Ontology Precedes Epistemology — References & Appendices (v6)

References

Andreas, J. (2022). Language models as agent models. In *Advances in Neural Information Processing Systems*, 35, 22966–22979.

Arendt, H. (1967). Truth and politics. *The New Yorker*. Reprinted in H. Arendt, *Between Past and Future* (pp. 227–264). Viking, 1968.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596.

Austin, J. L. (1962). *How to do things with words* (J. O. Urmson & M. Sbisà, Eds.). Oxford University Press.

Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073. Anthropic.

Becker, H. S. (1986). *Writing for social scientists: How to start and finish your thesis, book, or article*. University of Chicago Press.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). ACL.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT '21* (pp. 610–623). ACM.

Berlin, I. (1958). Two concepts of liberty. In I. Berlin, *Four Essays on Liberty* (pp. 118–172). Oxford University Press, 1969.

Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

BonJour, L., & Sosa, E. (2003). *Epistemic justification: Internalism vs. externalism, foundations vs. virtues*. Blackwell.

Borges, J. L. (1944). *Ficciones*. Sur.

Calvino, I. (1979). *Se una notte d'inverno un viaggiatore*. Einaudi.

Cammarano, C. (2018). *Le fake news e il marketing del vero*. De Agostini Libri.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford University Press.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.

Chang, T. A., & Bergen, B. K. (2023). Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1). https://arxiv.org/abs/2303.11504

Coady, C. A. J. (1992). *Testimony: A philosophical study*. Oxford University Press.

Eco, U. (1977). *Come si fa una tesi di laurea*. Bompiani.

Eco, U. (1984). *Semiotics and the philosophy of language*. Indiana University Press.

Fann, K. T. (1970). *Peirce's theory of abduction*. Martinus Nijhoff.

Ferraris, M. (2009). *Documentalità: Perché è necessario lasciar tracce*. Laterza. [English translation: *Documentality: Why It Is Necessary to Leave Traces* (R. Davies, Trans.). Fordham University Press, 2013.]

Feyerabend, P. K. (1975). *Against method*. New Left Books.

Floridi, L. (2011). *The philosophy of information*. Oxford University Press.

Floridi, L. (2015). Semantic conceptions of information. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/information-semantic/

Foot, P. (1972). Morality as a system of hypothetical imperatives. *The Philosophical Review*, 81(3), 305–316.

Frankfurt, H. G. (2005). *On bullshit*. Princeton University Press.

Gadamer, H.-G. (1960). *Truth and method* (J. Weinsheimer & D. G. Marshall, Trans., 2nd rev. ed.). Crossroad, 1989.

Gibbard, A. (2003). *Thinking how to live*. Harvard University Press.

Glanzberg, M. (2018). Truth. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/truth/

Goldman, A. I. (1999). *Knowledge in a social world*. Oxford University Press.

Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, T. B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., & Lindsey, J. (2026). Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*. https://transformer-circuits.pub/2026/workspace/index.html

Hacking, I. (1965). *Logic of statistical inference*. Cambridge University Press.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.

Haidt, J., & Rausch, Z. (2023). *The anxious generation*. Penguin Press.

Ichikawa, J. J., & Steup, M. (2018). The analysis of knowledge. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/knowledge-analysis/

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2022). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196–217.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press.

Kripke, S. A. (1980). *Naming and necessity*. Harvard University Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). University of Chicago Press.

Kvanvig, J. L. (2003). *The value of knowledge and the pursuit of understanding*. Cambridge University Press.

Lakatos, I. (1978). *The methodology of scientific research programmes* (J. Worrall & G. Currie, Eds.). Cambridge University Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

Liang, P., et al. (2022). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 3214–3252). ACL.

Lindsey, J. (2025). Emergent introspective awareness in large language models. *Transformer Circuits Thread*. https://transformer-circuits.pub/2025/introspection/index.html

Lindsey, J., et al. (2025). On the biology of a large language model. *Transformer Circuits Thread*. https://transformer-circuits.pub/2025/biology-of-a-lm/index.html

Loru, E., Nudo, J., Di Marco, N., Santirocchi, A., Atzeni, R., Cinelli, M., Cestari, V., Rossi-Arnaud, C., & Quattrociocchi, W. (2025). The simulation of judgment in LLMs. *Proceedings of the National Academy of Sciences*, 122(42), e2518443122. https://doi.org/10.1073/pnas.2518443122

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon Books.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13), e2215907120.

Moro, A. (1997). *The raising of predicates: Predicative noun phrases and the theory of clause structure*. Cambridge University Press.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.

OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, 3(2), 173–182.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Peirce, C. S. (1903). *Harvard lectures on pragmatism*. Reprinted in C. S. Peirce, *Collected Papers*, Vol. 5, §5.14–5.212. Harvard University Press, 1934.

Peirce, C. S. (1931–1958). *Collected papers of Charles Sanders Peirce* (Vols. 1–8; C. Hartshorne, P. Weiss, & A. W. Burks, Eds.). Harvard University Press.

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. arXiv preprint arXiv:2208.02957.

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson.

Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.

Quattrociocchi, W., Capraro, V., & Perc, M. (2025). Epistemological fault lines between human and artificial intelligence. arXiv preprint arXiv:2512.19466. https://arxiv.org/abs/2512.19466

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, 60(1), 20–43.

Quine, W. V. O. (1969). Epistemology naturalized. In W. V. O. Quine, *Ontological relativity and other essays* (pp. 69–90). Columbia University Press.

Reichenbach, H. (1949). *The theory of probability*. University of California Press.

Ryle, G. (1949). *The concept of mind*. Hutchinson.

Searle, J. R. (1976). A taxonomy of illocutionary acts. *Minnesota Studies in the Philosophy of Science*, 7, 344–369.

Searle, J. R. (1995). *The construction of social reality*. Free Press.

Sellars, W. (1956). Empiricism and the philosophy of mind. In H. Feigl & M. Scriven (Eds.), *Minnesota Studies in the Philosophy of Science* (Vol. 1, pp. 253–329). University of Minnesota Press.

Shapiro, S. (2000). *Thinking about mathematics: The philosophy of mathematics*. Oxford University Press.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Srivastava, A., et al. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Steup, M., & Neta, R. (2020). Epistemology. In E. N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/epistemology/

Strunk, W., & White, E. B. (1999). *The elements of style* (4th ed.). Longman.

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, 30. Curran Associates.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171. https://arxiv.org/abs/2203.11171

Williamson, T. (2007). *The philosophy of philosophy*. Blackwell.

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*, 29(4), 534–561. https://doi.org/10.1075/ijcl.23087.yu

Zagzebski, L. (2001). Recovering understanding. In M. Steup (Ed.), *Knowledge, truth, and duty* (pp. 235–252). Oxford University Press.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 36. https://arxiv.org/abs/2306.05685

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, 50(1), 237–291. https://doi.org/10.1162/coli_a_00502

Bibliography updates (v6): Gurnee, Sofroniew, Lindsey et al. (2026) added (§3.3, §6.4, §8.5), verified against the published page and positioned after Goldman. Lindsey (2025), *Emergent introspective awareness in large language models*, added as a distinct entry (§3.3): the v5 list contained only *On the biology of a large language model*, which is a different paper and does not support the claim §3.3 of v5 attached to it. The biology entry is retained, its venue corrected from "Anthropic Technical Report" to *Transformer Circuits Thread*, and it is no longer cited in §3.3. Prior updates (v5): Loru et al. (2025) — full nine-

author list restored (verified against PNAS DOI 10.1073/pnas.2518443122) and moved to correct alphabetical position; Quattrociocchi, Capraro & Perc (2025) — verified against arXiv:2512.19466 and repositioned before Quine; Sellars (1956) added (§4.6, Myth of the Given). Prior corrections (v3): Ji et al. (2022, not 2023); Chang & Bergen (2023); Ferraris (2009/2013); Chalmers (1995); Anthropic → Lindsey et al. (2025); Barbero removed; Wittgenstein §293 → §246, §201 → §150; new entries Wang et al. (2022), Zheng et al. (2023). Total: 67 entries.

Scope of this list. The References above are the works cited in the paper. They are not the project's full bibliography, which is maintained separately as epistemologia_ai_bibliografia_v6.bib and deposited with the open materials (Appendix G.1). The extended file records the works consulted across the research programme, including sources read and set aside, and is therefore longer than this list by design. The two are not intended to converge: the printed list is a citation record, the deposited file a reading record. Anyone replicating the study should work from the References; anyone extending the taxonomy will find the deposited file more useful.

Appendix A. Ontological Taxonomy v2.2 — Summary Reference

This appendix provides a condensed reference for the seven-dimensional framework. Full definitions, examples, and boundary conditions are in Section 5. The Coding Manual (Appendix B) provides the operational annotation protocol.

A.1 Gateway Procedures

Exclusion Rule — Pure Existence Predicates. Claims of the form 'X exists' or 'X does not exist' are outside the framework's scope. Rationale: existence is not a first-order predicate (Kant, Frege, Quine); the verb 'to be' conflates copula, identity, and existence in natural language (Moro, 1997); existence-negation claims are in principle unfalsifiable.

Pre-Step 0 — Illocutionary Act Verification (v2.2). Before any D1 classification is assigned, verify that the statement is being used as an assertion. Declarative form is treated as assertive function by default. When contextual markers clearly indicate non-assertive illocutionary force (imperative, performative, apophatic), assign code NA_ILLOCUTORIO and suspend validation. Theoretical grounding: Austin (1962). Founding case: Zen koan 'You must hear the sound of one hand clapping.'

A.2 D1 — Content Type (Primary Taxonomic Dimension)

Code	Category	Defining feature	Canonical validation outcome
D1.1	Historical-Factual	Singular past events, documentable	Variable (decidable / contested by evidence)
D1.2	Empirical-	Observation,	VERO / FALSO /

Code	Category	Defining feature	Canonical validation outcome
	Quantitative	measurement, data collection	PROB. FALSO (falsifiable)
D1.3	Logical-Formal	Truth guaranteed by logical structure	VERO (certain, a priori)
D1.4	Mathematical-Analytic	Truth from definitions / axioms	VERO (certain, a priori)
D1.5	Phenomenological-Subjective	First-person experiential states (cf. PI §246)	VERO for experiencer / VARIABILE intersubjectively
D1.6	Interpretive-Hermeneutic	Meaning and significance of texts / acts	PARZIALMENTE / DIPENDE (interpretation-relative)
D1.7	Aesthetic-Evaluative	Beauty, artistic quality, taste	DIPENDE / PARZIALMENTE (tradition-relative)
D1.8	Ethical-Normative	Moral obligation, justice, deontic force	DIPENDE (framework-dependent by type)
D1.9	Procedural-Instructional	How-to instructions, techniques (cf. PI §150)	VERO / FALSO (efficacy-based)
D1.10	Prescriptive-Policy	Institutional recommendations, governance	DIPENDE / PROB. (efficacy + values)
D1.14	Metaphysical-Ontological	Fundamental nature / structure of reality	DIPENDE / SOSPESO / STRUTTURATO
NA_ILLOCUTORIO	Non-Assertive	Not used as assertion (Pre-Step 0)	Validation suspended

A.3 D2–D7 — Attributive Dimensions

Dim.	Name	Question asked	Values	Redundancy re D1
D2	Temporal Scope	What time horizon?	D2.1 Past-Singular / D2.2 Past-General / D2.3 Present /	Semi-determined (7/11 fixed)

Dim.	Name	Question asked	Values	Redundancy re D1
			D2.4 Timeless	
D3	Certainty Level	How confident?	D3.1 Certain / D3.2 High-Confidence / D3.3 Probabilistic / D3.4 Speculative / D3.5 Unknowable	Independent but subjective (4/11 fixed)
D4	Causality	Is a causal claim asserted?	D4.1 No-Causal / D4.2 Correlational / D4.3 Causal-Mechanism / D4.4 Direction-Contested	Most independent (6/11 fixed; highest value)
D5	Framework-Dependency	Does truth depend on framework?	D5.1 Framework-Dependent / D5.2 Framework-Constitutive / D5.3 Framework-Independent	Semi-determined (7/11 fixed)
D6	Validation Method	What method is appropriate?	D6.1 Testimony / D6.2 Documentary / D6.3 Observation / D6.4 Social-Science / D6.5 Institutional / D6.6 Phenomenology / D6.7 Philosophical / D6.8 Aesthetic	High redundancy (7/11 fixed) — pedagogical retention
D7A	Falsifiability (Principle)	Can it be falsified in principle?	D7A.1 Principally Unfalsifiable /	Semi-determined (7/11 fixed)

Dim.	Name	Question asked	Values	Redundancy re D1
			D7A.2 Empirically Falsifiable / D7A.3 Philosophically Unfalsifiable	
D7B	Falsifiability (Practice)	Can it be tested now?	D7B.1 Currently Falsifiable / D7B.2 Conditionally / D7B.3 Not Currently	Maximal redundancy (11/11) — pedagogical retention

A.4 D4.3 Causal-Mechanism — Asymmetric Validation Procedure

Outcome	Conditions	Founding case
1 — FALSIFIED	Causal premise is false; mechanism absent, inverted, or non-operative; or true conclusion attributed to false / unwarranted premise	Monaco case: migration emergency → healthcare collapse. B true; A false; link undemonstrated.
2 — NON-CORROBORATED	Causal claim not falsified but unsupported by evidence of the mechanism. Neither A nor B necessarily false.	Social media use → adolescent depression. Both documented; causal mechanism actively debated (Orben & Przybylski, 2019).
3 — BEYOND FRAMEWORK SCOPE	Causal complexity exceeds what the protocol can adjudicate without specialized domain knowledge.	Globalization → decline of Western middle class. Multi-factorial; relative causal weights disputed across incompatible economic models.

A.5 Epistemic Responsibility Check (ERC)

Criterion	Sì (Yes)	Parziale (Partial)	No
Cita fonti/evidenze?	'According to a 2020 Nature study, X causes Y.'	'Studies show that X causes Y.' (vague)	'X causes Y.' (no justification)
Esprime incertezza?	'Evidence suggests X, though further	'X is probably true.'	'X is certainly true.' (when uncertainty

Criterion	Sì (Yes)	Parziale (Partial)	No
	research is needed.'		exists)
Distingue interp. da fatto?	'Data show correlation between X and Y, which I interpret as…'	'Data show that X causes Y.' (not flagged)	'X causes Y.' (correlation as causation without flag)
ESITO ERC	RESPONSABILE	PARZIALMENTE RESPONSABILE	IRRESPONSABILE

Appendix B. Coding Manual v2.1 — Operational Summary

This appendix provides an operational summary of the Coding Manual v2.1, which supersedes v2.0 and is the version deposited for Phase 2. The relation between the two versions is additive. The horizontal validation (Section 5) followed v2.0; v2.1 adds the executed six-claim scope, the ARlex / ARsem definitions, the Mode A (automated) / Mode M (manual) distinction with mandatory spot-validation, and the QC-logging requirement used in the vertical study. No annotation rule applied in the horizontal validation was revised, so the thirty-one annotations of Section 5 stand as conducted and are inherited by Phase 2 without a version caveat. The full document covers ambiguity resolution rules, quality checks, and troubleshooting. The manual applies to the horizontal validation (Section 5); the vertical validation used the structured prompt in Appendix C.

B.1 Validation Output Format

Section	Content	Format
PRE-STEP 0	Illocutionary act verification; composite claim flag; semantic ambiguity note	Narrative paragraph
PROSPETTO DI VALIDAZIONE	Six-row summary: Ontologia / Logica / Decidibilità / Esito (bold) / Pre-Step 0 / Sintesi rapida	Markdown table, exactly 6 rows
[EMPATIA]	Phenomenological and affective dimension of the claim	Narrative paragraph
[LOGICA]	Logical form analysis; vulnerabilities; defensible weaker versions	Narrative paragraph(s)
[INFORMAZIONE]	Substantive thesis; empirical data; counterexamples; main methodological objection	Narrative paragraph(s)
[EPISTEMIC	Four-row ERC table (see	Markdown table, exactly 4

Section	Content	Format
RESPONSIBILITY CHECK]	Appendix A.5)	rows
[SINTESI ESTESA]	Strong/weak form distinction; final verdict with conditions	Narrative paragraph(s)
FONTI	Bibliographic list	One entry per line

B.2 Primary Variables for Horizontal Annotation

Variable	Definition	Location	Format
V1 — REPLICATION_ID	Unique identifier	Annotator-assigned	[Claim]_R[N]
V2 — VERDICT	Final epistemic verdict	PROSPETTO: Esito row	VERO / FALSO / PROB_FALSO / DIPENDE / INCERTO / INDECIDIBILE / STRUTTURATO
V3 — CONFIDENCE	Explicit confidence %, if stated	Any section	Integer 0–100 or NA
V4–V6 — TOP 3 ARGUMENTS	Main arguments supporting verdict	SINTESI ESTESA or INFORMAZIONE	5–15 word paraphrase
V7–V9 — TOP 3 SOURCES	Principal sources cited	FONTI section	Author (Year)
V10–V12 — KEY QUANTITATIVE DATA	Up to 3 key numbers	Any section	Variable_name = Value (exact)
V13–V15 — FRAMEWORKS (D1.8 only)	Ethical frameworks named	Any section	Utilitarianism / Deontology / Virtue_Ethics / Rights_Based / Care_Ethics / Contractualism / Consequentialism / Natural_Law / Other / NA

B.3 Decision Rules for Ambiguous Cases

Scenario	Rule
A: Verdict conflicts between PROSPETTO and SINTESI ESTESA	PROSPETTO takes precedence.
B: Multiple verdicts for different frameworks (D1.8 claims)	Code ‘overall’ verdict. If none stated: code DIPENDE.

Scenario	Rule
C: Similar arguments across replications	Use consistent paraphrasing for genuinely similar arguments.
D: Numeric precision variance across replications	Code exactly as stated. Variance is the signal — do not normalize.
E: NA_ILLOCUTORIO detected at Pre-Step 0	Suspend validation. Code Esito = NA_ILLOCUTORIO.
F: D4.3 claim — three-outcome rule	Do not code true/false. Declare Outcome 1, 2, or 3 explicitly.

Appendix C. Vertical Validation — System Prompts and API Configuration

This appendix contains: (C.1) the six claims submitted; (C.2) the API configuration for the primary vertical validation; (C.3) the API configuration for the multi-model control (Section 6.6); (C.4) the naïf prompt for the uplift experiment (Section 6.7).

SYSTEM PROMPT — PRIMARY VALIDATION (verbatim)

You are an epistemic validation engine operating within the 'Validare Claude with Claude' framework (Claudio Cammarano, 2026). Apply the following locked format exactly, in this order:

1. PRE-STEP 0 — Narrative paragraph. Verify illocutionary act (Austin 1962): is the statement used as an assertion? If not, code NA_ILLOCUTORIO. If composite claim (A+B), flag it.
2. PROSPETTO DI VALIDAZIONE — Markdown table, 2 columns, exactly 6 rows: Ontologia | value / Logica | value / Decidibilità | value / Esito | value / Pre-Step 0 | value / Sintesi rapida | value. Do NOT use D1/D2/D3 labels.
3. [EMPATIA] — Phenomenological and affective dimension.
4. [LOGICA] — Logical form: form type, main vulnerabilities, weaker defensible versions.
5. [INFORMAZIONE] — Substantive thesis, empirical data, counterexamples, main methodological objection. NOT the claim's linguistic form.
6. [EPISTEMIC RESPONSIBILITY CHECK] — Markdown table, 2 columns, exactly 4 rows: Cita fonti/evidenze? | answer / Esprime incertezza? | answer / Distingue interp. da fatto? | answer / ESITO ERC | verdict + one sentence.
7. [SINTESI ESTESA] — Strong/weak form distinction. Final verdict and conditions.
8. FONTI — Bibliographic list.

USER MESSAGE (template, verbatim)

Valida questo claim: "[CLAIM TEXT]"

C.1 Six Claims Submitted (Primary Validation)

ID	Claim (Italian)	Claim (English)	D1
vaccini	I vaccini causano l'autismo	Vaccines cause autism	D1.2, D4.3
vegetarianismo	Il vegetarianismo è eticamente superiore	Vegetarianism is ethically superior	D1.8
bitcoin	Bitcoin raggiungerà $100k entro il 2026	Bitcoin will reach $100k by 2026	D1.4
prima_guerra	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
coscienza	La coscienza non è riducibile a processi neurali	Consciousness is not reducible to neural processes	D1.14
fumo	Fumare causa il cancro ai polmoni	Smoking causes lung cancer	D1.2, D4.3

C.2 API Configuration — Primary Validation (claude-sonnet-4-6)

Parameter	Value
Model	claude-sonnet-4-6
Max tokens	4000
Temperature	API default
System memory across runs	None (each run independent)
Replications per claim	30
Total runs	180
Skipped runs	0
Retry logic	5 retries per failed run; 30-second delay
Checkpoint-resume	Enabled (checkpoint.json)
Verdict extraction	Automated (Mode A), label-first rule with negation handling; deposited (analisi_varianza_v2.py)
Execution	25–26 June 2026; batch_validazione.py

C.3 API Configuration — Multi-Model Control [PHASE 1 — TO EXECUTE]

The multi-model control (Section 6.6) applies the identical system prompt to a second model. Configuration below is to be completed at execution.

Parameter	Value
Model	[MODEL-2 — to be specified: GPT-4o / open-weights equivalent]
API endpoint	[To be specified]
Max tokens	4000
Temperature	API default (same as primary)
System prompt	Identical to Appendix C — Primary (verbatim)
User message	Identical to Appendix C — Primary (verbatim)
Claims	vaccini + vegetarianismo only
Replications per claim	30
Total runs	60
Rationale for claim selection	vaccini: highest ARsem (100%) — tests whether determinacy replicates cross-model; vegetarianismo: SOSPESO-dominant D1.8 (70%) — tests whether the framework-dependent distributional signature replicates
Execution	[Date to be added]

C.4 Naïf Prompt — Uplift Experiment [PHASE 1 — TO EXECUTE]

The uplift experiment (Section 6.7) contrasts the framework system prompt (Appendix C, primary) with the following naïf prompt. The naïf prompt asks for evaluation without any ontological pre-classification, without the D4.3 asymmetric procedure, and without the ERC. It represents standard LLM use for claim evaluation.

NAÏF SYSTEM PROMPT (verbatim)

You are a careful and knowledgeable assistant. When given a claim, evaluate it as follows:

1. State whether the claim is true, false, or uncertain, and explain why.
2. Cite any relevant evidence or arguments.
3. Note any important qualifications or nuances.

Be clear, accurate, and balanced.

NAÏF USER MESSAGE (template, verbatim)

Evaluate this claim: “[CLAIM TEXT]”

Note on blind coding: the two-prompt outputs will be randomized before scoring. The scorer receives unlabelled outputs and codes the binary flag (does the output identify and flag a structural causal defect?) without knowledge of which prompt produced each output.

The coding key is committed to prior to scoring. See Appendix D.5 for the pre-registration of H_uplift.

Appendix D. Pre-Registration Summary

The vertical validation design was pre-registered with OSF prior to data collection. This appendix reproduces the summary. D.5 contains the pre-registration of the uplift experiment, deposited prior to execution.

D.1 Research Questions

RQ1: Are the epistemic validation outputs produced by Claude (claude-sonnet-4-6) reproducible across repeated identical-prompt replications?

RQ2: Does the degree of reproducibility vary systematically across ontological claim types as classified by the D1 taxonomy?

D.2 Hypotheses

H1: The model will show high semantic agreement (ARsem ≥ 70%) on claims with determinate, framework-independent epistemic outcomes (D1.2 massively refuted, D1.3/D1.4).

H2: Lexical agreement (ARlex) will be substantially lower than semantic agreement (ARsem), with a gap of at least 40 percentage points on average.

H3: Semantic agreement will be lower and more variable for claims whose epistemic indeterminacy is structural (D1.8, D1.14) than for claims with determinate outcomes.

D.3 Design

Element	Specification
Claims	6 (Appendix C.1); maximal epistemic contrast across D1 types
Replications	30 per claim = 180 total
Model	claude-sonnet-4-6, API default parameters
Prompt	Identical across all replications; no memory across runs
Primary metrics	ARlex, ARsem, $\bar{J}$ (pairwise Jaccard), IC composite index
Success criterion	Statistically meaningful difference in ARsem between framework-dependent and framework-independent claim types
Pre-registration date	Prior to data collection, Spring 2026
OSF repository	[DOI to be added at submission]

D.4 Pre-Registration Amendment

The original pre-registration specified Fleiss' κ as the primary agreement metric. Prior to data collection, it was determined that κ statistics assume a fixed coding scheme applied by multiple independent annotators, whereas the vertical validation involves a single model. The metrics were revised to ARlex, ARsem, $\bar{J}$, and IC. This amendment is documented here in full accordance with open science protocols.

D.5 Pre-Registration — Uplift Experiment [PHASE 1 — DEPOSIT PRIOR TO EXECUTION]

Research question: does the framework prompt elicit epistemically superior output — specifically, does it flag structural causal defects at a higher rate than the naïf prompt?

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 ($\Delta > 0$, Fisher's exact test, $p < 0.05$) for at least three of the four claims in the uplift corpus.

Element	Specification
Claims (uplift corpus)	4 D4.3 causal claims: (1) migrazione→sanità (Monaco case; Outcome 1); (2) vaccini→autismo (Outcome 1); (3) social media→depressione (Outcome 2); (4) globalizzazione→ceto medio (Outcome 3)
Conditions	Framework prompt (Appendix C, primary) vs. Naïf prompt (Appendix C.4)
Replications	20 per claim per condition = 160 total runs
Scoring	Binary (0/1): does the output identify the claim as a causal claim AND flag a structural defect in the causal premise?
Scoring procedure	Blind: outputs randomized and stripped of condition label before scoring; coder commits to coding key before examining any output
Primary metric	Δ flag rate = Framework% − Naïf% per claim, with 95% Wilson CI
Statistical test	Fisher's exact test per claim; Bonferroni correction for four tests ($\alpha = 0.0125$ per test)
Success criterion	$\Delta > 0$ and significant for ≥ 3/4 claims
Pre-registration date	[To be deposited at OSF prior to execution]
OSF repository	[DOI to be added]
Outcome handling	If Δ positive and significant: uplift confirmed — central contribution of Phase

Element	Specification
	1 empirical section. If null/negative: honest negative finding, reported as a limit discovered before peer review, does not invalidate the theoretical framework.

Appendix E. Horizontal Validation — Complete Outcome Table

Full annotations available in the companion document *Archivio 31 Validazioni v4 Estesa* (Cammarano, 2026), available from the author and at the OSF repository (Appendix G).

Note on Claim 27: The statement 'La Crimea è Russia' was attributed to the Italian historian Alessandro Barbero on the basis of a public television statement, not a published academic work. It is cited here as a motivating example illustrating the D1.1/D1.10 boundary: a recognized academic expert producing a structurally ambiguous geopolitical claim in a mass-media context, without specifying which sense of 'is' (factual, juridical, or de facto) they intend. The case is epistemically significant precisely because the speaker's authority in medieval history does not transfer to the geopolitical-legal domain the statement invokes. No bibliographic entry for Barbero is included in the References.

N	Claim (Italian / English)	D1	D4	Esito
1	L'aborto è moralmente sbagliato / Abortion is morally wrong	D1.8	D4.1	DIPENDE (framework etico)
2	È moralmente obbligatorio aiutare i rifugiati / It is morally obligatory to help refugees	D1.8	D4.1	DIPENDE (cosmopolitismo vs. comunitarismo)
3	La pena di morte è ingiusta / The death penalty is unjust	D1.8	D4.1	DIPENDE (teoria della giustizia)
4	L'AI sostituirà il 40% dei lavori entro il 2030 / AI will replace 40% of jobs by	D1.4 Pred.	D4.1	PROBABILMENTE FALSO

N	Claim (Italian / English)	D1	D4	Esito
	2030			
5	L’acqua è H_2O / Water is H_2O	D1.4/D1.2	D4.1	VERO (necessità a posteriori — Kripke)
6	I celibi non sono sposati / Bachelors are unmarried	D1.3	D4.1	VERO (analitico a priori)
7	La democrazia richiede libertà di stampa / Democracy requires freedom of the press	D1.5/D1.10	D4.1	CONDIZ./ PARZIALMENTE VERO
8	Il Barolo ha sentori di rosa e catrame / Barolo has notes of rose and tar	D1.5	D4.1	VERO per esperti / VARIABILE per neofiti
9	Il triangolo ha tre lati / A triangle has three sides	D1.3	D4.1	VERO (necessariamente, per definizione)
10	Il capitalismo sfrutta i lavoratori / Capitalism exploits workers	D1.8/D1.2	D4.2	DIPENDE (def. ‘sfruttamento’)
11	Il vegetarianismo è eticamente superiore / Vegetarianism is ethically superior	D1.8	D4.1	DIPENDE (framework morale)
12	Il rosso è più caldo del blu /	D1.5	D4.1	FALSO fisicamente /

N	Claim (Italian / English)	D1	D4	Esito
	Red is warmer than blue			VERO psicologicamente
13	Questo mondo è il migliore dei mondi possibili / This world is the best of all possible worlds	D1.14	D4.1	FALSO (problema del male)
14	Bitcoin raggiungerà $100k entro il 2026 / Bitcoin will reach $100k by 2026	D1.4 Pred.	D4.1	VERO (verificato a posteriori, dic. 2024)
15	La bellezza è negli occhi di chi guarda / Beauty is in the eye of the beholder	D1.7	D4.1	PARZIALMENTE VERO
16	Il terrapiattismo è una teoria credibile / Flat-earthism is a credible theory	D1.2	D4.1	FALSO (massimamente refutato)
17	Gli alieni hanno visitato la Terra / Aliens have visited Earth	D1.2	D4.1	PROBABILMENTE FALSO (P < 10%)
18	Il cambiamento climatico è una bufala inventata dalla Cina / Climate change is a hoax invented by China	D1.2+D1.1	D4.3	FALSO (refutato su A e B)
19	I vaccini	D1.2	D4.3	FALSO — D4.3

N	Claim (Italian / English)	D1	D4	Esito
	causano l’autismo / Vaccines cause autism			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	D4.1	DIPENDE (concezione vita buona)
21	I Vangeli non incoraggiano l’empatia universalistica / The Gospels do not encourage universalistic empathy	D1.6+D1.1	D4.1	PARZIALMENTE FONDATO
22	Chi non è flessibile fisicamente non lo è mentalmente / Those not physically flexible are not mentally flexible either	D1.2	D4.2	FALSO nella forma universale (controesempio: Hawking)
23	La matematica non è reale / Mathematics is not real	D1.14	D4.1	DIPENDE (framework ontologico)
24	Non ci sono fatti; solo interpretazioni / There are no facts; only interpretations	D1.14+D1.3	D4.1	FALSO (autoconfutazione performativa)
25	Devi sentire il	NA_ILLOCUTOR	—	NA_ILLOCUTOR

N	Claim (Italian / English)	D1	D4	Esito
	rumore che fa una mano sola / You must hear the sound of one hand clapping	IO		IO
26	Le decisioni della cartina politica sono reali / The decisions of the political map are real	D1.14	D4.1	VERO (fatti istituzionali — Searle)
27	La Crimea è Russia (Barbero*) / Crimea is Russia	D1.1+D1.10	D4.1	STRUTTURATO: FALSO giuridico / FONDATO storico / VERO de facto
28	Quando Gesù è morto aveva quasi cinquant'anni / When Jesus died he was nearly fifty	D1.1	D4.1	MOLTO PROBABILMENTE FALSO (consensus: 33–36 aa)
29	Gesù è veramente risorto / Jesus truly rose from the dead	D1.14+D1.2	D4.1	STRUTTURATO: INDECIDIBILE (stor.) / PROB. FALSO (natur.) / OLTRE PORTATA (metafis.)
30	La statua di Civitavecchia è un autentico miracolo / The Civitavecchia statue is an authentic miracle	D1.2+D1.14	D4.3	STRUTTURATO: fenomeno non spiegato / causa miracolosa OLTRE PORTATA

N	Claim (Italian / English)	D1	D4	Esito
31	Ogni esistenzialismo è un umanismo / Every existentialism is a humanism	D1.14	D4.1	MOLTO PROBABILMENTE FALSO (Heidegger, Kierkegaard)

* Claim 27 attributed to Alessandro Barbero on the basis of a public television statement; not a citable scholarly source. See note above.

Appendix F. Dimensional Redundancy Analysis — Matrice Integrata

The table below reports the redundancy analysis for dimensions D2–D7 relative to D1. D6 and D7B are retained for pedagogical reasons despite high redundancy; the case is argued in Section 5.3.

Dimension	Fixed values (predictable from D1)	Where variability exists	Ratio	Verdict
D2 Temporal Scope	D1.3/D1.4/D1.14 → Atemporal; D1.5 → Present; D1.1 → Past	D1.2 (present or timeless); D1.6 (any register); D1.8–D1.10 (margin)	7/11 fixed	Semi-determined
D3 Certainty Level	D1.3/D1.4 → Certain; D1.14 → Speculative; D1.5 → Certain/Unknowable	D1.2, D1.6, D1.7, D1.8, D1.10 (genuine variability; judgment subjective)	4/11 fixed	Independent but fragile (low κ expected)
D4 Causality	D1.3/D1.4/D1.5/D1.6/D1.7 → No-Causal; D1.9 → Causal-Mechanism	D1.2, D1.8, D1.10, D1.14 (all 4 D4 subtypes possible)	6/11 fixed; 5/11 variable	MOST INDEPENDENT — highest attributive value
D5 Framework-Dependency	D1.3/D1.4 → Constitutive; D1.7/D1.8/D1.14 →	Critical boundary D5.1/D5.3 for contested	7/11 fixed	Semi-determined; value at margins

Dimension	Fixed values (predictable from D1)	Where variability exists	Ratio	Verdict
	Dependent; D1.2/D1.5 → Independent	empirical claims		
D6 Validation Method	D1.3/D1.4/ D1.8/D1.14 → Philosophical; D1.5 → Phenomenological; D1.7 → Aesthetic; D1.9 → Observation	D1.1, D1.2, D1.6, D1.10 (multiple valid methods)	7/11 fixed	HIGH REDUNDANCY — retained for pedagogy
D7A Falsifiability (Principle)	D1.3/D1.4/ D1.7/D1.8 → Phil. Unfalsifiable; D1.5 → Principally Unfalsifiable; D1.2/D1.9 → Empirically Falsifiable	D1.6 and D1.14 variable; D1.1 and D1.10 marginally	7/11 fixed	Semi-determined
D7B Falsifiability (Practice)	D1.2/D1.3/ D1.4/D1.9 → Currently; D1.5/D1.7/D1.8/D1.14 → Not Currently; D1.1/D1.6/D1.10 → Conditionally	Near zero — practically all values predictable from D1	11/11 predictable	MAXIMAL REDUNDANCY — pedagogical retention only

Appendix G. Open Materials — OSF Repository

All Phase 1 materials are deposited openly at the Open Science Framework (OSF) repository to enable independent replication, taxonomy extension, and Phase 2 inter-annotator studies. The repository is licensed under CC BY 4.0, which explicitly invites reuse, adaptation, and derivative work with attribution.

G.1 Repository Contents

Item	Description	Format
Coding Manual v2.1	Full operational annotation	PDF + DOCX

Item	Description	Format
	protocol including all decision rules, boundary cases, ambiguity resolution guidance, and the Mode A / Mode M coding-mode specification	
Tassonomia v2.2	Complete seven-dimensional framework specification with all D1–D7 definitions, examples, and boundary conditions	PDF + DOCX
31-claim horizontal corpus	All 31 claims with full annotations: Pre-Step 0, D1–D7 coding, D4.3 procedure where applicable, ERC, and Sintesi Estesa	DOCX (archivio_31_validazioni_v4_estesa)
System prompt (primary)	Verbatim system prompt used in all 180 vertical validation runs (Appendix C)	TXT
Naïf prompt (uplift)	Verbatim naïf system prompt for the uplift experiment (Appendix C.4)	TXT
batch_validazione.py	Python script for the primary vertical validation: checkpoint-resume, 5-retry logic, API integration	PY
analisi_varianza.py	Original Python script for ARlex, ARsem, and Jaccard computation from batch output (deposited classifier)	PY
analisi_varianza_v2.py	Label-first verdict extractor (negation handling + self-label priority): the revised Mode A pipeline adopted in resolution of F-3, deposited alongside the original so both deposited and label-first figures regenerate from the raw file	PY

Item	Description	Format
QC Log (vertical)	Retrospective quality-control log for the 180-run corpus: checks QC1–QC4, spot-validation record, flags F-1/F-2/F-3, sensitivity analysis, aggregate results after QC	PDF
risultati_batch.json	Raw output from all 180 API runs	JSON
analisi_varianza.json / analisi_varianza_labelfirst.json	Computed metrics per claim under the deposited classifier and under the label-first rule: ARlex, ARsem, IC95%, $\bar{J}$, IC	JSON
campione_cieco_18.md	Blind spot-validation packet: 18 outputs, three per claim, stratified, randomized, labels stripped, seed 20260721; answer key held separately by the author	MD
Pre-registration (primary)	OSF pre-registration for the vertical validation design (Appendix D.1–D.4)	OSF link
Pre-registration (uplift)	OSF pre-registration for the uplift experiment (Appendix D.5)	OSF link
epistemologia_ai_bibliografia_v6.bib	Extended project bibliography, UTF-8, Zotero-ready. Broader in scope than the printed References (see note to the References): records works consulted across the programme, not only those cited	BIB
Annotation template	Blank annotation template for Phase 2 annotators, pre-formatted for the 50–100 claim IAA sub-corpus	XLSX

G.2 OSF Repository

Repository URL: [to be added at submission]

DOI: [to be added at submission]

All scripts and data files are version-controlled. The repository will be updated as Phase 2 data become available.

G.3 Call for Annotators

The Phase 2 inter-annotator reliability study requires a minimum of three independent annotators. Researchers and graduate students with background in epistemology, philosophy of language, linguistics, or AI are invited to participate. Participation involves annotating a sub-corpus of 20–30 claims following the Coding Manual v2.1. Full materials and training annotations are available in the OSF repository.

Aspect	Details
Commitment	20–30 claim annotations following Coding Manual v2.1; estimated 10–15 hours total
Training provided	Annotated walkthrough, decision tree, and 5 calibration examples with discussion
Deliverable	Completed annotation files in the standard template (Appendix G.1)
Acknowledgment	Named acknowledgment in the Phase 2 paper
Co-authorship	Co-authorship offered to annotators contributing a minimum of one full sub-corpus (20+ claims), subject to standard authorship criteria
Contact	[Author email to be added at submission]
OSF	[Repository link to be added at submission]

Note: participation in the annotation study does not require prior knowledge of the framework beyond what is provided in the Coding Manual and training materials. Annotators will be onboarded through a calibration session before independent annotation begins.
