Framework for Evaluating Reasoning Verification in Large Language Models

Keywords: Language Models, Reasoning Chains, Verification
Research Groups: Data, Intelligence, and Graph Team, Télécom Paris, France. BNP Paribas, France
Advisors: Mehwish Alam, Bérénice Jaulmes, Jean-Christophe Arouette.
Starting Date: Spring 2026

Scientific Context.

Recent progress in Large Language Models (LLMs) has led to remarkable advances in *Chain-of-Thought* (CoT) reasoning, the step-by-step generation of intermediate thoughts to reach a final answer. However, the reliability and interpretability of these reasoning chains remains an open challenge. This internship will contribute to the growing scientific effort to *verify* and *evaluate* CoT reasoning and its verifiers through the analysis and development of benchmark datasets and evaluation metrics. The goal is to strengthen our understanding of how LLMs reason, identify where they fail, and provide a basis for designing methods to measure reasoning quality more accurately.

The project will explore existing benchmarks, which provide large-scale annotated datasets for reasoning verification across domains like mathematics, physics, and commonsense reasoning. Each of these benchmarks introduces different verification methodologies; for instance, PRM800k uses fine-grained human annotations for every reasoning step, while THINK-Bench introduces precision and recall metrics for key logical steps. The intern will analyze the advantages and drawbacks of these datasets, such as scalability, annotation reliability, and domain coverage, and investigate how they can be extended or combined for more comprehensive reasoning evaluation.

In parallel, the internship will focus on evaluation metrics for reasoning quality, including Process Discernibility Score (PDS), ROSCOE, etc. These metrics capture different aspects of reasoning performance: PDS measures inter-chain consistency, ROSCOE quantifies logical and semantic coherence, and PRMScore evaluates precision in error detection. A key aspect of the internship will be to compare these metrics empirically, identify correlations with human judgment, and test their robustness under controlled perturbations of reasoning steps.

Finally, the intern will have the opportunity to contribute to the design of a new benchmark or metric aimed at evaluating *reasoning verification models*, such as PRMs or LLM-based critics. Depending on progress and interests, this may involve generating synthetic CoTs with controlled reasoning errors, building automatic evaluators, or testing LLM verifiers on multilingual or domain-specific reasoning datasets.

This work will contribute to the broader scientific goal of making LLM reasoning more interpretable, verifiable, and trustworthy, a cornerstone for next-generation AI research.

Objectives.

- Benchmark Analysis and Taxonomy Construction
- Metric Comparison and Integration
- Framework Design and Implementation
- Empirical Validation and Case Studies

Candidate Profile.

- Currently pursuing M2 in the field of Artificial Intelligence/Machine Learning
- Good programming skills, such as in Python (incl. Pytorch).
- Knowledge of Large Language Models is a plus but not required; however, interest in learning and keeping themselves up-to-date with upcoming trends in the field is required.
- Good communication skills, especially in English.

Outstanding candidates can be considered for a Thèse CIFRE based on the performance and availability of the funds.

Location.

The candidate will spend 80% of the time at Telecom Paris and 20% of the time at BNP Paribas. Télécom Paris ranks among the first engineering schools in France. It is the leading French school in digital technologies, including machine learning and artificial intelligence. The Data, Information, Graphs (DIG) team is focused on subjects ranging from database theory to knowledge bases and natural language processing. The team at BNP Paribas is mostly concerned with the topics related to Artificial Intelligence, such as stream mining, graph representation learning, trustworthy AI, etc.

Required Documents.

- A full CV
- A motivation letter expressing your interest in the position and relevant experience
- A transcript of records

Contacts.

Please send the complete required documents to Mehwish Alam (mehwish.alam@telecom-paris.fr), Bérénice Jaulmes (berenice.jaulmes@ip-paris.fr), and Jean-Christophe Arouete (jean-christophe.arouete@bnpparibas.com) in an email with the subject starting with “[M2Internship-CoT]”.

References.

[1] Y. He, S. Li, J. Liu, W. Wang, X. Bu, G. Zhang, Z. Peng, Z. Zhang, Z. Zheng, W. Su, et al. Can large language models detect errors in long chain-of-thought reasoning? arXiv preprint arXiv:2502.19361, 2025.

[2] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt. Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874, 2021.

[3] Z. Li, Y. Chang, Y. Wu. Think-bench: Evaluating thinking efficiency and chain-of-thought quality of large reasoning models. arXiv preprint arXiv:2505.22113, 2025.

[4] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, K. Cobbe. Let’s verify step by step. In The Twelfth International Conference on Learning Representations, 2023.

[5] X. Xu, S. Diao, C. Yang, Y. Wang. Can we verify step by step for incorrect answer detection? arXiv preprint arXiv:2402.10528, 2024.

[6] P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, Z. Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. arXiv preprint arXiv:2312.08935, 2023.

[7] O. Golovneva, M. Chen, S. Poff, M. Corredor, L. Zettlemoyer, M. Fazel-Zarandi, and A. Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reasoning. arXiv preprint arXiv:2212.07919, 2022.