

Can LLMs Abduce over Ontologies? A First Experience Report

Júlia Pukancová*, Martin Homola and Janka Boborová

Comenius University in Bratislava, Mlynská dolina, 842 41 Bratislava, Slovakia

Abstract

Abduction plays a foundational role in explanation-oriented reasoning across medicine, law, and other high-stakes domains, and recent work suggests that combining large language models (LLMs) with ontologies, knowledge graphs, and other structured knowledge resources may enable more interpretable and context-sensitive decision-support systems. Yet it remains unclear whether current LLMs can themselves perform the abductive reasoning required in such settings, or whether robust abductive inference must still be supplied by dedicated formal reasoning components. In this paper, we investigate this question through a first experience report on ontology-based ABox abduction. We evaluate several contemporary LLM-based systems on abductive reasoning tasks over ontologies of different complexity and compare their outputs against explanations computed by sound and complete symbolic abductive reasoners. Our results show that, while stronger cloud-based models can often recover a number of valid explanations, exact ontology-based abduction remains challenging, ontology- and prompt-sensitive, and error-prone for current LLMs. The main lesson is that LLMs are promising as generators and verbalizers of candidate explanations, but cannot yet be safely entrusted with ontology-based abduction without explicit symbolic verification.

Keywords

Abduction, ontologies, LLMs, eXplainable AI

1. Introduction and Motivation

Abduction is traditionally understood as a distinct mode of inference alongside deduction and induction. In Peirce’s formulation, abduction is the inferential step by which a reasoner moves from a surprising or incomplete observation to a plausible explanatory hypothesis—an act that initiates inquiry rather than simply verifies what is already known [1]. In this classical triad, (i) deduction derives consequences from hypotheses, (ii) induction generalized and evaluates consequences against experience, and (iii) abduction supplies the explanatory conjectures that make inquiry possible in the first place [1, 2]. Let us illustrate it by a simple example from a medical domain: (i) If we know that a patient has sensitive teeth and that sensitive teeth may cause toothache, we may deductively infer that toothache may likely occur (especially after a cold drink); (ii) From many observed cases, we may inductively learn the rule, that for patients with sensitive teeth having a cold drink is likely to cause toothache; (iii) If the said rule is accepted and a patient reports toothache, we may abductively hypothesise that it is plausible that they have sensitive teeth and have had a cold drink (among other plausible causes derived from other rules in our knowledge).

In philosophy of science, this Peircean insight was later developed into the now influential idea of *Inference to the Best Explanation* (IBE). Harman argued that many warranted non-deductive inferences are best understood as selecting the hypothesis that would provide the best explanation of the evidence, rather than as instances of simple enumerative induction [3]. Lipton subsequently elaborated this framework and made explanation quality central to theory choice, thereby reinforcing the view that explanatory considerations are not peripheral but constitutive of rational non-deductive inference [4].

Proceedings of the Joint Ontology Workshops (JOWO) - Episode XII: The Tropical Spring of Ontology, co-located with the 16th International Conference on Formal Ontology in Information Systems (FOIS 2026), September 21-22, 2026, Vitória (ES), Brazil

*Corresponding author.

✉ pukancova@fmph.uniba.sk (J. Pukancová); homola@fmph.uniba.sk (M. Homola); boborova@fmph.uniba.sk (J. Boborová)

🌐 <https://dai.fmph.uniba.sk/~pukancova/> (J. Pukancová); <https://dai.fmph.uniba.sk/~homola/> (M. Homola)

🆔 0009-0001-3505-0716 (J. Pukancová); 0000-0001-6384-9771 (M. Homola); 0009-0002-7487-9962 (J. Boborová)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The relevance of abduction is especially clear in high-stakes domains whose central cognitive task is to explain incomplete, noisy, or evolving evidence. In clinical diagnostics, abduction is recognised a critical component of medical reasoning, particularly for hypothesis generation in complex or atypical cases that do not fit routine protocols [5, 6]. In legal reasoning, abductive hypothesis formation is recognized as important for articulating arguments, interpreting norms, and developing legal theories [7, 8]. Other applications include e.g. system and fault diagnosis [9] and forensic investigation [10].

Because of this explanatory role, abduction has long been investigated not only philosophically but also computationally in knowledge representation and reasoning. In connections to ontologies and knowledge graphs it has been studied under the formal framework of description logics [11]. Notably *ABox abduction*, the reasoning problem focusing on extensional observations and explanations, with applications to diagnosis, ontology debugging, and other explanation-centred tasks [11, 12, 13, 14, 15, 16]. Dedicated systems such as the AAA [17], LETHE [18] or CATS [19], have been developed to compute explanations over expressive ontologies. More recently, abductive reasoning has also been studied over knowledge graphs, where the task is to generate logical hypotheses that explain a set of observed entities or facts [20]. This line of work has produced methods for complex logical hypothesis generation, controllable abductive hypothesis generation, and interactive refinement over graph-structured knowledge [20, 21].

These developments are now becoming relevant in the context of *LLM-based decision support systems*. As large language models are increasingly explored for healthcare, law, finance, and other critical societal domains, recent work has argued that their usefulness depends not merely on fluent language generation, but on their ability to provide explanations, justify conclusions, and operate safely under uncertainty [22]. In this emerging discussion, abduction has been identified as a promising mechanism for hypothesis generation, explanation, and trust calibration, especially because it offers a structured way to distinguish observed facts from inferred assumptions but as well plausible hypotheses from weakly supported hallucinations [23]. At the same time, recent analyses of LLM reasoning stress that current models often produce plausible but weakly grounded explanatory narratives, motivating greater attention to abductive structure as a criterion of trustworthiness [23, 24].

Recently it has been proposed to use abductive reasoning as a way to strengthen, verify, or correct LLM outputs. In particular, abductive logic has been used to frame hallucination not simply as factual error, but as a failure of explanatory support: an answer may be fluent and coherent while still relying on missing premises, unsupported intermediate steps, or hypotheses that collapse under comparison with better alternatives [23]. From this perspective, abduction is valuable not only for producing explanations but also for checking whether those explanations are genuinely warranted, making it directly relevant to correctness verification, hallucination detection, and the broader trustworthiness of LLM-supported decision-making [23].

Further more, such an approach may be reinforced by coupling abductive reasoning with external structured knowledge resources such as ontologies and knowledge graphs. Here, the goal is not merely to ask an LLM for an explanation, but to ground evolving user intent in formal semantic conditions, generate hypotheses that can be executed over structured knowledge, and refine failed hypotheses through explicit diagnosis and repair [21]. Such approaches suggest that abductive reasoning becomes more actionable, interpretable, and controllable when it is constrained by external knowledge structures rather than left to unconstrained language generation alone. In this sense, abduction forms a conceptual bridge between LLMs and formal knowledge resources in emerging decision-support architectures [21].

These developments motivate the question addressed in this paper. If abduction is foundational to explanation-oriented reasoning in medicine, law, forensics, and system diagnosis, and if ontologies and knowledge graphs provide the structured background knowledge over which such reasoning can be made explicit, then *an essential unresolved issue is whether contemporary LLMs can themselves perform the abductive work required in these settings*. Put differently, it remains unclear whether LLM-based tools can be entrusted with abduction over structured knowledge, or whether robust abductive inference must still be supplied by dedicated formal reasoning components and only complemented by LLMs.

In this paper, we investigate to what extent contemporary LLMs can perform ontology-based abduction, and how their behaviour compares to formal symbolic abductive reasoners. To this end, we

evaluate a set of LLM-based systems on several ABox abduction tasks over ontologies of different complexity and compare their outputs with explanations computed by sound and complete symbolic baselines. Our aim is not only to assess whether LLMs can generate plausible explanations, but to determine whether these explanations satisfy the formal requirements of proper abduction, and thus whether LLMs can be safely relied upon in structured decision-support settings.

We show that while stronger cloud-based models are often able to recover some valid abductive explanations, ontology-based abduction remains difficult for current LLMs and is highly sensitive to prompting, task structure, and the complexity of the ontology. The main lesson is therefore not that LLMs should replace formal abductive tools, but that they may be useful as front-end generators of candidate hypotheses, provided that these hypotheses are subsequently checked and grounded by explicit symbolic verification.

2. Illustrative Use Case

To illustrate the distinction between plausible language-model output and proper abductive reasoning, we use the following toy diagnostic scenario as a running example. A clinician consults an LLM-based diagnostic tool and enters the patient’s primary complaint:

John reports toothache.

The tool returns several candidate diagnoses together with short natural-language justifications. For instance, it may propose:

1. **Dental trauma:** *A toothache may be caused by dental trauma.*
2. **Sensitive teeth after a cold drink:** *Sensitive teeth may cause pain when a person drinks a cold drink.*
3. **Swollen gums:** *People with swollen gums often take painkillers to manage oral discomfort. Toothache also causes oral discomfort. If John may have taken a painkiller, he may also have swollen gums.*

Suppose the interaction then continues with an additional follow-up question from the tool:

Has John taken a painkiller?

John answers *yes*. The tool then strengthens its third hypothesis and presents *swollen gums* as the most likely explanation of the toothache. From the user’s perspective, this conclusion may seem reasonable: the tool has not only produced a candidate diagnosis, but also supplied a short textual rationale that appears to support it. However, we have no formal guarantee that the justification is valid and that the obtained explanations are really the most plausible ones given the available knowledge and the provided justifications.

Abductive verification. To verify the suggested diagnoses formally, we now perform knowledge-extraction and we subject the explanations to abductive verification [23]. We easily encode the explanations as description-logic axioms:

$$\begin{aligned}
 \text{DentalTrauma} &\sqsubseteq \text{Toothache}, \\
 \text{DrankColdDrink} \sqcap \text{SensitiveTeeth} &\sqsubseteq \text{Toothache}, \\
 \text{SwollenGums} &\sqsubseteq \text{OralDiscomfort}, \\
 \text{Toothache} &\sqsubseteq \text{OralDiscomfort}, \\
 \text{OralDiscomfort} &\sqsubseteq \text{TookPainkiller}.
 \end{aligned}$$

These axioms define a small ontology $\mathcal{O}$ that captures the reasoning structure implicit in the LLM’s own explanations. We now use ABox abduction to test whether the proposed candidate diagnoses genuinely explain the original complaint

$$\mathcal{O} = \{\text{Toothache}(\text{John})\}.$$

The three hypotheses suggested by the tool are represented as

$$E_1 = \{\text{DentalTrauma}(\text{John})\},$$

$$E_2 = \{\text{DrankColdDrink}(\text{John}), \text{SensitiveTeeth}(\text{John})\},$$

$$E_3 = \{\text{SwollenGums}(\text{John})\}.$$

Under formal abduction, a candidate hypothesis E is acceptable only if

$$\mathcal{O} \cup E \models \text{Toothache}(\text{John}).$$

With this criterion, the first two candidates are verified as genuine explanations. From $\text{DentalTrauma}(\text{John})$ and $\text{DentalTrauma} \sqsubseteq \text{Toothache}$, the observation follows immediately. Likewise, from $\text{DrankColdDrink}(\text{John})$ and $\text{SensitiveTeeth}(\text{John})$, together with $\text{DrankColdDrink} \sqcap \text{SensitiveTeeth} \sqsubseteq \text{Toothache}$, the observation is also entailed.

The third candidate, however, is ruled out. From $\text{SwollenGums}(\text{John})$ one may derive only $\text{TookPainkiller}(\text{John})$, but not $\text{Toothache}(\text{John})$. In other words, the explanation provided by the LLM relies on a semantically plausible association—both toothache and swollen gums are linked to painkiller use—but this does not amount to an abductive explanation of the complaint. The model has *drifted* from a shared consequence to a supposed cause. This is an example of a well documented issue in LLM reasoning known as semantic drift (or reasoning drift) [25, 9, 23].

Refinement through external knowledge. Following another relevant stream of research [21], we now show, how the explanations may be further refined by incorporating external domain knowledge from structured medical resources. Such knowledge may often come from ontologies (such as SNOMED-CT) or publicly available knowledge graphs, or it can be processed into this from clinical guidelines and other available medical resources. Suppose we add the DL axiom:

$$\text{Cavity} \sqsubseteq \text{SensitiveTeeth}.$$

This axiom is a simplified formalization of a clinically well-supported pattern: dental caries may expose dentin and thereby lead to sensitivity to hot, cold, or sweet stimuli. With this additional axiom, the earlier validated explanation E_2 can be refined into the more specific hypothesis E_4 :

$$\{\text{DrankColdDrink}(\text{John}), \text{Cavity}(\text{John})\}.$$

To obtain the respective human-readable representation, we may either refine the one respective to E_2 from which E_4 was obtained, or we may input the newly added axiom to the LLM and ask for a suitable justification. We obtain:

4. **Cavity after a cold drink:** Cavities can make teeth sensitive. Sensitive teeth may cause pain when a person drinks a cold drink. If John recently drank a cold drink, a cavity explains the toothache.

This explanation (and its justification) indeed verifies abductively: from $\text{Cavity}(\text{John})$ we derive $\text{SensitiveTeeth}(\text{John})$, and together with $\text{DrankColdDrink}(\text{John})$ the observation $\text{Toothache}(\text{John})$ still follows. This second step illustrates that external ontologies or knowledge graphs need not only serve as a verification layer [26, 27] for LLM-generated explanations; they can also strengthen the analysis by

refining generic symptom-level hypotheses into deeper and more informative explanatory candidates [21].

Altogether we have yielded three abductively verified explanations E_1 , E_2 , and E_4 . There are different options how to possibly compare and narrow them down. For instance the patient may be asked if an incident occurred that would support or rule out E_1 (dental trauma). E_2 and E_4 are related and they can either be compared using the established theoretical background on abduction: E_2 is semantically weaker than E_4 [11, 12], i.e. E_2 hypothesises less than E_4 and thus, if no external information is available, it is most typically preferred [12]. However, in this particular case, as potential cavity is a more specific diagnosis potentially requiring medical attention it may be assigned higher importance and the patient may be advised to consult a dentist. Note that comparison of abductively plausible diagnoses is not in the central focus of our paper and we include it as a side note.

We have demonstrated, that ontology-based abductive reasoning is a highly relevant tool for verification and grounding of LLM-based diagnostic explanations. In the reminder of the paper we look into the problem if it needs to be outsourced to external abductive reasoners, external to LLMs, or if at least to certain extent LLMs can be directly used also for this task.

3. Methodology

In order to test the abductive capabilities the LLMs are able to exert over ontological knowledge, we have created simple ontology-based inputs for ABox abduction problems and we have tested the performance of multiple LLMs against a formal sound and complete DL ABox abduction reasoner.

3.1. Benchmark

We have used three ontologies and multiple inputs in order to cover a range of domains and introduce several difficulty levels.

- The *toothache* ontology represents a small, synthetic ontology that captures one clear diagnostic scenario with simple explanation structure. It is a simplified version of the ontology $\mathcal{O}$ from Section 2 with three axioms. We have tested two distinct input observations yielding use cases *toothache* and *toothache_alice*.
- The *family* ontology introduces richer relational structure and role-based reasoning, making explanation generation more complex. We have paired it with one input observation.
- The *lubm* instances are derived from the standard Leigh University Benchmark (LUBM). It exhibits a more realistic and structurally complex knowledge bases. We have paired it with two observations yielding inputs *lubm-0_1_0* and *lubm-0_1_1*.

Together, these tasks allow us to assess model performance across varying levels of reasoning complexity, including the ability to generate minimal explanations, handle multiple competing hypotheses, and operate over both simple and structured ontological domains.

3.2. Symbolic Baseline

The idea of the symbolic baseline was to capture all explanations according to the definition of abduction: An ABox abduction problem is a pair $\mathcal{P} = (\mathcal{K}, O)$, where $\mathcal{K}$ is DL knowledge base and O (the observation) is a set of facts, such that $\mathcal{K} \not\models O$. An explanation E is a set of facts (ABox assertions) such that $\mathcal{K} \cup E \models O$. As usual in ABox abduction, we will assume that explanations are subset-minimal, consistent ($\mathcal{K} \cup E$ is consistent), and relevant ($E \not\models O$). (This rules out some counterintuitive explanations [11].)

As a symbolic reference systems, we used two description-logic-based abductive reasoning systems: CATS¹ [19] and AAA² [17]. CATS implements more efficient hybrid algorithms, but it currently does

¹<https://github.com/Comenius-Abduction-Team/CATS-Abduction-Solver>

²<https://dai.fmph.uniba.sk/~pukancova/aaa/>

Table 1

Evaluated models: local models were executed via Ollama; cloud models were accessed through their respective chat interfaces

Category	Model	Description
Cloud	Copilot	GPT 5.4 Thinking (via Microsoft Copilot Think Deeper mode)
	Claude	Anthropic Claude Sonnet 4.6 (low effort)
	Gemini	Google Gemini 3.5 Flash
Local	Gemma 3	Google Gemma 3 4B, Q4_K_M quantized, 131K context, served via Ollama
	Mistral	Mistral 7B, Q4_K_M quantized, 32K context, served via Ollama

not support role explanations (explanations that include relations between two individuals). When role explanations were also considered in comparisons, AAA was used instead.

Multiple distinct explanation sets may exist for a single task. Given an ontology and a target observation, these systems compute the set of all subset-minimal, relevant, and consistent abductive explanations according to the definition above. Each explanation is represented as a set of atomic assertions. These explanations are treated as the ground-truth solution space in our evaluation.

3.3. LLM Models

We evaluated two local open-weight models (Gemma 3 and Mistral) and three cloud-based systems (Claude, Gemini, Copilot), as summarized in Table 1. We intentionally did not enforce identical reasoning settings across all systems. Copilot (Think Deeper) was evaluated in a more advanced reasoning mode, whereas Claude (Low Effort) and Gemini (Flash) were used in more lightweight configurations, in order to reflect a wider pool of realistic usage scenarios and to capture variability in model behaviour across different levels of reasoning effort.

The inclusion of local models serves to assess the capabilities of fully locally deployable systems under realistic resource constraints; however, due to their smaller scale and open-weight nature, these models are expected to exhibit performance limitations compared to state-of-the-art cloud-based systems.

3.4. Prompt Variants

To study prompt sensitivity, we compared three prompt configurations, denoted **Prompt 1**, **Prompt 2**, and **Prompt 3**. These prompts are designed for the same task, but they differ in the way how the instructions are specified and the level of details supplied.

Prompt 1 represents a relatively balanced baseline, combining a clearly defined explanation structure with moderate output constraints, including a limit on the number of generated explanations. Prompt 2 introduces stronger constraints on symbol usage and output validity, explicitly prohibiting the use of symbols not present in the ontology and encouraging the generation of as many valid explanations as possible. Prompt 3 represents the most restrictive setting, emphasizing strict adherence to ontology symbols and output format, while adopting a more conservative generation strategy that discourages speculative outputs.

Overall, the prompts differ along three main dimensions: (i) strictness of output constraints, (ii) enforcement of symbol grounding in the ontology, and (iii) the trade-off between completeness and conservativeness in explanation generation. Refer to Appendix A for more prompt details.

In each experiment, each of these prompts was passed to each of the five models equally (i.e. without any model-specific modifications).

3.5. Evaluation Metrics

We report the following metrics:

- **Recall**: the fraction of CATS explanations recovered by the model.

- **Precision:** the fraction of model-generated explanations that exactly match a CATS explanation.
- **F1:** overall balanced performance, given as the harmonic mean of precision and recall.
- **Full match:** a binary indicator showing whether the model recovered all CATS explanations for a given task.

Formally, if E denotes the *ground truth* explanations set returned by the symbolic solver, and E_{LLM} denotes the set of explanations generated by an LLM model then:

$$\text{Precision} = \frac{|E_{LLM} \cap E|}{|E_{LLM}|} \quad (1)$$

$$\text{Recall} = \frac{|E_{LLM} \cap E|}{|E|} \quad (2)$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

$$\text{FullMatch} = \begin{cases} 1 & \text{if Recall} = 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Undefined F1 values (e.g., due to zero precision and recall or empty outputs) were treated as zero during aggregation, reflecting complete failure to produce a valid explanation.

3.6. Evaluation Setting and Output Post-processing

Because we were not able to obtain role explanations from the more efficient CATS solver, the main quantitative evaluation was conducted in a role-free setting. LLMs readily returned role assertions in explanations, therefore all explanation sets were post-processed and those containing at least one role assertion were discarded entirely before scoring. This ensured that all models were compared in the same explanation space and prevented a methodological mismatch between role-enabled and role-disabled hypothesis sets.

We subsequently conducted a more involved qualitative evaluation on the *family* ontology which is reported separately. Here the symbolic baseline was computed using AAA and therefore all outputs, including those with role explanations were included in the comparison.

In addition, model outputs were post-processed prior to evaluation to ensure compatibility with the CATS representation. In particular, auxiliary prefixes and formatting artefacts introduced by the LLMs were removed, and outputs were parsed into sets of atomic assertions. This normalization step ensured that comparisons were performed on a consistent symbolic representation across all systems.

4. Experiments and Results

This section presents the results of the quantitative evaluation, including the main model comparison and the effect of prompt design. In addition, we provide a qualitative case study on the *family* ontology to illustrate model behaviour in a role-enabled setting, which is not directly comparable to the main role-free evaluation.

4.1. Main Quantitative Comparison

Table 2 summarizes the main quantitative results across all models, macro-averaged over all benchmark tasks and prompt conditions. Overall, the cloud-based models substantially outperformed the local open-weight models, with Copilot achieving the strongest overall performance across the evaluated metrics.

Table 2

Macro-averaged quantitative comparison across models

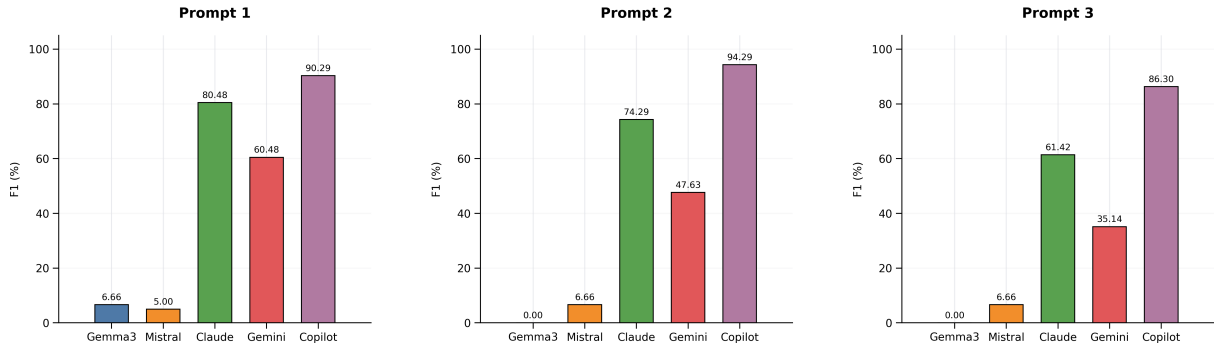
Model	Recall (%)	Precision (%)	F1 (%)	Full Match
Gemma 3	2.22	2.22	2.22	0.00
Mistral	6.66	5.77	6.11	0.00
Claude	81.11	66.67	72.06	0.73
Gemini	51.11	45.56	47.75	0.33
Copilot	97.78	85.56	90.29	0.93

Effect of Prompt Design. To evaluate prompt sensitivity, we compared Prompts 1, 2, and 3. Figure 1 shows the average F1 scores for all models under each prompt.

The results indicate that prompt design has a substantial impact on explanation quality, and that this effect is strongly model-dependent. Claude and Gemini achieve their best results with Prompt 1, whereas Copilot performs best with Prompt 2. In contrast, Prompt 3 consistently leads to lower performance across the cloud-based models.

For the local open-weight models (Gemma 3 and Mistral), performance remains consistently low across all prompt configurations, with only limited variation. In particular, these models frequently fail to generate valid explanation sets or produce outputs with zero overlap with the symbolic baseline, which reduces the impact of prompt design relative to the cloud-based systems.

Overall, Copilot appears to be the most robust model across prompt variants, maintaining high performance regardless of prompt configuration, whereas Claude and Gemini are more sensitive to prompt design and exhibit larger performance variations.

**Figure 1:** Average F1 scores in the role-free evaluation setting for Prompts 1, 2, and 3

Task-Level Behaviour. Task-level analysis revealed that some ontology-observation pairs were substantially easier than others. In particular, the *toothache* and *lubm-0_1_0* cases were handled much more reliably by the cloud models, whereas the *family* case and the role-sensitive variants proved more difficult. This suggests that explanation complexity and the structure of the available hypothesis space strongly affect model behaviour.

Error Analysis. Although cloud models often produced plausible explanations, many errors arose from *near misses* rather than fully irrelevant outputs. We observed several recurring failure modes:

- **Extra assertions:** explanation sets that were close to a correct CATS explanation but included one additional unsupported assertion.
- **Missing assertions:** explanation sets that were close to a correct CATS explanation but miss one assertion.
- **Malformed or weakly structured outputs:** especially frequent for local open-weight models.

Table 3

Illustrative qualitative comparison of model-generated explanation sets against the AAA baseline on a representative benchmark case (explanations matching the AAA baseline are shown in bold)

System	Example output
AAA	{Father(jack), Female(jane), hasChild(jane,jack)}; {Grandmother(jane)}; {hasChild(jack,jane), Female(jane), hasChild(jane,jack)}; {Person(jack), Female(jane), hasChild(jane,jack)}; {Female(jack), Female(jane), hasChild(jane,jack)}; {Grandfather(jack), Female(jane), hasChild(jane,jack)}; {Mother(jack), Female(jane), hasChild(jane,jack)}; {Grandmother(jack), Female(jane), hasChild(jane,jack)}; {Female(jane), Parent(jane)}; {Male(jack), Female(jane), hasChild(jane,jack)}; {Parent(jack), Female(jane), hasChild(jane,jack)}
Claude	{Female(jane), hasChild(jane,jack), Person(jack)} ; {Female(jane), hasChild(jane,jack), Male(jack)} ; {Female(jane), hasChild(jane,jack), Female(jack)} ; {Female(jane), hasChild(jane,jack), Parent(jack), hasChild(jack,jack)}
Gemini	{Female(jane), hasChild(jane,jack), Person(jack)} ; {Female(jane), hasChild(jane,child1), Person(child1)}
Copilot	{Grandmother(jane)} ; {Female(jane), Parent(jane)} ; {Female(jane), hasChild(jane,jack), Person(jack)}

Table 4

Metrics for illustrative comparison across models

Model	Recall (%)	Precision (%)	F1 (%)
Claude	27.30	75.00	40.00
Gemini	9.10	50.00	15.40
Copilot	27.30	100.00	42.90

These findings suggest that strict exact-set matching may undervalue explanations that are semantically related to the symbolic solution, but they also confirm that unconstrained generation introduces substantial noise.

4.2. Qualitative Case Study

To complement the aggregate metrics, we inspected representative examples manually. The *family* case is particularly informative because it features a semantically rich ontology, including role relations. At the same time the ontology is small enough, which enabled to use AAA to compute also role explanations. This setting enables to test the reasoning limitations of the LLM models and it exposes their tendency to generate semantically related but non-identical hypotheses.

We have included only the cloud models in this comparison and we have used Prompt 2, which was previously the most successful one. The explanations were computed for the *family* ontology and the observation Mother(jane). Table 3 shows the comparison between expected explanations (obtained by AAA) and the results obtained from the three models.

We observe distinct error patterns across models. Claude often produces near-correct explanations but includes additional unsupported assertions. Gemini tends to generate semantically plausible but structurally incorrect explanations, e.g. by introducing a non-existing individual. In contrast, Copilot produces fewer explanations, but these are typically well-formed and closely aligned with the symbolic baseline.

Notably, only a smaller subset of all syntactically minimal explanations (as computed by AAA) were returned by all three LLMs. On the one hand, this may point out the models’ general weakness in more

complex reasoning that requires to combine multiple axioms, on the other hand, especially the output of Copilot is heavily centered on semantically minimal explanations, which may actually be relevant in many use cases. This, however, requires further investigation.

For completeness Table 4 adds the quantitative measures obtained from Table 3. Here, Copilot achieves the highest precision, correctly producing only valid explanation sets. Claude achieves the same recall as Copilot at the cost of occasional overgeneration. Gemini exhibits both lower recall and precision, primarily due to hallucinated entities.

5. Limitations of the Study

We acknowledge that this is only a preliminary explorative study with a number of limitations. First, the benchmark contains only a small number of tasks, and limited particular configurations and prompts, which limits the statistical generality of the study. Second, role explanations could not be computed consistently across all symbolic runs, which motivated the overall role-free evaluation protocol for the main qualitative comparison. Third, exact-set matching is a strict metric that may assign zero score to explanations that are semantically related but not identical to the symbolic baseline – a more fine-grained comparison may bring deeper insights. Finally, the selection of models was not reasonably complete and balanced, more models need to be included in the future. Also, cloud models were evaluated through interactive interfaces rather than a fully controlled API pipeline, which may introduce minor variability across runs. Overall, these results are preliminary. The conclusions apply only to the specific models, configurations, and prompts tested under these limited conditions, rather than generally. This study is exploratory, and more rigorous findings are left for future work.

6. Discussion

The experiments show that ontology-based abductive explanation generation remains challenging for LLMs. The results indicate that performance significantly varies across models and also depends heavily on prompt design. As expected, cloud models clearly outperform lightweight local open-weight baselines, but their behaviour is still sensitive to evaluation protocol and task structure.

Among the cloud-based models, Copilot achieved the strongest overall performance, combining the highest recall, precision, F1, and the best robustness across prompt variants. Claude also performed well and often recovered a large portion of the valid explanation space, while Gemini showed clearly weaker and less stable behaviour across tasks and prompts. The local models rarely produced usable exact-set explanations, which suggests that direct ontology-based abduction remains difficult for smaller local models under raw ontology input. We conjecture (given our very limited qualitative evaluation) that even for the more powerful cloud-based models explanations that involved more complex semantic relations and higher numbers of axioms were difficult to find.

Overall, our findings show that LLMs can be useful for proposing and verbalizing candidate hypotheses, but robust abductive reasoning over ontologies still benefits from, and in many cases requires, explicit symbolic verification. Future work should expand the benchmark space, include richer ontologies and knowledge graphs, evaluate the outcomes also w.r.t. semantic minimality (e.g. using the LETHE [18] solver for ground truth), include richer and more balanced selection of LLM models, and investigate more systematically which abductive subtasks can be delegated to LLMs and which should remain under the control of formal reasoners.

Acknowledgments

The authors wish to thank to Ján Klůka for help with L^AT_EX. This research was sponsored by Slovak Republic under the SRDA grant APVV-23-0292 (DyMAX) and VEGA grant no. 1/0630/25 (eXSec).

Declaration on Generative AI

During the preparation of this work, the authors used Microsoft Copilot with GPT-5 and Google Gemini 3.5 for drafting content, content enhancement, text translation, paraphrasing and rewording, improving writing style, grammar and spelling check, formatting assistance, citation management, and generating of images (plotting the charts from supplied data). All substantive scientific content, argumentation, literature interpretation, and conclusions were written, verified, and approved by the authors. After using these tools, the authors reviewed and edited the content as needed, and take full responsibility for the publication's content.

References

- [1] C. S. Peirce, Illustrations of the logic of science VI: Deduction, induction, and hypothesis, *Popular Science Monthly* 13 (1878) 470–482.
- [2] I. Niiniluoto, Peirce's abduction and its interpretations, in: *The Oxford Handbook of Charles S. Peirce*, Oxford University Press, 2024. doi:10.1093/oxfordhb/9780197548561.013.19.
- [3] G. Harman, The inference to the best explanation, *The Philosophical Review* 74 (1965) 88–95. doi:10.2307/2183532.
- [4] P. Lipton, *Inference to the Best Explanation*, 2 ed., Routledge, London and New York, 2004.
- [5] D. E. Stanley, A.-V. Pietarinen, The logic of medical reasoning: Toward an integrated inductive, deductive, and abductive approach to clinical practices, *Philosophy, Ethics, and Humanities in Medicine* 20 (2025) 1–9. doi:10.1186/s13010-025-00178-y.
- [6] C. Martini, Abductive reasoning in clinical diagnostics, in: L. Magnani (Ed.), *Handbook of Abductive Cognition*, Springer, Cham, 2023, pp. 467–479. doi:10.1007/978-3-031-10135-9_13.
- [7] H.-T. Nguyen, R. Goebel, F. Toni, K. Stathis, K. Satoh, How well do SOTA legal reasoning models support abductive reasoning?, in: *Proceedings of the Workshop on Logic Programming and Legal Reasoning (LPLR 2023)*, 2023. URL: <https://ceur-ws.org/Vol-3437/paper1LPLR.pdf>.
- [8] B. Askeland, The potential of abductive legal reasoning, *Ratio Juris* 33 (2020) 66–81. doi:10.1111/raju.12268.
- [9] P. Zhang, Z. Liu, N. Zhou, H. Liu, Fault diagnosis for smart grids over knowledge graph and abductive reasoning, in: *Proceedings of the 2024 4th Power System and Green Energy Conference (PSGEC)*, 2024, pp. 1003–1008.
- [10] E. Kruger, Counterfeit drug investigations: Techniques, challenges, and the role of abductive reasoning, *Perspectives in Legal and Forensic Sciences* 2 (2025) 10005. doi:10.70322/plfs.2025.10005.
- [11] C. Elsenbroich, O. Kutz, U. Sattler, A case for abductive reasoning over ontologies, in: *Proceedings of the OWLED*06 Workshop on OWL: Experiences and Directions*, Athens, GA, US, volume 216 of *CEUR-WS*, CEUR-WS.org, 2006. URL: https://ceur-ws.org/Vol-216/submission_25.pdf.
- [12] J. Pukancová, M. Homola, Abductive reasoning with description logics: Use case in medical diagnosis, in: *Proceedings of the 28th International Workshop on Description Logics (DL 2015)*, volume 1350 of *CEUR Workshop Proceedings*, 2015. URL: <https://ceur-ws.org/Vol-1350/paper-60.pdf>.
- [13] M. Obeid, Z. Obeid, A. Moubaidin, N. Obeid, Using description logic and ABox abduction to capture medical diagnosis, in: *Advances and Trends in Artificial Intelligence. From Theory to Practice*, volume 11606 of *Lecture Notes in Artificial Intelligence*, Springer, 2019, pp. 376–388. doi:10.1007/978-3-030-22999-3_33.
- [14] S. M. Rashid, J. McCusker, D. Gruen, O. Seneviratne, D. L. McGuinness, A concise ontology to support research on complex, multimodal clinical reasoning, in: *The Semantic Web: ESWC 2023*, volume 13870 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 390–407. doi:10.1007/978-3-031-33455-9_23.
- [15] P. Lambrix, F. Wei-Kleiner, Z. Dragisic, V. Ivanova, Repairing missing is-a structure in ontologies

- is an abductive reasoning problem, in: Proceedings of the Second International Workshop on Debugging Ontologies and Ontology Mappings (WoDOOM 2013), volume 999 of *CEUR Workshop Proceedings*, 2013, pp. 33–44. URL: <https://ceur-ws.org/Vol-999/paper3.pdf>.
- [16] F. Wei-Kleiner, Z. Dragisic, P. Lambrix, Abduction framework for repairing incomplete el ontologies: Complexity results and algorithms, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014. doi:10.1609/aaai.v28i1.8858.
 - [17] J. Pukancová, M. Homola, The AAA ABox abduction solver, *KI - Künstliche Intelligenz* 34 (2020) 503–507. doi:10.1007/s13218-020-00685-4.
 - [18] P. Koopmann, W. Del-Pinto, S. Tourret, R. A. Schmidt, Signature-based abduction for expressive description logics, in: Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, KR 2020, Rhodes, Greece, 2020, pp. 592–602. URL: <https://doi.org/10.24963/krr.2020/59>. doi:10.24963/KR.2020/59.
 - [19] J. Kloc, J. Boborová, M. Homola, J. Pukancová, CATS solver: The rise of hybrid abduction algorithms, in: L. Tendera, Y. Ibáñez-García, P. Koopmann (Eds.), Proceedings of the 38th International Workshop on Description Logics - DL 2025, Opole, Poland, September 3-6, 2025, volume 4091 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025. URL: <https://ceur-ws.org/Vol-4091/paper34.pdf>.
 - [20] J. Bai, Y. Wang, T. Zheng, Y. Guo, X. Liu, Y. Song, Advancing abductive reasoning in knowledge graphs through complex logical hypothesis generation, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 1312–1329. doi:10.18653/v1/2024.acl-long.72.
 - [21] Y. Gao, Y. Cai, T. Zheng, J. Bai, Y. Song, Hypoagent: An agentic framework for interactive abductive hypothesis generation over knowledge graphs, arXiv preprint arXiv:2605.31370, 2026. Preprint.
 - [22] Z. Z. Chen, J. Ma, X. Zhang, N. Hao, A. Yan, A. Nourbakhsh, X. Yang, J. McAuley, L. Petzold, W. Y. Wang, A survey on large language models for critical societal domains: Finance, healthcare, and law, *Transactions on Machine Learning Research* (2024). URL: <https://openreview.net/forum?id=upAWnMgpnH>.
 - [23] B. A. Galitsky, Tackling LLM hallucination with abductive reasoning, Preprints.org preprint, 2025. doi:10.20944/preprints202511.1688.v1, not peer reviewed.
 - [24] P. Song, P. Han, N. Goodman, A survey on large language model reasoning failures, OpenReview, AI4Math@ICML 2025, 2025. URL: <https://openreview.net/forum?id=hsgMn4KBFG>.
 - [25] A. Spataru, E. Hambro, E. Voita, N. Cancedda, Know when to stop: A study of semantic drift in text generation, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, 2024, pp. 3656–3671. doi:10.18653/v1/2024.naacl-long.202.
 - [26] S. Zhou, M. Huang, J. Cole, C. Britton, M. Yin, J. Wolber, K. Li, Assessing automated fact-checking for medical llm responses with knowledge graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, AAAI Press, 2026, pp. 39646–39656. doi:10.1609/aaai.v40i46.41317.
 - [27] Y. Sui, Y. He, N. Liu, X. He, K. Wang, B. Hooi, Fidelis: Faithful reasoning in large language models for knowledge graph question answering, in: Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, 2025, pp. 8315–8330. doi:10.18653/v1/2025.findings-acl.436.

A. LLM Prompts Documentation

Prompt 1

You are given an ontology and an observation.

The ontology may be written in OWL or RDF-style syntax.
Your task is to use the ontology to generate abductive explanations for the observation.

Definition of explanation:

- An explanation is a set of ABox assertions.
- All assertions inside one explanation must hold together.
- Different explanations are alternative explanations.

Required output format:

- Return only explanation sets.
- Write each explanation inside curly braces { }.
- Separate alternative explanations with ';'.
- Separate assertions inside one explanation with ','.

Allowed assertion forms:

- Concept(individual)
- Role(individual1, individual2)

Rules:

1. Use the ontology as the main source of knowledge.
2. Use only symbols that appear in the ontology or in the observation.
3. Do not add comments, numbering, or explanations.
4. Return at most 5 alternative explanations.
5. Return nothing except the explanation sets.

Example of the required format:

{A(a)} ; {B(a), R(a,b)} ; {C(a), D(a), S(a,c)}

Prompt 2

You are given an ontology and an observation.

The ontology may be written in OWL or RDF-style syntax.
Use the ontology as the main source of knowledge.

Your task is to generate abductive explanations for the observation.

Definition:

- An explanation is a set of ABox assertions.
- All assertions inside one explanation must hold together.
- Different explanations are alternative explanations.

Required output format:

- Return only explanation sets.
- Write each explanation inside curly braces { }.
- Separate alternative explanations with ';'.
- Separate assertions inside one explanation with ','.

Allowed assertion forms:

- Concept(individual)

- Role(individual1, individual2)

Strict rules:

1. Use only symbols that explicitly occur in the ontology or in the observation.
2. Do not invent new concepts, roles, individuals, variables, or placeholders.
3. Do not output RDF/XML tags, namespaces, prefixes, or ontology syntax.
4. Do not add comments, explanations, numbering, or extra text.
5. Return as many valid alternative explanations as you can infer from the ontology.
6. If you are uncertain, return {}.

Before answering, internally verify that every symbol in the output appears in the ontology or observation.

Example of the required format:

{A(a)} ; {B(a), R(a,b)} ; {C(a), D(a), S(a,c)}

Prompt 3

You are given an ontology and an observation.

The ontology may be written in OWL or RDF-style syntax.
Use the ontology as the only source of knowledge.

Task:

Generate abductive explanations for the observation.

Explanation format:

- Each explanation is a set of ABox assertions.
- Write each explanation inside curly braces { }.
- Separate alternative explanations with ';'.
- Separate assertions inside one explanation with ','.

Allowed assertion forms:

- Concept(individual)
- Role(individual1, individual2)

Strict rules:

1. Use only symbols that explicitly appear in the ontology or the observation.
2. Do not invent concepts, roles, individuals, variables, or placeholders.
3. Do not output RDF/XML tags, namespaces, prefixes, or ontology syntax.
4. Do not add comments, reasoning, numbering, or any extra text.
5. If uncertain, return {}.

Output example:

{A(a)} ; {B(a), R(a,b)}