Meeting Summary for Public AI Seminar (2024/09/24)

The following notes were generated by Zoom and lightly edited to remove names and other attribution. AI-generated content may be inaccurate or misleading. Always check for accuracy.

Quick Recap

The meeting discussed the development of GPT-3, the first large language model for Scandinavian languages, and the challenges faced during its development, including data regulation and infrastructure limitations. The team also explored the trend of open models, the collaboration between different organizations, and the challenges posed by regulation to AI development, particularly in the context of copyright issues. The conversation ended with a discussion on the importance of legal certainty for AI projects and the need for collaboration to address these challenges.

Summary

Welcome and Meeting Structure Explanation
The meeting began with a welcome and an explanation of its structure, which included a 50-minute presentation followed by a 50-minute Q&A session. The first half-hour was recorded, while the second half was under Chatham House rules. Despite technical issues with the recording, the meeting proceeded as planned.

GPT-3 Development and AI Sweden Collaboration
The head of research for natural language understanding at AI Sweden discussed the development of GPT-3, the first large language model for Scandinavian languages. He shared insights on AI Sweden's role in fostering AI collaboration and highlighted the formation of a consortium with Research Institutes of Sweden to build the model. The project's mission, partnerships, and funding sources were also discussed.

Challenges in Large Parameter Model Development
The discussion shifted to the challenges of developing large models. Initially, the team aimed to build a model similar to GPT-3 but faced resistance from the National Library of Sweden. As a result, they created their own dataset, the Nordic pile. Issues with releasing data and models as open source were highlighted, due to their semi-private status. The speaker also mentioned that they chose not to go above 40 billion parameters due to insufficient high-quality data and challenges with releasing generative models under an Apache 2 license.

Open Models, Data Sets, and Multimodality Discussion
There was a discussion on the closing performance gap between proprietary and open models,

with a focus on Llama-based models being sufficient for many applications. The team is currently working on generating more high-quality data and exploring multimodality, such as an audio-to-audio model with text-to-audio transformation. A new dataset, Sweb, with over one trillion tokens, will be released soon. Collaboration with the Fraunhofer IA Group on open foundation models for European languages was also mentioned, with plans to train models for 24 official European languages and 46 minority languages. These models will be openly released.

Challenges in European Research Funding and Models
The speaker discussed challenges in European research funding, particularly in the Nordic region, emphasizing the need for large-scale infrastructure. The unclear implications of data protection laws on model training were a key concern. Despite this, the organization is shifting towards open-source models, and they continue to train models using Swedish data.

Challenges and Opportunities in Scandinavian Language Models
Challenges of building models for Scandinavian languages were discussed, particularly in relation to data regulations and infrastructure limitations. The importance of collaboration at a Nordic or European level was emphasized. Ongoing initiatives in Spain, Germany, and Finland were mentioned as examples of potential collaborative efforts.

AI Sweden's Reception and Legal Structure
The discussion touched on the cultural reception of AI Sweden, which has been positive despite budget cuts. AI Sweden provides a collaborative environment for organizations to share knowledge. However, some partners have dropped out due to perceived lack of benefit. The importance of keeping public AI from privatization was raised as a concern.

Addressing Copyright and Data Regulatory Issues
The speaker addressed the challenges of copyright and data regulations in Europe, particularly in Scandinavia. Legal uncertainty around training models on copyrighted material and the risk of lawsuits were highlighted. Although efforts to clarify the law are ongoing, the current legal framework remains unclear. Concerns about regulation stifling innovation were discussed, but it was noted that regulations are also necessary to counter cultural imperialism.

Addressing AI Regulation and Copyright Challenges
The conversation focused on the regulatory challenges faced by AI development, particularly concerning copyright. A proposal to connect legal experts with AI projects to build legal certainty was raised. The lack of legal clarity in the current environment was compared to a "wild West." A suggestion was made to explore stopgap measures to prevent innovation from being stifled, while others proposed hiring European copyright lawyers to manage legal risks. A shared reading group was suggested to bridge the gap between technical and legal aspects of public AI. The meeting concluded with a reminder to check the meeting summary on AI Companion and continue discussions on Slack.