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" [**kojima2022large**]. 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 [**schulhoff2024prompt**]. Furthermore, prompts themselves can be compressed and automatically optimized to improve efficiency and reduce costs [**chang2024efficient**]. [**ge2023context**] compress a long context 4x into memory slots, while [**weston2023system**] 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 [**schulhoff2024prompt**].

Soft prompting is an alternative to fine-tuning where the model is frozen while only a small number of "soft" prompt parameters are trained [**lester2021prompt**]. These parameters are used to guide the model in the right direction.

Agents An important step forward in using LLMs are agent-based architectures [**lin2024swiftsage**] [**ghafarollahi2024sciagents**]. 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) [**yu2023dataset**], or provide negative samples to show the student what

incorrect answers or reasoning paths it should avoid to improve task accuracy [**li2024turning**] (c.f. 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 [**magister2022teaching**], which is close to *response-based KD* where the student model mimics the output of the teacher. More effective in some cases 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 [**sepahvand2022teacher**], while the student model's structure may be a quantized, simplified, or condensed version of the teacher's architecture [**gou2021knowledge**].

In our project, we plan to develop an extended version of CoT knowledge distillation, where both the reasoning paths and the essence of the external knowledge the teacher agent has retrieved for the tasks are acting as training data for the student model. This way the teacher agent "defines" the "protocols" of solving complex tasks, while the student model tries to "follow" the learnt "instructions" and memorize important information for similar tasks. The student LLM should be small, capable of reasoning and capable of acquiring new distilled knowledge. If we succeed, such small models could be a valuable asset in practice in enterprise environments, to encode important protocols and parts of project documentation, replacing expensive intelligent agents for specific tasks.

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 [**lester2021prompt**].

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 [**magister2022teaching**] [**wei2022chain**], 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 [**lu2022learn**]. 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.