PEIRCE: Unifying Material and Formal Reasoning via LLM-Driven Neuro-Symbolic Refinement

Xin Quan*[1], Marco Valentino*[2,3], Danilo S. Carvalho[1,4], Dhairya Dalal[5], André Freitas[1,2,4]

[1]University of Manchester, United Kingdom
[2]Idiap Research Institute, Switzerland
[3]University of Sheffield, United Kingdom
[4]National Biomarker Centre, CRUK-MI, United Kingdom
[5]University of Galway, Ireland

https://github.com/neuro-symbolic-ai/peirce/

Abstract

A persistent challenge in AI is the effective integration of material and formal inference – the former concerning the plausibility and contextual relevance of arguments, while the latter focusing on their logical and structural validity. Large Language Models (LLMs), by virtue of their extensive pre-training on large textual corpora, exhibit strong capabilities in material inference. However, their reasoning often lacks formal rigour and verifiability. At the same time, LLMs' linguistic competence positions them as a promising bridge between natural and formal languages, opening up new opportunities for combining these two modes of reasoning. In this paper, we introduce PEIRCE, a neuro-symbolic framework designed to unify material and formal inference through an iterative conjecture–criticism process. Within this framework, LLMs play the central role of generating candidate solutions in natural and formal languages, which are then evaluated and refined via interaction with external critique models. These critiques include symbolic provers, which assess formal validity, as well as soft evaluators that measure the quality of the generated arguments along linguistic and epistemic dimensions such as plausibility, coherence, and parsimony. While PEIRCE is a general-purpose framework, we demonstrate its capabilities in the domain of natural language explanation generation – a setting that inherently demands both material adequacy and formal correctness.

1 Introduction

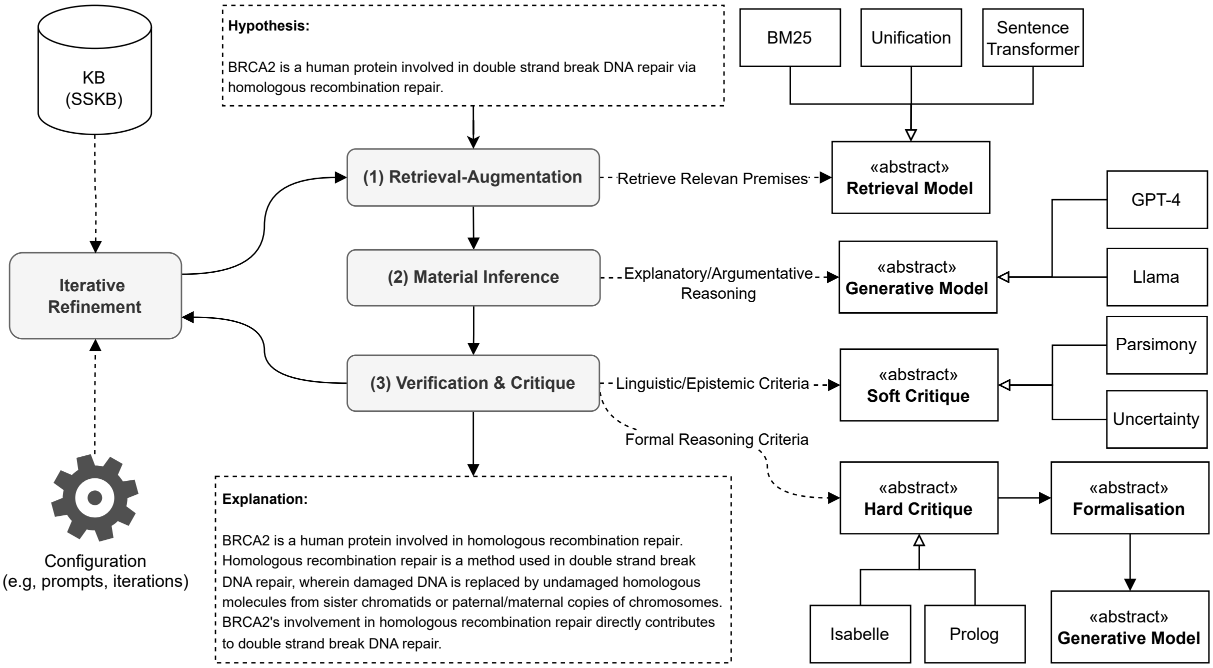
A core challenge in Artificial Intelligence (AI) is the integration of material and formal inference (Mahowald et al., 2024; Guo et al., 2025; Cheng et al., 2025; Dasgupta et al., 2022; Valentino and Freitas, 2024b; Hamilton et al., 2024; Kambhampati et al., 2024). Drawing from classical distinctions in logic and philosophy of science (Brandom, 1994; Haack, 1978), formal inference concerns the structural validity of arguments – whether conclusions follow necessarily from a set of premises according to fixed syntactic rules – while material inference is concerned with the plausibility of those arguments and their grounding in background knowledge, context, and domain-specific assumptions. Despite their complementary nature, these forms of inference are typically handled by distinct types of systems in AI: symbolic provers for formal reasoning, and statistical or neural models for material inference.

Recently, the advent of Large Language Models (LLMs) offers new opportunities for bridging these two modalities (Xu et al., 2024; Gandarela et al., 2024; Morishita et al., 2024; Ranaldi et al., 2025). Their linguistic fluency and access to broad world knowledge, in fact, enable them to generate candidate solutions that approximate material reasoning. Simultaneously, emerging work has shown that LLMs can support autoformalisation, translating natural language content into structured logical forms suitable for downstream symbolic verification (Quan et al., 2024b; Pan et al., 2023; Olausson et al., 2023; Jiang et al., 2024; Kirtania et al., 2024). This creates an opportunity for hybrid neuro-symbolic architectures that leverage the interpretive strengths of LLMs alongside the rigour of symbolic solvers.

This paper presents PEIRCE, a modular and extensible framework for modelling iterative reasoning workflows that unify material and formal inference. PEIRCE implements a conjecture–criticism cycle, in which LLMs generate candidate solutions in natural and formal languages, and a suite of external critique models – ranging from formal proof assistants to linguistic and semantic evaluators – assessing the quality of the generated solutions according to multiple criteria, including logical validity, plausibility, coherence, and parsimony.

*Equal contribution. For Marco Valentino, the work was done at Idiap under the NeuMath project.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 11–21
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics


KB
(SSKB)
Hypothesis:
BM25
Unification
Sentence Transformer
(1) Retrieval-Augmentation
Retrieve Relevan Premises
«abstract» Retrieval Model
GPT-4
Iterative Refinement
(2) Material Inference
Explanatory/Argumentative Reasoning
«abstract» Generative Model
Llama
Parsimony
(3) Verification & Critique
Linguistic/Epistemic Criteria
«abstract» Soft Critique
Uncertainty
Formal Reasoning Criteria
«abstract» Hard Critique
«abstract» Formalisation
Configuration
(e.g, prompts, iterations)
Explanation:
Isabelle
Prolog
«abstract» Generative Model

BRCA2 is a human protein involved in double strand break DNA repair via homologous recombination repair.

BRCA2 is a human protein involved in homologous recombination repair.
Homologous recombination repair is a method used in double strand break DNA repair, wherein damaged DNA is replaced by undamaged homologous molecules from sister chromatids or paternal/maternal copies of chromosomes.
BRCA2's involvement in homologous recombination repair directly contributes to double strand break DNA repair.

Figure 1: Overall architecture of PEIRCE. The framework provides an extensible and modular environment for unifying material and formal inference in natural language via a *conjecture-criticism* process. PEIRCE supports controllability and formal error correction mechanisms for implementing a complete end-to-end iterative refinement pipeline for explanatory arguments generated by LLMs.

To demonstrate the capabilities of PEIRCE, we focus on the task of natural language explanation generation as a representative case study. Explanations constitute a particularly useful testbed for reasoning, as they must simultaneously satisfy formal and material constraints (Valentino and Freitas, 2024a). We evaluated the framework across several domains and datasets spanning from textual entailment (Camburu et al., 2018), scientific question answering (Jansen and Ustalov, 2020; Dalvi et al., 2021), and clinical hypothesis verification, showing how PEIRCE effectively enables the generation, evaluation and refinement of high-quality explanatory arguments.

2 PEIRCE: Unifying Material and Formal Reasoning

PEIRCE provides an extensible and modular environment for modelling and unifying *material and formal reasoning* via a *conjecture-criticism* cycle. The overall architecture of PEIRCE is illustrated in Figure 1. The core functionality offered by the framework is the automation of an *iterative refinement* pipeline for *natural language inference* tasks in different domains. This pipeline is typically organised into three distinct stages implemented through the orchestration of customisable components – i.e., (1) *retrieval-augmentation*, (2) *material inference*, and (3) *verification and critique*.

Given an NLI problem as input (e.g., answering a question, predicting an entailment relation, verifying a scientific claim or a hypothesis, etc.), the first stage in the process involves querying external knowledge bases (Section 2.1) via retrieval models (Section 2.2) to select relevant premises to support reasoning. Subsequently, the retrieved knowledge can be provided in context to a generative model to generate an approximate solution in natural language (Section 2.3). The solution proposed by the generative model is then criticised by a suite of hard and soft critique models, which might use an internal formalisation stage (Section 2.4). The critiques' feedback can then be fed back to the generative model to refine the solution in the next iteration and improve its quality (Section 2.5).

PEIRCE provides abstract interfaces to instantiate and customise the iterative refinement pipeline, facilitating modularity and extensibility.

2.1 Data Model

PEIRCE integrates a data model interface designed for storing and retrieving knowledge from corpora of annotated premises. The data model is designed to be general, efficient, and extensible in order to cover a diverse set of knowledge bases supporting explanatory reasoning in different domains.

A knowledge base consists of a sequence of $statements$ that can be loaded and navigated as a collection. A $statement$ is a single fact, a sentence, or a claim (e.g., “The ‘(set) difference’ between two sets S and T is written $S \setminus T$, and means...”), which may refer to concrete $entities$, and may be linked to a set of premises (other $statements$) which together constitute an explanation of why the statement holds (see Figure 4).

This recursive structure facilitates access to multiple datasets in a unified format oriented towards explanatory reasoning. It is implemented in the form of the *Simple Statement Knowledge Bases* (SSKB) python package[1], illustrated in Figure 4. SSKB includes loaders for a few popular NLI datasets, such as e-SNLI (Camburu et al., 2018), WorldTree (Jansen et al., 2018), ProofWiki (Ferreira and Freitas, 2020), EntailmentBank (Dalvi et al., 2021), and NLI4CT (Jullien et al., 2023a,b, 2024) and also facilitates linguistic annotations through its compatibility with the *Simple Annotation Framework* (SAF)[2] NLP package.

2.2 Retrieval Models

In order to support the retrieval of relevant premises for reasoning from the knowledge base, PEIRCE provides an interface for implementing a suite of retrieval models, including sparse (i.e., BM25 (Robertson et al., 1995)), dense (i.e., Sentence-Transformers (Reimers and Gurevych, 2019)) and hybrid models specialised for explanatory inference (i.e., Unification and SCAR (Valentino et al., 2021b, 2022b)). The retrieval models are fully integrated with the data model to enable a dialogue with external corpora. Moreover, PEIRCE supports the creation of hybrid ensembles between retrieval models, allowing for a weighted ranking function (see Appendix B.2 for a concrete example).

2.3 Generative Models

PEIRCE implements a suite of classes to efficiently prompt and manage the adoption of different families of LLMs. In particular, PEIRCE supports full compatibility with OpenAI[3] and Huggingface[4] models. Different specialised classes following the same abstract interface facilitate reusability and extensibility for prompting LLMs for iterative refinement. The generative models internally use a class for dynamic prompting management that allows for the runtime instantiation of specific variables. This mechanism allows for the definition of a single prompt template that can be adapted at execution time to run experiments on different NLI problems (see Appendix B.3 for a concrete example).

2.4 Critique Models

The critique models are at the core of the iterative refinement process implemented in PEIRCE, representing the mechanism adopted to identify errors, inconsistencies and to determine the quality of the solutions generated by the LLMs. To facilitate their implementation and reuse, PEIRCE provides a suite of critique models, which can be instantiated and invoked through a common interface. In particular, PEIRCE provides the possibility of implementing both hard and soft critiques (Kambhampati et al., 2024; Dalal et al., 2024).

A hard critique model is responsible for verifying formal aspects of the reasoning, such as logical validity, and typically returns a discrete value (i.e., 1 or 0) that characterises the correctness of a specific aspect. Because of their formal nature, hard critique models may use an internal formalisation process to convert natural language into machine-verifiable languages (e.g., first-order logic). A soft critique model, on the other hand, is responsible for analysing linguistic and stylistic aspects of the generated solution (e.g., simplicity, uncertainty) and returns a normalised continuous score that quantifies the presence of a particular feature. Contrary to hard critique models, soft critiques do not typically require formalisation and operate directly on generated arguments in natural language.

A series of information can be returned within a critique model’s output depending on its nature, including a quality score in the case of a soft critique or the results of a formal verification (e.g., a logical proof) in the case of a hard critique. A concrete example of implementation is available in Appendix B.4.

2.4.1 Hard Critiques

Following recent work on the integration of LLMs and proof assistants for the verification and refinement of explanations (Quan et al., 2024b,a), PEIRCE provides a built-in implementation of hard

[1] https://github.com/neuro-symbolic-ai/SSKB

[2] https://github.com/dscarvalho/saf

[3] https://openai.com/index/openai-api/

[4] https://huggingface.co/models

	Science QA	Premise Selection
BM25	22.84	10.18
Unification	30.40	24.45
BM25 + Unification	**38.72**	**27.09**

Table 1: Explanation retrieval results (i.e., MAP) for science question answering (i.e., WorldTree) and natural language premise selection (i.e., ProofWiki).

critique models based on Isabelle[5] and Prolog[6].

These models use an internal formalisation process (through LLMs) to convert the NLI problem and the generated explanatory argument into a formal theory (through axioms and theorems) and verify, using a proof assistant or a symbolic solver, whether the generated solution logically entails the problem. If this is the case, the critique models will judge the solution as logically valid and will return the proof tactics found by the solver. If a proof cannot be found, the critique models return a detailed feedback describing the steps in which the proof construction has failed, allowing for error correction in a subsequent iteration.

The following is an example of proof tactics returned by the `IsabelleSolver` after successful verification:

```
'proof tactics': ['Sledgehammering
    ...', 'cvc4 found a proof...', '
    cvc4: Try this: using assms
    explanation_1 explanation_2 by
    blast (1 ms)', 'vampire found a
    proof...', 'vampire: Found
    duplicate proof', 'spass found a
    proof...', 'spass: Found
    duplicate proof', 'zipperposition
     found a proof...', '
    zipperposition: Found duplicate
    proof', 'Done']
```

2.4.2 Soft Critiques

Soft critiques are inspired by argumentation theory (van Eemeren et al., 2014) and philosophical accounts of inference to best explanation (Thagard, 1978; Lipton, 2017). Such methods can be adopted to qualify explanatory arguments and provide comparable selection criteria to identify the best solution amongst competing hypotheses. PEIRCE provides a built-in implementation of the parsimony, coherence, and uncertainty critique models introduced by Dalal et al. (2024).

Parsimony. Also known as Ockam's razor, parsimony favours arguments with the fewest assumptions and premises. This soft critique model is implemented computing the *concept drift*, which measures the number of new concepts and entities not present in the original NLI problem that are introduced in the generated solution.

Coherence Coherence evaluates the intermediate entailment relationships between the generated premises, favouring arguments that introduce conditional clauses that are more plausible. Specifically, this critique model adopts a pre-trained textual entailment model to measure the average entailment strength (through the predicted entailment score) over generated if-then clauses in an explanatory argument.

Uncertainty Uncertainty evaluates the plausibility of a generated argument via explicit linguistic signalling expressions. In particular, this critique models analyses hedging words such as *probably*, *might be*, and *could be* that typically signal ambiguity and are often used when the truth condition of a statement is unknown or probabilistic. This critique model adopts a fine-tuned model which analyses hedging language to establish the degree of uncertainty in the generated statements (Pei and Jurgens, 2021).

2.5 Iterative Refinement

Finally, PEIRCE provides a customisable class for iterative refinement that flexibly combines the components responsible for each intermediate stage.

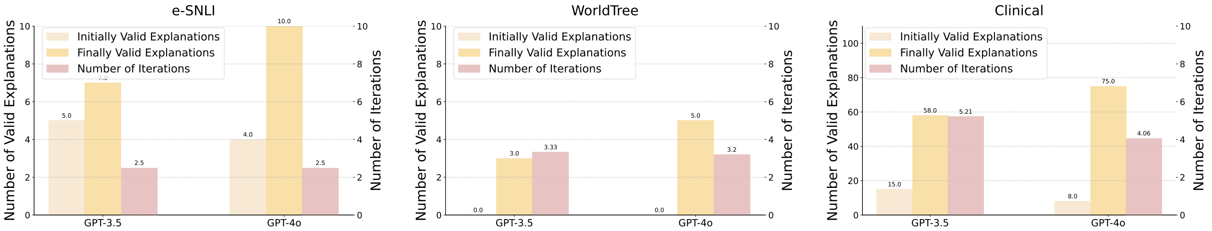
In particular, a class named `RefinementModel` is responsible for orchestrating retrieval models, LLMs, and critique models to perform solution refinement for a fixed number of iterations. If the critique model performs a hard critique (e.g., Isabelle), the refinement process ends when the generated argument can be formally verified (e.g., a proof is found). After the refinement, the output of the critique models, as well as the solution produced at each iteration step, will be returned. An example of implementation can be found in Appendix B.5.

3 Empirical Evaluation

We performed experiments to showcase PEIRCE's applicability to explanation-based NLI problems in different domains. In particular, we adopt PEIRCE to reproduce relevant models for natural language explanation generation, focusing on explanation retrieval, neuro-symbolic refinement of explanations for NLI, and inference to the best explanation with LLMs.

[5] https://isabelle.in.tum.de/

[6] https://www.swi-prolog.org/



(a) (b) (c)

Figure 2: Explanation refinement results via hard critique using GPT-4o and Isabelle (i.e., number of successfully verified explanations after a maximum of 10 iterations).

Dataset	Problem	Explanation	Iteration	Validity
e-SNLI	**Premise**: An infant is in a crib and crying. **Hypothesis**: A baby is unhappy.	if the infant is crying, it can be assumed that they are unhappy.	0	Invalid
		if the infant is crying, it can be assumed that they are unhappy. An infant is a type of baby.	1	Valid

Table 2: An example of how the explanations in e-SNLI can be refined via hard critique (i.e., GPT-4o and Isabelle).

3.1 Explanation Retrieval

For explanation retrieval, we measure the performance of BM25 (Robertson et al., 2009), the Unification-based retrieval model (Valentino et al., 2021b, 2022b), and an ensemble between the two on Science Question Answering (QA) and Natural Language Premise Selection. To this end, we measure the Mean Average Precision (MAP) of the retrieved explanatory premises on 50 randomly selected examples from the WorldTree corpus (for Science QA) (Jansen et al., 2018; Jansen and Ustalov, 2020; Thayaparan et al., 2021) and ProofWiki (for Premise Selection) (Ferreira and Freitas, 2020; Valentino et al., 2022a). The results, reported in Table 1, confirm the impact of the Unification-based retrieval model reported in previous work (Valentino et al., 2021a, 2022c,b), also demonstrating the benefit of performing an ensemble between the models.

3.2 Iterative Refinement via Hard Critique

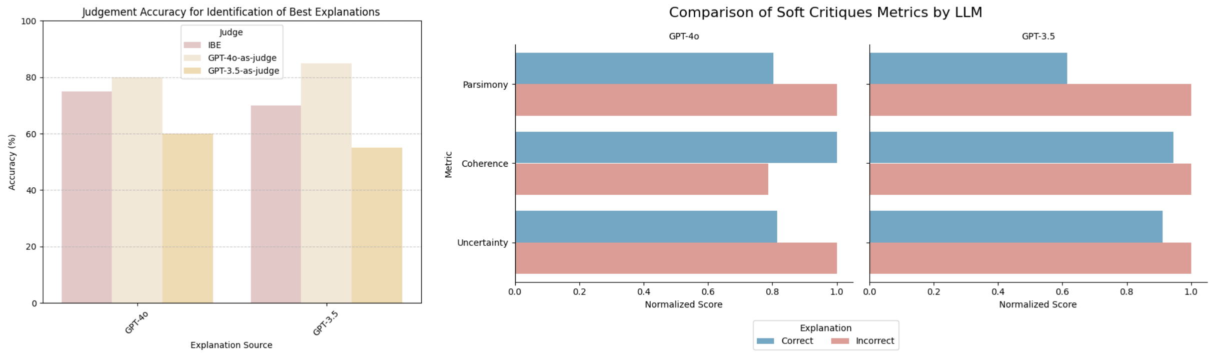
Using the built-in implementation of the refinement model and the hard critique based on Isabelle, we reproduced the iterative refinement pipeline introduced by (Quan et al., 2024b) on different domains (i.e., general textual entailment on e-SNLI (Camburu et al., 2018), science questions on Worldtree (Jansen et al., 2018), and clinical explanations annotated by domain experts). In particular, Figure 2 shows the number of natural language explanations that can be successfully verified and refined through the interaction of GPT-4o(Achiam et al., 2023) and Isabelle (Nipkow et al., 2002) after a maximum of 10 iterations. Qualitative examples of the results of the refinement process are provided in Tables 2 and 4.

3.3 Inference to the Best Explanation via Soft Critique

Finally, we demonstrate how soft critique models can be used to perform inference to the best explanation with LLMs (Dalal et al., 2024). Here, we consider the task of cause and effect prediction in a multiple-choice setting, where given a question and two competing candidates, the LLM must decide which is the most plausible answer. To this end, 20 causal questions were sourced from COPA (Gordon et al., 2012). GPT-4o and GPT-3.5 are then tasked with generating causal explanations for each candidate, which are then evaluated using the soft-critique criteria (Section 2.4.2). The best explanation is selected via a majority vote through the soft-critique scores (see example in Table 3). For comparison, LLM-as-judge baselines are provided in Figure 3a, with the results of the soft critique metrics reported provided in Figure 3b.

3.4 Related Work

Neuro-symbolic reasoning models integrate neural networks with symbolic solvers to provide a reliable and verifiable reasoning process for complex downstream tasks (e.g., multi-hop reasoning, scientific question-answering) involving large datasets



(a)

(b)

Figure 3: Accuracy in identifying the explanation associated with the correct answer via different soft critique models (i.e., parsimony, coherence and uncertainty in IBE vs. GPT-as-judge).

The cat chased the bird? What was the effect? A) The bird flew away. B) The bird caught a worm				
	Explanation 1	**Scores**	**Explanation 2**	**Scores**
GPT-4o	If a cat chases a bird, then the bird may perceive a threat. If the bird perceives a threat, then it is likely to take evasive action to escape. If the bird takes evasive action to escape, then it may fly away from the area. If the bird flies away from the area, then it will no longer be in the vicinity of the cat. Therefore, since the cat chased the bird, the bird perceived a threat and took evasive action by flying away to escape from the cat.	Coherence: 0.25 Parsimony: 6 Uncertainty: 1.03	If a cat chases a bird, then the bird may become startled or distracted. If the bird is startled or distracted, then it may stop focusing on its current activity, such as foraging for food. If the bird stops focusing on foraging, then it may inadvertently come across other food sources, such as worms. If the bird comes across a worm, then it may catch the worm as a food source. Therefore, since the cat chased the bird, causing it to become startled and stop foraging, the bird may have inadvertently caught a worm as a result of this distraction.	Coherence: 0.09 Parsimony: 9 Uncertainty: 2.33
GPT-3.5	If a cat chases a bird, then the bird may feel threatened. If the bird feels threatened, then it may try to escape. If the bird tries to escape, then it may fly away. Therefore, since the cat chased the bird, causing it to feel threatened, the bird likely flew away as a natural response to escape from the perceived danger.	Coherence: 0.06 Parsimony: 1 Uncertainty: 1.39	If the cat chased the bird, then the bird may have been alerted and flown away. If the bird flew away, then it may have landed on the ground to search for food. If the bird landed on the ground to search for food, then it may have caught a worm. Therefore, if the cat chased the bird, causing it to fly away and land on the ground to search for food, it is plausible that the bird caught a worm during its search.	Coherence: -0.05 Parsimony: 2 Uncertainty: 1.65

Table 3: An example of evaluating competing explanations via IBE using different soft critiques.

(Minervini et al., 2020; Kalyanpur et al., 2020; Shi et al., 2021; Wang and Pan, 2022; Weir et al., 2024).

Several studies have proposed differentiable solvers that enhance both the robustness of rule-based models and the interpretability of neural models (Rocktäschel and Riedel, 2017; Manhaeve et al., 2018; Weber et al., 2019; Thayaparan et al., 2022). More recently, integrating LLMs with logical reasoners has demonstrated significant effectiveness on natural language datasets (de Souza et al., 2025; Dalal et al., 2024; Lyu et al., 2023).

Research efforts have applied LLMs for autoformalisation, converting natural language into first-order logic forms, and subsequently employing symbolic provers on logical reasoning datasets (Pan et al., 2023; Olausson et al., 2023; Jiang et al., 2024). Quan et al. (2024b) integrated LLMs with external theorem provers for open-world natural language inference tasks to verify and refine natural language explanations.

Our research incorporates soft and hard critique models that uses various symbolic solvers and LLMs to evaluate logical and linguistic features, ensuring delivering logically valid, sound, and consistent explanations.

3.5 Conclusion & Future Work

This paper introduced PEIRCE, a framework that provides an extensible and modular environment for unifying material and formal inference in natural language via a *conjecture-criticism* process.

PEIRCE supports controllability and formal error correction mechanisms for implementing a complete iterative refinement pipeline for explanatory arguments generated by LLMs. We hope the release of PEIRCE will facilitate new research on neuro-symbolic applications driven by LLMs.

In future work, we plan to extend the suite of ready-to-use knowledge resources and critique models in the framework as well as integrate

PEIRCE with a supervised fine-tuning and reinforcement learning pipeline to leverage the feedback generated by the critique models and the refined solution for training.

Acknowledgments

This work was partially funded by the SNSF project NeuMath (200021_204617), by the EPSRC grant EP/T026995/1, "EnnCore" under Security for all in an AI-enabled society, by the CRUK National Biomarker Centre, and supported by the Manchester Experimental Cancer Medicine Centre and the NIHR Manchester Biomedical Research Centre.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Robert Brandom. 1994. *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard university press.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin. 2025. Empowering llms with logical reasoning: A comprehensive survey. *arXiv preprint arXiv:2502.15652*.

Dhairya Dalal, Marco Valentino, Andre Freitas, and Paul Buitelaar. 2024. Inference to the best explanation in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 217–235, Bangkok, Thailand. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051*.

João Pedro Gandarela de Souza, Danilo Carvalho, and André Freitas. 2025. Inductive learning of logical theories with llms: A expressivity-graded analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23752–23759.

Deborah Ferreira and André Freitas. 2020. Natural language premise selection: Finding supporting statements for mathematical text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2175–2182, Marseille, France. European Language Resources Association.

Joao Pedro Gandarela, Danilo S. Carvalho, and Andr'e Freitas. 2024. Inductive learning of logical theories with llms: A complexity-graded analysis. *ArXiv*, abs/2408.16779.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Susan Haack. 1978. *Philosophy of Logics*. Cambridge University Press.

Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. 2024. Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. *Semantic Web*, 15(4):1265–1306.

Peter Jansen and Dmitry Ustalov. 2020. TextGraphs 2020 Shared Task on Multi-Hop Inference for Explanation Regeneration. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, pages 85–97, Barcelona, Spain (Online). Association for Computational Linguistics.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. Worldtree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Dongwei Jiang, Marcio Fonseca, and Shay Cohen. 2024. LeanReasoner: Boosting complex logical reasoning with lean. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7497–7510, Mexico City, Mexico. Association for Computational Linguistics.

Mael Jullien, Marco Valentino, and André Freitas. 2024. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1947–1962, Mexico City, Mexico. Association for Computational Linguistics.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. 2023a. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16745–16764, Singapore. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023b. SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

Aditya Kalyanpur, Tom Breloff, and David A. Ferrucci. 2020. Braid: Weaving symbolic and neural knowledge into coherent logical explanations. In *AAAI Conference on Artificial Intelligence*.

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. 2024. Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning*.

Shashank Kirtania, Priyanshu Gupta, and Arjun Radhakrishna. 2024. LOGIC-LM++: Multi-step refinement for symbolic formulations. In *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*, pages 56–63, Bangkok, Thailand. Association for Computational Linguistics.

Peter Lipton. 2017. Inference to the best explanation. *A Companion to the Philosophy of Science*, pages 184–193.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. 2018. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette, and Tim Rocktäschel. 2020. Learning reasoning strategies in end-to-end differentiable proving. *Preprint*, arXiv:2007.06477.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2024. Enhancing reasoning capabilities of llms via principled synthetic logic corpus. In *Advances in Neural Information Processing Systems*, volume 37, pages 73572–73604. Curran Associates, Inc.

Tobias Nipkow, Markus Wenzel, and Lawrence C Paulson. 2002. *Isabelle/HOL: a proof assistant for higher-order logic*. Springer.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 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*, pages 5153–5176, Singapore. Association for Computational Linguistics.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

Jiaxin Pei and David Jurgens. 2021. Measuring sentence-level and aspect-level (un) certainty in science communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Xin Quan, Marco Valentino, Louise Dennis, and Andre Freitas. 2024a. Enhancing ethical explanations of large language models through iterative symbolic refinement. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–22, St. Julian's, Malta. Association for Computational Linguistics.

Xin Quan, Marco Valentino, Louise A. Dennis, and Andre Freitas. 2024b. 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*, pages 2933–2958, Miami, Florida, USA. Association for Computational Linguistics.

Leonardo Ranaldi, Marco Valentino, Alexander Polonsky, and André Freitas. 2025. Improving chain-of-thought reasoning via quasi-symbolic abstractions. *ArXiv*, abs/2502.12616.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, and 1 others. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Tim Rocktäschel and Sebastian Riedel. 2017. End-to-end differentiable proving. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 3791–3803.

Jihao Shi, Xiao Ding, Li Du, Ting Liu, and Bing Qin. 2021. Neural natural logic inference for interpretable question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3673–3684, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul R Thagard. 1978. The best explanation: Criteria for theory choice. *The journal of philosophy*, 75(2):76–92.

Mokanarangan Thayaparan, Marco Valentino, Deborah Ferreira, Julia Rozanova, and André Freitas. 2022. Diff-explainer: Differentiable convex optimization for explainable multi-hop inference. *Transactions of the Association for Computational Linguistics*, 10:1103–1119.

Mokanarangan Thayaparan, Marco Valentino, Peter Jansen, and Dmitry Ustalov. 2021. TextGraphs 2021 shared task on multi-hop inference for explanation regeneration. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 156–165, Mexico City, Mexico. Association for Computational Linguistics.

Marco Valentino, Deborah Ferreira, Mokanarangan Thayaparan, André Freitas, and Dmitry Ustalov. 2022a. TextGraphs 2022 shared task on natural language premise selection. In *Proceedings of TextGraphs-16: Graph-based Methods for Natural Language Processing*, pages 105–113, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Marco Valentino and André Freitas. 2024a. On the nature of explanation: An epistemological-linguistic perspective for explanation-based natural language inference. *Philosophy and Technology*, 37(3):1–33.

Marco Valentino and André Freitas. 2024b. Reasoning with natural language explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 25–31, Miami, Florida, USA. Association for Computational Linguistics.

Marco Valentino, Ian Pratt-Hartmann, and André Freitas. 2021a. Do natural language explanations represent valid logical arguments? verifying entailment in explainable NLI gold standards. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 76–86, Groningen, The Netherlands (online). Association for Computational Linguistics.

Marco Valentino, Mokanarangan Thayaparan, Deborah Ferreira, and André Freitas. 2022b. Hybrid autoregressive inference for scalable multi-hop explanation regeneration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11403–11411.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2021b. Unification-based reconstruction of multi-hop explanations for science questions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 200–211, Online. Association for Computational Linguistics.

Marco Valentino, Mokanarangan Thayaparan, and André Freitas. 2022c. Case-based abductive natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1556–1568, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. 2014. *Argumentation TheoryArgumentationtheory*, pages 1–49. Springer Netherlands, Dordrecht.

Wenya Wang and Sinno Pan. 2022. Deep inductive logic reasoning for multi-hop reading comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4999–5009, Dublin, Ireland. Association for Computational Linguistics.

Leon Weber, Pasquale Minervini, Jannes Münchmeyer, Ulf Leser, and Tim Rocktäschel. 2019. NLProlog: Reasoning with weak unification for question answering in natural language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6151–6161, Florence, Italy. Association for Computational Linguistics.

Nathaniel Weir, Peter Clark, and Benjamin Van Durme. 2024. Nellie: A neuro-symbolic inference engine for grounded, compositional, and explainable reasoning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 3602–3612. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association*

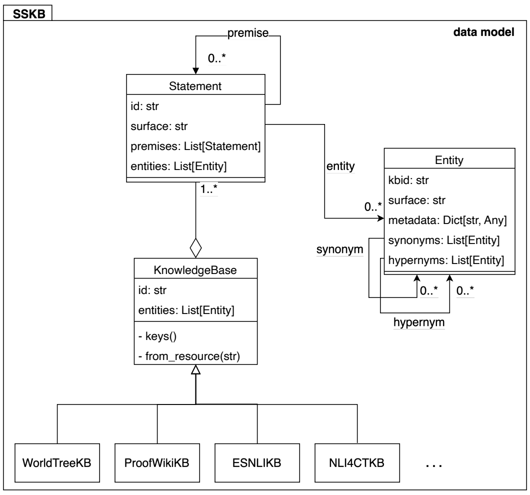


Figure 4: UML diagram of the *Simple Statement Knowledge Bases* (SSKB) package. The classes at the bottom implement loading facilities for popular NLI datasets.

for Computational Linguistics (Volume 1: Long Papers), pages 13326–13365, Bangkok, Thailand. Association for Computational Linguistics.

A Explanation Refinement Examples

Table 4 shows additional examples of iterative refinement via hard critique (i.e. GPT-4o and Isabelle) on Worldtree and clinical explanations.

B Implementation Details

B.1 Data Model

The following code snippet shows an example of how to use SSKB to load data from external explanation corpora (i.e., WordlTree):

```
from sskb import WorldTreeKB

kb = WorldTreeKB()

# Retrieve the individual facts in
    the corpus
facts_kb = [stt for stt in kb if (
    stt.annotations["type"] == "fact"
    )]

# Retrieve the questions in the test
     set
test_questions = [stt for stt in kb
    if (stt.annotations["type"] == "
    question" and stt.annotations["
    split"] == "test")]

# Retrieve a complete explanation
explanation = [p.surface for p in
    test_questions[42].premises]
```

B.2 Retrieval Models

An example of how to instantiate and query the data model via BM25 is presented below:

```
from retrieval.bm25 import BM25Model

# Initialize BM25 model
bm25 = BM25Model(facts_kb)

# Construct the list of queries
queries = [q.surface for q in
    test_questions]

# Compute BM25 ranking and scores
res_bm25 = bm25.query(queries)
```

An example of how to instantiate and query the data model using an ensemble model is presented below:

```
from retrieval.ensemble import
    EnsembleModel

# Initialise the ensemble model
ensemble_model = EnsembleModel(
        [bm25, unification],
        weights = [0.8, 0.2]
    )
```

B.3 Generative Models

An example of how to prompt GPT-4o for explanation generation is provided below:

```
from generation.generative_model
    import GPT

# Parameters for prompting
api_key = "personal key"
prompt_file = "explanation_prompt.
    txt"

# Input problem
hypothesis = "I pricked the baloon."
conclusion = "The balloon expanded."

# Initialise the model
llm = GPT('gpt-4o', api_key)

# Generate an explanation
explanation = llm.generate(
            prompt_file,
            hypothesis,
            conclusion
        )
```

An example of a dynamic prompt is provided below, with hypothesis and conclusion acting as variables that can be specified at runtime

```
You are an expert on causal
    reasoning
and explanation. You will use causal
knowledge and commonsense to provide
logical explanations for the
    provided
causal reasoning scenarios.

```

Dataset	Problem	Explanation	Iteration	Validity
WorldTree	**Hypothesis**: Being able to read is an example of a learned trait.	Usually people learn how to read and write in school. trait is synonymous with characteristic. inherited characteristics are the opposite of learned characteristics; acquired characteristics. inheriting is when a inherited characteristic is copied; is passed from parent to offspring by genetics;DNA.	0	Invalid
		Usually, people learn how to read and write in school, and activities learned in school are considered learned characteristics. Trait is synonymous with characteristic. Inherited characteristics are the opposite of learned characteristics. Learned characteristics are also known as acquired characteristics. Inheriting is when an inherited characteristic is copied or passed from parent to offspring by genetics or DNA. Learned characteristics are synonymous with learned traits.	2	Valid
Clinical	**Hypothesis**: BRCA2 is a human protein involved in double strand break DNA break repair via homologous recombination repair.	BRCA2 is a human protein involved in homologous recombination repair. Homologous recombination repair is a double strand break DNA repair process wherein damaged DNA is replaced by undamaged homologous molecules from sister chromatids or paternal/maternal copies of chromosomes.	0	Invalid
		BRCA2 is a human protein involved in homologous recombination repair. Homologous recombination repair is a method used in double strand break DNA repair, wherein damaged DNA is replaced by undamaged homologous molecules from sister chromatids or paternal/maternal copies of chromosomes. BRCA2's involvement in homologous recombination repair directly contributes to double strand break DNA repair.	2	Valid

Table 4: Examples of iterative explanation refinement for WorldTree and clinical explanations using GPT-4o and Isabelle.

```
For the hypothesis and conclusion
provided in the test example, let's
think step-by-step and generate an
explanation...

Test Example:

Hypothesis: {hypothesis}
Conclusion: {conclusion}
```

B.4 Critique Models

An example of how to instantiate a hard critique model via an external Isabelle solver and GPT-4o as formaliser is provided below:

```
from critique.isabelle import IsabelleSolver

# Example from e-SNLI
premise = "A couple playing with a little boy on the beach."
hypothesis = "A couple are playing with a young child outside."
explanation = "little boy is a young child."

# Initialise the model
llm = GPT('gpt-4o', api_key)

# Initialise the critique model
isabelle = IsabelleSolver(
    generative_model = llm,
    isabelle_session = 'HOL'
    )

# Perform the critique
res = critique_model.critique(
    hypothesis,
    premise,
    explanation
    )
```

B.5 Iterative Refinement

An example of how to instantiate a complete refinement process for 10 iterations is provided below:

```
from refinement.refinement_model import RefinementModel

# Initialise the refinement process
refinement_model = RefinementModel(
    generative_model = llm,
    critique_model = isabelle
)

# Perform refinement for 10 iterations
res = refinement_model.refine(
    hypothesis = hypothesis,
    premise = premise,
    explanation = explanation,
    iterations = 10
    )
```