Scientific Reasoning with LLMs

Stuttgart Team: Florian Dreyer (3667877) Ekaterina Kolos (3688474)
Daria Matiash (3668740)

October 29, 2024

1 Introduction

Scientific machine reasoning is the application of AI techniques to simulate scientific reasoning processes, such as data interpretation, hypothesis generation, and causal inference. It facilitates scientific inquiry and discovery by enabling machines to independently analyze scientific data, generate hypotheses, and make predictions.

For our project, we intend to compare different approaches around generative large language models to generate solutions for school science problems. On an existing benchmark with human solutions, we are planning to compare: a pure LLM with zero-shot and few-shot prompting (and possibly some more fine-grained prompting techniques), an agentic approach with ability to query external resources like textbooks in physics, maths etc, and soft-prompting techniques. We then plan to attempt knowledge distillation to a smaller LLM.

2 Background and Related Work

Prompting techniques A prompt is a command a user gives to a generative LLM used to guide its output for a downstream task, which contains several of the following: a role (persona) the LLM has to follow (e.g. "You are a helpful assistant"), a precise task description, examples (exemplars) for analogous learning (few-shot prompting), requirements on the process of reasoning, style instructions, output format requirements, additional information for in-context learning, emotion prompting components (highlighting why the task or a particular requirement is important). Prompts can be combined into sequences or graphs (with complex branching and parallelism). Of particular interest are small requirements on the reasoning process that can significantly improve performance when added to the prompt, such as "Let's think step by step" [6]. Chain-of-thought prompts improve reasoning capabilities by asking the model to speak its process of thinking out loud, with variations including self-ask, step-back prompting, thread-of-thought, least-to-most prompting (decomposing then solving), plan-and-solve prompting, tree-of-thought, recursion-of-thought and many more [11]. Furthermore, prompts themselves can be compressed and automatically optimized to improve efficiency and reduce costs [2]. [3] compress a long context 4x into memory slots, while [14] ask the LLM to first summarize the prompt and then execute it. Self-criticism and ensembling techniques can further be used to improve reasoning capabilities [11].

Soft prompting is an alternative to fine-tuning where the model is frozen while only a small number of "soft" prompt parameters are trained [1]. These parameters are used to guide the model in the right direction.

Agents An important step forward in using LLMs are agent-based architectures [8] [4]. They allow the LLM to act as an intelligent being capable of planning a solution of a complex task, while resorting to external resources on the way to acquire in-context the knowledge it does not have from pre-training. The retrieved information, results of invoking tools, such as calculators, and of interactions with the environment are appended to the prompt (or, in a more advanced scenario, to the summarized memory). Optionally, a reflexion step (explicit reasoning on all the accumulated information) is added before allowing further generation of the final response.

Knowledge distillation (KD) allows to obtain smaller models capable of successfully following the behavior of larger teacher models. This is particularly useful for privacy reasons (running AI applications on mobile devices) and for cases where access to very large models in the cloud is not guaranteed (e.g. in cars). The teacher model can be used to intelligently select examples on which the student model

is trained (dataset distillation) [15], or provide on negative samples to show the student what incorrect answers or reasoning paths it should avoid to improve task accuracy [7] (c.g. contrastive CoT). Training small models on a CoT reasoning path of a larger model was also shown to be a way to obtain a small student model replicating reasoning capabilities of teacher on downstream tasks [10], which is close to *response-based KD* where the student model mimics the output of the teacher. More effective can be *feature-based KD* where student also partially replicates the teacher on a feature-based level, when knowledge from specific layers of the teacher model is distilled into the student model [12], while the student model's structure may be a quantized, simplified, or condensed version of the teacher's architecture [5].

This way, the student model learns the "protocols" of how to solve different problems, as well as the essence of the relevant knowledge selected by the teacher LLM.

3 Methodology

We postulate the following research questions: RQ1: Can we build a LLM Agent to improve the LLMs performance on science questions? RQ2: Does Soft Prompting improve the performance of the LLM (Agent)? RQ3: Can an LLM-agent's behavior be distilled into a single model?

RQ1 We plan to start by prompting a multimodal model with reasoning capabilities with a zero-shot and few-shot prompting settings. A foundation model for this step could be a T4/T5 model or a multimodal Llama model. This simple generation will be compared with an agent-based approach using the same model, which will now include augmented retrieval of information on the scientific task from domain-specific texts.

RQ2 We plan to use Soft Prompting as one of the parameter-efficient fine-tuning techniques on the LLM used to guide it towards better reasoning. To achieve this we will add learnable prompt parameters to the base LLM we use. While training these parameters the rest of the model will be frozen [1].

RQ3 In order to distill the knowledge from the obtained model previous steps, we plan to do the model distillation with Chain-of-Thought Prompting for Reasoning approach. Following techniques described in [10] [13], the student model will be fine-tuned using these CoT responses to produce intermediate reasoning steps. At the same time, the teacher model generates multiple CoT responses, and the student learns from the aggregate (self-consistent) reasoning paths. The described approach can help the student model avoid common mistakes and output more consistent answers.

Dataset We plan to use the ScienceQA dataset [9]. The dataset includes a variety of science-related multimodal multiple-choice questions together with annotations of the answers that provide relevant lectures and explanations in Nature Science, Language Science and Social Science.

Metrics We'll evaluate model's performance with question answering accuracy domain-wise in order to have a fair comparison with leaderboards. A few metrics used in machine translation such as BLEU-1, BLEU-4 and BERTScore can also be used to evaluate the reasoning steps.

4 Approximate Timeline

1. **Now - mid November**:
   - Preliminary baseline tests: small LLM relying only on commonsense reasoning / only on pretraining (zero-shot, few-shot).
   - Optional: stronger baseline – same LLM with better prompting techniques like CoT/Self-Ask.
2. **mid November - mid December**:
   - Construct agentic / RAG pipeline with "big smart" model + additional resources.
   - Optional: soft-prompting for better workflow to distill from.
3. **mid December - mid January**:
   - Knowledge distillation experiments.

References

[1] Noah Constant Brian Lester Rami Al-Rfou. "The Power of Scale for Parameter-Efficient Prompt Tuning". In: *arXiv preprint arXiv:2104.08691* (2021).

[2] Kaiyan Chang et al. "Efficient Prompting Methods for Large Language Models: A Survey". In: *arXiv preprint arXiv:2404.01077* (2024).

[3] Tao Ge et al. "In-context autoencoder for context compression in a large language model". In: *arXiv preprint arXiv:2307.06945* (2023).

[4] Alireza Ghafarollahi and Markus J Buehler. "SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning". In: *arXiv preprint arXiv:2409.05556* (2024).

[5] Jianping Gou et al. "Knowledge distillation: A survey". In: *International Journal of Computer Vision* 129.6 (2021), pp. 1789–1819.

[6] Takeshi Kojima et al. "Large language models are zero-shot reasoners". In: *Advances in neural information processing systems* 35 (2022), pp. 22199–22213.

[7] Yiwei Li et al. "Turning dust into gold: Distilling complex reasoning capabilities from llms by leveraging negative data". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 17. 2024, pp. 18591–18599.

[8] Bill Yuchen Lin et al. "Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks". In: *Advances in Neural Information Processing Systems* 36 (2024).

[9] Pan Lu et al. "Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering". In: *The 36th Conference on Neural Information Processing Systems (NeurIPS)*. 2022.

[10] Lucie Charlotte Magister et al. "Teaching small language models to reason". In: *arXiv preprint arXiv:2212.08410* (2022).

[11] Sander Schulhoff et al. "The Prompt Report: A Systematic Survey of Prompting Techniques". In: *arXiv preprint arXiv:2406.06608* (2024).

[12] Majid Sepahvand, Fardin Abdali-Mohammadi, and Amir Taherkordi. "Teacher–student knowledge distillation based on decomposed deep feature representation for intelligent mobile applications". In: *Expert Systems with Applications* 202 (2022), p. 117474.

[13] Jason Wei et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[14] Jason Weston and Sainbayar Sukhbaatar. "System 2 Attention (is something you might need too)". In: *arXiv preprint arXiv:2311.11829* (2023).

[15] Ruonan Yu, Songhua Liu, and Xinchao Wang. "Dataset distillation: A comprehensive review". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).