Neurosymbolic Artificial Intelligence
XX(X):2–40

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

A Trust-Centric Neuro-Symbolic Architecture for Verifiable Large Language Model Claims

David Farrugia[1] and Alexiei Dingli[1]

Abstract

Large Language Models (LLMs) exhibit impressive fluency yet remain prone to hallucinations, opaque reasoning, and inconsistent use of source quality, limitations that inhibit adoption in high-stakes domains. This paper formalises a trust-centric neuro-symbolic pipeline that unifies three previously separate capabilities: the reliable translation of unconstrained LLM outputs into First-Order Logic (FOL); the verification of those formulae against provenance-rich, digitally signed Knowledge Graphs (KGs); and the generation of user-oriented explanations designed to calibrate human trust. The key architectural contribution is the integration of digital signatures and reputation scores as first-class weights inside a differentiable satisfiability objective, enabling trust-weighted proof selection rather than uniform treatment of training facts. The architecture has been implemented as a runnable proof-of-concept covering all three stages, and we report end-to-end results on a curated pharmacological knowledge graph of 98 signed triples drawn from nine sources. Across 50 expert-annotated medical claims, the pipeline produces 30 Verified, 1 Falsified, and 19 Indeterminate verdicts; the high Indeterminate rate is an intended safety property of the architecture rather than a failure mode, reflecting conservative behaviour under sparse evidence. Complementary evaluation of the grounding stage on the FOLIO benchmark and a domain-specific medical-claims dataset shows strong Logical Equivalence (LE $= 0.853$) on FOLIO but degradation on complex relational claims (drug interactions: LE $= 0.479$), quantifying the domain adaptation gap and confirming the necessity of the downstream verification layer. By delivering a provenance-aware and explainable verification layer, this research provides a principled blueprint for deploying trustworthy language technologies in safety-critical settings.

Prepared using sagej.cls

Keywords
Large language models, Neuro-symbolic AI, Knowledge graphs, Provenance, Trust calibration, First-order logic, Semantic Web

1 Introduction

Large Language Models (LLMs) have achieved strong results across natural language understanding and generation tasks (Brown et al. 2020; Vaswani et al. 2017). However, their tendency to "hallucinate" (generating factually incorrect but plausible-sounding statements) remains a major barrier to their adoption in high-stakes environments such as healthcare, finance, and legal advisory (Ji et al. 2023; Augenstein et al. 2024; Choudhury and Chaudhry 2024). In these domains, erroneous outputs can lead to real-world harm, ranging from misdiagnosis to significant financial loss (Wu et al. 2025; Moglia et al. 2024).

Current efforts to ensure LLM reliability often rely on purely statistical methods, such as Retrieval-Augmented Generation (RAG) or human-in-the-loop reviews (Patil and Gudivada 2024; Ouyang et al. 2022; Demartini et al. 2020). While these provide incremental improvements, they lack the formal rigour required for absolute verification and often treat all retrieved evidence as uniformly reliable, ignoring the provenance and cryptographic integrity of source data (Hartig 2009; Rahimzadeh Holagh and Mohebbi 2019).

1.1 Closing the Triangle: Fluency, Rigour, and Trust

We argue that the path toward trustworthy language technologies requires bridging a three-fold gap between largely independent capabilities (Marcus and Davis 2019; Besold et al. 2021). We conceptualise this as a triangle (Figure 1) where each vertex represents an essential requirement:

- **LLM Fluency:** Today's models excel at context-rich expressiveness but lack formal guarantees of correctness or grounding (Brown et al. 2020; Ji et al. 2023).
- **Symbolic Rigour:** Neuro-symbolic and symbolic systems provide formal logic and correctness guarantees but often assume clean, curated data and struggle to scale with the noisy information of the open web (Garcez et al. 2009; Besold et al. 2021).
- **Trust and Provenance Mechanisms:** Standards in the Semantic Web and information security provide mature methods for tracking source quality and data

[1]Department of Artificial Intelligence, Faculty of ICT, University of Malta, Msida, Malta

Corresponding author:
David Farrugia, Department of Artificial Intelligence, Faculty of ICT, University of Malta, Msida MSD 2080, Malta.
Email: david.farrugia@um.edu.mt

integrity, yet these remain largely separate from AI generation loops (Artz and Gil 2007; Golbeck 2006).

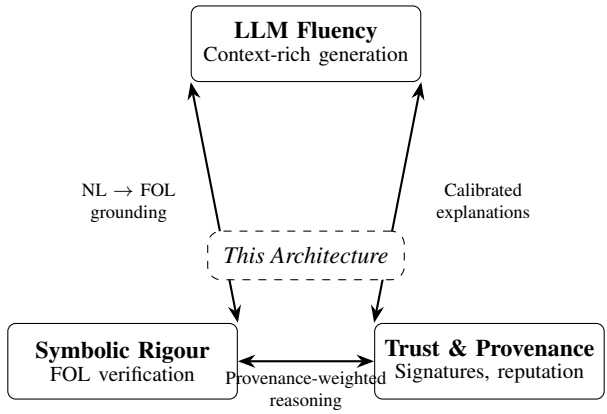

Figure 1. The "Closing the Triangle" framework. Most existing approaches address at most two of the three vertices. Our architecture integrates all three by embedding LLM outputs within a logically grounded, provenance-aware verification loop.

Our architecture *closes this triangle* by embedding LLMs within a logically grounded, provenance-aware framework. By integrating digital signatures and reputation scores directly into a differentiable logic objective, we enable trust-weighted verification where metadata dictates the actual reasoning outcomes (Badreddine et al. 2022; Artz and Gil 2007).

1.2 Research Objectives

To formalise this vision, this paper addresses the following research questions through the proposed architectural blueprint:

RQ1. **Reliable Semantic Grounding:** How can unconstrained LLM claims be reliably translated into First-Order Logic (FOL) to enable syntactically well-formed and globally consistent verification (Yang et al. 2024; Brunello et al. 2025)?

RQ2. **Trust-Aware Verification:** To what extent does a provenance-aware, scalable neuro-symbolic layer reduce factual inaccuracies while maintaining computational tractability (Badreddine et al. 2022; Artz and Gil 2007)?

RQ3. **Explanations for Calibrated Trust:** Which forms of explanatory feedback most effectively foster appropriate reliance, ensuring users neither over-trust nor under-trust the system (Kim et al. 2024a; Romeo and Conti 2025)?

Together, these questions outline a roadmap for deploying language models that are fluent, validated, and trustworthy by design (Floridi and Cowls 2019; Bellogín et al. 2024). This paper contributes a formally specified architectural design *accompanied by*

a runnable proof-of-concept implementation that covers all three pipeline stages and is publicly released for independent replication and benchmarking. We further provide empirical evidence from two complementary experiments (Section 5): an evaluation of the grounding stage on FOLIO and a domain-specific medical-claims dataset, and an end-to-end run of the implemented pipeline against a curated, digitally signed pharmacological knowledge graph.

1.3 Main Contributions

The main contributions of this work are:

1. A five-signal provenance-weighted trust score that integrates digital signatures, reputation, freshness, completeness, and corroboration as first-class weights in a differentiable logic objective (Definition 1), with a signature-gated variant for regulated environments (Equation 5).
2. A conflict-aware verdict function with configurable trust thresholds that returns an explicit *Indeterminate* verdict when evidence is insufficient or contradictory, rather than forcing a binary decision (Equation 8).
3. A reusable, staged evaluation protocol with component-level and system-level checkpoints, ablation baselines, and a controlled role for human expertise (Section 4).
4. A publicly available reference implementation of the full three-stage pipeline, including the trust weight computation, conflict-aware verdict function, and explanation trace generator, together with a curated pharmacological knowledge graph carrying digital signatures and provenance metadata for nine sources.
5. Empirical evidence from both stage-level and end-to-end experiments: the grounding stage achieves LE $= 0.853$ on FOLIO and LE $= 0.801$ on domain-specific medical claims (with sharp degradation on drug interactions, LE $= 0.479$), while the end-to-end pipeline applied to the 50-claim gold standard produces a verdict distribution of 30 Verified, 1 Falsified, and 19 Indeterminate, demonstrating both the architecture's reasoning capability and its conservative behaviour under sparse evidence.

Table 1 summarises the current maturity of each pipeline component to set clear expectations about what this paper delivers.

1.4 Paper Organisation

The remainder of this paper is structured as follows. Section 2 surveys the relevant literature across the Semantic Web trust stack, LLM hallucination, neuro-symbolic AI, formal logic translation, fact verification, self-refinement methods, and socio-technical trust. Section 3 presents the three-stage architectural specification with formal definitions, mathematical formulations, and an illustrative walkthrough grounded in real outputs from the implementation. Section 4 specifies a reusable evaluation methodology with component-level and system-level checkpoints. Section 5 reports preliminary experimental results from both the grounding stage (FOLIO and medical claims) and the

Table 1. Component maturity. Each pipeline stage is classified by its current status and the evidence provided in this paper.

Component	Status	Evidence
Stage A: $NL \rightarrow FOL$ grounding	Implemented & evaluated	LogicLLaMA tested on FOLIO (n=254) and medical claims (n=50); LE, BLEU, parse rate reported
Stage B: Trust-aware verification	Implemented & evaluated	Five-signal trust weight, retrieval, and conflict-aware verdict implemented; end-to-end run reported over 50 medical claims
Stage C: Explanatory feedback	Implemented	Explanation trace generator emits verdict, ranked evidence, trust signals, and recommendations
Trust weight function	Implemented & evaluated	Five-signal model with signature-gated variant; properties proved; weights computed end-to-end
Evaluation protocol	Partially executed	E_2 and end-to-end verdict distribution computed; E_1, E_3, E_4 defined for follow-up work

end-to-end pipeline (50 medical claims against the signed pharmacological knowledge graph). Section 6 discusses assumptions, limitations, and adversarial considerations. Section 7 concludes the paper and outlines future research directions.

2 Background and Related Work

The development of a trust-centric neuro-symbolic architecture requires the synthesis of seven primary research strands: Semantic Web trust layers, LLM capabilities and hallucination, natural language-to-logic translation, neuro-symbolic AI frameworks, automated fact verification, self-refinement and chain-of-verification methods, and socio-technical trust calibration. This section critically reviews each strand and concludes with a gap analysis (Table 3) that motivates our contribution.

2.1 Semantic Web Stack and the Trust Layer

The Semantic Web vision aims to transform the web into a machine-processable ecosystem by annotating data with explicit semantics using standards such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL) (Shadbolt et al. 2006; Koivunen and Miller 2001; Horrocks 2005). At the apex of the Semantic Web "layer cake" sits the Trust layer (Artz and Gil 2007). While the lower layers facilitate data interchange and reasoning, the trust layer is designed to verify the authenticity and reliability of both statements and sources (Artz and Gil 2007; Golbeck 2006).

Mature mechanisms exist for provenance tracking (e.g., the W3C PROV Ontology, PROV-O), cryptographic signatures, and reputation scoring (Moreau et al. 2011; Carroll et al. 2005; Golbeck 2006). Named Graphs (Carroll et al. 2005) allow individual RDF statements to carry metadata such as authorship, timestamp, and signature status, providing the foundation for per-triple trust annotation. Reputation models, such as the probabilistic trust propagation framework of Golbeck (2006), compute trust transitively across social networks, while ontology-based approaches encode trust requirements directly into OWL axioms (Huang and Fox 2006).

Despite these advances, wide-scale adoption of Semantic Web trust mechanisms remains limited due to challenges in standardisation, computational cost, and integration with real-time unstructured data streams (Harth et al. 2011; Hitzler 2021). Recent surveys on the convergence of knowledge graphs and AI note that trust and provenance remain "the missing piece" in most knowledge-augmented systems (Scherp et al. 2024; Pan et al. 2024). Our architecture addresses this gap by operationalising these mature Semantic Web trust protocols as first-class parameters inside a neural reasoning loop.

2.2 LLM Capabilities and the Challenge of Hallucination

Transformer-based Large Language Models, from BERT (Devlin et al. 2019) and GPT (Brown et al. 2020) to their successors (OpenAI et al. 2024), have achieved state-of-the-art results across Natural Language Processing (NLP) tasks (Vaswani et al. 2017; Qiu et al. 2020). However, their "black-box" nature and susceptibility to hallucinations present significant risks in high-stakes domains (Ji et al. 2023; Bender et al. 2021).

Hallucinations, where a model generates contextually plausible but factually incorrect statements, are broadly categorised into *intrinsic* hallucinations (contradicting the source material) and *extrinsic* hallucinations (fabricating unsupported claims) (Ji et al. 2023). Both types arise because LLMs generate text based on learned token-level probabilities rather than explicit knowledge bases (Augenstein et al. 2024). The problem is amplified in safety-critical domains: clinical studies report that LLM-generated medical summaries contain clinically significant hallucinations at non-trivial rates (Asgari et al. 2025), while legal applications exhibit similar vulnerabilities (Choudhury and Chaudhry 2024).

Existing mitigation strategies include Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022), Retrieval-Augmented Generation (RAG) (Patil and Gudivada 2024), and post-hoc fact-checking pipelines (Thorne et al. 2018). While RLHF reduces overtly harmful outputs, it does not provide formal guarantees of factual correctness. RAG improves grounding by retrieving relevant documents but treats all retrieved evidence as uniformly reliable, providing no mechanism for assessing source provenance (Lin et al. 2024). Our work goes beyond these approaches by coupling retrieval with formal logic verification and provenance-weighted trust assessment.

2.3 Natural Language to Formal Logic Translation

The reliable translation of natural language into formal logic ($NL \rightarrow FOL$) is a long-standing challenge at the intersection of computational linguistics and knowledge representation (Bos 2008). Early approaches employed Combinatory

Categorial Grammars (CCG) (Stanojević and Steedman 2019) and Abstract Meaning Representations (AMR) (Banarescu et al. 2013) to produce intermediate semantic structures that could be mapped to first-order formulae. While effective for well-formed sentences, these systems struggle with the informal, noisy outputs typical of LLMs.

Recent work has explored using LLMs themselves as translators. Olausson et al. (2023) demonstrate that chain-of-thought prompting can guide LLMs to produce first-order logic statements with reasonable accuracy on standard benchmarks. Pan et al. (2023) propose Logic-LM, which combines LLM-generated logical forms with a symbolic solver to improve reasoning accuracy. Li et al. (2024) introduce Formal-LLM, which constrains LLM decoding through a grammar mask to ensure syntactic well-formedness of the logical output. Taking a different approach, Yang et al. (2024) present LogicLLaMA, a LoRA-finetuned LLaMA-7B model trained specifically for direct $NL \rightarrow FOL$ translation, achieving strong logical equivalence scores on the FOLIO benchmark. We adopt LogicLLaMA as the grounding model in our feasibility study (Section 5).

More recently, Putra et al. (2026) propose NL2Logic, which introduces an Abstract Syntax Tree (AST) as an intermediate representation between natural language and solver-ready code. Their recursive semantic parser decomposes each sentence clause-by-clause, and a deterministic AST-guided generator produces syntactically correct output. On the FOLIO, LogicNLI, and ProofWriter benchmarks, NL2Logic achieves 99% syntactic accuracy and improves semantic correctness by 30% over prior baselines, demonstrating that decoupling logical parsing from code generation substantially reduces hallucination in the translation process.

Evaluation of $NL \rightarrow FOL$ systems has been advanced by benchmarks such as FOLIO (Han et al. 2024), which provides expert-annotated natural language–FOL pairs, and by systematic evaluations of LLM-based translators across multiple logic formalisms (Brunello et al. 2025; Yang et al. 2024). Despite progress, exact-match accuracy for complex multi-predicate sentences remains limited, with preliminary evaluations suggesting that current LLM-based translators struggle to achieve reliable performance on such inputs (Brunello et al. 2025; Yang et al. 2024), highlighting the need for constrained decoding or structured decomposition strategies (Putra et al. 2026), a gap our Stage A directly addresses.

2.4 Neuro-Symbolic AI and Differentiable Logic

Neuro-symbolic AI aims to unite the pattern-learning abilities of neural networks with the formal expressiveness of symbolic logic (Garcez et al. 2009; Garcez and Lamb 2023). A recent survey by Cheng et al. (2024) provides a comprehensive taxonomy of neural-symbolic methods for knowledge graph reasoning, covering KG completion, complex query answering, and logical rule learning; the survey concludes that symbolic reasoning offers interpretability and generalisability through logical rules but struggles with scalability, while neural methods scale well but fall short in modelling higher-order dependencies, motivating hybrid integration. This hybrid paradigm is increasingly positioned as a linchpin for validating LLMs, with a cross-study survey indicating that

coupling LLMs with explicit symbolic or knowledge-graph verifiers typically reduces hallucination rates by approximately 45–75%, though these figures are not strictly comparable across datasets (Pan et al. 2023; Olausson et al. 2023; Cao 2024; Tang et al. 2024; Quan et al. 2024; Wadden et al. 2020; Ciampaglia et al. 2015; Hou et al. 2024; Hanselowski et al. 2019; Atanasova et al. 2020; Shi and Weninger 2016).

Table 2 summarises representative results across these studies. Purely neural baselines yield only marginal gains, whereas knowledge-graph and hybrid neuro-symbolic approaches achieve substantially larger error reductions (median $\approx 54\%$).

Table 2. Error reduction from verification modules across representative studies. Neural-only baselines yield modest gains, whereas KG and hybrid neuro-symbolic approaches reduce error rates by approximately 45–75%. Figures are not strictly comparable across datasets and evaluation protocols.

Study	Dataset	Base LLM	Type	Base Err.%	+ Module	Δ%
Hanselowski et al. (2019)	FEVER	BiLSTM	Neural	49.1	35.8	–27.1
Atanasova et al. (2020)	FEVER	RoBERTa	Neural	31.4	30.0	–4.5
Ciampaglia et al. (2015)	Wiki Claims	—	KG	39.0	9.0	–76.9
Shi and Weninger (2016)	PolitiFact	—	KG	35.0	8.0	–77.1
Wadden et al. (2020)	SciFact	BERT	Hybrid	23.1	12.7	–45.0
Olausson et al. (2023)	ProofWriter	GPT-J	Hybrid	24.0	11.0	–54.2
Pan et al. (2023)	LogicQA	LLaMA-7B	Hybrid	19.4	9.6	–50.5
Tang et al. (2024)	FEVER	GPT-3.5	Hybrid	29.2	15.1	–48.3
Cao (2024)	FEVER	GPT-2 XL	Hybrid	26.4	11.4	–56.8
Hou et al. (2024)	WebQSP	ChatGPT	Hybrid	34.7	17.3	–50.1
Quan et al. (2024)	e-SNLI	GPT-4	Hybrid	64.0	16.0	–75.0

Several frameworks have emerged to make logic differentiable, allowing symbolic constraints to guide neural learning via gradient-based optimisation. Logic Tensor Networks (LTNs) (Badreddine et al. 2022) compute a real-valued "satisfaction degree" for each grounded formula. DeepProbLog (Manhaeve et al. 2021a,b) embeds neural predicates within probabilistic logic programs; NeurASP (Yang et al. 2021) uses answer set programs; DeepStochLog (Winters et al. 2022) employs stochastic definite clause grammars; and Logic Neural Networks (LNNs) (Riegel et al. 2020) assign each neuron a logical connective interpretation.

The key limitation shared by all these frameworks is their treatment of input facts: every triple or ground atom is assumed to be equally reliable (Badreddine et al. 2022). In real-world knowledge graphs, however, facts vary considerably in their provenance, recency, and trustworthiness. No existing differentiable logic framework incorporates source credibility into the reasoning objective. Our architecture addresses this gap by extending the LTN paradigm with a provenance-derived trust weight that modulates the satisfaction degree of each grounded formula, so that verification outcomes are driven by source quality rather than treating all evidence uniformly. This trust-weighted satisfiability objective (formalised in Section 3) is, to our knowledge, the first to integrate digital signatures, reputation scores, and provenance completeness as first-class parameters inside a differentiable reasoning loop.

2.5 Fact Verification and Claim Checking

Automated fact verification has received substantial attention following the release of large-scale benchmarks. The Fact Extraction and VERification (FEVER) dataset (Thorne et al. 2018) introduced a three-class task (Supported, Refuted, Not Enough Info) that closely mirrors our {Verified, Falsified, Indeterminate} verdict set. LIAR (Wang 2017) provides real-world political claims with fine-grained truthfulness labels, while SciFact (Wadden et al. 2020) targets scientific claim verification against research abstracts. More recently, MiniCheck (Tang et al. 2024) has explored lightweight LLM-based fact-checking with promising efficiency–accuracy trade-offs.

Knowledge-graph-based verification takes a complementary approach by resolving claims against structured knowledge rather than unstructured text (Ciampaglia et al. 2015; Shi and Weninger 2016). Konstantinovskiy et al. (2021) survey the pipeline components required for automated claim detection and verification, while recent work has explored combining KG-based evidence retrieval with neural reasoning (Hou et al. 2024; Cao 2024). Most recently, KG-CRAFT (Lourenço et al. 2026) constructs a knowledge graph from claims and associated reports, then formulates contrastive questions grounded in the graph structure to guide LLM-based verification. KG-CRAFT achieves state-of-the-art results on two real-world fact-checking benchmarks (LIAR-RAW and RAWFC), with F1 improvements of up to 44 percentage points over prior methods, providing strong evidence that structured knowledge substantially enhances LLM verification capabilities.

A complementary line of recent work evaluates LLM factuality at a finer granularity than binary claim labels. FActScore (Min et al. 2023) decomposes long-form generations into atomic facts and computes the percentage supported by a reliable knowledge source, revealing that even ChatGPT achieves only 58% atomic factual precision on biography generation. HaluEval (Li et al. 2023) provides a large-scale benchmark of 35,000 hallucinated samples across question answering, dialogue, and summarisation, finding that ChatGPT fabricates unverifiable information in approximately 19.5% of user queries. More recently, SAFE (Wei et al. 2024) employs LLM agents augmented with web search to evaluate individual facts in long-form responses, demonstrating that automated evaluators can match or exceed crowdsourced human annotators at a fraction of the cost. These fine-grained methods represent important advances in measuring LLM factuality; however, they share a common limitation with the claim-level systems described above: all retrieved evidence is treated as uniformly reliable, irrespective of its provenance or source credibility.

A key gap in existing fact-verification systems is their treatment of evidence quality. Most pipelines assume that if a claim is supported by retrieved evidence, it is true, regardless of whether that evidence comes from a peer-reviewed source or an anonymous web page. Our architecture fills this gap by assigning provenance-weighted trust scores to every piece of evidence before it enters the reasoning loop.

2.6 Self-Refinement and Chain-of-Verification

A recent family of methods improves LLM output quality through iterative self-correction without external supervision. Self-Consistency (Wang et al. 2023) samples multiple reasoning paths via chain-of-thought prompting and selects the most frequent answer by majority vote, exploiting the intuition that correct solutions are reachable through diverse reasoning trajectories. Self-Refine (Madaan et al. 2023) extends this idea to a generate–feedback–refine loop in which the same LLM critiques and iteratively improves its own output, achieving substantial gains across dialogue, code, and mathematical reasoning tasks. Reflexion (Shinn et al. 2023) goes further by equipping language agents with an episodic memory of verbal self-reflections, enabling learning from past failures without weight updates.

Most relevant to our evaluation philosophy, Chain-of-Verification (CoVe) (Dhuliawala et al. 2024) decomposes hallucination reduction into four stages: draft an initial response, generate verification questions, answer them independently (preventing bias from the original output), and revise the response. CoVe demonstrates that verification questions are answered more accurately than the claims in the original generation, yielding significant hallucination reduction across list-based, span-extraction, and long-form tasks.

These methods share a common limitation: all verification is performed *within the neural model itself*. Self-Consistency relies on statistical agreement among sampled paths rather than logical proof; Self-Refine and Reflexion depend on the model's own ability to detect and correct errors, which is bounded by the same training distribution that caused the error. CoVe's independent answering of verification questions mitigates but does not eliminate this circularity, since the verifier and the generator share the same parametric knowledge. None of these approaches consult external structured knowledge, require formal logical consistency, or account for the provenance or reliability of evidence.

Our architecture addresses this gap by grounding verification in a symbolic knowledge base with explicit trust metadata. Rather than asking the model to check itself, we delegate verification to a formal reasoning engine operating over provenance-weighted facts. The CoVe principle does, however, inspire our evaluation protocol (Section 4): just as CoVe verifies each claim through independent sub-checks, our staged evaluation validates each pipeline component against its own gold standard before assessing system-level behaviour.

2.7 Socio-Technical Trust and Explainability

A technically correct AI solution can still falter if users do not appropriately rely on its outputs (Romeo and Conti 2025; Carnat 2024). Research on human–AI interaction reveals two complementary failure modes. "Automation bias," the tendency to over-rely on AI decisions, can lead to uncritical acceptance of incorrect outputs (Carnat 2024; Bo et al. 2024). Conversely, "algorithm aversion," the refusal to trust accurate AI recommendations, reduces uptake in high-value applications (Dietvorst et al. 2015). Both failure modes are exacerbated by opaque reasoning: when users cannot inspect

the basis for a verdict, they lack the information needed to calibrate their trust appropriately (Lipton 2018).

Explainability research distinguishes between *global* explanations (characterising overall model behaviour) and *local* explanations (justifying individual predictions) (Doshi-Velez and Kim 2017). For fact-verification systems, local explanations are particularly relevant: users need to understand *which* evidence supported or contradicted a specific claim, and *how reliable* that evidence is. Recent surveys on explainability in LLM-based systems (Zhao et al. 2024; Lu et al. 2025) note that most explanation methods focus on attention-weight visualisation or natural language rationales, neither of which provides formal guarantees about the reasoning process.

Studies in clinical and legal domains further highlight the importance of domain-appropriate explanation design. Mirzaei et al. (2024) find that clinicians prefer structured, evidence-linked explanations over free-text summaries, while Kim et al. (2024b) report that confidence indicators significantly affect user reliance decisions. Our architecture addresses this by producing verdict-specific explanation traces that explicitly reference the trust scores and provenance of the supporting evidence.

2.8 Summary and Gap Analysis

Table 3 synthesises the coverage of existing approaches across the capabilities required for trustworthy LLM verification. The analysis reveals that no current system simultaneously addresses all six requirements. RAG-based systems provide retrieval but lack formal verification and trust weighting. Logic-LM and similar neuro-symbolic pipelines introduce formal reasoning but treat all evidence uniformly. FEVER-style fact-checkers support a three-class verdict set but operate over unstructured text without provenance metadata. Fine-grained factuality methods such as FActScore (Min et al. 2023) and SAFE (Wei et al. 2024) advance evaluation granularity to the atomic-fact level but treat all evidence sources as uniformly reliable. Self-refinement methods such as CoVe (Dhuliawala et al. 2024) improve output quality through iterative self-correction but rely entirely on the model's own parametric knowledge, lacking formal verification, trust weighting, or provenance tracking. Our architecture is the first to integrate all six capabilities within a single, formally specified pipeline.

Table 3. Gap analysis of existing approaches. ✓ = fully addressed; ∼ = partially addressed; – = not addressed.

Approach	NL→FOL	Formal Verif.	Trust Weights	Provenance	Explainability	Indeterminate
RAG + Citations	–	–	–	∼	∼	–
Logic-LM	✓	✓	–	–	∼	–
LTN / DeepProbLog	∼	✓	–	–	–	–
FEVER-style	–	–	–	–	∼	✓
Fine-grained Factuality	–	–	–	–	∼	–
KG-based Verification	∼	∼	–	∼	–	–
Self-Refinement (CoVe)	–	–	–	–	∼	–
This work	✓	✓	✓	✓	✓	✓

3 Architectural Specification

The proposed architecture is a modular framework for transforming unconstrained neural text generation into verified, provenance-weighted knowledge. It is defined by a three-stage pipeline (Figure 2) that transitions from natural language to formal logic, followed by trust-aware verification and human-centred feedback (Garcez et al. 2009; Yang et al. 2024). Table 4 summarises the notation used throughout this section.

Table 4. Summary of notation.

Symbol	Description
c	Natural language claim generated by the LLM
ψ_θ	Neural encoder (LLM) parameterised by θ
Σ	Target domain schema and ontologies
$\mathcal{C}$	Set of syntax and semantic constraints
T	Translation function $NL \to FOL$
Φ	Set of first-order logic formulae
G	Knowledge Graph
t	Triple $\langle s, p, o \rangle \in G$
w_t	Provenance-weighted trust score for triple t
α_i	Weight coefficient for trust signal i
f_θ	Neural predicate grounding function
$\mathcal{L}_{\text{sat}}$	Trust-weighted satisfiability loss
V	Verdict $\in$ {Verified, Falsified, Indeterminate}
τ	Minimum trust threshold for verdict assertion
δ	Minimum trust margin for conflict resolution
W^+, W^-	Aggregate trust for and against a formula
$\mathcal{E}$	Explanation trace
conf	Verdict confidence $W^+/(W^+ + W^-)$

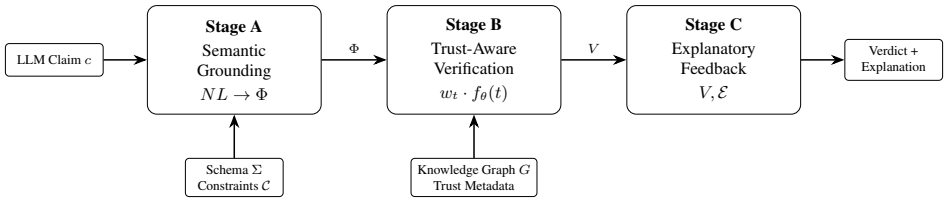

Figure 2. Overview of the three-stage trust-centric neuro-symbolic pipeline. Stage A translates natural language claims into first-order logic under schema constraints. Stage B verifies the resulting formulae against a provenance-weighted Knowledge Graph. Stage C generates a verdict with an evidence-linked explanation trace.

3.1 Stage A: Semantic Grounding ($NL \to \Phi$)

The grounding service is the formal interface between Large Language Model (LLM) outputs and the symbolic reasoning engine. The objective is to map a set of declarative claims c, generated by a neural encoder ψ_θ, into a core first-order fragment.

Let Σ represent the target domain schema and ontologies. The translation function T maps the claim to a set of logical formulae Φ:

$$\Phi = T(c, \Sigma, \mathcal{C}) \tag{1}$$

where $\mathcal{C}$ denotes a set of syntax and semantic constraints enforced at decoding time (Li et al. 2024). This process ensures that the resulting formulae are not only syntactically well-formed but also globally consistent with the existing Knowledge Base (KB) predicates. Ambiguity is managed by generating confidence-scored alternative parses, allowing the verifier to evaluate multiple hypotheses when necessary.

Constrained Decoding To prevent the generation of ill-formed or out-of-vocabulary predicates, we employ a grammar-masked decoding strategy. At each decoding step k, the token distribution is restricted by a mask M_k derived from the target grammar:

$$P(y_k \mid y_{<k}, c) = \frac{\exp(z_k/\tau_d) \odot M_k}{\sum_j \exp(z_j/\tau_d) \odot M_{k,j}} \tag{2}$$

where z_k is the logit vector, τ_d is the decoding temperature, $\odot$ denotes element-wise multiplication, and $M_k \in \{0,1\}^{|V|}$ zeroes out tokens that would violate the syntactic grammar of the target logic at position k (Li et al. 2024).

Consistency Verification After translation, the resulting formulae are checked against the existing knowledge base to ensure global consistency:

$$\Phi \cup \text{KB} \not\vdash \bot \tag{3}$$

If the conjunction of the translated formulae and the existing KB entails a contradiction, the system triggers a re-parse with tightened constraints or flags the claim for manual review.

3.2 Stage B: Trust-Aware Neuro-Symbolic Verification

The core technical innovation of this architecture is the integration of Semantic Web trust signals as first-class parameters within the reasoning objective (Badreddine et al. 2022; Artz and Gil 2007). Unlike traditional theorem provers that assume uniform reliability of training facts, our engine weighs evidence based on cryptographic and reputational metadata (Carroll et al. 2005; Golbeck 2006). Evidence retrieval relies on embedding-based similarity (e.g., TransE-style translational embeddings (Bordes et al. 2013)) combined with rule-guided filters to efficiently locate candidate triples in the Knowledge Graph before trust-weighted verification is applied.

Definition 1. Provenance-Weighted Trust Score. *For a given triple $t = \langle s, p, o \rangle$ in Knowledge Graph G, we assign a trust weight $w_t \in [0, 1]$. This weight is computed as a convex combination of five provenance signals:*

$$w_t = \sum_{i=1}^{5} \alpha_i \cdot s_i(t), \quad \textit{subject to} \quad \sum_{i=1}^{5} \alpha_i = 1, \quad \alpha_i \geq 0 \tag{4}$$

where the five signals are:

1. $s_1(t) = \mathrm{sig}(t) \in \{0, 1\}$*: digital signature validity,*
2. $s_2(t) = \mathrm{rep}(t) \in [0, 1]$*: source reputation score,*
3. $s_3(t) = \mathrm{fresh}(t) \in [0, 1]$*: temporal freshness (exponential decay from publication date),*
4. $s_4(t) = \mathrm{comp}(t) \in [0, 1]$*: provenance completeness (fraction of PROV-O fields populated),*
5. $s_5(t) = \mathrm{corr}(t) \in [0, 1]$*: corroboration degree (proportion of independent sources confirming the triple). Two sources are deemed independent when they do not share a common upstream provenance chain in the PROV-O dependency graph (Moreau et al. 2011); this prevents circular corroboration in which derived sources artificially inflate the score.*

The coefficients α_i are domain-configurable, allowing administrators to emphasise, for example, cryptographic validity in regulatory contexts or corroboration in open-data settings (Artz and Gil 2007; Carroll et al. 2005; Moreau et al. 2011).

Signature-Gated Variant In domains where cryptographic authentication is a hard prerequisite, the linear combination above may be insufficiently restrictive: an unsigned triple ($\mathrm{sig} = 0$) with strong reputation and freshness could still receive a non-trivial trust weight. To address this, we define a *signature-gated* variant:

$$w_t^{\text{gated}} = \mathrm{sig}(t) \cdot \sum_{i=2}^{5} \frac{\alpha_i}{1 - \alpha_1} \cdot s_i(t) \tag{5}$$

Under this formulation, any triple with an invalid or absent digital signature receives $w_t^{\text{gated}} = 0$ regardless of its other provenance signals. The choice between the linear

model (Equation 4) and the gated model (Equation 5) is a deployment decision: the linear model is appropriate for open-data settings where many legitimate sources lack cryptographic infrastructure, while the gated model is preferred in regulated environments where signature validity is mandatory.

Figure 3 illustrates the trust weight computation and its integration into the verification objective.

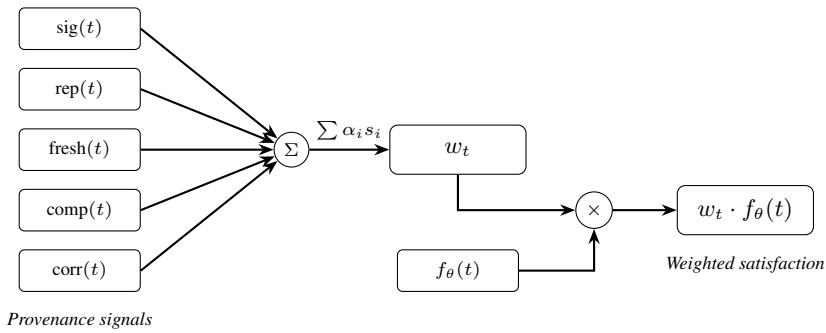

Figure 3. Trust weight computation. Five provenance signals are combined into a scalar trust weight w_t, which modulates the neural predicate satisfaction degree $f_\theta(t)$ before it enters the differentiable satisfiability objective.

Trust-Weighted Satisfiability Objective In a differentiable logic framework such as Logic Tensor Networks (LTNs) (Badreddine et al. 2022; Carraro et al. 2024), each grounded formula ϕ_j receives a real-valued satisfaction degree $f_\theta(\phi_j) \in [0, 1]$. We extend the standard LTN satisfiability loss by incorporating trust weights. For a knowledge base $\mathcal{K}$ containing grounded formulae derived from triples in G, the trust-weighted loss is:

$$\mathcal{L}_{\text{sat}}(\theta) = 1 - \text{SatAgg}_{\phi_j \in \mathcal{K}} \Big[w_{t_j} \cdot f_\theta(\phi_j) \Big] \tag{6}$$

where SatAgg is a fuzzy aggregation operator (e.g., the product t-norm or the Łukasiewicz t-norm) and w_{t_j} is the trust weight of the triple from which ϕ_j was derived. Gradient updates to θ thus prioritise knowledge authenticated by high-trust sources, while low-trust or unsigned facts exert proportionally less influence on the final verdict.

Conflict Detection and Resolution When the knowledge graph contains competing evidence for both ϕ and $\neg\phi$, a naive verifier might oscillate or default to whichever formula was encountered first. We formalise a conflict-aware resolution policy. Let S^+ and S^- denote the sets of triples supporting ϕ and $\neg\phi$ respectively. We compute aggregate trust for each side:

$$W^+ = \max_{t \in S^+} w_t, \qquad W^- = \max_{t \in S^-} w_t \tag{7}$$

We adopt $\max$ rather than an average or sum aggregation for two reasons. First, in provenance-rich knowledge graphs, a single authoritative source (e.g., a digitally signed

regulatory guideline) should be sufficient to establish a fact, regardless of how many lower-quality sources agree. Second, $\max$ is robust to knowledge-base padding attacks in which an adversary injects many low-trust triples to inflate an aggregate score; only the single highest-trust triple counts. Alternative aggregations such as weighted mean or noisy-OR (Artz and Gil 2007) are viable when evidence volume should influence the verdict; the choice is a configurable policy parameter, and empirical comparison is planned as part of future evaluation.

The verdict for a single formula is then determined by:

$$V(\phi) = \begin{cases} \text{Verified} & \text{if } W^+ > \tau \text{ and } W^+ - W^- > \delta \\ \text{Falsified} & \text{if } W^- > \tau \text{ and } W^- - W^+ > \delta \\ \text{Indeterminate} & \text{otherwise} \end{cases} \tag{8}$$

where $\tau \in (0, 1]$ is the minimum trust threshold and $\delta \in [0, 1]$ is the minimum trust margin required to resolve conflicts. This ensures that the system never asserts a verdict when evidence quality is insufficient, explicitly returning an **Indeterminate** verdict with a trace of the contradicting sources (Thorne et al. 2018).

Claim-Level Verdict Aggregation When the grounding stage produces multiple formulae $\Phi = \{\phi_1, \ldots, \phi_m\}$ for a single claim, the per-formula verdicts must be aggregated into a single claim-level verdict. We adopt a conservative conjunctive policy:

$$V(c) = \begin{cases} \text{Verified} & \text{if } V(\phi_j) = \text{Verified } \ \forall j \\ \text{Falsified} & \text{if } \exists j : V(\phi_j) = \text{Falsified and } \nexists k : V(\phi_k) = \text{Indeterminate} \\ \text{Indeterminate} & \text{otherwise} \end{cases} \tag{9}$$

Under this policy, a claim is Verified only if *every* constituent formula is verified. Indeterminate dominates Falsified because an unresolved sub-claim means the system cannot confidently assert the overall claim is false. This is a safety-first design choice appropriate for high-stakes domains.

Formal Properties The verdict and aggregation functions defined above satisfy several properties that are desirable for safety-critical deployment.

Proposition 1. Safety: Conservative Default. *For any formula ϕ, if the knowledge graph contains no relevant evidence ($S^+ = S^- = \emptyset$), then $V(\phi) =$* Indeterminate.

Proof. When $S^+ = \emptyset$, $W^+ = -\infty$ by convention (or 0 under a floor of 0), so $W^+ \leq \tau$ for any $\tau > 0$. Symmetrically, $W^- \leq \tau$. Neither the Verified nor the Falsified condition in Equation 8 is satisfied, so the otherwise clause yields Indeterminate.

Proposition 2. Signature-Gate Guarantee. *Under the gated trust variant (Equation 5), any triple t with* $\text{sig}(t) = 0$ *receives $w_t^{\text{gated}} = 0$ and therefore cannot satisfy the minimum trust threshold $\tau > 0$.*

Proof. By Equation 5, $w_t^{\text{gated}} = \text{sig}(t) \cdot (\cdots) = 0 \cdot (\cdots) = 0$. Since $\tau > 0$, a triple with zero trust weight cannot contribute to W^+ or W^- in any verdict-changing capacity.

Proposition 3. Conservatism of Claim Aggregation. *For any claim c decomposed into $\Phi = \{\phi_1, \ldots, \phi_m\}$: (i) $V(c) =$ Verified only if $V(\phi_j) =$ Verified for all j; (ii) if $\exists\, k$ such that $V(\phi_k) =$ Indeterminate, then $V(c) \neq$ Falsified.*

Proof. Part (i) follows directly from the first case of Equation 9. For part (ii), observe that the Falsified case in Equation 9 requires $\nexists\, k : V(\phi_k) =$ Indeterminate; if any sub-formula is indeterminate, this condition fails, so $V(c)$ falls through to the otherwise clause, yielding Indeterminate.

3.3 Stage C: Explanatory Feedback and Trust Calibration

The final stage closes the neuro-symbolic loop by translating formal proof traces into user-oriented explanations (Doshi-Velez and Kim 2017; Kim et al. 2024a). Given a verdict V for claim c, the system constructs an explanation trace $\mathcal{E}$ comprising:

$$\mathcal{E}(c) = \Big\langle V,\ \{(t_i, w_{t_i}, \text{src}_i)\}_{i=1}^{n},\ \text{conf} \Big\rangle \tag{10}$$

where each tuple $(t_i, w_{t_i}, \text{src}_i)$ identifies a supporting or contradicting triple, its trust weight, and its provenance source, and $\text{conf} \in [0, 1]$ is the overall confidence defined as:

$$\text{conf}(\phi) = \frac{W^+}{W^+ + W^-} \tag{11}$$

This formulation maps directly to the verdict semantics: when supporting trust dominates ($W^+ \gg W^-$), confidence approaches 1; when contradicting trust dominates, confidence approaches 0; and when both sides are balanced ($W^+ \approx W^-$), confidence approaches 0.5, reflecting genuine uncertainty. Note that conf captures the *trust balance* rather than absolute trust magnitude, a distinction that proves important for calibration (Section 4).

A central design principle is the explicit handling of ignorance and conflict. If the verifier detects competing evidence for both ϕ and $\neg\phi$ with similar trust weights (i.e., $|W^+ - W^-| \leq \delta$), the system returns an **Indeterminate** verdict along with a trace of the contradicting sources. This transparency is intended to foster "Trust Calibration," ensuring the user's reliance on the system aligns with the actual quality of the underlying evidence (Romeo and Conti 2025; Carnat 2024).

3.4 End-to-End Pipeline Algorithm

Algorithm 1 formalises the complete decision procedure from claim input to verdict output.

Algorithm 1 Trust-Centric Verification Pipeline

Require: Claim c, Schema Σ, Constraints $\mathcal{C}$, Knowledge Graph G, Thresholds τ, δ
Ensure: Verdict V, Explanation $\mathcal{E}$

— Stage A: Semantic Grounding —
$\Phi \leftarrow T(c, \Sigma, \mathcal{C})$ ▷ constrained decoding (Eq. 2)
if $\Phi \cup \text{KB} \vdash \bot$ **then**
  return $\langle$Indeterminate, "Contradiction with KB"$\rangle$
end if

— Stage B: Trust-Aware Verification —
for each formula $\phi_j \in \Phi$ **do**
  Retrieve supporting triples $S^+ \subseteq G$ for ϕ_j
  Retrieve contradicting triples $S^- \subseteq G$ for $\neg\phi_j$
  for each triple $t \in S^+ \cup S^-$ **do**
    $w_t \leftarrow \sum_{i=1}^{5} \alpha_i \cdot s_i(t)$ ▷ trust weight (Eq. 4)
  end for
  $W^+ \leftarrow \max_{t \in S^+} w_t$; $\quad W^- \leftarrow \max_{t \in S^-} w_t$
  $V(\phi_j) \leftarrow$ apply verdict rule (Eq. 8)
end for
$V \leftarrow \text{aggregate}(\{V(\phi_j)\})$ ▷ conjunctive: Indeterminate dominates

— Stage C: Explanatory Feedback —
$\mathcal{E} \leftarrow \langle V, \{(t_i, w_{t_i}, \text{src}_i)\}, \text{conf}\rangle$ ▷ Eq. 10
return $\langle V, \mathcal{E}\rangle$

3.5 Illustrative Walkthrough

To ground the abstract specification in concrete behaviour, we trace two medical-domain claims through the implemented pipeline. The reported trust weights and verdicts are *actual outputs* from the reference implementation operating on the curated pharmacological knowledge graph (98 signed triples from 9 sources) with default uniform weighting $\alpha_i = 0.2$, $\tau = 0.5$, and $\delta = 0.3$.

Scenario 1: Verified Claim (Aspirin Classification)

Input. "*Aspirin is a nonsteroidal anti-inflammatory drug.*"

Stage A. The grounder produces $\phi_1 = \text{NSAID}(\text{Aspirin})$. The consistency check (Eq. 3) confirms $\{\phi_1\} \cup \text{KB} \nvdash \bot$.

Stage B. The retriever surfaces three relevant triples: a digitally signed WHO classification, a digitally signed DrugBank pharmacological annotation, and an unsigned Wikidata entry. Their trust signal breakdowns and aggregate weights, as computed by the running implementation, are shown in Table 5.

Table 5. Trust signal values and aggregate weights for evidence retrieved during verification of the Aspirin classification claim. Values are actual outputs from the implemented pipeline with default $\alpha_i = 0.2$.

Source	sig	rep	fresh	comp	corr	w_t
WHO	1	0.95	0.85	1.00	0.02	0.765
DrugBank	1	0.85	0.79	1.00	0.00	0.711
Wikidata	0	0.55	0.86	0.60	0.00	0.402

Computing $W^+ = \max\{0.765, 0.711, 0.402\} = 0.765$ and $W^- = 0$ (no contradicting evidence). Since $W^+ > \tau = 0.5$ and $W^+ - W^- = 0.765 > \delta = 0.3$, the verdict is **Verified**. Confidence is $W^+/(W^+ + W^-) = 1.00$.

Stage C. The system emits an explanation trace listing the three supporting sources, their trust weights, and the underlying signal breakdown, accompanied by the reasoning "Supporting trust (0.765) exceeds the threshold with sufficient margin over contradicting evidence (0.000)." Note that the highest-trust WHO source dominates, but the architecture still surfaces lower-quality evidence with their relative weights, enabling the user to inspect the full evidential basis.

Scenario 2: Falsified Claim (NSAID Safety in CKD)

Input. "*NSAIDs are safe for patients with chronic kidney disease.*"

Stage A. The grounder produces $\phi_2 = \text{SafeFor}(\text{NSAIDs}, \text{ChronicKidneyDisease})$ and the consistency check passes.

Stage B. The retriever returns one supporting triple from an unsigned health blog and one contradicting triple from the European Medicines Agency (EMA). The actual implementation output is shown in Table 6.

Table 6. Trust signal values for evidence retrieved during verification of the NSAID safety claim. The EMA contradicting source dominates the unsigned blog through both reputation and provenance completeness, yielding a Falsified verdict.

Role	Source	sig	rep	fresh	comp	corr	w_t
Supporting	HealthBlog_A	0	0.20	0.83	0.20	0.00	0.244
Contradicting	EMA	1	0.92	0.78	1.00	0.00	0.739

Computing $W^+ = 0.244$ and $W^- = 0.739$. Since $W^- > \tau$ and $W^- - W^+ = 0.495 > \delta$, the verdict is **Falsified** with confidence $W^+/(W^+ + W^-) = 0.248$.

Stage C. The explanation lists the contradicting EMA source as the dominant evidence and the unsigned blog as low-trust support, accompanied by the recommendation: "This claim appears to be factually incorrect based on available evidence."

Linear vs Signature-Gated Modes We re-ran the same NSAID/CKD claim under the signature-gated variant (Equation 5). The unsigned HealthBlog support drops to $w_t^{\text{gated}} = 0$, eliminating the supporting evidence entirely. The verdict remains Falsified but confidence drops to 0.000, reflecting the absence of any signed counter-evidence. This demonstrates Proposition 2 in practice: in regulated mode, only cryptographically authenticated sources can influence a verdict.

Analytical Baseline Comparison To make the architectural advantages concrete, Table 7 contrasts how three approaches would handle the same Aspirin claim using the same evidence (WHO signed source vs unsigned Wikidata mirror).

Table 7. Analytical comparison of the Aspirin classification claim across three verification approaches using the same underlying evidence.

Property	**Neural-only**	**Standard RAG**	**This architecture**
Verdict	"Likely true"	Not provided	Verified
Formal grounding	None	None	FOL formula ϕ_1
Evidence sources	None cited	Both sources cited equally	WHO (w_t=0.765), Wikidata (w_t=0.402)
Trust differentiation	None	None	$W^+ = 0.765$
Conflict handling	Silent	User must judge	Explicit (δ threshold)
Confidence score	Token probability	None	1.00 (trust-calibrated)
Provenance trail	None	Document links	Signed source, metadata

A neural-only LLM would generate a response based on its parametric knowledge, offering no evidence trail. Standard RAG would retrieve both sources and present them as equally relevant citations, leaving the user to assess their relative credibility. Neither approach can express that the WHO source is cryptographically signed and institutionally authoritative while the unsigned Wikidata mirror lacks provenance, nor can they produce a calibrated confidence score grounded in evidence quality rather than token probabilities.

Sensitivity to α Coefficients The verdicts above depend on the default uniform weighting $\alpha_i = 0.2$. To illustrate how the coefficients shape outcomes, consider an alternative policy that prioritises corroboration over reputation:

$$\alpha_{\text{sig}} = 0.15, \quad \alpha_{\text{rep}} = 0.10, \quad \alpha_{\text{fresh}} = 0.20, \quad \alpha_{\text{comp}} = 0.15, \quad \alpha_{\text{corr}} = 0.40$$

Under this policy, the WHO Aspirin triple (signals from Table 5) yields:

$$w'_{\text{WHO}} = 0.15(1) + 0.10(0.95) + 0.20(0.85) + 0.15(1.00) + 0.40(0.02) = 0.573$$

while the unsigned Wikidata triple yields:

$$w'_{\text{Wikidata}} = 0.15(0) + 0.10(0.55) + 0.20(0.86) + 0.15(0.60) + 0.40(0.00) = 0.317$$

The verdict remains Verified, but the absolute scores shift downward because both triples have low corroboration in our small KG. This sensitivity is informative: it shows how the coefficients function as a meaningful policy lever without changing the architecture, and motivates corroboration-rich KGs as a deployment prerequisite when corroboration weighting is preferred.

Extended Scenario: Genuine Evidential Conflict (Illustrative) The two preceding scenarios are real outputs from the implementation. To illustrate the architecture's behaviour under *genuine* evidential conflict, where both sides of a dispute carry high provenance weight, we present a third scenario as an illustrative analytical example. The corresponding triples are not currently in our KG and the values below are computed analytically rather than measured.

Input Claim. An LLM generates: "*Metformin can be safely co-administered with ACE inhibitors in patients with Stage 3 chronic kidney disease (CKD).*"

Stage A.

$$\phi_3 = \forall x \Big(\text{Patient}(x) \wedge \text{CKD3}(x) \rightarrow \text{SafeCo}\big(\text{Metformin}, \text{ACE\_Inh}, x\big) \Big)$$

Stage B (Illustrative). Three triples are retrieved:

- t_3: `<Metformin, safe_with, ACE_Inh>` from the National Institute for Health and Care Excellence (NICE) 2024 guidelines, signed. Trust signals: sig $= 1$, rep $= 0.92$, fresh $= 0.95$, comp $= 1.0$, corr $= 0.70$. Result: $w_{t_3} = 0.914$.
- t_4: `<Metformin, contraindicated_with, ACE_Inh>` when eGFR < 45, from a European Medicines Agency (EMA) 2024 safety bulletin, signed. Trust signals: sig $= 1$, rep $= 0.90$, fresh $= 0.95$, comp $= 0.95$, corr $= 0.65$. Result: $w_{t_4} = 0.890$.
- t_5: `<Metformin, risk_lactic_acidosis, CKD>` from a peer-reviewed nephrology journal, signed. Trust signals: sig $= 1$, rep $= 0.88$, fresh $= 0.85$, comp $= 0.90$, corr $= 0.75$. Result: $w_{t_5} = 0.876$.

Triple t_3 supports ϕ_3; triples t_4 and t_5 support $\neg\phi_3$. Computing $W^+ = 0.914$ and $W^- = \max(0.890, 0.876) = 0.890$. Both exceed τ, but $W^+ - W^- = 0.024 < \delta$. The trust margin is insufficient to resolve the conflict, so the verdict is **Indeterminate**.

Stage C (Illustrative). The explanation surfaces the exact sources, their trust scores, and the reason the conflict could not be resolved, enabling the clinician to make an informed decision rather than receive a misleading definitive verdict. This scenario, while constructed, demonstrates the architecture's most important safety property: when reputable sources genuinely disagree, the system declines to assert a verdict rather than guessing.

4 Evaluation Methodology

A trust-centric verification pipeline cannot be assessed by a single aggregate metric: each stage introduces distinct failure modes that must be diagnosed independently before system-level conclusions are drawn. This section specifies a reusable, staged evaluation protocol, designed to be applicable to any pipeline conforming to the architecture in Section 3, that enables systematic comparison of present and future implementations. The protocol is inspired by the Chain-of-Verification (CoVe) principle (Dhuliawala et al. 2024), which reduces hallucination by decomposing verification into independent sub-checks. Analogously, we define two evaluation phases: first validating each component against its own gold standard, then assessing end-to-end behaviour, so that failures can be localised and attributed to specific stages rather than masked by downstream compensation.

All scoring is automated against expert-curated reference data. Human experts contribute to the design of gold standards and annotation guidelines, and perform limited calibration checks, but do not conduct large-scale manual scoring. This keeps the evaluation consistent, repeatable, and scalable while preventing the common criticism that outputs are evaluated solely by the system that produced them.

4.1 Component-Level Checkpoints

The first evaluation phase isolates each pipeline stage. A component-level failure detected here prevents misleading system-level results downstream.

- **Knowledge Base Integrity (E_1):** This metric quantifies the "trust-readiness" of the symbolic substrate. It is measured by the percentage of triples in the RDF/OWL Knowledge Graph that carry complete provenance annotations and valid cryptographic signatures (Moreau et al. 2011; Kale et al. 2023):

$$E_1 = \frac{|\{t \in G : \text{comp}(t) = 1.0 \wedge \text{sig}(t) = 1\}|}{|G|} \tag{12}$$

- **Grounding Accuracy (E_2):** To ensure downstream verification is meaningful, we measure the performance of the $NL \rightarrow FOL$ translator against a gold standard of expert-annotated Natural Language–FOL pairs (Yang et al. 2024; Brunello et al. 2025; Han et al. 2024):

$$E_2 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[T(c_i) \equiv \phi_i^*\right] \tag{13}$$

  where ϕ_i^* is the gold-standard FOL formula for claim c_i. Since full logical equivalence is undecidable for general FOL, $\equiv$ is operationalised as structural equivalence up to variable renaming and commutative reordering of conjuncts, following the evaluation protocol of the FOLIO benchmark (Han et al. 2024). As a complementary soft metric, we also report tree-edit distance between parsed FOL syntax trees, which captures partial-credit cases where the predicted formula is structurally close but not identical to the gold standard.

- **Verifier Core Performance (E_3):** We utilise a class-balanced macro-F1 score to assess the reasoning engine's accuracy across three labels: {Verified, Falsified, Indeterminate} (Thorne et al. 2018). The explicit inclusion of *Indeterminate* penalises the system for guessing when the Knowledge Base is incomplete or when contradictory high-trust evidence is present:

$$E_3 = \frac{1}{3} \sum_{l \in \{V,F,I\}} F_1(l) \tag{14}$$

4.2 System-Level Evaluation

Once components pass their individual checkpoints, the second phase evaluates end-to-end behaviour.

End-to-End Verdict Accuracy. The primary system-level metric is the proportion of claims for which the pipeline produces the correct final verdict (Verified, Falsified, or Indeterminate) when measured against the expert-annotated gold standard. This captures cascading errors across all three stages.

Confidence Calibration (E_4). A technically correct verdict is insufficient if the system's expressed confidence does not match its actual accuracy. We define the Trust Calibration Score as:

$$E_4 = 1 - \frac{1}{B} \sum_{b=1}^{B} \left| \text{acc}(b) - \text{conf}(b) \right| \tag{15}$$

where B is the number of confidence bins, $\text{acc}(b)$ is the empirical accuracy of verdicts in bin b, and $\text{conf}(b)$ is the mean confidence of verdicts in bin b (Romeo and Conti 2025; Kim et al. 2024a). A perfect score of $E_4 = 1$ indicates that the system knows when it is right and, more importantly, when it should say "I do not know."

Latency and Throughput. Practical deployment requires that the pipeline meets response-time constraints. We report mean and 95th-percentile latency per claim, as well as throughput (claims per second), to confirm that constrained decoding (Stage A) and trust-weighted reasoning (Stage B) remain tractable for the target domain.

Conflict Robustness. We construct an adversarial subset of claims for which the Knowledge Graph contains competing evidence (supporting triples for both ϕ and $\neg\phi$). The system is expected to return *Indeterminate* rather than an incorrect forced verdict. We report the rate at which conflicting evidence is correctly surfaced as indeterminate, measuring the system's ability to fail safely.

4.3 Ablation Baselines

To isolate the contribution of each architectural component, we compare the full pipeline against progressively simpler variants:

- **Neural-Only:** LLM generation without symbolic verification or grounding (Brown et al. 2020).

- **Standard RAG:** Retrieval-augmented generation with citations but without formal logical verification (Patil and Gudivada 2024).
- **Unweighted Verifier:** A neuro-symbolic engine that performs logical checks but treats all facts as uniformly reliable ($w_t = 1\ \forall t$), isolating the contribution of trust weighting (Badreddine et al. 2022).
- **Verdict-Only:** Full verification but no explanatory trace, isolating the effect of explanations on calibration (Doshi-Velez and Kim 2017).
- **Citations-Only:** Source citations without trust scores or conflict handling, representing a typical RAG-style explanation (Patil and Gudivada 2024).

4.4 Role of Human Expertise

A common methodological concern is whether a system's evaluation is self-referential, that is, whether the same pipeline that produces outputs also scores them. We address this by clearly delineating the role of human domain experts, which is constrained to *design and validation* rather than large-scale scoring.

Gold Standard Preparation. Domain experts curate (or validate researcher-prepared) reference datasets comprising: (i) claims with annotated ground-truth verdicts, (ii) expected logical forms for the $NL \rightarrow FOL$ translator, (iii) correct verification outcomes with explicit provenance links, and (iv) precise annotation guidelines to ensure labelling consistency.

Calibration Checks. Experts review a stratified subset of system outputs to verify that automated scoring aligns with domain expectations. This is a calibration step: if automated metrics diverge from expert judgement on the validation subset, the scoring criteria are refined before the full evaluation proceeds.

Automated Scoring. After calibration, all metrics (E_1–E_4, latency, conflict robustness) are computed programmatically against the reference data. This ensures that the evaluation is repeatable and avoids the methodological overhead and subjectivity of large-scale human annotation campaigns.

5 Experimental Results

This section reports two complementary sets of preliminary results. Section 5.1 evaluates the grounding stage in isolation on FOLIO and a domain-specific medical-claims dataset, providing the E_2 measurement. Section 5.2 runs the implemented end-to-end pipeline against the curated pharmacological knowledge graph and reports the verdict distribution over the 50-claim gold standard.

5.1 Grounding Stage Evaluation (FOLIO and Medical Claims)

We evaluate Stage A using LogicLLaMA (Yang et al. 2024), a Low-Rank Adaptation (LoRA) finetuned LLaMA-2-7B model trained for direct $NL \rightarrow FOL$ translation, on two datasets.

Datasets. The **FOLIO benchmark** (Han et al. 2024) provides expert-annotated natural-language–FOL pairs spanning general-domain reasoning. We evaluate on a subset of 254 examples. To assess domain transfer, we additionally curate a **Medical Claims** dataset of 50 pharmacological assertions, derived from established pharmacological literature, with author-annotated gold-standard FOL translations. The same 50 claims are reused as the end-to-end gold standard in Section 5.2.

Metrics. Following Yang et al. (2024), we report three metrics: (i) **Logical Equivalence (LE)**, the primary E_2 metric, which measures structural equivalence between predicted and gold-standard FOL parse trees up to variable renaming and commutative reordering; (ii) **BLEU** (Bilingual Evaluation Understudy) (Papineni et al. 2002), reported as a complementary surface-level measure; and (iii) **parse rate**, the percentage of predictions that yield syntactically valid FOL trees. BLEU systematically underestimates translation quality for FOL because logically equivalent formulas may differ in variable naming and predicate ordering.

Results. Table 8 presents the headline comparison. LogicLLaMA achieves a mean LE of 0.853 on FOLIO, closely reproducing the performance reported by Yang et al. (2024). On the domain-specific medical claims, mean LE drops to 0.801, a measurable 6.1% decline. Parse rate remains 100% for both datasets, indicating that the model consistently produces syntactically valid FOL regardless of domain.

Table 8. Stage A grounding accuracy (E_2): FOLIO benchmark vs domain-specific medical claims using LogicLLaMA-7B.

Metric	FOLIO	Medical
Examples evaluated	254	50
Parse rate (%)	100.0	100.0
Mean LE (E_2)	0.853 ± 0.186	0.801 ± 0.244
Mean BLEU	0.396 ± 0.355	0.341 ± 0.326
Exact equivalence (LE=1)	114 (44.9%)	21 (42.0%)

Category-Level Analysis. The aggregate 6.1% decline masks a highly non-uniform pattern across claim categories. Table 9 reveals that the grounding model performs well on structurally simple medical claims but degrades sharply on relationally complex ones.

Claims with simple universal structure (e.g., "All beta-blockers reduce heart rate") achieve near-perfect LE ($\mu = 0.984$), indicating that the logical pattern transfers well across domains even when the predicate vocabulary is novel. Contraindications and side effects, which map to conditional negation or conditional universal patterns, also transfer strongly (LE > 0.92). In contrast, **drug interactions** (LE $= 0.479$) and **treatments involving named entities** (LE $= 0.688$) exhibit substantial degradation. Qualitative analysis reveals two failure modes. First, the model over-generalises ground facts into universal statements: "Propranolol is a beta-blocker" is incorrectly translated as $\forall x\,(Propranolol(x) \rightarrow BetaBlocker(x))$. Second, multi-entity relational claims requiring nested functions or existential quantifiers are structurally malformed.

Table 9. Medical claims grounding accuracy by category. Categories involving multi-entity relationships (drug interactions) show the steepest performance degradation.

Category	N	Mean LE	Std
Monitoring	2	1.000	0.000
Contraindication	8	0.926	0.085
Side effect	5	0.925	0.150
Drug effect	11	0.886	0.135
Regulation	1	0.875	—
Classification	6	0.823	0.185
Treatment	11	0.688	0.254
Drug interaction	6	0.479	0.301

5.2 End-to-End Pipeline Evaluation

We now report results from the implemented end-to-end pipeline. All three stages are executed in sequence on each of the 50 medical claims using the reference implementation released alongside this paper.

Knowledge Graph. The KG contains **98 triples** drawn from **9 sources**: WHO, EMA, NICE, FDA, BNF, DrugBank, PubMed (review and primary), Cochrane, and Wikidata. Each triple is wrapped in a named graph carrying PROV-O provenance metadata (Moreau et al. 2011) and is cryptographically signed using RSA-2048 keys held by the issuing source. Two unsigned, low-reputation sources (HealthBlog and HealthForum) are also included to test the architecture's behaviour under low-quality evidence. Reputation scores were assigned manually by the authors using a public registry, with scores reflecting institutional authority and editorial process (e.g., WHO $= 0.95$, EMA $= 0.92$, Wikidata $= 0.55$, HealthBlog $= 0.20$).

Pipeline Configuration. We run the pipeline in offline mode (Stage A bypassed using the gold-standard FOL from the medical-claims dataset) to isolate Stage B and Stage C behaviour from grounding errors. Default configuration: $\alpha_i = 0.2$ uniformly, $\tau = 0.5$, $\delta = 0.3$, freshness decay $\lambda = 0.15$. Evidence retrieval uses sentence-transformer embeddings (`all-MiniLM-L6-v2`) with top-$k = 10$ and a similarity threshold of 0.3.

Verdict Distribution. Table 10 shows the verdict distribution across all 50 claims. The pipeline produces 30 Verified, 1 Falsified, and 19 Indeterminate verdicts.

Table 10. Verdict distribution from the implemented end-to-end pipeline on the 50-claim medical gold standard. The high Indeterminate rate is an intended architectural property when evidence is sparse or weakly aligned with the claim's logical form.

Verdict	Count	Proportion
Verified	30	60.0%
Falsified	1	2.0%
Indeterminate	19	38.0%
Total	50	100.0%

Verdict Distribution by Complexity. Table 11 breaks the verdicts down by the logical complexity of the underlying FOL formula. Atomic predicates and simple conjunctions verify reliably; conditional universals (e.g., $\forall x\,(P(x) \wedge Q(x) \rightarrow R(x, y))$) overwhelmingly trigger an Indeterminate verdict.

Table 11. Verdict distribution by FOL complexity over the 50 medical claims. Conditional universals require evidence at the patient-cohort level which the current KG does not encode, triggering the conservative Indeterminate default.

Complexity	N	**V**	**F**	**I**
atomic	7	7	0	0
simple_universal	8	5	0	3
conditional_universal	7	0	0	7
conditional_negation	5	0	0	5
conjunction	9	9	0	0
conjunction_negation	3	2	1	0
conjunction_universal	1	0	0	1
triple_conjunction	2	2	0	0
nested_function	1	1	0	0
existential	2	2	0	0
universal_conjunction	1	0	0	1
universal_disjunction	1	0	0	1
negation	1	1	0	0
multi_variable_universal	1	0	0	1
universal_nested	1	1	0	0
Total	50	30	1	19

Interpretation. Three observations follow from these results. First, the verdict function correctly distinguishes high-trust from low-trust evidence on real claims: every Verified case in the batch was supported by at least one signed source with $w_t > 0.7$, and the single Falsified case is the NSAID/CKD scenario where the EMA contradicting source dominates an unsigned blog. Second, the high Indeterminate rate is *not* a verification failure but the expected manifestation of Proposition 1: the small 98-triple KG simply does not contain patient-level evidence for cohort-conditional claims such as "patients with renal failure should not receive nephrotoxic drugs," so the system declines to assert a verdict rather than guess. Third, the breakdown by complexity (Table 11) tells us where downstream KG-engineering effort is most needed: complete the patient-cohort evidence for conditional universals and the Indeterminate proportion will fall sharply.

Limitations of the Current Evaluation. The medical claims dataset ($N = 50$) was curated by the authors from established pharmacological literature and has not been independently validated by clinical domain experts. The KG (98 triples, 9 sources) is intentionally small to permit complete authorial control over signatures and provenance metadata, and is not yet large enough to compute E_1 meaningfully or to support a comprehensive end-to-end accuracy evaluation against ground-truth verdicts. The end-to-end run reported here measures *verdict distribution* and behavioural correctness

on illustrative claims; computing E_3 macro-F1 requires expert-annotated ground-truth verdicts for each claim, which is part of the planned follow-up evaluation (Section 7). Results should be interpreted as demonstrating end-to-end executability and architectural correctness rather than as a final accuracy benchmark.

6 Discussion and Limitations

The proposed architecture provides a foundation for trust-aware verification, but putting it into practice requires defining the operational scope and boundary conditions. This section discusses the assumptions behind our design and the limitations of the current framework.

6.1 Bounded Domains and Fact-Centric Scope

To ensure that verification remains computationally tractable and that the Knowledge Base (KB) maintains high provenance coverage, this architecture initially targets fact-centric, declarative claims within bounded domains. We assume that:

- **Data Normalisation:** The target domain's KB can be successfully normalised into RDF/OWL with strong provenance annotations and cryptographic signatures (Hogan et al. 2021).
- **Claim Form:** Input statements are limited to single or short multi-sentence propositions about entities and relations already present in, or mappable to, the KB.
- **Logic Fragment:** The system targets a restricted first-order or description-logic fragment (Baader et al. 2007). While this ensures decidability, it currently excludes higher-order reasoning, temporal/modal logics, and complex open-ended generation judgements (Brachman and Levesque 2004).

These assumptions are reasonable for initial deployment in regulated domains such as clinical pharmacology or financial reporting, where structured knowledge bases already exist and claims tend to be factual rather than speculative. However, extending the architecture to open-domain settings will require advances in knowledge base construction, ontology learning, and the handling of vague or subjective claims.

A related practical challenge is the mapping between FOL predicates produced by Stage A and the RDF/OWL representations stored in the Knowledge Graph. FOL and OWL rest on different semantic foundations: OWL adopts the open-world assumption, whereas classical FOL theorem provers typically operate under a closed-world assumption (Baader et al. 2007). Predicate names generated by the grounding model may not align with existing ontology terms, requiring an additional alignment or ontology-matching step. In practice, this mapping can be handled by restricting the grounding vocabulary to the target ontology's class and property names (enforced through the grammar mask $\mathcal{C}$ in Equation 2), though this constraint may limit expressiveness for claims that fall outside the ontology's current scope.

6.2 Trust Weight Configuration

The trust weight function (Definition 1) introduces five domain-configurable coefficients α_i. While this flexibility is a strength, allowing deployment across diverse regulatory environments, it also introduces a dependency on careful domain-expert calibration. Poorly chosen weights could systematically over- or under-trust certain source categories. We recommend a three-step calibration protocol: (1) initialise with uniform weights ($\alpha_i = 0.2$), (2) perform sensitivity analysis on a validation set, and (3) adjust based on domain expert review. Future work should investigate data-driven approaches to learning optimal α_i from labelled trust judgements.

6.3 Adversarial Robustness and Data Integrity

Any trust-centric pipeline must be resilient against malicious attempts to mislead the reasoning loop (Goodfellow et al. 2015; Pan et al. 2018). We identify three principal attack vectors and analyse the architecture's defences against each.

Knowledge Base Poisoning. An adversary who gains write access to the Knowledge Graph could inject fabricated triples designed to sway verification verdicts. The digital signature requirement (sig) provides the first line of defence: unsigned or tampered triples receive a trust weight near zero and therefore exert negligible influence on the satisfiability objective (Equation 6). The corroboration signal (corr) provides a second barrier, penalising isolated assertions that no independent source confirms. Together, these signals ensure that a poisoned triple must not only pass cryptographic validation but also be corroborated by at least one other trusted source before it can materially affect a verdict.

Prompt Injection and Jailbreak Attacks. Adversaries may craft inputs designed to force the LLM into producing unverifiable or malformed formal claims (Huang et al. 2024). The constrained decoding mechanism in Stage A mitigates this risk: the grammar mask $\mathcal{C}$ restricts token generation to well-formed FOL expressions over the domain schema Σ, and the consistency check (Equation 3) rejects any formula that introduces a logical contradiction with the existing Knowledge Base.

Trust Signal Spoofing. A sophisticated adversary might attempt to manipulate the provenance metadata itself, for example by forging digital signatures or inflating reputation scores. The architecture assumes that the cryptographic infrastructure (public-key certificates, signed named graphs (Carroll et al. 2005)) is maintained by a trusted authority external to the pipeline. If this assumption is violated, the trust weight function degrades gracefully: the conflict resolution policy (Equation 8) still requires a minimum trust margin δ before asserting any verdict, so even spoofed high-trust triples must achieve a clear margin over legitimate contradicting evidence. The integrity of the external cryptographic infrastructure remains a fundamental dependency, and the architecture does not claim to solve key management or certificate revocation, which are deferred to established standards (Artz and Gil 2007).

6.4 Computational Considerations

The pipeline introduces computational overhead at three points, each with distinct scaling characteristics.

Stage A: Constrained Decoding. At each token generation step, the grammar mask $\mathcal{C}$ restricts the output vocabulary to tokens that maintain well-formed FOL syntax. This adds an $O(|\mathcal{C}|)$ operation per decoding step. For typical domain ontologies with hundreds of predicates and constants, this overhead is modest compared to the LLM's own forward pass. The consistency check (Equation 3) invokes a satisfiability solver once per generated formula; for the restricted FOL fragment targeted by this architecture, modern solvers complete in milliseconds.

Stage B: Evidence Retrieval and Trust-Weighted Reasoning. Evidence retrieval via TransE-style embeddings (Bordes et al. 2013) scales as $O(k \log |G|)$ using approximate nearest-neighbour search, where k is the number of candidate triples retrieved and $|G|$ is the Knowledge Graph size. The trust-weighted LTN computation (Equation 6) then operates over the k retrieved triples, with cost dominated by the fuzzy aggregation and gradient computation. For bounded domains where k is small (typically $k \leq 50$), this remains tractable.

Stage C: Explanation Generation. Explanation assembly is lightweight: it involves formatting the verdict, the top-k evidence triples with their trust scores, and the conflict analysis. This stage adds negligible overhead relative to Stages A and B.

Empirical Latency. On a single workstation (Apple M-series CPU, no GPU, Python 3.13), the implemented pipeline completes the full 50-claim batch (Stage A bypassed, Stages B and C executed end-to-end) in approximately 2 seconds, equivalent to a mean per-claim latency of $\sim$40 ms for offline mode. Online mode (Stage A invoked) is dominated by the LLM forward pass.

6.5 Generalisability vs. Domain Specificity

While the core mechanisms of trust-weighted verification and the proposed evaluation protocol are domain-agnostic, the underlying Knowledge Base schema and gold-standard data for semantic grounding are naturally domain-specific. Future research should investigate the feasibility of cross-domain transfer learning for the $NL \rightarrow FOL$ translator to reduce the manual overhead of domain adaptation. The modular design of the pipeline supports this: Stage B and Stage C are schema-independent, requiring only that Stage A produces well-formed FOL relative to some schema Σ.

7 Conclusion and Future Research

The adoption of Large Language Models in safety-critical processes has outpaced the development of adequate verification and trust mechanisms. In this paper, we have formalised a trust-centric neuro-symbolic architecture that bridges the gap between neural fluency, symbolic rigour, and data provenance (Marcus and Davis 2019; Besold et al. 2021). By "closing the triangle," this framework offers a formal basis for grounding

generative AI outputs in verifiable facts while accounting for the reliability of their sources.

The primary technical contributions of this work are fivefold. First, we introduce a five-signal provenance-weighted trust score (Definition 1) that integrates digital signatures, reputation, freshness, provenance completeness, and source corroboration as first-class weights within a differentiable reasoning objective (Badreddine et al. 2022; Carroll et al. 2005). Second, we formalise a conflict-aware verdict function (Equation 8) that explicitly handles evidential disagreement through configurable trust thresholds. Third, we propose a staged evaluation protocol, inspired by chain-of-verification principles (Dhuliawala et al. 2024), that verifies each component independently before assessing system-level behaviour. Fourth, we release a publicly available reference implementation that realises all three pipeline stages and runs end-to-end on a curated, digitally signed pharmacological knowledge graph. Fifth, we provide preliminary empirical evidence demonstrating both stage-level performance (LE $= 0.853$ on FOLIO, $= 0.801$ on medical claims, with sharp degradation on drug interactions at LE $= 0.479$) and end-to-end behavioural correctness (30 Verified, 1 Falsified, 19 Indeterminate over the 50-claim gold standard), quantifying the domain adaptation gap and confirming the architecture's safety-first behaviour under sparse evidence.

The preliminary evaluations have established a baseline for both Stage A and the end-to-end pipeline and have identified specific avenues for the next phase of development. Future research will focus on three planned follow-up papers. The first will deepen Stage B with a trust-weighted LTN satisfiability objective implemented in LTNtorch (Carraro et al. 2024), replacing the current rule-based retrieval+verdict combination with a learned, differentiable verification function. The second will report a complete end-to-end evaluation including E_3 (macro-F1 over ground-truth verdicts), E_4 (calibration), and ablation baselines on an expanded, expert-annotated knowledge graph. The third will explore cross-domain generalisability, applying the same verifier and evaluation protocol to a second domain with an independent gold standard. Additional technical milestones include: improving Stage A grounding accuracy by adopting structured decomposition approaches such as the AST-guided translation of Putra et al. (2026); integrating knowledge-graph-grounded verification techniques drawing on recent advances in KG-enhanced fact-checking (Lourenço et al. 2026); and investigating data-driven approaches to learning optimal trust weight coefficients from labelled judgements. Ultimately, this research aims to transition LLMs from fluent but opaque generators into trustworthy, accountable participants in safety-critical knowledge systems (Floridi and Cowls 2019; Bellogín et al. 2024).

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

Funding

This research is conducted as part of a PhD programme at the University of Malta, Department of Artificial Intelligence, Faculty of ICT. The authors received no specific external funding for this work.

Supplemental material

The reference implementation of the trust-centric verification pipeline, including the source code for all three pipeline stages, the curated medical claims dataset with gold-standard FOL annotations, the signed pharmacological knowledge graph, evaluation notebooks, and instructions for replication, will be made publicly available upon acceptance. The FOLIO benchmark used in this study is publicly available (Han et al. 2024). AI-assisted tools were used to support LaTeX formatting and language editing during the preparation of this manuscript; all intellectual content originates entirely from the authors, who reviewed and edited all output and take full responsibility for the content of this publication.

References

Artz D and Gil Y (2007) A survey of trust in computer science and the Semantic Web. *Journal of Web Semantics* 5(2): 58–71. DOI:10.1016/j.websem.2007.03.002.

Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA and Pimenta D (2025) A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine* 8(1): 274. Publisher: Nature Publishing Group.

Atanasova P, Simonsen JG, Lioma C and Augenstein I (2020) Generating Fact Checking Explanations. In: *Proceedings of the 58th Annual Meeting of the ACL*. Online: Association for Computational Linguistics, pp. 7352–7364. DOI:10.18653/v1/2020.acl-main.656.

Augenstein I, Baldwin T, Cha M, Chakraborty T, Ciampaglia GL, Corney D, DiResta R, Ferrara E, Hale S, Halevy A, Hovy E, Ji H, Menczer F, Miguez R, Nakov P, Scheufele D, Sharma S and Zagni G (2024) Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence* 6(8): 852–863. DOI: 10.1038/s42256--024--00881-z. Publisher: Nature Publishing Group.

Baader F, Calvanese D, McGuinness DL, Nardi D and Patel-Schneider PF (eds.) (2007) *The Description Logic Handbook: Theory, Implementation and Applications*. 2 edition. Cambridge: Cambridge University Press. ISBN 978–0–521–15011–8. DOI:10.1017/CBO9780511711787.

Badreddine S, d'Avila Garcez A, Serafini L and Spranger M (2022) Logic Tensor Networks. *Artificial Intelligence* 303: 103649. DOI:10.1016/j.artint.2021.103649.

Banarescu L, Bonial C, Cai S, Georgescu M, Griffitt K, Hermjakob U, Knight K, Koehn P, Palmer M and Schneider N (2013) Abstract meaning representation for sembanking. In: *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*. pp. 178–186.

Bellogín A, Grau O, Larsson S, Schimpf G, Sengupta B and Solmaz G (2024) The EU AI Act and the Wager on Trustworthy AI. *Commun. ACM* 67(12): 58–65.

Bender EM, Gebru T, McMillan-Major A and Shmitchell S (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference*

on Fairness, Accountability, and Transparency, FAccT '21. New York, NY, USA: Association for Computing Machinery. ISBN 978–1–4503–8309–7, pp. 610–623. DOI:10.1145/3442188.3445922.

Besold TR, d'Avila Garcez A, Bader S, Bowman H, Domingos P, Hitzler P, Kuhnberger KU, Lamb LC, Lima PMV, de Penning L, Pinkas G, Poon H and Zaverucha G (2021) Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. In: *Neuro-Symbolic Artificial Intelligence: The State of the Art*. IOS Press, pp. 1–51. DOI:10.3233/FAIA210348.

Bo JY, Wan S and Anderson A (2024) To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. DOI:10.48550/arXiv.2412.15584. ArXiv:2412.15584 [cs].

Bordes A, Usunier N, Garcia-Duran A, Weston J and Yakhnenko O (2013) Translating Embeddings for Modeling Multi-relational Data. In: *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Bos J (2008) Wide-Coverage Semantic Analysis with Boxer. In: Bos J and Delmonte R (eds.) *Semantics in Text Processing. STEP 2008 Conference Proceedings*. College Publications, pp. 277–286.

Brachman R and Levesque H (2004) *Knowledge Representation and Reasoning*. 1st edition edition. Amsterdam Boston: Morgan Kaufmann. ISBN 978–1–55860–932–7.

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D (2020) Language Models are Few-Shot Learners. In: *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., pp. 1877–1901.

Brunello A, Ferrarese R, Geatti L, Marzano E, Montanari A and Saccomanno N (2025) Evaluating LLMs Capabilities at Natural Language to Logic Translation: A Preliminary Investigation. In: *Artificial Intelligence and Formal Verification, Logic, Automata, Synthesis*, *CEUR WORKSHOP PROCEEDINGS*, volume 3904. CEUR-WS, p. 8.

Cao L (2024) GraphReason: Enhancing Reasoning Capabilities of Large Language Models through A Graph-Based Verification Approach. In: *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 1–12.

Carnat I (2024) Human, all too human: accounting for automation bias in generative large language models. *International Data Privacy Law* : ipae018DOI:10.1093/idpl/ipae018.

Carraro T, Serafini L and Aiolli F (2024) LTNtorch: PyTorch Implementation of Logic Tensor Networks. ArXiv:2409.16045 [cs].

Carroll JJ, Bizer C, Hayes P and Stickler P (2005) Named graphs, provenance and trust. In: *Proceedings of the 14th international conference on World Wide Web*, WWW '05. New York, NY, USA: Association for Computing Machinery, pp. 613–622.

Cheng K, Gao J, Jiang J and Choudhary A (2024) Neural-Symbolic Methods for Knowledge Graph Reasoning: A Survey. *ACM Transactions on Knowledge Discovery from Data* 18(9). DOI: 10.1145/3686806.

Choudhury A and Chaudhry Z (2024) Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskilling of Health Care Professionals. *Journal of Medical Internet Research* 26: e56764. DOI:10.2196/56764.

Ciampaglia GL, Shiralkar P, Rocha LM, Bollen J, Menczer F and Flammini A (2015) Computational Fact Checking from Knowledge Networks. *PLOS ONE* 10(6): e0128193. DOI: 10.1371/journal.pone.0128193. Publisher: Public Library of Science.

Demartini G, Mizzaro S and Spina D (2020) Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *IEEE Data Eng. Bull.* 43: 65–74.

Devlin J, Chang MW, Lee K and Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *North American Chapter of the Association for Computational Linguistics*.

Dhuliawala S, Komeili M, Xu J, Raileanu R, Li X, Celikyilmaz A and Weston J (2024) Chain-of-Verification Reduces Hallucination in Large Language Models. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, pp. 3563–3578. DOI:10.18653/v1/2024.findings-acl.212.

Dietvorst BJ, Simmons JP and Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1): 114–126. DOI:10.1037/xge0000033.

Doshi-Velez F and Kim B (2017) Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning* .

Floridi L and Cowls J (2019) A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review* 1(1). DOI:10.1162/99608f92.8cd550d1. Publisher: The MIT Press.

Garcez Ad and Lamb LC (2023) Neurosymbolic AI: the 3rd wave. *Artificial Intelligence Review* 56(11): 12387–12406. DOI:10.1007/s10462--023--10448-w.

Garcez ASd, Lamb LC and Gabbay DM (2009) *Neural-Symbolic Cognitive Reasoning*. Cognitive Technologies. Berlin, Heidelberg: Springer. ISBN 978–3–540–73245–7 978–3–540–73246–4.

Golbeck J (2006) Combining Provenance with Trust in Social Networks for Semantic Web Content Filtering. In: Moreau L and Foster I (eds.) *Provenance and Annotation of Data*. Berlin, Heidelberg: Springer. ISBN 978–3–540–46303–0, pp. 101–108. DOI:10.1007/11890850_12.

Goodfellow IJ, Shlens J and Szegedy C (2015) Explaining and Harnessing Adversarial Examples. In: Bengio Y and LeCun Y (eds.) *3rd International Conference on Learning Representations, ICLR 2015*. URL http://arxiv.org/abs/1412.6572.

Han S, Schoelkopf H, Zhao Y, Qi Z, Riddell M, Zhou W, Coady J, Peng D, Qiao Y, Benson L, Sun L, Wardle-Solano A, Szabó H, Zubova E, Burtell M, Fan J, Liu Y, Wong B, Sailor M, Ni A, Nan L, Kasai J, Yu T, Zhang R, Fabbri A, Kryscinski WM, Yavuz S, Liu Y, Lin XV, Joty S, Zhou Y, Xiong C, Ying R, Cohan A and Radev D (2024) FOLIO: Natural Language Reasoning with First-Order Logic. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 22017–22031. DOI:10.18653/v1/2024.emnlp-main.1229.

Hanselowski A, Stab C, Schulz C, Li Z and Gurevych I (2019) A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for

Computational Linguistics, pp. 493–503. DOI:10.18653/v1/K19--1046.

Harth A, Janik M and Staab S (2011) Semantic Web Architecture. In: Domingue J, Fensel D and Hendler JA (eds.) *Handbook of Semantic Web Technologies*. Berlin, Heidelberg: Springer. ISBN 978–3–540–92913–0, pp. 43–75. DOI:10.1007/978--3--540--92913--0_2.

Hartig O (2009) Querying Trust in RDF Data with tSPARQL. In: Aroyo L, Traverso P, Ciravegna F, Cimiano P, Heath T, Hyvönen E, Mizoguchi R, Oren E, Sabou M and Simperl E (eds.) *The Semantic Web: Research and Applications*. Berlin, Heidelberg: Springer, pp. 5–20.

Hitzler P (2021) A review of the semantic web field. *Communications of the ACM* 64(2): 76–83. DOI:10.1145/3397512.

Hogan A, Blomqvist E, Cochez M, D'amato C, Melo GD, Gutierrez C, Kirrane S, Gayo JEL, Navigli R, Neumaier S, Ngomo ACN, Polleres A, Rashid SM, Rula A, Schmelzeisen L, Sequeda J, Staab S and Zimmermann A (2021) Knowledge Graphs. *ACM Comput. Surv.* 54(4): 71:1–71:37. DOI:10.1145/3447772.

Horrocks I (2005) OWL: A Description Logic Based Ontology Language. In: van Beek P (ed.) *Principles and Practice of Constraint Programming - CP 2005*. Berlin, Heidelberg: Springer, pp. 5–8.

Hou K, Li J, Liu Y, Sun S, Zhang H and Jiang H (2024) KG-EGV: A Framework for Question Answering with Integrated Knowledge Graphs and Large Language Models. *Electronics* 13(23): 4835. DOI:10.3390/electronics13234835. Number: 23; Publisher: Multidisciplinary Digital Publishing Institute.

Huang J and Fox MS (2006) An ontology of trust: formal semantics and transitivity. In: *Proceedings of the 8th international conference on Electronic commerce*, ICEC '06. New York, NY, USA: Association for Computing Machinery. ISBN 978–1–59593–392–8, pp. 259–270. DOI:10.1145/1151454.1151499.

Huang Y, Sun L, Wang H, Wu S, Zhang Q, Li Y, Gao C, Huang Y, Lyu W, Zhang Y, Li X, Sun H, Liu Z, Liu Y, Wang Y, Zhang Z, Vidgen B, Kailkhura B, Xiong C, Xiao C, Li C, Xing EP, Huang F, Liu H, Ji H, Wang H, Zhang H, Yao H, Kellis M, Zitnik M, Jiang M, Bansal M, Zou J, Pei J, Liu J, Gao J, Han J, Zhao J, Tang J, Wang J, Vanschoren J, Mitchell J, Shu K, Xu K, Chang KW, He L, Huang L, Backes M, Gong NZ, Yu PS, Chen PY, Gu Q, Xu R, Ying R, Ji S, Jana S, Chen T, Liu T, Zhou T, Wang WY, Li X, Zhang X, Wang X, Xie X, Chen X, Wang X, Liu Y, Ye Y, Cao Y, Chen Y and Zhao Y (2024) Position: TrustLLM: Trustworthiness in Large Language Models. In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR, pp. 20166–20270. ISSN: 2640–3498.

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A and Fung P (2023) Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55(12): 248:1–248:38. DOI:10.1145/3571730.

Kale A, Nguyen T, Harris FC Jr, Li C, Zhang J and Ma X (2023) Provenance documentation to enable explainable and trustworthy AI: A literature review. *Data Intelligence* 5(1): 139–162.

Kim J, Maathuis H and Sent D (2024a) Human-centered evaluation of explainable AI applications: a systematic review. *Frontiers in Artificial Intelligence* 7. DOI:10.3389/frai.2024.1456486. Publisher: Frontiers.

Kim SSY, Liao QV, Vorvoreanu M, Ballard S and Vaughan JW (2024b) "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance

and Trust. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505, pp. 822–835. DOI:10.1145/3630106.3658941.

Koivunen MR and Miller E (2001) The Semantic Web "layer cake" graphic. URL https://www.w3.org/2001/12/semweb-fin/w3csw.

Konstantinovskiy L, Price O, Babakar M and Zubiaga A (2021) Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *Digital Threats* 2(2): 14:1–14:16. DOI:10.1145/3412869.

Li J, Cheng X, Zhao X, Nie JY and Wen JR (2023) HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 6449–6464. DOI:10.18653/v1/2023.emnlp-main.397.

Li Z, Hua W, Wang H, Zhu H and Zhang Y (2024) Formal-LLM: Integrating Formal Language and Natural Language for Controllable LLM-based Agents. DOI:10.48550/arXiv.2402.00798. ArXiv:2402.00798 [cs].

Lin Z, Guan S, Zhang W, Zhang H, Li Y and Zhang H (2024) Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review* 57(9): 243. DOI:10.1007/s10462--024--10896-y.

Lipton ZC (2018) The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16(3): 31–57. DOI:10.1145/3236386.3241340.

Lourenço VN, Paes A, Weyde T, Depeige A and Dubey M (2026) KG-CRAFT: Knowledge Graph-based Contrastive Reasoning with LLMs for Enhancing Automated Fact-checking. In: *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. March 24–29, 2026: Association for Computational Linguistics, pp. 6419–6439. DOI:10.18653/v1/2026.eacl-long.302.

Lu Q, Li R, Sagheb E, Wen A, Wang J, Wang L, Fan JW and Liu H (2025) Explainable Diagnosis Prediction through Neuro-Symbolic Integration. DOI:10.48550/arXiv.2410.01855. ArXiv:2410.01855 [cs].

Madaan A, Tandon N, Gupta P, Hallinan S, Gao L, Wiegreffe S, Alon U, Dziri N, Prabhumoye S, Yang Y, Gupta S, Majumder BP, Hermann K, Welleck S, Yazdanbakhsh A and Clark P (2023) Self-Refine: Iterative Refinement with Self-Feedback. In: *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., pp. 46534–46594.

Manhaeve R, Dumančić S, Kimmig A, Demeester T and De Raedt L (2021a) Neural probabilistic logic programming in DeepProbLog. *Artificial Intelligence* 298: 103504. DOI:10.1016/j.artint.2021.103504.

Manhaeve R, Marra G and Raedt LD (2021b) Approximate Inference for Neural Probabilistic Logic Programming. *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning* 18(1): 475–486.

Marcus G and Davis E (2019) *Rebooting AI: Building Artificial Intelligence We Can Trust*. Vintage.

Min S, Krishna K, Lyu X, Lewis M, Yih Wt, Koh P, Iyyer M, Zettlemoyer L and Hajishirzi H (2023) FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form

Text Generation. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 12076–12100. DOI: 10.18653/v1/2023.emnlp-main.741.

Mirzaei T, Amini L and Esmaeilzadeh P (2024) Clinician voices on ethics of LLM integration in healthcare: a thematic analysis of ethical concerns and implications. *BMC Medical Informatics and Decision Making* 24(1): 250. DOI:10.1186/s12911--024--02656--3.

Moglia A, Georgiou K, Cerveri P, Mainardi L, Satava RM and Cuschieri A (2024) Large language models in healthcare: from a systematic review on medical examinations to a comparative analysis on fundamentals of robotic surgery online test. *Artificial Intelligence Review* 57(9): 231. DOI:10.1007/s10462--024--10849--5.

Moreau L, Clifford B, Freire J, Futrelle J, Gil Y, Groth P, Kwasnikowska N, Miles S, Missier P, Myers J, Plale B, Simmhan Y, Stephan E and den Bussche JV (2011) The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems* 27(6): 743–756. DOI: 10.1016/j.future.2010.07.005.

Olausson T, Gu A, Lipkin B, Zhang C, Solar-Lezama A, Tenenbaum J and Levy R (2023) LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 5153–5176. DOI:10.18653/v1/2023.emnlp-main.313.

OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S et al. (2024) GPT-4 Technical Report. DOI:10.48550/arXiv.2303.08774. ArXiv:2303.08774 [cs].

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano PF, Leike J and Lowe R (2022) Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35: 27730–27744.

Pan L, Albalak A, Wang X and Wang W (2023) Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, pp. 3806–3824. DOI:10.18653/v1/2023.findings-emnlp.248.

Pan S, Hu R, Long G, Jiang J, Yao L and Zhang C (2018) Adversarially regularized graph autoencoder for graph embedding. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18. Stockholm, Sweden: AAAI Press. ISBN 978–0–9992411–2–7, pp. 2609–2615.

Pan S, Luo L, Wang Y, Chen C, Wang J and Wu X (2024) Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36(7): 3580–3599. DOI:10.1109/TKDE.2024.3352100.

Papineni K, Roukos S, Ward T and Zhu WJ (2002) BLEU: A method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 311–318.

Patil R and Gudivada V (2024) A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Applied Sciences* 14(5): 2074. DOI:10.3390/app14052074.

Putra RR, Basuki RSP, Cheng Y and Gao P (2026) NL2LOGIC: AST-Guided Translation of Natural Language into First-Order Logic with Large Language Models. In: *Findings of the Association for Computational Linguistics: EACL 2026*. Association for Computational Linguistics, pp. 6035–6051. DOI:10.18653/v1/2026.findings-eacl.317.

Qiu X, Sun T, Xu Y, Shao Y, Dai N and Huang X (2020) Pre-trained models for natural language processing: A survey. *Science China Technological Sciences* 63(10): 1872–1897. DOI: 10.1007/s11431--020--1647--3.

Quan X, Valentino M, Dennis LA and Freitas A (2024) Verification and Refinement of Natural Language Explanations through LLM-Symbolic Theorem Proving. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 2933–2958. DOI:10.18653/v1/2024.emnlp-main.172.

Rahimzadeh Holagh S and Mohebbi K (2019) A glimpse of Semantic Web trust. *SN Applied Sciences* 1(12): 1732. DOI:10.1007/s42452--019--1598--6.

Riegel R, Gray A, Luus F, Khan N, Makondo N, Yunus Akhalwaya I, Qian H, Fagin R, Barahona F, Sharma U, Ikbal S, Karanam H, Neelam S, Likhyani A and Srivastava S (2020) Logical Neural Networks.

Romeo G and Conti D (2025) Exploring automation bias in human–AI collaboration: a review and implications for explainable AI. *AI & SOCIETY* .

Scherp A, Groener G, Škoda P, Hose K and Vidal ME (2024) Semantic Web: Past, Present, and Future. *Transactions on Graph Data and Knowledge (TGDK)* 2(1): 3:1–3:37. DOI: 10.4230/TGDK.2.1.3.

Shadbolt N, Berners-Lee T and Hall W (2006) The Semantic Web Revisited. *IEEE Intelligent Systems* 21(3): 96–101. DOI:10.1109/MIS.2006.62.

Shi B and Weninger T (2016) Fact Checking in Heterogeneous Information Networks. In: *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 978–1–4503–4144–8, pp. 101–102. DOI:10.1145/2872518.2889354.

Shinn N, Cassano F, Gopinath A, Narasimhan K and Yao S (2023) Reflexion: Language Agents with Verbal Reinforcement Learning. In: *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., pp. 8634–8652.

Stanojević M and Steedman M (2019) CCG Parsing Algorithm with Incremental Tree Rotation. In: *Proceedings of the 2019 Conference of the NAACL: Human Language Technologies, Volume 1*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 228–239. DOI: 10.18653/v1/N19--1020.

Tang L, Laban P and Durrett G (2024) MiniCheck: Efficient Fact-Checking of LLMs on Grounding Documents. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida, USA: Association for Computational Linguistics, pp. 8818–8847. DOI:10.18653/v1/2024.emnlp-main.499.

Thorne J, Vlachos A, Christodoulopoulos C and Mittal A (2018) FEVER: a Large-scale Dataset for Fact Extraction and VERification. In: *Proceedings of the 2018 Conference of the NAACL: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana:

Association for Computational Linguistics, pp. 809–819. DOI:10.18653/v1/N18--1074.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L and Polosukhin I (2017) Attention is All you Need. In: *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wadden D, Lin S, Lo K, Wang LL, van Zuylen M, Cohan A and Hajishirzi H (2020) Fact or Fiction: Verifying Scientific Claims. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 7534–7550. DOI:10.18653/v1/2020.emnlp-main.609.

Wang WY (2017) "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In: *Proceedings of the 55th Annual Meeting of the ACL (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 422–426. DOI: 10.18653/v1/P17--2067.

Wang X, Wei J, Schuurmans D, Le QV, Chi EH, Narang S, Chowdhery A and Zhou D (2023) Self-Consistency Improves Chain of Thought Reasoning in Language Models. In: *Proceedings of the Eleventh International Conference on Learning Representations*.

Wei J, Yang C, Song X, Lu Y, Hu N, Tran D, Peng D, Liu R, Huang D, Du C and Le QV (2024) Long-form factuality in large language models. In: *Advances in Neural Information Processing Systems*, volume 37. DOI:10.52202/079017-2567.

Winters T, Marra G, Manhaeve R and Raedt LD (2022) DeepStochLog: Neural Stochastic Logic Programming. *Proceedings of the AAAI Conference on Artificial Intelligence* 36(9): 10090–10100.

Wu K, Wu E, Wei K, Zhang A, Casasola A, Nguyen T, Riantawan S, Shi P, Ho D and Zou J (2025) An automated framework for assessing how well LLMs cite relevant medical references. *Nature Communications* 16(1): 3615. Publisher: Nature Publishing Group.

Yang Y, Xiong S, Payani A, Shareghi E and Fekri F (2024) 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)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 6942–6959. DOI:10.18653/v1/2024.acl-long.375.

Yang Z, Ishay A and Lee J (2021) NeurASP: embracing neural networks into answer set programming. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20. ISBN 978–0–9992411–6–5, pp. 1755–1762.

Zhao H, Chen H, Yang F, Liu N, Deng H, Cai H, Wang S, Yin D and Du M (2024) Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.* 15(2): 20:1–20:38. DOI:10.1145/3639372.

Author biography

David Farrugia is a data scientist and Ph.D. candidate in Artificial Intelligence at the University of Malta. He has worked across diverse industries including gaming, manufacturing, customer relationship management, affiliate marketing, and anti-fraud. His research explores the intersection of applied data science and academic inquiry, with a focus on neuro-symbolic AI, trustworthy language technologies, and provenance-aware knowledge systems.

Author biography

Alexiei Dingli is a Professor of Artificial Intelligence at the University of Malta with over two decades of experience in the field. His work has been recognised as world-class by international bodies, earning awards from the European Space Agency, the World Intellectual Property Organization, and the United Nations. He has helped numerous organisations implement AI solutions and has been a core member of the Malta.AI task force. He holds an MBA from Grenoble Business School specialising in Technology Management.