M³FinMeeting: A Multilingual, Multi-Sector, and Multi-Task Financial Meeting Understanding Evaluation Dataset

Jie Zhu[1,2], **Junhui Li**[1*], **Yalong Wen**[2], **Xiandong Li**[3], **Lifan Guo**[2], **Feng Chen**[2]
[1]School of Computer Science and Technology, Soochow University
[2]Qwen DianJin Team, Alibaba Cloud Computing
[3]Nanjing University
zhujie951121@gmail.com, lijunhui@suda.edu.cn
{lifan.lg, betterman.chenf}@alibaba-inc.com

Abstract

Recent breakthroughs in large language models (LLMs) have led to the development of new benchmarks for evaluating their performance in the financial domain. However, current financial benchmarks often rely on news articles, earnings reports, or announcements, making it challenging to capture the real-world dynamics of financial meetings. To address this gap, we propose a novel benchmark called M³FinMeeting, which is a multilingual, multi-sector, and multi-task dataset designed for financial meeting understanding. First, M³FinMeeting supports English, Chinese, and Japanese, enhancing comprehension of financial discussions in diverse linguistic contexts. Second, it encompasses various industry sectors defined by the Global Industry Classification Standard (GICS), ensuring that the benchmark spans a broad range of financial activities. Finally, M³FinMeeting includes three tasks: summarization, question-answer (QA) pair extraction, and question answering, facilitating a more realistic and comprehensive evaluation of understanding. Experimental results with seven popular LLMs reveal that even the most advanced long-context models have significant room for improvement, demonstrating the effectiveness of M³FinMeeting as a benchmark for assessing LLMs' financial meeting comprehension skills.[1]

1 Introduction

Financial meetings, whether in person or virtual, serve as critical venues for decision-making, negotiation, and strategy formulation among various participants. Although large language models (LLMs) have demonstrated impressive performance across multiple natural language processing (NLP) tasks (OpenAI, 2023), it remains uncertain how they can effectively understand and process lengthy speech texts to help financial professionals expedite their work. Key capabilities, such as summarizing crucial points, responding to inquiries, and extracting question-answer pairs, are particularly beneficial for enhancing productivity and facilitating informed discussions in this context.

In the financial domain, there are various benchmarks available in different languages, such as FinQA (Chen et al., 2021) and ConvFinQA (Chen et al., 2022) in English, as well as CFLUE (Zhu et al., 2024) and the CCKS series shared tasks in Chinese (Tianchi, 2019, 2020, 2021, 2022). However, these datasets are primarily sourced from financial news and earnings reports, lacking content from real-world financial meetings. Additionally, they are monolingual, limited to English or Chinese. To address this gap, we introduce a new dataset called M³FinMeeting, designed for multilingual and multi-sector evaluation of financial meeting understanding, featuring multiple tasks. First, M³FinMeeting supports multiple languages, including English, Chinese and Japanese, enhancing the understanding of financial discussions across different linguistic contexts. Second, it encompasses all 11 industry sectors defined by Global Industry Classification Standard (GICS). Finally, M³FinMeeting includes multiple classical NLP tasks such as summarization, question answering, and question-answer pair extraction. Importantly, since financial meetings typically last one to two hours, M³FinMeeting provides long-context data, which is essential for assessing the ability of LLMs to handle complex tasks. Additionally, the tasks within M³FinMeeting require LLMs to generate long responses, allowing for a thorough assessment of their capabilities in producing coherent and relevant outputs in challenging scenarios.

Based on M³FinMeeting, we assess the effectiveness of seven representative LLMs including two OpenAI GPTs and five open-sourced LLMs. The experimental results indicate that Qwen2.5-72B-

*Corresponding Author
[1]We make our dataset and project available on https://github.com/aliyun/qwen-dianjin.

Findings of the Association for Computational Linguistics: ACL 2025, pages 244–266
July 27 - August 1, 2025

Instruct (Yang et al., 2024b) significantly outperforms other large language models (LLMs), achieving overall scores above 70 when evaluated by GPT-4 (OpenAI, 2023). However, this also suggests that even the most advanced LLMs currently available struggle with the tasks in M^{3}FinMeeting, revealing substantial rooms for performance improvement.

Our main contributions can be summarized as follows:

- M^{3}FinMeeting introduces a novel evaluation benchmark specifically designed for financial meetings, addressing the lack of real-world financial meeting data in existing benchmarks.

- M^{3}FinMeeting supports multilingual evaluation in English, Chinese, and Japanese, spans the 11 industry sectors defined by GICS, and includes three key NLP tasks: summarization, question-answer pair extraction, and question answering. All documents are carefully annotated by financial analysts to ensure high-quality and accurate evaluation.

- Through extensive experiments and detailed analyses, we thoroughly assess the performance of state-of-the-art long-context LLMs, offering valuable insights into their limitations and potential improvements for better long-context modeling in financial scenarios.

2 Related Work

2.1 Financial Evaluation Benchmarks

Financial NLP has become a key application area for LLMs, attracting increasing attention for its benchmarks. Table 1 summarizes recent benchmarks in this field. In English, FINQA by Chen et al. (2021) contains 8,281 question-answering pairs, with their numerical reasoning, from the earnings reports of S&P 500 companies. ECTSum by (Mukherjee et al., 2022) contains 2,425 transcripts of earnings calls, paired with short telegram-style bullet point summaries. FLUE by (Shah et al., 2022) and FLARE by Xie et al. (2024) offer heterogeneous benchmarks with various financial NLP tasks from existing datasets, including financial sentiment detection (Malo et al., 2014), named entity recognition (Alvarado et al., 2015), news headline classification (Sinha and Khandait, 2020), question answering (Maia et al., 2018), boundary detection tasks (Au et al., 2021), text summarization (Zhou et al., 2021; Mukherjee et al., 2022), and stock movement prediction (Xu and Cohen, 2018; Wu et al., 2018; Soun et al., 2022). FinTextQA by Chen et al. (2024) includes 1,262 long-form QA pairs from finance textbooks and government websites. BizBench by Krumdick et al. (2024) features eight quantitative reasoning tasks from professional exams, earnings reports, and other financial sources. FINANCEBENCH by Islam et al. (2023) includes 10,231 question-answer-evidence triplets from public filings and earnings reports.

In Chinese, the CCKS series has released several datasets specifically designed for various event extraction tasks (Tianchi, 2019, 2020, 2021, 2022). Additionally, significant effort has gone into creating evaluation datasets for document-level extraction from financial announcements, judgments, and news articles. These include DCFEE (Yang et al., 2018), DuEE-Fin (Han et al., 2022), Doc2EDAG (Zheng et al., 2019), and CFinDEE (Zhang et al., 2024a). Moreover, both BBT-CFLEB dataset (Lu et al., 2023) and CFLUE dataset (Zhu et al., 2024) serve as heterogeneous benchmarks that cover a wide range of NLP tasks, including multiple-choice question answering and reasoning, news and text classification, summarization, relation extraction, sentiment classification, question answering, and reading comprehension.

In addition to the benchmarks mentioned above, several datasets have been developed in the financial domain to support the training of financial LLMs, such as FLANG (Shah et al., 2022), Pixiu (Xie et al., 2024), InvestLM (Yang et al., 2023b), FinGPT (Yang et al., 2023a), and DianJin-R1 (Zhu et al., 2025). However, it is important to note that most of these datasets primarily rely on sources such as financial news, announcements, filings, and earnings reports. In contrast, our focus is on financial meetings, which provide unique insights and discussions that are often absent from existing benchmarks.

2.2 Summarization and Question-Answering Benchmarks

Summarization Benchmarks. Recent studies (Goyal et al., 2022; Zhang et al., 2024b; Pu et al., 2023) indicate that human preferences strongly favor summaries produced by LLMs over those generated by fine-tuned models or even reference summaries. This highlights the need to create new datasets that can comprehensively evaluate the summarization abilities of LLMs.

Language	Dataset	Sources	Tasks
English	FinQA (Chen et al., 2021)	Earnings reports	Question answering & reasoning
	ConvFinQA (Chen et al., 2022)	Reports	Question answering & reasoning
	FLUE (Shah et al., 2022)	Multiple	5 tasks
	ECTSum (Mukherjee et al., 2022)	Call transcripts	Summarization
	FLARE (Xie et al., 2024)	Multiple	8 tasks
	FinTextQA (Chen et al., 2024)	Long reports	Question answering
	BizBench (Krumdick et al., 2024)	Multiple	8 quantitative reasoning tasks
	FinanceBench (Islam et al., 2023)	Public filings, Earnings reports	Question answering
Chinese	CCKS Tianchi (2019, 2020, 2021, 2022)	Multiple	Multiple event-related tasks
	DCFEE (Yang et al., 2018)	Announcements	Document-level event extraction
	Doc2EDAG (Zheng et al., 2019)	Announcements	Document-level event extraction
	DuEE-Fin (Han et al., 2022)	Multiple	Document-level event extraction
	CFinDEE (Zhang et al., 2024a)	News articles	Document-level event extraction
	BBT-CFLEB (Lu et al., 2023)	Multiple	6 tasks
	CFLUE (Zhu et al., 2024)	Multiple	Multiple-choice Question Answering & Reasoning and 5 tasks
English, Chinese, Japanese	$\mathtt{M^3FinMeeting}$ (Ours)	Meetings	Summarization, Question answering, QA pair extraction

Table 1: Summary of recent financial benchmarks.

For instance, SumSurvey by Liu et al. (2024a) is a new dataset focused on summarizing long scientific survey papers. REFINESUMM by Patil et al. (2024) is tailored for image-text multimodal summarization. MovieSum by Saxena and Keller (2024) comprises 2,200 pairs of movie screenplays and their corresponding summaries. Additionally, HeSum by Paz-Argaman et al. (2024) is a novel benchmark dataset specifically designed for the low-resource Hebrew language. Our research also emphasizes long context, but we concentrate on meetings as our primary source, allowing us to analyze communication dynamics within this particular area. Several meeting speech summarization benchmarks are available, such as AMI (Carletta et al., 2005) and ICSI (Janin et al., 2003), which consist of English-language video (or audio) recordings of meetings, typically spanning tens of hours.

Question-Answering Benchmarks. Question Answering (QA) is a key task in NLP, with increasingly challenging benchmarks being developed, including those focused on the financial domain. For example, several benchmarks have been introduced to assess the long context understanding capabilities of LLMs (Li et al., 2024; Wang et al., 2024; Bai et al., 2024b). Additionally, many multimodal QA benchmarks have emerged to assess LLMs' capabilities in detecting and retrieving relevant information from various inputs, such as images, videos, texts, tables, and meetings (Li et al., 2023; Fu et al., 2023; Jin et al., 2024; Prasad et al., 2023).

3 $\mathtt{M^3FinMeeting}$: A Financial Meeting Benchmark

3.1 Overview

A formal financial meeting typically involves a few participants and lasts one to two hours. It promotes discussions among key participants, allowing for decision-making and strategic planning based on real-time information and reports. Participants express their opinions verbally, which distinguishes both the content and style of financial meetings from financial news, earnings reports, or announcements. The $\mathtt{M^3FinMeeting}$ benchmark, based on hundreds real financial meetings, includes three NLP tasks: summarization, question answering, and QA pair extraction. To reflect real-world sce-

Overview			
Language	**#Meeting**	**Avg hour**	**Avg token**
EN	100	0.96	10,086
ZH	400	1.15	11,740
JA	100	1.01	13,284
GICS Sector			
Sector	**Language**	**#Meeting**	**Avg token**
Com. Services	EN, ZH, JA	36	11,240
Con. Dis.	EN, ZH, JA	122	11,104
Con. Staples	EN, ZH, JA	52	11,699
Energy	EN, ZH, JA	32	15,841
Financials	EN, ZH, JA	49	12,835
Healthcare	EN, ZH, JA	47	13,937
Industrials	EN, ZH, JA	111	12,062
IT	EN, ZH, JA	98	13,430
Materials	EN, ZH, JA	32	11,393
RealEstate	EN, ZH, JA	13	15,270
Utilities	EN, ZH, JA	8	17,290
Length Set			
Length	**Language**	**#Meeting**	**Avg token**
Set1 (0-5K)	EN, ZH, JA	59	3,546
Set2 (5-10K)	EN, ZH, JA	164	7,509
Set3 (10-15K)	EN, ZH, JA	195	12,476
Set4 (15-20K)	EN, ZH, JA	124	17,419
Set5 (>20K)	EN, ZH, JA	58	25,281

Table 2: Data statistics of $\texttt{M}^3\texttt{FinMeeting}$ benchmark.

narios, we gather financial meetings from various sectors defined by GICS: Communication Services, Consumer Discretionary, Consumer Staples, Energy, Financials, Healthcare, Industrials, Information Technology (IT), Materials, Real Estate, and Utilities. Additionally, all audio recordings of the meetings are transcribed using a state-of-the-art automatic speech recognition (ASR) toolkit, followed by manual corrections. In total, $\texttt{M}^3\texttt{FinMeeting}$ includes 100 meetings in English (EN), 400 in Chinese (ZH), and 100 in Japanese (JA). Each meeting lasts an average of one hour. We employ the tiktoken tokenizer[2] to process all transcriptions. Table 2 presents detailed data statistics.

3.2 Evaluation Tasks

Given that financial meetings typically last one to two hours, users often need a summary and answers to specific questions. To address this, we propose three tasks that closely align with real-world needs.

3.2.1 Summarization

The summarization task aims to evaluate LLMs' ability to efficiently condense lengthy speeches while preserving the main ideas. Typically, these

[2] `https://platform.openai.com/tokenizer` (cl100k_base)

Language	AST	# SS	$\texttt{C}_{token}$	$\texttt{C}_{sent}$
EN	927	9.20	10.88	10.49
ZH	2,524	15.17	4.65	3.62
JA	1,149	8.24	11.56	11.92

Table 3: Statistics for the summarization task. Here, AST is for average summary token, #SS is for averaged number of section summary, $\texttt{C}_{token}$ and $\texttt{C}_{sent}$ for the compression ratio at token-level and sentence-level, respectively.

transcribed speech documents can be structured into sections based on discussion topics, with each section having its own summary, which we refer to section summary. We concatenate these section summaries sequentially to create a summary of the entire document. LLMs must implicitly identify and segment the document into various sections and then extract key point from each. Given transcribed speech documents and their reference summaries, we follow Koh et al. (2022) to compute the compression ratio of a source document length against its reference summary length at both token-level and sentence-level. As shown in Table 3, on average an English meeting contains 9.20 section summaries totaling 927 tokens.

3.2.2 QA Pair Extraction

The task of question-answer (QA) pair extraction involves identifying and extracting relevant QA pairs from transcribed financial meetings. This is crucial for analyzing discussions and making key insights readily accessible. To successfully perform this task, LLMs must recognize various types of questions posed during the meeting and accurately locate their corresponding answers. For example, questions like *What were we just talking about?* should be disregarded as they lack meaningful information. Additionally, participants may ask multiple questions at once, while responses may address them sequentially. This complexity requires LLMs to discern the structure of the dialogue, correctly pair each question with its answer. By employing natural language processing techniques, this task enhances information retrieval efficiency and facilitates subsequent analyses, providing structured and contextualized insights from the meeting. Table 4 shows statistics of QA pairs.

3.2.3 Question Answering

The question answering (QA) task evaluates the ability of LLMs to localize knowledge, which is essential for effective long-context processing (Wang

Lang.	$\texttt{QA}_{token}$	#QA	$\texttt{Q}_{token}$	$\texttt{A}_{token}$
EN	2,239	17.23	17.62	110.19
ZH	2,852	16.10	36.44	148.95
JA	2,934	10.84	34.55	178.99

Table 4: Statistics for both QA Pair Extraction and Question Answering Tasks. $\texttt{QA}_{token}$ is the average token length of all QA pairs per meeting, #QA is the average number of QA pairs per meeting, and $\texttt{Q}_{token}$ and $\texttt{A}_{token}$ denote the average token lengths of a question and answer, respectively.

et al., 2024). For simplicity, we use the QA pairs described above for this task. As mentioned, the transcribed speech text can be divided into multiple sections, the QA task tests the LLMs' capability to find evidence within that designated section, while other sections with similar but unrelated content act as noise. This setup ensures a focused assessment of the models' information retrieval skills.

3.3 Benchmark Construction

3.3.1 Data Collection

We have established four criteria for the manual collection of audio files from financial meetings, covering a range of events such as public roadshows, brokerage strategy meetings, industry exchanges, and earnings presentations: (1) Timeliness: Most meetings should be from recent years; (2) Length: Preference is given to longer audio files; (3) Categorizability: The audio files must align with categories defined in the GICS; (4) Authoritativeness: All audio files are sourced from our financial firm partners and are protected by our copyright.

All audio files are transcribed into text using ASR toolkit Whisper,[3] followed by a thorough manual correction process. Strict measures are used to ensure that no sensitive or personally identifiable information is included in the transcripts.

3.3.2 Annotation Process and Quality Control

The datasets for manual annotation are drawn from various projects that recruit experienced analysts fluent in different languages. These annotators follow the annotation guidelines (See Figure 10 in Appendix F), undergo comprehensive training, and have access to onboarding and guidance materials. Additionally, other analysts review the annotators' work and provide ongoing feedback.

- Transcribed Speech Correction: Annotators correct ASR-generated transcriptions with original audio files available for reference during the correction process.
- Summarization: The annotators segment each document into sections based on distinct topics, ensuring that only sections with clear boundaries are selected for summarization. For each valid section, annotators are encouraged to use simple sentences in the summary. Prior to the annotation process, annotators undergo extensive training and discussions to achieve a high level of consensus. In the annotation process, the original audio files are available for reference.
- QA Pair Extraction: With corrected speech documents, annotators manually extract financially relevant questions from the text. For each identified question, they search for corresponding answers in the subsequent content. Only questions with valid answers are retained for further analysis. This meticulous process guarantees the quality and relevance of the extracted QA pairs, significantly enhancing the dataset's value for financial insights.

Figure 4 in Appendix A shows a screenshot of an annotated example. For additional complete examples, please refer to the attached file.

4 Experimentation

4.1 Experimental Settings

Models. We evaluate seven advanced long-context LLMs with context windows ranging from 16K to 1000K, including two API-based LLMs: GPT-4o-2024-08-06-128K (OpenAI, 2023) and GPT-3.5.turbo-0125-16K, as well as five open-source LLMs: GLM4-9B-Chat-1000K (Zeng et al., 2022), Llama3.1-8b-Instruct-128K (Dubey et al., 2024), Qwen2-7B-chat-128K (Yang et al., 2024a), Qwen2-72B-Instruct-128K, and Qwen2.5-72B-Instruct-128K (Yang et al., 2024b). All models support the languages in $\texttt{M}^3\texttt{FinMeeting}$.

Prompts. We evaluate LLMs in a zero-shot setting. For summarization, we prompt the LLMs to implicitly identify document sections and generate individual summaries, which are then combined into a final document summary. For QA pair extraction, we first prompt the LLMs to extract all questions, then provide answers for each sequentially. For question answering, instead of addressing one

[3]https://openai.com/index/whisper/

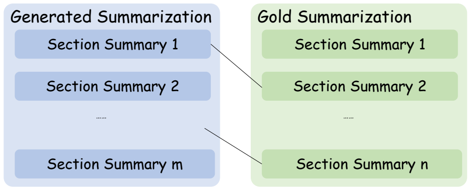


Figure 1: Illustration of the summarization evaluation.

question at a time, we combine related questions into a single prompt, allowing the LLM to produce a comprehensive response that includes all the answers. This better aligns with real-world tasks, like writing reviews or reports, and reduces API calls. Prompt examples are provided in Appendix C.

Metrics. For summarization, as shown in Figure 1, the final summary consists of multiple section summaries. We evaluate this using precision, recall, and F1 scores rather than traditional metrics like BLEU and ROUGE, as they better reflect how well the generated section summaries align with the reference (gold) summaries. Specifically, we align the automatic and gold section summaries using a cosine similarity score above 0.75, calculated with the OpenAI Embedding model[4]. Appendix B gives details of the three metrics. Meanwhile, recent studies (Wang et al., 2023; Zhang et al., 2023; Liu et al., 2024b) show that GPT-4 (OpenAI, 2023) aligns closely with human evaluations. Therefore, we use GPT-4 (gpt-4-turbo-2024-04-09) as the judge (GPT-4-Judge) to evaluate document-level summaries based on five criteria: Coverage, Redundancy, Readability, Accuracy, and Consistency, with scores ranging from 0 to 100. We report the average GPT-4-Judge score, based on the prompt in Appendix C.

For QA pair extraction, we assess the quality of generated questions using precision, recall, and F1 scores, comparing them to the reference questions, similar to the summarization evaluation. Additionally, GPT-4-Judge evaluates the generated QA pairs against the gold pairs using the same five criteria, and we report the average score.

For question answering, we group all questions together and prompt the LLMs to answer them sequentially, repeating each question before generating its corresponding answer. Therefore, we evaluate performance using the same metrics as in QA pair extraction, assessing both the quality of repeated questions and the overall performance of the generated QA pairs.

For the three tasks, Appendix D presents performance metrics using BLEU (Papineni et al., 2002) and ROUGE (Lin and Hovy, 2002). Appendix E includes detailed performance across languages, lengths, and GICS sectors. To address potential self-bias in LLM evaluations (e.g., GPT-4-Judge favoring GPT-4-generated answers), we follow Bai et al. (2024a) and use Qwen-plus[5] as an alternative judge model. Results, shown in Section 4.3, indicate minimal bias, as the performance trend from GPT-4-Judge and Qwen-plus-Judge are very consistent. Moreover, we perform a human evaluation and calculate Fleiss' Kappa (Scott, 1995) to measure agreement between GPT-4-Judge and human annotators, further supporting our conclusions. The results are presented in Section 4.4.

4.2 Results

Main Results. Table 5 shows the main results over all meetings.[6] From it, we have the following observations:

- **Overall performance**: The seven LLMs can be divided into three groups. Group 1 consists of Qwen2.5-72B-Instruct, Qwen2-72B-Instruct, and GPT-4o, all achieving an overall GPT-4-Judge score near or above 70.0. Among these, Qwen2.5-72B-Instruct performs best, followed by GPT-4o and Qwen2-72B-Instruct, which deliver comparable results. Group 2 includes Qwen2-7B-Instruct and GLM4-9B-Chat, both scoring around 60.0. Group 3 consists of GPT-3.5-turbo and LLaMA3.1-8B-Instruct, with LLaMA3.1-8B-Instruct outperforming GPT-3.5-turbo.

- **Summarization**: The precision, recall, and F1 scores for section-level summaries are all below 30%, indicating poor alignment between generated and gold section-level summaries. These low scores suggest that LLMs struggle both with semantic accuracy and document segmentation.

- **QA Pair Extraction**: The low precision, recall, and F1 scores suggest poor alignment between generated and gold questions.

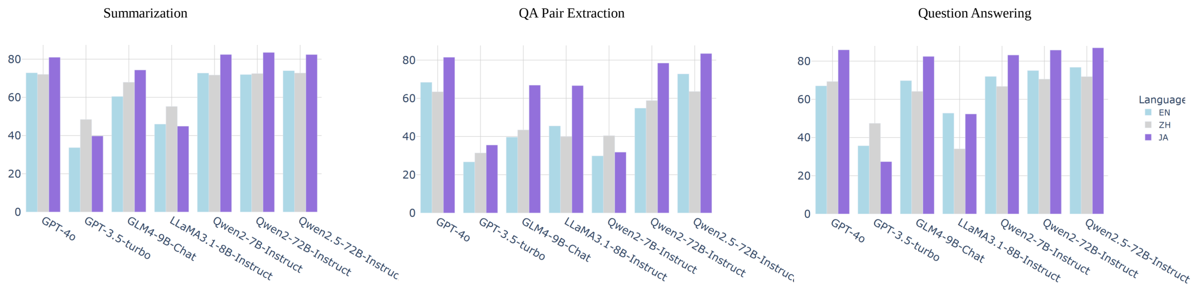
[4] https://platform.openai.com/docs/models/embeddings (text-embedding-3-small)

[5] https://help.aliyun.com/zh/model-studio/developer-reference/what-is-qwen-llm

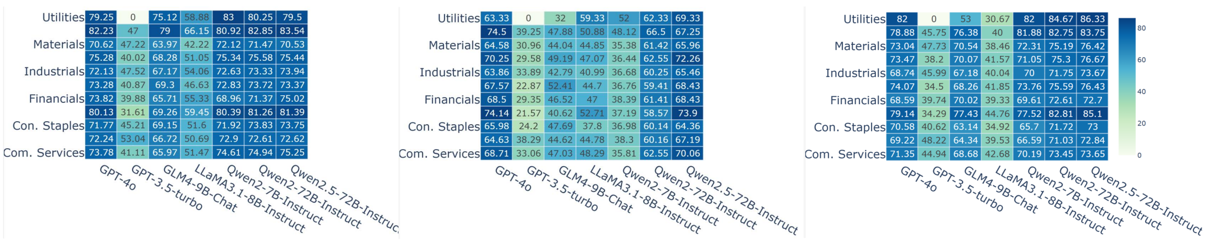
[6] The total cost for OpenAI API calls was about $2,500, while experiments with other LLMs use eight NVIDIA A100/80G GPUs.

Model	Summarization				QA Pair Extraction				Question Answering				Overall
	P.	R.	F1	GPT-4	P.	R.	F1	GPT-4	P.	R.	F1	GPT-4	GPT-4
GPT-4o	27.82	12.07	16.83	73.61	23.33	41.98	29.99	66.85	**93.93**	93.16	93.55	71.79	70.66
GPT-3.5-turbo	17.13	9.39	12.13	44.56	9.66	25.07	13.95	31.13	84.16	92.60	88.18	42.78	39.55
GLM4-9B-Chat	10.30	11.26	10.76	67.71	15.08	5.22	7.76	46.06	93.37	92.20	92.78	67.72	60.76
LLaMA3.1-8B-Instruct	6.24	8.34	7.14	52.01	13.05	7.30	9.37	44.64	57.21	39.97	47.06	40.01	45.76
Qwen2-7B-Instruct	28.67	19.39	23.14	73.59	11.98	16.40	13.85	37.33	89.83	93.01	91.40	69.99	60.71
Qwen2-72B-Instruct	**29.59**	**20.18**	**23.99**	74.17	22.43	28.82	25.22	60.85	93.27	93.58	93.42	73.50	69.66
Qwen2.5-72B-Instruct	28.98	15.56	20.25	**74.51**	**32.61**	**45.65**	**38.41**	**68.03**	93.75	**93.59**	**93.66**	**74.81**	**72.54**

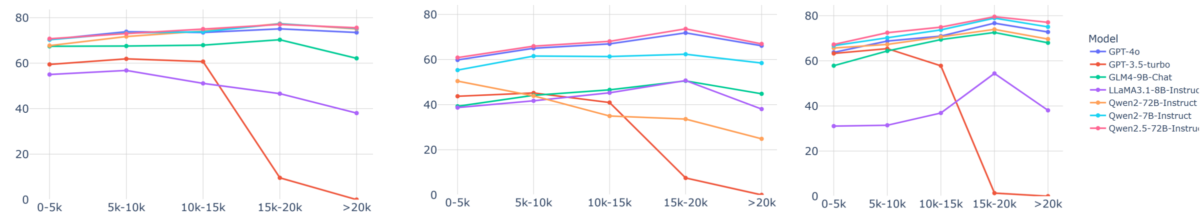
Table 5: Performance of LLMs on three evaluation tasks. The overall GPT-4-Judge score is the micro-average of its scores across all three tasks. Scores in **bold**/underline denote the top/second-best performances.



(a) GPT-4-Judge score over different languages.



(b) GPT-4-Judge score over different GICS sectors.



(c) GPT-4-Judge score over different input lengths.

Figure 2: Performance based on GPT-4-Judge scores across languages (a), GICS sectors (b), and input lengths (c).

For example, even the best-performing LLM, Qwen2.5-72B-Instruct, achieves only 45.65% recall, missing more than half of the gold questions. This highlights significant room for improvement in extracting relevant QA pairs.

- **Question Answering**: The performance of all LLMs—measured by precision, recall, F1, and GPT-4-Judge scores—is significantly higher than for QA pair extraction.[7] This discrepancy is not surprising, as in the question answering task, the questions are explicitly provided in the prompt. High F1 scores (over 90%) show that most LLMs can follow instructions well and properly repeat questions.

Effect over Different Languages. Figure 2 (a) shows the GPT-4-Judge scores across three languages. Most models perform best in Japanese, but there is no clear advantage in either Chinese or English. A closer look reveals that LLMs are more

[7]One exception is LLaMA3.1-8B-Instruct, which has difficulty following instructions and often fails to repeat the questions, resulting in lower GPT-4-Judge scores for the question answering task.

consistent in Japanese, likely because they adhere to the instructions more effectively in this language. The three Qwen models perform similarly in the Summarization task across all languages, followed by the Question Answering task. However, their performance varies most in the QA pair extraction task, where Qwen2.5-72B-Instruct obtains the highest scores, outperforming Qwen2-7B-Instruct with GPT-4-Judge scores of 72.76, 63.59, and 83.45 for English, Chinese, and Japanese, respectively.

Effect over Different Sectors. Figure 2 (b) compares model performance across sectors. Communication Services, Consumer Discretionary, and IT generally achieve higher GPT-4-Judge scores in summarization and question answering. However, the performance trend becomes more complex for the QA pair extraction task, showing increased variability across sectors. Overall, the performance gaps among sectors for GPT-4o, Qwen2-72B-Instruct, and Qwen2.5-72B-Instruct are much smaller compared to GPT-3.5-turbo and LLaMA3.1-8B-Instruct.[8] This suggests that the former models are less affected by sector differences.

Effect over Different Input Lengths. Figure 2 (c) compares the performance across varying input lengths. A key observation is the sharp drop in GPT-3.5-turbo's performance when the input exceeds 15K tokens, due to its 16K token context limit. In contrast, both Qwen2.5-72B-Instruct and GPT-4o demonstrate stable and competent performance across the three tasks, particularly excelling in handling longer contexts exceeding 15K tokens. On the other hand, Qwen2-72B-Instruct shows a declining trend in the QA pair extraction task as input length increases, indicating a reduced capability to maintain performance with longer inputs in this specific task. Future research could explore structured modeling, as outlined by (Zhu et al., 2019), to improve handling of long input contexts.

Effect of RAG-based Question Answering. Instead of prompting the LLMs to answer a list of questions in a single response, we also explore RAG-based question answering, where the LLM answers questions individually based on retrieved document chunks. Following Wang et al. (2024), we divide the document into 1,024-token chunks. For embedding, we utilize the OpenAI Embedding model. Figure 3 compares the performance of Qwen2.5-72B-Instruct before and after the integration of the RAG module. This comparison is based on a random selection of 10 meetings from each length set, resulting in a total of 50 meetings. The results indicate that for documents exceeding 15K tokens, answering all questions in a single response (Baseline 1) outperforms all other variants that answer questions one at a time. Additionally, for variants that respond to one question at a time in documents longer than 10K tokens, we observe that a larger context leads to better performance, specifically: Baseline 2 > RAG (top 5) > RAG (top 3) > RAG (top 1). Notably, RAG (with top 5) only surpasses the non-RAG variants for documents shorter than 10K tokens.

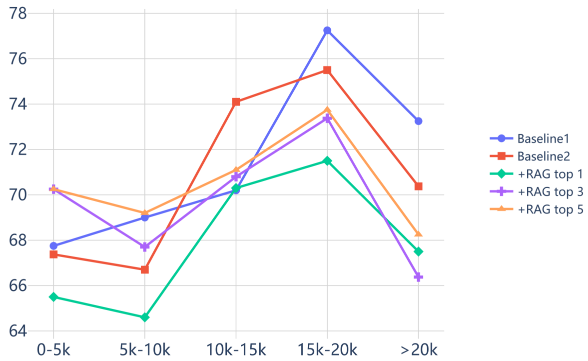


Figure 3: Performance (in GPT-4-Judge score) across different input lengths before or after adding RAG module. Baseline 1 refers to prompting the LLM to answer a list of questions, while Baseline 2 refers to answering one question per response.

4.3 Performance Evaluated by Qwen-plus-Judge

In addition to the performance assessed by the GPT-4-Judge, as presented in Table 5, Table 6 shows the performance evaluated by the Qwen-Plus-Judge. The prompt templates used for the Qwen-Plus-Judge are the same as those for GPT-4-Judge, as illustrated in Figure 8 and Figure 9. The performance trends in Table 6 align closely with those in Table 5, where Qwen2.5-72B-Instruct remains the top-performing model, followed by GPT-4o and Qwen2-72B-Instruct, which show similar performance. Next are Qwen2-7B-Instruct and GLM4-9B-Chat, both of which also exhibit comparable performance. This consistency further demonstrates that Qwen-Plus-Judge is a reliable alternative evaluator.

[8] GPT-3.5-turbo fails for the Utilities sector due to input length exceeding its 16K token limit.

Model	Summarization	QA Pair Extraction	Question Answering	Overall
GPT-4o	76.01	<u>67.02</u>	69.84	<u>71.12</u>
GPT-3.5-turbo	45.46	31.43	37.88	38.48
GLM4-9B-Chat	68.20	47.06	60.39	58.86
LLaMA3.1-8B-Instruct	51.46	43.90	32.05	42.75
Qwen2-7B-Instruct	75.39	42.65	67.87	62.38
Qwen2-72B-Instruct	<u>76.22</u>	65.58	<u>70.73</u>	71.01
Qwen2.5-72B-Instruct	**76.99**	**70.25**	**74.01**	**73.85**

Table 6: The results using Qwen-plus as the judge model.

Evaluation Method	Summarization	QA Pair Extraction	Question Answering
GPT-4-Judge	72.83	60.76	66.44
Human Annotators	3.68	3.13	3.36

Table 7: Comparison of average performance ratings between GPT-4-Judge and human annotators. Note that GPT-4-Judge uses a 1-100 scale, while human annotators use a 1-5 scale.

Agreement	Kappa
GPT-4-Judge & Human Annotators	0.701
Human Annotators	0.650

Table 8: Fleiss' Kappa score between GPT-4-Judge and human annotators.

4.4 Human Evaluation and Fleiss' Kappa Agreement Between GPT-4-Judge and Human Evaluators

We randomly select 100 meetings and recruit five expert human annotators to assess the overall quality of GPT-4o's responses, including summarization, QA pair extraction, and question answering. Each response is rated on a scale from 1 to 5, based on the same five criteria used to prompt GPT-4-Judge. The final score for each response is the average rating across the five annotators. Table 7 compares the performance assessed by GPT-4-Judge and human annotators.

Moreover, to assess the agreement among the six evaluators (five human annotators and GPT-4-Judge), we compute Fleiss' Kappa score, a metric for inter-annotator agreement. Specifically, we calculate two types of Kappa scores: 1) the agreement among all six evaluators, where GPT-4-Judge's ratings (on a 1-100 scale) are converted to a 1-5 scale by dividing by 20; and 2) the agreement among the five human annotators. As shown in Table 8, the agreement among GPT-4-Judge and human annotators is still higher than that among humans.

4.5 Performance in BLEU and ROUGE

The analysis of model performances, based on BLEU and ROUGE metrics, highlights notable differences across various tasks. Specifically, Qwen2-72B-Instruct excels in the summarization task, consistently generating concise and coherent summaries that underscore its strength in synthesizing information. Meanwhile, Qwen2.5-72B-Instruct leads in both QA pair extraction and question answering, demonstrating its adeptness at understanding queries and providing precise responses. This comparison underscores the unique advantages and task-specific expertise of each model, offering a more comprehensive insight into their capabilities. For detailed data and specific scores, please refer to the full table located in the appendix D.

5 Conclusion

In this paper, we have introduced M^{3}FinMeeting, a novel multilingual, multi-sector, and multi-task benchmark specifically designed to evaluate financial meeting understanding in large language models (LLMs). By incorporating real-world dialogue from financial meetings, our dataset fills significant gaps in existing benchmarks, which often rely on static sources like news articles and earnings reports. Supporting English, Chinese, and Japanese, and encompassing 11 industry sectors defined by GICS, M^{3}FinMeeting enables a comprehensive evaluation of LLMs through tasks such as summarization, question-answer pair extraction, and question answering. Experimental results with seven representative LLMs, including GPT-4o and Qwen2.5-72B-Instruct, demonstrate notable performance challenges, highlighting significant room for improvement and establishing M^{3}FinMeeting as a valuable resource for advancing research in financial language processing.

Acknowledgments

The authors express their heartfelt gratitude to the anonymous reviewers for their invaluable feedback and insightful comments, which greatly enhanced the quality of this work. We also offer our sincere appreciation to the members of the Qwen DianJin Team for their exceptional contributions, dedication, and hard work, which were instrumental in the success of this project. This work was supported by the Alibaba Innovative Research Program.

Limitations

Our work has several limitations.

- High Annotation Costs and Challenges: Both summarization and question-answer pair extraction require annotators to summarize content and extract question-answer pairs from audio recordings lasting 1-2 hours and text exceeding 10K tokens. This process relies heavily on the annotators' professional expertise. Summarization, in particular, demands a high level of skill from professional analysts and often involves open-ended responses. Consequently, the annotation process necessitates a significant investment of time and effort.

- Limited Coverage for Question Answering: In the question-answering task, we focus exclusively from the extracted QA pairs. This limits our ability to evaluate the LLMs' capacity to search for evidence within the document beyond the provided questions. As a result, the dataset may not fully capture the models' potential in more complex reasoning scenarios that require deeper comprehension of the content.

- Evaluation Issues: The performance of summary alignment in summarization and question alignment in both question-answer pair extraction and question answering relies on the embeddings used. Due to budget constraints, we have not conducted a more extensive evaluation, which may affect the robustness of our findings.

Ethics

Our dataset is sourced from publicly mandated disclosures, such as earnings calls and roadshows. While this information is publicly available, we anonymize it to respect corporate preferences for dissemination. This ensures both the academic utility and ethical integrity of our benchmark. Specifically, we remove company identifiers and sensitive details using GPT-4, followed by manual verification, to maintain privacy without compromising research value. All data annotators are part of the funded projects, ensuring consistent and responsible data handling. The dataset, excluding original audio files, will be available online.

References

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 84–90.

Willy Au, Abderrahim Ait-Azzi, and Juyeon Kang. 2021. Finsbd-2021: The 3rd shared task on structure boundary detection in unstructured text in the financial domain. In *Companion Proceedings of WWW*, page 276–279.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024a. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of ACL*, pages 7421–7454.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024b. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of ACL*, pages 3119–3137.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The ami meeting corpus: a pre-announcement. In *Proceedings of MLMI*, page 28–39.

Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. FinTextQA: A dataset for long-form financial question answering. In *Proceedings of ACL*, pages 6025–6047.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of EMNLP*, pages 3697–3711.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of EMNLP*, pages 6279–6292.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *Computing Research Repository*, arXiv:2407.21783.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Computing Research Repository*, arXiv:2306.1339.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *Computing Research Repository*, arXiv:2209.12356.

Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. 2022. Duee-fin: A large-scale dataset for document-level event extraction. In *Proceedings of NLPCC*, pages 172–183.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *Computing Research Repository*, arXiv:2311.11944.

Adam Janin, Don Baaron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chunk Wooters. 2003. The icsi meeting corpus. In *Proceedings of ICASSP*, pages 364–367.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of ACL*, pages 16077–16102.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.*, 55(8).

Michael Krumdick, Rik Koncel-Kedziorski, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. BizBench: A quantitative reasoning benchmark for business and finance. In *Proceedings of ACL*, pages 8309–8332.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. LooGLE: Can long-context language models understand long contexts? In *Proceedings of ACL*, pages 16304–16333.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023. M^3it: A large-scale dataset towards multimodal multilingual instruction tuning. *Computing Research Repository*, arXiv:2306.04387.

Chin-Yew Lin and Eduard Hovy. 2002. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of ACL*, page 311–318.

Ran Liu, Ming Liu, Min Yu, He Zhang, Jianguo Jiang, Gang Li, and Weiqing Huang. 2024a. SumSurvey: An abstractive dataset of scientific survey papers for long document summarization. In *Findings of ACL*, pages 9632–9651.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024b. Calibrating llm-based evaluator. In *Proceedings of LREC-COLING*, pages 2638–2656.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *Computing Research Repository*, arXiv:2302.09432.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of WWW*, page 1941–1942.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the American Society for Information Science and Technology*, 65:782–796.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of EMNLP*, pages 10893–10906.

OpenAI. 2023. Gpt-4 technical report. *Computing Research Repository*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, page 311–318.

Vaidehi Patil, Leonardo Ribeiro, Mengwen Liu, Mohit Bansal, and Markus Dreyer. 2024. REFINESUMM: Self-refining MLLM for generating a multimodal summarization dataset. In *Proceedings of ACL*, pages 13773–13786.

Tzuf Paz-Argaman, Itai Mondshine, Asaf Achi Mordechai, and Reut Tsarfaty. 2024. HeSum: a novel dataset for abstractive text summarization in Hebrew. In *Findings of ACL*, pages 6378–6388.

Archiki Prasad, Trung Bui, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt, and Mohit Bansal. 2023. MeetingQA: Extractive question-answering on meeting transcripts. In *Proceedings of ACL*, pages 15000–15025.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *Computing Research Repository*, arXiv:2309.09558.

Rohit Saxena and Frank Keller. 2024. MovieSum: An abstractive summarization dataset for movie screenplays. In *Findings of ACL*, pages 4043–4050.

William A. Scott. 1995. Reliability of content analysis: The case of nominal scale coding. *The Public Opinion Quarterly*, 19:321–325.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of EMNLP*, pages 2322–2335.

Ankur Sinha and Tanmay Khandait. 2020. Impact of news on the commodity market: Dataset and results. *Computing Research Repository*, arXiv:2009.04202.

Yejun Soun, Jaemin Yoo, Minyong Cho, Jihyeong Jeon, , and U Kang. 2022. Accurate stock movement prediction with self-supervised learning from sparse noisy tweets. In *Proceedings of CKIM*, pages 1691–1700.

Tianchi. 2019. Ccks2019 financial domain document-level event entity extraction dataset. https://tianchi.aliyun.com/dataset/111237.

Tianchi. 2020. Ccks2020 financial domain document-level event entity extraction dataset. https://https://tianchi.aliyun.com/dataset/111209.

Tianchi. 2021. Ccks2021 financial domain event causal relationship extraction dataset. https://tianchi.aliyun.com/dataset/110901.

Tianchi. 2022. Ccks2022 financial domain few-shot event extraction dataset. https://tianchi.aliyun.com/dataset/136800.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-qa evaluation. In *Proceedings of NeurIPs*, page 36.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *Computing Research Repository*, arXiv:2406.17419.

Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. 2018. Hybrid deep sequential modeling for social text-driven stock prediction. In *Proceedings of CKIM*, pages 1627–1630.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2024. Pixiu: a large language model, instruction data and evaluation benchmark for finance. In *Proceedings of NerIPS*, pages 33469–33484.

Yumo Xu and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of ACL*, pages 1970–1979.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. 2024a. Qwen2 technical report. *Computing Research Repository*, arXiv:2407.10671.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. 2024b. Qwen2.5 technical report. *Computing Research Repository*, arXiv:2412.15115.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. DCFEE: A document-level Chinese financial event extraction system based on automatically labeled training data. In *Proceedings of ACL: System Demonstrations*, pages 50–55.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023a. Fingpt: Open-source financial large language models. *Computing Research Repository*, arXiv:2306.06031.

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023b. Investlm: A large language model for investment using financial domain instruction tuning. *Computing Research Repository*, arXiv:2309.13064.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *Computing Research Repository*, arXiv:2210.02414.

Tian Zhang, Maofu Liu, and Bingying Zhou. 2024a. Cfindee: A chinese fine-grained financial dataset for document-level event extraction. In *Companion Proceedings of the ACM Web Conference*, pages 1511–1520.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024b. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. Wider and deeper llm networks are fairer llm evaluators. *Computing Research Repository*, arXiv:2308.01862.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. In *Proceedings of EMNLP-IJCNLP*, pages 337–346.

Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the event: Corporate events detection for news-based event-driven trading. In *Findings of ACL-IJCNLP*, pages 2114–2124.

Jie Zhu, Qian Chen, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. 2025. Dianjin-r1: Evaluating and enhancing financial reasoning in large language models. *arXiv preprint arXiv:2504.15716*.

Jie Zhu, Junhui Li, Yalong Wen, and Lifan Guo. 2024. Benchmarking large language models on CFLUE - a Chinese financial language understanding evaluation dataset. In *Findings of ACL*, pages 5673–5693.

Jie Zhu, Junhui Li, Muhua Zhu, Longhua Qian, Min Zhang, and Guodong Zhou. 2019. Modeling graph structure in transformer for better AMR-to-text generation. In *Proceedings of EMNLP*, pages 5459–5468, Hong Kong, China.

A Examples of $\mathtt{M^3FinMeeting}$

For brevity, Figure 4 presents a screenshot of an annotated $\mathtt{M^3FinMeeting}$ example. Complete examples are available in the attached file.

B Details of Precision, Recall and F1 Metrics

For summarization, as shown in Figure 1, we align the generated and gold section summaries based on a cosine similarity score above 0.75. Let us assume that there are m section summaries in the generated summarization and n section summaries in the reference (gold) summarization. After aligning the section summaries between the two, let m_a represent the number of generated section summaries aligned, and n_a represent the number of gold section summaried aligned. Note that m_a and n_a do not need to be equal, as one section summary from one side can be aligned to multiple section summaries on the other side. We compute the Precision, Recall, and F1 as follows:

$$\text{Precision} = \frac{m_a}{m}, \tag{1}$$

$$\text{Recall} = \frac{n_a}{n}, \tag{2}$$

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

The Precision, Recall, and F1 scores for QA pair extraction and question answering are computed similarly, with the generated and gold section summaries replaced by generated and gold questions.

C Prompt Examples for Tasks in $\mathtt{M^3FinMeeting}$

We use the same prompt templates, written in English, for English, Japanese, and Chinese. In these templates, we clearly specify that the output language must align with the meeting content.

Figure 5 and Figure 6 show the prompt template for the summarization task and the QA pair extraction task, respectively. Figure 7 displays the prompt template for the question answering task. In this template, all questions from the document are listed, and the LLMs are instructed to answer them in a single response.

When evaluating with GPT-4-Judge, Figure 8 presents the prompt template for the summarization task. Meanwhile, Figure 9 displays the prompt template for both the QA pair extraction task and the question-answering task.

D Performance in BLEU and ROUGE

Table 9 shows the overall performance in BLEU and ROUGE. It shows that Qwen2-72B-Instruct achieves the best performance in summarization while Qwen2.5-72B-Instruct leads in both QA pair extraction and question answering.

E Detailed Performance Across Languages, Lengths, and GICS Sectors

Table 10, Table 11, Table 12 show the performance of LLMs with respect to the languages, the length sets, and the GICS sectors, respectively.

F Annotation Guidelines

Figure 10 shows the annotation guideline of $\mathtt{M^3FinMeeting}$.

```
{
    "Meeting_ID": "M_EN_001",
    "GICS_Industry": "Industrials",
    "Language": "EN",
    "ASR_Text":
    {
        "content": "Honorable Investors, good evening. Thank you for attending the [Company1] 2022 Annual Results Conference. I am [Role1].
    },
    "Structured_Summary":
    {
        "content":
        [
            {
                "text": "In 2022, [Company1] achieved revenue of 267.5 billion, with a 29% year-on-year growth, and a net profit of 6.17 bil
            },
            {
                "text": "Consumer business growth was 17%, accounting for 52%, driven primarily by sectors like apparel, footwear, beauty, a
            },
            {
                "text": "Local business significantly reduced losses, improving net profit by 8.2 percentage points. International business
            },
            {…},
            {…}
        ]
    },
    "QA_List":
    {
        "content":
        [
            {
                "question": "I am [Role1] from [Company1], thankful for leaders' time and insights on the time-sensitive express business. T
                "answer": "Respected questioner, thank you for your focus. As outlined earlier, from both personal and company perspectives,
            },
            {
                "question": "I am [Role3] from [Company2], expressing gratitude to [Company1] executives [Role4], [Role5], [Role6], and [Rol
                "answer": "For this inquiry, let's have [Role9] respond. Firstly, regarding the overarching industry landscape, mergers like
            },
            {…},
            {…},
            {…},
            {
                "question": "Dear [Company1] leaders, good evening. I'm [Role18] from [Company6]. I'd like to ask about [Company1] and [Comp
                "answer": "At a strategic level, business core positioning aims not to align cycles to maintain group steadiness, as reflect
            },
            {
                "question": "Could the company provide us with guidance over this year's performance as a reference? Specifically, during Q1
                "answer": "Thank you for the inquiry. The question pertains to yearly and first-quarter directionality, addressed separately
            }
        ]
    }
},
```


Figure 4: Example of annotated M[3]FinMeeting.

Model	**Summarization**				**QA Pair Extraction**				**Question Answering**			
	B-4	**R-1**	**R-2**	**R-L**	**B-4**	**R-1**	**R-2**	**R-L**	**B-4**	**R-1**	**R-2**	**R-L**
GPT-4o	9.47	49.32	19.02	29.23	1.58	8.40	3.57	5.94	5.21	27.53	11.50	19.65
GPT-3.5-turbo	6.03	35.02	12.87	20.05	0.45	2.38	1.02	1.71	2.84	16.55	6.27	11.72
GLM4-9B-Chat	9.27	47.35	18.89	29.59	1.78	5.26	2.59	3.80	5.80	28.01	11.66	20.27
LLaMA3.1-8B-Instruct	4.77	36.37	11.82	21.38	1.16	3.97	1.93	2.98	2.76	14.20	5.98	10.58
Qwen2-7B-Instruct	14.72	55.63	24.12	35.98	0.34	2.00	0.75	1.36	6.67	29.78	11.87	20.04
Qwen2-72B-Instruct	**14.89**	**56.35**	**24.39**	**36.37**	1.65	6.66	3.04	4.81	6.90	30.39	12.22	21.15
Qwen2.5-72B-Instruct	11.71	51.92	20.84	32.76	**2.71**	**11.32**	**5.04**	**7.91**	**7.44**	**32.35**	**13.65**	**22.43**

Table 9: Performance (in BLEU and ROUGE) of LLMs on three evaluation tasks. Here B-4 is for BLEU-4, R-1/2/L for ROUGE-1/2/L.

Role:
You are a meeting content analyst responsible for summarizing the core information from meetings. Please analyze based on the following requirements.

Skill:
- Content Focus: Concentrate on substantive discussions, decisions, and key data from the meeting, ignoring procedural content, opening introductions, and host speeches.
- Information Extraction: Extract important financial metrics, market dynamics, management strategies, and interactive content from the provided meeting text. Digital information should be quoted accurately from the original text without any unauthorized additions.
- Formatted Output: Present the summary results in JSON format, ensuring that each summary element is complete and informative.
- Level of Detail: Ensure that the summary content is rich, and consolidate similar themes into a single paragraph to cover as many key insights as possible.

Constraints:
- Ignore procedural content, opening introductions, and host remarks that lack substantial significance, and focus on substantive discussions and decisions.
- The content should be as thorough as possible, merging similar themes for clarity.
- The output language should match the input language.

Meeting Text:
{meeting_note}

OutputFormat:
- The output should follow the specified JSON structure:
```
[
  {{
   "text": "Extracted key summary paragraph"
  }},
  ...
]
```

Figure 5: Prompt template used for the summarization task.

Skills:
- Precise Information Extraction: Distinguishing and extracting each specific question and answer from extensive meeting records.
- Content Rewriting Ability: Transforming colloquial meeting dialogue into clear, written expression.
- Structured Data Output: Strictly organizing and outputting all Q&A pairs according to JSON format standards, satisfying detailed format requirements.

Constraints:
- Maintain the integrity and accuracy of the information, omitting no questions or answers.
- Rewrite content to streamline questions while ensuring the richness of the answers.
- Must adhere to the specified JSON output format.
- Output language must be consistent with the input meeting notes.

Workflow:
- First, thoroughly review the meeting content to identify and separate each independent question and response paragraph.
- Next, rewrite each paragraph to ensure questions are concise and answers are detailed and formal.
- Finally, consolidate the processed information into JSON format, forming an organized list of Q&A pairs.

Output Format:
- The output should be a JSON array containing multiple Q&A objects, each with "question" and "answer" sections, for example:
```json
[
    {{
        "question": "",
        "answer": ""
    }},
    {{
        # second Q&A pair
    }}
]
```

Meeting Notes
{meeting_notes}

Figure 6: Prompt template used for the QA pair extraction task.

```
## Role:
Meeting Content Analysis Expert: Possesses the ability to deeply understand lengthy meeting texts and can
accurately extract information to answer a list of questions related to the meeting.

## Skill:
-- Efficient Text Summarization: Quickly distills key information and details from meeting texts.
-- Question Response Matching: Precisely identifies relevant answers in the meeting content for each question on the
list.

## Constraints:
-- Input Limitation: Long meeting text and question list have been provided.
-- Output Requirement: Answers to each question in the list formatted in JSON.
-- Output Language: Output language must remain consistent with the language of the meeting text.

## Workflow:
-- First, conduct a comprehensive analysis of the meeting text to extract core points.
-- Then, meticulously compare the content of the meeting against each question in the question list to find
corresponding answers.
-- Finally, organize the answers to ensure they are direct and accurate for each question.

## Meeting Text:
{meeting_note}

## Question List:
{question_list}

## Output Format:
-- The output language must be consistent with that of the meeting text.
-- The output should follow the JSON structure below:
[
  {{
    "question": "First question",
    "answer": "Answer to the first question"
  }},
  ...
]
```

Figure 7: Prompt template used for the question answering task. All questions from a document are listed, and the LLMs are prompted to answer them in a single response.

```
## Role:
Evaluation Expert: A professionally competent AI assistant evaluation specialist, proficient in accurately assessing the quality of AI assistant responses based on established criteria, ensuring the accuracy and objectivity of the evaluation.

## Gold Answer:
{gold_answer}

## Assistant's Predicted Answer:
{predicted_answer}

## Skill:
- Precision Matching: Accurately identify semantic consistency, numerical and sequential accuracy, and the presence of hallucinations in the AI assistant's response based on a gold standard answer.
- Key Point Extraction: Quickly identify and verify whether the response includes all necessary key information to meet the completeness requirements of the question answered.
- Unbiased Evaluation: Provide comprehensive and impartial evaluation explanations, strictly following scoring criteria to deliver a fair score.

## Constraints:
- Evaluation Criteria: The assessment is limited to the following core aspects:
1. Coverage: The summary should include key information from the source document, ensuring a comprehensive reflection of the original content.
2. Redundancy: The summary should avoid unnecessary repetition or lengthy expressions, such as repeated sentences or overused noun phrases.
3. Readability: The summary should remain fluent and understandable, with clear logic and well-organized information, avoiding ambiguities.
4. Accuracy: The truthfulness and correctness of the information must be guaranteed, ensuring that numbers, facts, and descriptions are consistent with the source document.
5. Consistency: The information in the summary should be logically consistent throughout, avoiding contradictions or conflicting information.

## Workflow:
- Analyze and compare the assistant's response with the gold standard answer.
- Determine the relevant compliance level according to the established evaluation criteria.
- Provide the final score, adhering to the specified format for output.

## Output Format:
- Scores should be integers from 1 to 100.
- Strictly follow the specified JSON format without any additional explanations.
{{
  "coverage": "",      // Coverage score
  "redundancy": "",    // Redundancy score
  "readability": "",   // Readability score
  "accuracy": "",      // Accuracy score
  "consistency": "",   // Consistency score
  "overall": ""        // Overall score
}}
```

Figure 8: Prompt template utilized for assessing summarization with GPT-4.

```
## Role:
- Evaluation Expert: A professionally competent AI assistant evaluation specialist, proficient in accurately assessing the quality of Q&A pairs extracted by the AI assistant from meeting texts based on established criteria, ensuring the accuracy and objectivity of the evaluation.

## Gold Answer:
{gold_answer}

## Assistant's Predicted Answer:
{predicted_answer}

## Skill:
- Precision Matching: Accurately identify semantic consistency, numerical and sequential accuracy, and the presence of hallucinations in the Q&A pairs extracted by the AI assistant based on the Gold Answer.
- Key Point Extraction: Quickly identify and verify whether the Q&A pairs contain all necessary information and context to meet the completeness requirements of the question answered.
- Unbiased Evaluation: Provide comprehensive and impartial evaluation explanations, strictly following scoring criteria to deliver a fair score.


## Constraints:
- Evaluation Criteria: The assessment is limited to the following core aspects:
1. Coverage: The Q&A pairs should include key information present in the Gold Answer, maintaining the same quantity and ensuring a comprehensive reflection of the original content.
2. Redundancy: The Q&A pairs should avoid unnecessary repetition, lengthy expressions, or content exceeding that of the Gold Answer, remaining concise and clear.
3. Readability: The Q&A pairs should be fluent and understandable, logically clear, with a well-structured information layout, facilitating comprehension.
4. Accuracy: The truthfulness and correctness of the information must be guaranteed, ensuring that the facts and descriptions in the Q&A pairs are consistent with the Gold Answer.
5. Consistency: The question and answer in the Q&A pairs should be logically consistent throughout, avoiding contradictions or conflicting information, and align in language with the Gold Answer.

## Workflow:
- Analyze and compare the Q&A pairs extracted by the assistant with the gold standard Q&A pairs.
- Determine the relevant compliance level according to the established evaluation criteria.
- Provide the final score, adhering to the specified format for output.


## OutputFormat:
- Scores should be integers from 1 to 100.
- Strictly follow the specified JSON format without any additional explanations.
{{
  "coverage": "",      // Coverage score
  "redundancy": "",    // Redundancy score
  "readability": "",   // Readability score
  "accuracy": "",      // Accuracy score
  "consistency": "",   // Consistency score
  "overall": ""        // Overall score
}}
```

Figure 9: Prompt template utilized for assessing both QA pair extraction and question answering with GPT-4-Judge.

Model	Summarization				QA Pair Extraction				Question Answering				Overall
	P.	**R.**	**F1**	**GPT-4**	**P.**	**R.**	**F1**	**GPT-4**	**P.**	**R.**	**F1**	**GPT-4**	**GPT-4**
English													
GPT-4o	5.58	3.62	4.39	72.83	**40.12**	19.21	25.98	68.37	98.25	**98.25**	**98.25**	67.02	69.41
GPT-3.5-turbo	6.84	4.17	5.18	33.68	19.32	3.31	5.65	26.70	**98.58**	36.27	53.03	35.68	32.02
GLM4-9B-Chat	0.63	3.29	1.06	60.45	2.54	7.31	3.77	39.69	98.31	97.62	97.96	69.76	56.63
LLaMA3.1-8B-Instruct	1.37	1.65	1.49	45.98	5.78	9.05	7.06	45.53	87.09	85.02	86.04	52.75	48.04
Qwen2-7B-Instruct	13.15	12.84	12.99	72.71	3.91	2.03	2.67	29.85	97.75	86.01	91.50	71.95	58.16
Qwen2-72B-Instruct	**15.43**	**16.02**	**15.72**	71.92	15.26	10.38	12.36	54.85	98.31	94.89	96.57	75.03	67.26
Qwen2.5-72B-Instruct	6.40	4.72	5.51	**73.94**	39.81	**22.57**	**28.81**	**72.76**	98.22	96.45	97.33	**76.76**	**74.48**
Chinese													
GPT-4o	31.49	11.55	16.90	72.03	40.76	23.22	29.58	63.43	91.49	91.86	91.68	69.33	68.36
GPT-3.5-turbo	18.34	4.52	7.24	48.42	25.11	6.64	10.51	31.41	92.20	59.67	72.45	47.42	42.57
GLM4-9B-Chat	21.24	10.79	14.31	67.90	5.31	15.18	7.86	43.45	90.32	91.61	90.96	64.14	58.81
LLaMA3.1-8B-Instruct	7.72	9.03	8.32	55.23	6.61	12.46	8.64	39.87	30.30	48.19	37.21	34.07	43.38
Qwen2-7B-Instruct	28.69	16.22	20.72	71.68	20.89	12.74	15.83	40.47	91.66	88.38	89.99	66.75	59.95
Qwen2-72B-Instruct	30.21	**16.99**	**21.75**	72.48	32.20	25.97	28.75	58.85	92.02	91.83	91.92	70.57	67.43
Qwen2.5-72B-Instruct	**32.38**	14.63	20.16	**72.74**	**45.82**	**34.47**	**39.34**	**63.59**	**92.22**	**92.48**	**92.35**	**71.88**	**69.49**
Japanese													
GPT-4o	37.03	25.03	29.87	80.06	61.04	35.51	44.89	81.50	90.95	**90.86**	**90.91**	85.88	82.62
GPT-3.5-turbo	24.14	16.77	19.79	39.75	41.95	5.53	9.77	35.55	**91.98**	17.98	30.09	27.36	34.68
GLM4-9B-Chat	33.65	22.15	26.71	74.32	19.85	29.33	23.67	66.88	90.56	90.31	90.43	82.43	74.52
LLaMA3.1-8B-Instruct	3.12	6.51	4.22	44.90	39.34	26.56	31.71	66.66	91.07	90.40	90.74	52.30	53.79
Qwen2-7B-Instruct	**43.32**	**50.81**	**46.77**	82.42	13.92	6.64	8.99	31.82	91.07	90.40	90.74	83.17	67.20
Qwen2-72B-Instruct	41.41	48.93	44.86	**83.49**	36.87	25.00	29.79	78.46	90.89	90.22	90.55	85.78	82.65
Qwen2.5-72B-Instruct	35.67	34.91	35.29	82.40	**67.74**	**46.49**	**55.14**	**83.45**	89.75	89.66	89.70	**86.95**	**84.10**

Table 10: Performance of LLMs on three evaluation tasks with different languages.

Model	Summarization				QA Pair Extraction				Question Answering				Overall
	P.	**R.**	**F1**	**GPT-4**	**P.**	**R.**	**F1**	**GPT-4**	**P.**	**R.**	**F1**	**GPT-4**	**GPT-4**
						Set1 (0-5K)							
GPT-4o	25.00	28.52	26.64	69.75	36.90	24.29	29.30	**69.55**	93.70	**93.70**	**93.70**	64.33	65.25
GPT-3.5-turbo	18.25	28.52	22.26	58.76	27.06	15.29	19.54	44.38	93.70	**93.70**	**93.70**	62.67	55.89
GLM4-9B-Chat	23.10	26.17	24.54	66.76	14.27	16.19	15.17	39.20	**93.77**	90.26	91.98	57.33	56.24
LLaMA3.1-8B-Instruct	10.63	19.14	13.67	54.35	16.94	17.24	17.09	39.90	15.84	38.23	22.40	31.81	42.96
Qwen2-7B-Instruct	25.25	**39.84**	30.91	67.18	26.33	17.84	21.27	50.95	93.54	93.40	93.47	65.62	61.93
Qwen2-72B-Instruct	23.82	39.45	29.71	70.29	25.78	23.69	24.69	56.48	93.55	93.55	93.55	66.24	64.74
Qwen2.5-72B-Instruct	**28.02**	37.11	**31.93**	**70.40**	**41.60**	**33.43**	**37.07**	60.93	93.70	**93.70**	**93.70**	**67.07**	**66.74**
						Set2 (5-10K)							
GPT-4o	26.01	17.01	20.57	**73.51**	44.78	26.76	33.50	65.37	91.97	92.31	92.14	69.10	69.42
GPT-3.5-turbo	16.22	11.67	13.57	62.07	27.47	9.10	13.68	45.45	91.98	79.26	85.15	64.69	57.71
GLM4-9B-Chat	16.34	13.75	14.93	68.46	17.11	15.05	16.01	45.31	90.55	90.75	90.65	63.73	59.10
LLaMA3.1-8B-Instruct	9.90	12.22	10.94	56.25	4.63	13.27	6.87	41.96	25.73	42.29	31.99	30.91	43.95
Qwen2-7B-Instruct	25.59	26.94	**26.25**	71.97	21.98	14.79	17.68	43.90	91.90	88.48	90.16	66.61	61.47
Qwen2-72B-Instruct	25.41	**27.01**	26.19	73.39	37.96	30.70	33.93	61.36	92.20	92.20	92.20	70.09	68.57
Qwen2.5-72B-Instruct	**27.60**	22.29	24.66	72.87	**48.25**	**35.38**	**40.82**	**65.91**	**92.27**	**92.27**	**92.27**	**71.93**	**70.59**
						Set3 (10-15K)							
GPT-4o	29.80	11.30	16.39	73.34	43.45	24.04	30.96	**67.56**	91.90	**93.48**	**92.68**	72.54	70.49
GPT-3.5-turbo	17.76	6.59	9.61	61.43	22.18	6.31	9.83	38.86	**93.90**	65.30	77.03	56.94	53.22
GLM4-9B-Chat	10.15	9.92	10.03	68.80	3.03	17.21	5.15	48.28	91.86	**93.48**	92.67	68.94	61.42
LLaMA3.1-8B-Instruct	9.93	6.80	8.07	51.80	8.85	14.01	10.85	44.69	38.50	52.85	44.55	36.16	44.48
Qwen2-7B-Instruct	28.48	17.65	21.79	74.21	14.55	10.11	11.93	36.04	92.26	88.86	90.53	70.89	60.06
Qwen2-72B-Instruct	**30.68**	**18.97**	**23.44**	73.43	32.47	25.39	28.49	61.37	92.62	91.72	92.17	74.14	69.66
Qwen2.5-72B-Instruct	29.57	13.97	18.98	**74.41**	**47.71**	**35.19**	**40.50**	67.44	92.30	91.83	92.06	**75.89**	**72.69**
						Set4 (15-20K)							
GPT-4o	23.84	8.96	13.02	74.62	42.40	23.14	29.94	71.32	94.76	94.96	94.86	75.26	74.54
GPT-3.5-turbo	15.31	0.79	1.51	19.00	31.25	0.79	1.54	14.36	**100.00**	0.68	1.36	11.16	6.19
GLM4-9B-Chat	15.27	10.92	12. 73	67.41	8.49	15.01	10.85	50.09	93.82	94.81	94.31	71.86	64.54
LLaMA3.1-8B-Instruct	7.13	6.31	6.69	45.04	18.48	13.27	15.45	48.97	70.59	78.44	74.30	52.89	50.52
Qwen2-7B-Instruct	32.26	18.60	23.60	76.08	11.14	4.77	6.68	31.29	94.41	87.72	90.94	72.85	61.83
Qwen2-72B-Instruct	**34.69**	**19.87**	**25.27**	76.60	20.87	16.11	18.18	62.49	94.90	93.65	94.27	**76.97**	72.91
Qwen2.5-72B-Instruct	27.47	12.98	17.63	**76.67**	**48.23**	**33.58**	**39.59**	**73.37**	94.97	**95.02**	**94.99**	76.65	**76.73**
						Set5 (>20K)							
GPT-4o	37.81	9.24	14.86	75.01	44.14	17.68	25.25	65.42	92.36	89.04	90.67	69.99	70.91
GPT-3.5-turbo	0	0	0	0	0	0	0	0	0	0	0	0	0
GLM4-9B-Chat	3.90	8.02	5.25	65.84	6.18	11.42	8.02	41.12	90.19	92.27	91.22	70.89	58.63
LLaMA3.1-8B-Instruct	1.89	5.73	2.84	52.82	9.45	10.68	10.03	45.14	77.85	77.35	77.60	47.42	47.13
Qwen2-7B-Instruct	34.06	**11.99**	**17.74**	75.52	10.93	2.49	4.05	31.36	91.87	83.24	87.34	71.71	57.21
Qwen2-72B-Instruct	32.30	11.92	17.41	75.70	16.67	8.84	11.55	59.15	92.63	90.34	91.42	76.33	69.83
Qwen2.5-72B-Instruct	**34.93**	11.15	16.91	**76.79**	**46.97**	**25.69**	**33.21**	**68.59**	**93.09**	**93.09**	**93.09**	**78.75**	**73.28**

Table 11: Performance of LLMs on three evaluation tasks with different length sets.

Model	Summarization				QA Pair Extraction				Question Answering				Overall
	P.	R.	F1	GPT-4	P.	R.	F1	GPT-4	P.	R.	F1	GPT-4	GPT-4
Communication Services													
GPT-4o	23.72	11.02	15.05	73.78	37.23	24.70	29.70	68.71	94.02	96.23	95.11	71.35	71.41
Qwen2-72B-Instruct	23.05	**18.63**	**20.61**	74.94	30.57	26.11	28.17	62.55	**95.29**	**95.29**	**95.29**	73.45	70.55
Qwen2.5-72B-Instruct	**24.13**	14.69	18.27	**75.25**	**40.46**	**28.47**	**33.42**	**70.06**	**95.29**	**95.29**	**95.29**	**76.65**	**73.10**
Consumer Discretionary													
GPT-4o	23.82	10.85	14.91	72.24	41.86	22.16	28.98	64.63	93.04	94.49	93.76	69.22	68.74
Qwen2-72B-Instruct	**25.04**	**17.08**	**20.31**	72.61	29.95	23.38	26.26	60.16	**93.68**	**93.55**	**93.61**	71.03	68.00
Qwen2.5-72B-Instruct	22.06	12.34	15.83	**72.62**	**44.60**	**30.92**	**36.52**	**67.19**	93.17	93.83	93.51	**72.84**	**73.10**
Consumer Staples													
GPT-4o	**27.43**	10.90	15.60	71.77	41.91	18.51	25.68	65.98	91.27	**92.38**	91.82	70.58	69.47
Qwen2-72B-Instruct	26.63	**16.92**	**20.70**	**73.83**	32.41	22.08	26.27	60.14	91.76	91.07	91.41	71.71	68.63
Qwen2.5-72B-Instruct	24.34	13.05	17.23	73.75	**44.79**	**29.13**	**35.30**	**64.36**	**92.20**	92.29	**92.24**	**73.00**	**70.41**
Energy													
GPT-4o	48.60	29.09	36.40	80.13	**55.80**	38.54	45.59	**74.14**	**95.40**	**95.03**	**95.21**	79.14	78.12
Qwen2-72B-Instruct	50.53	**47.49**	**48.96**	81.26	26.93	25.19	26.03	58.57	95.38	94.65	95.01	82.81	75.17
Qwen2.5-72B-Instruct	**53.81**	40.13	45.97	**81.39**	52.88	**45.41**	**48.87**	73.90	**95.40**	**95.03**	**95.21**	**85.10**	**80.30**
Financials													
GPT-4o	24.13	8.68	12.77	73.81	41.34	29.62	34.52	**68.50**	**95.49**	**95.49**	**95.49**	68.59	70.37
Qwen2-72B-Instruct	**29.62**	**16.87**	**21.50**	71.37	28.38	24.31	26.19	61.41	94.47	90.82	92.61	72.61	68.52
Qwen2.5-72B-Instruct	26.92	12.15	16.75	**75.02**	**46.91**	**36.71**	**41.19**	68.43	95.41	90.49	92.89	**72.70**	**72.11**
Healthcare													
GPT-4o	24.00	9.25	13.36	73.28	41.51	25.72	31.76	67.57	**93.52**	**93.52**	**93.52**	74.07	71.63
Qwen2-72B-Instruct	28.67	**16.97**	**21.32**	**73.72**	22.22	20.71	21.44	58.41	93.41	89.48	91.40	75.59	69.57
Qwen2.5-72B-Instruct	**29.07**	12.76	17.73	73.37	**44.52**	**37.54**	**40.73**	**68.43**	93.37	**93.52**	93.45	**76.43**	**72.74**
Industrials													
GPT-4o	27.15	10.56	15.20	72.13	43.81	23.08	30.23	63.86	92.22	89.85	91.02	68.74	68.28
Qwen2-72B-Instruct	28.93	**18.15**	**22.30**	73.33	31.95	23.60	27.15	60.25	92.52	92.37	92.44	71.75	68.50
Qwen2.5-72B-Instruct	**29.35**	13.53	18.52	**73.94**	**50.84**	**36.27**	**42.34**	**65.46**	**92.53**	**92.53**	**92.53**	**73.67**	**71.05**
Information Technology (IT)													
GPT-4o	23.82	11.02	15.07	75.28	46.21	23.66	31.30	70.25	92.25	**93.07**	92.66	73.47	73.05
Qwen2-72B-Instruct	**28.78**	**21.13**	**24.37**	**75.58**	29.32	20.02	23.79	62.55	92.59	91.72	92.15	75.30	71.25
Qwen2.5-72B-Instruct	27.38	16.29	20.42	75.44	**49.45**	**32.00**	**38.85**	**72.26**	**92.66**	92.77	**92.72**	**76.67**	**74.80**
Materials													
GPT-4o	30.17	13.71	18.85	70.63	42.80	27.95	33.81	64.58	91.34	91.56	91.45	73.04	69.5
Qwen2-72B-Instruct	**31.63**	**25.00**	**27.92**	**71.47**	38.03	29.87	33.46	61.42	**91.58**	**91.80**	**91.69**	75.19	69.51
Qwen2.5-72B-Instruct	29.36	18.54	22.73	70.53	**49.67**	**36.86**	**42.32**	**65.96**	91.56	91.56	91.56	**76.42**	**70.94**
Real Estate													
GPT-4o	**66.66**	38.16	48.54	82.23	38.80	40	39.39	**74.50**	89.70	**93.84**	91.72	78.88	**79.17**
Qwen2-72B-Instruct	48.00	36.64	41.55	82.85	16.92	16.92	16.92	66.50	**93.84**	**93.84**	**93.84**	82.75	77.35
Qwen2.5-72B-Instruct	61.29	**43.51**	**50.89**	**83.54**	**48.27**	**43.07**	**45.52**	67.25	**93.84**	**93.84**	**93.84**	**83.75**	79.10
Utilities													
GPT-4o	**57.77**	43.33	49.52	79.25	39.28	22.44	28.57	63.33	**83.67**	**83.67**	**83.67**	82.00	76.42
Qwen2-72B-Instruct	48.64	**60.00**	**53.73**	**80.25**	22.72	20.40	21.50	62.33	**83.67**	**83.67**	**83.67**	85.67	77.35
Qwen2.5-72B-Instruct	53.70	48.33	50.87	79.50	**65.95**	**63.26**	**64.58**	**69.33**	**83.67**	**83.67**	**83.67**	**86.33**	**78.78**

Table 12: Performance of LLMs on three evaluation tasks with different GICS sectors. To save space, we only report the performance of three LLMs.

Annotation Guidelines

1. Purpose

These guidelines are intended to provide financial analysts with a clear framework for consistently annotating audio transcripts from financial roadshow meetings, extracting comprehensive and professional structured summaries and Q&A information. All information should be saved in JSON format to ensure consistency and repeatability.

2. Annotation Process

2.1 Listen to the Audio

Complete Listening: Analysts should listen to the entire meeting audio to ensure a full understanding of the content.

Note-taking: While listening, jot down key themes and important details for use in subsequent annotation steps.

2.2 Load ASR Text

Obtain and Verify: Obtain the automatic speech recognition (ASR) text of the audio.

Text Proofreading: Compare the ASR text with the actual audio to identify and correct apparent errors.

2.3 Professional Structured Summary Annotation

Theme and Information Extraction

Independent Analysis: Two analysts independently extract important themes and relevant key information from the text.

Themes and Information: Ensure each theme contains its key information.

JSON Format Saving

Save Format: Format each theme and its key information into a JSON structure, with each JSON element representing a thematic paragraph.

Merging and Review

Senior Analyst Review: A senior analyst reviews and consolidates the two reports to ensure completeness and accuracy.

Annotation Steps

Create Theme Object: Create a JSON object for each theme, specifying the theme name.

Record Key Information: Save all crucial information entries in the "Key Information" array.

2.4 Q&A Extraction Annotation

Question and Answer Identification

Mark Interaction: Mark the Q&A interaction parts in the text, ensuring the questioner and answerer are identified (e.g., speaker and audience).

Summarization and Extraction

Refine Summaries: Extract and summarize each Q&A interaction from the meeting.

JSON Format Saving

Save Format: Save the Q&A pairs in JSON format, with each Q&A pair being an individual JSON element.

Merging and Review

Comprehensive Integration: Two analysts independently annotate and a senior analyst consolidates the final results.

Annotation Steps

Create Q&A Object: Create a JSON object for each Q&A pair, recording the questioner and answerer.

Record Content: Record detailed content in the "Question" and "Answer" fields, respectively.

3. Annotation Check

Review and Confirm: A senior analyst reviews the annotations to ensure consistency, accuracy, and completeness.

Communicative Adjustment: If inconsistencies arise, the senior analyst should discuss with annotators to finalize the standard version.

4. Notes

Maintain Neutrality: Stay objective during the annotation process, avoiding personal views.

Team Collaboration: Record and discuss difficult-to-judge information with the team promptly.

Data Confidentiality: Ensure compliance with confidentiality and data protection regulations for the meeting content.

Figure 10: Annotation guidelines of M^{3}FinMeeting.