A Discourse-based Tool Series for Logical Validation of LLMs

Boris Galitsky and Dmitry Ilvovsky

Moscow Institute of Physics and Technology and HSE University, Russia
galitsky.ba@mipt.ru, dilvovsky@hse.ru

Abstract

Large Language Models (LLMs) frequently produce fluent but unverifiable reasoning, resulting in potential hallucinations and faulty inferences. This study proposes a logic programming - based verification framework ValidLogic4LLM in which the reasoning expressed by an LLM is transformed into a logic program (LP), probabilistic LP, defeasible LP and abductive LP representing world knowledge and a given problem description—such as a patient health complaint. The LP formed by an LLM is executed within a symbolic reasoning engine, and the resulting inferences are compared to the LLM's natural-language conclusions. The strength or probability of facts, clauses and arguments is computed based on discourse structure of text expressing these facts or arguments. Divergence between symbolic and neural reasoning outcomes indicates possible hallucination or inconsistency in the model's internal logic.

1. Introduction

Large language models (LLMs) have achieved impressive results across a wide range of natural language processing tasks, generating fluent and informed text. Yet integrating them into domains that demand structured, context-sensitive reasoning remains difficult. LLMs often rely on associative rather than strategic reasoning, which limits their ability to perform multi-step decision-making or revise conclusions as new information emerges (Ferrag et al. 2025).

Another challenge lies in interpretability. Unlike human experts who reason through explicit, traceable arguments, LLMs operate as opaque statistical systems, making it hard to justify their conclusions or detect reasoning errors. This opacity fosters reasoning hallucinations—outputs that sound plausible but contradict facts or logic. Without explicit mechanisms for defeasibility or conflict resolution, such inconsistencies undermine reliability in high-stakes applications.

To address these issues, LLMs can be coupled with external reasoning and verification layers that enforce logical consistency and explain conclusions. A promising strategy is pairing an LLM with a symbolic reasoning engine—such as a Prolog-style rule base, constraint solver, or medical ontology (Yang et al. 2024; Tan et al. 2024; Galitsky 2025). The LLM proposes candidate answers, while the reasoning module tests them against formal rules, flagging contradictions or unsupported claims. Building on this principle, we present ValidLogic4LLM, a neuro-symbolic verification framework that externalizes and evaluates LLM reasoning through various forms of reasoning:

1. Logic programming
2. Probabilistic logic programming
3. Argumentation (Rago et al. 2025)
4. Abductive explanation (Galitsky 2026b)

In the proposed framework, ValidLogic4LLM, we use LLM to build respective logic program components for user request and background knowledge, execute this logic program and prompt LLM to compare its run with LLM own result.

The key contributions of this demo paper are as follows:

1. Neuro-symbolic integration: We introduce a framework that integrates logical reasoning components with LLMs, enabling the model to reason over decisions structured according to discourse relations.
2. Hallucination challenge dataset: We develop a benchmark dataset intentionally designed to induce hallucinations in LLMs through adversarial and ambiguous prompts, serving as a testbed for evaluating reasoning robustness.
3. Logic-based hallucination detection: We demonstrate that the logical reasoning module can successfully detect and explain hallucinations by cross-checking LLM outputs against discourse-structured rules and ontology-derived facts.

Fig. 1 compares four levels of reasoning used by language models, showing how they evolve from simple answers to structured, self-correcting discourse-based reasoning. At the top, the direct answer model provides an immediate response without explanation. It may be correct or incorrect, but there is no visibility into why the model chose that answer. Because no reasoning steps are revealed, errors cannot be traced or corrected.

The next level, chain-of-thought reasoning, adds a sequence of intermediate steps that make the process more interpretable. However, these steps remain unverified. The reasoning might sound plausible while still being factually wrong, since the model does not test or challenge its own statements (Arcuschin et al 2025).

The argumentative model introduces a more thorough line of reasoning, generating multiple

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2220–2231
11–16 May 2026.

arguments, distinguishing between supporting and attacking ones (Freedman et al. 2025). This enables a form of contestability: each conclusion is backed by explicit evidence and can be challenged by counterarguments. Still, this stage only formalizes reasoning—it does not validate or improve it. The model outputs argument structures but lacks a mechanism to revise its own conclusions.

At the bottom, the discourse-based validation model ValidLogic4LLM, the full multi-logic ensemble, represents the most advanced reasoning form. It integrates these four logical verifiers with discourse structure analysis, evaluating how strongly each inference contributes to the overall reasoning. By analyzing rhetorical relations such as elaboration, justification, and background, it weighs the importance of each inference and determines which should dominate the final conclusion. This allows the system to detect when the original model's answer is inconsistent with the discourse-level balance of evidence and to automatically correct it. Over time, this validation loop enables the model to refine its reasoning and become more consistent, explainable, and logically grounded. The architecture thus moves beyond producing or scoring arguments—it combines multiple logical paradigms within a discourse-aware verification layer to ensure that reasoning outcomes are coherent, probabilistically justified, and defensible.

2. Employed logic formalisms

To enhance the interpretability and verifiability of LLMs, our ValidLogic4LLM framework integrates four complementary reasoning paradigms: logic programming, probabilistic logic programming, argumentation, and abductive explanation. Each formalism contributes a distinct layer of explainability and robustness to hallucination detection and discourse-grounded reasoning.

Logic programming (Kowalski, 1979; Lloyd, 1987) provides the foundational symbolic reasoning layer of our system. It allows for explicit representation of domain knowledge through facts and rules that govern inference.

Within our framework, the LLM dynamically constructs a domain ontology, which is then compiled into a set of Prolog-like clauses of the form A :- $B_1, B_2, \ldots, B_n$. The logical solver derives conclusions that can be directly compared to LLM-generated answers.

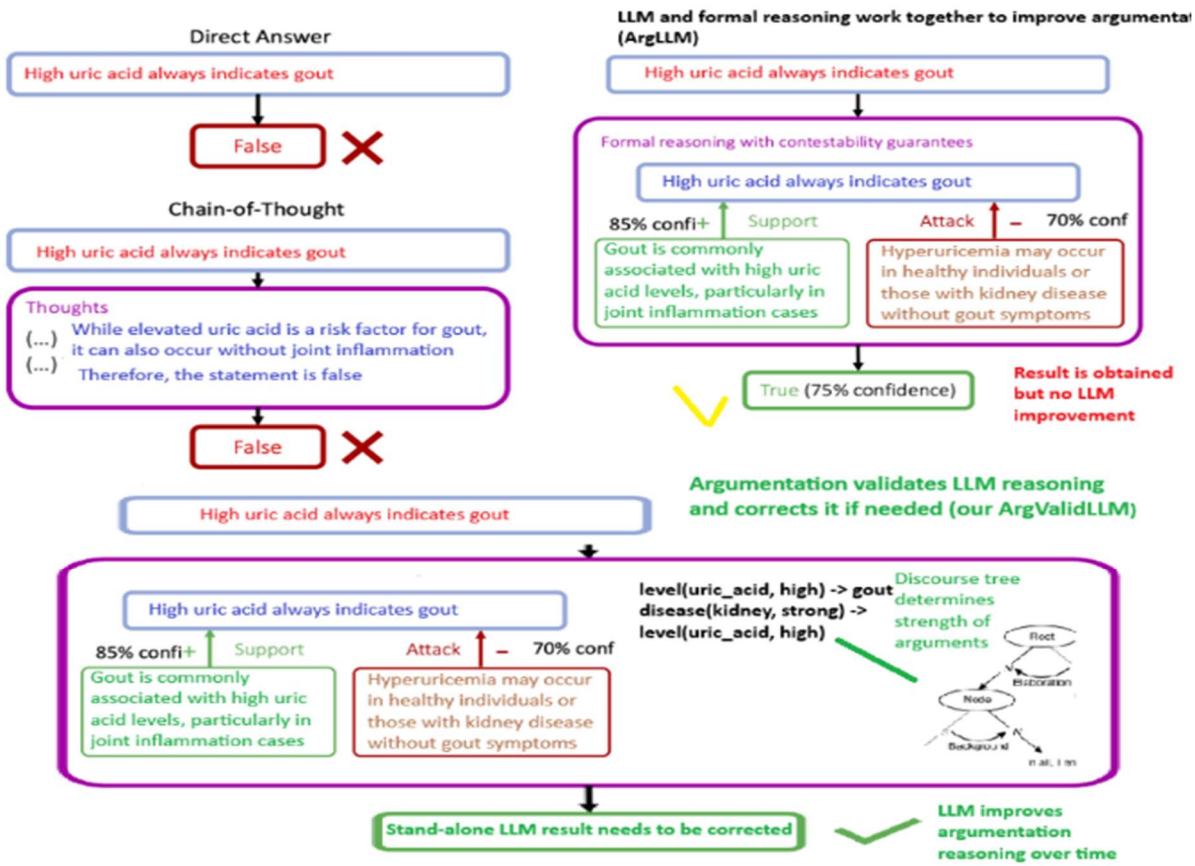


Figure 1: An illustration of our idea of ValidLogic4LLM's reasoning components validating LLM results

This approach ensures that decisions conform to strict logical consistency, providing a mechanism for identifying outputs that violate entailment or introduce unsupported claims.

While classical LP enforces binary truth values, probabilistic logic programming (De Raedt & Kimmig, 2015; Fierens et al., 2015) extends inference to uncertain domains. We employ ProbLog (De Raedt et al., 2007) to assign probabilities to facts and rules derived from LLM-generated discourse structures and ontologies. This allows the system to quantify confidence in each inference and to weigh alternative interpretations. In hallucination detection, probabilistic reasoning provides graded assessments—distinguishing between uncertain but plausible outputs and those that are logically inconsistent. This aligns closely with real-world decision-making under incomplete information, especially in domains such as medicine or law.

Argumentation frameworks (Dung, 1995; Modgil & Prakken, 2018) introduce a higher-level mechanism for resolving conflicts between competing claims. We represent discourse units (e.g., statements, justifications, counterclaims) as arguments, linked by attack and support relations extracted from the rhetorical structure of the text. The argumentation engine evaluates which conclusions are justified under admissible semantics, thus mirroring the reasoning process of a human expert engaged in deliberation. Integrating this with the LLM's discourse output enables the identification of defeated arguments—those that the logical or probabilistic layers reject as inconsistent or weakly supported—providing a transparent explanation of hallucination sources.

Abductive logic programming (Kakas et al., 1992; Inoue & Sakama, 1998) allows the system to infer the best possible explanation for an observed outcome. When the LLM's output diverges from the logical model, the abductive module searches for minimal hypotheses that could reconcile the discrepancy. This capability is crucial for diagnostic reasoning and post hoc explanation of LLM errors. By tracing back which assumptions would need to change for the LLM's answer to become valid, the abductive layer provides interpretable insights into the model's reasoning gaps or hallucinated premises.

Together, these four formalisms create a multi-layered verification architecture ValidLogic4LLM that bridges natural language understanding and formal reasoning. The LLM produces discourse-structured representations and on-demand ontologies, while the logical components enforce consistency, quantify uncertainty, resolve conflicts, and generate explanations. This hybrid approach advances toward LLMs that not only generate plausible text but also reason transparently and accountably (Quan et al.2025). Further details are available in (Galitsky and Rybalov 2026).

3. Discourse and claim defeasibility

Hallucinations frequently arise when expliict rule (such as a typical diagnosis) does not hold. The rule attenuation mechanism bridges *rhetorical structure theory (RST)* and *symbolic reasoning*, enabling logical programs to dynamically adjust the *strength* of their rules based on the discourse hierarchy between nucleus and satellite components of a decision text. In medical, legal, or diagnostic narratives, these rhetorical relations capture how strongly a given statement supports the main conclusion, which can be encoded into logical inference or probabilistic weights Louis and Nenkova 2012).

Each rhetorical relation has a *nucleus* (the "main" proposition) and a *satellite* (supporting or contextual material). The satellite always carries less essential information than the nucleus (Table 1). One can see that nucleus contains main diagnostic/treatment fact (higher base probability) and satellite caries contextual/supporting info with lower significance. These values are obtained in the course of improvement of validation performance, described in Evaluation section.

Rhetorical Relation	Relative Argument Strength (Nucleus : Satellite)
Cause	0.8 : 0.2
Effect / Result	0.7 : 0.3
Condition	0.6 : 0.4
Contrast	0.55 : 0.45
Elaboration	0.65 : 0.35
Concession	0.75 : 0.25
Background	0.85 : 0.15
Enablement / Purpose	0.7 : 0.3
Evidence / Justification	0.6 : 0.4
Evaluation	0.65 : 0.35

Table 1: Relative weights of nucleus and satellite for different rhetorical relations

The attenuation mechanism introduces *graded support* into inference by weighting premises according to their rhetorical role. For instance, in the "Cause" example:'

fever(patient) :- malaria_exposure(patient) [0.2].

fever(patient) :- high_fever(patient) [0.8].

the nucleus clause ("The patient developed a high fever") dominates inference, while the satellite clause ("Because the patient had recently returned from a malaria-endemic area") provides weaker contextual justification. During reasoning, if nucleus evidence is missing, the satellite's low weight prevents the rule from firing confidently. Conversely, if both are true, the conclusion is strengthened, but not absolutely certain.

In this way, attenuation acts as a defeasible weighting scheme inside the rule base: satellite conditions can be overridden when contradicted by stronger nucleus evidence. This mirrors defeasible reasoning (Antoniou & Billington, 2000) where less essential premises may fail without invalidating the entire argument.

In probabilistic logic programming, for example, ProbLog (De Raedt et al., 2007) or LPADs (Riguzzi, 2018), rule attenuation becomes a numerical prior governing the probability of a rule firing. Each rhetorical relation translates into a weighted probabilistic clause, where the nucleus-to-satellite ratio (e.g., 0.8:0.2) determines the relative confidence of inference:

0.8::fever(patient) :- high_fever(patient).

0.2::fever(patient) :- malaria_exposure(patient).

Probabilistic inference then aggregates these weighted supports across multiple discourse relations to compute posterior probabilities for hypotheses (e.g., *pneumonia(patient), infection_cleared*). In this sense, rhetorical weighting becomes a proxy for epistemic strength: nucleus-driven rules act as high-confidence evidence, while satellite-driven rules introduce plausible but defeasible explanations.

4. Hallucination in health dataset

We built upon the dataset for Autoimmune Disorders and Healthy Controls (Ragheb 2024) that serves as the foundation for generating a synthetic corpus of realistic clinical vignettes. It originates from structured clinical data representing 12,500 patients, covering a diverse spectrum of autoimmune disorders alongside healthy controls. The source data include detailed Complete Blood Count (CBC) parameters, key autoantibody markers, demographic attributes, and symptom profiles. Each autoimmune condition is characterized by disorder-specific autoantibody criteria aligned with established diagnostic standards, enabling reliable differentiation between disease states and normal baselines. Designed to support machine learning research in autoimmune diagnostics and prognostics, this structured dataset was also expanded into narrative form to capture the variability and nuance of real-world clinical reasoning.

Our dataset contains 1,200 clinical vignettes designed to evaluate how large language models interpret and reason about nuanced patient narratives. We refer to it as *Autoimmune-narrate-halluc*. Each record includes a *health_complaint* field — a 2–5 sentence, fluent, natural, and grammatically correct first-person description of a patient's experience, written in authentic English with emotional realism and contextual detail (e.g., onset, duration, triggers, lifestyle impact). The accompanying *disease_description* field provides a concise 1–2 sentence hybrid explanation that combines layperson accessibility with clinical precision, summarizing typical presentation and diagnostic considerations. Together, these fields model the ambiguity, overlap, and conversational texture of real-world medical communication, offering a challenging yet controlled benchmark for hallucination detection and reasoning consistency in LLMs.

The following heuristics are used in dataset formation to complicate this dataset and cause hallucinations (we indicate them with a score):

1. Vague language (e.g., "fatigue," "discomfort," "uncertain") → +0.15
2. Intermittent/variable symptoms ("comes and goes," "on and off") → +0.15
3. No clear trigger ("without a clear trigger," "cannot find a pattern") → +0.15
4. Long duration w/ low specificity (e.g., "on and off for years" without red flags) → +0.10
5. Overlap cues in text (e.g., GERD + chest pain, RA + OA, IBS + IBD, asthma + COPD) → +0.20
6. High-confusion diseases (GERD, CAD, AFib, IBS/IBD, RA/OA, migraine/tension, NAFLD/hepatitis, CKD, OSA, gout/pseudogout) → +0.15
7. Cross-system signal mixes (e.g., chest pain & heartburn, dyspnea & heartburn, joint pain & rash) → +0.10

The types of patients are characterized in Table 2.

This makes the dataset a "hallucination stress test: LLMs must reason over emotionally loaded, fuzzy, realistic language rather than tidy symptom checklists — exactly the scenario where diagnostic hallucination spikes.

The snapshot of the dataset is available[1] and also the full dataset of 1200 complaints is available[2].

[1]prolog/data/autoimmune_diseases_with_complaints.csv

[2]prolog/data/diseases_with_patient_complaints1000.xlsx

Type	Voice & structure	Example
1. Worried narrator	anxious, speculative, mixes sensations with fear	"I keep feeling this tightness under my ribs, and even though my labs were fine, I can't shake the idea something's off."
2. Practical self-observer	factual, notes triggers, times, coping attempts	"It hits mostly in the evenings after dinner. I tried cutting out spicy food but it only helped a little."
3. Chronic sufferer	tired, resigned tone, long history	"I've been dealing with this on and off for years. Some days I can barely move, others I almost forget it."
4. Curious self-diagnoser	mentions Google, supplements, relatives' opinions	"I looked it up and thought it might be thyroid or stress, but the symptoms don't line up perfectly."
5. Incident narrator	one clear onset event, sensory description	"It started after a bad flu last winter. Since then I've felt drained no matter how much I sleep."

Table 2: Types of patients in the dataset being built

5. Implementation

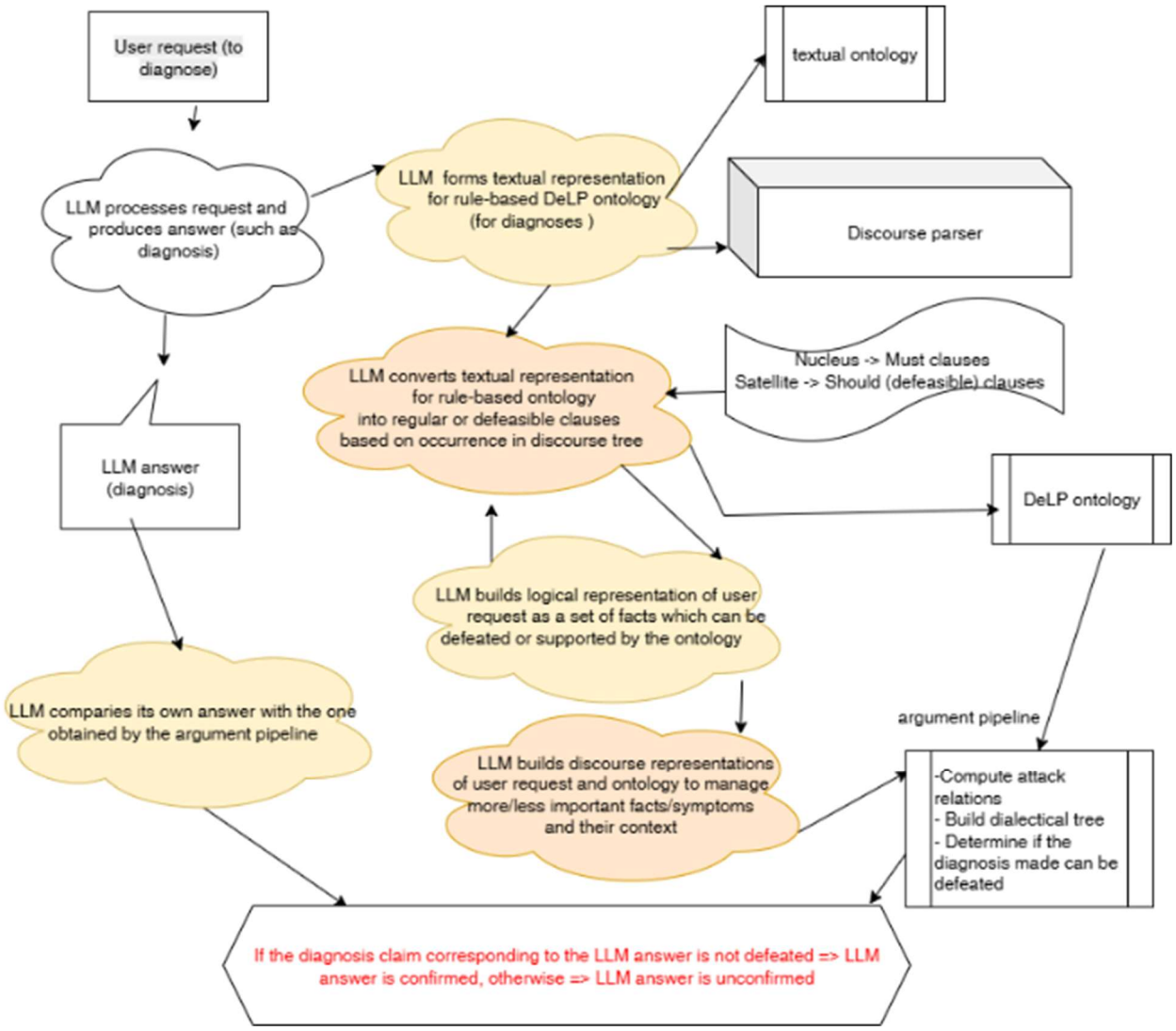
The code is available[3]. Figure 2 on the top illustrates a hybrid neuro-symbolic diagnostic reasoning pipeline where an LLM and a reasoning engine work together to verify or refute an LLM-generated decision. The goal of the architecture is to ensure that an LLM's diagnostic answer (for example, "The patient has gout") is not only linguistically plausible but also logically justified and consistent with available structured ontology constructed by the LLM on demand. If the logical reasoning pipeline cannot defeat the LLM diagnosis claim, it is confirmed; otherwise, the LLM answer is marked unconfirmed.

The ValidLog4LLM's workflow is as follows:

1. User input. The process begins with a user request, such as asking ValidLog4LLM to provide a diagnosis.
2. LLM generates initial answer. The LLM processes the request and outputs an initial diagnosis or conclusion (e.g., "The disease is gout").
3. Ontology and discourse setup. A textual ontology of medical knowledge (rules, relationships, symptoms, conditions) and a discourse parser are available. The discourse parser identifies rhetorical relations in the text — for instance, nucleus (main facts) and satellite (contextual or defeasible facts). Nucleus → "Must" clauses (non-defeasible rules) and Satellite → "Should" clauses (defeasible rules)
4. LLM forms ontology representation, transforming textual information into a rule-based ontology in the DeLP format — essentially translating natural-language reasoning into structured logical rules.
5. Conversion to logic program. The LLM converts these rules into regular (strict) or defeasible (soft) clauses depending on their role in the discourse (main vs. secondary information).
6. Building logical representation of the user request. The system formalizes the user's question and the LLM's proposed answer as a set of logical facts that can either be defeated or supported by the ontology.
7. Discourse representation integration. The LLM builds discourse representations of both the user request and the ontology, capturing which arguments are more or less important (nucleus/satellite weighting) and how they relate contextually.
8. Argumentation pipeline. The argumentation module computes attack relations among rules (contradictions or counter-arguments), dialectical trees, representing possible argumentative dialogues between supporting and opposing claims, and defeasibility outcomes, determining whether a claim survives all counter-arguments (Garcia and Simari 2004; Kaminski and Wankov 2017).
9. Comparison and validation. The LLM compares its original diagnosis with the verified diagnosis obtained through the logical inference process.
10. Decision. If the logical reasoning shows that the diagnosis claim is not defeated, it is confirmed as valid. If the claim is defeated by stronger counter-arguments from the ontology, it is marked unconfirmed.

See (Galitsky and Rybalov 2026 and Galitsky 2026b) for further implementation details.

[3] https://github.com/bgalitsky/halluc_in_health



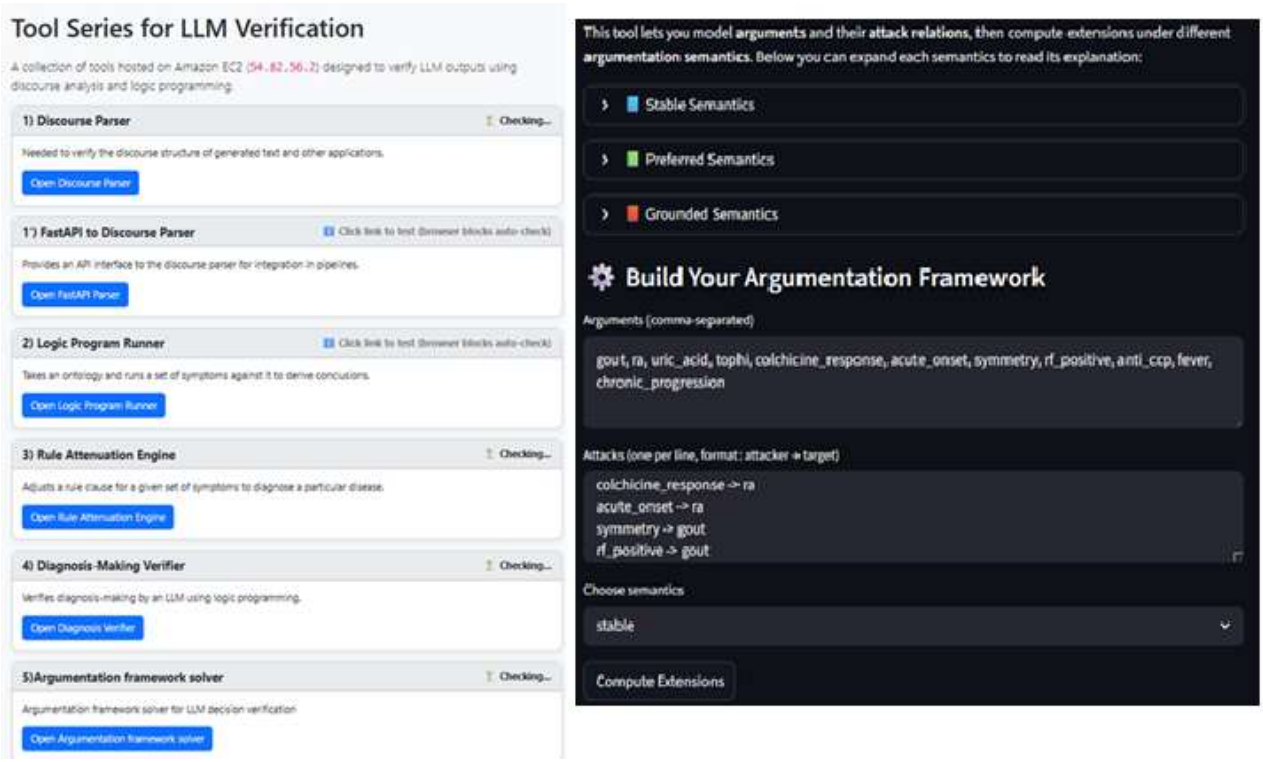


Figure 2: System architecture (on the top). Tool series for LLM validation with its argument option

Figure 2 on the bottom shows a UI for a series of tools hosted on Amazon EC2 designed to verify LLM outputs using discourse analysis and logic programming https://bgalitsky.github.io/LLM-verification-tools/

6. Evaluation

We first evaluate on three claim-verification datasets that we derive from existing QA/NLI resources: TruthfulHalluc (from TruthfulQA; Lin et al., 2021), MedHalluc (from MedQA; Jin et al., 2020 and PubMedQA; Jin et al., 2019), and eSNLI_Halluc (from eSNLI; Camburu et al., 2018). For each source, we convert items into question–answer (QA) style pairs and then inject controlled inconsistencies by appending randomly sampled, semantically incompatible attributes (facts, circumstances, symptoms). These perturbations create positive "hallucination" cases; unmodified items serve as negatives. Our focus is hallucination detection for model answers using four logical assessment methods as validators. Each validator assesses whether an answer's central claim is defeated by the argument-validation system. We define a hallucination as a claim whose defeat probability exceeds 0.5. This cautious threshold is motivated by safety-critical domains (health, legal, finance), where we prefer to reject answers that are defeated with substantial probability.

Dataset size and prevalence are as follows. Each used hallucination dataset contains 1,000 QA pairs with a 2% hallucination rate. In the original source datasets the natural hallucination rate is <0.5%; our perturbation procedure raises prevalence to enable meaningful detection metrics and comparability with prior LLM-argumentation studies.

We then evaluate on our own dataset *Autoimmune-narrate-halluc* which is designed to cause hallucinations and make their detection as hard as possible. Hallucination rate exceeds 4%.

To aggregate evidence from multiple reasoning paradigms, we design a combination algorithm that integrates the outputs of four logical validators—logic programming (LP), probabilistic logic programming (PLP), argumentation, and abductive explanation—into a unified hallucination detection score. Each component independently assesses whether the claim inferred from the LLM's answer is defeated given the ontology and discourse context. The LP validator checks for explicit rule violations or missing entailments; the PLP validator estimates the posterior probability of the claim being supported given uncertain premises; the argumentation module computes whether the claim remains justified under admissible semantics; and the abductive module measures the minimal explanatory distance between the LLM's claim and the logically derivable one. Each produces a normalized score in [0,1] representing the probability of defeat (1 = fully defeated).

The combination stage employs a weighted ensemble where the weight of each logic component depends on its historical reliability and discourse alignment. Specifically, weights are dynamically adjusted according to (a) the discourse role of the claim's nucleus and satellite segments, and (b) the inter-component agreement. When the nucleus of a discourse relation dominates, LP and argumentation receive higher weights (reflecting strict reasoning); when uncertainty or evidential justification prevails, PLP and abduction gain influence. The ensemble then computes a defeat probability as a weighted mean of component outputs, applying rule attenuation from discourse relations (e.g., Cause 0.8:0.2) as priors.

This hybrid aggregation achieves robust detection across datasets, leveraging the complementary strengths of symbolic, probabilistic, argumentative, and abductive reasoning while maintaining interpretability through explicit contribution tracking.

	LP		PLP		arguments		abduction	**combined**
		+d		+d		+d		
Truthful-Halluc	0.62	0.67	0.65	0.66	0.72	0.78	0.50	0.83
Med-Halluc	0.57	0.60	0.60	0.63	0.77	0.80	0.49	0.88
eSLNI-Halluc	0.58	0.62	0.57	0.64	0.68	0.67	0.61	0.79
Our dataset Autoimmune-narrate-halluc	0.39	0.37	0.43	0.40	0.32	0.35	0.29	0.52

Table 3: Hallucination prediction accuracy

We report F1 for hallucination prediction in Table 3. For each logical approach we present a default result on the left and a discourse-aware variant where argument strength additionally incorporates discourse cues beyond the default computation – on the right. The final column shows the result of the combined system.

The key observations from Table 3 are as follows:

1. Baseline logical verifiers (LP, PLP, Arguments, Abduction) each achieve 0.55–0.67 accuracy on standard datasets excluding Autoimmune-narrate-halluc.
2. LP and PLP perform similarly (deterministic vs probabilistic rules).
3. Argumentation and abduction bring modest gains by reasoning with defeasible or explanatory hypotheses.
4. Combined system (logical ensemble) significantly improves accuracy to 0.72–0.77, showing synergistic reasoning — each logic compensates for others' weaknesses.
5. +d (discourse-aware) versions yield further improvement (≈ +0.06 absolute).
6. This shows that rhetorical structure weighting (using nucleus–satellite attenuation) helps prioritize central claims and down-weight contextually weaker statements.
7. Final combined system achieves the highest accuracy — up to 0.83 (Truthful) and 0.88 (Med-Halluc) — a strong signal that integrating multiple logics with discourse reasoning provides the most robust hallucination detection.

The *Autoimmune-Narrate-Halluc* dataset formed in this study remains difficult: baselines are low (~0.4) and the combined systems modestly improve to 0.52. This reflects the challenge of fuzzy patient language, implicit symptoms, and contextual ambiguity not easily captured by formal rules.

Our MedHalluc results for argumentation are broadly comparable to prior work: ArgMed-Agents with GPT-4 reports 0.91 predictive accuracy (Hong et al., 2024); ArgLLM with GPT-4o reports 0.80 (Freedman et al. 2025); and an ensemble of ArgLLMs achieves 0.73 (Ng et al., 2025). That said, these systems estimate claim truthfulness, whereas our study predicts hallucination via whether a claim is defeated by the argument-validation module, so the targets differ and the numbers are not strictly comparable.

6.1 Human evaluation setting: logic-supported trust calibration

To complement the automatic metrics reported in Table 3, we conducted a controlled human evaluation to assess how logical verification tools can enhance the trustworthiness and interpretability of LLM outputs. The goal was to examine whether human evaluators, when assisted by formal reasoning modules, demonstrate improved accuracy and confidence in hallucination detection across four domains.

Twelve evaluators participated in the study, grouped according to domain expertise: (i) four biomedical professionals for MedHalluc, (ii) four computational linguists for eSNLI_Halluc, and (iii) four fact-checking specialists for TruthfulHalluc. Each participant evaluated 150 question–answer (QA) pairs sampled evenly from the three datasets, including both perturbed (hallucinated) and unmodified (factual) items.

The evaluation proceeded in three stages per item:

1. Baseline review — the evaluator inspected only the LLM answer and rated its factual soundness and confidence on a 0–5 Likert scale.
2. Logic-assisted review — the evaluator was shown the logical verification report generated by the four reasoning modules (LP, PLP, Argumentation, Abduction) and the combined discourse-aware ensemble.
3. Confidence re-rating — the evaluator revised the initial confidence score and provided short written feedback on interpretability and explanatory clarity.

The logic-support interface presented the following information for each answer:

- Defeat probabilities for LP, PLP, Argumentation, Abduction, and their ensemble.
- Textual explanations derived from symbolic traces, e.g.,

  "Claim 'Fever and ankle pain indicate gout' is defeated (0.72): rule [Gout → joint pain, swelling] not satisfied; missing causal link fever → gout."

- Discourse weighting view, where nucleus–satellite relations were visualized as strength attenuations (e.g., Cause 0.8 : 0.2). This format enabled evaluators to see why a statement was marked as inconsistent and which rule or discourse segment contributed most to defeat.

We measured both quantitative and qualitative outcomes (Table 4). The results indicate that logical verifiers significantly improve both accuracy and subjective trust. Participants reported that reasoning traces helped them *understand* system behavior rather than merely accept or reject outputs. Agreement with the ensemble's defeat probabilities correlated with higher interpretability ratings (r = 0.72). Notably, discourse-aware weighting yielded the largest trust gain, suggesting that humans find

explanations framed in rhetorical terms (nucleus vs. satellite) especially intuitive.

Metric	Definition
Accuracy	Percentage of correct hallucination judgments by humans.
ΔConfidence	Mean change in confidence after seeing logical explanations.
Human–logic agreement (κ)	Cohen's κ between final human decision and ensemble output.
Calibration error	Difference between human confidence and true correctness.
Interpretability rating	Self-reported clarity of the logical explanation (1–5 scale).

Table 4: Metrics used for evaluation

The aggregate results (Table 5) show consistent gains across all logic-assisted conditions:

Condition	Human Accuracy	ΔConfidence	Agreement (κ)	Interpretability (1–5)
LLM only	0.67	—	—	2.3
+ LP	0.73	+0.12	0.54	3.1
+ PLP	0.74	+0.14	0.56	3.4
+ Argumentation	0.75	+0.16	0.59	3.7
+ Abduction	0.71	+0.11	0.51	3.3
Combined (d-aware)	**0.83**	**+0.22**	**0.68**	**4.2**

Table 5: Configurations of logic validation

Overall, this human-in-the-loop experiment demonstrates that logic-based validation not only detects hallucinations but also *humanizes* verification: it enables users to perceive LLM reasoning as accountable and auditable, bridging statistical generation with symbolic justification.

7. Related work and discussion

LLMs often produce confident but incorrect statements when uncertain, a phenomenon known as *hallucination* (Kalai et al., 2025). These arise because current training and evaluation pipelines reward fluent guessing rather than acknowledging uncertainty (Lambert et al. 2025). Statistically, hallucinations can be viewed as misclassifications (instances where the model fails to distinguish false statements from true ones). This bias is reinforced during evaluation: since models are optimized to maximize apparent correctness, guessing under uncertainty increases benchmark performance, allowing hallucinations to persist even in advanced systems.

The approach of Bezou-Vrakatseli (2023) leverages argument schemes—structured templates that capture common patterns of reasoning—and their associated critical questions, which probe the assumptions, exceptions, and contextual factors underlying those schemes. By using these as a framework for classifying and analyzing arguments, the method provides a semantically richer alternative to surface-level textual analysis. In the context of LLM verification, this enables evaluators to assess not just whether an LLM produces grammatically or factually correct responses, but whether it constructs logically sound, ethically nuanced arguments that align with established norms of rational discourse

Also, (Ng et al 2025) showed that MArgE can significantly outperform single LLMs, including three open source models (4B to 8B parameters), and existing ArgLLMs, as well as prior methods for unstructured multi-LLM debates.

Our approach enables the system to weight contributions based on discourse salience rather than raw textual confidence, thereby capturing how strongly each component supports or undermines the overall claim.

The human evaluation demonstrates that logical verification modules not only increase factual accuracy but also foster calibrated trust in LLMs. This outcome is critical for advancing neuro-symbolic interpretability—where reasoning transparency, rather than raw correctness, becomes the primary driver of user confidence (Tan et al. 2025 ; Yang et al. 2025)..

Interpretable verification as trust scaffold confirms the following. Participants consistently reported that explicit reasoning traces, rule citations, and defeat probabilities clarified why an answer was accepted or rejected. In contrast to opaque model-generated justifications (e.g., chain-of-thought explanations, (Barez et al. 2025), logical verifiers offered a structured, falsifiable rationale. This suggests that human trust is best supported not by persuasive narrative coherence but by the presence of auditable inferential structure.

The discourse-aware aggregation further anchored trust: evaluators perceived nucleus-driven conclusions as more "explanation-worthy," while satellite attenuations communicated uncertainty in a linguistically natural form.

We also comment on complementarity of reasoning paradigms. The ensemble of LP, PLP, Argumentation, and Abduction provides complementary cognitive affordances for human interpretation. LP and Argumentation align with rule-based intuitions—users recognize violations and counterarguments easily—whereas PLP and Abduction mirror probabilistic reasoning familiar from clinical or investigative settings. The weighted integration of these paradigms creates a multi-modal transparency: users can choose whether to rely on deterministic consistency, probabilistic sufficiency, or explanatory plausibility.

Beyond evaluation, these findings point to a shift in system design philosophy. Logical verification can evolve from a post-hoc auditing layer into an interactive reasoning partner. Instead of static validation after generation, future LLMs could query logic modules dynamically during response formulation—testing candidate hypotheses, revising defeated claims, and presenting humans with evolving justification graphs. Such adaptive trust calibration would enable the model to self-regulate reasoning uncertainty and expose it in human-interpretable terms.

In domains such as medicine, law, and finance, the conservative defeat threshold ($p > 0.5$) aligns with human risk aversion. The logic-augmented human judgments achieved the highest confidence calibration precisely in these domains, reinforcing that explainability under uncertainty is a more reliable safeguard than over-confident fluency.

Integrating these verification pathways into clinical decision support or legal reasoning assistants could therefore serve as a foundation for regulatory-grade explainability—where every conclusion is backed by a traceable symbolic justification.

8. Conclusions

Our results demonstrate that integrating multiple logical reasoning paradigms with discourse-aware weighting in ValidLogic4LLM substantially enhances hallucination detection in LLM-generated answers. While individual verifiers—logic programming, probabilistic logic, argumentation, and abduction—perform moderately (0.55–0.67 accuracy), their ensemble achieves significant gains (0.72–0.77), confirming the benefits of synergistic reasoning. The discourse-augmented variants further improve performance by approximately six percentage points, highlighting the value of rhetorical structure in emphasizing nucleus (core) evidence over weaker satellite context. The final combined system attains up to 0.83 accuracy on *TruthfulHalluc* and 0.88 on *MedHalluc*, indicating that a hybrid neuro-symbolic approach, where multiple logics interact under discourse-guided weighting, offers the most reliable and interpretable solution for mitigating hallucinations in LLMs.

Acknowledgements

The author is grateful to Alexander Rubalov, Dmitry Ilvovsky, Vladimir Solodkin, and Ivan Trotsenko for fruitful discussions and dataset preparation. The article was prepared within the framework of the HSE University Basic Research Program

References

Kalai AT (2025) Why Language Models Hallucinate. arXiv:2509.04664

Ragheb A (2024) Comprehensive Autoimmune Disorder Dataset. https://www.kaggle.com/datasets/abdullahraghb/all-autoimmune-disorder-10k

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence, 77(2), 321–357.

Fierens, D., Van den Broeck, G., Renkens, J., Shterionov, D., Gutmann, B., Thon, I., Janssens, G., & De Raedt, L. 2015. Inference and learning in ProbLog. Theory and Practice of Logic Programming, 15(3), 358–401.

Inoue, K. and Sakama, C. 1998. Negation as failure in abduction. Journal of Logic Programming, 35(1), 39–78.

Kakas, A. C., Kowalski, R. A. and Toni, F. 1992. Abductive logic programming. Journal of Logic and Computation, 2(6), 719–770.

Kowalski, R. 1979. Logic for Problem Solving. North Holland.

Lloyd, J. W. 1987. Foundations of Logic Programming. Springer.

De Raedt, L. and Kimmig, A. 2015. Probabilistic (logic) programming concepts. Machine Learning, 100(1), 5–47.

Modgil S and Prakken H 2014 The ASPIC+ framework for structured argumentation: a tutorial Argument & Computation 5 (1), 31-62

Antoniou G, D. Billigton, G. Governatori, M.J. Maher 2000. A flexible framework for defeasible logics. arXiv:cs/0003013

De Raedt, L., Kimmig, A., & Toivonen, H. (2007). ProbLog: A probabilistic Prolog and its application in link discovery. IJCAI, p 2468.

Riguzzi, F. 2022. Foundations of probabilistic logic programming. River Publishers. New York

Bezou-Vrakatseli E. 2023. Evaluation of LLM Reasoning via Argument Schemes. Online Handbook of Argumentation for AI, Vol.4 p 1

Louis A and Nenkova A. 2012. A Coherence Model Based on .Syntactic Patterns. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1157–1168, Jeju Island, Korea. Association for Computational Linguistics.

Freedman, G.; Dejl, A.; Gorur, D.; Yin, X.; Rago, A.; and Toni, F. 2025. Argumentative Large Language Models for Explainable and Contestable Claim Verification. Proceedings of the AAAI Conference on Artificial Intelligence, 39(14): 14930–14939.

Rago, A.; Toni, F.; Aurisicchio, M.; and Baroni, P. 2016. Discontinuity-Free Decision Support with Quantitative Argumentation Debates. In Principles of Knowledge Represen tation and Reasoning: Proceedings of the Fifteenth Interna tional Conference, KR 2016, 63–73.

Ng MP and Junqi Jiang and Gabriel Freedman and Antonio Rago and Francesca Toni. MArgE: Meshing Argumentative Evidence from Multiple Large Language Models for Justifiable Claim Verification, arxiv 2508.02584

Arcuschin, I.; Janiak, J.; Krzyzanowski, R.; Rajamanoharan, S.; Nanda, N.; and Conmy, A. 2025. Chain-of-thought reasoning in the wild is not always faithful. ICLR 2025 Reasoning and Planning for LLMs Workshop.

Barez, F.; Wu, T.-Y.; Arcuschin, I.; Lan, M.; Wang, V.; Siegel, N.; Collignon, N.; Neo, C.; Lee, I.; Paren, A.; et al. 2025. Chain-of-thought is not explainability. Preprint, arXiv.

Lambert, N.; Pyatkin, V.; Morrison, J.; Miranda, L. J. V.; Lin, B. Y.; Chandu, K. R.; Dziri, N.; Kumar, S.; Zick, T.; Choi, Y.; Smith, N. A.; and Hajishirzi, H. 2025. Reward Bench: Evaluating Reward Models for Language Modeling. In Findings of the Association for Computational Linguistics: NAACL 2025, 1755–1797.

Kaminski, R and Wankov P. 2017. A Tutorial on Hybrid Answer Set Solving with clingo. In: Reasoning Web. Semantic Interoperability on the Web (pp.167-203)

Garcia, A. and Simari, G. 2004. Defeasible logic programming: an argumentative approach. Theory Pract. Log. Program. 4, 95–138.

Ferrag MA, Norbert Tihanyi, Merouane Debbah, Reasoning beyond limits: Advances and open problems for LLMs, ICT Express, 2025.

Galitsky B. 2025. Enabling large language model with plug-and-play symbolic reasoning components. In Health Apps of Neuro-symbolic AI, Elsevier. pp 59-80.

Galitsky B. 2026a. Discourse based argumentation analysis for LLM verification. CMNA workshop, 31-41

Galitsky, B. 2026b. An Information–Theoretic Model of Abduction for Detecting Hallucinations in Explanations" Entropy. 28, no. 2: 173. https://doi.org/10.3390/e28020173

Lin S, Hilton J., and Owain E. 2021. TruthfulQA: Measuring how models mimic human falsehoods. CoRR, abs/2109.07958.

Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, 2019. PubmedQA A dataset for biomedical research question answering.

Jin D, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. CoRR, abs/2009.13081

Hong S, Xiao L, Zhang X, Chen J. 2024. ArgMed-Agents: Explainable Clinical Decision Reasoning with LLM Discussion via Argumentation Schemes

Camburu O-M, Rocktäschel T, Lukasiewicz T, Blunsom P. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. Advances in Neural Information Processing Systems 31

Mellgren N, Schneider-Kamp P, and Galke Poech L. 2025. Training Language Models to Use Prolog as a Tool. arXiv:2512.07407

Yang S., Li X., Cui L., Bing L., and Lam W. 2025. Neuro-Symbolic Integration Brings Causal and Reliable Reasoning Proofs. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 5747–5759, Albuquerque, New Mexi-co. ACL.

Yang X., Chen B., and Tam Y.-C. 2024. Arithmetic Reasoning with LLM: Prolog Generation & Permutation. In Proceedings of the 2024 Conference of the North American Chapter of the Asso-ciation for Computational Linguistics: Human Language Technolo-gies (Volume 2: Short Papers), pages 699–710, Mexico City, Mexico. ACL

Tan, X., Deng Y., Qiu X., Xu W., Qu C, Chu W, Xu Y. and Qi Y. 2024. Thought-Like-Pro: Enhancing Reasoning of Large Language Models through Self-Driven Prolog-based Chain-of-Though." ArXiv abs/2407.14562

Tan X, Li B., Xu W, Qu C, Chu W, Xu Y, Qi Y, and Qiu X. 2025. Prolog-Driven Rule-Based Diagnostics with Large Lan-guage Models for Precise Clinical Decision Support. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference, Daejeon, South Korea, Sep-tember 23–27, 2025, Proceedings, Part X. Springer-Verlag, Berlin, Heidelberg, 413–423.

Quan X, Valentino M, Carvalho D, Dalal D, and Freitas A. 2025. PEIRCE: Unifying material and formal reasoning via LLM-driven neuro-symbolic refinement. ACL System Demonstrations, pages 11–21.