ELITR-Bench: A Meeting Assistant Benchmark for Long-Context Language Models

Thibaut Thonet [1] **Jos Rozen** [1] **Laurent Besacier** [1]

Abstract

Research on Large Language Models (LLMs) has recently witnessed an increasing interest in extending models' context size to better capture dependencies within long documents. While benchmarks have been proposed to assess long-range abilities, existing efforts primarily considered generic tasks that are not necessarily aligned with real-world applications. In contrast, our work proposes a new benchmark for long-context LLMs focused on a practical meeting assistant scenario. In this scenario, the long contexts consist of transcripts obtained by automatic speech recognition, presenting unique challenges for LLMs due to the inherent noisiness and oral nature of such data. Our benchmark, named ELITR-Bench, augments the existing ELITR corpus' transcripts with 271 manually crafted questions and their ground-truth answers. Our experiments with recent long-context LLMs on ELITR-Bench highlight a gap between open-source and proprietary models, especially when questions are asked sequentially within a conversation.

1. Introduction

The context window of Large Language Models (LLMs) has recently undergone a significant expansion, scaling from a few thousand tokens to tens or even hundreds of thousands (Chen et al., 2023; Xiong et al., 2023; Liu et al., 2023a; Chen et al., 2024; Bai et al., 2024). As a consequence, benchmarks have emerged to assess LLMs' long-range abilities, tackling the specific challenges of Question Answering (QA) on long documents (An et al., 2023; Li et al., 2023a; Bai et al., 2023; Li et al., 2023b; Maharana et al., 2024; Zhang et al., 2024). However, while previous datasets focusing on long-context models offer longitudinal evaluations across different tasks, they often provide only superficial analyses of each task. The covered tasks are also often generic – e.g., questions on Wikipedia (Li et al., 2023b) – and thus not particularly suitable for realistic, focused applications.[1]

In contrast, our work advocates for a situated evaluation of long-context LLM performance within specific, real-world scenarios. As a practical illustration, consider a meeting assistant that allows users to inquire about meetings they did not attend. Given that hour-long meeting transcripts must fit within the agent's context window, proficient handling of long contexts is a prerequisite. This paper then introduces the first benchmark – to the best of our knowledge – for evaluating long-context LLMs on a realistic meeting assistant task. Our benchmark, named ELITR-Bench,[2] is built upon the meeting transcripts of the past ELITR project (Nedoluzhko et al., 2022). These transcripts have been obtained by Automatic Speech Recognition (ASR) with minimal human corrections – yielding long, noisy documents of oral nature that present unique challenges for LLMs. Our extensive experiments on ELITR-Bench with 9 recent long-context LLMs showed a gap between proprietary and open-source models that is emphasized when questions are asked sequentially within a conversation rather in a QA mode. The remainder of the paper is structured as follows. We introduce ELITR-Bench in Section 2. We then describe our experimental setup and results in Sections 3 and 4, respectively. Finally, Section 5 concludes the paper.[3]

2. ELITR-Bench

We build our benchmark on top of the ELITR Minuting Corpus (Nedoluzhko et al., 2022).[4] This corpus contains transcripts of meetings conducted in both Czech and English,

[1]NAVER LABS Europe, France. Correspondence to: Thibaut Thonet <thibaut.thonet@naverlabs.com>.

Proceedings of the 1st Workshop on Long-Context Foundation Models, Vienna, Austria. 2024.

[1]Aside from this very recent work (Lin et al., 2024) (WildBench), which offers an automated evaluation framework for assessing LLMs with complex real-world user queries.

[2]We release the data for our benchmark at https://github.com/utter-project/ELITR-Bench.

[3]We addtionally provide: a review of related literature (Appendix A); experimental setup details (Appendix B); additional experimental results (Appendix C), and an in-depth assessment of our LLM-based evaluation methodology (Appendix D).

[4]Accessible at: https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-4692

along with manually crafted summaries referred to as 'minutes'. The meeting durations range from 10 minutes to over 2 hours, with the majority lasting around one hour. Although transcripts have been corrected from ASR outputs, they still contain noise and reflect various oral language phenomena such as interjections. Each transcript is de-identified[5] and accompanied by one or multiple corresponding minutes files. However, in the benchmark described here, we only use the verbatim transcripts and exclude the minutes. In the current version of ELITR-Bench, our focus is on English meetings. Specifically, we utilized the official *dev* and *test2* sets, consisting of 10 and 8 meetings respectively, both sourced from ELITR-English. These meetings focus on discussions related to the computer science domain, with a particular emphasis on Natural Language Processing (NLP) topics. For every meeting within this corpus, we meticulously formulated a series of questions that can be directly addressed using the corresponding transcript, and provided their corresponding ground-truth answers. We present in Appendix F (Table 13) a snippet of a meeting transcript, and showcase examples of Q&As introduced in ELITR-Bench.

Question type and answer position. The questions we defined span various types, including: **Who** questions, **What** questions (that also cover *Why* questions), **When** questions, and **How many** questions. Additionally, we annotated the position of the answer within the meeting transcript, categorizing it as either in the **Beginning** (first third), **Middle** (second third), **End** (final third), or spanning **Several** passages throughout the transcript. This annotation was conducted to verify the findings of Liu et al. (2023b), which suggest that LLMs may face challenges in processing information located in the middle of long contexts, potentially leading to performance degradation.

QA and Conversation settings. The proposed ELITR-Bench is available in two settings. In **ELITR-Bench-QA**, we designed for each meeting a set of stand-alone questions (along with their answers) that can be addressed solely based on the meeting transcript, without additional context. We also designed a modified **ELITR-Bench-Conv** version where questions are to be asked in sequence, in a pre-defined order within a conversation. In this setting, some of the questions contain pronominal references or ellipses, for which previous conversational context (i.e., previous questions and answers) must be used to answer properly. For example, the question "*What is challenging about testing the demo system at the students firm fair?*" from the QA setting is replaced in the Conv setting with "*What is challenging about this event?*", where the answer to the previous question in the conversation was "*The students firm fair*". Such questions have been obtained by manually re-writing the Conv questions into QA questions by resolving coreferences. The number of QA/Conv differentiating questions is 16 (out of 141) for the dev set and 17 (out of 130) for the test set.

Table 1 provides a summary of the statistics for our benchmark. In the upcoming sections, we will showcase the performance of long-context LLMs on ELITR-Bench, particularly in their ability to handle hour-long meetings – which requires processing extended contexts of more than 10K tokens on average.

3. Experimental setup

Evaluation protocol. The evaluation on ELITR-Bench is conducted as follows. For each meeting, a prompt containing the transcript and detailing the assistant's task is formed. Then, questions are appended to the initial prompt to drive the conversation about the corresponding meeting. We consider two ways to do this: (i) the *single-turn mode*, where only a single question is tackled in the conversation (i.e., the prompt is re-initialized for each new question), or (ii) the *multi-turn mode*, where all the questions related to a meeting are asked successively within a single conversation. Given the stand-alone nature of questions in ELITR-Bench-QA, one can adopt either the single-turn or multi-turn modes for this setting, whereas for ELITR-Bench-Conv it only makes sense to use the multi-turn mode as some questions are inter-dependent. In our evaluation methodology, given a question integrated in the aforementioned prompt, the response generated by an LLM is evaluated automatically using a GPT-4 judge,[6] following the standard practice in LLM evaluation (as discussed in Appendix A). Specifically, we adopted a score rubric-based evaluation methodology (Kim et al., 2024) in which a generated response is evaluated on its proximity to the ground-truth answer, given the associated question and a score rubric that details the quality criteria expected at each score level (ranging from the lowest score of 1 to the perfect score of 10). The prompt used for the evaluation as well as our manually defined score rubric are provided in Appendix G (Figs. 5 and 6, respectively). Although our core experiment results rely on automatic LLM-based evaluation (Section 4), we also confirm the validity of this methodology against human judgement (see Appendix D).

Compared models. In our experiments on ELITR-Bench, we compared responses generated by 9 recent LLMs with long-context capabilities. We included both commercial

[5] Nedoluzhko et al. (2022) ensured the removal or masking of any personally identifiable information (PII), such as names, addresses, or other details from the transcripts. Moreover, they de-identified any project or organization-related information, as its inclusion could indirectly reveal the individuals involved.

[6] Our GPT-4 judge is based on the gpt-4-0613 checkpoint, for its cheaper cost compared to gpt-4-turbo models. Pilot experiments with different GPT-4 judges led to similar evaluation scores.

Split	#Meetings	#Questions	#Questions by question type		#Questions by answer position		#Tokens per meeting: average [min; max]
Dev	10	141	What	59	Begin	45	11,339 [5,152; 17,410]
			Who	51	Middle	29	
			When	21	End	32	
			How many	10	Several	35	
Test	8	130	What	57	Begin	43	12,562 [4,779; 17,615]
			Who	45	Middle	34	
			When	20	End	22	
			How many	8	Several	31	

Table 1. Statistics for the ELITR-Bench dataset: all questions and answers are annotated by question type (*What*, *Who*, *When*, *How many*) and by the position of the answer within the meeting transcript (*Beginning*, *Middle*, *End*, or spanning *Several* sections). The number of tokens per meeting is counted using a LLaMA-2 tokenizer.

models and open-source long-context models based on LLaMA-2 in our benchmarking:

- **GPT-3.5** and **GPT-4** (OpenAI, 2023), which respectively enable a context length of 16K and 128K tokens,
- **LongAlpaca-7B** and **LongAlpaca-13B** both introduced in Chen et al. (2024), with context size 32K,
- **LongChat-7B-v1.5** the LLaMA-2 version of the original LongChat-7B model (Li et al., 2023a), with a context of at most 32K tokens,
- **Vicuna-7B-v1.5** and **Vicuna-13B-v1.5** obtained by fine-tuning LLaMA-2 similarly to the original Vicuna model (Chiang et al., 2023). Their context length is 16K – which we extrapolate to 32K at inference time using RoPE (Su et al., 2024),
- **LongAlign-7B** and **LongAlign-13B** based on the LongAlign recipe (Bai et al., 2024), limited to 64K tokens.

We provide more details on the compared models in Appendix B.1. Additionally, we describe the search conducted to select the best configuration (including inference hyperparameters and prompt formatting) for each model in Appendix B.2.

4. Experimental results

4.1. Main results

Main results of the benchmarking on ELITR-Bench are reported in Table 2. The compared models are evaluated in three settings that combine the ELITR-Bench-QA or ELITR-Bench-Conv question set with the single-turn mode (i.e., one question asked per conversation) or multi-turn mode (i.e. all questions related to one meeting asked in a single conversation).[7] For each of the considered settings, we report results on the dev set, results on the test set, and their mean. Given the extensive cost of GPT4-based evaluation, we performed a single seeded run for the dev set and three seeded runs for the test set. For the latter we report average score over the three runs. In Appendix C, we provide more details about the seeded runs as well as their standard deviations.

Looking at the three settings in Table 2, we observe that GPT-4 clearly dominates over all other approaches with an average score that is always above 8.[8] GPT-3.5 obtained a slightly lower average score – around 7 – that still outperforms open-source LLMs. Among these, differences are smaller with scores close to 6 on the single-turn setting, and ranging from 4 to 6 on the multi-turn settings. Nonetheless, we note that Vicuna-13B-v1.5 is the open-source approach that performed the most favorably overall on the three settings. Interestingly, results in the single-turn and multi-turn modes show large discrepancies for open-source models – even when the question set is exactly the same, for ELITR-Bench-QA. This seems to indicate that open-source long-context LLMs get distracted by the previous questions and answers, which affects their performance. In contrast, GPT-4 is instead able to increase its performance between the single-turn mode and the multi-turn mode. Comparing results of the QA and Conv settings in the multi-turn mode, we found only minimal differences. This can be explained by the small number of questions that differ between QA and Conv (16 for the dev set and 17 for the test set). In Appendix C.3, we analyze the results on this subset of differentiating questions to better understand the impact of the benchmark setting (QA vs Conv) and inference mode (single-turn vs multi-turn).

[7] Single-turn ELITR-Bench-Conv is omitted as some questions in the Conv setting are context-dependent (i.e., rely on previous questions or answers) and thus could not be asked independently.

[8] While one might argue that GPT-4 is unfairly advantaged due to the use of a GPT-4 judge, we show in Appendix D.2 that the dominance of this model is observed for other evaluators as well.

Model	Single-turn			Multi-turn					
	ELITR-Bench-QA			ELITR-Bench-QA			ELITR-Bench-Conv		
	Dev	Test	Mean	Dev	Test	Mean	Dev	Test	Mean
GPT-3.5	7.04	7.44	7.24	-	-	-	-	-	-
GPT-4	**8.21**	**8.39**	**8.30**	**8.53**	**8.42**	**8.47**	**8.53**	**8.36**	**8.45**
LongAlpaca-7B	5.89	5.60	5.75	4.53	4.84	4.68	4.70	4.58	4.64
LongAlpaca-13B	6.17	6.25	6.21	4.76	4.71	4.73	4.74	4.74	4.74
LongChat-7B-v1.5	**6.60**	5.78	6.19	**5.85**	4.17	5.01	5.21	4.31	4.76
Vicuna-7B-v1.5	5.42	5.61	5.51	4.68	4.61	4.65	4.67	4.69	4.68
Vicuna-13B-v1.5	5.92	**6.52**	6.22	5.52	**5.67**	**5.60**	**5.42**	**5.78**	**5.60**
LongAlign-7B	6.11	6.46	6.28	5.43	4.47	4.95	5.04	5.06	5.05
LongAlign-13B	6.27	6.33	**6.30**	4.65	5.33	4.99	4.81	4.95	4.88

Table 2. Results on different ELITR-Bench settings. The reported numbers correspond to the average scores from 1 to 10 (higher is better) obtained by a GPT-4 evaluator, on a single seeded run for the dev set and 3 seeded runs for the test set. Boldface numbers correspond to the best performance among proprietary or open-source models. The results for GPT-3.5 are omitted in the multi-turn setting as the context length exceeded the 16K limit of this model.

Model family	Question type				Answer location			
	Who (N=45)	What (N=57)	When (N=20)	How many (N=8)	Begin (N=43)	Middle (N=34)	End (N=22)	Several (N=31)
GPT	8.24	7.62	7.98	8.04	7.85	7.87	8.04	7.97
LLaMA-2	6.50	5.88	6.00	5.29	6.31	5.84	6.00	6.07

Table 3. Results by question type and answer location for the GPT family (2 models) and the LLaMA-2 family (7 models) on the test set of ELITR-Bench-QA in single-turn mode. The number N below a subset indicates the corresponding subset size.

4.2. Impact of question type and answer position

As each question in ELITR-Bench is characterized by its type (*Who*, *What*, *When*, *How many*) and answer location (*Beginning*, *Middle*, *End*, *Several*), we sought to identify whether these impact the models' ability to answer correctly. We show in Table 3 the results restricted to each question type and answer location, obtained on the test set of ELITR-Bench-QA in single-turn mode. Due to space limitations, we aggregate the results by family of models (GPT models or LLaMA-2 models) to look for general trends among comparable models. The detailed, per-model results are available in Appendix C.2 (Table 8). Looking at the question type results, we find that for both model families *Who* questions are the easiest to answer. In contrast, *What* questions were the most challenging for GPT models and the second most challenging for LLaMA-2 models. This is not surprising as *What* questions sometimes require complex answers that go beyond simply listing entities, dates or numbers. Interestingly, LLaMA-2 models struggled the most with *How many* questions. Although the amount of such questions is very limited (8 in the test set) which calls for caution on tentative interpretations, this suggests that LLaMA-2 models are notably less proficient at dealing with quantities and numbers than GPT models. In contrast, the results by answer location in Table 3 do not seem to show any general trend. In particular, we do not notice at first glance any "lost in the middle" effect (Liu et al., 2023b) which posits that information located in the middle section of long contexts is harder to access for LLMs. To further verify this, we conducted a one-tailed Welch's t-test (Welch, 1947) for each model to investigate the hypothesis stating that the model's average score for questions with middle-position answers is lower than that of other questions. We found that this hypothesis is only verified for two models: LongChat-7B-v1.5 (p-value = 0.032) and Vicuna-7B-v1.5 (p-value = 0.046) – the full results are available in Appendix C.2 (Table 9). This suggests that all models may not be affected in the same way by the location of information in the context.

5. Conclusion

This paper introduced ELITR-Bench, a new benchmark for long-context LLMs focused on the meeting assistant task. We augmented the meeting transcripts from existing ELITR corpus with 271 manually crafted questions and their respective ground-truth answers. Our experiments showcase the performance of recent long-context LLMs on ELITR-Bench, highlighting a gap between proprietary OpenAI models and LLaMA-2-based long-context models – in particular when dealing with multi-turn conversations.

Acknowledgments

This paper was partially funded by the European Commission through the UTTER project under grant number 101070631.

Impact statement

This paper presents work whose goal is to advance the fields of Machine Learning and Natural Language Processing by introducing a new benchmark. As such, there are potential societal consequences of our work, but none which we feel must be specifically highlighted here.

With respect to the ethical considerations for this work and in particular for the crowdsourcing study we conducted, the data collection and evaluation process rigorously adhered to the guidelines established by the UTTER EU project. In accordance with EU project policies, we regularly report to an ethics panel, with the most recent Ethical Review meeting held on November 2nd, 2023. Notably, for the human evaluation of LLMs, we chose Prolific, a crowdsourcing platform tailored for academic research. We meticulously followed Prolific's guidelines for human experiments, deviating only in terms of compensation for human labelers. While Prolific sets a minimum compensation of \$6.50 per hour, we offered a significantly higher rate of £9 per hour (equivalent to \$11.5 per hour).

References

An, C., Gong, S., Zhong, M., Zhao, X., Li, M., Zhang, J., Kong, L., and Qiu, X. L-eval: Instituting standardized evaluation for long context language models, 2023.

Apel, R., Braude, T., Kantor, A., and Kolman, E. Meeqa: Natural questions in meeting transcripts. *CoRR*, abs/2305.08502, 2023. doi: 10.48550/ARXIV.2305.08502. URL https://doi.org/10.48550/arXiv.2305.08502.

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., Dong, Y., Tang, J., and Li, J. Longbench: A bilingual, multitask benchmark for long context understanding, 2023.

Bai, Y., Lv, X., Zhang, J., He, Y., Qi, J., Hou, L., Tang, J., Dong, Y., and Li, J. Longalign: A recipe for long context alignment of large language models, 2024.

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., and Testoni, A. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks, 2024. URL https://arxiv.org/abs/2406.18403.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020. URL https://arxiv.org/abs/2004.05150.

Bulatov, A., Kuratov, Y., Kapushev, Y., and Burtsev, M. S. Scaling transformer to 1m tokens and beyond with rmt, 2024.

Chen, S., Wong, S., Chen, L., and Tian, Y. Extending context window of large language models via positional interpolation, 2023.

Chen, Y., Qian, S., Tang, H., Lai, X., Liu, Z., Han, S., and Jia, J. Longlora: Efficient fine-tuning of long-context large language models, 2024.

Chevalier, A., Wettig, A., Ajith, A., and Chen, D. Adapting language models to compress contexts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 3829–3846, 2023. URL https://aclanthology.org/2023.emnlp-main.232.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/, March 2023.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *CoRR*, abs/1904.10509, 2019. URL http://arxiv.org/abs/1904.10509.

Choromanski, K. M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J. Q., Mohiuddin, A., Kaiser, L., Belanger, D. B., Colwell, L. J., and Weller, A. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285.

Dong, Z., Tang, T., Li, J., Zhao, W. X., and Wen, J.-R. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models, 2024.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces, 2023.

He, Z., Huang, C.-Y., Ding, C.-K. C., Rohatgi, S., and Huang, T.-H. If in a crowdsourced data annotation pipeline, a gpt-4. *arXiv preprint arXiv:2402.16795*, 2024.

Kamradt, G. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2024. GitHub.

Katharopoulos, A., Vyas, A., Pappas, N., and Fleuret, F. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. URL https://arxiv.org/pdf/2006.16236.pdf.

Khandve, S. I., Wagh, V., Wani, A., Joshi, I., and Joshi, R. Hierarchical neural network approaches for long document classification. *CoRR*, abs/2201.06774, 2022. URL https://arxiv.org/abs/2201.06774.

Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., Yun, S., Shin, S., Kim, S., Thorne, J., and Seo, M. Prometheus: Inducing evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8euJaTveKw.

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023.

Koo, T. K. and Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

Levy, M., Jacoby, A., and Goldberg, Y. Same task, more tokens: the impact of input length on the reasoning performance of large language models, 2024.

Li, D., Shao, R., Xie, A., Sheng, Y., Zheng, L., Gonzalez, J. E., Stoica, I., Ma, X., and Zhang, H. How long can open-source llms truly promise on context length? https://lmsys.org/blog/2023-06-29-longchat/, June 2023a.

Li, J., Wang, M., Zheng, Z., and Zhang, M. Loogle: Can long-context language models understand long contexts?, 2023b.

Lin, B. Y., Deng, Y., Chandu, K., Brahman, F., Ravichander, A., Pyatkin, V., Dziri, N., Bras, R. L., and Choi, Y. Wildbench: Benchmarking llms with challenging tasks from real users in the wild, 2024. URL https://arxiv.org/abs/2406.04770.

Liu, H., Zaharia, M., and Abbeel, P. Ring attention with blockwise transformers for near-infinite context, 2023a.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts, 2023b.

Liu, Y., Liu, J., Chen, L., Lu, Y., Feng, S., Feng, Z., Sun, Y., Tian, H., Wu, H., and Wang, H. Ernie-sparse: Learning hierarchical efficient transformer through regularized self-attention, 2022.

Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., and Fang, Y. Evaluating very long-term conversational memory of llm agents, 2024.

Martins, A., Farinhas, A., Treviso, M., Niculae, V., Aguiar, P., and Figueiredo, M. Sparse and continuous attention mechanisms. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20989–21001. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f0b76267fbe12b936bd65e203dc675c1-Paper.pdf.

Nedoluzhko, A., Singh, M., Hledíková, M., Ghosal, T., and Bojar, O. ELITR minuting corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 3174–3182, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.340.

OpenAI. GPT-4 Technical Report. pp. 1–100, 2023. URL http://arxiv.org/abs/2303.08774.

Pal, A., Karkhanis, D., Roberts, M., Dooley, S., Sundararajan, A., and Naidu, S. Giraffe: Adventures in expanding context lengths in llms, 2023.

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Derczynski, L., Du, X., Grella, M., Gv, K., He, X., Hou, H., Kazienko, P., Kocon, J., Kong, J., Koptyra, B., Lau, H., Lin, J., Mantri, K. S. I., Mom, F., Saito, A., Song, G., Tang, X., Wind, J., Woźniak, S., Zhang, Z., Zhou, Q., Zhu, J., and Zhu, R.-J. RWKV: Reinventing RNNs for the transformer era. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14077, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.936. URL https://aclanthology.org/2023.findings-emnlp.936.

Peng, B., Quesnelle, J., Fan, H., and Shippole, E. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

Prasad, A., Bui, T., Yoon, S., Deilamsalehy, H., Dernoncourt, F., and Bansal, M. Meetingqa: Extractive question-answering on meeting transcripts. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 15000–15025. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.837. URL https://doi.org/10.18653/v1/2023.acl-long.837.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., and Metzler, D. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

Wang, J., Gangavarapu, T., Yan, J. N., and Rush, A. M. Mambabyte: Token-free selective state space model, 2024.

Wang, S., Li, B. Z., Khabsa, M., Fang, H., and Ma, H. Linformer: Self-attention with linear complexity. *CoRR*, abs/2006.04768, 2020. URL https://arxiv.org/abs/2006.04768.

Welch, B. L. The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947. URL http://www.jstor.org/stable/2332510.

Wu, C., Wu, F., Qi, T., and Huang, Y. Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 848–853, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.107. URL https://aclanthology.org/2021.acl-short.107.

Xiong, W., Liu, J., Molybog, I., Zhang, H., Bhargava, P., Hou, R., Martin, L., Rungta, R., Sankararaman, K. A., Oguz, B., Khabsa, M., Fang, H., Mehdad, Y., Narang, S., Malik, K., Fan, A., Bhosale, S., Edunov, S., Lewis, M., Wang, S., and Ma, H. Effective long-context scaling of foundation models, 2023.

Xu, P., Ping, W., Wu, X., McAfee, L., Zhu, C., Liu, Z., Subramanian, S., Bakhturina, E., Shoeybi, M., and Catanzaro, B. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xw5nxFWMlo.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. Big bird: Transformers for longer sequences. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf.

Zhang, X., Chen, Y., Hu, S., Xu, Z., Chen, J., Hao, M. K., Han, X., Thai, Z. L., Wang, S., Liu, Z., and Sun, M. ∞bench: Extending long context evaluation beyond 100k tokens, 2024.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

A. Related work

Long-context LLMs and techniques. Numerous techniques have emerged to address the challenge of long-context modeling.[9] While an exhaustive survey of these methods is beyond the scope of this paper, they can generally be categorized into three main groups (excluding other distinct approaches such as retrieval-augmented generation (Xu et al., 2024) and context compression (Chevalier et al., 2023)): (a) the development of efficient transformer architectures to address the quadratic attention challenge, including sparse transformers (Child et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020; Martins et al., 2020), linear transformers (Katharopoulos et al., 2020; Wang et al., 2020; Choromanski et al., 2021), and hierarchical transformers (Khandve et al., 2022; Wu et al., 2021; Liu et al., 2022); (b) approaches like recurrent attention networks (Dai et al., 2019; Peng et al., 2023; Bulatov et al., 2024) and state-space models (Gu & Dao, 2023; Wang et al., 2024); (c) length extrapolation or position embedding interpolation, where LLMs are fine-tuned or adapted at inference time to adjust tokens' positions to match the new context length (Chen et al., 2023; Xiong et al., 2023; Peng et al., 2024; Pal et al., 2023; Liu et al., 2023a; Chen et al., 2024; Bai et al., 2024). These techniques also contributed to the context length expansion in proprietary models like GPT-4 (32K-128K), Claude-3 (200k), and Gemini-1.5 (128K-1M).

Long-context benchmarks. Several benchmarks have recently emerged with the growing interest in evaluating techniques that extend the context length of LLMs. Long Range Arena (Tay et al., 2021) was proposed to assess the quality of efficient transformer models in long-context scenarios, covering 1K-16K tokens sequences through different data types and modalities. L-Eval (An et al., 2023) offers a comprehensive evaluation suite with 20 sub-tasks and over 2,000 human-labeled query-response pairs, aggregating pre-existing datasets like NarrativeQA (Kočiský et al., 2018). LongEval (Li et al., 2023a) proposes synthetic tasks of varying difficulty, while LongBench (Bai et al., 2023) and LongBench-Chat (Bai et al., 2024) aggregate several datasets in English and Chinese. Other recent benchmarks appeared, such as: LongAlpaca (Chen et al., 2024), Loogle (Li et al., 2023b), LoCoMo (Maharana et al., 2024), BAMBOO (Dong et al., 2024), FLenQA (Levy et al., 2024), and ∞Bench (Zhang et al., 2024) that proposes an average data length over 100K tokens. Recently, the needle-in-the-haystack test was proposed by (Kamradt, 2024), in which a long-context LLM must retrieve a short text (the needle) from a long document (the haystack). This initial test has since inspired several subsequent works that propose more and more complex tasks. Our contribution, ELITR-Bench, distinguishes itself from existing benchmarks in several ways: (a) it focuses on a real use-case – meeting assistants,[10] (b) it challenges models by requiring them to make inferences from noisy ASR-based documents, and (c) it offers both question answering and conversation versions (see Section 2), enabling the analysis of different prompt modes.

Evaluation with LLMs. Recent works explored the use of LLMs such as GPT-4 as judges to evaluate responses on open-ended questions. Zheng et al. (2023) measured agreement between LLM and human evaluators while introducing two datasets (MT-bench and Chatbot Arena). They showed that LLM judges like GPT-4 can match both controlled and crowdsourced human annotations, achieving over 80% agreement – the same level of agreement between humans. He et al. (2024) evaluated the performance of GPT-4 against 415 crowdsourcing human labelers. Despite employing best quality control practices, the highest labeling accuracy achieved through crowdsourcing was 81.5% whereas GPT-4 obtained 83.6%. As in certain scenarios, employing proprietary LLMs as evaluators can pose challenges due to their closed-source nature, Kim et al. (2024) introduced Prometheus, an open-source LLM fine-tuned for evaluation. Recently, Bavaresco et al. (2024) introduced Judge-Bench, a collection of 20 NLP datasets with human annotations for evaluating LLMs' ability to replicate human judgments. In this work, we compare LLMs-as-a-judge (GPT-4 and Prometheus) with expert and crowdsourcing-based human evaluators to assess responses generated by several long-context models on ELITR-Bench.

B. Experimental setup details

B.1. Compared models and hardware details

We summarize the details of the different long-context LLMs compared in our experiments in Table 4. We provide for each model its context size limit in tokens, its backbone model (i.e., the pre-trained model used for the fine-tuning), and the link

[9] For a comprehensive collection of resources on this subject, we let the reader refer to `https://github.com/Xnhyacinth/Awesome-LLM-Long-Context-Modeling`.

[10] The existing MeeQA (Apel et al., 2023) and MeetingQA (Prasad et al., 2023) datasets have been proposed to study QA on meeting transcripts. However, the questions and answers in these datasets have been directly extracted from the transcripts, which implies that (i) they are more likely to be of poor quality than our manually crafted questions and answers, and (ii) answering in MeeQA and MeetingQA only requires an extraction from the transcript without any inference, making the task much less challenging than in ELITR-Bench.

Model	Context limit	Backbone	Link
GPT-3.5 (turbo-16k-0613)	16K	-	https://platform.openai.com/docs/models/gpt-3-5-turbo
GPT-4 (1106-preview)	128K	-	https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo
LongAlpaca-7B	32K	LLaMA-2-7B	https://huggingface.co/Yukang/LongAlpaca-7B
LongAlpaca-13B	32K	LLaMA-2-13B	https://huggingface.co/Yukang/LongAlpaca-13B
LongChat-7B-v1.5	32K	LLaMA-2-7B	https://huggingface.co/lmsys/longchat-7b-v1.5-32k
Vicuna-7B-v1.5	16K*	LLaMA-2-7B	https://huggingface.co/lmsys/vicuna-7b-v1.5-16k
Vicuna-13B-v1.5	16K*	LLaMA-2-13B	https://huggingface.co/lmsys/vicuna-13b-v1.5-16k
LongAlign-7B	64K	LLaMA-2-7B	https://huggingface.co/THUDM/LongAlign-7B-64k
LongAlign-13B	64K	LLaMA-2-13B	https://huggingface.co/THUDM/LongAlign-13B-64k

Table 4. Summary of the long-context models compared in Section 4. *Vicuna models are provided with a 16K context limit, but it was extended to 32K using RoPE extrapolation (Su et al., 2024).

to the model checkpoint on Huggingface for open-source models or the link to the relevant OpenAI documentation for proprietary models.

The inference was done on a single A100 GPU with 80GB memory. In preliminary experiments, we also attempted to include the Mistral-7B-Instruct-v0.2[11] model in our study, as this model supports a context of up to 32K tokens. However, running this model on ELITR-Bench led to a GPU out-of-memory error on the A100, and thus we discarded it.

B.2. Configuration search on ELITR-Bench-QA's dev set

In our pilot experiments, we noted that the open-source models retained for our study tended to be fairly impacted by the choice of the prompt and the inference hyperparameters. Therefore, we conducted a search on the inference configuration space to select appropriate hyperparameters for each open-source model.[12] The configuration search was carried out in two steps on the dev set of ELITR-Bench-QA, in the single-turn mode. The evaluation was performed using GPT-4 as the evaluator, as described in the evaluation protocol in Section 3.

In the first step of the search – whose results are given in Table 5 – we varied three dimensions in the inference:

- The decoding method, which was either greedy decoding or nucleus sampling with a temperature of 0.6 and top-p of 0.9;
- The use (or absence) of a chat template,[13] which modifies the prompt to integrate the same tags used in the fine-tuning stage – those tags varying across models;
- The use (or absence) of question-answer markers, which introduces to the prompt the tokens 'QUESTION:' and 'ANSWER:' before the question and the expected answer, respectively.

The specific chat template we adopted for each model is based on the one used during the model's fine-tuning: the LLaMA2 template for LongAlpaca-7B and LongAlpaca-13B; the Vicuna template for LongChat-7B-v1.5, Vicuna-7B-v1.5 and Vicuna-13B-v1.5; and the LongAlign template for LongAlign-7B and LongAlign-13B.

In the second step of the search, we used the configuration that yielded the best results on the first step for each model and tested the impact of setting the repetition penalty hyperparameter to 1.1 (instead of the default 1.0 value) in the inference. The results of step 2 are provided in Table 6.

Ultimately, the following configurations were retained for each model:

- **LongAlpaca-7B:** greedy decoding with a chat template and QA markers;
- **LongAlpaca-13B:** greedy decoding with a chat template;

[11] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[12] In comparison to open-source models, we found that GPT-4 and GPT-3.5 were more robust to differences in the inference configuration. Therefore, given the extensive cost of doing a large number of runs for commercial models, we did not conduct a configuration search on these.

[13] https://huggingface.co/docs/transformers/main/en/chat_templating

Decoding	Chat templ.	QA mark.	LongAlpaca-7B	LongAlpaca-13B	LongChat-7B-v1.5	Vicuna-7B-v1.5	Vicuna-13B-v1.5	LongAlign-7B	LongAlign-13B
Greedy	Y	Y	**5.89**	6.13	6.22	4.94	5.19	6.04	6.16
Greedy	Y	N	5.55	**6.17**	**6.60**	5.38	5.13	**6.11**	6.16
Greedy	N	Y	5.89	5.87	6.23	5.05	4.71	5.43	5.94
Nucleus	Y	Y	5.18	5.91	6.19	4.99	**5.70**	5.67	6.25
Nucleus	Y	N	5.61	6.11	5.85	5.33	5.00	6.06	**6.27**
Nucleus	N	Y	5.58	5.96	5.91	**5.42**	4.89	5.18	5.99

Table 5. Results of step 1 for our configuration search on ELITR-Bench-QA's dev set, in the single-turn mode. The configuration corresponding to using neither a chat template nor QA markers is not included as this was shown to severely underperform in our preliminary experiments.

Repetition penalty	LongAlpaca-7B	LongAlpaca-13B	LongChat-7B-v1.5	Vicuna-7B-v1.5	Vicuna-13B-v1.5	LongAlign-7B	LongAlign-13B
Y	5.80	5.73	6.11	4.90	**5.92**	5.90	6.21
N	**5.89**	**6.17**	**6.60**	**5.42**	5.70	**6.11**	**6.27**

Table 6. Results of step 2 for our configuration search on ELITR-Bench-QA's dev set, in the single-turn mode. In the cases where we include a repetition penalty, we set the corresponding hyperparameter to 1.1 (instead of 1.0, the default value corresponding to no repetition penalty).

- **LongChat-7B-v1.5:** greedy decoding with a chat template;
- **Vicuna-7B-v1.5:** nucleus sampling with QA markers;
- **Vicuna-13B-v1.5:** nucleus sampling with a chat template, QA markers, and repetition penalty;
- **LongAlign-7B:** greedy decoding with a chat template;
- **LongAlign-13B:** nucleus sampling with a chat template.

For the proprietary models, **GPT-3.5** and **GPT-4**, we used nucleus sampling (temperature = 0.6 and top-p = 0.9) with the standard OpenAI chat template.

The cost of the two-step configuration search amounted to approximately \$150.[14] To limit excessive expenses, we used the same model configuration for the different settings we experimented in (single-turn ELITR-Bench-QA, multi-turn ELITR-Bench-QA, and multi-turn ELITR-Bench-Conv).

C. Additional experimental results

C.1. Variance over seeded run results

To account for the seed-dependent variability in the evaluation, we performed 3 seeded runs on the test set. The set of seeds used is $\{2023, 2024, 2025\}$. The results are reported in Table 7, where we indicate for each (model, setting) pair the mean score over the 3 seeds as well as the sample standard deviation.

Note that the same seed is used both for the response generation part and the GPT-4-based evaluation part, as both can be sources of variance in the reported results. Based on our configuration search (see Appendix B.2), some of the response generation models were set to use greedy decoding: LongAlpaca-7B, LongAlpaca-13B, LongChat-7B-v1.5, LongAlign-7B. For such models, the response generation is deterministic and the only source of variance is that of the GPT-4 evaluator.

The results from Table 7 show that the variance across settings is fairly different. In the single-turn ELITR-Bench-QA setting, the standard deviation for all models remain relatively low, even for the models that use nucleus sampling (GPT-3.5,

[14] We assessed the cost of performing the evaluation of a single model on the 141 dev set questions to \$3 approximately. As we evaluated 7 models on 6 configurations in the first step, and 7 models on 1 configuration in the second step, this yields \$147.

Model	Single-turn	Multi-turn	
	ELITR-Bench-QA (test set)	ELITR-Bench-QA (test set)	ELITR-Bench-Conv (test set)
GPT-3.5	7.44 ± 0.12	-	-
GPT-4	**8.38 ± 0.07**	**8.42 ± 0.09**	**8.36 ± 0.12**
LongAlpaca-7B	5.60 ± 0.06	4.84 ± 0.02	4.58 ± 0.04
LongAlpaca-13B	6.25 ± 0.05	4.71 ± 0.01	4.74 ± 0.06
LongChat-7B-v1.5	5.78 ± 0.06	4.17 ± 0.07	4.31 ± 0.07
Vicuna-7B-v1.5	5.61 ± 0.17	4.61 ± 0.26	4.69 ± 0.34
Vicuna-13B-v1.5	**6.52 ± 0.16**	**5.67 ± 0.10**	**5.78 ± 0.13**
LongAlign-7B	6.46 ± 0.07	4.47 ± 0.01	5.06 ± 0.03
LongAlign-13B	6.33 ± 0.09	5.33 ± 0.47	4.95 ± 0.22

Table 7. Results for the seeded runs on the test set for different ELITR-Bench settings. The reported numbers correspond to the mean score ± sample standard deviation computed over 3 seeds. Boldface numbers correspond to the best performance among proprietary or open-source models. The results for GPT-3.5 are omitted in the multi-turn setting as the context length exceeded the 16K limit of this model.

Model	Question type				Answer position			
	Who (N=45)	What (N=57)	When (N=20)	How many (N=8)	Begin (N=43)	Middle (N=34)	End (N=22)	Several (N=31)
GPT-3.5	7.91	6.94	7.68	7.79	7.33	7.45	7.76	7.37
GPT-4	**8.56**	**8.29**	**8.28**	**8.29**	**8.36**	**8.29**	**8.32**	**8.57**
LongAlpaca-7B	5.35	5.37	6.35	**6.79**	5.81	5.80	4.97	5.53
LongAlpaca-13B	**7.19**	5.47	**6.47**	6.00	5.93	5.95	**6.85**	6.59
LongChat-7B-v1.5	6.88	4.94	6.33	4.17	6.41	4.91	5.89	5.77
Vicuna-7B-v1.5	6.13	5.65	5.40	2.88	5.89	5.21	4.96	6.12
Vicuna-13B-v1.5	6.96	6.68	5.48	5.54	6.35	**6.41**	6.55	**6.87**
LongAlign-7B-64k	6.93	6.33	6.00	5.88	**7.09**	6.39	6.47	5.66
LongAlign-13B-64k	6.08	**6.74**	5.97	5.75	6.71	6.21	6.33	5.95

Table 8. Results by question type and answer position on the test set of ELITR-Bench-QA in single-turn mode. The number N below a subset indicates the corresponding subset size.

GPT-4, Vicuna-7B-v1.5, Vicuna-13B-v1.5, LongAlign-13B). However, in the multi-turn settings, we observe an increased standard deviation for those same models overall, in particular for Vicuna-7B-v1.5 and LongAlign-13B. We hypothesize that the sequence of questions asked in the same conversation in the multi-turn setting causes different seeded runs to cumulate errors and slightly diverge along the course of the conversation.

C.2. Additional results on question type and answer position

In this section, we provide the full results per question type and answer position to expand the compact results of the GPT and LLaMA-2 model families given in Section 4.2. The results are given in Table 8 and were obtained on ELITR-Bench-QA's test set in the single-turn setting. Looking at the global model performance over the different question types and answer positions, we do not identify any clear trend highlighting a question type or position answer as notably easier or harder.

In constrast, past work reported a "lost in the middle" effect (Liu et al., 2023b), stating that the middle of a model's context tends to be overlooked more often than the beginning or end of the context. To further investigate this phenomenon in our dataset, we conducted a statistical hypothesis test on the scores obtained by each individual model. Specifically, we ran a one-tailed Welch's t-test (Welch, 1947) with the following alternative hypothesis: "The average score for questions with middle-position answers is lower than the average score of other questions". The p-values obtained for each model's set of scores are given in Table 9. Interestingly, we observe that the "lost in the middle" hypothesis is statistically verified (p-value < 0.05) for only two models: LongChat-7B-v1.5 (p-value $= 0.032$) and Vicuna-7B-v1.5 (p-value $= 0.046$). While we

Model	p-value
GPT-3.5	0.466
GPT-4	0.372
LongAlpaca-7B	0.713
LongAlpaca-13B	0.265
LongChat-7B-v1.5	**0.032**
Vicuna-7B-v1.5	**0.046**
Vicuna-13B-v1.5	0.469
LongAlign-7B	0.409
LongAlign-13B	0.413

Table 9. Results of a one-tailed Welch's t-test on the alternative hypothesis "The average score for questions with middle-position answers is lower than the average score of other questions", to verify the presence or absence of a "lost in the middle" effect (Liu et al., 2023b). Boldface numbers denote statistically significant results (p-value < 0.05).

do not have a clear explanation about which of these two models' characteristics caused that effect, these models have in common that they are based on LLaMA-2-7B and were trained by the same LMSYS organization (Chiang et al., 2023; Li et al., 2023a). It is then possible – although purely hypothetical – that the specific fine-tuning recipe followed by LMSYS on LLaMA-2-7B for these two models led to the "lost in the middle" effect.

C.3. Results on QA/Conv differentiating questions

As introduced in Section 2, some of the questions differ between ELITR-Bench-QA and ELITR-Bench-Conv and typically contain pronominal references or ellipses in the Conv setting, which makes them particularly challenging to tackle. In this section, we look at the results on this subset of differentiating questions – both in their QA and Conv versions – and study the impact of using the single-turn or multi-turn mode. The results are provided in Fig. 1, which compares 3 settings: single-turn mode with QA questions, multi-turn mode with QA questions, and multi-turn mode with Conv questions. The reported scores are averaged over the dev and test sets' differentiating questions (respectively, 16 and 17 questions) to make up for the limited size of these subsets.

Similarly with what we observed in Table 2, we notice again a clear difference between GPT-4 and open-source models: the performance of the former improves (slightly) from single-turn to multi-turn, whereas the performance of the latter notably degrades. In contrast with our previous findings that showed little to no difference between the results on ELITR-Bench-QA and ELITR-Bench-Conv for the multi-turn mode, we observe this time that the average score decreases from QA to Conv for open-source models. While the difference is small, this trend is present for all open-source models except LongChat-7B-v1.5. This trend was expected as the Conv questions in this subset are more challenging to answer. However, interestingly, GPT-4 results did not show the same trend. We hypothesize that the opposite trends identified for GPT-4 and open-source models might be explained by a 'snowballing' effect that causes an error propagation in lower-performing open-source models and instead provides additional helpful context for GPT-4.

C.4. Updated results with recent long-context LLMs

In this section, we provide updated results to include some of the latest long-context LLMs that have been released after the core experiments of this paper have been conducted. In particular, we added the three following models: OpenAI's GPT-4o,[15] GradientAI's LLaMA-3-8B-Instruct-262K[16] which was fine-tuned from the original LLaMA-3 model,[17] and Microsoft's Phi-3-Small-128K-Instruct.[18] For GPT-4o, we used the same hyperparameters as for GPT-4 and GPT-4.5: nucleus sampling with a temperature of 0.6 and top-p of 0.9. For LLaMA-3-8B-Instruct-262K and Phi-3-Small-128K-Instruct, inference was done by greedy decoding and we used their original chat templates. QA markers and repetition penalty have not been included.

[15] https://platform.openai.com/docs/models/gpt-4o
[16] https://huggingface.co/gradientai/Llama-3-8B-Instruct-262k
[17] https://llama.meta.com/llama3/
[18] https://huggingface.co/microsoft/Phi-3-small-128k-instruct

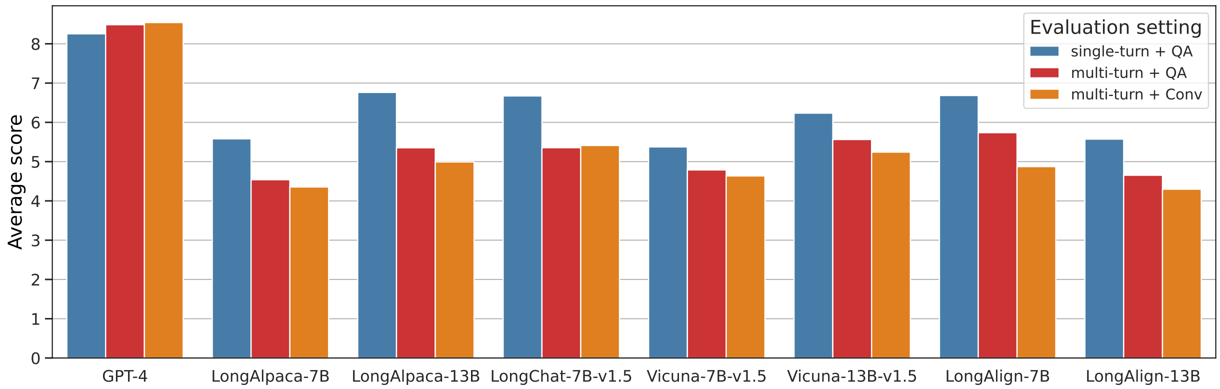


Figure 1. Results restricted to QA/Conv differentiating questions. The score reported for each model and evaluation setting corresponds to the average of the scores obtained on the dev subset (16 questions) and the test subset (17 questions). Best viewed in color.

Model	Single-turn ELITR-Bench-QA		
	Dev	Test	Mean
GPT-3.5	7.04	7.44	7.24
GPT-4	8.21	8.39	8.30
GPT-4o	**8.50**	**8.44**	**8.47**
LongAlpaca-7B	5.89	5.60	5.75
LongAlpaca-13B	6.17	6.25	6.21
LongChat-7B-v1.5	6.60	5.78	6.19
Vicuna-7B-v1.5	5.42	5.61	5.51
Vicuna-13B-v1.5	5.92	6.52	6.22
LongAlign-7B	6.11	6.46	6.28
LongAlign-13B	6.27	6.33	6.30
LLaMA-3-8B-Instruct-262K	6.83	6.51	6.67
Phi-3-Small-128K-Instruct	**7.37**	**7.34**	**7.36**

Table 10. Results on single-turn ELITR-Bench-QA updated with recent long-context LLMs: GPT-4o, LLaMA-3-8B-Instruct-262K, and Phi-3-Small-128K-Instruct. The reported numbers correspond to the average scores from 1 to 10 (higher is better) obtained by a GPT-4 evaluator, on a single seeded run for the dev set and 3 seeded runs for the test set. Boldface numbers correspond to the best performance among proprietary or open-source models.

The comparison of these additional models and previously used approaches is provided in Table 10. These results show that GPT-4o performed essentially on par with GPT-4, with a small gain on the dev set. Interestingly, LLaMA-3-8B-Instruct-262K remains close in performance to LLaMA-2 models and only slightly outperforms the best among these, while Phi-3-Small-128K-Instruct beats both LLaMA-2 and LLaMA-3 models by a good margin. It even outperforms GPT-3.5 despite having only 7B parameters. This suggests that this Phi-3 model is a good open-source alternative to more powerful models such as GPT-4 and GPT-4o for our meeting assistant task.

D. LLM-based evaluation assessment

In this appendix, we seek to verify the validity of the LLM-based (namely, GPT-4-based) evaluation methodology introduced in Section 3 and applied in Section 4. In Appendix D.1, we define the LLM-based and human-based evaluators that we considered for comparison. Then, Appendix D.2 presents our results and findings on the evaluator comparison.

D.1. Compared evaluators

Our evaluation assessment experiment consists in checking the validity of the numeric scores (from 1 to 10) assigned for each tuple composed of a question, its ground-truth answer, and an LLM response to evaluate. For that purpose, we

Model	Evaluator			
	GPT-4	**Prometheus**	**Gold Human**	**Silver Human**
GPT-4	8.33	5.68	7.93	7.21
Vicuna-13B-v1.5	6.69	4.80	6.19	5.80
LongAlpaca-7B	5.57	4.46	4.55	4.72

Table 11. Comparison of the scores obtained by different evaluators for the responses generated by GPT-4, Vicuna-13B-v1.5, or LongAlpaca-7B. The evaluation was performed on ELITR-Bench-QA's test set in the single-turn mode, and for a single seeded run.

compared the score annotations obtained through two LLM-based evaluators and two human-based evaluators:

- **GPT-4** (OpenAI, 2023): This evaluator corresponds to the one detailed in the evaluation protocol in Section 3 and is based on the gpt-4-0613 model.
- **Prometheus** (Kim et al., 2024): This fine-tuned model was originally proposed to provide an open-source alternative to using GPT-4 for score rubric-based evaluation. We used the Prometheus-13B-v1.0[19] model, with a prompt similar to the one adopted for GPT-4 – the only difference is that the score rubric is re-scaled to a 1-5 range to fit Prometheus' expected format and multiplied by 2 in post-processing to be comparable to other scores. The prompt is available in Appendix G (Figs. 7 and 8).
- **Gold Human**: This expert human annotation was done by one of the authors. The scores were assigned following the same 10-point score rubric as the one used for the GPT-4 evaluator (given in Appendix G, Fig. 6), to enforce consistency across questions.
- **Silver Human**: This evaluator is based on a crowdsourcing study with the Prolific[20] platform where we averaged the scores assigned by 10 human annotators for each question. The annotators were provided with the same score rubric as for GPT-4 and Gold Human. We give more details on this evaluation in Appendix E.

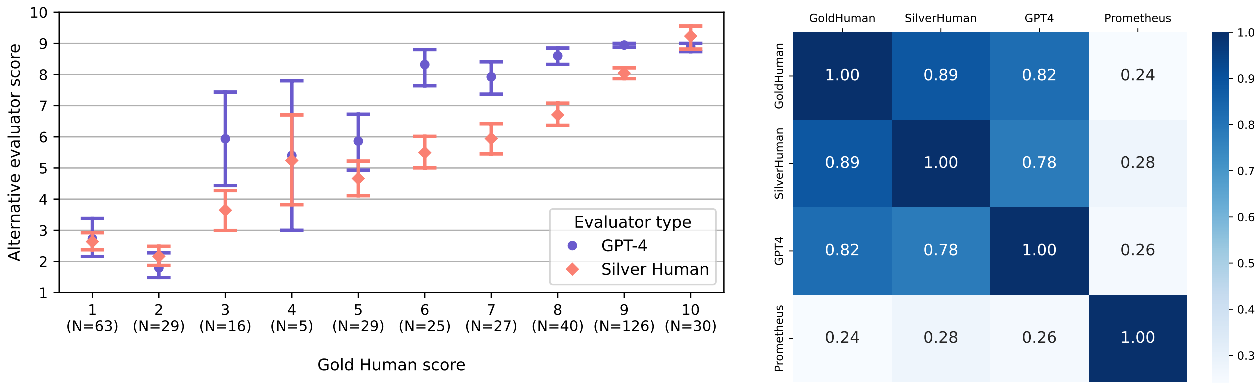
Given the costly nature of human annotations, we performed our evaluation assessment on a small subset of the experiments described in Section 4.1. We specifically focused on the results of ELITR-Bench-QA's test set in the single-turn mode. We looked at the results of 3 models that performed diversely in this setting: the best proprietary model (GPT-4), the best open-source model (Vicuna-13B-v1.5), and the worst open-source model (LongAlpaca-7B).

D.2. Evaluator comparison results

Model-level comparison. To get a high-level, coarse-grained comparison of the different evaluators introduced above, we applied each of them to the responses generated by GPT-4, Vicuna-13B-v1.5, and LongAlpaca-7B. The results of the corresponding evaluations are presented in Table 11. We can first observe that the ranking of the three models to evaluate is the same for all the evaluators: GPT-4, Vicuna-13B-v1.5, and LongAlpaca-7B (from the most highly rated model to the most poorly rated one). However, we found that the range of scores was more diverse: Prometheus' scores were overall fairly low (from 4 to 6), while GPT-4's scores are much higher (from 5 to 9). In comparison, the human scores from the Gold Human and Silver Human evaluators were more similar to GPT-4 with scores between 4 and 8.

Correlation analysis. To get a deeper understanding of how evaluators compare to one another, we calculated the Pearson correlation for every evaluator pair on the responses aggregated over the 8 meetings of the test set and generated by the three retained models. The results are displayed in Fig 2b. GPT-4 shows a strong correlation with the two human-based evaluators (0.82 with Gold Human and 0.78 with Silver Human), which is in agreement with the findings from previous studies on GPT-4 judges (Kim et al., 2024; Bai et al., 2024). Prometheus, on the other hand, yielded a weak correlation (between 0.2 and 0.3) with all the other evaluators. We hypothesize that this could be due to a domain shift with respect to what Prometheus was fine-tuned on, caused by the nature of the meeting-related questions and the presence of anonymized entities (e.g., [PERSON3]). Turning to the two human-based evaluators, Gold Human and Silver Human obtained a very

[19] https://huggingface.co/kaist-ai/prometheus-13b-v1.0
[20] https://www.prolific.com/



(a) (b)

Figure 2. **(a)** Distribution of GPT-4 and Silver Human scores with respect to each Gold Human score bin (1-10); the N below a score bin indicates the bin size. **(b)** Pearson correlation between evaluators.

strong correlation of 0.89 which confirms the validity of the crowdsourcing study and the feasibility of the annotation task by non-expert judges.

Comparison of score distribution across evaluators. So far in this section, we have found that GPT-4 and human-based evaluators lead to scores that are highly correlated (see Figure 2b) but with slightly different score ranges (see Table 11). This led us to investigate how scores are distributed for different evaluators, and to study to what extent score levels match across evaluators. For that purpose, we considered the pool of (question, response, score) tuples obtained with the Gold Human evaluator on the responses from GPT-4, Vicuna-13B-v1.5, and LongAlpaca-7B for the 130 questions of the test set, i.e., 390 instances in total. We split these instances into 10 bins based on their score value from 1 to 10. Then, for all the instances in a bin, we check the distribution of the scores obtained by other evaluators on the bin's (question, response) pairs. In practical terms, we seek to highlight through this procedure how Gold Human and alternative evaluators align at the grade level. The results are plotted in Fig. 2a where we describe the score distribution of the alternative evaluator through its means and 95% confidence intervals. Interestingly, we observe that the scores for the GPT-4 evaluator seem to fall into 3 distinct clusters, corresponding respectively to the intervals [1, 2], [3, 5] and [6, 10] in the Gold Human scores. This suggests that despite the use of a 10-point score rubric to align the GPT-4 evaluator's scores with detailed desiderata, this evaluator is only able to distinguish between three levels of response quality. This finding then leads us to question the common practice of using LLM-based evaluator scores on a 5-point or 10-point scale. In contrast, the scores from Silver Human show a more linear relationship with the Gold Human scores, suggesting that implementing the 10-point score rubric in the crowdsourcing study aided in achieving a closer alignment between external human annotators and the evaluation criteria set by the organizers.

E. Crowdsourcing study details

Our *Silver Human* evaluation is based on a crowdsourcing study using the Prolific[21] platform. A task in this study consists in scoring the responses of the 3 considered models (GPT-4, Vicuna-13B-v1.5, and LongAlpaca-7B) for all the questions of a single meeting – out of 8 meetings in the test set. For each meeting, we hired 10 annotators, without constraining the 10 annotators to be the same across meetings. Participants were screened based on their primary language (English) and domain expertise (including Computer Science, Information Technology, Engineering, or Mathematics). Each participant received £9 per hour when completing a task (with each task comprising approximately 40-50 questions for assessment). We estimated the task duration to be around 30 minutes – our post-analysis indicated a median time spent per study ranging between 16 and 29 minutes depending on the meeting. We discarded the annotations that were flagged as too inconsistent with Gold Human scores, and hired new annotators when needed until we had a satisfactory set of 10 annotators per meeting. In total, the crowdsourced Prolific evaluation cost was £400.

[21] https://www.prolific.com/

Meeting ID	ICC
01	0.872
02	0.964
03	0.912
04	0.941
05	0.906
06	0.940
07	0.936
08	0.942
all	0.965

Table 12. Intra-class correlation (ICC) coefficients across annotators from the Prolific crowdsourcing study, corresponding to the Silver Human evaluator.

The guidelines provided for this study start with general information about the task as well as the 10-point score rubric given in Fig. 6, in order to help annotators calibrate their scores with concrete criteria. Then the interface presents a tuple composed of a question, its ground-truth answer, and an LLM response to evaluate. From this tuple, the annotator is asked to grade the LLM response with a score ranging from 1 to 10, following the provided score rubric. A screenshot of our interface is shown in Fig. 3.

To measure the inter-annotator agreement, we used the intra-class correlation (ICC) coefficient (Koo & Li, 2016) which assesses how consistent annotators' scores are for every (question, ground-truth answer, LLM response) tuple. The ICC results are detailed in Table 12 for each individual meeting and overall. For individual meetings, we report the two-way coefficient ICC(2,k) as the set of hired annotators is the same across all the questions of a given meeting. For the result over all meetings, we used instead the one-way coefficient ICC(1,k) since the set of annotators differs across meetings. Most of the ICC coefficients being above 0.9 suggests an excellent inter-annotator agreement, following the interpretation guidelines from (Koo & Li, 2016).

F. ELITR-Bench excerpt

We provide in Table 13 an excerpt of meeting 010 from the dev set of the ELITR corpus (Nedoluzhko et al., 2022). Entities, such as (PERSON10), (PERSON19), and [ORGANIZATION11], have been de-identified in the original work for the sake of anonymization. Below the excerpt, we provide 4 questions (and their respective answers) related to the same meeting, which have been added through the proposed ELITR-Bench. For each question, we indicate its type between brackets (i.e., *Who*, *What*, *When*, or *How many*).

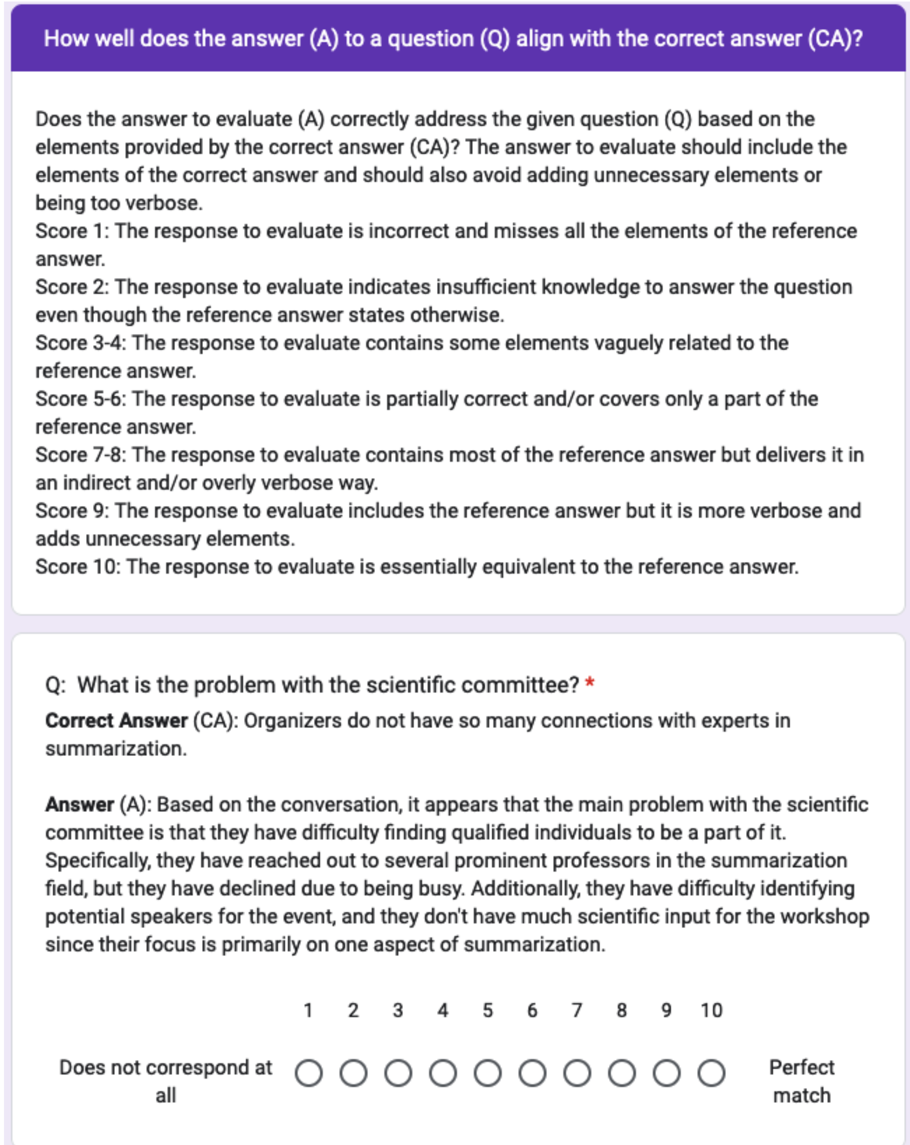
G. Prompts

In this section, we list the different prompts used in the paper, both for response generation and evaluation. The prompt for response generation follows the same general template given in Fig. 4 for every evaluated model – both proprietary and open-source models. Then, questions and answers are appended to the prompt as described in Section 3 – either a single question per conversation in the single-turn mode, or all the questions of a meeting in sequence in the multi-turn mode. As detailed in Appendix B.2, we slightly modify this base prompt depending on the model-specific selected configuration. As a reminder, these alterations may take two forms: the use of a chat template (which only adds special tags to the prompt) and the use of question-answer markers (which add 'QUESTION:' before a question and 'ANSWER:' before an answer).

The prompts that we used for evaluation are inspired from the prompt originally proposed in (Kim et al., 2024) and include: the question, the response to evaluate, the ground-truth answer, and a score rubric. Note that the transcript is not included in the evaluation prompt as the question and ground-truth answer should provide sufficient information to assess the correctness of the response to evaluate. The full prompts are given in Fig. 5 for the GPT-4 evaluator and in Fig. 7 for the Prometheus evaluator. Their score rubrics are shown in Fig. 6 and Fig. 8, respectively. For Prometheus, we had to adapt the 10-point score scale to a 5-point scale to match the format used when this model was fine-tuned (Kim et al., 2024). The 5-point rubric was defined to retain the main criteria expressed in the 10-point rubric and minimally alter it to enable a fair comparison between the two evaluators.

Transcript excerpt	... (PERSON19) Just <unintelligible/> like a virtual machine image. (PERSON10) Yeah, yeah. (PERSON19) You just fire up, an- anyone can fire up, it's not like you have to you have to call- (PERSON10) Yeah. (PERSON19) Like [ORGANIZATION11], get them to run it. (PERSON10) Yeah. (PERSON19) I I don't know that's easier, but I mean it it's more more flexible. (PERSON10) Yeah, yeah. I haven't since I haven't really done it, it's uh, it's hard for me to access, so we- (PERSON19) I know, I know. (PERSON10) You know. Uh, okay, so that's good, we know what to do. I don't know whether we'll manage to have these systems package before the demo, but hopefully uh, there won't be any power outage an our uh, at our site. (PERSON19) <laugh/> (PERSON10) <laugh/> So that was the 1 thing, that I've learnt, that we must not uh, that that we must have uh, rep-replicated uh, components across the site. ...
Question (What)	Which risk, related to the demo, was discussed?
Answer	Power outages at [ORGANIZATION2]
Question (Who)	Which entity is running the translation module?
Answer	[ORGANIZATION11]
Question (What)	What should be frozen 1 or 2 weeks before the demo?
Answer	The stable components of the systems should be frozen 1-2 weeks before the demo
Question (When)	When should the recorded demo be provided?
Answer	17th of June

Table 13. Small excerpt of meeting 010 from ELITR's dev set, with sample questions and answers related to the same meeting from ELITR-Bench.



How well does the answer (A) to a question (Q) align with the correct answer (CA)?

Does the answer to evaluate (A) correctly address the given question (Q) based on the elements provided by the correct answer (CA)? The answer to evaluate should include the elements of the correct answer and should also avoid adding unnecessary elements or being too verbose.
Score 1: The response to evaluate is incorrect and misses all the elements of the reference answer.
Score 2: The response to evaluate indicates insufficient knowledge to answer the question even though the reference answer states otherwise.
Score 3-4: The response to evaluate contains some elements vaguely related to the reference answer.
Score 5-6: The response to evaluate is partially correct and/or covers only a part of the reference answer.
Score 7-8: The response to evaluate contains most of the reference answer but delivers it in an indirect and/or overly verbose way.
Score 9: The response to evaluate includes the reference answer but it is more verbose and adds unnecessary elements.
Score 10: The response to evaluate is essentially equivalent to the reference answer.

Q: What is the problem with the scientific committee? *

Correct Answer (CA): Organizers do not have so many connections with experts in summarization.

Answer (A): Based on the conversation, it appears that the main problem with the scientific committee is that they have difficulty finding qualified individuals to be a part of it. Specifically, they have reached out to several prominent professors in the summarization field, but they have declined due to being busy. Additionally, they have difficulty identifying potential speakers for the event, and they don't have much scientific input for the workshop since their focus is primarily on one aspect of summarization.

	1	2	3	4	5	6	7	8	9	10	
Does not correspond at all	○	○	○	○	○	○	○	○	○	○	Perfect match

Figure 3. Interface for our Prolific crowdsourcing study to collect Silver Human score annotations.

The following is the transcript of a meeting with multiple participants, where utterances start with the speaker's anonymized name (for instance (PERSON4)) and may span over several lines.

{transcript}

As a professional conversational assistant, your task is to answer questions about the meeting by making inferences from the provided transcript.

Figure 4. Answer prompt used to obtain LLMs' responses. Questions are appended to this prompt as described in Section 3. The element in blue and enclosed in curly brackets corresponds to a meeting-specific text span that is dynamically adapted.

Task description:
You are provided below with a question, a response to evaluate, a reference answer that gets the maximum score of 10, and a score rubric representing evaluation criteria.
1. Write a detailed feedback that assess the quality of the response strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between 1 and 10. You should refer to the score rubric.
3. The output format should first include the feedback and then indicate the integer score in \boxed{}.
4. Please do not generate any other opening, closing, and explanations.

Question:
{question}

Response to evaluate:
{response}

Reference answer (score 10):
{reference}

Score rubric:
{rubric}

Feedback:

Figure 5. Evaluation prompt for the GPT-4 evaluator, inspired from Kim et al. (2024). The elements in blue and enclosed in curly brackets correspond to question-specific text spans that are dynamically adapted.

[Does the response to evaluate correctly address the given question based on the elements provided by the reference answer? The response should include the elements of the reference answer and should also avoid adding unnecessary elements or being too verbose.]
Score 1: The response to evaluate is incorrect and misses all the elements of the reference answer.
Score 2: The response to evaluate indicates insufficient knowledge to answer the question even though the reference answer states otherwise.
Score 3-4: The response to evaluate contains some elements vaguely related to the reference answer.
Score 5-6: The response to evaluate is partially correct and/or covers only a part of the reference answer.
Score 7-8: The response to evaluate contains most of the reference answer but delivers it in an indirect and/or overly verbose way.
Score 9: The response to evaluate includes the reference answer but it is more verbose and adds unnecessary elements.
Score 10: The response to evaluate is essentially equivalent to the reference answer.

Figure 6. Score rubric for the GPT-4 evaluator. Boldface is added for the sake of readability and is not included in the actual prompt.

Task description:
You are provided below with a question, a response to evaluate, a reference answer that gets the maximum score of 5, and a score rubric representing evaluation criteria.
1. Write a detailed feedback that assesses the quality of the response strictly based on the given score rubric, not evaluating in general.
2. After writing a feedback, write a score that is an integer between 1 and 5. You should refer to the score rubric.
3. The output format should look as follows: "Feedback: (write the quality assessment feedback) [RESULT] (an integer number between 1 and 5)".
4. Please do not generate any other opening, closing, and explanations.

Question:
{question}

Response to evaluate:
{response}

Reference answer (score 5):
{reference}

Score rubric:
{rubric}

Feedback:

Figure 7. Evaluation prompt for the Prometheus evaluator, inspired from Kim et al. (2024). The elements in blue and enclosed in curly brackets correspond to question-specific text spans that are dynamically adapted.

[Does the response to evaluate correctly address the given question based on the elements provided by the reference answer? The response should include the elements of the reference answer and should also avoid adding unnecessary elements or being too verbose.]
Score 1: The response to evaluate is incorrect and misses all the elements of the reference answer.
Score 2: The response to evaluate contains some elements vaguely related to the reference answer.
Score 3: The response to evaluate is partially correct and/or covers only a part of the reference answer.
Score 4: The response to evaluate contains most of the reference answer but delivers it in an indirect and/or overly verbose way.
Score 5: The response to evaluate is essentially equivalent to the reference answer.

Figure 8. Score rubric for the Prometheus evaluator. Boldface is added for the sake of readability and is not included in the actual prompt.