Meeting 0426

Meeting Topic: Technical Architecture and Model Data Discussion

Meeting Time: April 27 (Sunday), 12:55 – 13:22 (GMT+08)
Participants: @Zhou Junwei @Chen Wenhao @Jiang Qiyi @Zhu Qicheng @Luke Anderson
Note: This intelligent meeting summary was generated by AI and may contain inaccuracies. Please review carefully before use.

Summary

The meeting discussed various topics including frontend/backend selection, database choices, file import and retrieval workflows, models, and SQL-based subset retrieval. Key discussion points included:

- **Technical Architecture Plan Discussion**
  - **Frontend Selection**: Gradio, a lightweight framework, was chosen for the frontend due to its fast development capabilities and suitability for form-based interfaces.
  - **Backend Decision**: FastAPI was selected for the backend because it enables quick service deployment.
  - **Database Choice**: MySQL will be used to store table data, and Milvus will store vector embeddings.
  - **File Import Workflow**: The process includes storing raw file info in MySQL, extracting text from various file types, handling images within text and tables, performing label recognition, keyword extraction, and text embeddings.
  - **Retrieval Workflow**: Users input a query, which is first rewritten. Then both SQL-based and vector-based searches are performed. The results are intersected or subsetted, followed by reranking. Topics such as reranking methods and number of returned results were also discussed.
  - **Model Discussions**: Covered query rewriting and intent recognition models—whether they're needed, model size, functions, and complexity.
  - **SQL Subset Search**: Discussed the feasibility and challenges of performing Milvus retrieval within subsets obtained via SQL search, along with potential solutions.
- **Model and Data Discussion**

- **NL to SQL Model Fine-Tuning**: On March 10, during his internship, Nan shared with Jiang Qiyi the fine-tuning details of an NL-to-DSL model. It wasn't used in production, and SQL outputs might perform better than DSL.
- **Models Required for the Task**: Small models may be suitable for label recognition, keyword extraction, and intent detection. The reranking solution is still under consideration—whether to use large models or traditional methods.
- **Database Design**: Designed two MySQL tables: one for raw files and one for extracted files, recording file names, paths, types, etc. Milvus will be linked to MySQL IDs and also log file types.
- **Data Compilation Status**: Zhou Junwei asked Jiang Qiyi, Chen Wenhao, and Zhu Qicheng to review the data compilation results. The three had some uncertainties about the types of data needed and the collection methods. They decided to wait for Zhou Junwei’ s return to clarify.

Note: This content is temporarily unavailable for display outside of Feishu Docs.

To-Do Items

- **Dataset Coordination**: Jiang Qiyi, Chen Wenhao, and Zhu Qicheng should first communicate among themselves to finalize the results of data compilation and identify common file types. Once Zhou Junwei is available, discuss with him to clarify needs and feasible approaches to populating the dataset.
  (From Zhou Junwei (Frank))
  Participants: @Jiang Qiyi @Chen Wenhao @Zhu Qicheng @Luke Anderson

Note: This content is temporarily unavailable for display outside of Feishu Docs.

Related Links

- **Text Record**
  - *Technical Architecture and Model Data Discussion – April 27, 2025*