Factual Recall in Financial QA: Between SLM model SALMONN vs. Text-only LM Vicuna

Kevin Wang, Sam Huang

Overview

Spoken language models (SLMs) process input in the form of audio rather than text, offering the potential to understand language in a more natural and human modality. However, does incorporating speech compromise the model's performance on core language tasks? How well do SLMs handle long-context reasoning compared to LLMs that operate solely on text?

In this Project, we investigate these questions through financial question answering (QA), originate from earnings call transcripts. We benchmark the performance of a state-of-the-art SLM, **SALMONN**, against a competitive text-only model, **Vicuna**, on tasks that require factual recall and multi-step reasoning.

Method

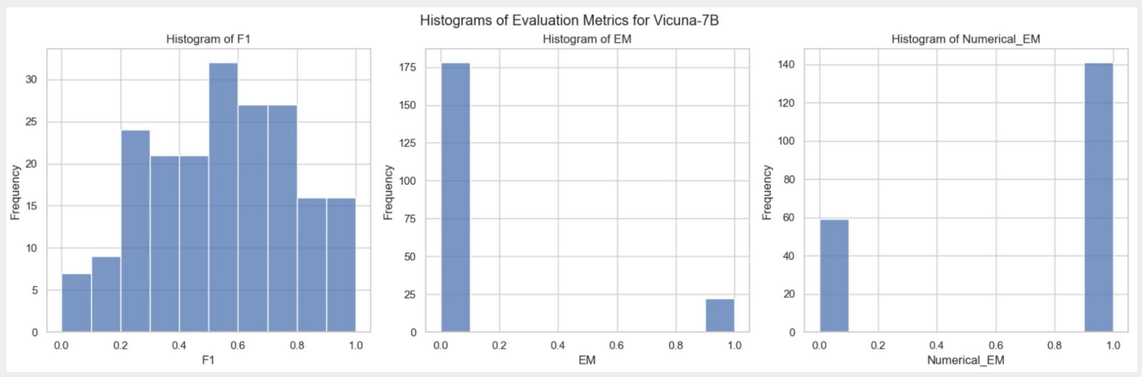
To compare the different capabilities of SLMs vs LLMs, we compare SALMONN-7b and Vicuna-7b on the Lamini Earnings Call dataset [1], containing:

- A *audio/transcript* of each earning call
- A *question* that asks about each transcript
- A natural language reference *answer*

We begin by applying traditional evaluation benchmarks and then extend our analysis to incorporate more nuanced, model-based judgment methods.

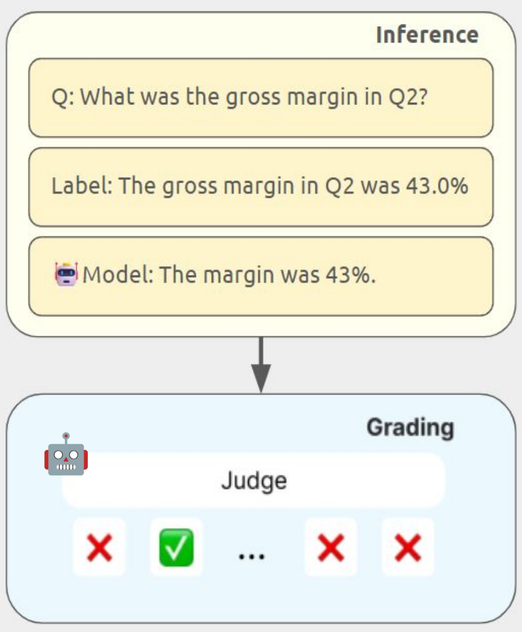
Task Type	Description	Example	Expected Answer	Evaluating Metrics
Factual Recall	Retrieve specific facts or values explicitly stated in context	"What was the net revenue in Q2 2022?"	$20 million	**Exact Match (EM)**, **F1 Score**, optionally **BERTScore**
Numerical Reasoning	Perform arithmetic or multi-hop reasoning over numeric facts	"How much did revenue increase from Q1 to Q2?"	$1 million (e.g., 20 - 19)	**Numerical EM**,

However, when we evaluate F1, EM, and Numerical EM, we found the result is not very direct and easy to compare.

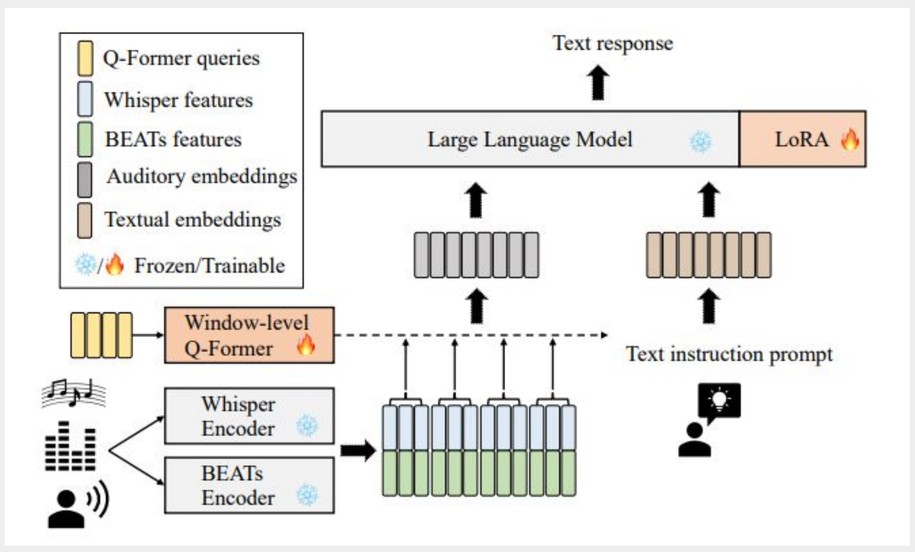


This is because,since the answer is in natural language, it is difficult to assign a binary correctness label

- We make an LLM judge using OpenAI's GPT-4.1 model.
- This method of using high-end models as "judges" for lesser models is common for LLM benchmarking research [2].



SALMONN & Vicuna



SALMONN (speech audio language music open neural network) is able to process speech and audio input, conditioned by a text prompt. [3] We use:

- Transcript audio as audio input
- Question as text prompt

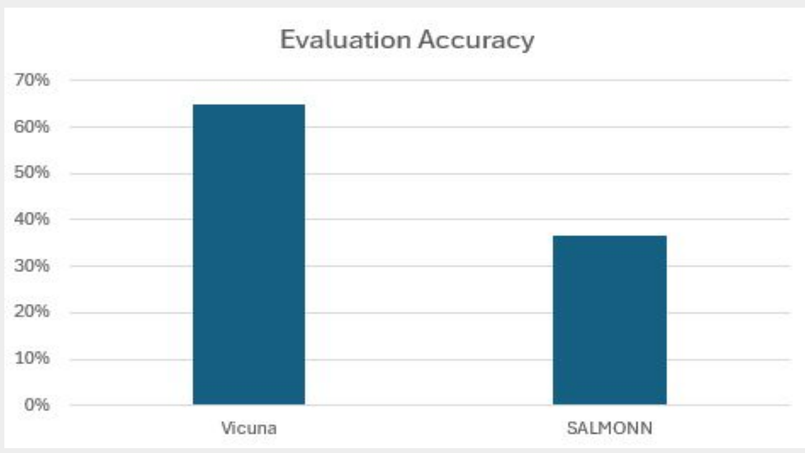
Vicuna-7B is an open-weight llm based on LLaMA-7B. We here use transcript text is passed in as input context. But question is often long and exceed the max token limit.

- We use retrieval-augmented generation (RAG):
  - The transcript is chunked.
  - A retriever (E5) ranks chunks based on similarity to the question.
    Top-ranked chunks are included in the Vicuna prompt to simulate long-context understanding.

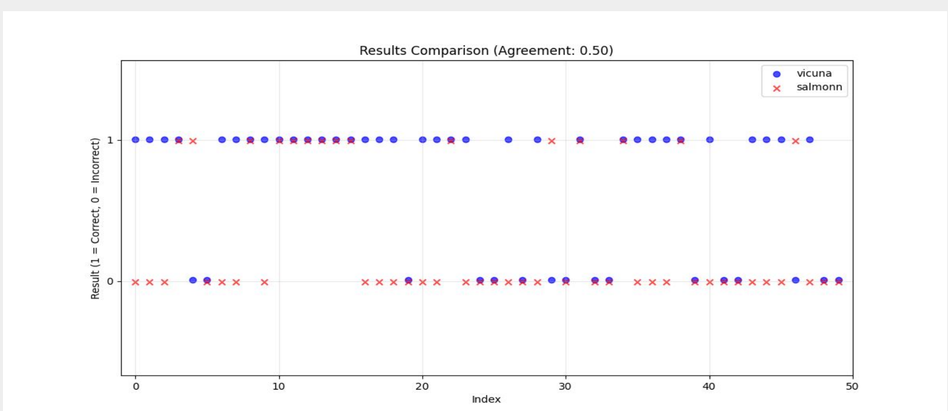
Results

Due to computation resource constraints, we tested on 200 samples of the 1000-sample dataset. Judging from GPT4.1

- Vicuna has a correct rate of 65%
- SALMONN has a correct rate of 36.5%



To understand why SALMONN performs worse, we analyze the correlation between the correctness of answers from the two models. However, the agreement rate is consistently close to random.



[1] lamini-ai earning calls, https://github.com/lamini-ai/lamini-earnings-calls
[2] Starace, Giulio, et al. "PaperBench: Evaluating AI's Ability to Replicate AI Research." *arXiv preprint arXiv:2504.01848* (2025).
[3] Tang, Changli, et al. "Salmonn: Towards generic hearing abilities for large language models." *arXiv preprint arXiv:2310.13289* (2023).