Master internship in Robust **K**nowledge **E**xtraction from **LLM**s

Context. Research efforts have expanded to improve large language models (LLMs) logical reasoning precision and transparency through combination of neuro-symbolic system components and self-checking methods as well as argumentation building frameworks. These procedures intend to build better logical coherence within LLM-generated reasoning as well as reducing statistical pattern recognition errors. Structured prompting through reasoning templates provides a decomposition method for logical reasoning that helps create systematic analysis of inference processes. Logical coherence in LLMs receives support through the combination of FOL theorem provers with LLMs in both LOGIC-LM [6] and LINC [5] to maintain outputs within formal logical frameworks. Moreover, the SelfCheck [4] functionality embedded in LLMs allows them to monitor their reasoning chain structure automatically making it possible to perform internal cheques for errors before generating outputs which reduces hallucinations and inferential inconsistencies. DetermLR [7] provides LLMs with a structured reasoning framework which determines premises as either definite or ambiguous so LLMs can produce decisive logical outputs. The decision-making procedure reinforced with LLM-based solutions becomes more dependable through logical consistency frameworks which enforce rules for transitivity and commutativity and negation invariance [3]. The development of fact verification systems that use knowledge graphs enhances retrieval-augmented generation (RAG) systems [2] thus improving LLM capabilities for propositional logic queries. Argumentative LLMs evolved from studies on argumentation theory to create decision-supporting systems that apply structured argumentation frameworks for debate-based choices. More recently, argumentative LLMs [1] establish structured argument frameworks to create an organized reasoning process that enhances public debate about AI-generated conclusions and bolsters their validity. Advanced LLMs continue to face obstacles in applying real logical reasoning because their statistical heuristics need enhancement through fundamental studies of formalised AI reasoning frameworks.

Despite these recent advances in improving the logical reasoning capabilities of LLMs, several limitations persist. A major challenge remains the ability of LLMs to show consistent and reliable reasoning, as their outputs are prone to hallucinations, biases, and logical contradictions. This is particularly concerning in critical domains such as law, medicine and finance. LLMs may generate outputs that conflict with background knowledge or input text, undermining their ability. Explainability, or understanding the reasoning behind LLM outputs, also remains another significant challenge. In particular, LLMs are currently limited in their ability to reliably provide outputs that are both explainable and contestable, which is crucial for transparency and trust in AI systems. Although the combined deployment of these methods could improve LLM reasoning deficiencies and establishes new potential applications for complex decision systems, LLM still face significant challenges when confronted with complex logical and mathematical problems. These limitations can hinder their reliability, particularly in complex and dynamic scenarios, where stable and accurate reasoning is crucial for practical applications.

Research Objectives. The ultimate goal of this internship is to develop principled and rigorous explainable techniques for dealing with inconsistency, improving the robustness of LLMs, and guaranteeing the faithfulness of their reasoning. This project is a first step towards a long term ambitious aim of designing tractable methods for dealing with inconsistency in NLP to draw (high level) explanatory conclusions using social choice techniques.

Throughout this internship, we aim to address the following three fundamental challenges in a broader context:

- Managing LLMs inconsistency by employing different logics to maintain coherent reasoning in LLMs despite contradictions, ensuring more robust and reliable decision-making processes.
- Enhancing the transparency of AI decision-making in LLMs by integrating new logical reasoning into current frameworks and handling logical inconsistency, thereby improving their ability to emulate human reasoning.
- Improving knowledge extraction from LLMs through the application of new techniques in the evaluation and synthetis of model responses, thereby enhancing the quality and reliabiliy of extracted information.

References

[1] Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. Argumentative large language models for explainable and contestable decision-making. *CoRR*, abs/2405.02079, 2024.

[2] Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, and Arijit Khan. Logical consistency of large language models in fact-checking.

[3] Yinhong Liu, Zhijiang Guo, Tianya Liang, Ehsan Shareghi, Ivan Vulić, and Nigel Collier. Aligning with logic: Measuring, evaluating and improving logical preference consistency in large language models.

[4] Ning Miao, Yee Whye Teh, and Tom Rainforth. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. In *ICLR*, 2024.

[5] Theo Olausson, Alex Gu, Benjamin Lipkin, Cedegao E. Zhang, Armando Solar-Lezama, Joshua B. Tenenbaum, and Roger Levy. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *EMNLP*, pages 5153–5176, 2023.

[6] Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In *EMNLP*, pages 3806–3824, 2023.

[7] Hongda Sun, Weikai Xu, Wei Liu, Jian Luan, Bin Wang, Shuo Shang, Ji-Rong Wen, and Rui Yan. Determlr: Augmenting llm-based logical reasoning from indeterminacy to determinacy. In *ACL*, pages 9828–9862, 2024.