Query-OPT: Optimizing Inference of Large Language Models via Multi-Query Instructions in Meeting Summarization

Anonymous ACL Rolling Review Submission

Abstract

This work focuses on the task of query-based meeting summarization in which the summary of a context (meeting transcript) is generated in response to a specific query. When using Large Language Models (LLMs) for this task, a new call to the LLM inference endpoint/API is required for each new query even if the context stays the same. However, repeated calls to the LLM inference endpoints would significantly increase the costs of using them in production, making LLMs impractical for many real-world use cases. To address this problem, in this paper, we investigate whether combining the queries for the same input context in a single prompt to minimize repeated calls can be successfully used in meeting summarization. In this regard, we conduct extensive experiments by comparing the performance of various popular LLMs: GPT-4, PaLM-2, LLaMA-2, Mistral, and FLAN-T5 in single-query and multi-query settings. We observe that while most LLMs tend to respond to the multi-query instructions, almost all of them (except GPT-4), even after fine-tuning, could not properly generate the response in the required output format. We conclude that while multi-query prompting could be useful to optimize the inference costs by reducing calls to the inference endpoints/APIs for the task of meeting summarization, this capability to reliably generate the response in the expected format is only limited to certain LLMs.

1 Introduction

One of the key advantages of recent LLMs is demonstrating strong emergent abilities[1] (Wei et al., 2022a), which helps them to achieve impressive instruction-following capabilities even in zero-shot scenarios (Laskar et al., 2023a; Qin et al., 2023; Bang et al., 2023). These emergent abilities of LLMs have also led to an increase in utilizing LLMs in numerous real world use cases (Laskar et al., 2023b). However, despite their impressive performance, deploying LLMs in the real world is not trivial, with one of the primary obstacles being significant inference costs. Thus, optimizing the inference cost alongside ensuring high accuracy and efficiency is important in real world settings.

While several optimization techniques (Zhu et al., 2023), such as knowledge distillation, post-training quantization, etc. are utilized to minimize the cost associated with LLMs in production, these techniques cannot be applied to the closed-source LLMs like GPT-4 (OpenAI, 2023) or PaLM-2 (Google, 2023). For open-source LLMs (Touvron et al., 2023), these techniques may come with different trade-offs. For instance, while applying quantization (Zhu et al., 2023) can reduce the GPU requirement and make it possible to do LLM inference on less expensive hardware, it may also result in slower inference speed and reduced accuracy.

Another limitation is that the cost associated with LLM inference depends on the number of tokens processed by LLMs, which is true for both closed-source[2] and open-source LLMs. This issue can be addressed by reducing either the context size or the number of calls to the inference endpoints. For the task of query-focused meeting summarization, the latter is highly preferable as it can address the limits on API calls by enforcing the model to process larger contexts at once.

To address the above, in this paper, we study whether we can minimize calls to the LLM inference APIs/endpoints by optimizing the prompts. More specifically, we study whether we can combine the queries for the same context in a single prompt to minimize calls to the same inference endpoints for different query-context pairs. For this purpose, we created a new version of the QMSUM dataset for query-focused meeting summarization task (Zhong et al., 2021) by combin-

[1] The emerging behavior of language models refers to certain capabilities that are only present in larger language models but not in smaller language models (Wei et al., 2022a)

[2] https://openai.com/pricing

ing all queries for the same context in a single prompt. We fine-tuned several open-source models and compared their performance to the popular closed-source models on single-query and multi-query versions of the dataset. Our experimental results show that most open-source LLMs, even after fine-tuning, fail to properly follow multi-query instructions to generate the response in the required format. We also find a similar trend in many larger zero-shot LLMs. However, one notable exception in this regard is GPT-4, which demonstrates better instruction-following capabilities to generate summaries for multiple queries from meeting transcripts. Our extensive experiments give insights about which LLMs are capable of this emerging ability to follow multi-query instructions and their respective limitations to ensure the optimization of production cost in building real-world LLM-powered systems to handle multiple queries. Our major contributions are as follows:

(1) We conduct an extensive evaluation of various LLMs in the multi-query settings for query-focused meeting summarization to investigate their performance in following multi-query instructions and compare it to the single-query scenario.

(2) We observe that while most LLMs successfully respond to multiple queries in a single prompt, they could not generate the responses in the required format. To ensure fair evaluation, we propose a new evaluation criteria based on ROUGE where all the query-summary pairs for the same transcript are combined to compute the score.

(3) The findings from our experimental evaluation will provide insights into optimizing prompts to reduce production costs while deploying LLMs for real-world usage.

2 Related Work

The impressive instruction-following capabilities of LLMs have led to their utilization in the real world in various tasks, which includes generating summaries from meeting transcripts (Laskar et al., 2023b). However, in many scenarios, users may require extracting other information from the transcripts rather than a generic summary of the meetings. In such cases, one straightforward way is to call the LLM inference API/endpoint for the given query-transcript pair. However, this approach is not cost-effective, since the same transcript for a different query would be given as input again to the LLM in different calls. Thus, it will lead to a non-optimal usage cost for processing the same tokens in a transcript multiple times.

One possible solution in this regard could be combining the queries in a single prompt, similar to the work of Laskar et al. (2023a) where they evaluated ChatGPT[3] (i.e., GPT-3.5) in the open-domain question-answering task in about 100 samples in the Natural Questions (Kwiatkowski et al., 2019) and the WebQuestions (Berant et al., 2013) datasets. While there evaluation demonstrates that instruction-following LLMs like GPT-3.5 can respond to multiple queries in a single prompt, they did not investigate the following research questions: (i) *Are LLMs capable of responding answers to a given input text that requires understanding long conversation context?* (ii) *Can LLMs generate the response in a specified format to ensure easier parsing of the LLM-generated output?* (iii) *Do smaller*[4] *open-source LLMs also posses this capability similar to larger closed-source LLMs like ChatGPT?*

To address the above questions, in this paper, we conduct a comprehensive evaluation of popular closed-source and open-source LLMs in the QMSUM (Zhong et al., 2021) dataset for query-focused meeting summarization task to investigate their performance in following multi-query instructions to extract information from long conversations.

3 Our Methodology

The objective of this research is to study whether LLMs are capable of following multi-query instructions to extract information from the given source text depending on the input queries. For this purpose, we utilize the QMSUM dataset (Zhong et al., 2021) and convert it to a multi-query instruction dataset for query-focused meeting summarization. The original dataset consists of query-transcript pairs, with the same transcript appearing multiple times for different queries. In our modified multi-query instruction version of the QMSUM dataset, we combine all the queries for the same input transcript to construct our prompt. More specifically, for the input transcript T, we combine the queries $q_1, q_2, \ldots q_n$ to construct the multi-query set Q. Then, we add the instruction I that explains the task as well as the required output format. This results in the reduced number of samples in the multi-query version of QMSUM: **162/35/35** instances in train/valid/test sets, whereas the original dataset has **1257/272/281** instances, respectively.

[3] https://openai.com/chatgpt

[4] The smaller open-source LLMs were not evaluated for multi-query instructions by (Laskar et al., 2023a).

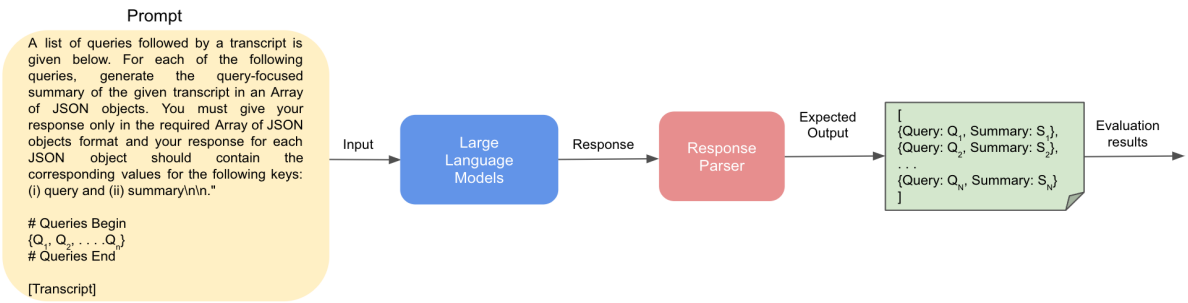

Figure 1: An overview of our Methodology to Evaluate Multi-Query Instructions in the Query Focused Summarization Task.

Model	Correctness
GPT-4	100%
Mixtral-8x7B	60%
PaLM-2	20%
LLaMA-2-13B-Fine-Tuned	20%

Table 1: LLMs performance in generating responses in the required format. Other LLMs have 0% accuracy.

To evaluate the performance of LLMs, for the given sample in the multi-query instruction format, at first the response is generated by the respective LLM. Then we build a parser to extract the summary for each query given in the multi-query input. Finally, we compute the ROUGE score across the whole dataset. An overview of our proposed multi-query instruction evaluation is shown in Figure 1.

4 Experiments

In this section, we first present the LLMs used for evaluation, then followed by the performance metrics achieved by the models.

4.1 Models

Below, we describe the models that we study in this work. For all models, we use their default parameters. We set the *maximum output tokens* limit to 1000 to allow enough token count for the multi-query responses, and a maximum of 3000 input tokens for the open-source models to meet their input size limits. We implement the open-source LLMs using HuggingFace (Wolf et al., 2020).

GPT-4: It is the most powerful LLM released by OpenAI which also has multi-modal capability (OpenAI, 2023). In several benchmarks[5], GPT-4 is found to be the best-performing LLM. We use the *gpt-4-0125-preview* model in this work.

PaLM-2: PaLM-2 is an LLM (Google, 2023) developed by Google that leverages the mixture of objectives technique (Google, 2023) and significantly outperforms the original PaLM (Chowdhery et al., 2022) model. We use the *text-bison-32k@002* model in *Google's VertexAI*[6]

LLaMA-2: LLaMA-2 (Touvron et al., 2023) is an open-source LLM developed by Meta. We use its respective Chat versions of various sizes: 7B, 13B, and 70B.

Mistral-7B-Instruct: The Mistral 7B (Jiang et al., 2023) is an LLM proposed by MistralAI. It leverages grouped-query attention along with sliding window attention to effectively handle sequences of arbitrary length. We use its instruction fine-tuned Mistral-7B-instruct-v0.2 model.

Mixtral-8x-7B: The Mixtral 8x7B (Jiang et al., 2024) is a Sparse Mixture of Experts language model which has the same architecture as Mistral 7B (Jiang et al., 2023), but with the difference that each layer is composed of 8 feed-forward blocks. This leads to an improved performance.

FLAN-T5-XL: FLAN-T5 (Chung et al., 2022) is an extension of the T5 (Raffel et al., 2020) model. While the architecture of FLAN-T5 is similar to the original T5 model, it leverages instruction fine-tuning instead of traditional fine-tuning.

4.2 Results and Discussions

For performance evaluation, we first investigate whether LLMs could properly generate the output in the required JSON format. Further, we report the results based on ROUGE-1, 2, L scores (Lin, 2004) and compare the performance in both multi-query and single-query settings.

4.2.1 Evaluating the Response Format

We surprisingly found that most LLMs could not generate the response (see Appendix A for sample responses) in the expected JSON format or in a consistent format, even if incorrect such that it can be parsed easily without accounting for numerous edge cases. From Table 3, we find that only GPT-4 could generate the output in the required format for

[5] https://lmsys.org/blog/2023-06-22-leaderboard/

[6] https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/text

Models	Multi-Query			Single-Query		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
GPT-4 (Zero-Shot)	53.5	26.3	32.9	32.8	11.5	15.8
PaLM-2-text-bison@002 (Zero-Shot)	50.8	24.2	30.5	31.7	11.6	15.3
LLaMA-2-70B (Zero-Shot)	48.9	26.2	32.3	31.2	12.1	15.7
Mixtral-8x7B-Instruct-v0.1 (Zero-Shot)	44.6	22.6	28.1	39.7	13.1	18.6
LLaMA-2-7B (Fine-Tuned)	48.3	24.9	31.5	50.0	16.6	22.4
LLaMA-2-13B (Fine-Tuned)	50.2	26.1	32.8	50.6	17.2	23.7
Mistral-7B-Instruct-v0.2 (Fine-Tuned)	40.2	19.4	25.0	46.5	15.3	21.2
FLAN-T5-XL (Fine-Tuned)	45.0	22.6	28.1	48.0	17.1	23.5

Table 2: Performance of LLMs on the QMSUM dataset using combined reference summaries per transcript for evaluation.

all test samples, followed by Mixtral-8x7B (60%), and fine-tuned LLaMA-2-13B and PaLM-2 (20%).

4.2.2 Performance Evaluation

Intuitively, we would only be interested in evaluating the summary responses and will thus need to separate the queries from summaries in the output. However, it is challenging to do so since most LLMs could not produce the result in the required format. Therefore, we propose the following evaluation procedure based on ROUGE scores:

- In the multi-query scenario, for each transcript, we compute the ROUGE score between the reference, which combines summaries for each query (with the queries removed), and the entire output returned by the model.

- To ensure a fair comparison between the models in multi-query and single-query settings, we similarly combine all output summaries for a given transcript in the single-query scenario and compute ROUGE scores against the combined reference summaries.

We show our results for both cases in Table 2. Below, we summarize our observations:

(i) We find that GPT-4 is the best on the multi-query dataset, whereas the fine-tuned LLaMA-2-13B model is the best in the single-query setting. We also find that zero-shot larger LLMs perform on par or better than fine-tuned smaller LLMs in multi-query settings.

(ii) We also find that most zero-shot LLMs perform poorer in single-query scenarios than in multi-query scenarios. Our investigation shows that LLMs used in the zero-shot way tend to generate longer summaries than gold summaries, resulting in lower ROUGE scores. However, in multi-query settings, due to the maximum output token limit, the average length of the summary is closer to the reference, which is reflected in better ROUGE scores. For instance, the average gold summary length is 64.7 words, and GPT-4 generates 72.4 words per summary on average in multi-query settings and 191.7 words in single-query settings.

Type	ROUGE-1	ROUGE-2	ROUGE-L
Multi-Query	30.9	7.3	19.2
Single-Query	24.1	6.0	15.0

Table 3: Case study results after parsing GPT-4 generated responses *without* combining reference summaries.

(iii) Interestingly, we find that the performance of each fine-tuned model in single-query settings is much better than zero-shot LLMs. This could be due to the fact that fine-tuning helps LLMs to better understand the required summary length, resulting in an improved ROUGE score.

4.2.3 Case Study

Since GPT-4 could generate the output in the required format, we parse the GPT-4 generated responses and construct distinct query-summary pairs from the multi-query response and compute pairwise ROUGE scores against the references. We show our results in Table 3 and observe that, similar to our findings in the *combined reference summaries* setting, it performs poorer in the single-query scenario. However, since the length of the reference and the model-generated summaries are smaller in this setting, the overall scores are also lower than in the combined setting in both multi-query and single-query scenarios.

5 Conclusion

In this paper, our experimental findings involving various LLMs led to several key insights on building an efficient meeting summarization system for real-world usage. Although we find that most LLMs failed to generate the output in the required format, relatively high ROUGE scores in the multi-query settings hint at the LLMs' ability to handle multi-query prompts. While we show that opting for multi-query prompting is a viable way to optimize inference costs for LLM-based systems, future work should investigate ways to make LLMs generate better formatted responses.

Limitations

One of the limitations of this work is that we couldn't ensure completely fair comparison between the models in the multi-query setting as the models could not produce the outputs in the required format, and future work should focus on improving instruction following for various LLMs in terms of output format. For instance, the effects of the size of the datasets used for fine-tuning LLMs were left out of the scope of this work and should be considered further. It is expected that larger training data used in fine-tuning LLMs will aid in generating the responses in the required format.

Further, more extensive prompt engineering could be beneficial. While we tried several prompts which still did not help most LLMs to generate properly formatted output, our experiments were limited. Leveraging techniques such as Chain-of-Thought (CoT) (Wei et al., 2022b), few-shot learning, etc. may help improve the response generation quality of the LLMs.

Finally, ROUGE scores, although standard across summarization tasks, don't always fairly reflect the quality of generated summaries and do not ensure the quality criteria such as factual correctness or informativeness are met. Applying extensive human evaluation to the LLM-generated responses will provide better insight on the performance of the LLMs and their ability to generate meaningful high-quality summaries.

Ethics Statement

We maintained the licensing requirements accordingly while using different tools from the providers (e.g., OpenAI, Google, Meta, Mistral, HuggingFace). Only a publicly available academic dataset has been used which is available for both research and commercial purposes.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Google. 2023. Palm 2 technical report. *Goole AI*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023a. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023b. Building real-world meeting summarization systems using large language models: A practical perspective. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 343–352, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

OpenAI. 2023. Gpt-4 technical report.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2023. A survey on model compression for large language models. *arXiv preprint arXiv:2308.07633*.

A Appendix

In this section, we show some example responses of different LLMs for a sample prompt.

- See Table 4 for GPT-4.
- See Table 5 for PaLM-2.
- See Table 6 for LLaMA-2-70B.
- See Table 7 for Mixtral-8x7B.
- See Table 8 for Fine-tuned LLaMA-2-13B.
- See Table 9 for Fine-tuned LLaMA-2-7B.
- See Table 10 for Fine-tuned Mistral-7B.
- See Table 11 for FLAN-T5-XL.

Prompt	GPT-4 Response	Comments
A list of queries followed by a transcript is given below. For each of the following queries, generate the query-focused summary of the given transcript in an Array of JSON objects. You must give your response only in the required Array of JSON objects format and your response for each JSON object should contain the corresponding values for the following keys: (i) query and (ii) summary. #Queries Begin 1. Summarize the meeting 2. What's the thesis about? 3. What was said about EML talk? 4. What was said about the talk and neurons? #Queries End #Transcript Begin Grad A: Ah , so comfortable . Grad F: Smooth . Grad A: Mm - hmm . Good . I know that he 's going to like , Taiwan and other places to eat . So . Grad D: On ? Am I on ? Grad A: Yep . Yep . Grad D: I think I 'm on ? Grad B: Yeah . Grad D: Good . Good . Grad A: Bye . Grad B: Actually disfmarker Grad F: I just had one of the most frustrating meetings of my career . Grad A: It 's definitely not the most frustrating meeting I 've ever had . Grad D: You a You 're disfmarker you remember you 're being recorded at this point . Grad A: Oh , yeah , so , w we didn't yet specify with whom . . . . Grad C: We c we can show people pictures of objects and then have then ask the system about the objects and engage in conversation on the history and the art and the architecture and so forth . Professor E: Mm - hmm . OK . So why don't we plan to give you feedback electronically . Wish you a good trip . All success . Grad D: For some reason when you said " feedback electronically " I thought of that disfmarker you ever see the Simpsons where they 're disfmarker like the family 's got the buzzers and they buzz each other when they don't like what the other one is saying ? Grad A: Yeah . That 's the disfmarker first one , I think . The very very first one . Grad D: It was a very early one . I don't know if it 's the first one . Grad A: Mmm . Mmm . #Transcript End	[{ "query": "Summarize the meeting", "summary": "The meeting involved discussions on various topics, including the frustration of a previous meeting, the logistics of coordinating discussions, presenting a thesis proposal, contemplating examples of construal and polysemy within a tourism domain, and preparing talks for an upcoming event at EML. It included technical discussions on linguistic concepts like metaphor, metonymy, and the application of these concepts in computational linguistics, specifically within a tourism information context. Suggestions were made for improving the thesis proposal with an emphasis on construal's role in language understanding." }, { "query": "What's the thesis about?", "summary": "The thesis discussed in the meeting proposes building a probabilistic relational model based on general domain rules for how things are construed. It aims to instantiate elements in probabilistic relational model classes using ontology, situation, user, and discourse models to make inferences about what is being construed in the tourism domain, particularly focusing on understanding language through construal mechanisms." }, { "query": "What was said about EML talk?", "summary": "During the EML talk, the presenter plans to start with a slide highlighting the most relevant aspects of their stay and then ask the audience to imagine being asked about a well-known discotheque in Heidelberg early in the morning to illustrate how ontological knowledge and the time of day affect language interpretation. The talk will also cover work done on the SmartKom system, data about how people mimic synthesis when speaking to computers, and discussions on the rephrasing problems and preliminary observations on human-computer interaction." }, { "query": "What was said about the talk and neurons?", "summary": "The talk will discuss the embodied and simulation approach to language understanding, as a prelude to this approach, they will discuss the research on monkeys in Italy. This will include discussing mirror neurons and the plan to explore how stories and narratives might activate similar neurons in humans. The aim is to see how language understanding might be linked to motor neurons and embodied experiences." }]	Correct Response

Table 4: GPT-4 generated response.

Prompt	PaLM-2 Response	Comments

A list of queries followed by a transcript is given below. For each of the following queries, generate the query-focused summary of the given transcript in an Array of JSON objects. You must give your response only in the required Array of JSON objects format and your response for each JSON object should contain the corresponding values for the following keys: (i) query and (ii) summary.
#Queries Begin
1. Summarize the meeting 2. What's the thesis about? 3. What was said about EML talk? 4. What was said about the talk and neurons?
#Queries End
#Transcript Begin
Grad A: Ah , so comfortable .
Grad F: Smooth .
Grad A: Mm - hmm . Good . I know that he 's going to like , Taiwan and other places to eat . So .
Grad D: On ? Am I on ?
Grad A: Yep . Yep .
Grad D: I think I 'm on ?
Grad B: Yeah .
Grad D: Good . Good .
Grad A: Bye .
Grad B: Actually disfmarker
Grad F: I just had one of the most frustrating meetings of my career .
Grad A: It 's definitely not the most frustrating meeting I 've ever had .
Grad D: You a You 're disfmarker you remember you 're being recorded at this point .
Grad A: Oh , yeah , so , w we didn't yet specify with whom .
. . . . . .
. . . . . .
. . . . . .
Grad C: We c we can show people pictures of objects and then have then ask the system about the objects and engage in conversation on the history and the art and the architecture and so forth .
Professor E: Mm - hmm . OK . So why don't we plan to give you feedback electronically . Wish you a good trip . All success .
Grad D: For some reason when you said " feedback electronically " I thought of that disfmarker you ever see the Simpsons where they 're disfmarker like the family 's got the buzzers and they buzz each other when they don't like what the other one is saying ?
Grad A: Yeah . That 's the disfmarker first one , I think . The very very first one .
Grad D: It was a very early one . I don't know if it 's the first one .
Grad A: Mmm . Mmm .
#Transcript End
	# Array of JSON objects [{
	"query": "Summarize the meeting", "summary": "The meeting was about the progress of Robert's thesis proposal. They discussed the construction parser, tablet-based parser, and generation outputter. Fey is currently working on the synthesis stuff. The EML talk, FMRI stuff, and NTL story were also brought up." },
	{
	"query": "WhatŠ2019s the thesis about?", "summary": "Robert's thesis is about building a probabilistic relational model that has some general, domain-general rules of how things are construed. The idea is to use ontology, situation, user, and discourse model to instantiate elements in the classes of the probabilistic relational model to do some inferences in terms of what is being construed as what in the tourism domain." },
	{
	"query": "What was said about EML talk?", "summary": "They are going to present a talk at EML about what they have done in the project so far. They will start with the slide that shows the most relevant aspects of their stay in Heidelberg. Then, they will ask the audience to imagine that they are standing somewhere in Heidelberg and someone asks them where the Cave Forty-Five is. They will use this example to motivate their work on construal and how it can be used to make inferences about what people are asking for." },
	{
	"query": "What was said about the talk and neurons?", "summary": "Professor E mentioned that the plan is to see to what extent they get the same phenomena with stories about this, so that they can tell if the monkeys in Italy will have the same reaction to stories about kicking and neurons." }
	]
		Additional text added at the beginning.

Table 5: PaLM-2 generated response.

Prompt	LLaMA-2-70B Response	Comments
A list of queries followed by a transcript is given below. For each of the following queries, generate the query-focused summary of the given transcript in an Array of JSON objects. You must give your response only in the required Array of JSON objects format and your response for each JSON object should contain the corresponding values for the following keys: (i) query and (ii) summary. #Queries Begin 1. Summarize the meeting 2. What's the thesis about? 3. What was said about EML talk? 4. What was said about the talk and neurons? #Queries End #Transcript Begin Grad A: Ah , so comfortable . Grad F: Smooth . Grad A: Mm - hmm . Good . I know that he 's going to like , Taiwan and other places to eat . So . Grad D: On ? Am I on ? Grad A: Yep . Yep . Grad D: I think I 'm on ? Grad B: Yeah . Grad D: Good . Good . Grad A: Bye . Grad B: Actually disfmarker Grad F: I just had one of the most frustrating meetings of my career . Grad A: It 's definitely not the most frustrating meeting I 've ever had . Grad D: You a You 're disfmarker you remember you 're being recorded at this point . Grad A: Oh , yeah , so , w we didn't yet specify with whom . . . . Grad C: We c we can show people pictures of objects and then have then ask the system about the objects and engage in conversation on the history and the art and the architecture and so forth . Professor E: Mm - hmm . OK . So why don't we plan to give you feedback electronically . Wish you a good trip . All success . Grad D: For some reason when you said " feedback electronically " I thought of that disfmarker you ever see the Simpsons where they 're disfmarker like the family 's got the buzzers and they buzz each other when they don't like what the other one is saying ? Grad A: Yeah . That 's the disfmarker first one , I think . The very very first one . Grad D: It was a very early one . I don't know if it 's the first one . Grad A: Mmm . Mmm . #Transcript End	Query-focused summary of the given transcript: Query 1: Summarize the meeting Summary: The meeting discussed the thesis proposal, specifically Robert's thesis proposal, and the talk about EML. The conversation also touched on the topic of ontological knowledge and how it can be used to answer questions about the location of a discotheque. Query 2: What's the thesis about? Summary: Robert's thesis proposal is about constructing a probabilistic relational model that has general, domain-general rules for how things are construed. The model will use ontology, situation, user, and discourse model to instantiate elements in the classes of the probabilistic relational model. Query 3: What was said about EML talk? Summary: Grad C mentioned that they are going to present a little talk at EML about what they have done here, and they will start out with a slide showing the most relevant aspects of their stay here. They also mentioned that they will ask the audience to imagine that they are standing somewhere in Heidelberg and someone asks them in the morning where a certain discotheque is. Query 4: What was said about the talk and neurons? Summary: The conversation did not touch on the topic of neurons. Overall, the meeting discussed Robert's thesis proposal, the talk at EML, and ontological knowledge.	Wrong format.

Table 6: LLaMA-2-70B generated response.

Prompt	Mixtral-8x7B Response	Comments

A list of queries followed by a transcript is given below. For each of the following queries, generate the query-focused summary of the given transcript in an Array of JSON objects. You must give your response only in the required Array of JSON objects format and your response for each JSON object should contain the corresponding values for the following keys: (i) query and (ii) summary.
#Queries Begin
1. Summarize the meeting 2. What's the thesis about? 3. What was said about EML talk? 4. What was said about the talk and neurons?
#Queries End
#Transcript Begin
Grad A: Ah , so comfortable .
Grad F: Smooth .
Grad A: Mm - hmm . Good . I know that he 's going to like , Taiwan and other places to eat . So .
Grad D: On ? Am I on ?
Grad A: Yep . Yep .
Grad D: I think I 'm on ?
Grad B: Yeah .
Grad D: Good . Good .
Grad A: Bye .
Grad B: Actually disfmarker
Grad F: I just had one of the most frustrating meetings of my career .
Grad A: It 's definitely not the most frustrating meeting I 've ever had .
Grad D: You a You 're disfmarker you remember you 're being recorded at this point .
Grad A: Oh , yeah , so , w we didn't yet specify with whom .
. . . . . .
. . . . . .
. . . . . .
Grad C: We c we can show people pictures of objects and then have then ask the system about the objects and engage in conversation on the history and the art and the architecture and so forth .
Professor E: Mm - hmm . OK . So why don't we plan to give you feedback electronically . Wish you a good trip . All success .
Grad D: For some reason when you said " feedback electronically " I thought of that disfmarker you ever see the Simpsons where they 're disfmarker like the family 's got the buzzers and they buzz each other when they don't like what the other one is saying ?
Grad A: Yeah . That 's the disfmarker first one , I think . The very very first one .
Grad D: It was a very early one . I don't know if it 's the first one .
Grad A: Mmm . Mmm .
#Transcript End
	[{
	"query": "Summarize the whole meeting",
	"summary": "The group discussed building influence links, starting with path ones, and laying out the influence structure of what influences what. They also talked about belief nodes needed for computation and decided to sort out the basic design of what influences what before deciding exactly how to compute it. Go-there decision, user budget, and probability computation were some of the main topics discussed." },
	{
	"query": "Summarize the discussion about value for user and situation models", "summary": "The group discussed the user and situation models in the context of belief nodes and their influence on decisions. They talked about writing down the values these nodes can take and how to update them externally. The concept of propagating beliefs every time a node is updated or only when needed was also brought up." },
		Missing ending brackets.

Table 7: Mixtral-8x7B generated Response.

Prompt	LLaMA-2-13B Response	Comments
A list of queries followed by a transcript is given below. For each of the following queries, generate the query-focused summary of the given transcript in an Array of JSON objects. You must give your response only in the required Array of JSON objects format and your response for each JSON object should contain the corresponding values for the following keys: (i) query and (ii) summary. #Queries Begin 1. Summarize the meeting 2. What's the thesis about? 3. What was said about EML talk? 4. What was said about the talk and neurons? #Queries End #Transcript Begin Grad A: Ah , so comfortable . Grad F: Smooth . Grad A: Mm - hmm . Good . I know that he 's going to like , Taiwan and other places to eat . So . Grad D: On ? Am I on ? Grad A: Yep . Yep . Grad D: I think I 'm on ? Grad B: Yeah . Grad D: Good . Good . Grad A: Bye . Grad B: Actually disfmarker Grad F: I just had one of the most frustrating meetings of my career . Grad A: It 's definitely not the most frustrating meeting I 've ever had . Grad D: You a You 're disfmarker you remember you 're being recorded at this point . Grad A: Oh , yeah , so , w we didn't yet specify with whom . . . . Grad C: We c we can show people pictures of objects and then have then ask the system about the objects and engage in conversation on the history and the art and the architecture and so forth . Professor E: Mm - hmm . OK . So why don't we plan to give you feedback electronically . Wish you a good trip . All success . Grad D: For some reason when you said " feedback electronically " I thought of that disfmarker you ever see the Simpsons where they 're disfmarker like the family 's got the buzzers and they buzz each other when they don't like what the other one is saying ? Grad A: Yeah . That 's the disfmarker first one , I think . The very very first one . Grad D: It was a very early one . I don't know if it 's the first one . Grad A: Mmm . Mmm . #Transcript End	"query": "Summarize the meeting", "summary": "The group discussed Robert's thesis proposal, which focused on the construction of a probabilistic relational model that could infer the construal of a given sentence. The group also talked about the ontology and situation models, which would be used to instantiate elements in the classes of the probabilistic relational model. Next, the group talked about the EML talk, which would be about the work done in Heidelberg. The talk would be about the ontology and the situation models, and the group agreed to work on the slides for the talk. Finally, the group discussed the progress of the SmartKom project. The parser and the generation outputter were done, but the strategies needed to be changed. The group also talked about the synthesis stuff, which was being done by Fey." , "query": "WhatŽ019s the thesis about?", "summary": "Robert's thesis proposal was about the construction of a probabilistic relational model that could infer the construal of a given sentence. The model would be based on ontology, situation, user, and discourse model to do some inferences in terms of what is being construed as what. The thesis would be about the probabilistic relational model, the ontology and the situation models, and the instantiation of the elements in the classes of the probabilistic relational model." , "query": "What was said about EML talk?", "summary": "The EML talk would be about the work done in Heidelberg. The talk would be about the ontology and the situation models, and the group agreed to work on the slides for the talk. The group agreed that the talk would be about the most relevant aspects of the work done in Heidelberg, and the group agreed to start with the slide about the ontology and the situation models." , "query": "What was said about the talk and neurons?", "summary": "The group agreed that the talk would be about the ontology and the situation models, and the group agreed to start with the slide about the ontology and the situation models. The group agreed that the talk would be about the most relevant aspects of the work done in Heidelberg, and the group agreed to work on the slides for the talk."]	Missing required brackets

Table 8: LLaMA-2-13B generated response.

Prompt	LLaMA-2-7B Response	Comments
A list of queries followed by a transcript is given below. For each of the following queries, generate the query-focused summary of the given transcript in an Array of JSON objects. You must give your response only in the required Array of JSON objects format and your response for each JSON object should contain the corresponding values for the following keys: (i) query and (ii) summary. #Queries Begin 1. Summarize the meeting 2. What's the thesis about? 3. What was said about EML talk? 4. What was said about the talk and neurons? #Queries End #Transcript Begin Grad A: Ah , so comfortable . Grad F: Smooth . Grad A: Mm - hmm . Good . I know that he 's going to like , Taiwan and other places to eat . So . Grad D: On ? Am I on ? Grad A: Yep . Yep . Grad D: I think I 'm on ? Grad B: Yeah . Grad D: Good . Good . Grad A: Bye . Grad B: Actually disfmarker Grad F: I just had one of the most frustrating meetings of my career . Grad A: It 's definitely not the most frustrating meeting I 've ever had . Grad D: You a You 're disfmarker you remember you 're being recorded at this point . Grad A: Oh , yeah , so , w we didn't yet specify with whom . . . . Grad C: We c we can show people pictures of objects and then have then ask the system about the objects and engage in conversation on the history and the art and the architecture and so forth . Professor E: Mm - hmm . OK . So why don't we plan to give you feedback electronically . Wish you a good trip . All success . Grad D: For some reason when you said " feedback electronically " I thought of that disfmarker you ever see the Simpsons where they 're disfmarker like the family 's got the buzzers and they buzz each other when they don't like what the other one is saying ? Grad A: Yeah . That 's the disfmarker first one , I think . The very very first one . Grad D: It was a very early one . I don't know if it 's the first one . Grad A: Mmm . Mmm . #Transcript End	"query": "Summarize the meeting", "summary": "The group discussed Robert's thesis proposal, which focused on the construction of a probabilistic relational model that could infer the construal of a given sentence. The group also talked about the ontology and situation models, which would be used to instantiate elements in the classes of the probabilistic relational model. Next, the group talked about the EML talk, which would be about the work done in Heidelberg. The talk would be about the ontology and the situation models, and the group agreed to work on the slides for the talk. Finally, the group discussed the progress of the SmartKom project. The parser and the generation outputter were done, but the strategies needed to be changed. The group also talked about the synthesis stuff, which was being done by Fey." , "query": "WhatŽ019s the thesis about?", "summary": "Robert's thesis proposal was about the construction of a probabilistic relational model that could infer the construal of a given sentence. The model would be based on ontology, situation, user, and discourse model to do some inferences in terms of what is being construed as what. The thesis would be about the probabilistic relational model, the ontology and the situation models, and the instantiation of the elements in the classes of the probabilistic relational model." , "query": "What was said about EML talk?", "summary": "The EML talk would be about the work done in Heidelberg. The talk would be about the ontology and the situation models, and the group agreed to work on the slides for the talk. The group agreed that the talk would be about the most relevant aspects of the work done in Heidelberg, and the group agreed to start with the slide about the ontology and the situation models." , "query": "What was said about the talk and neurons?", "summary": "The group agreed that the talk would be about the ontology and the situation models, and the group agreed to start with the slide about the ontology and the situation models. The group agreed that the talk would be about the most relevant aspects of the work done in Heidelberg, and the group agreed to work on the slides for the talk."	Missing brackets.

Table 9: LLaMA-2-7B generated response

Prompt	Mistral-7B Response	Comments
A list of queries followed by a transcript is given below. For each of the following queries, generate the query-focused summary of the given transcript in an Array of JSON objects. You must give your response only in the required Array of JSON objects format and your response for each JSON object should contain the corresponding values for the following keys: (i) query and (ii) summary. #Queries Begin 1. Summarize the meeting 2. What's the thesis about? 3. What was said about EML talk? 4. What was said about the talk and neurons? #Queries End #Transcript Begin Grad A: Ah , so comfortable . Grad F: Smooth . Grad A: Mm - hmm . Good . I know that he 's going to like , Taiwan and other places to eat . So . Grad D: On ? Am I on ? Grad A: Yep . Yep . Grad D: I think I 'm on ? Grad B: Yeah . Grad D: Good . Good . Grad A: Bye . Grad B: Actually disfmarker Grad F: I just had one of the most frustrating meetings of my career . Grad A: It 's definitely not the most frustrating meeting I 've ever had . Grad D: You a You 're disfmarker you remember you 're being recorded at this point . Grad A: Oh , yeah , so , w we didn't yet specify with whom . . . . Grad C: We c we can show people pictures of objects and then have then ask the system about the objects and engage in conversation on the history and the art and the architecture and so forth . Professor E: Mm - hmm . OK . So why don't we plan to give you feedback electronically . Wish you a good trip . All success . Grad D: For some reason when you said " feedback electronically " I thought of that disfmarker you ever see the Simpsons where they 're disfmarker like the family 's got the buzzers and they buzz each other when they don't like what the other one is saying ? Grad A: Yeah . That 's the disfmarker first one , I think . The very very first one . Grad D: It was a very early one . I don't know if it 's the first one . Grad A: Mmm . Mmm . #Transcript End	{ "query": "Summarize the meeting", "summary": "The meeting began with a brief discussion about the logistics of the upcoming EML conference. The group then moved on to discussing the details of Robert's thesis proposal. The proposal outlines a probabilistic relational model for representing and reasoning about the construal of tourism domain concepts. The group then discussed the details of the model and the various components that would need to be built to complete the work. Finally, the group discussed the logistics of the upcoming SmartKom evaluation and the various tasks that would need to be completed in order to prepare for it." }, { "query": "WhatŽ019s the thesis about?", "summary": "The thesis proposes a probabilistic relational model for representing and reasoning about the construal of tourism domain concepts. The model includes a set of general, domain-independent rules for how things are construed, as well as a set of domain-specific ontology, situation, user, and discourse models that can be used to instantiate elements in the classes of the probabilistic relational model. The model is intended to be used to do some inferences in terms of what is being construed as what in the tourism domain." }, { "query": "What was said about EML talk?", "summary": "The group discussed the details of the upcoming EML conference, including the logistics of travel and accommodation. The group also discussed the content of the talks that would be given at the conference, including the talk that Robert would be giving about the work that had been done in the group. The group agreed that the talk would be a good opportunity to present the work to a wider audience and to get some feedback on it." }, { "query": "What was said about the talk and neurons?", "summary": "The group discussed the details of the talk that Robert would be giving at EML. The talk would be about the work that had been done in the group and would include a brief overview of the probabilistic relational model for representing and reasoning about the construal of tourism domain concepts. The group also discussed the possibility of incorporating some information about the neural basis of language processing into the talk. The group agreed that this would be an interesting addition to the talk, but that it would require some additional research to be done beforehand." }]	Missing starting brackets.

Table 10: Mistral-7B generated response.

Prompt	FLAN-T5-XL Response	Comments
A list of queries followed by a transcript is given below. For each of the following queries, generate the query-focused summary of the given transcript in an Array of JSON objects. You must give your response only in the required Array of JSON objects format and your response for each JSON object should contain the corresponding values for the following keys: (i) query and (ii) summary. #Queries Begin 1. Summarize the meeting 2. What's the thesis about? 3. What was said about EML talk? 4. What was said about the talk and neurons? #Queries End #Transcript Begin Grad A: Ah , so comfortable . Grad F: Smooth . Grad A: Mm - hmm . Good . I know that he 's going to like , Taiwan and other places to eat . So . Grad D: On ? Am I on ? Grad A: Yep . Yep . Grad D: I think I 'm on ? Grad B: Yeah . Grad D: Good . Good . Grad A: Bye . Grad B: Actually disfmarker Grad F: I just had one of the most frustrating meetings of my career . Grad A: It 's definitely not the most frustrating meeting I 've ever had . Grad D: You a You 're disfmarker you remember you 're being recorded at this point . Grad A: Oh , yeah , so , w we didn't yet specify with whom . . . . Grad C: We c we can show people pictures of objects and then have then ask the system about the objects and engage in conversation on the history and the art and the architecture and so forth . Professor E: Mm - hmm . OK . So why don't we plan to give you feedback electronically . Wish you a good trip . All success . Grad D: For some reason when you said " feedback electronically " I thought of that disfmarker you ever see the Simpsons where they 're disfmarker like the family 's got the buzzers and they buzz each other when they don't like what the other one is saying ? Grad A: Yeah . That 's the disfmarker first one , I think . The very very first one . Grad D: It was a very early one . I don't know if it 's the first one . Grad A: Mmm . Mmm . #Transcript End	["query": "Summarize the meeting", "summary": "The meeting was about Robert's thesis proposal. The thesis was about a probabilistic relational model that could be used to infer what people construe as interesting in the tourism domain. The thesis was not yet finished, but the group was able to give feedback on it. The group was also preparing for a talk at EML about what they had done here. The talk was about the use of a neural network to model the interaction between neurons and the underlying context. The group was also preparing for a trip to Taiwan.", "query": "What's the thesis about?", "summary": "The thesis was about a probabilistic relational model that could be used to infer what people construe as interesting in the tourism domain. The thesis was not yet finished, but the group was able to give feedback on it. The group was also preparing for a trip to Taiwan.", "query": "What was said about EML talk?", "summary": "Grad C was going to give a talk at EML about what they had done here. He was also going to meet with Wolfgang Wahlster to sell this idea. The group was also preparing for a trip to Taiwan.", "query": "What was said about the talk and neurons?", "summary": "Grad C was going to give a talk about what they had done here. He was also going to meet with Wolfgang Wahlster to sell this idea.", "query": "What was said about the talk and neurons?", "summary": "Grad C was going to give a talk about what they had done here. He was also going to meet with Wolfgang Wahlster to sell this idea.", "query": "What was said about the talk and neurons?", "summary": "Grad C was going to give a talk about what they had done here. He was also going to meet with Wolfgang Wahlster to sell this idea."]	Wrong format.

Table 11: FLAN-T5-XL generated Response.