Feb 2, 2025

LLM Agents Study Session

Summary

Arunesh Mishra presented on inference-time reasoning techniques for LLMs, covering prompting strategies, self-consistency, and iterative self-improvement, with a focus on the DeepSeek models (V3, R1, R10) and their architecture, training, and performance compared to OpenAI models. Kevin Toronto Developer contributed insights on model advancements and market impact, while Jin inquired about distilled model variations. Future steps include deeper exploration of reinforcement learning techniques and potential collaborations on research projects and conference submissions.

Details

- **Meeting Introduction and Logistics:** Sairamakrishna Nampalli and Arunesh Mishra discussed scheduling. They agreed to reschedule the meeting to a time more convenient for participants in India. Arunesh Mishra identified themself as a software engineer based in California, originally from Rajasthan, India. Sairamakrishna Nampalli works on cloud technologies. They also noted the presence of AI note-takers in the meeting and the lack of a formed group for a hackathon project. They agreed that forming a team for the hackathon would be beneficial.
- **Lecture Overview and Reasoning Techniques:** Arunesh Mishra presented on inference-time reasoning techniques for LLMs, reviewing material from a previous lecture. They highlighted reasoning as a key area of innovation for LLMs , noting improvements in math and coding benchmarks with increased token budget and inference time. They discussed various prompting techniques, including basic prompting, instruction prompting, and analogical prompting , explaining that embedding reasoning in examples or using instructions like "let's

think step by step" improves accuracy. They also covered LLM-assisted prompt engineering and the intuition behind chain-of-thought prompting, emphasizing its variable computation.

- **Multiple Solution Strategies and Self-Consistency:** Arunesh Mishra described strategies for exploring multiple solution candidates, including generating multiple solutions and self-consistency. Self-consistency involves running the model multiple times and selecting the most frequent answer. Universal self-consistency extends this to open-ended problems by using another LLM to assess consistency. They discussed challenges in selecting the best output when dealing with multiple solutions, especially for subjective tasks, suggesting presenting multiple options to the user. A participant, Kevin Toronto Developer, noted that these techniques might be outdated due to advancements in models like DeepSeek.
- **Iterative Self-Improvement and DeepSeek:** Arunesh Mishra explained iterative self-improvement techniques, including reflection and self-refinement.   They highlighted the need for an independent feedback mechanism, noting that relying on another LLM for feedback can be detrimental. They stated that self-consistency generally outperforms multi-agent approaches. Arunesh Mishra then gave an overview of DeepSeek R1 and R10, emphasizing their comparable performance to OpenAI's GPT-1 on math and coding benchmarks, due to reinforcement learning techniques. They also discussed DeepSeek V3 base, highlighting its mixture-of-experts architecture and multi-head latent attention.
- **DeepSeek V3 Architecture and Training:** Arunesh Mishra described DeepSeek V3's architecture, which uses a router (a multilayer neural network) to select experts based on input and partial output. The model saves computation by avoiding quadratic scaling with input length , and it performs multi-token prediction with a multiple-head output design. It was trained on approximately 14-15 trillion tokens, with a focus on math, programming, and reasoning, and fine-tuned on 1.5 million instruction-tuned instances. Kevin Toronto Developer noted that regardless of the exact training cost, the model's impact is significant, evidenced by the drop in Nvidia's stock value. They also pointed out that major cloud providers (AWS and Azure) adopted the model, suggesting its performance is real, though the reported training cost may be exaggerated.
- **DeepSeek R1 and R10: Reinforcement Learning and Reasoning:** Arunesh Mishra explained DeepSeek R1 and R10. R10 uses reinforcement learning with generalized proximal policy optimization (GPO) and a rule-based reward function,

aiming to let the model learn reasoning strategies automatically. They use a sampling method across multiple outputs (similar to best-of-K in self-consistency), feeding the best output back into the training process. The reward function biases the model towards more thoughtful responses , leading to emergent behaviors like self-reflection and self-consistency. R1 is a version of R10 improved for non-reasoning tasks. The model's ability to extend its test-time compute by reevaluating its approach emerged spontaneously.

- **DeepSeek R1 Training and Distillation:** Arunesh Mishra presented a flowchart illustrating the complex training process of DeepSeek R1, which involves generating both reasoning and non-reasoning data from DeepSeek V3, then applying supervised fine-tuning and reinforcement learning. They also discussed distilling R1 into smaller models (starting with a smaller base model like Llama instead of DeepSeek V3) to allow for easier use on less powerful hardware. Jin inquired about the differences between these distilled models, and Arunesh Mishra explained that larger models generally perform better on reasoning and related tasks. Arunesh Mishra highlighted that the ability to distill the model to smaller sizes is a key advantage.
- **DeepSeek R1 Performance and Limitations:** Arunesh Mishra summarized DeepSeek R1's performance, noting that it achieves results comparable to OpenAI's 01 model in some aspects and surpasses GPT-4 in others , with improvements on math benchmarks. However, they also highlighted limitations: R1 underperforms DeepSeek V3 due to the RL training's emphasis on reasoning, negatively impacting tasks like JSON or SQL output and multi-turn conversations. Prompt engineering is detrimental since R1 already employs complex reasoning strategies.
- **Research Project Ideas:** Arunesh Mishra suggested research project ideas, such as applying the same RL-based reasoning to smaller models, comparing RL-based reasoning with inference-time methods (like tree of thought and Monte Carlo Tree Search), and determining in which domains and situations each method performs better. They emphasized the feasibility of these projects given readily available resources.
- **Collaboration and Future Discussions:** Participants discussed potential collaborations on research projects and conference submissions. Arunesh Mishra proposed to delve deeper into reinforcement learning details in a future meeting, based on the interest of the participants.

You should review Gemini's notes to make sure they're accurate. [Get tips and learn how Gemini takes notes]

Please provide feedback about using Gemini to take notes in a [short survey.]