Beyond Answer Correctness

Evaluating warranted commitment through post-training and prompting

Felipe A. Ly
Entel AI Research

2026-10-16

Abstract A correct answer can lack adequate grounds, and a correct assessment can precede an unwarranted final answer. We connect this philosophical distinction to a finite evaluation of language-model commitments and examine two empirical investigations. The criterion identifies when available evidence warrants an answer, when it requires suspension, and when support survives evidence withdrawal. An exploratory training pilot showed higher joint correctness and fewer unsupported commitments under certificate supervision, alongside retention and suspension costs. A later Certificate-versus-Copy comparison used three paired seeds: mean joint correctness increased by 1.03 percentage points over 260 panels, while unsupported commitments increased by 0.55 points over 243 unresolved states and harm safeguards failed. In the workflow comparison, ordinary analysis improved selected outcomes relative to direct answering, with false-suspension costs; the structured package failed its declared screens. A retained-assessment diagnostic found seven incorrect final answers following twelve fully correct assessments. Further exploratory diagnostics examined presentation, wording and supported cases. The contribution is a worked audit that separates evidence sufficiency, record correctness and final-answer warrant, making gains, costs and failed safeguards jointly inspectable. The studies retain their separate controls and decision rules. Their findings concern specified behavior, without identifying a common mechanism or establishing a general theory of understanding.

Working paper v1.0.

1 Introduction: What a correct answer leaves open

An analyst reports three indicators correctly and declares an assessment complete. A colleague asks whether a condition required by the assessment rule has been checked. It has not. The reports remain correct, but they do not settle the assessment. Later, a reliable audit establishes the missing condition and warrants the answer initially announced. The answer's value stays the same; its grounds change.

This hypothetical encounter captures the question from which the work developed: **how can another perspective advance inquiry, and what would justify calling that inquiry complete?** The colleague contributes no new measurement. Instead, the challenge exposes a requirement that already belonged to the task. Recognizing that gap corrects the claim that the assessment is settled while leaving its answer open. The later audit supplies the missing ground. An evaluation concerned only with whether the final answer changed would miss both contributions.

We distinguish **a true answer**, **adequate grounds**, **competent use of those grounds**, and **justified completion of a bounded task**. These are separate objects of assessment. A correction can improve the inquiry before resolution; supplying a justified answer does not establish another person's ability to explain or apply its grounds. Completion additionally depends on which questions the task requires and whether that scope is defensible. This makes the original question constructive: identify what an encounter repairs, what it preserves and what remains unresolved.

1.1 Perspectives, grounds and understanding

The distinctions have substantive philosophical precedents. Massimi describes scientific perspectives through the practices, resources and principles by which communities produce claims and assess their reliability. She distinguishes truth from the perspectival justification of a claim (Massimi 2022, 5–9). Our encounter concerns a narrower act of scrutiny: shifting attention from the reported indicators to the conditions for using them. It does not require the participants to represent different scientific perspectives or test perspectival realism. It shares the concern that assessing grounds can matter even when the claim's truth does not change.

Grimm distinguishes accepting a dependence claim from grasping how the dependence works, using competent responses to relevant variations to illuminate that difference (Grimm 2006, 530–34). This motivates our working phrase *dependency-sensitive understanding*: a capacity to identify relevant premises, consider alternatives and preserve qualifications in use. The phrase names the bounded capacity of interest, rather than an established new theory. Grimm argues that understanding is a species of knowledge; our evaluation does not decide that classification. Receiving or producing one correct account of the grounds is insufficient to establish the capacity.

Elgin supplies a further qualification. Her account connects understanding with competent use while allowing idealizations to make selected features accessible (Elgin 2007). An omission may therefore be productive or may defeat the intended inquiry. In the encounter, the unchecked condition matters because the assessment rule already requires it. The issue is the relevance of what has been left out, rather than a demand to represent everything. Our exact finite test below concerns selected answers under explicit assumptions; it does not make literal truth of every element a general requirement for scientific understanding.

1.2 From the inquiry question to model evaluation

The connection to machine learning concerns the treatment of conclusions when their support changes. Here, a *commitment* is a definite answer; *warrant* is its justification by the permitted evidence under the task's assumptions. *Suspension* means reporting that this evidence leaves the question unresolved. A model should suspend when the sole adequate ground is lost, retain an answer when independent support survives, and answer when the evidence settles the query. We use *alignment* only for conformity to this specified evidential norm.

This is a chosen standard for categorical reports that the evidence settles a question. If an admissible alternative gives the opposite answer, that report is too strong. The requirement to give a supported answer when available is the complementary part of the task contract, preventing indiscriminate suspension. These requirements do not say that every rational belief or action under uncertainty needs deductive certainty. Nor do they measure understanding as a whole: varying the available observations is different from testing a participant's grasp of causal or explanatory dependencies.

We ask whether the training and prompting interventions examined here improved conformity to this standard. Post-training varied auxiliary supervision under common final-answer targets. The workflow study varied the assessment requested before answering. A diagnostic in the latter study found correct assessments followed by unwarranted final answers, making the difference between record correctness and final-answer warrant visible (see the assessment diagnostic). The studies have distinct controls, units and decision rules; they are not matched replications.

The contribution is a worked audit of warranted commitment: an explicit task contract, exact support labels and separate checks of assessment records and final answers. For ML evaluation, this distinguishes a favorable score from meeting an intervention's benefit and harm requirements. Philosophically, it makes a limited behavioral consequence of adequate grounds testable without equating it with understanding. Appendix A illustrates why successive resolutions need not establish comprehensive completion. The finite criterion and empirical conclusions do not depend on it.

1.3 Relation to machine-learning research

These targets have substantial precedents. Truth maintenance records reasons for beliefs and supports their revision when assumptions change (Doyle 1979). RuleTaker tests premise removal and alternative proofs; ProofWriter combines open-world unknown labels with proof generation (Clark et al. 2020; Tafjord et al. 2021).

Sufficient Context separates context sufficiency from answer correctness and studies abstention, including reported selective-generation benefits (Joren et al. 2025). Selective question answering evaluates risk and coverage; predicting model error differs from detecting unresolved evidence (Kamath et al. 2020).

Process supervision also predates this work. The inspected *Let's Verify Step by Step* compares reward models selecting from a fixed generator; our study adapts the generator (Lightman et al. 2023). Evidence-sufficiency training and paired evidence diagnostics already examine answering transitions, redundant support and evidence changes (Sato et al. 2026; Deswal 2026). These precedents rule out presenting those targets as new here.

Two further distinctions locate the contribution. Explanation faithfulness concerns the relation to the prediction process (Jacovi and Goldberg 2020); our record audit checks semantic content and subsequent behavior without identifying that relation. Vardi and Sedoc's reliance certificates characterize recommendation stability and contextual scope without objective ground truth (Vardi and Sedoc 2026). Here, warrant means support under stipulated evidence, scored by an exact finite oracle. These are different operationalizations, not competing intervention results.

2 A finite test of evidential support

We now isolate the evidential part of the argument. The guiding test is simple: consider every situation still compatible with the permitted evidence. If all give the same answer, that answer is supported. If they give different answers, the evidence leaves the question unresolved. This is a test of what follows from the evidence, not of how confident the model sounds.

To examine both missing and surviving support, consider a variant of the opening assessment: both checks x and z must pass, unless an independently sufficient condition c holds. Write the rule as `(x AND z) OR c`. The following cases show why missing information need not require suspension.

Table 1: Evidence sufficiency is relative to the query.

Available observations	Warranted status	Reason
x=1, z=1, c=0	supported 1	The first branch suffices.
x=1, c=0	unresolved	Different values of z give different answers.
x=1, c=1	supported 1	The c branch survives without z.
x=0, c=0	supported 0	Neither branch can hold, whatever z is.

Formally, let Ω be a finite set of possible states, q a binary query and $K(E)$ the states compatible with permitted observations E. The answers still possible are

$$A(E) = \{q(w) : w \in K(E)\}.$$

For nonempty $K(E)$, the warranted status is `supported_0` when $A(E) = \{0\}$, `supported_1` when $A(E) = \{1\}$, and `unresolved` when both values remain possible. An empty compatible set flags incompatible inputs, not vacuous support. Interpreting this exact score as warrant assumes reliable observations, fixed meanings and rule, and a state space representing the actual situation. The oracle, an exact reference procedure, computes the consequences of those assumptions; it does not validate their suitability for a real decision.

In the experiments, **answer correctness** means agreement with this evidence-status label, rather than guessing the Boolean value in an unobserved actual state. A definite answer can match that value and still be incorrect for an unresolved task. **Record correctness** concerns the content of an intermediate assessment and is evaluated separately. These distinctions connect the task to the motivating encounter without treating the scores as a general measure of epistemic warrant.

Pure removal of constraints enlarges the compatible set. A query constant on the smaller nonempty set can remain constant or become unresolved; it cannot become constant at the opposite value. An opposite

warranted answer requires changing or replacing constraints, or changing the rule. Derived records follow the same dependency discipline: a certificate based on withdrawn evidence is not an independent surviving ground.

The studies operationalize these obligations through distinct endpoints. **Joint correctness** requires both responses in a designated evidence-change pair to be correct. **Retention** requires both to remain correct when the oracle preserves the same supported answer. **Unsupported commitment** counts definite answers on unresolved cases. **Unnecessary suspension** counts unresolved answers on supported cases. **Total unresolved error** also includes invalid responses, so rejection is not counted as successful suspension by the combined criteria. Original eligible denominators and weights remain study-specific.

For workflows, three further questions matter: was the record admitted by the format gate, was its content semantically correct, and was the final answer warranted? Structural acceptance establishes none of the later properties. Conversely, following an incorrect record faithfully can preserve its error. The target is warranted final behavior, with record quality evaluated separately.

3 Post-training: auxiliary tasks and their tradeoffs

3.1 Comparison and controls

The training studies used Llama 3.2 3B Instruct with fresh low-rank adaptation (LoRA) adapters, trainable additions to a frozen base model. Final-only was trained to predict warranted final statuses. Copy added reproduction of visible evidence and missingness markers. Certificate added sufficient active-observation sets for supported cases and two compatible worlds with opposite answers for unresolved cases. Certificates supplied training targets, not additional permitted evidence at final evaluation.

The auxiliary arms matched primary exposure and order, sequence shapes, response-token budgets, loss coefficients and update counts. They differed in annotation information and learning difficulty: Certificate annotations included complete witness states beyond the selected primary states, all within the full training partition (Appendix B). Consequently, the comparison estimates the specified auxiliary-task contrast rather than an isolated effect of support semantics.

Each adapter saw 312 primary training states, forming 156 panels, twice. Development used 520 states, 260 panels, 71 retention pairs, 243 unresolved states and 277 supported states. Evaluation requested only final statuses in fresh contexts, with no gold certificate, auxiliary output or arm identity. A static grammar permitted all three labels without choosing the correct one. The recipe and declarations are summarized in Appendix B.

3.2 Developmental signal and costs

PLT002 was a recipe-adequacy pilot with three paired seed triplets, comparing Final-only, Copy and Certificate. Its equal-rule-class development means showed higher joint correctness and fewer unsupported commitments for Certificate. Relative to Final-only, those changes accompanied lower retention and greater unnecessary suspension.

Table 2: PLT002 means under the original equal-class endpoint weights. These are weighted fractions, not pooled counts over 260 or 71 panels. Suspension means unnecessary suspension on supported states.

Arm	Joint %	Retention %	Unsupported %	Suspension %
Final-only	36.21	71.58	71.45	5.70
Copy	36.27	64.10	57.30	18.66
Certificate	48.34	65.38	35.08	21.42

All nine technical cycles completed, but none passed the full declared fitting conjunction. The favorable development endpoints and adequacy failures both remain findings. Their coexistence makes the result informative about the combined recipe, without establishing a successful general intervention or the cause of incomplete fitting.

3.3 The fixed Certificate-versus-Copy comparison

CMP001 used three fresh paired seeds with a fixed recipe. It compared Certificate with Copy only, with no unadapted comparator. Means used raw eligible-unit fractions and equal seed-pair weights, rather than PLT002′s equal-class summaries. All six adapters completed; all 3,120 constrained development outputs were valid. Historical fitting passes were 0/6, with no seed excluded or recovered.

Table 3: Each cell gives Certificate / Copy counts. Higher joint and retention are better; lower unsupported commitment is better.

Seed	Joint /260	Retention /71	Unsupported /243
104729	122 / 132	46 / 51	93 / 80
130363	132 / 139	44 / 44	52 / 43
155921	155 / 130	55 / 50	71 / 89

Table 4: Paired count differences. Denominators are 260, 243, 71 and 243, respectively. Positive values in the last two columns indicate harm.

Seed	Joint gain	Commitment reduction	Retention loss	Unresolved-error rise
104729	−10	−13	+5	+13
130363	−7	−9	0	+9
155921	+25	+18	−5	−18

Mean joint correctness increased by 1.03 percentage points, while unsupported commitments increased by 0.55 points. Both mean benefit requirements, gains of at least 13 panels and reductions of at least 13 unresolved-state commitments, failed. Seed 104729 lost five retention successes, exceeding the allowed one; unresolved errors increased in the first two pairs. Equal mean retention would conceal that adverse pair.

The declared classification was mixed and not promising. The predeclared all-positive sensitivity also failed. The result is a complete finite comparison with heterogeneous seed outcomes. It does not estimate improvement over the starting checkpoint, population significance, future-seed reliability or what additional training would achieve.

4 Prompting: assessments and final answers

4.1 Workflow comparison

The workflow study used the retained Llama 3.2 3B Instruct checkpoint without study-specific adaptation. Direct answering (D) requested a final status. Ordinary analysis (O) requested free-text analysis followed by a final answer. Structured assessment (S) requested observations, missing variables, a status and either sufficient support or two opposite-answer witness worlds. S thus externalizes at inference time the support and witness structure supervised by Certificate in Section 3.1; this is a structural parallel, without establishing a common mechanism.

The controller checked record structure, not semantic correctness. Malformed or truncated S records became assigned invalid outcomes without repair or retry; O required nonempty, untruncated analysis. Both later requests received the original task and prior text as an escaped JSON string in a fresh context. The primary S-versus-O contrast therefore includes instructions and formatting burden. Equal caps did not equalize realized computation.

Each full census crossed four rules with all 81 partial assignments over four binary variables: 324 states, comprising 130 supported-negative, 68 supported-positive and 126 unresolved cases. Its 864 overlapping removal pairs reused endpoint outputs. These are evidence-state comparisons, not conversational revision or independent replications. The census had been historically inspected; the revised run reused it after observing earlier outcomes.

4.2 Gains, costs and failed safeguards

Ordinary analysis improved aggregate correctness and reduced unsupported commitments relative to direct answering in both censuses. It also introduced false suspension and reduced correct supported-negative answers. Structured assessment failed its prospective screens.

Table 5: Assigned census outcomes. Invalid responses remain in their original denominators.

Run	Arm	Correct /324	Unsupported /126	Invalid /324
EBC001R1	D	163	121	0
EBC001R1	O	244	21	6
EBC001R1	S	74	56	186
EBC002	D	171	122	0
EBC002	O	237	42	1
EBC002	S	0	0	324

Table 6: Warranted answers and false suspension on supported cases. Zero false suspension does not credit invalid outcomes.

Run	Arm	Negative /130	Positive /68	False unresolved /198
EBC001R1	D	97	61	5
EBC001R1	O	87	56	51
EBC001R1	S	42	32	0
EBC002	D	108	59	0
EBC002	O	90	64	42
EBC002	S	0	0	0

EBC001R1 was non-promising and failed all seven gates. EBC002 clarified the variable universe and schema instructions together. Every S record contained outer Markdown fences and was rejected by the strict gate, so no S final answer was generated. The frozen classifier called this mixed because commitments fell arithmetically; all six safeguards failed. Its zero commitments describe universal invalidation.

4.3 What a correct assessment establishes

EBC003 reused retained assessments on 48 hash-selected cases, without selecting for semantic quality. The full assigned comparison gave 25/48 correct task-only answers and 24/48 record-conditioned answers. Two record rejections remained invalid assignments. Among 46 admitted pairs, correctness was still 25 versus 24. Four of five gates failed. Seven of twelve fully correct assessments preceded incorrect finals, all seven on unresolved cases; this was a post-hoc finding.

One case makes the discrepancy concrete. For `(x AND z) OR (c AND y)` with no observations, the model-produced assessment supplied all-zero and all-one assignments and correctly stated that the case was unresolved. When that assessment was included in a later request, the final answer was `supported_0`, a definite negative answer. The available record was correct; the final commitment was not. This is an observable discrepancy between outputs, without identifying how the model internally used the record.

EBC004 changed the presentation of 24 retained records from JSON to field lines. Correct totals were 11/24 and 10/24; three final labels changed. Neither presentation produced an unresolved final among the 16 records stating unresolved, or a correct answer on the 11 gold-unresolved cases. Reuse of historical JSON controls was conditional on six exact live token replays. Three of six gates failed. Presentation affected some outputs without yielding the declared normative improvement.

Fresh-rule pilots B and C explored wording and record design. Each arm had 240 assignments: 24 unresolved cases under ten wordings. B used four rules and O/S; C used eight further rules and O/S1/S2. S1 explicitly requested the record's status in the final wrapper; S2 simplified the record. Most records passing the format gate were fully correct in B/S (117/133) and C/S1 (103/129). Yet every admitted record in those two arms was

followed by an unwarranted committed final answer. Counting commitments and rejections, unresolved errors were 155/240 for B/O versus 240/240 for B/S; for C they were 177/240 for O and 240/240 for each structured variant. C selected neither variant; B had no advancement gate. These unresolved-only pilots cannot assess preservation of warranted answers.

Pilot D supplied that missing supported-case check, using S alone on Llama, Gemma 2 2B it and Phi-3.5 mini instruct. Each checkpoint received 60 assignments, including four distinct supported cases repeated over ten wordings. Here the failure was already in the record: 74 of 80 admitted records on supported cases falsely stated unresolved. Admitted-and-warranted supported counts were 0/40 for ungated Llama, 0/40 for Gemma and 6/40 for Phi; neither gated checkpoint passed. None of Phi's six successes was a fully correct record. The gate concerned the record's status. With no ordinary arm, D estimates no supported-case workflow effect.

5 Discussion: What the combined evaluation adds

5.1 A useful target for post-training and alignment

The audit asks three practical questions: does a gain preserve warranted answers, does reduced commitment reflect successful suspension rather than rejection, and does a correct assessment lead to a warranted final answer? The finite standard makes answering, suspension and retention assessable together. PLT002 and the ordinary-versus-direct workflow contrast show favorable endpoint changes with costs. CMP001 shows why a positive mean joint gain does not settle an intervention judgment when another benefit endpoint and per-pair safeguards fail.

For post-training, the results distinguish adding auxiliary supervision from improving the final behavior that motivated it. The Certificate-versus-Copy control matches several exposures while leaving information and difficulty unequal. The comparison therefore informs evaluation of that training package. It does not identify support semantics as the cause, and incomplete fitting does not identify why the declared benefit requirements were not met.

For prompting, admission, record semantics and the final answer locate different failures. In EBC002, rejection of every structured record does not establish successful suspension. Pilots B and C expose correct records followed by unwarranted commitments; Pilot D exposes false suspension within admitted records on supported cases. The distinction directs attention to the stage requiring investigation without identifying a remedy. These observations do not show that the model internally ignored the record, or that larger models or another prompt would exhibit the same pattern.

5.2 Returning to understanding and completion

The encounter helps distinguish the repairs these evaluations can reveal. A wrong answer calls for correction; an unsupported commitment calls for examination of its grounds; unnecessary suspension calls for recognizing support already available. The original answer can remain true while the appropriate response changes after evidence withdrawal. These differences concern the relation between claims and their support, not merely the number of answers produced.

The stronger capacity question remains separate. A record may state the right dependencies without establishing that its producer or recipient can use them competently. The workflow discrepancies show that even a correct record did not ensure a warranted final answer under the tested procedures. They provide a reason to evaluate those outcomes separately. They do not explain the model's internal relation to the record or prove that it lacks understanding.

Bounded completion is likewise an achievement with conditions. If the declared questions are adequately settled and the task boundary is defensible, leaving other features unrepresented need not defeat completion. Elgin's attention to selective idealization helps articulate this distinction (Elgin 2007). It supplies no automatic vindication of our chosen rule or observations. Their reliability and relevance need separate grounds. Resolving a specified query under these assumptions does not decide whether further investigation is warranted: an inquiry may be discontinued while a question remains open, or reopened when its grounds are challenged. The finite oracle scores resolution under a contract, not a universal stopping policy.

5.3 Limits of the synthesis

These studies share a question, not a common estimand. Training and workflow interventions were not compared against each other. They used different weights, denominators, controls and declarations, with reused development material and exploratory diagnostics. We retain exact finite counts without pooling effects, treating overlapping cases as independent, or claiming population significance and achieved power. The wider model, task and stochastic generalizations remain unmeasured.

PLT002′s methods and adequacy criteria are documented in its report and tables; its original operational declaration is not included in the study records examined here.

The literature assessment used targeted primary-source comparisons across inquiry, evidence sufficiency, support revision and explanation faithfulness. It is selective and version-bound, not an exhaustive priority search. The contribution is the worked evaluation: explicit assumptions, separate checks of records and answers, and intervention judgments that retain gains, costs and failed safeguards. The philosophical comparisons locate this use without resolving disputes about knowledge and understanding or establishing a new theory.

6 Conclusion

Evaluating warranted commitment asks whether a model answers, suspends and retains conclusions as the available evidence requires. The paper makes that judgment inspectable by separating evidence sufficiency, assessment quality and final behavior. The retained studies show selected gains, meaningful costs, mixed seed outcomes and correct assessments followed by unwarranted finals. Neither closed intervention comparison passed its declared success criteria. The practical contribution is a worked way to identify what an improvement achieves and where a procedure fails, while leaving understanding, broader alignment and the effectiveness of other interventions open.

Acknowledgements

During the development of this work, the author used OpenAI's Codex and Anthropic's Claude to support critical discussion and manuscript development, and Google Gemini Research to assist with literature discovery and evidence organization. Codex also supported writing, code reviewing and technical checks. Scholarly claims are supported by the cited literature and retained study records. The author directed the research, thoroughly vetted the AI-assisted outputs for scientific accuracy, and assumes full responsibility for the selection and interpretation of evidence and the integrity of the final manuscript.

References

Clark, Peter, Oyvind Tafjord, and Kyle Richardson. 2020. "Transformers as Soft Reasoners over Language." *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 3882–90. https://doi.org/10.24963/ijcai.2020/537.

Deswal, Suryadeep Singh. 2026. *EviScope: Paired Counterfactual Evidence Diagnostics for Faithful and Efficient Grounded Language Models*. https://arxiv.org/abs/2609.17081v1.

Doyle, Jon. 1979. "A Truth Maintenance System." *Artificial Intelligence* 12 (3): 231–72. https://doi.org/10.1016/0004-3702(79)90008-0.

Elgin, Catherine Z. 2007. "Understanding and the Facts." *Philosophical Studies* 132: 33–42. https://doi.org/10.1007/s11098-006-9054-z.

Grimm, Stephen R. 2006. "Is Understanding a Species of Knowledge?" *The British Journal for the Philosophy of Science* 57 (3): 515–35. https://doi.org/10.1093/bjps/axl015.

Jacovi, Alon, and Yoav Goldberg. 2020. "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–205. https://doi.org/10.18653/v1/2020.acl-main.386.

Joren, Hailey, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. "Sufficient Context: A New Lens on Retrieval Augmented Generation Systems." *The Thirteenth International Conference on Learning Representations*. https://arxiv.org/abs/2411.06037v3.

Kamath, Amita, Robin Jia, and Percy Liang. 2020. "Selective Question Answering Under Domain Shift." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5684–96. https://doi.org/10.18653/v1/2020.acl-main.503.

Lightman, Hunter, Vineet Kosaraju, Yura Burda, et al. 2023. *Let's Verify Step by Step*. https://arxiv.org/abs/2305.20050v1.

Massimi, Michela. 2022. *Perspectival Realism*. Oxford Studies in Philosophy of Science. Oxford University Press. https://doi.org/10.1093/oso/9780197555620.001.0001.

O'Connor, Russell. 2005. "Essential Incompleteness of Arithmetic Verified by Coq." In *Theorem Proving in Higher Order Logics*, edited by Joe Hurd and Tom Melham, vol. 3603. Lecture Notes in Computer Science. Springer. https://doi.org/10.1007/11541868_16.

Sato, Haruto, Yuki Tanaka, Ren Nakamura, Aoi Kobayashi, and Mei Ito. 2026. *Learning Evidence Sufficiency Boundaries for Selective Answering in Grounded Multi-Hop QA*. https://arxiv.org/abs/2609.01687v1.

Tafjord, Oyvind, Bhavana Dalvi Mishra, and Peter Clark. 2021. "ProofWriter: Generating Implications, Proofs, and Abductive Statements over Natural Language." *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3621–34. https://doi.org/10.18653/v1/2021.findings-acl.317.

Vardi, Shai, and João Sedoc. 2026. *Epistemic Warrant for LLM Recommendations: Characterizing the Basis for Reliance When Ground Truth Is Unavailable*. https://arxiv.org/abs/2609.04127v1.

Appendix A. Successive resolution and the limits of completion

The opening question invites an extrapolation: if each newly exposed gap can be addressed, will inquiry eventually be complete? Three achievements must be distinguished: progress at finite stages, an effective procedure for continuing, and completeness of an infinite collection. Establishing one does not establish the others.

Arithmetic supplies a restricted illustration. Fix classical first-order logic with finite proofs in $L_{NN} = \{0, S, +, \times, <\}$, with equality, and O'Connor's finite arithmetic base NN, whose axioms hold in the standard natural numbers. Here, completeness means proving every sentence or its negation. Enumerate all sentences of this language. Starting from NN, add each sentence if true in that interpretation, otherwise add its negation. Every finite stage has finitely many true axioms, so it is consistent and its axiom membership is decidable. O'Connor's incompleteness result applies to these recursive extensions of NN: each stage remains incomplete (O'Connor 2005, secs. 2.8, 4). Yet their union contains a true choice for every sentence and is therefore complete.

This construction separates eventual resolution of each question from completion at a common finite stage. It assumes access to arithmetic truth; it supplies neither an effective selection procedure nor epistemic warrant for the additions. Some additions may already be derivable, so it does not guarantee new progress at every step. The example limits an extrapolation from successive resolution; it establishes no general restriction on model learning or on completing a bounded empirical task. Evidence withdrawal is also different: the construction retains earlier axioms, whereas correction must reassess surviving grounds. The finite criterion and empirical conclusions in this paper are independent of the illustration.

Appendix B. Declarations, units and retained detail

Training. Certificate targets exposed 257 complete object identities, including 220 outside the selected primary states but within the full training partition. The fixed base revision was `0cb88a4f764b7a12671c53f0838cd831a0843b95`. Ordinary LoRA used rank 16, alpha 32 and zero dropout on q/k/v/o/gate/up/down projections, with an unquantized FP16 base and FP32 adapters. AdamW used a constant 10^{-4} rate, no warmup or weight decay, and norm clipping at 1. Primary and auxiliary response-token mean losses each had coefficient 1; their backward passes accumulated before one update. Only final adapters were assessed. PLT002 used seeds 314159, 271828 and 161803; CMP001 used 104729, 130363 and 155921.

PLT002′s fitting conjunction required at least 297/312 correct training states, 95% balanced accuracy, 149/156 joint panels, 50/52 retention pairs and at least half the retention pairs per rule. Label counts had to reach 81/90, 78/86 and 123/136, with at least 90% balanced recall for each label. Rule/label cells required ceiling(3n/4) successes, with perfect fitting in five scarce cells. Unsupported commitments and unnecessary suspensions were capped at 13/136 and 17/176; free validity and agreement each required 312/312. All seeds had to fit, with within-arm raw/balanced state, joint and retention ranges at most 2 percentage points. These were correlated engineering criteria. Development endpoint weights were $1/(26n_{r,e})$, where $n_{r,e}$ counts eligible units for rule r and endpoint e.

CMP001 required six technical completions, mean joint gain at least 13/260, mean commitment reduction at least 13/243, retention loss at most 1/71 in every pair and no unresolved-error increase in any pair. These prospective descriptive thresholds are not significance tests. Training-correct counts for Certificate/Copy were 245/245, 255/252 and 277/236 of 312; development-correct counts were 356/365, 359/367 and 393/350 of 520, in seed order.

Workflows. Greedy decoding used declared seed 104729, with 512-token caps for analysis or assessment and 32 for final answers. The four census rules were `x AND z AND c`, `(x AND z) OR c`, `x AND (z OR c)` and `(x AND z) OR (c AND y)`. Three ignore y. Their 864 removal pairs comprise 243 supported-to-unresolved, 366 supported-retention and 255 unresolved-retention pairs. Full-run S-versus-O gates required at least seven fewer unsupported commitments, at least 98/130 negative and 51/68 positive correct answers, losses of at most two negative and one positive successes, no unresolved-error rise and at most three invalids.

EBC003 comprised 17 negative, 12 positive and 19 unresolved cases. Its gates required four net additional correct answers, no loss in either supported stratum and no increase in unresolved errors or invalids. EBC004′s main set contained 8 negative, 5 positive and 11 unresolved cases. It required four additional unresolved answers among the 16 records stating that status, four additional correct answers on the 11 unresolved cases, zero invalid field-line outputs, no supported-stratum loss and no unresolved-error rise. Eight of those 16 recorded unresolved statuses were wrong, so copying alone could not meet the full criterion.

Pilot C required admission of at least 3/4 separately for every wording and for every rule, and at least 1/10 reduction in unresolved error relative to O over the matched grid. Neither variant qualified. Pilot D required at least 30/40 admitted-and-warranted supported records and at least 2/4 within eight of ten wordings for Gemma or Phi. Neither passed. D had only four distinct supported cases per checkpoint; label stratum also coincided with missing-variable count. Pilot A was an exploratory fit/template/format check, and B had no advancement gate.

The precursor EBC001 stopped before model loading and produced no outcomes; it is a technical stop, not another census.