Apr 5, 2025

LLM Agents Study Session

Summary

Arunesh Mishra presented Meta's Llama 2 launch, detailing its three models (Scout, Maverick, and Behemoth), architecture (iROPE for long context windows), and technical specifications, while Yatharth Piplani and Sumanth Meenan contributed discussions on inference temperature scaling and data extraction challenges, respectively. Sumanth Meenan suggested exploring Donut and Copali for data extraction, while Arunesh Mishra proposed creating an evaluation set to compare models, integrating project updates into future meetings.

Details

- **Meeting Reschedule and Agenda:** Yatharth Piplani requested to postpone their presentation on the Deepsea reward model to the following week due to extensive prerequisites. Arunesh Mishra agreed and suggested focusing on the Llama 2 launch announcement instead (00:00:00).
- **Team Introductions and Locations:** Participants introduced themselves: Arunesh Mishra (California), Yatharth Piplani (Delhi), and Sumanth Meenan (Atlanta). They discussed the time zone confusion caused by a previous participant (00:02:33).
- **Llama 2 Launch Overview:** Arunesh Mishra presented an overview of Meta's Llama 2 launch, highlighting three models: Scout, Maverick, and the still-under-training Behemoth. They emphasized the models' mixture-of-experts architecture and significantly increased context length (00:05:05).
- **Llama 2 Model Details:** The presentation detailed the technical specifications of Scout (16 experts, 109 billion parameters) and Maverick (128 experts, 400 billion

parameters), noting their shared 17 billion active parameters and multimodal capabilities (00:06:24). Arunesh Mishra highlighted the models' performance on various benchmarks and their cost-effectiveness (00:07:32).

- **Llama 2 Architecture and iROPE:** The discussion covered the Llama 2 architecture, focusing on the novel iROPE (interleaved layers and rotary positional embedding) method for achieving a long context window (00:10:12). Arunesh Mishra explained how iROPE combines local attention with a global attention layer to handle longer sequences (00:11:35).
- **Inference Time Temperature Scaling:** They discussed the inference-time temperature scaling used in Llama 2 to address the issue of losing information in long contexts (00:13:10). Yatharth Piplani noted that this scaling is applied to the global attention layers (00:18:30).
- **Model Availability and Costs:** Arunesh Mishra confirmed that Scout and Maverick models are downloadable, with an API available (though not free for extended context lengths) (00:14:20) (00:16:40). They discussed the models' sizes and computational requirements (00:15:24).
- **Data Extraction from Documents:** Sumanth Meenan raised the challenge of extracting data from documents, particularly from complex layouts (00:24:02). They mentioned exploring models like Donut and Copali for this task, which do not require Optical Character Recognition (OCR) (00:26:15).
- **Model Comparison for Data Extraction:** Arunesh Mishra suggested creating an evaluation set to compare different models for data extraction, highlighting the importance of evaluation in agentic applications (00:29:43). They also mentioned using LLMs to classify document types before processing (00:34:30).
- **Agentes Projects and Future Meetings:** They discussed the Agentes projects and the process of forming teams (00:37:01). Arunesh Mishra suggested integrating project updates into future meetings (00:38:10).
- **Scalable Softmax and Attention Fading:** Yatharth Piplani discussed the scalable softmax approach used in Llama 2 to mitigate attention fading in long sequences (00:40:49). They noted that the order of layers (global attention before positional embeddings) is reversed from traditional transformers (00:39:18).

Suggested next steps

- [ ] Yatharth Piplani will review the Meta Llama 4 paper and write an explanation of the iROPE architecture and inference-time temperature scaling.
- [ ] Arunesh Mishra will prepare a lecture on coding vulnerabilities for a future meeting, including examples.
- [ ] Yatharth Piplani will prepare a session discussing the evolution of transformers, including positional embeddings, self-attention, multi-query attention, and the impact of different architectural changes on context window size.