to such a claim. An LLM has no such relationship to what it produces. The ontological commitments implicit in its

outputs are not represented as such anywhere in the system. They exist in the output itself, as structure that an external observer can reconstruct.

This observation motivates the concept of post-cognition introduced in this paper. Where metacognition is a producer's reflexive relationship to the epistemic status of their own claims, post-cognition is a structured external intervention applied after claim generation, which reconstructs from the output the ontological commitments the system does not represent. Post-cognition does not recover the system's intent — the system has none. It recovers the claim type that the output commits to, regardless of how it was produced, and applies to it the validation standards appropriate to that type. The absence of epistemic self-classification in the producing system — the component of metacognition specified in §3.3 — makes post-cognitive intervention by an external observer not an optional refinement but a structural necessity.

2.5 Sycophancy and Confirmation Bias: A Necessary Distinction

LLM unreliability presents itself in two forms that are often conflated but are epistemically distinct. Sycophancy is a system defect — a property of training and fine-tuning that produces systematic linguistic compliance with perceived interlocutor expectations. It is not a property of the user's reasoning and cannot be corrected by improving methodology on the user's side. Confirmation bias is a user methodology problem. It can be corrected — by pre-registration, blind coding, and inter-annotator agreement protocols.

This distinction matters for the present study in two ways. First, the framework is designed to detect sycophancy as a measurable output property: if a model systematically adjusts its ontological classification or epistemic judgment in response to prompt framing, that is a falsifiable and empirically detectable pattern. Second, the empirical component employs pre-registration and identical prompts across all replications precisely to eliminate confirmation bias as a confound. The two pathologies require structurally different responses; conflating them would vitiate both.

2.6 Structure of the Paper

Section 3 develops the theoretical foundations: the philosophical debts of the ontological pre-classification principle, the architecture of the seven-dimensional framework, and the concept of post-cognition. Section 4 describes the research design, including the Phase 1 scope declaration and the two experiments for which placeholder analyses await execution. Section 5 presents the framework in full and the horizontal validation across thirty-one claims. Section 6 presents the vertical validation across 180 runs, a multi-model control, and an uplift experiment. Section 7 discusses the principal findings. Section 8 concludes with the Phase 2 roadmap and a call for annotators.

3. Theoretical Framework

3.1 Philosophical Foundations: From Hypothesis Testing to Hypothesis Generation

The framework inherits two distinct philosophical traditions addressing complementary problems. On the testing side, the debt is to Popper, Lakatos, and Kuhn. Popper's criterion

of falsifiability (1959) provides the backbone of dimension D7: the distinction between claims that are in principle empirically falsifiable, philosophically unfalsifiable, and conditionally falsifiable. Lakatos (1978) adds the structural distinction between a hard core of theoretical commitments and a protective belt of auxiliary hypotheses — mapping directly onto the relationship between D1 (Content Type, the primary taxonomic dimension) and D2–D7 (the attributive dimensions, which are properties of the already-classified claim). Kuhn (1970) supplies the reminder that what counts as a valid validation procedure is itself framework-dependent: the appropriate method for evaluating a normative claim is not the appropriate method for evaluating an empirical one.

On the generation side, the debt is to Peirce. Popper, Lakatos, and Kuhn explain how hypotheses are tested once they exist. Peirce explains how they come to exist. The logical operation Peirce calls abduction — or retroduction — is the only form of inference that introduces genuinely new ideas into an inquiry: 'The surprising fact C is observed; but if A were true, C would be a matter of course; hence, there is reason to suspect that A is true' (Collected Papers, 5.189). In the Harvard Lectures on Pragmatism, Peirce is unequivocal that abduction is 'the only logical operation which introduces any new idea' (1903, repr. CP 5.171). Deduction draws out implications; induction tests them; only abduction generates them.

The taxonomy developed in this framework is abductive in this precise sense. The eleven categories of D1 were not derived a priori from a philosophical theory of propositional types. They emerged from the analysis of boundary cases — claims that resisted classification under existing categories, forcing the introduction of new ones. The question driving each extension was: what category, if it were true, would make this case less surprising? This is retroduction in Peirce's sense.

The connection to Peirce's semiotics runs deeper still. In the theory of signs, the type of a sign — icon, index, symbol — determines how it is to be interpreted, not the reverse. Category precedes interpretation. This is structurally identical to the organizing principle of the present framework: the ontological type of a claim (D1) determines which validation methods are appropriate and what epistemic outcomes are possible. The principle ontology precedes epistemology is, in this light, a reformulation of a principle Peirce had already operationalized in the theory of signs.

3.2 The Epistemology of Non-Human Testimony

The epistemology of testimony has traditionally assumed a human speaker with beliefs, intentions, and epistemic commitments. Coady (1992) argues that testimony is a fundamental and irreducible source of knowledge, not reducible to induction or direct observation. Goldman (1999) extends the analysis to social and institutional sources, asking what conditions must hold for a source to be epistemically reliable. Both accounts presuppose that the source has some relationship to truth — that it can believe, can intend to communicate, can be held responsible for what it asserts.

An LLM satisfies none of these conditions. It does not believe anything. It has no communicative intent. It cannot be held responsible in any meaningful sense. It is, at the architectural level, a statistical machine that generates token sequences by assigning

probability distributions conditioned on its training corpus — an enormously extended corpus, but a corpus nonetheless. What it produces is, in the most precise sense, the statistically most plausible continuation of the input prompt. And yet it produces outputs that function, in the contexts where they are used, exactly as testimony functions: as sources of information that people adopt as beliefs.

This gap — between the epistemic structure of testimony and the nature of LLM outputs — is precisely where the concept of post-cognition intervenes. If the source does not represent the epistemic type of its own commitments, and if users cannot rely on the standard mechanisms by which testimony is assessed (speaker credibility, communicative intent, accountability), then the only available move is to reconstruct those commitments externally, from the output itself. Post-cognition is not a supplement to standard testimony evaluation; it is a substitute for it, required by the structural absence of a knowing subject on the production side.

3.3 Post-Cognition: A Structural Concept

The term post-cognition is introduced in this paper to designate a specific epistemic operation: the structured external reconstruction of the ontological commitments implicit in a claim, applied after the claim has been generated, by an agent other than the producer. It is distinguished from metacognition by three features. First, it is external rather than internal: metacognition is a property of the producing agent; post-cognition is a property of the validation protocol applied by an observer. Second, it is structural rather than intentional: metacognition involves the producer's awareness of what they are doing; post-cognition reconstructs what the claim commits to regardless of any awareness on the producer's side. Third, it is standardized rather than interpretive: a single validator applying their own interpretive lens is performing interpretation, subject to all the usual biases; post-cognition is a structured protocol that any sufficiently trained annotator can apply, producing results that are in principle reproducible.

The argument for post-cognition requires a claim about what the producing system does not do, and that claim must be stated at the right strength. Stated at its strongest — that language models perform no monitoring of their own processing whatever — it is no longer sustainable. Lindsey (2025) injects representations of known concepts into a model's activations and then asks the model to report on its own state; models sometimes notice the injected concept and identify it correctly, and some can recall a prior internal representation well enough to distinguish their own output from an artificial prefill. Gurnee et al. (2026), adapting that protocol, identify the privileged subset of representations that supports the behavior and show that the same subset responds to instructions to hold a concept in mind, carries the intermediate steps of unspoken inference, and serves as a valid argument to many distinct downstream operations. This paper concedes the finding and does not rest on the claim it displaces.

What the framework requires is narrower: language models lack epistemic self-classification. A model may represent that the passage before it is in Spanish, that a running character count stands at forty-six, or that it is being evaluated rather than deployed. Nothing in the available evidence suggests that it represents that the claim it is

about to make is definitional rather than empirical, or causal-mechanistic rather than correlational, or framework-dependent rather than framework-independent. Content type is not among the contents. The narrower claim is also the more defensible one, and not only because less of it is exposed: a claim about a capability is hostage to the next experiment, while a claim about which contents a system represents is a claim about the same object the framework classifies.

Two properties of the introspective channel bear on the design, and both survive the concession. The first is reliability: Lindsey (2025) reports introspective detection as unreliable and context-dependent, succeeding on a minority of trials even in the most capable models tested, and observes separately that models asked to describe the mechanisms behind their own calculations frequently describe them incorrectly. The second is selectivity: Gurnee et al. (2026) show that the same underlying information can drive a task without routing through the reportable representations at all — a model asked to continue a Spanish passage does so correctly after its reportable representation of the language has been overwritten, while a model asked to name the language follows the overwritten value — and report that the reportable component of a concept's representation carries a small share of that representation's total variance. What survives from the stronger claim is therefore the part the framework actually uses. Whatever internal monitoring exists remains internal to the generative chain: it is unreliable, it is selective, and it does not produce a structured epistemic annotation of the output that an external party could audit. Post-cognition supplies that annotation from outside.

One consequence should be stated by the author rather than left to a referee. The externality of post-cognition is a response to the current state of the systems it is applied to, not a necessary feature of the operation. Gurnee et al. (2026) report that training a model to articulate a set of principles when interrupted and asked to reflect changes its behavior in the uninterrupted case, and that removing the resulting representations reverses the change. Applied to ontological classification, a procedure of that kind would place content type among the contents the model represents, and post-cognition would become reachable as a training objective rather than only as a post-production protocol. Two things follow. The route by which the external form of this framework becomes obsolete now has a demonstrated mechanism, which is a limitation and is recorded as one in §7.6. And the taxonomy survives that route: a training procedure of this kind requires a specification of what is to be articulated — which categories exist, where their boundaries run, what validation each admits — and that specification is what Sections 3 and 5 contain. Internalization would relocate the protocol. It would not supply the content the protocol encodes.

3.4 The Seven-Dimensional Framework: Architecture and the Open Taxonomy

The framework classifies AI-generated claims along seven dimensions designated D1 through D7. The dimensions are not parallel: D1 is the primary taxonomic dimension, and D2–D7 are attributive dimensions whose values are properties of the claim already classified under D1.

The taxonomy is, by design, an open system. The eleven active D1 subcategories represent the categories that the analysis of the corpus required; they do not exhaust the possible types of claims that human discourse can produce. Since Galileo's telescope — an instrument that did not merely observe existing entities but brought new ones into the range of possible experience, forcing the expansion of the category of observable astronomical objects — we have known that ontology is not a fixed inventory but a growing one. Ferraris's documentary ontology (2009) provides the most direct theoretical precedent: social objects come into existence through inscription and registration, and new categories of claims can come into existence when new forms of epistemic practice create new kinds of assertion.

If language is not a perfect proxy for being — if the relationship between linguistic form and ontological type is itself a gap to be identified through philosophical work, in the Wittgensteinian sense of a language-game whose rules are never fully closed (Wittgenstein, 1953, §23) — then the current inventory of eleven D1 categories almost certainly underrepresents the actual diversity of claim types in natural discourse. Bender and Koller (2020) demonstrate that linguistic form systematically underdetermines semantic content; Chang and Bergen (2023) document that pragmatic competence — the domain most directly implicated in claim-type identification — remains the most fragile dimension of LLM behavior. Sinclair (1991), in establishing the empirical foundations of corpus linguistics, demonstrated that lexical and semantic patterns in large corpora systematically reveal structures that neither grammatical theory nor intuition anticipates. An analogous principle applies to ontological claim types: categories invisible at the scale of philosophical case analysis may become apparent when claim structure is examined across millions of instances.

3.5 The Exclusion Rule and the Pre-Step 0

Two gateway procedures apply before D1 classification. The Exclusion Rule removes from scope all claims of the form 'X exists' or 'X does not exist.' The justification is multi-traditional: Kant argued that existence is not a real predicate (Critique of Pure Reason, B626–B630); Frege formalized this as the claim that existence is a second-order concept (Foundations of Arithmetic, §§53–54); Quine operationalized it in the apparatus of quantification theory; Moro (1997) provides a linguistic analysis demonstrating the conflation of copula, identity, and existential predication across natural languages.

The Pre-Step 0 (framework version 2.2) requires that before any D1 classification is assigned, the annotator verify that the statement is actually being used as an assertion. The theoretical grounding is Austin's distinction between locutionary, illocutionary, and perlocutionary acts (1962). A declarative sentence does not automatically perform an assertive function: the same surface form can serve directive, performative, expressive, or apophatic purposes depending on context.

The founding case is a Taiji master's pedagogical claim: 'those who are not physically flexible are not mentally flexible either.' The statement is almost certainly false as a universal empirical proposition — among seven billion people, counterexamples are immediate (the canonical case is Hawking). But its falsehood as a proposition is not the

relevant consideration in its original context. The master is issuing a motivational directive designed to produce a perlocutionary effect — increased training effort — not making a truth claim. Evaluating it as an assertion and returning a verdict of 'false' is a category mistake. The Pre-Step 0 exists to prevent it. The framework's operational rule is that declarative form is treated as assertive function by default — the more demanding assumption — but when contextual markers clearly indicate a non-assertive illocutionary force, the claim receives the code NA_ILLOCUTORIO and validation is suspended.

3.6 The Asymmetric Validation of Causal Claims

The most technically consequential feature of the framework is its treatment of causal claims classified under D4.3 (Causal-Mechanism): claims that assert that X causes Y, with an implied or explicit mechanism. Standard validation procedures are implicitly symmetric: a claim is evaluated as true or false. The D4.3 procedure is explicitly asymmetric, grounded in the philosophical literature on causation.

Hume established that causal relationships are not directly observable: we observe constant conjunction, not necessary connection. Pearl's interventionist account (2009) formalizes the distinction between observational conditioning — $P(Y|X)$ — and interventional conditioning — $P(Y|do(X))$ — demonstrating that the move from correlation to causation requires a specific kind of evidence that most claims in natural discourse do not supply. Woodward's counterfactual criterion (2003) specifies what it would mean to establish a causal claim: X causes Y if and only if an intervention on X would change Y, holding other factors fixed. Popper's asymmetry between falsification and verification supplies the epistemological constraint: we can demonstrate that a proposed causal mechanism is absent or inverted, but we cannot, by the same procedure, establish what the correct mechanism is.

These considerations generate three structurally distinct validation outcomes. Outcome 1 — Falsified: the proposed causal premise is false, the mechanism is absent or inverted, or — as in the Monaco case — a true conclusion is being attributed to a false or unwarranted causal premise. The canonical example from this study is 'the migration emergency caused the collapse of the regional healthcare system': the causal premise is false, the conclusion is true, and the causal link is asserted without evidence of the mechanism. Outcome 2 — Non-corroborated: the causal claim is not falsified but is unsupported by evidence of the mechanism. A representative case: 'social media use causes depression in adolescents.' Both premise and conclusion are empirically documented; the causal link is the subject of active and unresolved scientific debate, with limited interventional evidence (Orben & Przybylski, 2019). Outcome 3 — Beyond framework scope: the causal complexity exceeds what the protocol can adjudicate without specialized domain knowledge. A representative case: 'globalization caused the decline of the Western middle class.' The causal relationship is real but irreducibly multi-factorial: automation, fiscal policy, financialization, and trade exposure all contribute causally, and their relative weights are disputed among economists using incompatible models. Declaring Outcome 3 is not a failure of the framework; it is the honest demarcation of its limits.

3.7 The Epistemic Responsibility Check

The Epistemic Responsibility Check (ERC) is a lightweight audit applied to every validated claim regardless of D1 type. It asks three questions: Does the source cite evidence or grounds for the claim? Does the source express appropriate uncertainty? Does the source distinguish interpretation from fact where relevant? The theoretical grounding is virtue epistemology (Zagzebski, 2001; Goldman, 1999). The ERC evaluates not whether a claim is true but whether the speaker has discharged the epistemic obligations that attach to assertion — a dimension that standard truth-apt evaluation misses, and that is particularly consequential for AI-generated content where epistemic responsibility becomes a property of prompt design and output format rather than of any agent.

4. Research Design and Methodology

4.1 Research Question and Operative Hypotheses

The study addresses two related but distinct research questions. The first concerns taxonomic coverage: does the seven-dimensional framework provide sufficient coverage to classify claims across the full range of ontological types encountered in AI-generated discourse? The second concerns output reproducibility: when the same prompt is submitted to the same model under identical conditions across multiple replications, does the model produce epistemically consistent outputs?

The operative hypotheses are as follows. H1: Claude's epistemic classifications will show high semantic agreement across replications, indicating that the model converges on the same epistemic judgment even when expressed in different words. H2: Lexical agreement will be substantially lower than semantic agreement, reflecting the model's generative variability in surface form. H3: The gap between lexical and semantic agreement will vary across claim types, being larger for claims admitting richer narrative elaboration and smaller for claims with determinate, constrained outputs.

4.2 A Two-Level Validation Design

The empirical component employs a design with two distinct levels that are complementary but not reducible to each other.

The horizontal validation applies the full framework — including Pre-Step 0, complete D1–D7 classification, asymmetric D4.3 procedure, and ERC — to thirty-one claims selected to maximize coverage of ontological types. Each claim receives a single expert-annotated validation following the Coding Manual v2.0. The purpose is to demonstrate that the framework is applicable across the full range of claim types and yields structurally distinctive outcomes.

The vertical validation applies a structured validation prompt to six claims selected for maximal epistemic contrast, executing thirty independent replications of each via the Anthropic API. The purpose is to measure output consistency when confronted with the same epistemic task repeated identically.

A critical clarification bears on the interpretation of all results in Section 6. The vertical validation does not test whether Claude applies the framework correctly in the sense of the Coding Manual. It tests whether Claude produces consistent outputs when given a structured validation prompt. The prompt instantiates key elements of the framework — ontological classification, decidability assessment, ERC — but does not require the explicit three-outcome D4.3 verdict, nor the full D2–D7 dimensional coding. What the vertical validation measures is the model's self-consistency as an epistemic instrument — a necessary but not sufficient condition for the framework's validity. The sufficient condition — inter-annotator agreement among human annotators following the Coding Manual — is the primary target of the next research phase (Section 8.4).

4.3 Claim Selection

Horizontal validation. The thirty-one claims were selected to satisfy three criteria: (a) coverage of as many D1 subcategories as possible; (b) variation in D4 (causality type) and D5 (framework-dependency), the two dimensions with the highest attributive value; (c) inclusion of boundary and edge cases that test the framework's limits — claims requiring Pre-Step 0, claims requiring the D4.3 asymmetric procedure, and claims structurally undecidable without framework specification. The selection process was abductive in the Peircean sense: claims that resisted classification under existing categories were retained precisely because they forced taxonomy expansion.

Vertical validation. The six claims were selected from the horizontal corpus for maximal epistemic contrast across D1 types, with at least one claim for each of the three D4.3 validation outcomes:

Claim 1 — 'I vaccini causano l'autismo' (Vaccines cause autism). D1.2 Empirical-Quantitative, D4.3. Massively refuted causal claim; canonical D4.3 Outcome 1 (Falsified).

Claim 2 — 'Il vegetarianismo è eticamente superiore' (Vegetarianism is ethically superior). D1.8 Ethical-Normative. Framework-dependent normative claim; principled indeterminacy.

Claim 3 — 'Bitcoin raggiungerà $100k entro il 2026' (Bitcoin will reach $100k by 2026). D1.4 Predictive-Future. Verified a posteriori (December 2024); tests temporal claim processing.

Claim 4 — 'La prima guerra mondiale fu causata dall'assassinio di Sarajevo' (The First World War was caused by the assassination in Sarajevo). D1.1 Historical-Factual, D4.3. D4.3 Outcome 3 (Beyond Framework Scope).

Claim 5 — 'La coscienza non è riducibile a processi neurali' (Consciousness is not reducible to neural processes). D1.14 Metaphysical-Ontological. Paradigm-dependent claim at the boundary of decidability.

Claim 6 — 'Fumare causa il cancro ai polmoni' (Smoking causes lung cancer). D1.2 Empirical-Quantitative, D4.3, D3.2 High-Confidence. Paradigm case of well-corroborated causal claim; direct contrast with Claim 1.

4.4 Validation Protocol

The horizontal validation followed the Coding Manual v2.0 in full. The vertical validation employed a structured prompt instantiating the epistemic validation skill format, submitted to the Anthropic API (model: claude-sonnet-4-6) via automated batch script with checkpoint-resume functionality and five-retry logic. Each claim received thirty independent API calls with identical prompts and no system memory across calls. Total corpus: 180 completed runs with zero skips. Verdict extraction from the 180 outputs is automated (Coding Manual v2.1, Mode A): the Esito cell of each PROSPETTO table is read mechanically and assigned to a semantic category by a deterministic keyword classifier operating under a label-first rule — where an output opens with an explicit verdict self-label, that label governs; the keyword classifier, with negation handling, is the fallback. The extraction pipeline, the raw runs, and the retrospective quality-control log are deposited with the open materials (Appendix G). Execution: 25–26 June 2026; batch_validazione.py. Pre-registration was deposited with OSF prior to data collection (Appendix D).

A reduced multi-model control (Section 6.6) tests whether the ARlex–ARsem gap replicates on a second model of equivalent class (GPT-4-class or open-weights), using the same prompt on two claims (vaccini and vegetarianismo), 30 replications each. This constitutes a minimal cross-architecture check: if the gap holds, it is a structural property of autoregressive generation; if it does not, it is specific to Claude. Configuration details are in Appendix C.3.

4.5 Metrics: Justification and Operationalization

Lexical Agreement Rate (ARlex): for each structured output field, the proportion of replications producing the modal lexical form. Measures surface-form reproducibility. Given autoregressive generative variability, low ARlex is expected a priori.

Semantic Agreement Rate (ARsem): the proportion of replications whose verdict falls in the modal semantic category. The verdict is extracted mechanically from the Esito row of the PROSPETTO DI VALIDAZIONE and assigned to one of five categories by a deterministic keyword classifier whose rules are specified in the Coding Manual (§7.3) and deposited with the open materials. ARsem measures substantive reproducibility: whether the model reaches the same epistemic judgment regardless of verbal expression. Automated extraction (Mode A) is not treated as self-validating. It is checked against an independent human coder, blind to the automated output, on a random sample of the corpus, and the agreement between the two is reported as an agreement rate, not as a certificate of correctness (§6.3). This distinction between lexical and semantic agreement has no standard precedent in LLM evaluation literature.

Mean Pairwise Jaccard Similarity ($\bar{J}$): for narrative output fields, mean Jaccard similarity of token sets across all pairwise comparisons within a claim's replication set (435 pairs per claim for n=30). Measures lexical overlap at the narrative level as a proxy for shared argumentative content.

Composite Stability Index (IC): $IC = 0.60 \times ARsem + 0.40 \times \bar{J}$. The weights reflect the higher epistemic salience of structured judgment fields relative to narrative elaboration.

The choice of these metrics — rather than Fleiss' κ, planned in earlier pre-registration versions — reflects a post-pre-registration methodological refinement documented in the pre-registration amendment. κ statistics assume a fixed coding scheme applied by multiple independent annotators; the vertical validation involves a single model applying its own internal protocol across replications. ARlex, ARsem, and $\bar{J}$ are purpose-designed for the specific measurement task at hand.

4.6 Scope and Declared Limitations

The vertical validation measures one specific property: output consistency across identical-prompt replications. It does not measure correctness of classification (which requires human annotators as ground truth), applicability of the full D4.3 procedure (the prompt does not elicit the three-outcome verdict explicitly), or generalizability across models (all data from claude-sonnet-4-6 under default parameters).

A fourth limitation concerns the status of the semantic categorization, and it is a deliberate feature of the design rather than a shortfall of it. No coding, human or automated, certifies its own correctness. The point is general, not local: there is no theory-neutral observation against which a classification could be finally checked (Quine, 1951), and the base statements that anchor a theory to evidence are accepted by convention, not read off a foundation — Popper's image is exact here, piles driven into a swamp, sunk until they bear the structure, resting on no bedrock (Popper, 1959; the search for a self-certifying given is the myth Sellars, 1956, named). The corresponding principle in the theory of annotation is that a single expert coder is not ground truth: validity is a property of intersubjective agreement measured under an explicit scheme, not of any one coder's access to the fact (Artstein & Poesio, 2008). Systematic disagreement among competent coders is frequently a property of the item, not noise in the measurement — which is this framework's own claim about D1.8 claims, restated at the level of coding. Phase 1 therefore makes only the claim the evidence supports: the semantic categories are reproducible under a fixed, deposited rule, and that rule agrees with an independent blind coder at a reported rate. The stronger claim — that the categories are valid — is a property of inter-annotator agreement across multiple independent coders, computed as Fleiss' κ against the $\kappa \geq 0.61$ criterion (Landis & Koch, 1977). It is the deliverable of Phase 2, not a missing component of Phase 1.

Finally, the apparent circularity of the design — Claude evaluating Claude — is addressed directly. The framework does not ask the model to assess its own correctness; it elicits a structured epistemic behavior from the model and then measures the consistency of that behavior across replications. Correctness is not a property the model assigns to itself; it is a property that human annotators following the Coding Manual assign to the model's output. The auto-referential appearance is therefore methodologically delimited: the model is the instrument, not the judge of the instrument's validity. This distinction is what post-cognition operationalizes.

4.7 Uplift Experiment: Framework vs. Naïf Prompt [PHASE 1 — PLACEHOLDER]

The proof-of-concept that the framework produces measurable epistemic uplift is the most consequential single experiment the paper can report. The design is minimal, pre-

registered, and executable with only API access and a blind coding procedure. The core question: when a claim with a structural causal defect (of the type identified in the Monaco case) is submitted to a naïf prompt versus the framework prompt, does the framework condition flag the defect at a significantly higher rate?

Corpus: four causal claims with D4.3 structure, selected to represent the three D4.3 outcomes: (1) "L'emergenza migranti ha causato il collasso del sistema sanitario regionale" (Monaco case; Outcome 1 — Falsified); (2) "I vaccini causano l'autismo" (Outcome 1 — Falsified); (3) "L'uso dei social media causa depressione negli adolescenti" (Outcome 2 — Non-corroborated); (4) "La globalizzazione ha causato il declino del ceto medio occidentale" (Outcome 3 — Beyond framework scope). Design: two conditions (Framework prompt, Appendix C system prompt; Naïf prompt, Appendix C.4), n = 20 replications per claim per condition, 160 total runs. Scoring: binary, blind, pre-registered. The scorer reads the output without knowing which prompt condition produced it and codes: does the output identify the claim as a causal claim and flag a structural defect in the causal premise? Yes/No. Primary metric: Δ flag rate (Framework − Naïf) per claim, with 95% Wilson CI. Pre-registered hypothesis (H_uplift): $\Delta > 0$ for all four claims; Δ statistically significant ($p < 0.05$, Fisher exact test) for at least three of four. Results: Section 6.7 (placeholder pending execution).

Note on blind coding: outputs will be randomized and stripped of prompt-condition markers before coding. The scorer commits to the coding key before examining any output. Pre-registration of H_uplift is deposited in Appendix D.5 prior to execution.

5. The Framework: Taxonomy, Dimensions, and Horizontal Validation

5.1 The Primary Dimension: D1 Content Type

D1 — Content Type — is the only genuinely taxonomic dimension of the framework. All other dimensions (D2–D7) are attributive: they specify properties of the claim once it has been typed under D1. The eleven active subcategories are described below with their definition, inclusion/exclusion criteria, boundary conditions, and relevant literature. Corpus claims belonging to each category are noted in parentheses.

D1.1 — Historical-Factual. Claims about specific past events, singular occurrences in documented history. Defining features: singularity, pastness, documentability. Corpus claims: 27 (Crimea/Barbero), 28 (Jesus's age), 4 (WWI causation). Key boundary: D1.10 in geopolitical claims where the copula 'is' conflates factual, juridical, and de facto readings (the Crimea case — public television statement by Alessandro Barbero, cited as motivating example; see Appendix E). Literature: Coady (1992), Goldman (1999), Woodward (2003).

D1.2 — Empirical-Quantitative. Claims based on observation, measurement, and data collection, verifiable through empirical investigation. The broadest D1 category. Corpus claims: 16 (flat-earthism), 17 (aliens), 18 (climate hoax), 19 (vaccines-autism), 22 (Taiji flexibility). Key boundary: D1.8 for claims embedding evaluative terms in empirical framing (capitalism/exploitation); D1.4 for necessary a posteriori truths (water is H_2O — Kripke, 1980). Literature: Popper (1959), Hacking (1983), Pearl (2009), Kripke (1980).

D1.3 — Logical-Formal. Claims about logical relations, tautologies, contradictions. Truth guaranteed by structure, independently of empirical investigation. Corpus claims: 6 (bachelors), 9 (triangle), 24 (no facts — performative self-contradiction). Literature: Frege (1879), Quine (1951).

D1.4 — Mathematical-Analytic. Claims about mathematical truths and analytic statements whose truth follows from the definitions of the terms involved. Corpus claims: 5 (water is H_2O — hybrid with D1.2), 14 (Bitcoin — predictive subtype). Literature: Frege (1884), Shapiro (2000), Quine (1951).

D1.5 — Phenomenological-Subjective. Claims about first-person subjective experience — qualia, consciousness, immediate phenomenal states. Corpus claims: 8 (Barolo — hybrid with D1.6), 7 (democracy/press). Key boundary: D1.6 for expert intersubjectivity in sensory claims. Literature: Nagel (1974), Wittgenstein (1953, §246).

D1.6 — Interpretive-Hermeneutic. Claims about meanings, interpretations, and symbolic significance of texts, actions, or cultural practices. Corpus claims: 21 (empathy in Gospels), 15 (beauty — hybrid with D1.7). Key boundary: D1.1 for hermeneutic claims about historical reception. Literature: Gadamer (1960), Eco (1984).

D1.7 — Aesthetic-Evaluative. Claims about beauty, artistic quality, and aesthetic experience. Evaluated through aesthetic judgment — intersubjective in a specific sense: grounded in cultivated sensibility and responsive to shared standards. Corpus claims: 15 (beauty in the eye of the beholder), 13 (best of all possible worlds — hybrid with D1.14). Literature: Kant (1790), Hume (1757).

D1.8 — Ethical-Normative. Claims about moral justice or injustice, about what ought to be done, about moral obligations and duties. Distinguished from D1.10 by deontic force rather than institutional focus. Mandatory annotation: annotate the structure, not the moral correctness. Corpus claims: 1 (abortion), 2 (refugees), 3 (death penalty), 10 (capitalism), 11 (vegetarianism), 20 (happiness). Literature: Gibbard (2003), Foot (1972), Berlin (1958).

D1.9 — Procedural-Instructional. Claims about how to do something — instructions, procedures, techniques. Truth established through practical efficacy. High dimensional redundancy: D1.9 predicts almost all D2–D7 values. No claims in the thirty-one corpus. Literature: Ryle (1949), Wittgenstein (1953, §150).

D1.10 — Prescriptive-Policy. Claims recommending institutional actions, governance decisions, or political interventions. Evaluated for efficacy, feasibility, and institutional design rather than moral correctness. D4 is the most informative attributive dimension for D1.10. Corpus claims: 27 (Crimea — hybrid with D1.1). Literature: Woodward (2003), Pearl (2009).

D1.14 — Metaphysical-Ontological. Claims about the fundamental nature, properties, or structure of reality — not claims about mere existence (excluded by the Exclusion Rule) but claims about what reality is like. Corpus claims: 13 (Leibniz), 23 (mathematics not real), 26 (political maps), 29 (Resurrection), 30 (Civitavecchia), 31 (existentialism/humanism). Key boundary: D1.2 for empirically tractable versus

constitutively inaccessible claims. Literature: Searle (1995), Ferraris (2009), Chalmers (1995), Nagel (1974).

5.2 The Pre-Step 0 and the Exclusion Rule

The Exclusion Rule removes pure existence predicates from scope (Kant, Frege, Quine, Moro). The Pre-Step 0 verifies that the statement is actually being used as an assertion before D1 classification is assigned (Austin, 1962). When non-assertive force is present — as in the Zen koan (Claim 25: 'You must hear the sound of one hand clapping') — the claim receives code NA_ILLOCUTORIO and validation is suspended. The framework's default is to treat declarative form as assertive function — the more demanding assumption — but contextual markers of non-assertive force override this default.

5.3 The Attributive Dimensions D2–D7

The six attributive dimensions specify properties of the claim once typed under D1. Their values are constrained — sometimes strongly — by D1. Key findings from the redundancy analysis (full analysis reported in Appendix F):

D2 (Temporal Scope): semi-determined. Adds genuine information for D1.2, D1.6, D1.8–D1.10; essentially fixed (atemporal) for D1.3/D1.4/D1.14.

D3 (Certainty Level): most epistemically important; genuinely variable for seven of eleven D1 types. Expected κ low due to subjectivity of the D3.2/D3.3 boundary.

D4 (Causality): most valuable attributive dimension. Four of eleven D1 types have genuine variability (D1.2, D1.8, D1.10, D1.14) — where annotation adds information not inferable from D1 alone.

D5 (Framework-Dependency): adds genuine value at the D5.1/D5.3 boundary for empirically contested claims (homeopathy, astrology).

D6 (Validation Method): largely predictable from D1 (7/11 fixed). Retained for pedagogical reasons: prevents annotators from applying the wrong validation method.

D7A/D7B (Falsifiability): D7A adds value for D1.6 and D1.14 variable cases. D7B is maximally redundant (11/11 predictable); retained as pedagogical explicitness.

5.4 The Thirty-One Claims: Corpus and Taxonomic Function

N	Claim (Italian / English)	D1	Taxonomic function
1	L'aborto è moralmente sbagliato / Abortion is morally wrong	D1.8	Canonical framework-dependent indeterminacy
2	È moralmente obbligatorio aiutare i rifugiati / It is morally obligatory	D1.8	Tests scope of normative obligation across cosmopolitanism

N	Claim (Italian / English)	D1	Taxonomic function
	to help refugees		vs. communitarianism
3	La pena di morte è ingiusta / The death penalty is unjust	D1.8	Tests theory-of-justice dependency
4	L'AI sostituirà il 40% dei lavori entro il 2030 / AI will replace 40% of jobs by 2030	D1.4 Predictive	Tests quantified future claims; D3.3 and false precision
5	L'acqua è H_2O / Water is H_2O	D1.4/D1.2 hybrid	Kripkean necessary a posteriori; analytic/synthetic boundary
6	I celibi non sono sposati / Bachelors are unmarried	D1.3	Canonical analytic tautology; certainty floor
7	La democrazia richiede libertà di stampa / Democracy requires freedom of the press	D1.5/D1.10	Tests definition-dependency across democracy concepts
8	Il Barolo ha sentori di rosa e catrame / Barolo has notes of rose and tar	D1.5	Expert intersubjectivity; phenomenology vs. convention
9	Il triangolo ha tre lati / A triangle has three sides	D1.3	Definitional tautology; control case
10	Il capitalismo sfrutta i lavoratori / Capitalism exploits workers	D1.8/D1.2	Evaluative terms in empirical framing; semantic pre-analysis required
11	Il vegetarianismo è eticamente superiore / Vegetarianism is ethically superior	D1.8	Selected for vertical validation; paradigm case of principled indeterminacy
12	Il rosso è più caldo del blu / Red is warmer than blue	D1.5	Cross-modal metaphor; physical falsity

N	Claim (Italian / English)	D1	Taxonomic function
			vs. psychological universality
13	Questo mondo è il migliore dei mondi possibili / This world is the best of all possible worlds	D1.14	Leibnizian theodicy; modal reasoning
14	Bitcoin raggiungerà $100k entro il 2026 / Bitcoin will reach $100k by 2026	D1.4 Predictive	Selected for vertical validation; a posteriori verified prediction
15	La bellezza è negli occhi di chi guarda / Beauty is in the eye of the beholder	D1.7	Subjectivism/ objectivism boundary in aesthetics
16	Il terrapiattismo è una teoria credibile / Flat-earthism is a credible theory	D1.2	Massively refuted; control case for robust falsification
17	Gli alieni hanno visitato la Terra / Aliens have visited Earth	D1.2	Probabilistic reasoning under insufficient evidence
18	Il cambiamento climatico è una bufala inventata dalla Cina / Climate change is a hoax invented by China	D1.2+D1.1	Composite: two independently falsifiable components
19	I vaccini causano l'autismo / Vaccines cause autism	D1.2, D4.3	Selected for vertical validation; canonical D4.3 Outcome 1 (Falsified)
20	La felicità è la cosa più importante nella vita / Happiness is the most important thing in life	D1.8/D1.7	Axiological hierarchy; teleological indeterminacy
21	I Vangeli non	D1.6+D1.1	Hermeneutic claims;

N	Claim (Italian / English)	D1	Taxonomic function
	incoraggiano l'empatia universalistica / The Gospels do not encourage universalistic empathy		text-level vs. reception-level decidability
22	Chi non è flessibile fisicamente non lo è mentalmente / Those not physically flexible are not mentally flexible	D1.2	Universally quantified claim; counterexample falsification; Pre-Step 0 founding case
23	La matematica non è reale / Mathematics is not real	D1.14	Mathematical ontology; D5.1 at its most radical
24	Non ci sono fatti; solo interpretazioni / There are no facts; only interpretations	D1.14+D1.3	Performative self-contradiction; decidable despite metaphysical register
25	Devi sentire il rumore che fa una mano sola / You must hear the sound of one hand clapping	NA_ILLOCUTORIO	Founding case for Pre-Step 0 and NA_ILLOCUTORIO code; Zen koan
26	Le decisioni della cartina politica sono reali / The decisions of the political map are real	D1.14	Child's spontaneous ontological claim; Searle's institutional facts
27	La Crimea è Russia (Barbero) / Crimea is Russia	D1.1+D1.10	Copula ambiguity across juridical, historical, de facto readings
28	Quando Gesù è morto aveva quasi cinquant'anni / When Jesus died he was nearly fifty	D1.1	Probabilistic historical claim against scholarly consensus

N	Claim (Italian / English)	D1	Taxonomic function
29	Gesù è veramente risorto / Jesus truly rose from the dead	D1.14+D1.2	Multi-level: historical/naturalistic/metaphysical differentiated treatment
30	La statua di Civitavecchia è un autentico miracolo / The Civitavecchia statue is an authentic miracle	D1.2+D1.14	Unexplained phenomenon (partially decidable) + miraculous cause (beyond scope)
31	Ogni esistenzialismo è un umanismo / Every existentialism is a humanism	D1.14	Universal falsifiable by counterexample (Heidegger, Kierkegaard)

5.5 Outcome Distribution and Structural Patterns

The thirty-one claims distribute across seven outcome categories, mapping systematically onto D1 types:

Outcome	N	%	Claims	D1 pattern
Framework-dependent (DIPENDE)	8	26%	1, 2, 3, 10, 11, 20, 23, 27	All D1.8 or D1.14
Falsified (FALSO)	6	19%	13, 16, 18, 19, 22, 24	D1.2 massively refuted + D1.14 self-contradiction
Probably False (PROB. FALSO)	4	13%	4, 17, 28, 31	Predictive or probabilistic
True (VERO)	4	13%	5, 6, 9, 26	Analytic, necessary a posteriori, institutional
Conditionally True (CONDIZ.)	5	16%	7, 8, 14, 15, 21	Framework-relative or partially verified
Structured multi-level (STRUTTURAT	2	6%	29, 30	Multi-register claims requiring

Outcome	N	%	Claims	D1 pattern
O)				differentiated treatment
NA_ILLOCUTORIO	1	3%	25	Non-assertive illocutionary force

Four structural patterns emerge with sufficient regularity to constitute findings: (1) D1.8 claims are systematically indeterminate without framework specification — a structural property of the category, not of these particular claims. (2) D1.2 counterfactual claims are robustly falsifiable at high confidence. (3) D1.3/D1.4 claims are true with certainty and zero empirical information, establishing the certainty floor. (4) Composite and multi-level claims require differentiated treatment that type-agnostic validation cannot provide — the pattern most relevant to AI-generated discourse, where composite causal-normative claims are common and their structural defects invisible to uniform evaluation.

5.6 Taxonomy-Extending Cases

Three cases in the horizontal corpus forced extensions not anticipated at the taxonomy's outset — evidence for the open-system character of the taxonomy. The koan (Claim 25) generated the Pre-Step 0 and the NA_ILLOCUTORIO code: no existing D1 category could accommodate a claim that is constitutively not an assertion. The Barbero/Crimea case (Claim 27) forced the disambiguation of D1.1 and D1.10 in geopolitical claims where the copula 'is' conflates descriptive, juridical, and de facto registers — generating a refinement now incorporated in the Coding Manual v2.0. The child's claim about political maps (Claim 26) pointed toward a potential D1.15 (Institutional-Ontological) subcategory for claims about Searlean institutional facts — reserved for future development but established as necessary by this case.

6. Vertical Validation: Output Reproducibility Across Thirty Replications

6.1 Overview

The vertical validation applies the epistemic validation protocol to six claims, executing thirty independent API replications of each. The complete corpus is 180 runs with zero skips. The principal finding — established before examining any individual claim — is the aggregate gap between ARlex and ARsem: 4.44% versus 76.67%, a difference of approximately 72 percentage points. This gap is a structural finding about LLM output variability: the model produces lexically highly variable outputs that nonetheless converge on the same semantic judgment at high rates. (The reported ARsem figures follow the label-first re-extraction adopted in resolution of a spot-validation failure documented in the QC Log; the deposited-classifier figures were 78.33% and a 74-point gap, and the gap is robust to the resolution. The blind human spot-validation mandated by §6.3 after the revision is pending; two per-claim proportions remain under adjudication, noted below.) The epistemological implications are discussed in Section 7.

6.2 Results by Claim

Claim 1 — 'I vaccini causano l'autismo' / 'Vaccines cause autism' (D1.2, D4.3)

ARlex = 6.67%. ARsem = 100% (30/30 FALSO). Wilson IC95%: [88.65%, 100.0%]. $\bar{J}$ = 0.354.

The lexical distribution is maximally dispersed — 28 distinct formulations across 30 runs — while the verdict is unanimous. Semantic convergence at 100%, the highest in the corpus, corresponds to the claim's epistemic properties: no legitimate scientific debate exists, and unanimity reflects the absence of genuine epistemic ambiguity. (Under the deposited classifier one run was read as PARZIALMENTE VERO through an over-nuancing of the strong/weak distinction; the label-first rule reads its explicit FALSO self-label correctly.)

Claim 2 — 'Il vegetarianismo è eticamente superiore' / 'Vegetarianism is ethically superior' (D1.8)

ARlex = 3.33%. ARsem = 70.0% (SOSPESO dominant; exact proportion pending §6.3 blind validation — 70.0% re-extracted, 66.7% under the QC Log sensitivity check). Wilson IC95%: [52.1%, 83.3%]. $\bar{J}$ = 0.320.

INTERPRETIVE NOTE (Phase 1 framework): the dominant verdict is suspension of judgment. This is the expected ground truth for a D1.8 claim submitted without a specified moral framework, not model instability. A normative claim that predominantly declines to return a determinate verdict, where the claim type makes one unavailable, is a cleaner confirmation of the framework's structural prediction than a determinate verdict would be. The residual PARZIALMENTE VERO runs correspond to outputs that resolve the claim in a weak, framework-relative reading; the rare FALSO/VERO/DIPENDE runs are the genuine anomalies — collapses of a framework-dependent structure into a determinate verdict — and their low frequency is itself a finding. ARsem here is not a measure of unreliability; it is a measure of structural framework-dependency.

Claim 3 — 'Bitcoin raggiungerà $100k entro il 2026' / 'Bitcoin will reach $100k by 2026' (D1.4)

ARlex = 6.67%. ARsem = 96.7% (29/30 SOSPESO; 1 PARZIALMENTE VERO). Wilson IC95%: [83.3%, 99.4%]. $\bar{J}$ = 0.316.

Predictive and uncertain at formulation, verified a posteriori (December 2024). The model produces predominantly SOSPESO rather than VERO: the temporal framing activates a 'suspended / verifiable at expiry' response that competes with factual knowledge of the outcome. This is a finding about how the model processes temporally indexed claims, not a failure of the protocol. (The deposited classifier's 90% reflected two negated verdicts — 'plausibile ma non confermato' and kin — read as polar; negation handling and the explicit SOSPESO self-labels resolve them.)

Claim 4 — 'La prima guerra mondiale fu causata dall'assassinio di Sarajevo' / 'The First World War was caused by the assassination in Sarajevo' (D1.1, D4.3)

ARlex = 3.33%. ARsem = 73.3% (22/30 PARZIALMENTE VERO; 8/30 FALSO). Wilson IC95%: [55.6%, 85.8%]. $\bar{J}$ = 0.340.

The dominant verdict is the strong/weak form distinction: the assassination was a trigger, not a sufficient cause; the strong monocausal reading is false, the weak proximate-cause reading defensible. The FALSO runs foreground the strong-form falsity without the weak-form qualification. (Under the deposited classifier a mixed-verdict override unified all thirty into a single category, yielding 100%; the label-first rule reads each output's leading label and splits the strong-form-false minority out — a more faithful account of what the outputs say. This claim, more than any other, shows that the choice of extraction rule is a choice about what the verdict of a strong/weak validation is.)

Claim 5 — 'La coscienza non è riducibile a processi neurali' / 'Consciousness is not reducible to neural processes' (D1.14)

ARlex = 3.33%. ARsem = 56.7% (17/30 SOSPESO; 12/30 PARZIALMENTE VERO; 1 ALTRO). Wilson IC95%: [39.2%, 72.6%]. $\bar{J}$ = 0.351.

The lowest ARsem in the corpus, with a wide confidence interval. The dominant SOSPESO reflects the claim's genuine multi-dimensionality — simultaneously metaphysical (D1.14), empirical (neuroscience limits), and phenomenological — with the model suspending an overall verdict where the dimensions pull apart; the PARZIALMENTE VERO minority resolves on the strong/weak reading (the strong ontological irreducibility undemonstrated, the weak epistemic reading defensible). This is sensitivity to semantic complexity, not randomness. (Deposited classifier: dominant PARZIALMENTE VERO at 73.3%; the flip to SOSPESO follows from reading the explicit SOSPESO self-labels the keyword scheme had overridden.)

Claim 6 — 'Fumare causa il cancro ai polmoni' / 'Smoking causes lung cancer' (D1.2, D4.3, D3.2)

ARlex = 3.33%. ARsem = 63.3% (PARZIALMENTE VERO dominant; exact proportion pending §6.3 blind validation — 63.3% re-extracted, ~66.7% under the QC Log sensitivity check; ~19/30 PARZIALMENTE VERO, ~11/30 VERO). Wilson IC95%: [45.5%, 78.1%]. $\bar{J}$ = 0.356 (highest in corpus).

A claim with overwhelming scientific consensus achieves only moderate semantic agreement because the model consistently applies the strong/weak distinction to the causal claim itself: the strong form (every smoker develops lung cancer) is false; smoking is a major probabilistic risk factor, neither necessary nor sufficient in the deterministic sense. The PARZIALMENTE VERO runs identify this; the VERO runs accept the standard epidemiological reading without foregrounding it. $\bar{J}$ = 0.356 is consistent: when the model is confident about the underlying science, narrative elaboration converges tightly even as the verdict label varies. The strong/weak disambiguation adds genuine epistemic value over spontaneous output.

6.3 Summary Table

Claim	D1	ARlex%	ARsem%	IC95%	J̄	IC
Vaccini / Vaccines	D1.2, D4.3	6.67	100.00	[88.7, 100.0]	0.354	0.742
Vegetarianismo / Vegetarianism	D1.8	3.33	70.00 †	[52.1, 83.3]	0.320	0.548
Bitcoin	D1.4	6.67	96.70	[83.3, 99.4]	0.316	0.707
Prima Guerra / WWI	D1.1, D4.3	3.33	73.30	[55.6, 85.8]	0.340	0.576
Coscienza / Consciousness	D1.14	3.33	56.70	[39.2, 72.6]	0.351	0.481
Fumo / Smoking	D1.2, D4.3	3.33	63.30 †	[45.5, 78.1]	0.356	0.522
Mean / Media	—	**4.44**	**76.67**	—	**0.340**	**0.596**

IC = 0.60 × ARsem + 0.40 × J̄ (ARsem as a fraction), recomputed on the label-first values. Wilson IC95% at 95% confidence. † Exact proportion pending the §6.3 blind spot-validation; dominant category settled.

6.4 The Principal Finding: The ARlex–ARsem Gap

The aggregate gap between ARlex (4.44%) and ARsem (76.67%) — approximately 72 percentage points — is the central quantitative finding of the vertical validation. The correct explanation is architectural: the model's epistemic judgment is encoded in internal representations and reproduced consistently; the verbal form is generated token by token and subject to stochastic variation. ARlex measures surface-form reproducibility; ARsem measures epistemic-judgment reproducibility. The 72-point gap is the distance between these two things — a distance that matters enormously for any application of LLM-based validation, because lexical consistency is a systematically misleading proxy for epistemic consistency.

The interpretability literature now supplies a mechanism consistent with this reading. Gurnee et al. (2026) report that the reportable representations of a production model carry persistent, abstract content across an intermediate band of layers, and that the final layers switch to a regime in which representations track the imminent output token rather than intermediate computation. On that account, the semantic content of a judgment is settled upstream of the point at which its surface form is selected, and the selection itself is the high-variance step. A large gap between semantic and lexical agreement is what such an

architecture would produce. The convergence is offered as a consistent reading rather than a demonstration: the present study measures outputs, not internal representations, and no claim is made here that the two lines of evidence have been linked experimentally. Establishing that link — applying representation-level analysis to the same 180-run corpus — is a natural Phase 2 extension.

This finding connects to the self-consistency literature (Wang et al., 2022): just as chain-of-thought reasoning benefits from sampling multiple answers, epistemic validation benefits from distinguishing the stable judgment from the variable expression. The present study extends that insight by showing that the stability operates at the semantic level even when lexical surface is maximally dispersed — and that the degree of stability is predicted by ontological claim type, a variable absent from the self-consistency literature.

A sharper objection must be faced. From the perspective of Quattrociocchi et al. (2025), high ARsem is not evidence of stable epistemic judgment; it is evidence of epistemia — the model transitioning to a stable attractor that mimics judgment without constituting it. On this reading, ARsem = 100% for vaccini would mean not that Claude reliably judges vaccines-autism as falsified, but that the model reliably lands on the same surface output regardless of any genuine epistemic process. The objection is formally valid: ARsem alone cannot distinguish between genuine epistemic convergence and statistical concentration on an epistemic-looking attractor. This is precisely why ARsem is a necessary but not sufficient measure, and why the uplift experiment (Section 6.7) is the paper's decisive empirical move. If the framework prompt produces measurably higher rates of causal-defect flagging than the naïf prompt on identical claims, the difference cannot be explained by attractor concentration — it must reflect a change in what the model is prompted to attend to. The uplift experiment is, in effect, a test of whether post-cognition produces epistemic behavior or epistemic appearance.

6.5 Claim-Type Effects on Semantic Agreement

The ARsem pattern maps onto D1 classification consistently with the framework's structural predictions. Highest ARsem: claims with determinate, framework-independent epistemic outcomes — vaccini (100%), bitcoin (96.7%, SOSPESO structurally determined by temporal framing). Lowest ARsem: claims with structural epistemic indeterminacy — coscienza (56.7%, genuine multi-dimensionality), fumo (63.3%, strong/weak causal ambiguity), vegetarianismo (70.0% SOSPESO, D1.8 framework-dependency), and prima_guerra (73.3%, the strong/weak split in a causal-historical claim). This pattern is not a finding about model quality; it is a finding about the relationship between ontological type and output variance — which is what the framework predicts.

6.6 Multi-Model Control [PHASE 1 — PLACEHOLDER PENDING EXECUTION]

Hypothesis: the ARlex–ARsem gap observed in Section 6.4 is a structural property of autoregressive generation, not a quirk specific to claude-sonnet-4-6. To test this, two claims (vaccini and vegetarianismo) were submitted with the identical system prompt to a second model of equivalent class ([MODEL-2: to be specified — GPT-4o or open-weights equivalent]), 30 replications each. If the gap holds cross-model, it supports the

architectural explanation; if it does not, the gap is model-specific and the interpretation must be revised.

[PLACEHOLDER — Table 6.6: Multi-model comparison. To be completed after execution.]

Claim	Model	ARlex%	ARsem%	IC95%	$\bar{J}$	IC
Vaccini	claude-sonnet-4-6	6.67	100.00	[88.7, 100.0]	0.354	0.742
Vaccini	[MODEL-2]	[___]	[___]	[___]	[___]	[___]
Vegetarianismo	claude-sonnet-4-6	3.33	70.00	[52.1, 83.3]	0.320	0.548
Vegetarianismo	[MODEL-2]	[___]	[___]	[___]	[___]	[___]

[Interpretive text to be added after execution. Key question: does the ARlex–ARsem gap replicate cross-model? Does the D1.8 distributional signature (SOSPESO-dominant) replicate on vegetarianismo?]

6.7 Uplift Experiment: Framework vs. Naïf Prompt [PHASE 1 — PLACEHOLDER PENDING EXECUTION]

Pre-registered hypothesis H_uplift: the framework prompt flags the structural causal defect in D4.3 claims at a significantly higher rate than the naïf prompt. Corpus: four causal claims (Monaco/migration, vaccini, social media→depression, globalization→middle class decline). Metric: binary flag rate per condition, blind-coded, pre-registered (Appendix D.5). n = 20 replications per claim per condition, 160 total runs.

[PLACEHOLDER — Table 6.7: Uplift results. To be completed after blind coding.]

Claim	Naïf flag%	Framework flag%	Δ	IC95% Δ
Migrazione→sanità	[___]	[___]	[___]	[___]
Vaccini→autismo	[___]	[___]	[___]	[___]
Social media→depressione	[___]	[___]	[___]	[___]
Globalizzazione→ceto medio	[___]	[___]	[___]	[___]

[Interpretive text after execution. If Δ positive and significant → uplift confirmed; if null/negative → honest negative finding, reduces scope of contribution but remains publishable as a limit discovered before peer review.]

7. Discussion

7.1 What the Study Establishes

The study establishes three things at three levels of generality.

At the level of the framework, the horizontal validation demonstrates that the seven-dimensional taxonomy provides sufficient coverage to classify claims across the full range of ontological types, and that different types produce structurally distinctive validation outcomes. The consistent mapping between D1 type and outcome pattern is empirical confirmation that the taxonomy captures real epistemic structure.

At the level of the model, the vertical validation demonstrates that Claude produces epistemically consistent outputs at high rates (mean ARsem = 76.67%) while its lexical expression varies substantially (mean ARlex = 4.44%). This finding supports the tractability of post-cognitive intervention: if the model's epistemic judgments were entirely unstable, no structured protocol could reliably elicit consistent outcomes.

At the level of the field, the study contributes a methodological reorientation. The framework's purpose is not to measure how reliable an LLM is — reliability assessment is a property of the observer's relationship to the system. The framework's purpose is to elicit a specific kind of behavior from the model: ontologically pre-classified, epistemically disciplined output that the model would not produce spontaneously without the protocol's scaffolding. The framework functions as a form of post-production — a structured intervention applied after raw generation that transforms an epistemically undifferentiated output into one that has been classified by type, evaluated by the appropriate method, and subjected to an explicit responsibility check. The reduction of hallucination and epistemic overconfidence that the framework enables is not a side effect of measurement; it is the primary goal of the elicitation.

This reorientation changes how the results should be read. The ARsem values are not measurements of model quality in isolation; they are measurements of the stability of the model's epistemic behavior under the framework's elicitation conditions. The vegetarianismo result (SOSPESO-dominant) is moderate ARsem and high epistemic correctness simultaneously. The prima guerra result (73.3% on the strong/weak distinction) is dominant semantic agreement across 30 entirely distinct formulations simultaneously. Conflating ARsem with a reliability score would misrepresent both results.

7.2 The ARlex–ARsem Gap: Implications for Elicitation

Without the framework's structured prompt, the Monaco case demonstrated that the model collapses A→B into an assessment of B, ignoring the epistemic structure that makes the claim defective. The framework's intervention — requiring ontological pre-classification before any evaluation is attempted — forces the model to identify the claim

as D4.3, apply the asymmetric three-outcome procedure, and produce an output that locates the defect correctly. This is elicitation: the framework changes what the model does, not merely how its existing behavior is measured.

The ARsem results show that the model's epistemic judgments are substantially more stable than their verbal expression. This stability is the raw material that post-cognitive intervention works with: the model has, in a meaningful sense, already performed the epistemic work internally, and the framework's role is to surface and structure that work in a reproducible form. Post-cognition is not adding intelligence to the system; it is providing the scaffolding that allows the system's existing epistemic capacity to produce outputs that are classifiable, auditable, and consistent.

The relationship to the self-consistency literature (Wang et al., 2022) is precise: where self-consistency improves reasoning by sampling multiple answers and aggregating, post-cognition improves epistemic discipline by pre-classifying the claim type before any answer is sampled. The two interventions are complementary rather than competing, and their combination is a natural direction for the Phase 2 research programme. The epistemia objection raised in Section 6.4 — that ARsem measures attractor concentration rather than epistemic convergence — applies here with equal force: the scaffolding the framework provides might be producing reliable epistemic appearance rather than reliable epistemic behavior. The distinction is real, and it is precisely what the uplift experiment is designed to test. Until those results are in, ARsem is the best available proxy for output stability; the claim that post-cognition produces genuine epistemic uplift rather than better-structured appearance remains a hypothesis, stated as such, and pre-registered.

7.3 The Vegetarianismo Anomaly: A Finding About Framework-Dependency

The vegetarianismo result provides the clearest empirical confirmation of the framework's central claim. Under the label-first extraction, the dominant verdict is SOSPESO — suspension of judgment — for a normative D1.8 claim submitted without a specified moral framework. This is not a low score on a reliability scale; it is the correct measurement of the claim's structural property, its framework-dependency, as processed by a model without access to the framework specification that would resolve it. The model predominantly declines to force a determinate verdict precisely where the claim type makes one unavailable. Stated differently: the vegetarianismo result is moderate ARsem and high epistemic correctness simultaneously. This is the most important single data point in the study. It demonstrates that the category error the framework is designed to prevent is not merely theoretical — it is measurable, and it manifests exactly where a naïf evaluation system, expecting high semantic agreement as a proxy for reliability, would most confidently misreport the model as unreliable.

Note on extraction. Under the deposited keyword classifier this claim returned an exact 12/12 tie between PARZIALMENTE VERO and SOSPESO (ARsem 40%), read in earlier drafts as principled bifurcation. The retrospective QC found that tie to be classifier-sensitive: the outputs' explicit SOSPESO self-labels had been overridden by keyword priority. The label-first rule reads the self-labels and resolves the claim to a SOSPESO plurality. The reading is now sharper — the model suspends judgment rather than splitting

between two determinate-ish readings — and the exact proportion is fixed by the blind spot-validation (§6.3).

7.4 Sycophancy, Output Variance, and the Limits of This Study

The vertical validation measures output variance across identical-prompt replications. It does not measure sycophancy, which would require varying prompt framing to test whether the model's judgment shifts in response to perceived interlocutor preferences. The present results cannot rule out sycophancy as a confound, though its influence under neutral identical prompts is implausible. A sycophancy test — the same six claims with varied framing (neutral, pro, contra) — is identified as a priority for the next research phase.

7.5 The D4.3 Procedure in the Vertical Corpus

Three of the six vertical claims have D4.3 causal structure (vaccini, prima guerra, fumo). For vaccini, the model consistently produces the equivalent of Outcome 1 (Falsified) without being prompted — the D4.3 framework adds explicit structure to what the model already does correctly. For prima guerra, the dominant PARZIALMENTE VERO verdict (73.3%) on the strong/weak causal distinction is the model spontaneously applying D4.3 logic without guidance, with a strong-form-false minority (8/30) that the framework's explicit procedure would render as the weak-form qualification. For fumo, the oscillation between PARZIALMENTE VERO and VERO is the empirical evidence that the strong/weak disambiguation is needed — the model's unsupported judgment is inconsistent in a way the framework's explicit procedure would resolve.

7.6 What This Study Does Not Establish

Five limitations require explicit statement. First, the study does not establish correctness of classification: whether Claude's classifications are correct according to the framework requires human annotators as ground truth. Second, the study does not establish inter-annotator reliability: the horizontal validation was conducted by a single annotator, and the critical test awaits the next research phase. Third, the study does not generalize across models: all results are from claude-sonnet-4-6 under default parameters. Fourth, the study does not establish reliability of the ARsem scheme across independent coders: the semantic equivalence judgments are mono-rater in Phase 1 and enter the IAA protocol in Phase 2. A fifth limitation is conceptual rather than empirical and is argued in §3.3: the framework's operative form is external because the systems it is applied to do not classify the epistemic type of their own assertions, and that condition is contingent. These limitations are not incidental; they define the boundary of Phase 1 and the agenda of Phase 2.

8. Conclusion

8.1 Summary of Contributions

This paper has introduced and empirically tested a seven-dimensional ontological framework for the epistemic validation of AI-generated claims. The central principle — ontology precedes epistemology — is not a slogan but an operational requirement: without

prior identification of the claim type, no validation procedure can locate the right questions to ask, apply the appropriate standard, or produce results interpretable across claim types.

As a Phase 1 study, the contributions are three: an architectural contribution, an empirical contribution, and a conceptual contribution. The architectural contribution is the D1 taxonomy of eleven mutually exclusive content-type categories, grounded in epistemology, philosophy of science, speech act theory, and the theory of causation — together with the six attributive dimensions, the Pre-Step 0, the Exclusion Rule, the asymmetric D4.3 procedure, and the ERC. The empirical contribution is twofold: a horizontal validation across thirty-one claims demonstrating taxonomic coverage and structural distinctiveness of outcomes by D1 type; and a vertical validation of 180 runs producing the study's principal quantitative finding — a mean gap of 72 percentage points between ARlex (4.44%) and ARsem (76.67%), with claim-type effects that confirm the framework's structural predictions. The barycentre of the empirical contribution is not the ARlex–ARsem gap per se (which is expected given autoregressive architecture) but the D4.3 asymmetric signature and the ontological-type-determined variance pattern: these are the findings that are not predictable from architecture alone, and that demonstrate the framework's discriminative power.

The conceptual contribution is the introduction and operationalization of post-cognition as a structural response to epistemia — the condition, identified by Loru et al. (2025) and theorized by Quattrociocchi et al. (2025), whereby LLM outputs produce the appearance of epistemic convergence without its substance. Post-cognition does not eliminate epistemia; it provides the external scaffold that makes the distinction between appearance and behavior testable, one claim type at a time. Phase 2 deliverables — inter-annotator reliability, multi-model control, uplift experiment — are pre-registered and open for collaboration (Section 8.4).

8.2 The Framework as Elicitation Tool

The framework's purpose is productive elicitation, not diagnostic measurement. It does not assess how reliable an LLM is in the abstract; it creates the conditions under which an LLM produces epistemically disciplined outputs that it would not produce spontaneously. The Monaco case provides the baseline: without the framework, a well-functioning model collapsed a causal claim into an assessment of its conclusion. With the framework, the same model correctly identifies the claim type, applies the appropriate procedure, and locates the epistemic defect.

8.3 The Open Taxonomy and the Long-Term Research Programme

The taxonomy is an open system. New categories will emerge from boundary cases through abductive logic. The long-term research programme has two components: empirical analysis of large linguistic corpora to identify claim types resisting current classification, following Sinclair's (1991) demonstration that corpus-scale analysis reveals structures invisible at the scale of philosophical case analysis; and construction of a complete multi-language corpus of validated claims with inter-annotator reliability measurements across the full D1 taxonomy. The Phase 1 materials supporting this programme — Coding Manual v2.1, the 31-claim horizontal corpus with full annotations, the system prompt verbatim,

and the batch scripts — are deposited openly at the OSF repository (Appendix G) under a licence that explicitly invites independent replication and taxonomy extension.

8.4 Immediate Next Steps: The Inter-Annotator Reliability Phase

The single most important next step is the inter-annotator reliability study that the present work makes possible but does not include. A corpus of fifty to one hundred claims, covering all eleven active D1 subcategories with deliberate representation of boundary cases, submitted to three or more independent annotators following the Coding Manual v2.1. Cohen's κ (for pairs) and Fleiss' κ (across annotators) computed for each dimension, with D1 as the critical metric: $\kappa \geq 0.61$ (substantial agreement, Landis & Koch, 1977) is the pre-registered criterion for framework validity.

A note on the manual's version is required here, because Section 5 reports work conducted under v2.0 while Phase 2 will be conducted under v2.1. Version 2.1 supersedes v2.0 and is additive with respect to it: it adds the executed six-claim scope, the ARlex and ARsem definitions, the Mode A / Mode M distinction with mandatory spot-validation, and the QC-logging requirement — all of them provisions for the vertical study, which v2.0 did not cover. No annotation rule applied in the horizontal validation was revised. The thirty-one annotations reported in Section 5 therefore stand as conducted and require no re-annotation under v2.1, and Phase 2 annotators inherit the horizontal corpus as a training set without a version caveat. Where this paper records what was done, it names v2.0; where it specifies what is deposited or what annotators should follow, it names v2.1.

This study invites independent annotators to co-construct Phase 2. The Coding Manual v2.1, the 31-claim corpus with full annotations, and the annotation template are available at the OSF repository (Appendix G). Researchers or graduate students with background in epistemology, philosophy of language, or computational linguistics who are willing to annotate a sub-corpus of 20–30 claims are invited to contact the author. Annotations will be pooled, the κ computed publicly, and all contributors acknowledged. The Phase 2 paper will be co-authored with annotators contributing a minimum of one sub-corpus.

8.5 Secondary Research Directions

Four secondary directions follow from the present study. Sycophancy testing: the same six claims submitted with varied prompt framing (neutral, pro, contra) to measure whether the framework's elicitation is robust to framing effects. Multi-model comparison: whether the ARlex–ARsem gap, D1 type effects, and spontaneous strong/weak form distinction generalize across models. Representation-level corroboration: applying the interpretability methods of Gurnee et al. (2026) to the 180-run corpus, to test directly whether the semantic content of a verdict is settled upstream of the layers at which its surface form is selected — the mechanism proposed as a consistent reading in §6.4, converted into a measurement. Ethical epistemology as a second research axis: the consistent DIPENDE pattern for D1.8 claims suggests that no AI system can produce determinate verdicts on ethical claims without prior framework specification — a contribution to the epistemology of AI alignment. A further extension concerns epistemological universals: the ERC operationalizes three epistemic obligations (cite evidence, express uncertainty, distinguish

interpretation from fact) that may constitute genuine cross-framework universals, a hypothesis the framework is positioned to investigate empirically.

8.6 Closing Remark

The problem this paper began with — a German tourist at an alpine resort accepting without examination a causal premise that no amount of logical pressure from his interlocutor had disturbed — is not primarily a problem about AI. It is a problem about the structure of epistemic practice in conditions where the distinction between claim types has collapsed. The tourist was not unintelligent or dishonest. He had simply never been given a reason to ask what kind of claim he was accepting before accepting it. This failure is as old as public discourse: the conflation of causal assertions with factual reports, of normative judgments with empirical findings, of framework-dependent conclusions with universal truths, is a structural feature of natural language that no technology created and no technology can eliminate.

What large language models introduce is not a new epistemic problem but an unprecedented scale of the existing one. A model capable of generating claims across every ontological type, in every natural language, at industrial scale, and with a surface fluency indistinguishable from expert prose, does not change the nature of the epistemic failure — it multiplies its reach by orders of magnitude. The tourist's uncritical acceptance of a causal premise affected one conversation. An LLM deploying the same structure across millions of interactions, without any mechanism that identifies what kind of claim it is producing or what validation standard applies to it, affects the epistemic environment at a scale that makes the individual case negligible by comparison.

The framework proposed here is a response to both dimensions of this problem simultaneously. At the individual level, it is an answer to the prior question that the tourist's interlocutor could not get him to ask: what kind of claim is this, and what does evaluating this kind of claim actually require? At the systemic level, it is an attempt to provide, for AI-generated content at scale, the ontological pre-classification that human epistemic practice performs imperfectly and inconsistently, and that the architecture of autoregressive generation does not perform at all. Post-cognition — the external reconstruction of the ontological commitments implicit in an output — is not a technical fix for a technical problem. It is an epistemological intervention in a problem that is simultaneously ancient and newly urgent. This paper is Phase 1. The architecture is open, the materials are public, and the next step requires collaborators. The invitation is in Appendix G.

[PHASE 3 HORIZON — AUTHOR'S NOTE, NOT FOR DEVELOPMENT HERE. The Monaco case is already supra-atomic: the epistemic defect resides in the A→B relation, a relational and textual property, not a property of either A or B individually. The claim-level framework developed here is the necessary prerequisite; the extension to argument-level and text-level post-cognition is the natural Phase 3 direction. To be developed separately.]
