

Response to the Reviewers

Manuscript NAI-980: “A Trust-Centric Neuro-Symbolic Architecture for Verifiable Large Language Model Claims”

Dear Editors and Reviewers,

We thank the three reviewers for careful and constructive reports. They identified the same central weaknesses: the evaluated system did not include the differentiable component that the paper presented as central; the experiments reported a verdict distribution rather than accuracy; no baseline isolated trust weighting; the “end-to-end” run bypassed Stage A; and the FOL→KG step was not specified. We agree with these criticisms. Below we respond to every comment in turn, quoting each one verbatim. After each response we give the location of the corresponding change in the revised manuscript; a version with changes marked is also provided.

Summary of the main changes.

1. **Rescoping (R2.1, R3.1, R3.2).** The trust-weighted LTN objective is now presented as the specified target of the architecture, not as part of the evaluated system, which is a rule-based verifier. The abstract, contributions, Table 1 and conclusion are aligned with this; the title makes no claim about differentiability and is unchanged.
2. **Formal FOL→KG mapping (R2.2, R1.10).** A new subsection defines how formulae become KG evidence obligations, with a worked example, an explicit open-world rule and updated pseudocode.
3. **Reference verdicts and E_3 (R1.2, R1.3, R3.4).** All 50 claims have KG-relative reference verdicts. We report accuracy, E_3 , a complexity breakdown and an explanation of every error, together with the limitation that the labels contain no false claim.
4. **Unweighted Verifier ablation (R1.1, R2.6, R3.5).** Trust-weighted accuracy is 0.72 and uniform-weight accuracy 0.66. The difference is not significant ($p = 0.375$), and we report it as such.
5. **Cascaded run on predicted FOL (R1.4, R3.3).** Accuracy falls to 0.16, entirely through abstention; grounding is now identified as the main bottleneck.
6. **Sensitivity analysis (R1.8, m2.3)** over thresholds, coefficients and reputation errors.
7. **Corrections to the trust methodology (R3.7, R2.3):** signatures are Ed25519 and simulated; five of nine graphs are signed; reputation values are hand-assigned by one author; retrieval in the walkthrough is TF-IDF.

Why the verdict numbers differ from the submitted version. The submitted paper reported 30 Verified, 1 Falsified and 19 Indeterminate verdicts. Writing down the FOL→KG mapping that Reviewer 2 requested (comment R2.2) exposed two defects in the submitted verifier.

- *Typing atoms.* Unary typing atoms such as *Patient*(x) were checked as evidence obligations; they found no evidence and forced every conditional claim to Indeterminate.
- *Negation polarity.* A negated atom was scored with the wrong polarity, which produced the single Falsified verdict. That verdict was a correct claim about penicillin, not the NSAID example the submitted text described.

The revised verifier implements the published mapping. It was frozen, with recorded file hashes, before any reference verdict was written, and it was not changed afterwards. Against the same reference verdicts, the submitted verdicts reach accuracy 0.58 and the revised verifier 0.72. Because these are different implementations, we do not attribute the difference to trust weighting. The submitted results remain reproducible from an archived copy of the submitted code, which we can provide on request.

What we have not done. We have not implemented the learned LTN objective, enlarged the dataset or KG, obtained clinical validation of the claims, or run the Neural-Only, RAG, Verdict-Only and Citations-Only baselines or a comparison with CoVe or KG-CRAFT. Reviewers 1 and 3 did not require these, and Reviewer 2 listed some of them for a resubmission. We have rescoped the paper so that it claims no more than the evidence supports, and the conclusion states that a larger, independently annotated benchmark including false claims is required. The main new results are also less favourable than the submitted framing suggested: on these labels a constant Verified answer would score 0.98 accuracy, and the full cascade reaches 0.16. We report both prominently.

Reference labels and use of AI. The candidate reference verdicts were drafted by an LLM from a written guideline and the KG catalogue, without access to any system output. Each claim was then reviewed and adjudicated by the first author, who changed 2 of the 50. We disclose this protocol in §5.2 and report no inter-annotator agreement, because the review was not independent of the draft. The guideline, labels and per-claim rationales will be released with the code on acceptance.

Reviewer 1

Comment R1.1.

Here’s my main problem: the experiments don’t test the paper’s central claim. The whole argument is that trust-weighted verification beats the alternatives, but not one of the five ablation baselines from Section 4.3 was ever run. The Unweighted Verifier is the one that counts, since it’s the only comparison that isolates what the trust weights buy you. Without it, the core claim rests on the design argument alone.

Response. We agree that this is the comparison that isolates what trust weighting contributes, and we have run it. The Unweighted Verifier uses the same FOL→KG mapping and verdict rule with $w_t = 1$ for every triple, on the same 50 claims and reference verdicts. It reaches accuracy 0.66 against 0.72 for the trust-weighted verifier (Table 10). The two differ on five claims, four in favour of trust weighting and one against (exact McNemar $p = 0.375$), so the difference is not statistically significant at this scale and we say so. The mechanism is visible in the four claims: a signed EMA contraindication contradicted by an unsigned blog or forum triple. Trust weighting lets the EMA evidence decide; uniform weighting produces a tie and the verifier abstains. In the opposite case (claim 11), support comes only from an unsigned review below τ . We report the signature-gated variant as well; it agrees with the linear score on all 50 claims. Following your remark that this is “the one that counts”, we have not run the other four baselines; §4.3 now states which baselines are and are not evaluated, and why.

Changes: §5.2 (“Unweighted Verifier”, Table 10); §4.3 (final paragraph)

Comment R1.2.

The headline number is also close to unfalsifiable as reported. Nineteen of fifty claims come back Indeterminate, and the paper reads that as an intended safety property. Maybe so, but without ground-truth verdicts, which the authors put in future work, “appropriately cautious” and “couldn’t figure it out” look

exactly the same in the data. Table 11 makes this sharper. Every conditional universal comes back Indeterminate, 7 out of 7. Every conditional negation, 5 out of 5. Meanwhile the 30 Verified cases sit almost entirely in atomic predicates (7 of 7) and conjunctions (9 of 9). So on the evidence here, the verdict function has been shown to work on the easy logical forms and nowhere else, with exactly one Falsified case in the entire batch. Annotating ground-truth verdicts for these same 50 claims and reporting E3, even at this small scale, would turn a distribution into a result. If the authors do only one thing from this review, it should be that.

Response. We agree, and this was the change we prioritised. Each of the 50 claims now has a reference verdict relative to the 98-triple KG, with the supporting triple identifiers recorded. The verdicts were written after the revised verifier was frozen and its hashes recorded, and the verifier was not changed afterwards. On gold FOL, accuracy is 0.72 and E_3 is 0.479 (Table 10), with per-class results, a confusion analysis, a breakdown by complexity (Table 11), and an explanation of every error. The results are less favourable than the submitted framing. The reference distribution is 49 Verified, 1 Indeterminate and 0 Falsified, so a constant Verified answer scores 0.98 accuracy. We now report that row in Table 10 and state in §5.2, the abstract and the conclusion that the evaluation measures how much KG-contained support the verifier recovers (35 of 49 claims, with no incorrect Verified verdict), not its ability to reject false claims.

To answer your point directly: of the 15 Indeterminate verdicts, one is the claim whose reference verdict is Indeterminate, and the other 14 are claims the KG supports but the mapping could not reach. On these data the Indeterminates are therefore mostly “couldn’t figure it out”, not “appropriately cautious”. We report them as such and no longer describe the Indeterminate rate as an intended safety property.

On annotation: the candidate verdicts were drafted by an LLM from a written guideline and the KG catalogue, without access to any system output, and every claim was then reviewed and adjudicated by the first author, who changed 2 of the 50. Because the review was not independent of the draft, we report no agreement statistic. We disclose this protocol in §5.2, and the labels, guideline and rationales will be released on acceptance.

The revised numbers differ from the submitted ones for a reason we should state openly. Preparing the reference verdicts required an explicit FOL→KG mapping (the concern raised by Reviewer 2, comment 2). Writing that mapping exposed two defects in the submitted verifier. First, unary typing atoms such as *Patient*(*x*) were retrieved as evidence obligations, found nothing, and forced every conditional universal to Indeterminate. Second, a negated atom was scored with the wrong polarity: for $\neg \textit{Treats}(\textit{Penicillin}, \textit{ViralInfection})$, the KG triple stating that penicillin does not treat viral infections was counted as contradicting the claim rather than supporting it. This produced the single submitted Falsified verdict. That verdict was claim 7 (penicillin), not the NSAID/CKD example as the submitted text stated. The revised verifier implements the mapping of §3.2. Scored against the same reference verdicts, the submitted verdicts reach accuracy 0.58 and the revised verifier 0.72. This comparison involves different implementations, so we do not attribute it to trust weighting.

Changes: §5.2 (“Reference Verdicts”, “Results”, “Results by Complexity and Error Analysis”; Tables 10–11); §3.2 (“From FOL Formulae to KG Evidence”); abstract; §7

Comment R1.3.

Table 11 makes this sharper. Every conditional universal comes back Indeterminate, 7 out of 7. Every conditional negation, 5 out of 5. Meanwhile the 30 Verified cases sit almost entirely in atomic predicates (7 of 7) and conjunctions (9 of 9). So on the evidence here, the verdict function has been shown to work on the easy logical forms and nowhere else, with exactly one Falsified case in the entire batch.

Response. This observation motivated the error analysis in §5.2. In the submitted verifier, all 12 conditional universals and conditional negations contain at least one typing or cohort atom that was checked as an evidence obligation and found no evidence (for example *Patient*(*x*)), and by the conjunctive rule this alone made each claim Indeterminate. The pattern you identified therefore came from the typing atoms described in our response to comment 2, not from missing cohort evidence. With the mapping of §3.2, 4 of 7 conditional universals and 4 of 5 conditional negations are now Verified, and each agrees with its reference verdict

(Table 11). The remaining errors are grouped by cause: lexical coverage (6 claims), universals without a safe mapping (4), structures on which the mapping abstains by design, namely nested function terms and a disjunction (3), and one claim below τ . There is still no Falsified claim in the reference set, so the concern that the verdict function is shown working only in easy cases is reduced but not eliminated. We state this limitation in §5.2.

Changes: §5.2 (Table 11 and “Results by Complexity and Error Analysis”)

Comment R1.4.

The staged evaluation also makes the system look better than it is. Stage A gets measured on FOLIO (LE 0.853) and the medical set (LE 0.801), but then the end-to-end run bypasses Stage A entirely and feeds in gold-standard FOL. Given that Stage A scores 0.479 on drug interactions and 0.688 on treatments, a pipeline that actually strings the stages together will do noticeably worse than that verdict distribution suggests. Push the same 50 claims through Stage A’s own output and you get a number for error propagation, which to my mind is the most interesting question this architecture raises.

Response. We ran this experiment. LogicLLaMA’s predicted FOL for the same 50 claims was passed through the identical frozen verifier and scored against the same reference verdicts; we did not extend the relation lexicon to fit the predictions. Accuracy is 0.16 on predicted FOL, compared with 0.72 on gold FOL (exact McNemar $p < 10^{-6}$). Thirty claims move from Verified to Indeterminate, none becomes Falsified, and all 7 Verified verdicts on predicted FOL are correct, so in this run grounding errors propagate as abstention rather than as wrong verdicts. The main cause is the over-generalisation failure already visible in §5.1: in 20 predictions a named drug is written as a unary predicate ($\forall x (Aspirin(x) \rightarrow \dots)$), so the formula no longer refers to a KG constant. As you anticipated, this is a much lower figure than the gold-FOL run, and we now identify grounding as the main bottleneck of the pipeline. We also renamed the gold-FOL experiment throughout, so that “end-to-end” refers only to this run.

Changes: new §5.3; Table 10 (last row); Table 11 (last column); abstract; §1.3; §7

Comment R1.5.

One inconsistency needs fixing. The abstract says “50 expert-annotated medical claims.” Section 5.2 says the set was curated by the authors and has not been independently validated by clinical experts. Section 5.2 is the accurate version and the abstract needs to match it. That matters more than usual when the claims are clinical.

Response. Corrected. The abstract and Section 4.2 now describe the claims and gold standard as author-curated and reference-based, consistent with Section 5.

Changes: abstract; §4.2 (“End-to-End Verdict Accuracy”); §5.1; §5.2

Comment R1.6.

On originality I land somewhere in the middle. The parts are all off the shelf: LogicLLaMA for grounding, LTN-style differentiable satisfiability, PROV-O and signed named graphs for provenance. What’s new is the composition, plus dropping trust weights into the satisfiability loss. That’s a sensible piece of integration, but it’s integration, not a new mechanism. I’d also soften Table 3, where the authors pick the comparison dimensions and, unsurprisingly, only their own system checks every box. Those tables tend to convince the people who made them more than anyone else.

Response. We agree. We removed the “first” novelty claim, narrowed it to the integration of provenance signals into the verification decision, and added a discussion of ClaimEval [Samadi et al., 2016], which already combines source credibility with soft logical reasoning. The Table 3 caption now states that the capability dimensions were chosen by the authors and that the table frames our requirements rather than

ranking systems, and the accompanying text says the contribution is integration rather than a new individual mechanism.

Changes: §2.4 (ClaimEval); §2.8 and Table 3 caption; §1.3

Comment R1.7.

Two smaller technical notes. Propositions 1 and 2 fall straight out of the definitions, no evidence gives you Indeterminate, a zero signature gives you zero weight, and they'd sit better as stated design properties than as propositions with proofs. As written they promise a depth the statements don't deliver.

Response. We agree. Propositions 1–3 are now stated as design properties P1–P3, without proof environments.

Changes: §3.2 (“Design Properties”)

Comment R1.8.

Second, the trust score is a hand-tuned convex combination, uniform coefficients at alpha 0.2, tau 0.5, delta 0.3, with reputations assigned by hand. Since the entire architecture hinges on this one quantity, it needs a sensitivity analysis. How much do the verdicts move when you vary the coefficients and thresholds, and what happens when a reputation score is simply wrong? That's cheap to run with what's already built.

Response. We ran the sensitivity analysis over the 50 claims, varying one factor at a time around the frozen configuration and scoring against the reference verdicts.

- *Thresholds.* Over $\tau \in \{0.3, \dots, 0.7\}$ and $\delta \in \{0.1, \dots, 0.5\}$, accuracy stays between 0.68 and 0.74 and at most 2 verdicts change.
- *Coefficients.* Removing any one signal changes at most one verdict. Over 500 Dirichlet draws of α , median accuracy is 0.72, but 9% of draws fall below 0.50; these place most of the weight on corroboration, which is almost zero on this KG.
- *Wrong reputation scores.* Noise of ± 0.1 or ± 0.2 on every source changed no verdict. Of the 36 pairwise swaps of source reputations, only swapping EMA with a low-reputation web source changes anything: 2 claims.

We interpret the stability cautiously, because it partly reflects a small KG with few conflicts. The frozen configuration was not re-selected from these runs.

Changes: new §5.4; §3.5 (the single α example now points to §5.4); §6.2

Comment R1.9.

Two small things: the caption on Table 10 already interprets the result as an intended property, which tells the reader what to think before they've looked, and would read better as plain description.

Response. Changed. The table the comment refers to has been replaced. The new result tables (Tables 10 and 11) have descriptive captions, and the interpretation is in the text.

Changes: §5.2 (Tables 10–11)

Comment R1.10.

And the open-world versus closed-world gap between FOL and OWL, noted in Section 6.1, deserves more than a paragraph, since it goes to whether Stage A's output can be checked against the KG at all.

Response. We agree that this deserved more than a paragraph, and your question “can Stage A output be checked against the KG at all?” is now answered with data. §3.2 states how the mapping operationalises the open-world reading: absence of a triple never falsifies; only an explicit negative-polarity triple does, and output outside the lexicon and KG vocabulary is Indeterminate by construction. The cascaded run (§5.3) measures the consequence: most LogicLLaMA output is not KG-checkable under this mapping, because predicate and constant vocabulary, not logical shape, is what fails.

Changes: §3.2 (“Open world”); §5.3

Reviewer 2

Comment R2.1.

Section 3.2 presents the integration with LTNs as the central innovation; Equation 6 defines a trust-weighted satisfiability loss, $w_t \cdot f_\theta(\varphi_j)$, where f_θ is the neural predicate grounding function proper to the LTN framework (Badreddine et al. 2022). But: - in none of the reported examples does a value of f_θ appear. All calculations shown use exclusively w_t . - Section 5.2 confirms that the actual implementation uses “rule-guided filters” and embedding similarity retrieval (sentence-transformers), not a differentiable neural component. The authors themselves acknowledge in Section 7 that the learned, differentiable LTN objective is not yet implemented and will be the subject of the “first follow-up paper,” replacing the current “rule-based retrieval+verdict combination.” Consequently, the title, abstract, and Eq. 6 promise a differentiable neuro-symbolic system, but the system actually built and evaluated is a rule-based system with semantic retrieval and trust thresholds.

Response. We accept this criticism. Of the two routes you offer, we have taken the rescoping. Equation 5 (the trust-weighted satisfiability objective) is now presented as the *specified target objective* of the architecture. A new paragraph (“Status of Equation 5”) states that the evaluated verifier contains no learned predicate f_θ and no gradient training, and that the trust weight enters the decision rule rather than a loss. The abstract, introduction, contribution list, Figures 2–3, the notation table, Table 1 (a new row: “LTN objective — specified only”) and the conclusion now say the same. We note two things about the system description in your comment. First, no embedding-based retrieval was used. The submitted implementation used TF-IDF, which the submitted text misdescribed as sentence-transformer and TransE-style retrieval; this is now corrected. The revised verifier evaluated in §5 uses no similarity retrieval at all: it locates evidence by exact lookup through the declared mapping of §3.2, while the walkthrough in §3.5 is labelled as coming from the TF-IDF path. Second, the verifier is rule-based. The paper no longer claims a differentiable verifier among its evaluated contributions. Implementing and evaluating Equation 5 is stated as future work.

Changes: §3.2 (“Trust-Weighted Satisfiability Objective” and “Status of Equation 5”); abstract; §1.1, §1.3; Table 1; Figures 2–3; §7

Comment R2.2.

The passage from a logical formula to KG evidence is described in an extremely vague manner. Algorithm 1 condenses it into a single line (“Retrieve supporting triples $S^+ \subseteq G$ for φ_i ”), and Section 3.2 mentions only generically “embedding-based similarity (TransE-style)... combined with rule-guided filters.” It is not clear how an FOL predicate is transformed into a retrieval query, nor how the semantic misalignment between the closed-world assumption (FOL) and the open-world assumption (OWL/RDF) is concretely resolved — a problem the authors themselves acknowledge in Section 6.1 but for which they offer only a preventive mitigation (constraining the grounding vocabulary to the target ontology), not a mapping algorithm. An explicit formal description of this step is recommended, with pseudocode or at least an end-to-end example showing the predicate→query→retrieved triples transformation.

Response. We agree. The mapping was implicit, and making it explicit changed the results. §3.2 now contains a subsection, “From FOL Formulae to KG Evidence”, that defines it: a declared relation lexicon

(FOL predicate $\rightarrow$ KG predicates, argument order and polarity); exact entity matching, so similarity alone is never evidence; how atoms, negated atoms, type restrictors, conditional universals (a stated group-level policy reading), and existentials become evidence obligations; and when the mapping abstains (disjunction, nested terms, unbound variables, out-of-lexicon predicates, parse failures). It also gives a worked example that traces predicate $\rightarrow$ query $\rightarrow$ triple (“Patients with chronic kidney disease should avoid NSAIDs”, met by a signed EMA triple), and Algorithm 1 is updated to match. The open-world treatment is stated as an operational rule rather than a mitigation: a missing triple never falsifies. The lexicon and verifier were frozen, with recorded hashes, before the reference verdicts were written. As explained in our response to Reviewer 1 (comment 2), writing the mapping down exposed two defects in the submitted verifier, and the revised results come from the verifier that implements this specification.

Changes: §3.2 (“From FOL Formulae to KG Evidence”, “Open world”, “Worked example”); Algorithm 1

Comment R2.3.

The dataset of 50 medical claims was curated by the authors themselves and not validated by independent clinical experts (explicitly admitted in Section 5.2). The KG of 98 triples is admittedly built ad hoc “to permit complete authorial control over signatures and provenance metadata,” is not representative of a real KG, and does not allow for a meaningful computation of E_1 nor an E_3 (macro-F1) evaluation against ground-truth verdicts; and Tables 5–6 and Scenario 3 present trust coefficients (rep, fresh, comp, corr) manually assigned by the authors “using a public registry,” without a description of the procedure or any inter-annotator reliability checks. These limitations are declared by the paper itself, which is commendable for transparency, but they make the numerical results (30/1/19 verdicts) demonstrative of architectural executability rather than of real accuracy a point that should be reflected more explicitly in the abstract and conclusion, where the numbers are instead presented with a weight that suggests broader validity.

Response. We agree with each point and address the three we can within this revision.

- *Reputation procedure.* §5.2 now describes it: each source was assigned a type and then a score ordered by institutional authority and editorial oversight. We also say plainly that the values are the authors’ judgement, with no inter-rater check, and that no external registry was used. The phrase “using a public registry” was inaccurate and has been removed.
- *The other signals.* §5.2 describes how each is obtained from the graph. For corroboration, the implementation treats distinct institutions as independent sources. This is a proxy for the PROV-O independence criterion in the definition, and Definition 1 now says so. In checking this, we found that several per-signal values printed in Tables 5–6 and in the α example of §3.5 did not match the saved output of the implementation, although the resulting trust weights and verdicts did. These values are now taken from the saved output.
- *Sensitivity.* Because the values are hand-assigned, we measured how much the verdicts depend on them (§5.4). Noise of ± 0.2 on every source changes no verdict, while swapping the most and least trusted sources changes two.

We compute E_1 (0.755) and E_3 , and state in §5.2, the abstract and the conclusion what the 50 claims can and cannot show.

We have not enlarged the dataset or the KG, and the claims are not clinically validated. We do not think these can be done properly within this revision. The paper is therefore rescoped to a proof of concept evaluated at small scale, and the conclusion states that a larger benchmark with independently annotated verdicts, sources of mixed reliability and false claims is required before general conclusions can be drawn.

Changes: §5.2 (“Knowledge Graph”, “Reference Verdicts”); §5.4; §4 (opening paragraphs; the former §4.4 on the role of human expertise was removed as it described a protocol not applied here); abstract; §7

Comment R2.4.

The dataset of 50 medical claims was curated by the authors themselves and not validated by independent clinical experts (explicitly admitted in Section 5.2). The KG of 98 triples is admittedly built ad hoc “to permit complete authorial control over signatures and provenance metadata,” is not representative of a real KG, and does not allow for a meaningful computation of E1 nor an E3 (macro-F1) evaluation against ground-truth verdicts; and Tables 5-6 and Scenario 3 present trust coefficients (rep, fresh, comp, corr) manually assigned by the authors “using a public registry,” without a description of the procedure or any inter-annotator reliability checks. These limitations are declared by the paper itself, which is commendable for transparency, but they make the numerical results (30/1/19 verdicts) demonstrative of architectural executability rather than of real accuracy a point that should be reflected more explicitly in the abstract and conclusion, where the numbers are instead presented with a weight that suggests broader validity.

Response. We agree. The distribution no longer appears as a result in the abstract, introduction or conclusion. Those sections now report accuracy against reference verdicts, the comparison with the Unweighted Verifier, the cascaded run, and the constant-answer reference point. The submitted table of the distribution has been replaced by accuracy results, and we have removed the claim that the Indeterminate rate was an intended safety property. §5.2 accounts for the 14 Indeterminate verdicts on claims the KG supports: 13 are mapping failures (lexical coverage, universals without a safe mapping, and structures the mapping abstains on by design), and one is a claim whose only support falls below τ . The description of the verdict rule in §3.2 no longer says that the system never asserts a verdict when evidence is insufficient; it now states that the rule applies only to the evidence the mapping has located.

Changes: abstract; §1.3; §3.2 (text after Equation 7); §5.2; §7

Comment R2.5.

Several components presented in the architectural flow are adopted from external works without always being clearly distinguished from the original contributions in the text (although correctly cited): - the grammar-masked decoding (Eq. 2) is from Li et al. (2024), not the authors. - the NL→FOL translator (LogicLLaMA) is from Yang et al. (2024). - the differentiable logic framework (LTN) is from Badreddine et al. (2022).

Response. We agree. A paragraph after the contribution list now separates adopted from original components. Adopted: LogicLLaMA [Yang et al., 2024]; the grammar mask of Equation 2 [Li et al., 2024]; the LTN-style objective [Badreddine et al., 2022]; PROV-O and signed named graphs; and the FEVER three-way labels. Original: the five-signal trust score and gated variant, the verdict and aggregation rules, the FOL→KG mapping, the evaluation protocol, and their integration. The text and Table 1 also state that the grammar mask is specified but not implemented: the evaluated Stage A uses unconstrained LogicLLaMA generation.

Changes: §1.3 (paragraph after the contribution list); §3.1 (“Constrained Decoding”); Table 1; §6.3

Comment R2.6.

Section 4.3 defines five ablation variants explicitly designed to isolate the contribution of each architectural component; however, none of these ablations is executed or reported. Thus, the implicit claim that trust weighting is responsible for the quality of the reported verdicts (Section 5.2, “Interpretation”) remains empirically unverified

Response. We ran the Unweighted Verifier, which isolates the effect of trust weighting against a uniform treatment of evidence (§5.2, Table 10). The trust-weighted verifier is more accurate (0.72 against 0.66), but the difference rests on five discordant claims and is not significant (McNemar $p = 0.375$). We report it as a small, mechanism-level effect, with the four claims in which a signed source overrides an unsigned contradiction, rather than as confirmation of the thesis. Details are in our response to Reviewer 1 (comment 1); §4.3 states which of the remaining baselines are not run and why.

Changes: §5.2 (“Unweighted Verifier”); §4.3

Comment m2.1.

The $\langle s, p, o \rangle$ notation for triples (Table 4, p. 14) is never made explicit as the standard RDF convention (subject-predicate-object), making the paper less accessible to readers who are not Semantic Web specialists.

Response. Added: $\langle s, p, o \rangle$ is now defined as an RDF subject–predicate–object triple, both in the notation table and where it first appears in the text.

Changes: §3 (notation table, Table 4; Definition 1)

Comment m2.2.

f_θ appears in Table 4 and in Eq. 6 without an explicit reference to where/how it is defined in the originating LTN framework, leaving the reader to infer the connection.

Response. Added. Under Equation 5, f_θ is now defined as the LTN grounding that maps a formula to a truth degree through learnable neural predicates, with a citation to Badreddine et al. [2022]. The notation table and the Figure 3 caption also state that it is specified but not implemented.

Changes: §3.2 (“Trust-Weighted Satisfiability Objective”); notation table (Table 4); Figure 3 caption

Comment m2.3.

The manual assignment of the α_i coefficients (Section 3.5, “Sensitivity to α Coefficients”) shows only one example of sensitivity analysis; systematic criteria or data for calibrating these hyperparameters are missing, despite Section 6.2 recommending a three-step calibration protocol.

Response. The single example has been replaced by a systematic analysis over thresholds, coefficients and reputation errors on all 50 claims (§5.4; see our response to Reviewer 1, comment 8). The worked example in §3.5 is kept only as an illustration, and §6.2 now refers to the analysis. We do not claim to have calibrated the coefficients: with the present reference set, calibrating them would risk fitting the 50 labels. The protocol in §6.2 remains a recommendation.

Changes: §5.4; §3.5; §6.2

Comment m2.4.

Table 7 (comparison with neural-only and standard RAG) is useful but remains a qualitative/analytical comparison, not an empirical comparative evaluation on the same data.

Response. We agree it is analytical. Its subsection title and caption already identified it as such, but the text beside it asserted how the two baselines would behave as though this had been observed. We have rewritten that paragraph. It now states that neither baseline was run, describes each by design, and states that whether trust weighting gives better-calibrated confidence than either baseline is not tested, because our confidence score is not evaluated for calibration. The empirical comparison we added is with the Unweighted Verifier (§5.2), which isolates trust weighting. We did not run Neural-Only or standard RAG, for the reasons given in §4.3.

Changes: §3.5 (paragraph after Table 7); §4.3; §5.2

Reviewer 3

Comment R3.1.

1. The role and contribution of differentiability are unclear Equation 6 introduces a trust-weighted differentiable satisfiability objective and states that gradient updates to θ prioritise high-trust evidence. However, it is not clear what f_θ is trained to represent, how it is trained, or how its learned satisfaction values influence the final verdict. The verdict in Equations 7 and 8 appears to depend directly on provenance-derived trust weights through max aggregation and fixed thresholds, without an obvious role for the differentiable component. The authors should clarify the connection between the differentiable objective, the learned neural predicates, and the final verification decision, and explain what differentiability contributes beyond a symbolic verifier with provenance-weighted evidence.

Response. You read the submitted system correctly, and we agree that the paper did not make this clear. The verdict depends only on the provenance-derived weights, through max aggregation and the fixed thresholds; the differentiable component played no role, because it was not implemented. We have rescoped accordingly. The new paragraph “Status of Equation 5” states that Equation 5 is the *specified target* objective, and that the evaluated verifier has no learned f_θ and no training, so that the trust weight enters the decision rule rather than a loss. It also states what differentiability would add if implemented: the ability to learn predicate groundings, so that evidence could be matched softly instead of through an exact lexicon, and to learn the trust coefficients from labelled data. Neither is claimed for the evaluated system. Our cascaded results (§5.3) show the exact-match mapping to be the main bottleneck, which is the case such learned grounding is meant to address; we present this as motivation, not as a result.

Changes: §3.2 (“Trust-Weighted Satisfiability Objective”, “Status of Equation 5”); Table 1; abstract; §1.3

Comment R3.2.

2. The differentiable verifier does not appear to be part of the evaluated system Relatedly, the conclusion states that a future version will implement the trust-weighted LTN objective in LTNtorch, replacing the current rule-based retrieval and verdict mechanism with a learned differentiable verification function. This makes the status of the central technical contribution unclear and Table 1 is not really transparent on this. Ideally, the differentiable verifier should be implemented and evaluated in this paper rather than deferred to follow-up work.

Response. We took the repositioning route your overall assessment allows. Table 1 now has a separate row, “LTN objective — specified only”, states for each stage whether it is implemented, evaluated, or only specified (including that the Stage A grammar mask is specified but not implemented), and lists the evidence reported for each. The contribution list, abstract and conclusion now describe the evaluated verifier as rule-based. We did not implement the differentiable verifier in this revision, and the conclusion lists it as future work without implying that it is complete.

Changes: Table 1; §1.3; §3.2; §7

Comment R3.3.

3. The reported “end-to-end” experiment bypasses Stage A Section 5.2 states that all three stages are executed in sequence, but the pipeline configuration subsequently states that Stage A is bypassed and gold-standard FOL representations are used in order to isolate Stage B and Stage C. I therefore do not think the resulting experiment should be described as a full end-to-end evaluation of the proposed architecture. One of the central challenges motivating the paper is precisely the possibility that natural-language claims are grounded incorrectly. These errors cannot propagate into the reported system-level results because gold FOL is supplied directly. The authors should either run the complete pipeline using predicted Stage A outputs or consistently describe the current experiment as an evaluation of Stages B and C conditioned on gold-standard logical forms.

Response. Agreed. We did both. The gold-FOL experiment is now described consistently as Stages B–C conditioned on gold-standard FOL, and we added a cascaded run that feeds LogicLLaMA’s predicted FOL through the same frozen verifier (§5.3). This lets grounding errors propagate into the verdicts, as you asked. On predicted FOL, accuracy is 0.16, compared with 0.72 on gold FOL, entirely through abstention, and we analyse the causes. Please see our response to Reviewer 1 (comment 4) for details.

Changes: §5.2 opening; new §5.3; abstract; §1.3; §7

Comment R3.4.

4. No end-to-end accuracy result is currently reported Section 4 identifies end-to-end verdict accuracy and E3 macro-F1 as important evaluation criteria. However, the current experiment only reports the distribution of outputs: 30 Verified, 1 Falsified, and 19 Indeterminate. The authors explicitly acknowledge that the current dataset does not contain independently expert-validated ground-truth verdicts sufficient to compute E3. A verdict distribution does not establish verification accuracy. For example, producing Indeterminate frequently may reflect desirable conservatism, but without an independent ground truth it is not possible to determine whether these abstentions are appropriate. This substantially limits claims about factual verification performance, “behavioural correctness,” or reduction in hallucination. At minimum, these claims should be narrowed. Preferably, the authors should provide independently validated verdict labels and report accuracy, macro-F1, and error analysis.

Response. We took your preferred route. We report accuracy, E_3 , per-class results, a breakdown by complexity and an error analysis against KG-relative reference verdicts for all 50 claims (§5.2, Tables 10–11). The claims about “behavioural correctness” and hallucination reduction have been removed. Your example also turned out to apply: of the 15 Indeterminate verdicts, 14 are claims the KG supports but the mapping could not reach, so they are coverage failures rather than conservatism, and we say so.

On independence, the reference verdicts are defined relative to the knowledge graph rather than as clinical judgements, and they are not independently validated. They were drafted by an LLM without access to system outputs and adjudicated claim by claim by the first author. We state this protocol in §4 and §5.2, and report no agreement statistic because the adjudication was not independent of the draft. We also report that the label distribution (49 Verified, 1 Indeterminate, 0 Falsified) means a constant Verified answer would score 0.98, so the evaluation does not measure rejection of false claims. The abstract and conclusion are narrowed accordingly. Please see our response to Reviewer 1 (comment 2), which also explains why the revised numbers differ from the submitted ones.

Changes: §5.2; Tables 10–11; abstract; §1.3; §4 (opening paragraphs; the former §4.4 on the role of human expertise was removed as it described a protocol not applied here); §7

Comment R3.5.

5. The proposed ablations and relevant baselines are not evaluated Section 4.3 defines Neural-Only, Standard RAG, Unweighted Verifier, Verdict-Only, and Citations-Only baselines. However, these comparisons are not reported in Section 5 and are explicitly deferred to future work. Without the Unweighted Verifier baseline in particular, the current experiments do not establish that provenance weighting improves verification outcomes relative to the same symbolic verifier with uniform evidence weights. The related-work section also discusses strong existing approaches such as CoVe and KG-CRAFT. KG-CRAFT is described as achieving state-of-the-art results on two real-world fact-checking datasets, yet there is no empirical comparison to it or another competitive KG-based verification approach. I do not necessarily expect all of these systems to be directly comparable on the authors’ medical dataset, but the paper should contain at least one meaningful empirical baseline that isolates the contribution of the proposed trust mechanism.

Response. We have run the baseline you identify as essential: the same symbolic verifier with uniform evidence weights, on the same claims and labels (§5.2, Table 10). Trust weighting is ahead (0.72 against 0.66), but the difference rests on five discordant claims and is not significant (McNemar $p = 0.375$), so we report

it as a mechanism-level observation rather than an established improvement. On CoVe and KG-CRAFT, we have not added an empirical comparison, and §4.3 now explains why. Both operate on unstructured text or on graphs built from claim reports, whereas our verifier takes a formula and a fixed 98-triple KG, so results on our 50 claims would not be comparable. The Neural-Only, RAG, Verdict-Only and Citations-Only baselines are stated as not evaluated.

Changes: §5.2 (“Unweighted Verifier”); §4.3

Comment R3.6.

6. The novelty claim should be narrowed and the related-work discussion expanded. The statement that “no existing differentiable logic framework incorporates source credibility into the reasoning objective” appears too broad. In particular, ClaimEval (Samadi et al., AAAI 2016) jointly estimates source credibility and claim correctness using Probabilistic Soft Logic, representing credibility relations as first-order rules with continuous truth values and Łukasiewicz soft logic. This appears closely related to the proposed use of trust within a soft logical reasoning objective and should be discussed explicitly. The authors should position their contribution more precisely, perhaps around the particular combination of provenance signals, cryptographic signatures, and LTN-style reasoning rather than source credibility in soft reasoning more generally.

Response. Thank you for this reference; it is closely related and we should have discussed it. We have removed the statement you quote, together with the accompanying “to our knowledge, the first” claim, and §2.4 now discusses ClaimEval [Samadi et al., 2016]. ClaimEval jointly estimates source credibility and claim correctness with Probabilistic Soft Logic, so the combination of credibility with soft logical reasoning is not new. We now position our work more narrowly. Credibility is not learned from agreement between claims but derived from per-triple provenance metadata: signature validity, reputation, freshness, PROV-O completeness and corroboration. It weights the KG evidence consulted when verifying formulae translated from LLM output. The §2.8 gap analysis now states that our contribution is the integration of existing mechanisms, not a new individual mechanism.

Changes: §2.4; §2.8 and Table 3 caption

Comment R3.7.

7. The trust scores require more methodological detail The reputation values used in the experiment are manually assigned by the authors using a “public registry.” The paper should explain how these values are derived and how the other trust signals would be obtained in practice. It should also clarify whether the RSA signatures attributed to sources such as WHO and EMA are genuine source-issued signatures or simulated signatures created for the experiment. This distinction seems important for assessing the practical applicability of the method.

Response. We agree that this distinction matters, and we should have stated it. The signatures are *simulated*. They are Ed25519 (not RSA-2048, which the submitted text stated in error), generated with key pairs we created for this study, not issued by WHO or EMA. They therefore certify the integrity of our curated graph, not institutional attestation. §5.2 now states this, together with the corrected list of signed sources (five of the nine graphs, 74 of the 98 triples). It also states that the verifier reads each graph’s declared signed status rather than re-verifying signatures at run time; we checked that all five stored signatures verify against their public keys. On derivation, §5.2 now describes how the reputation values were assigned: the authors’ type-ordered judgement, with no external registry, and “public registry” has been removed. It also explains how freshness, completeness and corroboration are computed from the graph’s provenance metadata. How much the verdicts depend on the hand-assigned reputations is measured in §5.4.

Changes: §5.2 (“Knowledge Graph”); §5.4

Comment R3.8.

8. Stage A’s consistency check appears to conflate incorrect grounding with factual contradiction. After translation, the paper checks whether $\Phi \cup \text{KB}$ entails a contradiction. If so, it triggers reparsing or manual review. Algorithm 1 instead returns an Indeterminate verdict before Stage B. This raises a conceptual concern. If a claim is correctly parsed but is contradicted by the knowledge base, this seems like precisely the type of evidence that should be handled by Stage B as negative evidence and potentially produce a Falsified verdict. For example, if the claim is $P(a)$ and the KB contains $\neg P(a)$, the contradiction does not necessarily imply a parsing error. Did the authors consider this case, or did I misunderstand something there?

Response. You did not misunderstand, and we agree with the concern. A correctly parsed claim $P(a)$ that the KB contradicts is negative evidence, not a grounding error, and treating it as a Stage A failure conflated the two. We have changed the design. Stage A now performs only a well-formedness check (does the formula parse under the target grammar), and KB contradiction is handled in Stage B, where an explicit negative-polarity triple is counted as contradicting evidence and can produce a Falsified verdict. Algorithm 1 and the text of §3.1 are updated accordingly. The implementation matches: the evaluated verifier returns Falsified in this case. The walkthrough in §3.5 shows an example, a claim of safety contradicted by a signed EMA contraindication.

Changes: §3.1 (“Well-Formedness Check”); Algorithm 1; §3.5 (Scenario 2)

Comment R3.9.

9. The claim-level aggregation rule requires additional justification. Equation 9 states that an overall claim is Falsified only if at least one constituent formula is Falsified and none is Indeterminate. If one constituent is Falsified and all others are Verified, the resulting Falsified verdict is straightforward. However, if one constituent is clearly Falsified while another is Indeterminate, the overall claim becomes Indeterminate. For a conjunctive claim, this differs from classical logic, where one false conjunct is sufficient to falsify the conjunction regardless of whether another conjunct is unknown. I understand that this may be an intentional safety-oriented or epistemic policy, but the rationale for allowing Indeterminate to dominate Falsified should be explained more clearly.

Response. It is intentional, and §3.2 now explains the rationale and states the departure from classical conjunction explicitly. A verdict reports what the KG establishes about the claim as stated. Because the constituent formulae come from an imperfect Stage A translation, an Indeterminate constituent may reflect a grounding or coverage gap rather than a genuine unknown. We therefore reserve Falsified for claims whose every constituent has been resolved against the KG, which we judge the safer error in high-stakes use. The cost is lower Falsified recall on compound claims. A classical policy, under which any Falsified conjunct falsifies the claim, is a one-line configuration change, and we name it as an alternative. We do not evaluate it here, because the reference set contains no Falsified claim.

Changes: §3.2 (“Claim-Level Verdict Aggregation”)

Comment R3.10.

The selection procedure for the 254 FOLIO examples should be reported, including whether they overlap with LogicLLaMA’s fine-tuning data.

Response. Now reported in §5.1. The 254 examples are the first 254 items, in file order, of the pre-parsed FOLIO release distributed by the LogicLLaMA authors (`folio_parsed.json` in MALLS-v0); the evaluation run was stopped at that point, and the subset was not otherwise selected. Because this release comes from the model’s own authors, overlap with LogicLLaMA’s fine-tuning data is likely and we could not rule it out. §5.1 therefore describes the FOLIO score as a reproduction check that may be optimistic, not as a held-out estimate.

Changes: §5.1 (“Datasets”)

Comment R3.11.

The paper also states that tree-edit distance is reported as a soft grounding metric, but I could not find this result in Section 5.

Response. Now reported. On the 50 medical claims, the mean normalised tree-edit distance between predicted and gold FOL is 0.581 ± 0.360 over the 47 predictions our parser can read, with 3 exact matches. We used the Zhang–Shasha algorithm with unit costs and positional variable names, normalised by the larger tree. We report it alongside the cascaded run, because it helps explain why a high logical-equivalence score (0.801) coexists with low KG-checkability.

Changes: §5.3; §4.1 (metric definition now points to §5.3)

Comment R3.12.

Finally, there is some inconsistency between the abstract’s description of the medical claims as ”expert-annotated” and the later statement that they were author-annotated and not independently clinically validated.

Response. Corrected. The abstract and Section 4.2 now describe the claims and gold standard as author-curated and reference-based, consistent with Section 5.

Changes: abstract; §4.2 (“End-to-End Verdict Accuracy”); §5.1; §5.2

We thank the reviewers again. Their reports substantially improved the paper, not least by prompting the experiments that showed where the system falls short.

Sincerely,

David Farrugia and Alexiei Dingli

Department of Artificial Intelligence, Faculty of ICT, University of Malta

References

- Samy Badreddine, Artur d’Avila Garcez, Luciano Serafini, and Michael Spranger. Logic Tensor Networks. *Artificial Intelligence*, 303:103649, February 2022. ISSN 0004–3702. doi: 10.1016/j.artint.2021.103649.
- Zelong Li, Wenyue Hua, Hao Wang, He Zhu, and Yongfeng Zhang. Formal-LLM: Integrating Formal Language and Natural Language for Controllable LLM-based Agents, August 2024. arXiv:2402.00798 [cs].
- Mehdi Samadi, Partha Talukdar, Manuela Veloso, and Manuel Blum. ClaimEval: Integrated and flexible framework for claim evaluation using credibility of sources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, pages 222–228, 2016. doi: 10.1609/aaai.v30i1.9996.
- Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. Harnessing the Power of Large Language Models for Natural Language to First-Order Logic Translation. In *Proceedings of the 62nd Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 6942–6959, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.375.