Is long context helpful for dialog response generation?

Anonymous ACL submission

Abstract

Personalization has been a key challenge in building engaging conversational agents, necessitating models to effectively utilize long-range context to maintain coherence and consistency over extended interactions. In this work, we investigate the potential of large language models (LLMs) to generate coherent and personalized responses in long-term human-human conversations. We experiment with *fixed context* and *retrieval-based* approaches to use the dialogue history between two speakers. We evaluate our methods and perform analysis on four long-term conversational datasets. Our results indicate that including only a few preceding utterances is generally sufficient for response generation. Retrieval or more extended contexts from past dialogues provide minimal benefits for personalizing model responses. Further analysis of instances that benefited most from retrieval reveals that these cases typically involve either explicit references to previously shared information or scenarios requiring stylistic consistency, such as farewell messages.[1]

1 Introduction

Over the past year or so, the amount of context that can be effectively handled by language models has scaled rapidly from several thousand tokens to hundreds of thousands or even millions. Various approaches have contributed to this increase in practical context length, including approximate or sparse attention methods (Choromanski et al., 2022; Tay et al., 2020; Child et al., 2019), methods that encode many shorter chunks at a time (Liu et al., 2023a; Bertsch et al., 2023), and retrieval-based methods that sub-select context (Lewis et al., 2020; Jiang et al., 2023; Yao et al., 2023).

These long-context transformers are evaluated on a variety of benchmarks. (Tay et al., 2020) introduced the Long Range Arena (LRA) benchmark, which, although it includes six tasks across various modalities with sequences ranging from 1K to 16K tokens, has practical limitations due to its artificial elongation of sequences and limited use of natural language text. (Liu et al., 2023b) employ multi-document question answering and key-value retrieval to assess performance; however, this dataset is primarily suited for analytical purposes. Other benchmarks, such as those in SCROLLS (Shaham et al., 2022), offer a more realistic evaluation of language models on longer contexts but are constrained to specific use cases like summarization and question answering.

While these evaluation methods test various abilities of language models over longer sequences, they fall short in measuring a crucial use case: personalization. Personalization is particularly important when an LM functions as a chatbot, engaging in extended conversations with humans. These long-term interactions present a challenging problem because i) LMs often struggle with managing longer contexts (Liu et al., 2023b), ii) human-to-human conversations are inherently complex, involving external references, digressions, and the constant introduction of new information (Wei et al., 2024), and iii) maintaining a consistent persona, tone, and pragmatic understanding over time is difficult for current LMs. Existing research proposes achieving personalization by either fine-tuning language models on personalized data or using retrieval augmentation (Salemi et al., 2024). Retrieval augmentation involves retrieving relevant personal items (e.g., past writings, preferences) from a user's profile and incorporating them into the language model's input prompt to enable personalized generation.

In this work, we curate a task specifically designed to evaluate how well LMs can handle long human-to-human conversations. These extended dialogues provide a testbed for evaluating personalization, as replying consistently requires LMs to

[1] Code and Data are available at https://anonymous.4open.science/r/Long_Context_Dialog-0431/

enact a user's speaking style, remember past interactions, and adapt to evolving discourse. We evaluate these models on datasets containing long conversations over multiple sessions between two speakers. The overall task is to use the conversation history to generate the best response to the speaker's most recent utterance. Our work examines the following set of research questions about LM's abilities in modeling long conversations:

- Q1 Does providing more previous utterances between two speakers benefit dialogue generation? (section 3)
- Q2 Is it helpful to retrieve past utterances relevant to the current utterance and then use them to make predictions? (section 4)
- Q3 For which conversations is having a retrieval augmented LM the most useful? (section 5)

Surprisingly, our findings suggest that retrieval or a longer context window over the conversation history provides only limited improvement in the personalization of dialogue generation. These improvements are mostly seen in instances that either require a direct reference, such as a fact or proper noun shared in the past, or in utterances that demand some stylistic consistency, such as farewell messages. This calls into question the effectiveness of using long contexts or retrieval-based methods for dialogue personalization and underscores the importance of domain-specific considerations. Identifying when to make direct references and understanding the nuances of conversational contexts are crucial for improving the performance of language models in extended human-to-human interactions.

2 Experimental Setup

2.1 Datasets

We evaluate LMs using conversation data that consists of alternating utterances between two speakers. For each conversation, we test the model on each dialogue of the second speaker. Hence, the task is to predict the response to the most recent turn by the first speaker (the *query* utterance), utilizing the context provided by past utterances. Many of the datasets we considered were curated in a multi-session format, where humans (or language models acting as humans) engaged in conversations across multiple sessions. In these scenarios, participants were asked to continue the conversation from where it was left off in the previous session, sometimes after a significant time gap. This setup helps in evaluating the model's ability to maintain context and coherence over extended periods and across multiple interactions. There are very few well-annotated datasets explicitly curated to test the personalization of model responses over long conversations. Therefore, we identified four datasets with long conversational context, some originally intended for other purposes, and curate them into a format conducive for testing personalization (statistics shown in Table 1):

Natural Customer Service (NatCS) (Gung et al., 2023) This dataset contains natural customer support conversations, designed to train AI models for customer service applications. It was created by carefully eliciting dialogues from crowd workers to mimic real-world customer support scenarios.

Beyond Goldfish Memory (BGM) (Xu et al., 2021) This dataset features human-human conversations across multiple chat sessions, where speaking partners learn about each other over time. It was designed to facilitate research on open-domain conversation models that can maintain consistent personas and memories across sessions.

Switchboard (SB) (Godfrey et al., 1992) This dataset includes approximately 2,400 two-sided telephone conversations among 543 speakers from various regions of the United States. The conversations cover about 70 topics, with each speaker participating only once per topic. We used a cleaned version of the dataset,[2] addressing transcription errors present in the original version.

Locomo (Maharana et al., 2024) This dataset introduces a machine-human pipeline to generate high-quality, long-term dialogues using LLM-based agent architectures, grounding the dialogues on personas and temporal event graphs. Human annotators verify and edit the generated conversations for long-range consistency and alignment with the event graphs. Although the Locomo dataset includes images, we evaluate in a text-only setting where the images are replaced by their captions.

[2]https://convokit.cornell.edu/documentation/switchboard.html

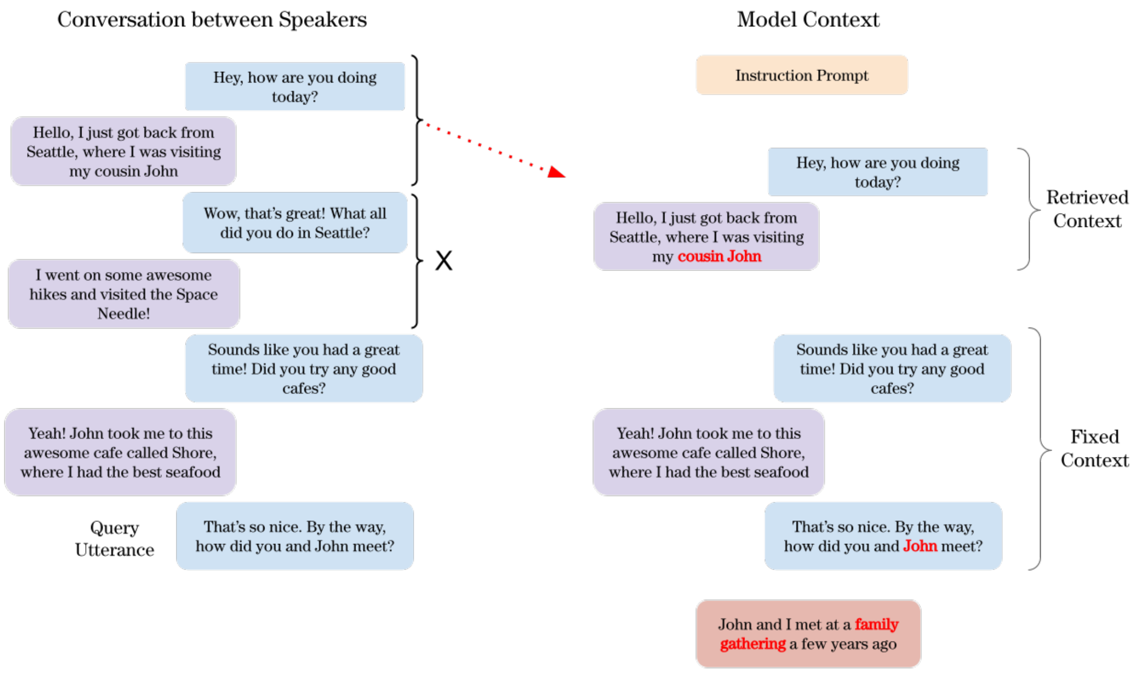

Figure 1: The diagram shows the retrieval-based approach to select the model context for a toy example. As can be seen, we use a fixed and retrieval-based context window of size 2 utterances and retrieve on the basis of query utterance.

Table 1: Dataset statistics

Dataset	NatCS	BGM	SB	Locomo
Conversations	930	1000	1150	35
Avg. utterances/conversation	70.5	60.8	37.9	421.1
Avg. words/utterance	9.8	22.2	32.2	22.9
Avg. tokens/utterance	14.7	30.7	46.5	33.8
Avg. Sessions	1	4	1	19
Avg. utterances/Session	36	14	38	15

2.2 Modeling Methods

Base Model We perform most of our experimentation with Vicuna-7b-1.1 (Chiang et al., 2023) as the base LM. This model is a fine-tuned version of LLaMA (Touvron et al., 2023) on user-shared conversations collected from ShareGPT, and has demonstrated performance on par with other LLMs on the Alpaca benchmark (Li et al., 2023), making it a reasonable candidate for modeling human-to-human natural dialogues.

Context Modeling Because the conversational context is long, it is necessary to use methods to handle these long contexts; we experiment with two main approaches. The first is a recency-based *fixed context* approach, where we use only the utterances immediately preceding the query utterance as the model context. The second is a *retrieval-based* approach (Figure 1), where we include some retrieved utterances as well as the most immediate prior utterances. Details of each approaches are discussed in §3 and §4, respectively.

2.3 Evaluation Metrics

In order to perform uniform evaluation over heterogeneous datasets, we evaluate the performance of various methods by comparing the similarity of the predicted response with that of the ground truth user dialogue using a diverse set of metrics.

BERTScore (Zhang et al., 2020) measures the semantic similarity between texts based on BERT embedding. We also employ its variant, BERTScore-idf, which incorporates inverse document frequency (idf) weights to emphasize rare words.

Character n-gram F-score (chrF) (Popović, 2015) measures the overlap between the generated and reference texts at the character level, capturing fluency and subtle differences in word forms.

In conversational dialogues, BERTScore and chrF complement each other by evaluating the intended meaning and the expression of the speaker, respectively. These metrics are crucial for capturing a speaker's personality. We observe that BERTScore and chrF generally follow similar trends, so we primarily report results using the BERTScore metric.

3 Effect of Longer Context Windows

Methodology Given that natural dialogues maintain continuity in their content, the utterances that immediately precede the query utterance are highly relevant for generating an appropriate response. Thus, a natural way to utilize dialog context is to append the most recent utterances to the LM's prompt. In this section, we examine providing the language model with increasing numbers of previous utterances as context and evaluate the effect on accuracy. Due to the limitations in context window size (e.g., 2048 tokens for Vicuna-7B), it is often not possible to include more than a certain number of utterances, and we truncate the context in such cases, fitting as many recent utterances as possible.

We experiment with context lengths of $n \in \{1, 2, 4, 8, 16, 32\}$, with these counts including the query utterance. For a context length of 1, we provide only the query utterance without any past context. Near the beginning of the conversation, if there are fewer than n total utterances, we provide all available utterances. We also assess the dialogue generation capabilities of advanced models, including Llama-2-7B-32K-Instruct[3] (TogetherAI) and GPT-3.5-turbo[4] (OpenAI), to determine their performance relative to our baseline Vicuna model.

Results Table 2 shows the BERTScore performance of fixed context models with increasing context lengths on NatCS, Switchboard, and Locomo datasets. The numbers are also visualized in the plot Figure 2.

As observed from the table and graph, longer context does improve model performance. However, the performance gain diminishes as the context window size increases, leading to saturation in the BERTScore. This indicates that while a short context preceding the current utterance is beneficial in almost all instances, a longer context provides additional help only in a few instances.

Table 3 compares the performance of the Vicuna model with the Llama and GPT models. The results indicate that the Vicuna models slightly outperform the others, likely due to their explicit fine-tuning for chatbot settings. We, therefore, use the vicuna model for most of our experimentation.

	Locomo Dataset			BGM Dataset	
Context Length	Vicuna-7B	Llama-7B-32K	Context Length	Vicuna-7B	GPT-3.5-turbo
8	0.593	0.587	2	0.535	0.534
16	0.597	0.598	4	0.546	0.537
32	0.598	0.593	8	0.549	0.541

Table 3: The tables present the BertScore results of the fixed context approach with Vicuna-7B, Llama-7B-32K, and GPT-3.5-turbo on the Locomo and BGM datasets

[3]https://www.together.ai/blog/llama-2-7b-32k-instruct
[4]https://platform.openai.com/docs/models/gpt-3-5-turbo

4 Moving beyond Fixed Context

Consider the example shown in 1, where two speakers discuss a recent trip to Seattle. Suppose one speaker mentions visiting their cousin John and trying out some great cafes. Later in the conversation, they might refer back to these details. If these references are far back in the conversation, fixed context models might not capture this useful information when responding to a current query. In such cases, a retrieval model that can fetch relevant past conversations can be very useful. For instance, in the current dialogue, if one speaker asks, "How did you and John meet?", a retrieval model can fetch the earlier mention of John, aiding in producing a more coherent and contextually appropriate response.

4.1 Retrieval-based Context Augmentation

Methodology Based on this idea, we develop a retrieval-based approach to leveraging previous dialog context. Intuitively, the most recent context (examined in section 3) will remain important in most cases, so we develop a method that uses a hybrid of (1) most recent utterances and (2) retrieved utterances based on their predicted relevance to the query utterance. Figure 1 illustrates this approach with a toy example.

The retrieved utterances are selected according to the vector similarity between the query utterance and the retrieved utterance. For most experiments we use the BM25 (Robertson and Zaragoza, 2009) retrieval method, which is a simple lexical-based metric that uses term frequencies and inverse term frequencies to rank documents. Despite its simplicity, BM25 is a strong baseline and robust to a variety of settings.

We then select the top k most relevant utterances. These retrieved utterances, along with the fixed context utterances, are passed to the model as context. We always keep utterances in pairs to preserve the flow of the dialogue. To maintain coherence, we also ensure the order of these utterances remains as they appear in the conversation. The algorithm 1 summarizes the steps we follow for each instance of generating user utterances. To compare this approach with the fixed context models, we split the context in half between the fixed and retrieved utterances; for example, if we have a context window of 16 utterances, we keep 8 preceding utterances and retrieve 8 utterances from the past.

Dataset	1	2	4	8	16	32
NatCS	0.4785	0.5068	0.5270	0.5422	0.5486	0.5523
BGM	0.5223	0.5351	0.5462	0.54887	0.5443	0.5447
Switchboard	0.4549	0.4641	0.4797	0.4855	0.4895	0.4919
Locomo	0.5767	0.5864	0.5906	0.5931	0.5969	0.5981

Table 2: BERTScore values for different datasets with varying context lengths

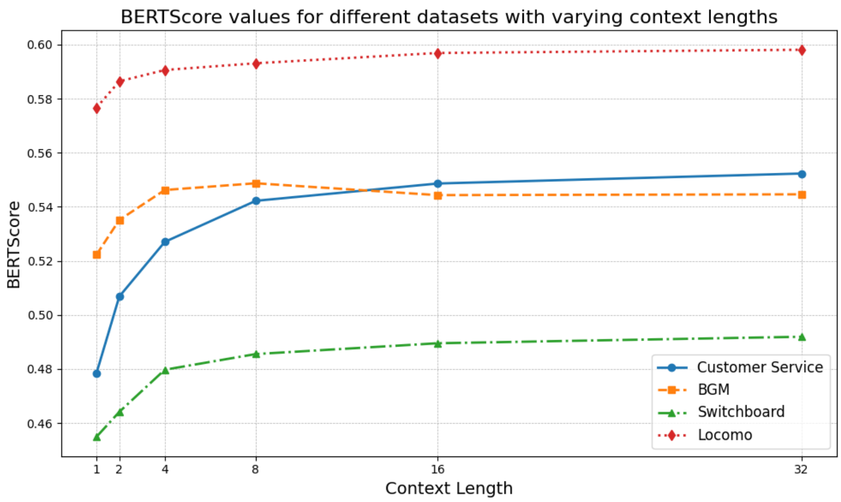

Figure 2: The above plot presents the variation of the BertScore metric with increasing context lengths and different datasets

Results Table 4 presents the comparison of the retrieval augmented approach with the fixed context approach across various datasets.

From this table, we can see that there is a negligible difference between fixed context and retrieval-based methods. This raises questions about the necessity of retrieval for improving the performance of conversational language models.

4.2 Oracle Retrieval

To assess the potential upper bound of retrieval-based approaches, we employed an oracle-based method. This method considers not only the query utterance but also the ground truth response for retrieval. By using both the query utterance and the ground truth response as the query document, we aim to fetch the most relevant past utterances. BM25, being a sparse retrieval method, relies on term frequencies to determine relevance. To complement this, we also evaluated oracle methods using dense retrieval models that operate in the embedding space.

In our first experiment, we utilized sentence embedding models to encode semantic information into dense vectors. Specifically, we used the SFR-Embedding-Mistral model (Rui Meng, 2024), noted for its top performance on the MTEB (Muennighoff et al., 2023) benchmark. Adopting a bi-encoder retriever framework, we encoded both the query utterances and past candidate utterances with the same encoder, ensuring consistent representation. We then ranked past utterances based on two distance metrics: norm distance and cosine similarity.

Method	Context Length	
	8	16
NatCS		
- Baseline	0.5422	0.5486
- BM25 Retrieval	0.5418	0.5481
Switchboard		
- Baseline	0.4855	0.4895
- BM25 Retrieval	0.4862	0.4897
Locomo		
- Baseline	0.5928	0.5960
- BM25 Retrieval	0.5957	0.5981

Table 4: Comparison of Baseline and Retrieval-based Methods for Multiple Datasets

Our second experiment employed the retrieval method proposed by (Fernandes et al., 2023), which ranks documents based on the probability of generating the second speaker's response. Utterances leading to higher probabilities of the ideal

Algorithm 1 Retrieval Augmented Generation for Long Conversations

Ensure: Utterances $u_1, a_1, u_2, a_2, \ldots, u_n, a_n, u_{n+1}$, $u_i's$ and $a_i's$ are first and second speaker utterances respectively. The LM's task is to predict the turns of the second speaker - a_{n+1}

Require: Derive contextually rich k utterances that help in predicting a_{n+1}

1: $query \leftarrow u_{n+1}$
2: Define fixed context with last p-pair utterances: $C \leftarrow (u_{n-p+1}, a_{n-p+1}), \ldots, (u_n, a_n)$
3: Create $n - p$ pairs as documents to retrieve from: $D \leftarrow \{(u_1, a_1), (u_2, a_2), \ldots, (u_p, a_p)\}$
4: Select k-p top-scored documents from D based on the retrieval strategy
5: Use the selected documents in the order they appear in the conversation, followed by the fixed context model and the query as the full prompt to the model.

response were ranked higher. Due to the computational complexity of this approach, we tested it on a randomly selected subset of the dataset.

Comparing the performance of oracle retrieval with fixed context models provides insights into the efficacy of retrieval-based approaches. If oracle retrieval significantly outperforms fixed context models, it indicates potential for improving retrieval methods. Conversely, if improvements are marginal, the focus should shift to enhancing the language model's ability to utilize recent context more effectively.

Table 5 compares the fixed context model with various retrieval methods, including BM25, MTEB embedding-based retrieval, and CXMI retrieval, across different context lengths for the Locomo dataset. The results show that oracle retrieval methods, which use the ground truth label to fetch optimal past responses, offer only slight improvements over the fixed context model. For instance, with a context length of 8, the fixed context model achieves a score of 0.593, while the oracle BM25 retrieval method achieves 0.598, a marginal improvement of 0.005. Similarly, with a context length of 16, the fixed context model scores 0.597, compared to 0.602 for the Oracle BM25 retrieval method, an improvement of 0.005. Interestingly, the MTEB embedding-based retrieval and CXMI retrieval methods perform slightly worse than the fixed context model, even with oracle retrieval. This suggests that these retrieval methods may not be as effective as BM25 for this dataset.

The slight improvements observed suggest that while retrieval can offer some benefits, the primary focus should remain on improving the language model's ability to leverage recent context effectively.

5 Conversation Instances where having retrieval helps the most

Since we observed a very small difference in the average BERTScore between the retrieval-based and fixed context approaches, even with the oracle method, it raises the question of whether the observed difference is due to mere randomness or if retrieval actually improves model performance. To evaluate this hypothesis, we conducted a one-sided t-test on the Locomo dataset with a context window size of 32. The null hypothesis is that retrieval-based methods perform equally or worse compared to the fixed context approach, while the alternative hypothesis posits that retrieval methods perform better. We treated each instance of the data as an individual hypothesis test, generating four instances of responses for each method: fixed context, fixed retrieval, and oracle retrieval. We aimed to determine if each group's BERTScore distribution differs significantly. We consider three different alternative hypotheses:

- Fixed < Retrieval: The performance of the normal retrieval method is better than the fixed context approach.
- Fixed < Oracle: The performance of the Oracle retrieval method is better than the fixed context approach
- Retrieval < Oracle: The performance of the Oracle retrieval method is better than the normal retrieval method

Table 6 presents the percentage of instances rejecting the null hypothesis at a 5% significance level for each alternative hypothesis and evaluation metric. The results indicate that only a small percentage of the dataset shows statistically significant improvements when using retrieval methods compared to the fixed context approach. Based on these results, we conclude that retrieval is not always necessary for improving the performance of conversational language models. The fact that less than 10% of the dataset demonstrates statistically

Table 5: Comparison of Retrieval Results Across Different Context Lengths for Locomo Dataset

Context Length	Fixed Context	BM25 Retrieval		MTEB Embedding Retrieval	CXMI Retrieval
		Normal	Oracle	Oracle	Oracle
8	0.593	0.596	0.598	0.588	0.588
16	0.597	0.598	0.602	0.590	0.590
32	0.598	0.600	0.602	0.594	0.593

Table 6: Percentage of the samples that reject the null hypothesis with the 5% significance level

Metric	**Fixed < Retrieval**	**Fixed < Oracle**	**Retrieval < Oracle**
BertScore	8.3%	9.25%	7.2%
Chrf	8.3%	8.93%	7.15%

significant improvements with retrieval methods suggests that, for a large portion of the data, there is no significant difference between the fixed context approach and retrieval-based methods. This reinforces the idea that the most recent context may be sufficient for many conversational scenarios, and the potential gains from retrieval methods may be limited.

5.1 Locomo Select

To further investigate instances where retrieval methods prove beneficial, we selected the 8.3% of instances from the Locomo dataset that rejected the null hypothesis for the "Fixed < Retrieval" hypothesis, creating a subset called the "Locomo Select" dataset. Upon analyzing the conversations in the "Locomo Select" dataset, we observed that retrieval models are particularly necessary when current utterances directly reference specific information from the past, such as proper nouns or facts mentioned earlier in the conversation. In these cases, retrieving relevant context from the conversation history allows the model to generate more accurate and contextually appropriate responses. Additionally, a substantial portion of the selected examples involved maintaining stylistic consistency, such as farewell messages. This is likely because many people use similar farewell phrases, which the model can fetch from previous sessions, enhancing response coherence. Appendix section B illustrates two instances where the retrieval model performs better than the fixed context model. Table 7 presents the BERTScore-idf results for the fixed context and oracle retrieval models across different datasets and context lengths. The performance gap between the fixed context and oracle retrieval models is more pronounced in the Locomo Select dataset. This increased difference is due to BERTScore-idf's higher weighting of rare words, and most instances in the Locomo Select dataset involve either repetitive stylistic messages or factual information containing proper nouns.

6 Related Work

Long Context Models Various approaches have been explored to develop language models capable of handling long input sequences. (Liu et al., 2023a) propose chunking the input context into blocks and performing self-attention on these individual blocks, then passing key-value pairs in a ring-like fashion to produce the final output. Similarly, Bertsch et al. (2023) suggests chunking the input during the encoding phase, followed by the decoder heads performing a KNN search over the encoder output of each chunk. (Fu et al., 2024) employ a data engineering approach that involves continual pretraining on a balanced mix of domains with 1-5 billion tokens and length upsampling, enabling LMs to handle contexts up to 128K tokens long effectively. Additionally, retrieval-based methods have been investigated to provide relevant context to the model. (Shi et al., 2023) introduces a framework that augments an LM with a tunable retrieval model by prepending retrieved documents to the input. (Jiang et al., 2023) propose iteratively using the prediction of the upcoming sentence to retrieve relevant documents and regenerate low-confidence tokens.

Benchmarks & Evaluation Recent advancements in long-context language models have necessitated the development of specialized bench-

Table 7: Comparison of Fixed Context and Oracle Results Across Datasets using bertscore-idf

Context	Switchboard Corpus		Locomo		Locomo Select	
	Fixed Context	Oracle	Fixed Context	Oracle	Fixed Context	Oracle
8	0.4218	0.4240	0.5388	0.5453	0.5436	0.5514
16	0.4251	0.4254	0.5427	0.5471	0.5457	0.5548
32	0.4358	0.4260	0.5444	0.5477	0.5453	0.5585

marks to evaluate their performance. The Long Range Arena (LRA) benchmark introduced by (Tay et al., 2020) assesses models on tasks requiring long-context understanding across multiple modalities, including text, images, and math. However, LRA's reliance on artificially elongated sequences limits its practical applicability. Alternatively, the SCROLLS benchmark by (Shaham et al., 2022) provides naturally long sequences from diverse domains, requiring models to synthesize dispersed information. Another significant benchmark is the work by (Liu et al., 2023b), which investigates the degradation of model performance when critical information is located in the middle of a long sequence, highlighting challenges in long-context comprehension. These benchmarks collectively aim to test and improve the ability of models to handle extended contexts effectively.

Personalization and Dialogue Systems Personalization in dialogue systems focuses on generating responses that align with individual users' preferences, traits, and backgrounds. (Kasahara et al., 2022) introduce a method using prompt-tuning on large pre-trained language models to create personalized dialogue systems capable of producing natural, persona-consistent responses. Their evaluations indicate that prompt-tuned models outperform fine-tuned models in maintaining consistent personas. Work by (Li et al., 2021) presents a personalized hybrid matching network (PHMN) that leverages user-specific dialogue history to enhance response selection by extracting personalized wording behaviors and employing a customized attention mechanism for improved context-response interaction. (Wang et al., 2023) propose an unsupervised approach to infuse personality traits into large language models through personalized lexicons derived from unlabeled data. These studies highlight various strategies for achieving personalization in dialogue systems, contributing to more coherent and engaging conversational agents.

7 Conclusion

In this work, we investigated the ability of large language models to utilize long-range conversational contexts for generating coherent and personalized responses. Through experiments on multiple datasets, we found that expanding the context window improves response quality, but most of the benefit comes from the immediately preceding utterances, with diminishing returns from more distant contexts. We also explored various retrieval-based methods to select relevant contexts from the conversation history. Surprisingly, even oracle retrieval using the ground-truth response provided only minor gains over a fixed window of recent context. The benefits of retrieval were largely confined to specific scenarios, such as repeating factual information or names mentioned earlier in the conversation or maintaining stylistic consistency for certain utterance types like greetings and farewells.

In conclusion, our study found that larger context windows and retrieval approaches provide limited benefits in modeling long-term conversations. This underscores the need for further research into conversational agent models and methods that can enhance personalization. While BERTScore and chrF metrics were used to evaluate personalization in this study, more refined metrics are necessary to better capture the nuances of personalization in extended dialogues. Future work should focus on developing these metrics and improving models to effectively handle long-range dependencies and user-specific contexts in conversations.

8 Limitations

While our study provides valuable insights into the effectiveness of long-range context and retrieval methods for personalized dialogue generation, it is important to acknowledge its limitations. First, although we experimented with context lengths up to 32 utterances, it would be beneficial to explore the performance of models capable of handling even longer contexts, such as the entire conversation history. Such experiments could shed light on the potential benefits and challenges of utilizing the full conversational context for personalized response generation. Second, the datasets used in this study, although diverse, may not capture all the nuances and challenges of real-world conversational scenarios. Future work should explore the generalizability of our conclusions across a broader spectrum of conversational domains and styles. Finally, the evaluation metrics employed, such as BERTScore and chrF, while widely used and informative, may not fully capture all aspects of personalization and coherence in generated responses. Developing more fine-grained and targeted metrics for assessing personalization could provide a more comprehensive understanding of the strengths and limitations of different context modeling approaches.

9 Ethical Considerations

The use of personalized language models raises significant ethical concerns regarding the potential for effectively impersonating individuals. As these models become adept at mimicking personal traits and conversational styles, there is a heightened risk of misuse in generating misleading or harmful content that appears to originate from real individuals. Ensuring robust safeguards and ethical guidelines are paramount to prevent the exploitation of such technology for identity theft, fraud, or other malicious activities.

References

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R. Gormley. 2023. Unlimiformer: Long-range transformers with unlimited length input. *Preprint*, arXiv:2305.01625.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *Preprint*, arXiv:1904.10509.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. 2022. Rethinking attention with performers. *Preprint*, arXiv:2009.14794.

Patrick Fernandes, Kayo Yin, Emmy Liu, André F. T. Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. *Preprint*, arXiv:2109.07446.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data engineering for scaling language models to 128k context. *Preprint*, arXiv:2402.10171.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA. IEEE Computer Society.

James Gung, Emily Moeng, Wesley Rose, Arshit Gupta, Yi Zhang, and Saab Mansour. 2023. Natcs: Eliciting natural customer support dialogues. *Preprint*, arXiv:2305.03007.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. *Preprint*, arXiv:2305.06983.

Tomohito Kasahara, Daisuke Kawahara, Nguyen Tung, Shengzhe Li, Kenta Shinzato, and Toshinori Sato. 2022. Building a personalized dialogue system with prompt-tuning. *Preprint*, arXiv:2206.05399.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Juntao Li, Chang Liu, Chongyang Tao, Zhangming Chan, Dongyan Zhao, Min Zhang, and Rui Yan. 2021. Dialogue history matters! personalized response selectionin multi-turn retrieval-based chatbots. *Preprint*, arXiv:2103.09534.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023a. Ring attention with blockwise transformers for near-infinite context. *Preprint*, arXiv:2310.01889.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *Preprint*, arXiv:2402.17753.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Shafiq Rayhan Joty Caiming Xiong Yingbo Zhou Semih Yavuz Rui Meng, Ye Liu. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. *Preprint*, arXiv:2304.11406.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. Scrolls: Standardized comparison over long language sequences. *Preprint*, arXiv:2201.03533.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *Preprint*, arXiv:2301.12652.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020. Long range arena: A benchmark for efficient transformers. *Preprint*, arXiv:2011.04006.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jian Wang, Yi Cheng, Dongding Lin, Chak Tou Leong, and Wenjie Li. 2023. Target-oriented proactive dialogue systems with personalization: Problem formulation and dataset curation. *Preprint*, arXiv:2310.07397.

Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging large language models to power chatbots for collecting user self-reported data. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *Preprint*, arXiv:2107.07567.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Example Retrieval Prompt

```
SYSTEM: You are a chatbot tasked with making small-talk with people.

SYSTEM: The below utterances are retrieved from the past conversation.

USER: He's found his happy place! Reminds me I need to too.

ASSISTANT: Hey Jeff, finding our happy places is key. What makes you feel relaxed and joyful?

USER: Drumming, walking my dogs and nature give me joy and chillax.

ASSISTANT: That's an awesome place to chill! What's your favorite thing about it?

SYSTEM: The below utterances are from the current conversation between the same two speakers.

USER: I love sitting on a bench, taking in nature and hearing the creek flow. It's so peaceful!

ASSISTANT: That looks so peaceful! You go there often? The creek sounds so calming.

USER: I try to walk at least once a month. But I would like to do it more often.

ASSISTANT: Nature calms me and reminds me of my beloved pets my dog and horse which bring me so much joy. Truly a blessing!

USER: Yeah, Megan, nature is amazing. It brings us so much joy! It's a great gift.

ASSISTANT:
```

B Instances where Retrieval is Necessary

```
--- Past Relevant Instances --
USER: Wow, that's impressive! How did you get into watercolor painting?
ASSISTANT: My friend got me into it and gave me some advice, and I was hooked right away!
USER: Wow! I hope I can find something I'm as passionate about as you are with watercolor painting.

. . .

--- Current Utterance to be Answered--
USER: You're already good at drawing, why did you decide to join a class? And why did you start painting then years ago anyway?

LABEL: Thanks, Sam! It all started when a friend of mine gave me this painting one day, it inspired me a lot and that's when I started painting. And I joined the classes to find like-minded people and show them what I can do, you can always improve your skills.
```

```
RETRIEVAL MODEL RESPONSE: I took a class because I wanted to improve my skills and
    ↪ learn new techniques. And I started painting years ago because my
    friend introduced me to it and it became a hobby that I enjoy. It's a great way
    ↪  to express my creativity and find peace.

FIXED CONTEXT MODEL RESPONSE: I've always enjoyed drawing, but I never really had
    ↪ the chance to explore it as much as I wanted to. So, I decided to join the
    ↪ class to learn new techniques and improve my skills. As for painting, I
    ↪ always found it to be a form of self-expression. It's a way to relax, unwind
    ↪  and express myself in a different way. And honestly, I find it to be
    ↪ therapeutic.
```

```
--- Past Relevant Instances --
USER: Hey Ned! Nice to see you. Things are going great. Yesterday we went to the
    ↪ theater. It was awesome!
ASSISTANT: Sounds fun! What show? Why'd you go?
USER: It was a "The Importance of Being
    Earnest" - my daughter's Anna friend was in it.

. . .

--- Current Utterance to be Answered--
USER: Yeah, get it. Work and family demand a lot of time. That connection to
    ↪ something you love is great, even if you can't take part as much anymore.
    ↪ Plus, seeing our kids developing their own passions is awesome - like your
    ↪ son with his school play! Which show did you go to?

LABEL: Watching my kids do their thing brings me so much joy! It was "The
    Importance of Being Earnest".

RETRIEVAL MODEL RESPONSE: I went to see "The Importance of Being
    Earnest" - it was so witty and entertaining. I loved it!

FIXED CONTEXT MODEL RESPONSE: I went to a musical called "West Side Story" and it
    ↪ was incredible! The music, choreography, and performances were so powerful.
    ↪ I had goosebumps the entire time.
```

C Ablation Studies for Retrieval

Table 8: Performance of Different Retrieval Selection Methods on Locomo Dataset

Experiment	Normal			Oracle		
	8	16	32	8	16	32
Normal Retriever	0.5964	0.5991	0.5995	0.5989	0.6004	0.6021
Worst Selection	0.5920	0.5958	0.5973	0.5915	0.5942	0.5968
Random Selection	0.5931	0.5962	0.5978	0.5939	0.5960	0.5984

To further investigate the effectiveness of our retrieval methods, we conduct ablation studies on the Locomo dataset. We compare the performance of our normal retriever with two alternative selection strategies: worst selection and random selection. In the *Worst Selection* method, we deliberately choose the past utterances with the lowest retrieval scores, indicating their predicted irrelevance to the current context. This approach allows us to assess the impact of selecting the least relevant information on the model's performance. The *Random Selection* method, on the other hand, randomly selects past utterances without considering their retrieval scores. This serves as a baseline to evaluate the effectiveness of our targeted retrieval strategies compared to a random approach. Table 8 presents the results of these ablation experiments. We observe that the normal retriever consistently outperforms both the worst selection and random selection methods across all context lengths. Though the performance difference is only marginal indicating that just having the fixed context is sufficient.

D Using Other Evaluation Metrics

We further evaluate the performance of different retrieval methods, we employ ROUGE scores and the UniEval framework. Table 9 presents the ROUGE scores for the MTEB Embedding, CXMI, and BM25 retrieval methods on the Locomo dataset, with the BM25 retriever achieving the highest scores across all ROUGE variants. Additionally, we use the UniEval framework to assess the impact of retrieval on various dimensions of response quality. Table 10 shows the UniEval results for the Switchboard dataset, comparing the performance of the model with and without retrieval.

Method	rouge1	rouge2	rougeL	rougeLsum
MTEB Embedding	0.2093	0.0448	0.1613	0.1613
CXMI	0.2085	0.0447	0.1603	0.1604
BM25 Retriever	0.2194	0.0507	0.1686	0.1686

Table 9: ROUGE scores for different retrieval methods on Locomo Dataset

Dimension	Without Retrieval	With Retrieval
Naturalness	0.5624	0.5565
Coherence	0.6806	0.6944
Engagingness	1.551	1.6432
Groundedness	0.6652	0.6953
Understandability	0.5550	0.5491
Overall	0.8028	0.8272

Table 10: UniEval Results for the Switchboard Dataset

E Visualizing Generation Probabilities of Query with different utterances throughout the generation

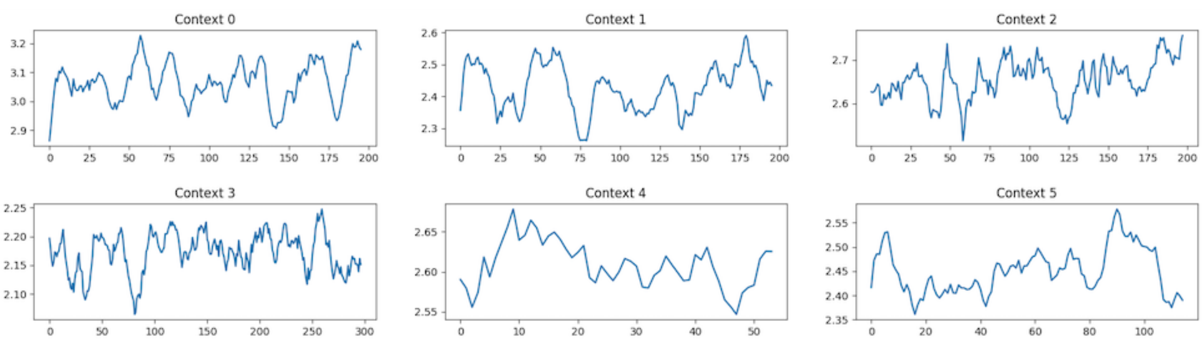

Figure 3: The above image shows the plot for -log(Prob) for the ground truth label response with different candidate utterances throughout the conversation. The x-axis has the utterance number in the conversation

Figure 3 visualizes the relevance of each utterance through the conversational history with respect to the ground truth label response. The relevance is checked by calculating the negative logarithm value of the probability of generating the ground truth sequence conditional on the given context. A lower negative log value indicates a higher relevance as the conditional probability is higher and vice versa. We observe that the plots have various local minima at various utterance numbers indicating that relevant utterances can be found anywhere in the past conversation history.