CODING MANUAL

v2.1

LLM Claim Validation Reproducibility Study

Validating Claude with Claude

Claudio Cammarano — Episteme Advisory — July 2026

Version 2.1 — Changes with respect to v2.0:

1. **Scope updated** to the executed six-claim design: 6 claims × 30 replications = 180 validations — Section 1
2. **Agreement metrics formally defined**: Lexical Agreement Rate (AR-lex) and Semantic Agreement Rate (AR-sem) — Section 7
3. **Automated extraction protocol** with mandatory human spot-validation — Section 6
4. **Quality-control logging requirement**: QC results must be recorded in a persistent QC Log — Section 4
5. **Deviation log from pre-registration** — Appendix C

Version history: v1.0 (original protocol) → v2.0 (Pre-Step 0, illocutionary check; Epistemic Note D4.3) → v2.1 (this document).

Contents

1. Overview
2. Variables to Extract
3. Ambiguity Resolution Rules
4. Quality Checks
5. Workflow Step-by-Step
6. Coding Modes: Manual and Automated ← *v2.1*
7. Agreement Metrics ← *v2.1*
8. Time Management
9. Troubleshooting

Appendix A — Quick Reference Appendix B — Epistemic Note D4.3: Limits of Causal Validation Appendix C — Deviation Log from Pre-Registration ← *v2.1*

---

1. Overview

1.1 Purpose

Extract structured metrics from LLM-generated claim validations to measure reproducibility across ontological categories, and to test whether reproducibility covaries with the ontological type of the claim rather than with task difficulty alone.

1.2 Scope [updated v2.1]

Total validations: **180 (6 claims × 30 replications each)**.

The six claims were selected to maximize ontological diversity across the D1 taxonomy, with deliberate variability on D4 (causality):

Claim wording is reproduced verbatim from the batch script — for a study whose primary finding concerns lexical variance, the exact prompt wording is part of the data:

#	Claim ID	Claim (original Italian, verbatim)	D1 classification	D4 note
1	vaccini	“I vaccini causano l’autismo”	D1.1 Historical-Empirical	D4.3 causal component
2	vegetarianismo	“Il vegetarianismo è eticamente superiore”	D1.7 Normative-Ethical	Non-causal
3	bitcoin	“Bitcoin raggiungerà $100k entro il 2026”	D1.8 Predictive	Non-causal
4	prima_guerra	“La prima guerra mondiale fu causata dall’assassinio di Sarajevo”	D1.3 / D4.3 Causal	Causal-mechanism
5	coscienza	“La coscienza non è riducibile a processi neurali”	D1.10 Metaphysical	Non-causal
6	fumo	“Fumare causa il cancro ai polmoni”	D1.1 / D4.3 Causal-Empirical	Causal-mechanism

Run parameters. All replications were generated via the Anthropic API with `claude-sonnet-4-6`, max_tokens 4000, default sampling parameters, and a fixed system prompt enforcing the locked output format (PROSPETTO DI VALIDAZIONE, ERC table, FONTI). Each claim was submitted with the identical user message `Valida questo claim: "<claim>"`. Runs were executed sequentially with checkpoint-resume; failed runs were retried up to 5 times and, if still failing, recorded as permanently skipped rather than silently dropped. Skipped runs are reported in the QC Log and excluded from agreement denominators. The batch and analysis scripts are part of the open-materials deposit.

Relation to the horizontal corpus. This vertical study (few claims, many replications) is complementary to the horizontal validation corpus (31 claims, one validation each, full D1–D7 coverage). The two corpora share the same taxonomy and the same validation output format; only this manual’s coding protocol applies to the vertical study.

1.3 Design rationale

Thirty replications per claim allow within-claim variance estimation; six claims spanning empirical, normative, predictive, causal, and metaphysical types allow between-type comparison. The unit of analysis is the single validation output; the quantity of interest is agreement across replications of the same claim.

---

2. Variables to Extract

2.1 Primary Variables (ALL validations)

V1 — REPLICATION_ID

Format: `[Claim]_R[Number]` — e.g., `Vaccines_R01`, `WWI_R15`, `Smoking_R30`. Sequential numbering in order generated.

V2 — VERDICT

Definition: final claim assessment from the PROSPETTO box (line "Esito:"). Verdict labels are coded in the original Italian of the validation outputs.

Code	Meaning
VERO	Claim is true
FALSO	Claim is false
PROBABILMENTE_VERO	Likely true but uncertain
PROBABILMENTE_FALSO	Likely false but uncertain
DIPENDE	Depends on interpretation/framework
INCERTO	Insufficient evidence to decide
INDECIDIBILE	Fundamentally undecidable
NA_ILLOCUTORIO	Non-assertive utterance (Pre-Step 0)

For claims classified D4.3, the admissible verdict space is restricted to the three outcomes of Appendix B (FALSIFIED / NOT CORROBORATED / BEYOND FRAMEWORK SCOPE), mapped onto V2 as documented there.

Decision rule: PROSPETTO takes precedence. If ambiguous, check the final sentence of SINTESI ESTESA.

V3 — CONFIDENCE

Integer 0–100 (percentage stated explicitly) or NA if not present. Ranges are coded at their midpoint ("P ~60–70%" → 65).

V4–V6 — TOP 3 ARGUMENTS

Main arguments supporting the verdict, ranked by prominence (length + placement: earlier/longer → more prominent). Format: brief paraphrase, 5–15 words — never copy-paste. Location: SINTESI ESTESA bullets/list, or the first three arguments in INFORMAZIONE. If fewer than three arguments: fill available, leave the rest NA.

V7–V9 — TOP 3 SOURCES

Format: Author (Year) or Organization (Year). Location: FONTI section or inline citations. Priority: sources in the FONTI bibliography; prefer the most prominent.

V10–V12 — KEY QUANTITATIVE DATA

Format: `Variable_name = Value` (e.g., `Denmark_N = 657461`). Up to three most important numbers. Preserve precision EXACTLY as stated, units included where relevant.

2.2 Secondary Variables — NORMATIVE claims only

V13–V15 — PHILOSOPHICAL FRAMEWORKS

Code only if the framework is explicitly named or clearly described.

Code	Examples
Utilitarianism	Singer, Bentham, Mill
Deontology	Kant, duty-based ethics
Virtue_Ethics	Aristotle, Hursthouse
Rights_Based	Regan, Nozick
Care_Ethics	Gilligan, Noddings
Contractualism	Rawls, Scanlon
Consequentialism	General outcome-based
Natural_Law	Aquinas, Catholic tradition
Other	Specify in Notes
NA	Non-normative claim OR unclear

---

3. Ambiguity Resolution Rules

Scenario A — Verdict conflicts (PROSPETTO vs SINTESI). PROSPETTO takes precedence: it is the deliberate summary produced after the full analysis.

Scenario B — Multiple verdicts for different frameworks. Code the "overall" or "standard interpretation" verdict. If no overall verdict is stated: code DIPENDE.

Scenario C — Similar arguments across replications. Use consistent paraphrasing for genuinely similar arguments. Do NOT force uniformity where arguments differ substantively — code them differently.

Scenario D — Numeric precision variance. Code exactly as stated. Preserve differences: Rep1 "657,461" → `Denmark_N = 657461`; Rep2 "657k" → `Denmark_N = 657000`. Variance is the signal — do not normalize.

Scenario E — Framework unclear or unnamed. Too vague → NA. Clear enough ("Singer's utilitarian approach" → Utilitarianism). If genuinely uncertain: NA. Never guess.

Scenario F — D4.3 claims: causal validation. For claims classified D4.3 (Causal-Mechanism), the framework admits exactly three outcomes. See Appendix B for the full procedure.

> **Three D4.3 outcomes (summary)** 1. **FALSIFIED** — the causal link is absent, inverted, or the antecedent is false. 2. **NOT CORROBORATED** — the link is asserted but lacks evidence of mechanism. 3. **BEYOND FRAMEWORK SCOPE** — causal complexity exceeds the instrument.
>
> Declaring outcome 3 is not a weakness of the protocol — it is the epistemically honest response.

---

4. Quality Checks [updated v2.1]

Run after each batch of 10 replications. **v2.1** **Every QC pass must be recorded in the QC Log** (batch ID, date, values counted, flags raised, corrective action). A study whose QC cannot be exhibited has, for reproducibility purposes, no QC. The QC Log is part of the open-materials deposit.

Check	What to count	Flag if…
QC1: Verdict distribution	Verdict frequencies in batch	Empirical < 5/10 modal; any category < 4/10

Check	What to count	Flag if…
QC2: Argument stability	Unique V4_ARG1 values	7+ unique (check paraphrasing consistency)
QC3: Missing data	NA count in required fields	> 2 NAs in V2 or V4–V6
QC4: Spot verification	Every 10th replication: re-read original	Discrepancy found → correct + review previous 9

Retrospective application v2.1. When QC is applied after data collection (as for the present 180-run corpus), this must be declared explicitly in the QC Log header. Retrospective QC preserves the audit function (errors are found and corrected before analysis is finalized) but not the interruption function (a drifting batch cannot be stopped mid-collection). Both facts are stated, not hidden.

---

5. Workflow Step-by-Step

5.1 Setup (before first replication)

1. Read this manual completely.
2. Open the coding spreadsheet.
3. Complete the walkthrough practice validation.
4. Verify coding accuracy on the practice example.

5.2 Coding a Single Replication

> **PRE-STEP 0 — Illocutionary act check [v2.0] (30 sec)**
>
> Before classifying the claim, verify that the utterance is used as an ASSERTION.
>
> RULE: declarative linguistic form presupposes assertive function, absent explicit markers of a different illocutionary force.
>
> Markers indicating NON-assertive function (suspend validation): — explicit imperative mood: "Run faster" / "Be flexible" — explicit performative: "I now pronounce you husband and wife" — direct interrogative: "Are you really convinced?"
>
> Ambiguous declarative form → treat as assertion. The framework does not recover contexts the speaker did not supply: where decontextualization occurred on the speaker's side, the validator takes the utterance as asserted.
>
> If NON-assertive: code Verdict = NA_ILLOCUTORIO. Record in Notes (mandatory).

Step 1 — Read the entire validation (5 min). Skim-read the full text; get the gestalt of the argument structure. Do NOT code yet.

Step 2 — Extract V2 Verdict (30 sec). Locate the PROSPETTO box → "Esito:" line → code.

Step 3 — Extract V3 Confidence (1 min). Scan for a percentage. Code if present, NA if not.

Step 4 — Extract V4–V6 Arguments (10–15 min). Read SINTESI ESTESA carefully. Identify the top three by prominence. Paraphrase consistently.

Step 5 — Extract V7–V9 Sources (5 min). Scan FONTI. Top three by prominence. Author (Year) format.

Step 6 — Extract V10–V12 Quantitative data (5 min). Select the three most important numbers. Preserve precision exactly.

Step 7 — Extract V13–V15 Frameworks (5 min, normative only). Standardized names. NA for non-normative claims.

Step 8 — Notes (2 min). Optional, except: mandatory if Verdict = NA_ILLOCUTORIO or D4.3 outcome = BEYOND FRAMEWORK SCOPE.

5.3 Batch Workflow

Code 10 replications → run QC1–QC4 → record in QC Log → correct if flagged → proceed to next batch.

5.4 Claim Completion

After all 30 replications: overall consistency review → export data → checkpoint with collaborator → proceed to next claim.

---

6. Coding Modes: Manual and Automated v2.1

The v2.0 protocol assumed a human annotator. v2.1 admits two coding modes and specifies the validation bridge between them.

6.1 Mode M — Manual coding

The workflow of Section 5, executed by a human annotator. Mode M is the reference standard: automated extraction is valid only insofar as it agrees with Mode M.

6.2 Mode A — Automated extraction

Structured fields (V1–V3, V7–V12) may be extracted by script from the validation outputs, exploiting the locked output format (PROSPETTO table, ERC table, FONTI section). The verdict pipeline used for the reported results — table-row extraction of the Esito cell, string normalization, and keyword-based semantic categorization — is specified in Section 7 and deposited as code with the open materials. Interpretive fields (V4–V6 argument paraphrase, V13–V15 frameworks) require either manual coding or LLM-assisted extraction; in the latter case the extraction prompt is fixed, versioned, and deposited with the open materials.

6.3 Mandatory spot-validation of Mode A

Automated extraction must be validated against human coding as follows:

1. Draw a random sample of **10% of replications (18 of 180), with at least 3 per claim**.
2. A human coder codes the sample blind to the automated output, following Section 5.
3. Compute per-variable agreement between Mode A and Mode M.
4. **Acceptance threshold: ≥ 90% agreement on V2 (verdict)** and ≥ 80% on V4–V6. Below threshold: the automated extractor is revised and the full corpus re-extracted; the failed round is recorded in the QC Log.
5. Spot-validation results (sample IDs, agreement rates, discrepancies, resolutions) are recorded in the QC Log.

6.4 Rationale

Mode A removes fatigue-driven error and makes coding trivially re-runnable; Mode M anchors the semantic judgments no parser can certify for itself. Neither substitutes for the other: A scales, M grounds.

---

7. Agreement Metrics v2.1

This section formalizes the agreement metrics that constitute the study's primary outcome measures, as implemented in the deposited analysis script. Fleiss' κ is retained as a chance-corrected complement.

7.1 Unit of analysis and verdict extraction

The unit of analysis is the single replication. From each replication, the verdict is extracted mechanically from the PROSPETTO DI VALIDAZIONE markdown table: the row whose first cell is exactly "Esito" (which excludes the "Pre-Step 0", "Sintesi rapida", and "ESITO ERC" rows). Replications recorded as skipped are excluded; replications where no Esito row can be extracted are counted and reported, and excluded from AR-lex denominators.

Both agreement rates are **modal**: the claim-level rate is the share of replications carrying the most frequent value, and the study-level rate is the unweighted mean of the six claim-level rates.

7.2 Lexical Agreement Rate (AR-lex)

The verbatim Esito string is normalized minimally: markdown emphasis markers removed, whitespace collapsed, case-folded. No synonym mapping, no paraphrase tolerance.

AR-lex(c) = (frequency of the modal normalized Esito string) / (replications with an extracted Esito). AR-lex = mean over the six claims.

AR-lex measures **surface reproducibility**: does the model say the same thing in the same words? Note that under the modal definition the floor for 30 replications is 1/30 ≈ 3.33%, not 0.

7.3 Semantic Agreement Rate (AR-sem)

Each normalized Esito string is assigned to one of five semantic categories by a deterministic keyword classifier:

Category	Trigger keywords (whole-word match)
FALSO	falso, confutato, refutato, infondato, non supportato
PARZIALMENTE VERO	parzialmente, parziale, riduttivo, incompleto, forma debole, fuorviante
VERO	vero, confermato, verificato, supportato
DIPENDE	dipende, contestuale, condizionale
SOSPESO	indeterminato, indecidibile, sospeso, dipende, non decidibile, metafisico, non validabile, strutturale, non verificabile

Disambiguation rules, applied in this order: (1) **mixed-verdict override** — if an Esito contains both FALSO keywords and PARZIALMENTE VERO keywords (or the word "debole"), it is classified PARZIALMENTE VERO, so that verdicts of the form "false in the strong form, true in the weak form" do not collapse into FALSO; (2) otherwise, the first matching category in the priority order **FALSO > PARZIALMENTE VERO > VERO > DIPENDE > SOSPESO** wins. Esiti matching no keyword are classified ALTRO, reported as unclassified, and retained in the AR-sem denominator (they can never inflate agreement, only depress it).

AR-sem(c) = (frequency of the modal category) / (all categorized replications, ALTRO included). AR-sem = mean over the six claims.

Ties between modal categories are reported explicitly as ties, never resolved arbitrarily. Each claim-level AR-sem is accompanied by a **Wilson 95% confidence interval** on the modal proportion, appropriate for small n and proportions near the extremes.

AR-sem measures **verdict reproducibility**: does the model reach the same assessment, however phrased?

Relation to the V2 coding table. The five classifier categories are a coarsening of the V2 verdict space of Section 2.1 (e.g., VERO and PROBABILMENTE_VERO fall in one lexical field; INCERTO and INDECIDIBILE fall under SOSPESO). The classifier is an instance of Mode A (Section 6.2) and is therefore subject to the spot-validation of Section 6.3: a human coder verifies on the 10% sample that the automated category matches the category a Mode M coding of V2 would induce.

7.4 Full-text lexical similarity (Jaccard)

As a complement to AR-lex — which looks only at the verdict line — the mean pairwise **Jaccard similarity** over bag-of-words token sets of the complete validation texts is computed for each claim (all C(n,2) pairs of valid runs). This captures how much the whole argumentative text, not just the verdict, varies at the surface level.

7.5 The lexical–semantic gap

Gap = AR-sem − AR-lex.

A large gap indicates that variability lives at the surface-linguistic layer while the underlying verdict is stable — i.e., that agreement is a property of the semantic layer, not of the lexical one. This is the operational content of the study's thesis that validation must address what a claim *is* before what it *says*.

Reported values (six-claim corpus): AR-lex = 4.44%, AR-sem = 78.33%, gap = 73.89 points. Per-claim breakdown, confidence intervals, Jaccard values, and the QC underlying these figures are given in the companion QC Log.

7.6 Chance-corrected agreement

For each dimension D1–D7 coded on the corpus, Fleiss' κ is computed across the replications per claim. Success criterion: **$\kappa \geq 0.61$** (substantial agreement, Landis & Koch 1977). κ complements AR-sem by correcting for category prevalence; both are reported.

7.7 What these metrics are not

AR-sem is not a measure of *correctness*: 30 replications may agree on a wrong verdict. Agreement rates measure reproducibility, a necessary but not sufficient condition for reliability. The framework remains an elicitation tool, not a reliability certificate.

---

8. Time Management

Manual coding estimates (Mode M), per claim of 30 replications: first 10 replications ~40–50 min each (learning), next 10 ~30–35 min (fluency), final 10 ~25–30 min (efficiency) — roughly 15–17 h per claim, ~95 h for the full six-claim corpus. Batch in 2–3 h sessions (5–7 replications); avoid marathon sessions; break every 10 replications.

Under Mode A with spot-validation (Section 6), human effort reduces to the 18-replication validation sample plus QC review: approximately 12–15 h total.

---

9. Troubleshooting

Can't find the verdict in PROSPETTO. Check the final paragraph of SINTESI ESTESA. If still unclear: flag in Notes, consult collaborator.

More than 3 equally important arguments. Apply the prominence rule (length + placement). If tied, select the first three. Note in comments.

Number stated multiple ways. Use the most precise version (prefer "657,461" over "657k").

Framework mentioned but unclear which. Code NA if genuinely ambiguous. Never guess or infer.

Validation seems incomplete/truncated. Flag in Notes. Do NOT exclude — all 180 replications are included.

Imperative or performative utterance [v2.0]. Apply Pre-Step 0 (Section 5.2). Code NA_ILLOCUTORIO. Record in Notes.

Causal claim of excessive complexity [v2.0]. Apply the D4.3 procedure (Appendix B). Code outcome = BEYOND FRAMEWORK SCOPE. Apply the Epistemic Responsibility Check.

Automated extractor fails a field v2.1. Do not patch values by hand silently. Fix the extractor or code that field in Mode M; either way, record the intervention in the QC Log.

When to pause and consult the collaborator:

- First 10 replications of any claim show < 30% verdict agreement
- Systematic coding confusion not resolved by this manual
- Technical issues (corrupted text, spreadsheet errors)
- Before excluding ANY replication from the dataset
- Mode A spot-validation below threshold v2.1

Appendix A — Quick Reference

Variables checklist

Variable	Required?	Notes
V1 Replication_ID	YES	Pre-populated
V2 Verdict	YES	Never NA (except NA_ILLOCUTORIO)
V3 Confidence	If stated	Integer 0–100 or NA
V4–V6 Arguments	YES	Paraphrase, 5–15 words each
V7–V9 Sources	YES	Author (Year) format
V10–V12 Quantitative	If present	Variable = Value format
V13–V15 Frameworks	Normative only	NA otherwise
Notes	Optional*	*Mandatory for NA_ILLOCUTORIO and BEYOND FRAMEWORK SC

Common mistakes to avoid

- Copy-pasting arguments verbatim (paraphrase instead)
- Coding 5+ arguments when asked for the top 3
- Normalizing numbers (657,461 → 657,000)
- Guessing a framework when unclear (use NA)
- Skipping quality checks after each batch of 10
- Classifying an illocutionary act as an assertion without Pre-Step 0 [v2.0]
- Producing a "corrected causal link" for a D4.3 claim instead of declaring the appropriate outcome [v2.0]
- Running QC without recording it in the QC Log v2.1

- Treating automated extraction as self-validating v2.1

Appendix B — Epistemic Note D4.3: Limits of Causal Validation

Context and motivation

This appendix formalizes an epistemic distinction that emerged during the development of the framework: the structural limits of validating D4.3 (Causal-Mechanism) claims. The distinction was elaborated from the Munich case (immigration → healthcare collapse), a key moment in the project's development.

The Munich case presented the geometry: A false + B true + causal link A→B undemonstrated. The manipulation consisted in presenting A as a necessary or sufficient condition of B when it was neither. The validation separated the three levels asymmetrically and appropriately: it did not replace the link with a "correct" one — it falsified the one proposed.

Fundamental principle

> **Causal validation is structurally asymmetric.**
>
> The framework can FALSIFY an erroneous causal link. It cannot REPLACE it with a correct one without falling into the same epistemic problem.
>
> This limit is structural, not contingent: it follows from the fundamental asymmetry between falsification and verification (Hume, Popper) and from the impossibility of observing causation directly (Pearl, Woodward).
>
> Operational consequence: declaring outcome 3 (beyond framework scope) is not a weakness of the Coding Manual. It is the epistemically honest response — and what distinguishes this instrument from one that pretends to more than it can demonstrate.

D4.3 validation procedure — three outcomes

For every claim classified D4.3, the annotator must conclude with exactly one of the following. No intermediate or ambiguous conclusions are admitted.

Outcome 1: link FALSIFIED

Conditions: A is false, or the proposed causal mechanism is absent, inverted, or non-operative under the described conditions. Action: document which element is falsified (A, the mechanism, or both). If A is false and B is true, state that the truth of B is independent of A and does not confirm it. Canonical example: "migrant emergency → healthcare collapse" (Munich case). A = false; B = true (for independent reasons); link = undemonstrated. Outcome: FALSIFIED.

Outcome 2: link NOT CORROBORATED

Conditions: the causal assertion is present but unsupported by evidence of mechanism. Neither A nor B need be false, but the causal connection is undocumented. Action: report the absence of mechanism evidence. Do not conclude the link is false — only that it is not corroborated. Distinguish from Outcome 1. Example: "Social media use causes depression in adolescents." A plausible, B plausible, link debated: correlations exist, mechanism contested.

Outcome 3: link BEYOND FRAMEWORK SCOPE

Conditions: causal complexity exceeds what the framework can resolve; the link is not falsifiable with the annotator's available instruments, or requires specialist domain competence outside the protocol. Action: flag the claim as problematic; transfer epistemic responsibility to the speaker; apply the Epistemic Responsibility Check (below). Note: Outcome 3 is not a surrender — it is the instrument's honest boundary.

Special case: composite claims A→B

In claims with conditional or optative structure ("If A, then B"; "Since A, we must do B"), D1 classification follows B, not A. B defines the ontological type of the assertion; A specifies its applicability conditions.

Structure	D1 type	If A false → effect
If A [empirical], then B [policy]	D1.10 (B determines type)	Non sequitur: B not applicable. Does not falsify B absolutely.
Since A [empirical], B [normative]	D1.8 or D1.10 (B)	Non sequitur + epistemic irresponsibility if A undeclared.
A [declarative causal] → B [declarative]	D1.2 + D4.3	Apply the three-outcome D4.3 procedure.

Epistemic Responsibility Check

Mandatory for Outcome 3. Is the speaker entitled to assert this causal link?

Marker	Yes	Partial	No
Cites sources/evidence?	"According to a 2020 study in Nature, X causes Y."	"Studies show that X causes Y." (vague)	"X causes Y." (no justification)
Expresses appropriate uncertainty?	"The evidence suggests X; further research is needed."	"X is probably true."	"X is certainly true." (where uncertainty exists)
Distinguishes interpretation from fact?	"The data show correlation, which I interpret as…"	"The data show that X causes Y." (unflagged)	"X causes Y." (correlation as causation)

Theoretical references

- Hume, D. — Fundamental asymmetry between observation and causation.
- Pearl, J. (2009). *Causality* — P(Y|X) vs P(Y|do(X)): correlation vs intervention.
- Woodward, J. (2003). *Making Things Happen* — Counterfactual criterion of causation.
- Popper, K. (1959). *The Logic of Scientific Discovery* — Falsification/verification asymmetry.

Provenance: derived from the Addendum to Coding Manual v1.0, integrated in v2.0 (February 2026); carried unchanged into v2.1. **Founding case:** Munich case (A false, B true, link undemonstrated).

Appendix C — Deviation Log from Pre-Registration v2.1

The study was pre-registered with a 3-claim × 30-replication design (90 validations) and % agreement + Fleiss' κ as planned metrics. The following deviations occurred and are declared here in full, consistent with the anti-HARKing function of pre-registration: a deviation declared is a datum; a deviation hidden is a defect.

D-1. Design expansion: 3 → 6 claims (90 → 180 validations). *Nature:* three claims were added (WWI/Franz Ferdinand, consciousness/matter, smoking/cancer) to extend D1 coverage to causal-historical, metaphysical, and causal-empirical types, and to include two D4.3 claims

permitting within-study test of the three-outcome causal procedure. *Timing:* decided after pre-registration, before analysis of the added claims. *Risk assessment:* the expansion adds observations without dropping any pre-registered cell; no pre-registered claim was removed or re-run. Selective-reporting risk: none identified — all 180 runs are included.

D-2. Metric addition: AR-lex / AR-sem. *Nature:* the pre-registered metrics (% agreement, Fleiss' κ) did not distinguish lexical from semantic agreement. Inspection of the raw outputs showed near-zero verbatim agreement coexisting with high verdict agreement; AR-lex and AR-sem (Section 7) were defined to capture this separation. *Timing:* defined after first inspection of raw outputs — this is a data-driven metric addition and is flagged as such. The pre-registered metrics are still computed and reported; AR-lex/AR-sem are reported alongside, not instead.

D-3. Coding mode: introduction of automated extraction (Mode A). *Nature:* the pre-registered protocol assumed fully manual coding. Structured-field extraction was automated; the human-anchored spot-validation of Section 6.3 was added as the validity bridge. *Timing:* operational decision during data collection.

D-4. Retrospective QC. *Nature:* QC1–QC4 were applied to the completed corpus rather than batch-by-batch during collection. See the companion QC Log for the full retrospective register.

Coding Manual v2.1 — Episteme Advisory Lab. Published as part of the open-materials set of the study "Validating Claude with Claude: Reproducibility and Consistency in LLM Claim Validation."