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

Kevin Wang[1], *Sam Huang*[1]

[1]University of Chicago

haochuanwang@uchicago.edu, binruih@uchicago.edu

Abstract

Spoken language models (SLMs) offer the promise of processing audio inputs directly, providing a more natural interface for human communication. Yet, their effectiveness on standard language tasks compared to traditional text-based language models (LLMs) remains underexplored. This paper investigates the factual recall and reasoning capabilities of a state-of-the-art SLM (SALMONN) against a competitive text-only model (Vicuna) in the domain of financial question answering (QA), where the inputs is drawn from lengthy and information dense earnings calls. We find that SALMONN significantly slack behind when dealing with factual recall with long contexts.

Index Terms: spoken language models, large language models, financial question answering, factual recall, long-context reasoning, retrieval-augmented generation, earnings calls, SALMONN, Vicuna

1. Introduction

Spoken language models (SLMs) extend the capabilities of conventional language models by processing input as audio rather than text alone. This shift in modality offers the promise of more natural, human-like understanding—capturing not only words but also the nuances of speech. However, it also raises key questions: does incorporating raw speech signal compromise performance on core language understanding tasks? And more specifically, how well do SLMs handle long-context reasoning and factual recall compared to text-only large language models (LLMs)?

In this project, we investigate these questions in the context of financial question answering (QA) by leveraging both transcripts and audio from corporate earnings calls. We benchmark SALMONN—a state-of-the-art SLM built on Vicuna-7B that directly ingests spoken input—against the original Vicuna-7B text-only model. Our evaluation targets tasks demanding precise factual recall and multi-step numerical reasoning, capabilities that are essential for extracting actionable insights from dense financial disclosures.

1.1. Nature of Financial Data

Earnings calls provide a particularly rich testbed. Held immediately after quarterly reports, these calls feature senior executives (typically the CEO or CFO) presenting performance metrics and fielding analysts' questions. Beyond raw numbers, the discussion offers qualitative insights into corporate strategy, market outlook, and leadership sentiment.

The very factors that make earnings calls valuable also make them challenging for LMs. Calls often exceed an hour in length, with key facts and figures scattered throughout. For instance, information about a company's revenue growth over the past two years may be dispersed across multiple conversations and QA sessions and information needed to answer this question spanning ten minutes or more. An effective model must not only locate these dispersed data points but also integrate them to form coherent, multi-step inferences.

Moreover, simply transcribing audio into text discards prosodic cues—tone, emphasis, and pacing—that convey confidence or caution. In one call, a CEO might exult, "We achieved 5 % growth last quarter!"; the transcript records "5 % growth," but often face the issue of missing the excitement and conviction in the speaker's voice. Later, when asked to forecast future performance, an SLM could leverage those vocal inflections to gauge management's true confidence—something a text-only model cannot. Altogether, earnings-call QA encompasses a variety of question types and reasoning challenges, making it an ideal benchmark to assess whether SLMs like SALMONN can match or surpass the performance of text-based models such as Vicuna in both information extraction and deep reasoning.

2. Related Works

2.1. Multimodal Spoken Language Models

Recent work has extended text-only large language models (LLMs) to handle spoken or audio inputs, enabling question answering (QA) directly from speech. SALMONN (Speech Audio Language Music Open Neural Network) integrates a Whisper speech encoder and a BEATs audio encoder with Vicuna-7B via a Q-Former, allowing the LLM to ingest raw audio tokens [1]. SALMONN can "listen" to audio and has achieved competitive performance on tasks such as auditory question answering (spoken QA) [1]. Notably, it also demonstrates emergent abilities not present in its training data—such as zero-shot spoken-query QA and multimodal co-reasoning—highlighting its capacity to handle open-ended spoken queries. However, one limitation of SALMONN is its input token window: because it is built on Vicuna-7B, the maximum context length is 2 048 tokens, so even with prompt engineering, the combined length of question and instruction cannot exceed roughly 1 500 words.

In this project, we run all models on Google Colab's NVIDIA A100 GPU. Due to RAM and VRAM constraints, we have not tested larger speech-enabled models, but the following represent promising avenues for future exploration. AudioChatLlama (Meta AI, 2024) extends Llama-2 with end-to-end speech understanding, accepting audio prompts in place of text and supporting sustained spoken dialogue [2]. Unlike earlier task-specific systems, AudioChatLlama handles general spoken QA, speech translation, and audio summarization in both closed and open domains. Similarly, Qwen-Audio (Alibaba, 2023) is a unified audio-language model built on a Whisper + Qwen architecture with large-scale multi-task training, and Qwen offers an 8 192-token context window—four times that of Vicuna—making it a worthwhile subject for further study.

2.2. Financial QA Datasets

To evaluate QA on long financial transcripts, new datasets have emerged. In this project, we use the Lamini Earnings Calls QA dataset (2023). This is a large collection of Q&A pairs generated from real earnings call transcripts, created by using an LLM to read transcripts and pose questions about the content. The dataset contains on the order of 100 thousands of QA pairs in 4GB covering various companies and financial metrics, and is openly available (CC BY-4.0) for fine-tuning models. The Lamini dataset provides a testbed closely aligned to the project at hand, as it features the same type of input: long CEO/CFO monologues and analyst Q&A sessions, with questions targeting facts like revenue figures, guidance, or strategic comments in the call. [3]

3. Dataset

We use the Lamini Earnings Call QA dataset, which contains short (3–10 minute) excerpts of earnings-call transcripts paired with audio. Each example in the dataset consists of:

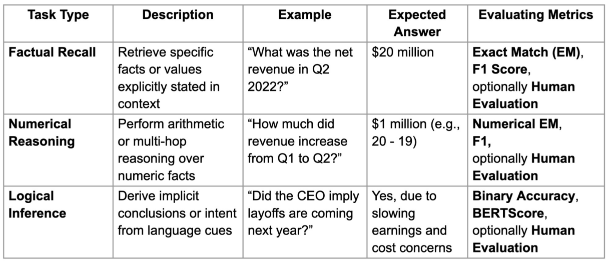
- An audio recording and corresponding transcript excerpt
- A natural-language question about that excerpt
- A reference answer in prose

Each question falls into one of three types, each requiring progressively deeper reasoning:

1. **Factual Recall.** The model must locate a specific fact or value explicitly stated in the transcript.
2. **Numerical Reasoning.** In addition to locating numeric information, the model must perform simple calculations (e.g. computing growth rates or differences).
3. **Logical Inference.** The model must draw implicit conclusions or evaluate performance based on contextual cues, requiring higher-level understanding and analysis.

For example, a logical-inference question might ask:“What is the expected delivery schedule for certain tools in TSMC’s 2023 capacity plan?” Even if no exact schedule is given, a correct answer could be:“The transcript does not state a precise delivery date. It notes that TSMC is working closely with tool suppliers to address supply-chain challenges and may defer some capex into 2023. ”

And to discourage irrelevant hallucinations, we also include “unanswerable” prompts. For instance, if asked “What were Apple’s earnings last quarter?” when presented with a Tesla call excerpt, the model should reply that the context does not provide that information.



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 **Human Evaluation**
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**, **F1,** optionally **Human Evaluation**
Logical Inference	Derive implicit conclusions or intent from language cues	“Did the CEO imply layoffs are coming next year?”	Yes, due to slowing earnings and cost concerns	**Binary Accuracy**, **BERTScore**, optionally **Human Evaluation**

Figure 1: *Evaluation Metrics*

4. Method and Evaluation

4.1. SALMONN

SALMONN (Speech Audio Language Music Open Neural Network) is a multimodal spoken-language model built on a Vicuna7B backbone. It integrates a Whisper-based speech encoder and a BEATs audio encoder via a Q-Former module, allowing the transformer decoder to attend jointly to raw audio tokens and text prompts. In our setup (Figure 2), each earnings-call excerpt is fed in two streams:

- *Audio input*: the call’s audio waveform, tokenized by the Whisper and BEATs encoders
- *Text prompt*: the natural-language question appended after a special audio-split token

This design enables SALMONN to leverage prosodic and paralinguistic cues—such as emphasis, speaker sentiment, and intonation—in addition to the literal transcript, which can improve performance on long-context reasoning and nuanced interpretation tasks.

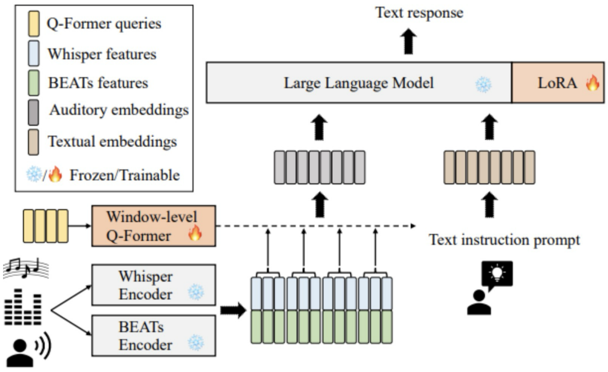


Figure 2: *SALMONN architecture: audio and text are jointly processed via a Q-Former into Vicuna-7B.*

4.2. Vicuna-7B with Retrieval-Augmented Generation

Vicuna-7B is a LLM trained from Meta’s LLaMA-7B that operates solely on text. In our text-only pipeline, we pass the transcript and question together as the input context. Since this often exceeds Vicuna’s 2 048-token limit, we adopt a retrieval-augmented generation (RAG) strategy:

1. Split the full transcript into overlapping chunks.
2. Use an E5 dense retriever to rank chunks by similarity to the question.
3. Concatenate the top-k chunks with the question to form Vicuna’s prompt.

This approach simulates long-context comprehension while respecting Vicuna’s token window. We evaluate Vicuna-7B on 200 QA pairs across factual recall, numerical reasoning, and logical inference, using Exact Match (EM), F1, BERTScore, Numerical EM, unit-consistency checks, and binary accuracy. However, based on the result, its still very difficult to interpret the results as each metrics has corner cases that fails to provide a standardized evaluation benchmark. For example EM is extremely strict and only counts if our answer is a exact same match as the preference answer. And for Numerical EM, we only check if the number present in the preference answer also shows in our output, which has other corner cases that our model fails to capture.

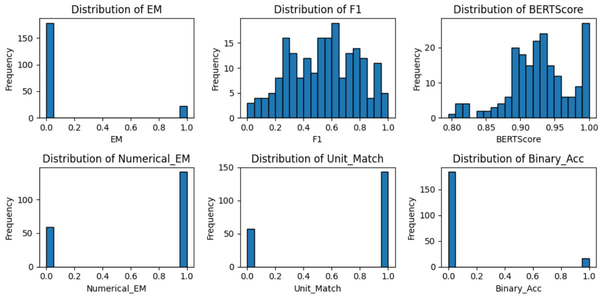


Figure 3: *Vicuna-7B performance on 200 QA pairs across the three task types.*

4.3. Evaluating Metrics

Since our answers are expressed in natural language, assigning a simple binary correctness label becomes difficult. For example, an answer of ”one thousand” and an answer of ”1,000” is equivalent, but appears as different words. To address this, we employ OpenAI’s GPT-4.1 as an automated judge, prompting it to evaluate and score each response. For each question, we provide the dataset label and the machine-generated response and ask GPT 4.1 if the two are equivalent. This method of using higher-end state-of-the-art models to evaluate the answers of smaller models is common practice in NLP benchmarking research [4]. We also manually screen 20 question and answer pairs and observe that the judge agrees with human judgment 100% of the time, making it a reasonable substitute for human evaluation on large a testing corpus.

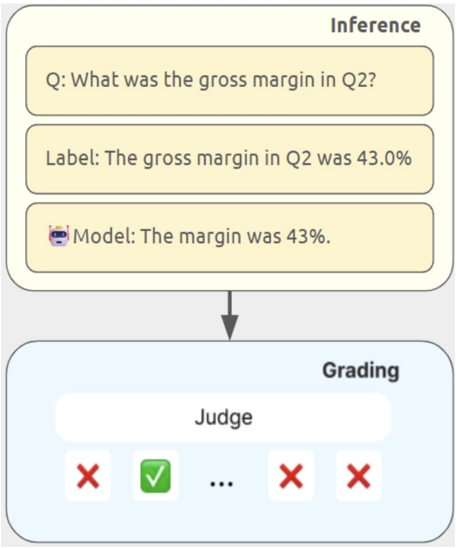

Inference
Q: What was the gross margin in Q2?
Label: The gross margin in Q2 was 43.0%
Model: The margin was 43%.
Grading
Judge
✗ ✓ ... ✗ ✗

Figure 4: *Evaluating our Answer with OpenAI’s GPT-4.1*

4.4. Performance

We evaluated both models on 200 QA pair and employed GPT-4.1 as an automated judge to adjudicate fine-grained correctness. Under these conditions, Vicuna-7B RAG achieved an overall accuracy of 65.0%, substantially outperforming SALMONN’s 36.5% on the same subset.

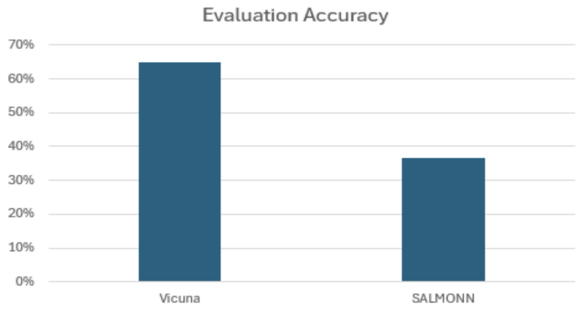


Figure 5: *Enter Caption*

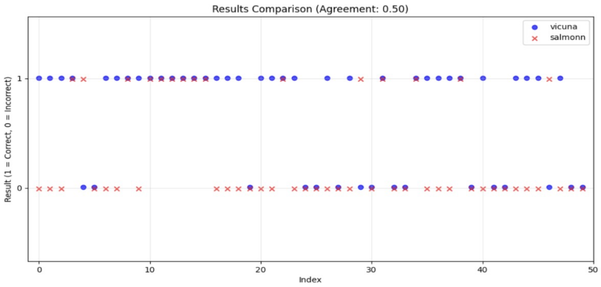


Figure 6: *Enter Caption*

5. Conclusions

Several factors explain this gap between 65% and 35%. Vicuna-7B RAG benefits from direct access to complete textual context once retrieved, whereas SALMONN’s Whisper encoder processes audio in 30-second windows, which can fragment long passages and impair cross-segment reasoning—especially for numerical and logical inference tasks. Furthermore, in our project, not satisfied with the edge cases in using traditional metrics (like the strictness of EM versus the leniency of Numerical EM), we used GPT 4.1 as judge. After manual screening the GPT 4.1 shows fair evaluation as judge, but we still need a more unified benchmarking framework for future works besides manual checking or using GPT 4.1 as judge.

Looking forward, we plan to (1) refine SALMONN’s audio segmentation to better aggregate information across longer contexts, finding the equivalent for text RAG for audio (2) explore hybrid fine-tuning of Vicuna on aligned audio–text pairs to enrich its embeddings with prosodic features, and (3) develop or adopt a standardized evaluation suite that balances strict string match, numerical tolerance, and semantic equivalence. And given that our research is run on google collab’s A100 GPU cluster, we didn’t explore with more computationally expensive model. But its worth trying with other SLM like Qwen and expanding our experiments to the full sample dataset and incorporating human annotations will further validate model performance and inform improvements for future multi-modal QA systems in finance.

6. Acknowledgements

This project was completed as part of TTIC 31110 at the University of Chicago. We sincerely thank Professor Karen Livescu and TAs Ju-Chieh Chou and Chung-Ming Chien for having us this quarter.

7. References

[1] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, “SALMONN: Towards generic hearing abilities for large language models,” arXiv preprint arXiv:2310.13289, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.13289

[2] Y. Fathullah, C. Wu, E. Lakomkin, K. Li, J. Jia, Y. Shangguan, J. Mahadeokar, O. Kalinli, C. Fuegen, and M. Seltzer, “AudioChatLlama: Towards general-purpose speech abilities for LLMs,” arXiv preprint arXiv:2311.06753, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.06753

[3] Lamini AI, “Lamini earnings calls qa dataset,” https://github.com/lamini-ai/lamini-earnings-calls, 2025, accessed: 2025-05-30.

[4] G. Starace, O. Jaffe, D. Sherburn, J. Aung, J. S. Chan, L. Maksin, R. Dias, E. Mays, B. Kinsella, W. Thompson *et al.*, “Paperbench: Evaluating ai’s ability to replicate ai research,” *arXiv preprint arXiv:2504.01848*, 2025.