AskAlice chatbot evaluation

INITIAL RAG PIPELINE EVALUATION ON A LABELED DATASET

Evaluation details

DATASET

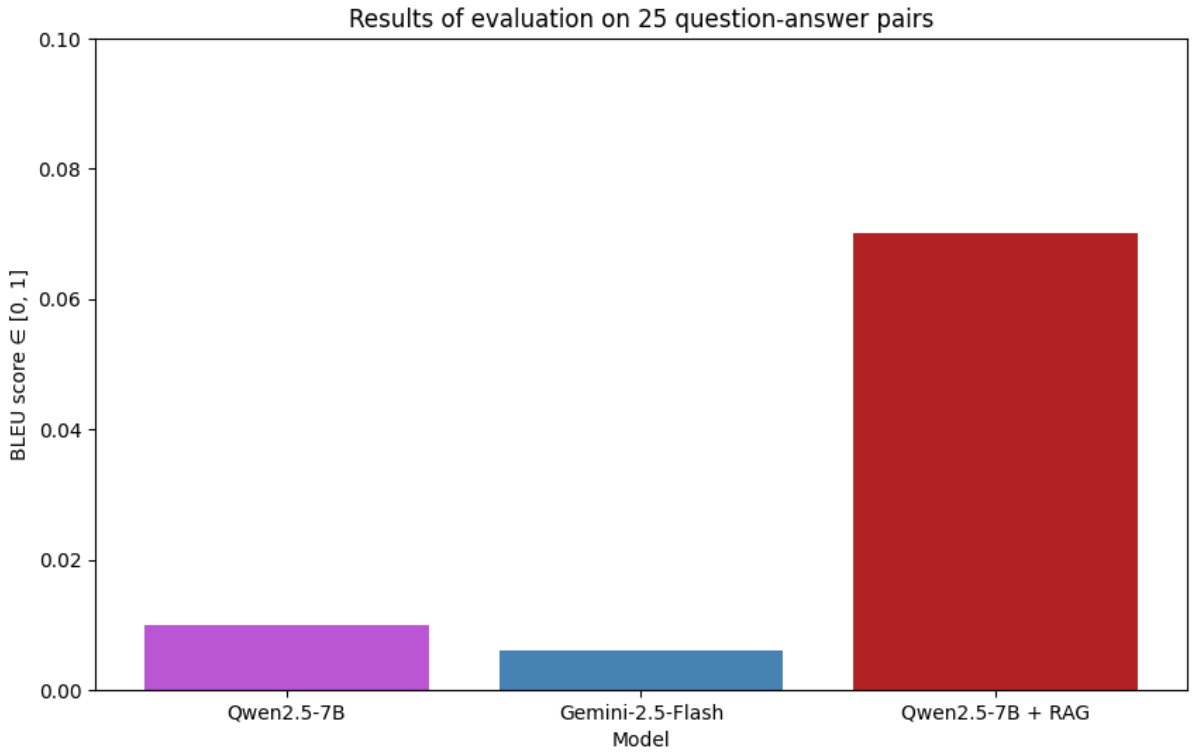
- **25 question-answer pairs** provided by Sandro
- Final answers constructed by ChatGPT

METRICS

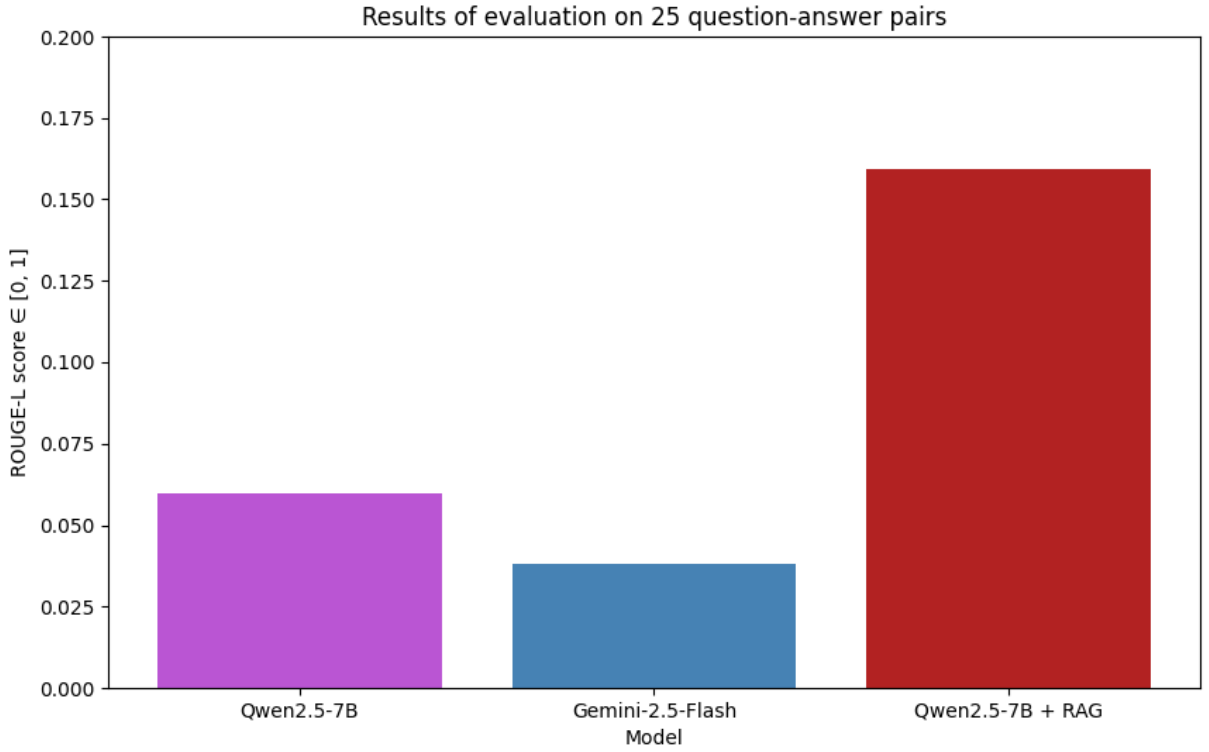
- Similarity **between generated and correct answer**
- Lexical overlap: BLEU and ROUGE scores
- Embedding based: Cosine similarity
- Learned judgement: LLM-as-judge

BLEU score comparison

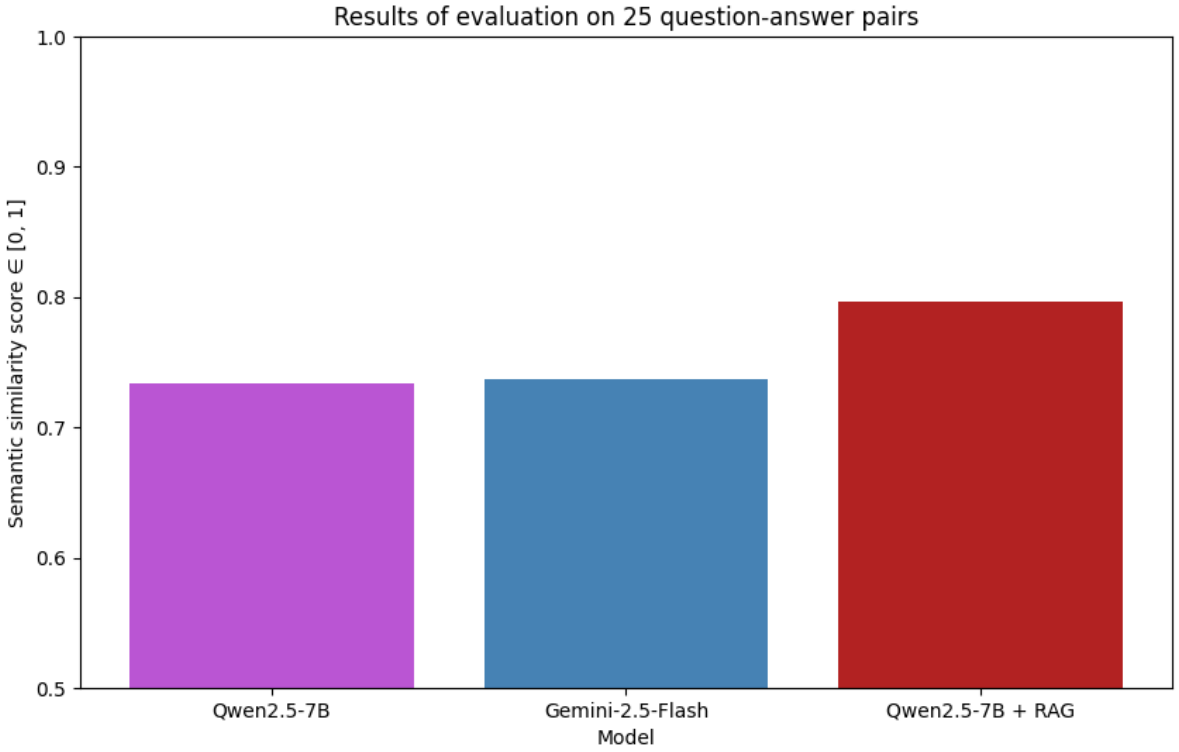
- (Probably the least appropriate metric, as it is literal and precision based)



ROUGE-L score comparison

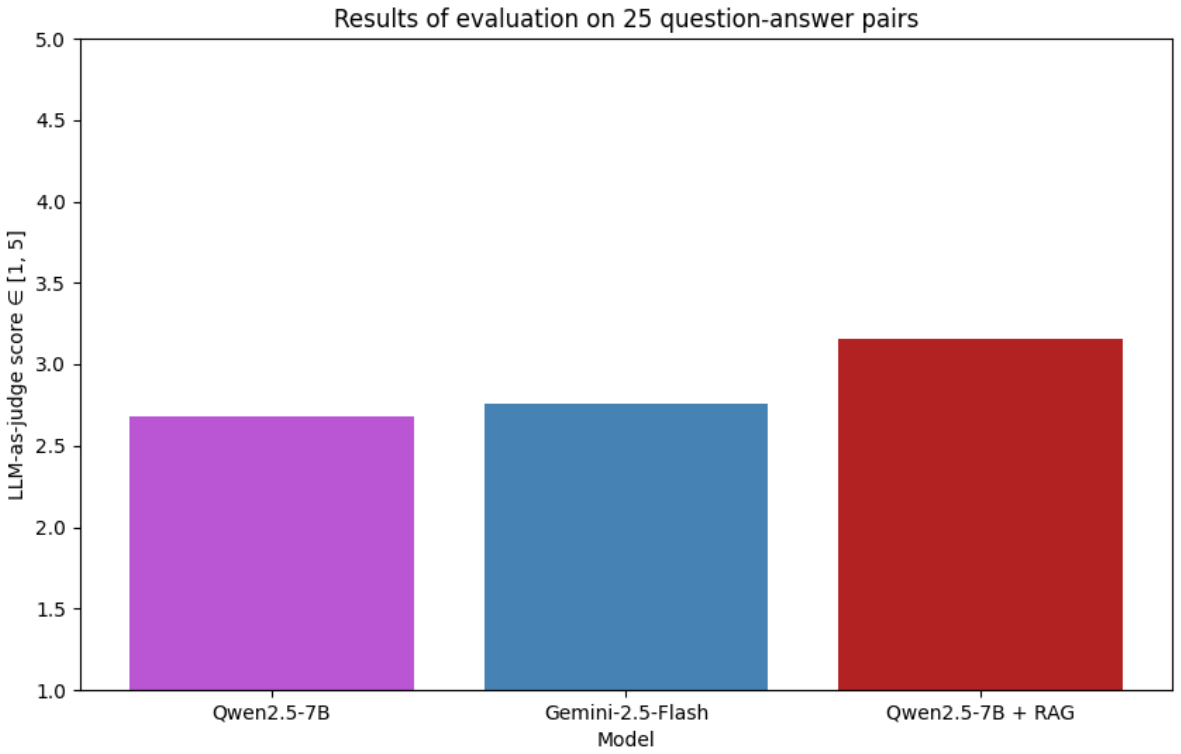


Semantic similarity comparison



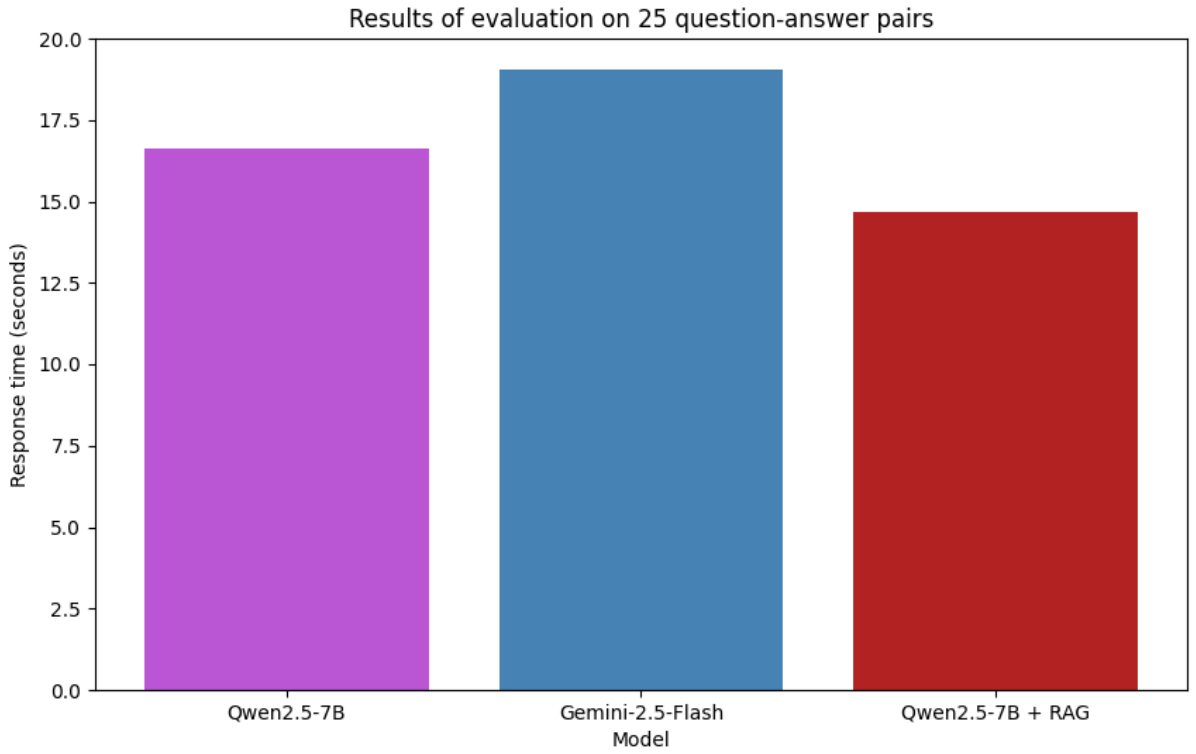
LLM-as-judge score comparison

- We should aim for 3.5 or higher (note that 5 is only given to almost exact matches)



Average response time comparison

- The RAG chatbot generates responses only based on the retrieved context!



Next steps

- Extend the dataset with more question-answer pairs
- Evaluate more base models (Mistral, Llama, Gemma)
- Optimize RAG parameters