Feb 9, 2025

LLM Agents Study Session

Summary

This meeting, led by Arunesh Mishra and attended by Raj Murtinty, Abdulhakeem Adefioye, Joydeep Ghatak, Yatharth Piplani, and Shristi Gautam, focused on improving reasoning in LLMs, particularly through System 1 (intrinsic improvement) and System 2 (agentic frameworks) approaches as discussed in Professor Jason Weston's lecture. DeepSeek's reinforcement learning approach, utilizing GRPO and chain-of-thought prompting, was detailed, with a future "deep dive" session scheduled for Saturday evening Pacific Time (Sunday morning India time) to cover its code, reward functions, and cost-effectiveness; Yatharth Piplani will present on GRPO. The meeting also covered various reasoning techniques like chain-of-thought prompting, chain-of-verification, and self-rewarding LLMs, including the introduction of IRPO and meta-rewarding, along with challenges in reward design and LLM judgment; Arunesh Mishra will share code and resources on Discord.

Details

- **Meeting Overview and Attendance:** Arun opened the meeting, noting that attendance might be low due to the Super Bowl. The meeting's purpose was to discuss lecture two, focusing on improving reasoning with LLMs, a key area for DeepSeek in 2025.
- **Lecture Two Summary: System 1 vs. System 2:** Arunesh Mishra summarized lecture two, presented by Professor Jason Weston, which explored how to create self-improving AI. They described two approaches: System 1 (intrinsic LLM improvement) and System 2 (agentic frameworks with multiple LLM calls).

System 1 techniques face challenges like hallucination and jailbreaking, while System 2 addresses these issues through techniques like planning and chain of thought. The lecture primarily focused on iterative model improvement, where the model generates responses, ranks them using a different prompt, and uses preference optimization (DPO) for training.

- **Historical Context of LLMs:** Arunesh Mishra provided a brief history of language modeling, tracing its development from early neural probabilistic models to the rise of transformers and the scaling hypothesis. They noted the significance of attention mechanisms for reasoning and the impact of reinforcement learning with human feedback (RLHF) on instruction following in models like InstructGPT.
- **DeepSeek and Reinforcement Learning:** Raj Murtinty and Abdulhakeem Adefioye discussed DeepSeek, clarifying that it uses reinforcement learning with a reward function and gradient-based preference optimization (GRPO), but not human feedback. They explained that DeepSeek generates multiple responses, compares them to ground truth, and assigns rewards based on correctness. A participant shared a LinkedIn post explaining DeepSeek's approach using the analogy of a chess game where the model learns from mistakes. Joydeep Ghatak highlighted that DeepSeek builds upon basic reinforcement learning principles but incorporates chain-of-thought prompting. A further discussion of DeepSeek's cost-effectiveness and optimization techniques was tabled for a future "deep dive" session.
- **DeepSeek Deep Dive and GRPO:** Multiple participants expressed interest in a detailed session on DeepSeek and GRPO. Arunesh Mishra proposed dedicating time during a future meeting for a detailed explanation of DeepSeek's methods, including its code, reward functions, and cost-effectiveness. Yatharth Piplani offered to present on GRPO.
- **System 2 Reasoning Techniques:** Arunesh Mishra discussed several System 2 reasoning techniques. Chain-of-thought prompting, initially demonstrated by giving examples, later evolved to prompting the model directly to reason step-by-step. Chain-of-verification was introduced as a method to reduce hallucination by verifying responses through reverse questioning. Addressing the issue of sycophancy (biased answers due to biased prompts), they presented system 2 attention, which uses an LLM to rephrase prompts into unbiased forms before generating answers. For complex tasks, they suggested a method of generating plans and subproblems using LLMs to break down the larger task.

- **Self-Rewarding LLMs:** The discussion shifted to self-rewarding LLMs, where the model evaluates its own outputs and assigns rewards to guide its improvement. Arunesh Mishra highlighted the transition from human-based evaluation to LLM-based evaluation as models become increasingly sophisticated. They described a self-improving loop involving generating responses, using the model (with a different prompt) to judge them via DPO training, and iteratively improving the model. An example using Llama 2 70B and data from Open Assistant was presented, showing comparable performance to GPT-4 after three iterations.
- **Iterative Reasoning Preference Optimization (IRPO):** Dealing with the challenge of LLMs ineffectively judging reasoning tasks, especially math problems, Arunesh Mishra introduced IRPO. This approach uses verifiable rewards (correct answers) to create preference pairs involving chain-of-thoughts and final answers. This was highlighted as similar to DeepSeek's method. Raj Murtinty summarized DeepSeek as removing supervised fine-tuning (SFT) and relying solely on reinforcement learning for certain training phases. Yatharth Piplani added that DeepSeek's process involves pre-processing chain-of-thoughts to mimic human thought before using SFT.
- **Addressing Reward Challenges:** Shristi Gautam raised questions about rewarding models based on non-integer answers. Arunesh Mishra suggested using an LLM verifier and potentially fine-tuning the judge model for specialized datasets.
- **Generating Chain of Thoughts:** The final discussion covered methods for prompting LLMs to generate chain-of-thoughts. Arunesh Mishra presented a prompt DeepSeek uses that encourages step-by-step reasoning and a specific final answer format. The meeting concluded with a summary of how this process incorporates rewarding models for adherence to the specified format and utilizing the correct final answer to implicitly gauge the correctness of reasoning steps. The use of temperature and self-consistency to generate diverse chain-of-thoughts were also discussed. The meeting ended with a comparison of the DeepSeek approach and the OpenAI approach to reasoning.
- **Verifiable Rewards and LLM Training:** Arunesh Mishra discussed verifiable rewards, a method to evaluate model outputs without relying solely on LLMs. They explained a technique using an LLM as a judge to generate preference pairs for training, initially resulting in worse performance but improving after iterations. The method applies to various instruction-following tasks, and OpenAI's work on RL suggests a similar approach, although specifics are limited. The same LLM

can act as both the main model and the judge with specific prompting, and the model can output its thoughts even for simple tasks.

- **Meta-Rewarding LLMs:** Arunesh Mishra introduced meta-rewarding, a technique to improve both the LLM's performance and its judging capabilities. This involves using the same LLM as a meta-judge to evaluate its own judgments. They described how this produces preference pairs for both the main task and the judging task, improving the LLM's performance and evaluation skills. The meta-rewarding model outperforms self-rewarding models because it enhances the judge's capabilities, which is crucial for the DPO training. Future work may integrate chain-of-thought (CoT) prompting into the judging process. They also discussed how synthetic data generation helps create verifiable rewards for judging, addressing the recursive problem of using an LLM to judge another LLM.
- **Improving LLM Judgment and Understanding:** Joydeep Ghatak raised concerns about the model's understanding of the question. Arunesh Mishra suggested using the LLM to improve the input and potentially explore reasoning within the LLM's encoder-decoder layers, or prompting it to output its thoughts on the question before generating a solution.
- **Code and Quiz:** Arunesh Mishra offered to share code and resources on Discord, and they conducted a quiz covering concepts discussed during the meeting . Participants discussed different approaches to chain-of-verification, specifically comparing methods that focused on error reduction and robustness with those emphasizing logical consistency .
- **DeepSeek Deep Dive and Future Plans:** The participants scheduled a DeepSeek deep dive for Saturday evening Pacific Time (Sunday morning India time), covering topics like PPO, RLHF, and GRPO, using a provided notebook and code. They also discussed training models using the provided DeepSeek code and evaluating them on a novel dataset. The potential use of the J-Benchmark for evaluation was also suggested. Several participants mentioned challenges running the code and the time it took to train models. The session concluded with discussion of the Super Bowl and plans to share recordings on a YouTube channel.