
Retrieval Augmented Generation for Dialog Modeling

Lilly Kumari *
University of Washington
lkumari@uw.edu

Usama Bin Shafqat
Google
ushafqat@google.com

Nikhil Sarda
Google
nikhilsarda@google.com

Abstract

In this work, we explore the use of Large Language Models (LLMs) for the challenging task of long-range dialog modeling. While LLMs have excelled in various Natural Language Processing (NLP) tasks, adapting them for extended dialog contexts poses challenges due to computational overhead and data requirements. LLMs often struggle with fixed context window sizes, limiting their application in lengthy conversations. In this work, we leverage LLMs’ contextual learning capabilities using instruction prompts and retrieval-based context augmentation, without any fine-tuning. We focus on long-term dialog modeling, addressing challenges like data independence, avoiding fine-tuning, and accommodating the context of long conversations within shorter windows. Our empirical experiments on two datasets, namely Multi-Session Chat and MultiDoc2Dial demonstrate how including relevant information in LLMs’ input context affects dialog generation performance while reducing computational costs associated with longer contexts.

1 Introduction

Transformers-based pre-trained Large Language Models (LLMs) [44, 9, 4, 43, 30] have demonstrated exceptional performance across a spectrum of Natural Language Processing (NLP) tasks. However, fine-tuning these LLMs for different use cases such as long-range dialog modeling poses significant challenges due to (1) computational overhead associated with updating the model parameters in the order of billions and (2) dependence on the availability of labeled task-specific data. Additionally, these LLMs are trained using pre-defined context window size, thus, limiting their direct application to dialog modeling when the conversation context goes beyond the pre-set context window [12, 32]. Besides adaptation to longer context windows, processing longer sequences is computationally intensive due to self-attention scaling quadratically with context length during training and inference [57, 3].

Due to extensive pre-training on large corpora, most LLMs exhibit a remarkable ability for In-Context Learning (ICL) via instruction prompting [33, 15, 46, 47, 11]. This ability empowers their versatile deployment across various end tasks without the need for any additional fine-tuning. Recent research advances in retrieval-augmented LLMs [24] have explored the augmentation of parametric LLMs with external knowledge sources, demonstrating their effectiveness in knowledge-intensive NLP tasks [23, 39, 18] while still eliminating the need for further fine-tuning.

In this work, we focus on the application of LLMs for long-range dialog modeling. In particular, we aim to model long-term dependencies in conversational data for person-grounded and knowledge-grounded dialog modeling while restricting ourselves to the following settings in line with the challenges covered earlier: (1) no labeled data dependence, (2) no additional LLM fine-tuning, and (3) using shorter conversation context to reduce the pressure on the context window available to the LLM.

Drawing inspiration from retrieval-augmented generation techniques and emergent capabilities of LLMs such as ICL via instruction prompting, we utilize LLMs to retrieve (1) relevant persona details

*work done during internship at Google

from the entire historical context for facilitating persona-aware dialog modeling and (2) relevant context from external knowledge sources for knowledge-grounded dialog modeling. This retrieved information is subsequently integrated (or augmented) with an instruction prompt and the latest conversation context to assist the LLM in generating the next dialog.

We conduct empirical experiments on (1) Multi-Session Chat (MSC) dataset [50] for persona-aware dialog modeling of conversations spanning over multiple sessions and (2) MultiDoc2Dial dataset [16] for knowledge-grounded dialog modeling, using LLMs of different complexities. We thoroughly examine how the inclusion of varying amounts of relevant information in the LLMs’ input context affects the final performance for dialog generation while taking into consideration the computational costs associated with longer context lengths when more amount of relevant information is included.

2 Related Work

In the context of our problem setting, existing research can be divided into three main groups: (1) Transformers and their long-range versions, (2) Retrieval Augmented LLMs, and (3) LLMs for Dialog Modeling. A comprehensive discussion of these groups can be found in the Appendix A.

3 Retrieval Augmented Generation

3.1 Notations & Definitions

Long-term conversations can span over extended time periods, often divided into multiple sessions, each addressing specific segments of the discussion. Denoting the user as u and the chatbot agent as a , the entire history conversation for session s at turn t can be denoted as follows: $H_t^s = \{u_1^1, a_1^1, u_2^1, a_2^1, \dots, u_1^s, a_1^s, u_2^s, a_2^s, \dots, u_t^s\}$. Here, the superscript denotes the session and the subscript denotes the dialog turn in a particular session. Given the entire conversation H_t^s , the goal is to generate the agent’s response a_t^s . Fitting the entire history in LLMs’ context window for later sessions can be extremely challenging due to the limited context size, such as 4K for gpt-3.5. In this work, we aim to summarize the dialogues in H_t^{s-1} (sessions prior to the current session s) relevant to the local conversational context. In the case of generating knowledge-grounded responses, prior history H_t^{s-1} can be supplemented/replaced by a knowledge datastore such as wikipedia article segments.

3.2 Methodology

For modeling the next dialog of the chatbot agent given the current context, we consider two different approaches to extract relevant information from the past dialog sessions (or knowledge base) instead of using the entire previous conversation as a part of the input context to the LLM.

kNN-based context selection: The first approach involves performing a simple kNN-based retrieval of dialogues (or knowledge-base passages) that are semantically related to the most recent conversation between the user and the agent. To perform similarity-based retrieval, we compute the text embeddings of all previous dialogues along with the current session dialogues as they get recorded. Using the user’s most recent dialog vector as the query, we perform a semantic search in the embedding space to select the most relevant dialogues from the past dialog sessions that can assist the agent in generating its next response.

Submodular Span Summarization for context selection: The second approach involves generating an extractive conditional (or query-focused) summary of the past sessions’ dialogues conditioned on the most recent conversation history. To compute the query-focused summary, we use a two-stage conditional submodular summarization framework, Submodular Span Summarization (S3) [21] that uses a single submodular function to model both properties of a good query-focused summary: (1) query-relevance and (2) diversity. Using the same set of text embeddings as described earlier, we instantiate a facility function using a cosine similarity-based metric. The most recent conversation utterances are used as candidates for the query set.

Mathematically, given the query set Q , the ground set V (where $V \setminus Q$ contains the past sessions dialogues or knowledge base), and a monotone, non-decreasing submodular function $f : 2^V \rightarrow \mathbb{R}$, we aim to compute a query-focused summary of $V \setminus Q$ in two stages. The first stage minimizes

the submodular conditional gain i.e., $\min_{A \subseteq V \setminus Q, |A| \geq k_1} f(A|Q)$ resulting in a set of dialogues A_Q^* relevant to the current context. The second stage summarizes the redundant output of the first stage by maximizing the submodular function i.e., $\max_{A \subseteq A_Q^*, |A| \leq k_2} f(A)$ using the greedy algorithm [29].

To instruct the LLM to respond to the user’s dialog and follow the provided summary or relevant information, we design instruction prompts that are concatenated along with the summary and the most recent conversation context.

4 Experiments & Results

4.1 Datasets

We empirically validate the effectiveness of our proposed methods for summarizing the long conversation context on two publicly available dialog modeling datasets: (1) Multi-Session Chat (MSC) [50], (2) MultiDoc2Dial [16]. A detailed description of these datasets can be found in Appendix B.

4.2 Models & Metrics

In this work, we work with the following LLMs: (1) PaLM [4] variants: 1B, 24B, and 340B (2) FLAN-T5-XXL [13]. For evaluating the performance of knowledge-grounded and persona-grounded dialogues generated using different methods, we use the following metrics similar to [16, 54]: BLEURT score [38], METEOR [6], F1 score, BLEU [31], and ROUGE-L [25].

4.3 Results on Multi-Session Chat (MSC) Dataset

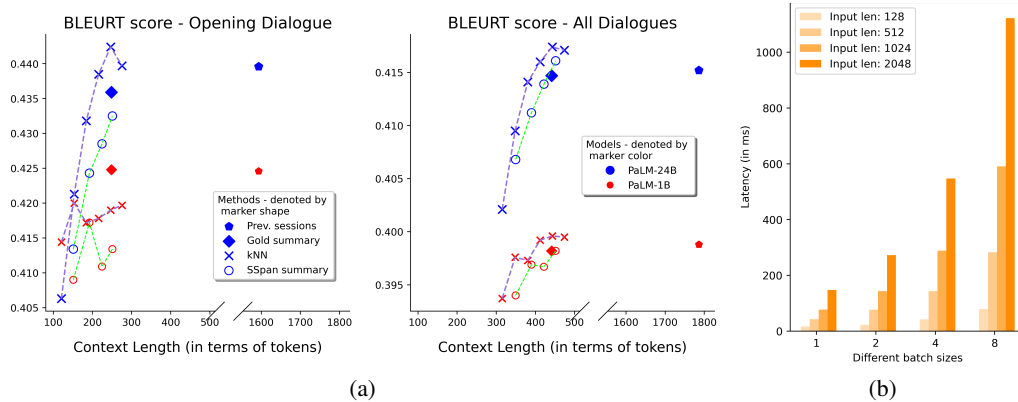


Figure 1: (a): Performance of different methods in terms of BLEURT score on the test set of MSC dataset for predicting the response to different dialogues of session-5 using PaLM-24B and PaLM-1B. Both legends are shared by all figures. For kNN-based augmentation, we report results for $k \in [1, 6]$. The context length grows as k increases. For the submodular span summary-based method, we report results for summary budget $k_2 \in [3, 6]$. (b): Latency (in ms) for different input lengths when using FLAN-T5-XXL (11B) on an NVIDIA A100 GPU.

On the MSC dataset, we compare our proposed retrieval-based methods to following baselines: (1) **Previous sessions**: LLM has access to all previous sessions’ dialogues along with the current session (2) **Gold summary**: LLM generates the next dialog using the human-annotated gold summary of the previous four sessions along with the current session context.

In Fig. 1a, 3a, 3b, and 3c (in Appendix C.4), we compare our proposed kNN and submodular span (SSpan) summary-based retrieval methods for context augmentation to the baselines mentioned above. As pointed out in the MSC [50] paper, the session openings look quite different from other dialogues in the dataset as they require the users to re-engage after a period of time and utilize the information previously exchanged. Therefore, we separately compare different methods on the fifth session opening dialogues.

For PaLM-1B and PaLM-24B, we can see from Fig. 1a that kNN and submodular span summary-based context augmentation for dialog generation perform better than the strong baseline of using

gold summary (human-annotated) in the input context to the LLM across all four metrics. For the session opening dialog, we observe that increasing the retrieval/summarization budget k leads to a considerably larger improvement in all metrics for both PaLM-1B and 24B. This shows that for generating the next response to session-opening dialogues, augmenting the input context with more relevant information provides useful context to the LLM. We also provide some examples showing the generated responses against different input augmentation methods in Table 2 and 3 in Appdx C.2.

The *previous sessions* baseline containing more than 1,600 tokens performs better than our proposed methods only in terms of the F1 score (Fig. 3a) which relies on unigrams overlap. Moreover, due to the quadratic complexity of self-attention, the latency costs increase significantly when processing much longer inputs encompassing dialogues from all previous sessions, as shown in Fig. 1b.

4.4 Results on MultiDoc2Dial Dataset

On the MultiDoc2Dial dataset, we study how varying the retrieval (or summarization) budget i.e., k (or k_2) affects the final performance for knowledge-grounded response generation across three different models: **Flan-T5-XXL**, **PaLM-24B**, and **PaLM-340B**. Before evaluating our context augmentation techniques for response *generation*, we first assess the recall performance of our PaLM 2-based kNN *retrieval* module, which performs on par with other supervised fine-tuned methods in terms of document-level recall (Table 1 in Appendix C.1).

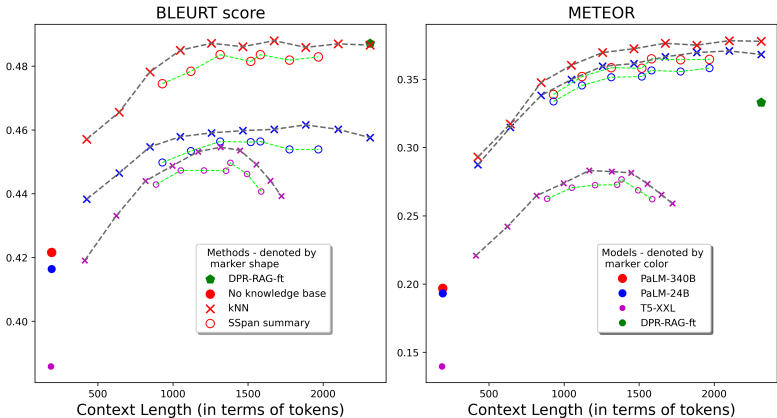


Figure 2: Performance of different methods on MultiDoc2Dial test set for agents’ dialog prediction task using **FLAN-T5-XXL**, **PaLM-24B**, and **PaLM-340B**. Both legends are shared by all figures. For kNN-based augmentation, we report results for $k \in [1, 10]$. The context length grows as k increases. For the Submodular Span (SSpan) summary-based method, we report results for summary budget $k_2 \in [4, 10]$.

In Fig. 2 and 4 (in Appendix C.5), we compare our proposed retrieval and summarization-based input context augmentation methods to two baselines: (1) **No knowledge base**: LLM solely relies on the conversation context with no access to any external knowledge base (2) **DPR-RAG-ft** [16]: DPR-based retrieval augmented generation model fine-tuned on the MultiDoc2Dial dataset. For the two models with a 4,096 context window (**PaLM-24B** and **PaLM-340B**), we observe that beyond a certain k , performance saturates across all recorded metrics, irrespective of increased retrieval or summarization budgets.

When using **FLAN-T5-XXL**, the performance corresponding to the response generation task plateaus at around $k = 6$ and subsequently declines due to its smaller context window size of 2,048. The pre-trained **PaLM-340B** model outperforms the fine-tuned **DPR-RAG-ft** model in all metrics except BLEU-4. This improvement is achieved by incorporating retrieved knowledge chunks into the LLMs’ input context, along with an instructive prompt that guides the PaLM model to utilize the provided knowledge for grounding its response. In Appendix C.3, we provide sample examples where submodular span summarization method (diversity + query relevance) outperforms simple kNN retrieval.

5 Conclusion

In this work, we utilize large language models’ ability to learn from context through instruction prompting for the challenging task of long-range persona-grounded and knowledge-grounded dialog modeling. Instead of fine-tuning, our research explores retrieval-augmented approaches focusing on query relevance and diversity to leverage external knowledge sources as well as previous conversations for the purpose of LLMs’ input context augmentation. We empirically demonstrate the effectiveness of our proposed retrieval and summarization-based context augmentation techniques on two dialog modeling datasets: MSC (persona-grounded) and MultiDoc2Dial (knowledge-grounded).

References

- [1] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.
- [2] Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. Etc: Encoding long and structured inputs in transformers. *arXiv preprint arXiv:2004.08483*, 2020.
- [3] Sotiris Anagnostidis, Dario Pavllo, Luca Biggio, Lorenzo Noci, Aurelien Lucchi, and Thomas Hoffmann. Dynamic context pruning for efficient and interpretable autoregressive transformers. *arXiv preprint arXiv:2305.15805*, 2023.
- [4] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.
- [5] Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yui Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. Keep me updated! memory management in long-term conversations. *arXiv preprint arXiv:2210.08750*, 2022.
- [6] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- [7] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.
- [8] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.
- [9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [10] Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091, 2022.
- [11] Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891, 2022.
- [12] Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

- [13] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- [14] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.
- [15] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.
- [16] Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. Multidoc2dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, 2021.
- [17] Curtis Hawthorne, Andrew Jaegle, Cătălina Cangea, Sebastian Borgeaud, Charlie Nash, Mateusz Malinowski, Sander Dieleman, Oriol Vinyals, Matthew Botvinick, Ian Simon, et al. General-purpose, long-context autoregressive modeling with perceiver ar. In *International Conference on Machine Learning*, pages 8535–8558. PMLR, 2022.
- [18] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.
- [19] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.
- [20] Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*, 2021.
- [21] Lilly Kumari and Jeff Bilmes. Submodular span, with applications to conditional data summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12344–12352, 2021.
- [22] Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. Prompted llms as chatbot modules for long open-domain conversation. *arXiv preprint arXiv:2305.04533*, 2023.
- [23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [24] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*, 2022.
- [25] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [26] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*, 2018.
- [27] Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*, 2021.
- [28] Pedro Henrique Martins, Zita Marinho, and André FT Martins. ∞ -former: Infinite memory transformer-former: Infinite memory transformer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5468–5485, 2022.
- [29] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

- [30] OpenAI. Gpt-4 technical report, 2023.
- [31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [32] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2023.
- [33] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [34] Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.
- [35] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*, 2023.
- [36] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, 2021.
- [37] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68, 2021.
- [38] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.
- [39] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.
- [40] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*, 2022.
- [41] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey, 2022.
- [42] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.
- [43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [45] Shuohang Wang, Luwei Zhou, Zhe Gan, Yen-Chun Chen, Yuwei Fang, Siqi Sun, Yu Cheng, and Jingjing Liu. Cluster-former: Clustering-based sparse transformer for question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3958–3968, 2021.
- [46] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.
- [47] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

- [48] Qingyang Wu, Zhenzhong Lan, Kun Qian, Jing Gu, Alborz Geramifard, and Zhou Yu. Mem-former: A memory-augmented transformer for sequence modeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*, pages 308–318, 2022.
- [49] Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. *arXiv preprint arXiv:2203.08913*, 2022.
- [50] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*, 2021.
- [51] Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory. *arXiv preprint arXiv:2203.05797*, 2022.
- [52] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33: 17283–17297, 2020.
- [53] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, 2018.
- [54] Shiwei Zhang, Yiyang Du, Guanzhong Liu, Zhao Yan, and Yunbo Cao. G4: grounding-guided goal-oriented dialogues generation with multiple documents. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 108–114, 2022.
- [55] Tong Zhang, Yong Liu, Boyang Li, Zhiwei Zeng, Pengwei Wang, Yuan You, Chunyan Miao, and Lizhen Cui. History-aware hierarchical transformer for multi-session open-domain dialogue system. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3395–3407, 2022.
- [56] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, 2020.
- [57] Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H₂O: Heavy-hitter oracle for efficient generative inference of large language models, 2023.

A Related Work

A.1 Transformers and their long-range versions

Prior works on transformers modeling long-range dependencies can be broadly categorized into four groups:

1. Transformers using fixed or learnable attention patterns for sparsification to enable long-range coverage of the input sequence. These include Compressed Transformer [26], Longformer [7], BigBird [52], ETC [2], Cluster-former [45], Routing Transformers [37], Reformer [19], etc. A more detailed and comprehensive overview of these sparse and efficient transformer models is covered in [41].
2. Transformers that store computed hidden states from previous segments and utilize them as memory during the processing of the current segment, potentially utilizing compression techniques. These include Compressive Transformers [34], Transformer-XL [14], ∞ -former [28]
3. Transformers extending the context by applying cross-attention over a significantly wider span at the input layer, such as Perceiver-AR [17].
4. Transformers that utilize an external dynamic memory module to efficiently handle long sequences such as Memformer [48], Recurrent Memory Transformer [10].

A.2 Retrieval Augmented LLMs

This line of research broadens the capabilities of Large Language Models (LLMs) by augmenting their parametric knowledge with an external knowledge base and a retrieval module. The retrieval module fetches relevant context, such as passages or documents, and integrates them into the LLM’s input context. Examples of such retrieval-augmented LLMs include but are not limited to RAG [23], Memorizing Transformers [49], TRIME [55], RETRO [8], ATLAS [18], in-context RALM [35].

A.3 LLMs for Dialog Modeling

In the field of open-domain generation-based dialogue modeling, transformers-based LLMs such as Meena [1], BlenderBot [36], and DialoGPT [56] have made substantial advancements. Chatbots such as LaMDA [42] and BlenderBot3 [40] involve further fine-tuning of pre-trained LLMs on existing dialogue and question-answering datasets along with incorporating external knowledge sources [20] to enhance both model safety and factual grounding.

For dialog modeling over multiple conversation sessions [50], specialized works such as HAHT [55] maintain a long-term memory of past conversations, use it to understand the current context and generate responses by dynamically switching between a general vocabulary and a history-aware vocabulary. PLATO-LTM [51] augments open-domain dialogue models with a plug-and-play long-term memory system, enabling real-time extraction and utilization of persona information to enhance long-term persona conversations. This memory module is also explored in [5] where the memory consists of unstructured text about the user’s key information for generating more consistent and engaged responses. Furthermore, [27, 22] investigate the roles of few-shot in-context learning and chain-of-thought style prompting in the context of open-domain dialog generation.

B Dataset Description

The MSC dataset comprises human-human crowdworker conversations spanning over five sessions, each session containing up to 14 exchanges, where participants re-engage after some time has passed. The dataset also includes crowdsourced gold summaries of the user conversations preceding the fifth session. Both individuals engaged in a conversation are allocated personas from the pool of 1,155 personas provided in [53] and the objective is to model the dialogues of the second user based on the complete conversation history up to that point.

The MultiDoc2Dial dataset is designed for modeling goal-oriented dialogues that rely on information from multiple documents. It includes dialogues grounded in diverse documents from four domains,

with the task being to generate the agent’s next response given the user’s current utterance, dialogue history, and the entire set of documents from all four domains (knowledge base for this dataset is constructed from passages extracted directly from these documents).

C Additional Results

C.1 MultiDoc2Dial: Retrieval Results

Prior to evaluating the performance of our proposed input context augmentation techniques for the response generation task, we first assess the retrieval aspect, specifically the grounding performance of our retrieval module in terms of recall metric. This module utilizes PaLM 2 embeddings to encode both the knowledge chunks/segments and the dialog queries. As shown in Table 1 below, the results demonstrate that our PaLM 2-based kNN retriever achieves performance comparable to that of other supervised fine-tuned methods at document level.

In Table 1, we compare our pre-trained retrieval module to other methods:(1) **DPR-ft** [16], (2) **ANCE-ft** [54], both of which are fine-tuned on the training set of MultiDoc2Dial, hence outperforming the PaLM embeddings based module on the passage retrieval task.

Table 1: Performance of different retrieval methods on MultiDoc2Dial validation dataset

Method	Doc Recall@1	Doc Recall@5	Doc Recall@10	Passage Recall@1	Passage Recall@5	Passage Recall@10
PaLM 2	49.56	77.51	84.43	26.52	56.72	68.01
DPR-ft	49.61	75.43	82.69	38.09	65.70	74.03
ANCE-ft	-	-	-	39.54	68.46	77.27

C.2 Multi-Session Chat (MSC): Sample Responses

<p>Current session context: <i>first user</i> : How are your dogs? Have you seen a bearded dragon in real life yet? <i>second user</i> : My dogs are happy and well! I haven't seen a bearded dragon yet, but I'm looking forward to it! How are your reptiles doing? <i>first user</i> : They are doing great! I cannot wait to get more. But I was also thinking maybe I should start focusing on life other than my reptiles. You know, get out of the house more. I suppose dogs are good for that. <i>second user</i> : I'm glad to hear it! Have you decided that you're definitely getting another? Yes, dogs are good for getting out of the house! I guess you couldn't take your lizard on a walk very easily. <i>first user</i> : I haven't decided, maybe I will go the dog route instead. I was looking at golden retriever puppies the other day, they are soooooo cute! <i>second user</i> : Yes, they are! Puppies are very fun! I recommend getting one if you've got the free time to take care of it. Remember that they're a lot of work though! <i>first user</i> : Yes, it seems like you are almost taking care of a baby with a puppy! Did you get your dogs from a breeder or a shelter? <i>second user</i> :</p>
<p>Gold Summary: Summarized personas of <i>second user</i> : I work as an assistant at a doctor's office. My job has great benefits and paid vacation. I am an assistant for an eye specialist. I find it unnerving to see what kind of diseases can effect you vision. I would like to see a bearded dragon. I like reptiles. I do not want reptiles as a pet. I have a dog. I wouldn't take reptiles out of their cages if I had them as pets. I might be willing to own a bearded dragon. I have dogs. I worry that getting another dog will make mine jealous. If I get another dog it would be a rescue. I adopted my dogs from a friend who couldn't keep them anymore. My dogs are mixed breed Scottish terrier and something else. My dogs are bigger than most terriers.</p>
<p>kNN based context: Summarized personas of <i>second user</i>: I guess am lucky. what do you feed your pet with? wow how big is your pet now? They do get lonely at times, but they mostly get by. They seem to enjoy my company the most, so I think getting another dog might make them jealous if they see me giving them less attention than normal. That is true, that voucher could easily come in handy. I think if I do want to get another dog, it would be a rescue as I would want to adopt a dog that is in need of a good home. Would you be fine adopting a bearded dragon as opposed to raising it from a hatchling? They aren't a rescue from a shelter, but I did adopt them from a friend who couldn't keep them anymore. They were moving to an apartment that didn't allow for pets, and I was willing to take them and give them a good home.</p>
<p>Submodular Span Summary: Summarized personas of <i>second user</i>: I guess am lucky. what do you feed your pet with? I like reptiles, but would not like one as a pet! I do have a dog. That would be a great idea. I wonder how much they cost and how easy they are to get. They do get lonely at times, but they mostly get by. They seem to enjoy my company the most, so I think getting another dog might make them jealous if they see me giving them less attention than normal. They aren't a rescue from a shelter, but I did adopt them from a friend who couldn't keep them anymore. They were moving to an apartment that didn't allow for pets, and I was willing to take them and give them a good home. I know they are a mixed breed, they seem to be a mix of a scottish terrier with another type of dog, as they are much bigger than most terriers I have seen.</p>
<p>Gold Response: I actually adopted my dogs from a friend who could no longer care for them. If you're getting a dog though, I recommend rescuing one from a shelter!</p>
<p>Generated Responses: Gold Summary: I got my dogs from a friend who couldn't keep them anymore. I've heard good things about adopting from shelters, but I've never done it myself. kNN based context: That's true! They are a lot of work, but they're so worth it. I got my dogs from a friend who couldn't keep them anymore. They were moving to an apartment that didn't allow for pets, and I was willing to take them and give them a good home. Submodular Span Summary: I got my dogs from a friend who couldn't keep them anymore. They were moving to an apartment that didn't allow for pets, and I was willing to take them and give them a good home.</p>

Table 2: Responses generated by PaLM-24B corresponding to different context augmentation methods on an example conversation derived from the MSC dataset.

<p>Current session context: <i>first user</i> : I went and saw that Acura, I think I'm going to bite the bullet and get it! <i>second user</i> : That's great! I'd love to see it when you get it. <i>first user</i> : Yes, I can take you out for a ride in it! Or maybe if I am feeling nice that day, you can practice your driving in it haha! <i>second user</i> : That sounds like fun! I'd be happy to as long as we don't have to go on the interstate. How is your landscape art coming along? <i>first user</i> : It is going great, I went up to a lakehouse on the weekend and got some inspiration to draw the nature scene there! <i>second user</i> : That sounds really nice! I wish I was good at painting so that I could do things like that too. <i>first user</i> : You can definitely learn now! Some of it is natural talent but a lot of it is learned skill. I can send you some amazing instructional videos if you would like, <i>second user</i> :</p>
<p>Gold Summary: Summarized personas of <i>second user</i> : I'm in high school. I want to write songs. I am busy with the track team. I am writing one now but haven't finished it as I am having trouble with the melody. I am a runner. I have my mom and dad. I just got my drivers license. I'm scared of interstate driving. I prefer to write love songs as they are easier and it is hard to find a good other topic. I think it would be cool to do something more creative. Making money from music is not my priority. I can appreciate a good looking car. I like how the Acura NSX looks. I don't own my own car. I am able to drive. I don't think minivans are cool. I am a hopeless romantic. I am not good at drawing or painting. I wish I was better at making visual art.</p>
<p>kNN based context: Summarized personas of <i>second user</i>: I do not really do art, I am more interested in writing songs. What is the cool car you found? Sounds amazing. I just got my license and could only dream of owning an Acura NSX. I have little artistic talent. I can draw stick figures. A landscape of stick figures would not be very interesting. I wish I could express myself visually. Landscapes are quite powerful. Plus, wherever you go in the world, there's a subject waiting for you to paint it. Can you draw cartoon style stuff? You could get a job in video games. They always need artists. You might have to draw the same stuff a million times, but it's a job doing art. It's awfully hard to get rich in music, but it is possible to make a living. Once you take a paycheck, you have to do someone else's music, not yours.</p>
<p>Submodular Span Summary: Summarized personas of <i>second user</i>: I would love to write songs, yes. I stay busy with the track team too.. What do you plan on painting? I started one but haven't finished. It's about thinking you love someone, but finding out you didn't really know them well. I'm having trouble with the melody though Yeah I think they're easier. It's hard to find a good other topic. There's a reason most songs are about love! But it would be cool to do something more creative. Do you have any ideas for topics? The NSX is a beautiful car! I would love to have one myself. I have little artistic talent. I can draw stick figures. A landscape of stick figures would not be very interesting. I wish I could express myself visually. Landscapes are quite powerful. Plus, wherever you go in the world, there's a subject waiting for you to paint it.</p>
<p>Gold Response: Yes, that sounds good! I would love to become more creative in the field of visual art.</p>
<p>Generated Responses: Gold Summary: That would be great! I'd love to learn how to paint. kNN based context: That would be great! I'd love to learn more about it. I've always been more interested in music, but I think it would be really cool to be able to paint landscapes. Submodular Span Summary: That would be great! I would love to learn how to paint.</p>

Table 3: Responses generated by PaLM-24B corresponding to different context augmentation methods on an example conversation derived from the MSC dataset.

C.3 MultiDoc2Dial: Sample Responses

On the MultiDoc2Dial dataset, we examine user-agent conversations where submodular span summarization-based method focusing on both diversity and query relevance performs better than the simple kNN retrieval method. Table 4 and 5 show responses generated by PaLM-340B when using kNN retrieval and submodular span summarization for LLMs' context augmentation. In both examples, we set the summary size (k_2) and kNN retrieval budget (k) to five.

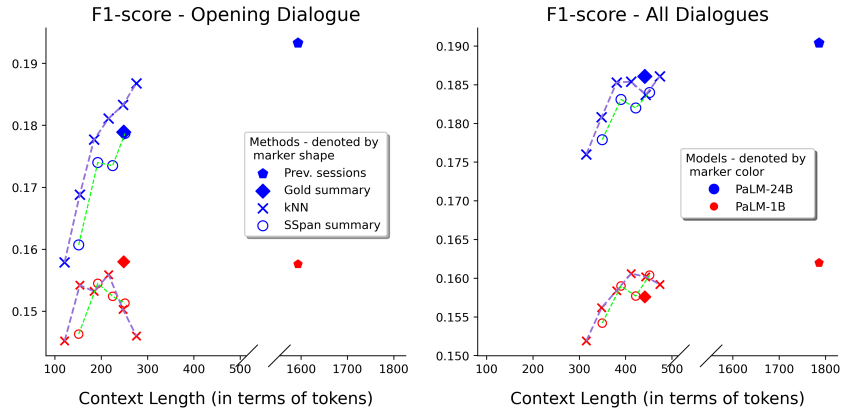
<p>Current conversation context: Here is the conversation history between the user and the agent:</p> <p><i>user:</i> I'm 17 and ready for a driver license <i>agent:</i> Would you like to know more about the Graduated license law? <i>user:</i> nope, I just want a license <i>agent:</i> Did you want to know more about getting a senior driver license? <i>user:</i> yes, vroom vroom let's go <i>agent:</i> so you are 17 years old, correct? <i>user:</i> yes <i>agent:</i> Okay, you are eligible for a senior driver license, Class D or M, if you have a junior driver license or limited junior driver license and have completed a state-approved high school or college driver education course. <i>user:</i> Another question, regarding going to the DMV, should I take any tests? <i>agent:</i> You have to take the vision test or prove that you meet our vision requirement. <i>user:</i> Why do I need a REAL ID? <i>agent:</i> If you plan on traveling you will be required to have it. You are the agent in this conversation. Given the background knowledge and the conversation history between you and the user, respond to the following dialog spoken by the user by matching the style and tone of the agent. <i>user:</i> Is a license required to become a taxi driver? <i>agent:</i></p>
<p>kNN based context: Here are some relevant background knowledge that can be helpful in responding to the user's current conversation: Get a license to drive a taxi or livery vehicle // If you do not have New York State driver license If you do not have a driver license , learn how to apply for your first driver license [5]. Apply for a Class E driver license instead of a Class D driver license. The Class D license and the Class E license both require the same vision, written and road tests. The New York City Taxi and Limousine Commission T&LC regulates taxi drivers and livery drivers in New York City. If you plan to drive a taxi or livery vehicle in New York City, contact the Taxi and Limousine Commission [6]. Starting in October 2020 , you will need an Enhanced or REAL ID to board a domestic flight unless you have a passport. If your license or permit was issued before October 30 , 2017 and is not Enhanced , you have a Standard document. Newly issued Standard license and permits state : ' NOT FOR FEDERAL PURPOSES ' Get a license to drive a taxi or livery vehicle // If you plan on traveling We recommend you come to the DMV now to get an Enhanced or REAL ID, and avoid a second office visit later. 1 Learn about Enhanced, REAL ID and Standard documents [1]. <i>Get a license to drive a taxi or livery vehicle // Step 2: Determine what identification you need to bring to the DMV if you want a REAL ID or Enhanced, complete this document guide to know what proofs of identity to bring to an office Find your proofs [3] On the first page , select : Change my existing NY State license / ID to a REAL ID or Enhanced if you want a Standard that will display " NOT FOR FEDERAL PURPOSES " , bring your current NY State license as proof of identity</i> <i>Get a license to drive a taxi or livery vehicle // Step 3: Go to the DMV surrender your Class D license, apply and pay for your Class E license the fee depends on the expiration date of your current driver license - we cannot calculate your fee until you apply take the vision test or prove that you meet our vision requirements [4]</i> <i>How to renew a license // Renew at an office and get an Enhanced or REAL ID You don't have to get an Enhanced or REAL ID. 4 However , if you have a Standard , you will not be allowed to board a domestic flight starting in October 2020 unless you have a Passport. Complete this guide to make sure you have the proofs of identification and paperwork you need before you come to DMV. You can make a reservation at some DMV locations after completing the guide. You do not need a reservation to visit the DMV. Find your proofs [3] You can take a vision test at the DMV, for no additional charge. You will receive a temporary license to use until your new license arrives by mail within 10 business days.</i></p>
<p>Submodular Span Summary: Here are some relevant background knowledge that can be helpful in responding to the user's current conversation: <i>Get a license to drive a taxi or livery vehicle // Step 2: Determine what identification you need to bring to the DMV if you want a REAL ID or Enhanced, complete this document guide to know what proofs of identity to bring to an office Find your proofs [3] On the first page , select : Change my existing NY State license / ID to a REAL ID or Enhanced if you want a Standard that will display " NOT FOR FEDERAL PURPOSES " , bring your current NY State license as proof of identity</i> <i>Get a license to drive a taxi or livery vehicle // Step 3: Go to the DMV surrender your Class D license, apply and pay for your Class E license the fee depends on the expiration date of your current driver license - we cannot calculate your fee until you apply take the vision test or prove that you meet our vision requirements [4]</i> <i>How to renew a license // Renew at an office and get an Enhanced or REAL ID You don't have to get an Enhanced or REAL ID. 4 However , if you have a Standard , you will not be allowed to board a domestic flight starting in October 2020 unless you have a Passport. Complete this guide to make sure you have the proofs of identification and paperwork you need before you come to DMV. You can make a reservation at some DMV locations after completing the guide. You do not need a reservation to visit the DMV. Find your proofs [3] You can take a vision test at the DMV, for no additional charge. You will receive a temporary license to use until your new license arrives by mail within 10 business days.</i> Get a driver license // Step 1: Determine what license class and type you need Class : most people get a class D - Operator if 18 or older or a DJ - Junior Operator if under 18. See instructions to get a class A , B , C - commercial license CDL [2] class M - motorcycle license [3] class E - taxi or limousine livery license [4] Type : DMV offers 3 types of licenses. Learn about Enhanced, REAL ID and Standard documents [5]. Get a license to drive a taxi or livery vehicle // To drive taxis and other livery vehicles in New York State , you must be 18 years of age or older and you must have a driver license that is one of these classes Class A commercial driver license CDL Class B CDL Class C CDL Class E driver license</p>
<p>Gold Response: Yes, you must have a driver license that is either a Class A commercial driver license CDL, Class B CDL or one of two others.</p>
<p>Generated Responses: kNN based context: Yes, you need a Class E driver license to drive a taxi or livery vehicle. Submodular Span Summary: Yes, you must be 18 years of age or older and you must have a driver license that is one of these classes Class A commercial driver license CDL Class B CDL Class C CDL Class E driver license.</p>

Table 4: Responses generated by PaLM-340B corresponding to different context augmentation methods on an example conversation derived from the MultiDoc2Dial dataset. Passages common to both the kNN-based context and the submodular span summary are italicized.

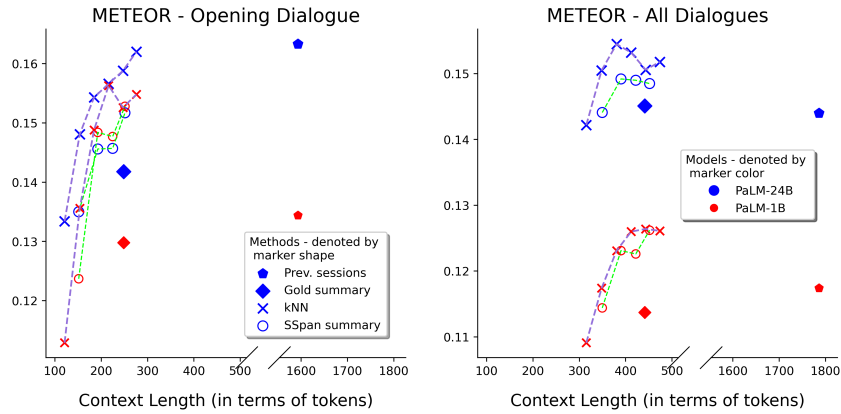
<p>Current conversation context: Here is the conversation history between the user and the agent:</p> <p><i>user:</i> Does everyone who files a claim need to have a VA claim exam? <i>agent:</i> No. We'll ask you to have a claim exam only if we need more information to decide your claim . <i>agent:</i> Do you have enough medical evidence to support your claim? <i>user:</i> Yes I believe so <i>agent:</i> we won't ask you to have a claim exam . Medical evidence may include doctor and hospital reports , test results , and other documents . <i>user:</i> Now I'd like to ask you a random question. How can I qualify for these benefits. Also, can you tell me more about the testing? <i>agent:</i> If you were part of Projects 112 or SHAD from 1962 to 1974 and want to know more about a certain test site, ship, or unit, you can view the declassified Department of Defense fact sheets. We can give you a link to the declassified Department of Defense fact sheets You are the agent in this conversation. Given the background knowledge and the conversation history between you and the user, respond to the following dialog spoken by the user by matching the style and tone of the agent. <i>user:</i> Got it, I also want to know about the VA claim exam (C&P exam), what happens after I take the VA claim exam? <i>agent:</i></p>
<p>kNN based context: Here are some relevant background knowledge that can be helpful in responding to the user's current conversation:</p> <p><i>Exposure Through Project 112 Or Project SHAD Veterans Affairs // How do I get these benefits? You'll need to file a claim for disability compensation. You can call the VA Special Issues Helpline at 800 - 749 - 8387. Or you can file a claim online. Find out how to file a claim for disability compensation If you were part of Projects 112 or SHAD from 1962 to 1974 and want to know more about a certain test site, ship, or unit , see the declassified Department of Defense fact sheets. Get declassified Department of Defense fact sheets If you have a question about the tests , if you have any information that can help show you were part of them including whether you may have been part of them or contact the Department of Defense at 800 - 497 - 6261.</i> <i>VA claim exam (C&P exam) Veterans Affairs // What to expect at your VA claim exam You don't need to bring anything to your exam. If you have any new non - VA medical records like records from a recent surgery or illness - please be sure to submit them before your appointment. The health care provider can't review new information during the exam. If you have children , you'll need to plan ahead for childcare during your exam. Because you and the doctor may discuss sensitive topics , children shouldn't be in the room with you. On the day of the exam , you'll want to wear comfortable clothes so you can move freely while the doctor exams you. A VA claim exam isn't like a normal medical exam or other VA health care visits. The doctor won't treat you for any illness or injury, give you referrals to other health care providers, or prescribe medicine. That's because the purpose of the exam is to gather information that will help us make a decision on your claim. Each exam is different, depending on the information we need. During your exam , the doctor may do any or all of these things : Review your claim file with you Ask you questions based on the medical records in your claim file. These may include questions from the Disability Benefits Questionnaire for each service - connected condition you're claiming. Take a look at the Disability Benefits Questionnaires DBQs Perform a basic physical exam Ask you to get other tests like X - rays or blood work if needed If you have any questions about what's happening during your exam , feel free to ask the doctor. But keep in mind that the doctor's job is to examine you, not to make any decisions about your claim. They won't be able to answer questions about the claims process. After your exam , the doctor will write up a report and send it to the VA claims processor to be added to your claim file. Your exam may be very short , or it may last an hour or more. It all depends on the conditions you've claimed and whether we need more information to make a decision. For example , the required questions the doctor asks about a knee injury may take only a few minutes. But the questions for a more complex illness or injury could take much longer. A mental health evaluation usually lasts 2 to 4 hours. It's important to remember that you in - person visit is only one part of the claim review process. The doctor will also spend time outside the exam carefully reviewing your records. Either a VA doctor or a non - VA doctor who we've contracted with will do the exam. You can also have your own doctor examine you and fill out the necessary forms. Get instructions for your doctor to fill out VA claim exam forms No. The doctor will give the exam results to our VA claims processors. Our claims processors will decide your claim based on these results, plus your other medical and military records. If you have questions about the claims process , call your nearest VA regional benefit office. Find contact information for your nearest VA regional benefit office If you miss your claim exam , contact us right away to let us know why you didn't make it. We can help you figure out what to do next. If you have what we consider to be a good reason for missing your exam called good cause , we'll work with you to reschedule your exam. Examples of good cause might be an illness or the death of someone in your immediate family. Find contact information for your nearest VA regional benefit office</i> <i>VA claim exam (C&P exam) Veterans Affairs // Does everyone who files a claim need to have a VA claim exam? No. We'll ask you to have a claim exam only if we need more information to decide your claim. If you have enough medical evidence in your file to support your claim , we won't ask you to have a claim exam. Medical evidence may include doctor and hospital reports, test results, and other documents.</i> <i>VA claim exam (C&P exam) Veterans Affairs // VA claim exam (C&P exam) After you file your disability benefits claim , we may ask you to have a claim exam also known as a compensation and pension, or C&P, exam. Get answers to commonly asked questions about the VA claim exam below.</i> <i>VA claim exam (C&P exam) Veterans Affairs // Download helpful PDFs VA claim exam fact sheet A quick guide on the claim exam process , what you can expect during and after your exam and to help make the claim exam process easier and less stressful VA claim exam tips Tips VA claim exam FAQs Answers to frequently asked questions about the claim exam process Step - by - step VA claim process a quick, step - by - step overview of the claim process and exam Mental health exam fact sheet Information about how we evaluate mental health claims for compensation , why it's important to go to your exam and Review exams fact sheet Information about why and when we request review exams, what will happen at the exam, and why it's important to go to your exam</i></p>
<p>Submodular Span Summary: Here are some relevant background knowledge that can be helpful in responding to the user's current conversation:</p> <p><i>Exposure Through Project 112 Or Project SHAD Veterans Affairs // How do I get these benefits? You'll need to file a claim for disability compensation. You can call the VA Special Issues Helpline at 800 - 749 - 8387. Or you can file a claim online. Find out how to file a claim for disability compensation If you were part of Projects 112 or SHAD from 1962 to 1974 and want to know more about a certain test site, ship, or unit , see the declassified Department of Defense fact sheets. Get declassified Department of Defense fact sheets If you have a question about the tests , if you have any information that can help show you were part of them including whether you may have been part of them or contact the Department of Defense at 800 - 497 - 6261.</i> <i>VA claim exam (C&P exam) Veterans Affairs // Does everyone who files a claim need to have a VA claim exam? No. We'll ask you to have a claim exam only if we need more information to decide your claim. If you have enough medical evidence in your file to support your claim , we won't ask you to have a claim exam. Medical evidence may include doctor and hospital reports, test results, and other documents.</i> <i>VA claim exam (C&P exam) Veterans Affairs // Download helpful PDFs VA claim exam fact sheet A quick guide on the claim exam process , what you can expect during and after your exam and to help make the claim exam process easier and less stressful VA claim exam tips Tips VA claim exam FAQs Answers to frequently asked questions about the claim exam process Step - by - step VA claim process a quick, step - by - step overview of the claim process and exam Mental health exam fact sheet Information about how we evaluate mental health claims for compensation , why it's important to go to your exam and Review exams fact sheet Information about why and when we request review exams, what will happen at the exam, and why it's important to go to your exam Fully Developed Claim For A VA Pension Veterans Affairs // Standard claim process With the standard claim process , we'll handle the evidence - gathering steps listed below. We'll: that you identify and authorize us to get from a federal facility, like a VA medical center Request relevant records Schedule a health exam or get a medical opinion for you , if we decide we need it for your claim Make every reasonable effort to request relevant records that you authorize us to get that aren't at a federal facility. These may include records from state or local governments, documents from a private doctor, hospital records, or records from current or former employers.</i> <i>VA claim exam (C&P exam) Veterans Affairs // After your VA claim exam We'll review all the evidence in your file, assign your disability rating, and send you a decision notice a letter letting you know your disability rating. Each claim is different , but it usually takes us about 3 to 4 months to process a claim from start to finish. The processing time for your claim depends on how complex your claim is and how many conditions you've claimed. See our most recent estimate for the average number of days to complete a claim Sign in to track your claim Yes. To get a copy of the final report from your exam , you can : Contact your nearest VA regional benefit office , or Call us at 800 - 827 - 1000 and request an appointment to view your file Our VA claims processors will review : you've given us The medical records The report from the doctor who handled your claim exam, and the results of any ordered medical tests Statements from you and others about your claim Your military medical and personnel records We may ask you to have a claim exam if you appeal your disability benefits decision. Learn more about appealing a decision</i></p>
<p>Gold Response: We'll review all the evidence in your file , assign your disability rating , and send you a decision notice a letter letting you know your disability rating .</p>
<p>Generated Responses:</p> <p>kNN based context: After your exam , the doctor will write up a report and send it to a VA claims processor to be added to your claim file. Submodular Span Summary: After your VA claim exam, we will review all the evidence in your file, assign your disability rating, and send you a decision notice a letter letting you know your disability rating.</p>

Table 5: Responses generated by PaLM-340B corresponding to different context augmentation methods on an example conversation derived from the MultiDoc2Dial dataset. Passages common to both the kNN-based context and the submodular span summary are italicized.

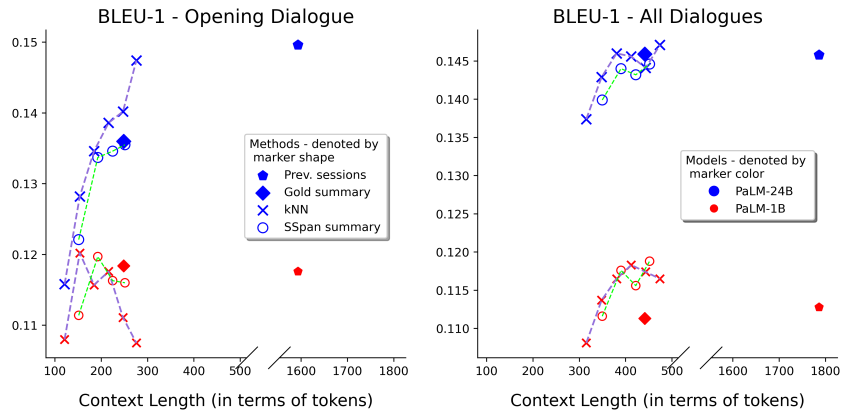
C.4 Multi-Session Chat (MSC): Additional Results



(a) F1 score



(b) METEOR



(c) BLEU-1

Figure 3: Performance of different methods in terms of F1 score, METEOR, and BLEU-1 on the test set of MSC dataset for predicting the response to different dialogues of session-5 using PaLM-24B and PaLM-1B. Both legends are shared by all figures. For kNN-based augmentation, we report results for $k \in [1, 6]$. The context length grows as k increases. For the submodular span summary-based method, we report results for summary budget $k_2 \in [3, 6]$.

C.5 MultiDoc2Dial: Additional Results

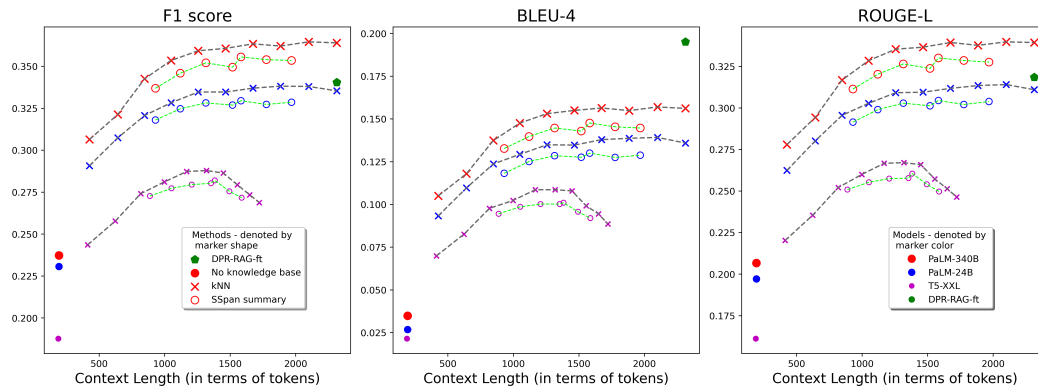


Figure 4: Performance of different methods on MultiDoc2Dial test set for agents’ dialog prediction task using **FLAN-T5-XXL**, **PaLM-24B**, and **PaLM-340B**. Both legends are shared by all figures. For kNN-based augmentation, we report results for $k \in [1, 10]$. The context length grows as k increases. For the Submodular Span (SSpan) summary-based method, we report results for summary budget $k_2 \in [4, 10]$.