DocTalk: Scalable Graph-based Dialogue Synthesis for Enhancing LLM Conversational Capabilities

Jing Yang Lee[1]*, Hamed Bonab[2], Nasser Zalmout[2], Ming Zeng[2], Sanket Lokegaonkar[2]
Colin Lockard[2], Binxuan Huang[2], Ritesh Sarkhel[2], Haodong Wang[2]
Nanyang Technological University[1], Amazon[2]
jingyang001@e.ntu.edu.sg, {hamedrab, nzalmout, minzen, sllokega}@amazon.com
{clockard, binxuan, rssarkhe, wanghaod}@amazon.com

Abstract

Large Language Models (LLMs) are increasingly employed in multi-turn conversational tasks, yet their pre-training data predominantly consists of continuous prose, creating a potential mismatch between required capabilities and training paradigms. We introduce a novel approach to address this discrepancy by synthesizing conversational data from existing text corpora. We present a pipeline that transforms a cluster of multiple related documents into an extended multi-turn, multi-topic information-seeking dialogue. Applying our pipeline to Wikipedia articles, we curate *DocTalk*, a multi-turn pre-training dialogue corpus consisting of over 730k long conversations. We hypothesize that exposure to such synthesized conversational structures during pre-training can enhance the fundamental multi-turn capabilities of LLMs, such as context memory and understanding. Empirically, we show that incorporating *DocTalk* during pre-training results in up to 40% gain in context memory and understanding, without compromising base performance. *DocTalk* is available at https://huggingface.co/datasets/AmazonScience/DocTalk.

1 Introduction

Users expect conversational AI assistants to deliver dynamic and natural dialogue, driving a significant increase in Large Language Model (LLM) deployment. These multi-turn, multi-topic conversations involve users asking questions and LLMs providing relevant answers while maintaining context and adapting to topic shifts. However, LLMs are primarily pre-trained on large datasets of prosaic web-sourced text, including news, books, and papers with only a small portion comprising conversational data like chat logs or forums (Longpre et al., 2024; Zheng et al.). The limited amount of conversational data creates a gap between pre-training and typical LLM usage. Existing dialogue datasets largely focus on single-topic conversations, unlike real-world interactions, which span 1 to 10 topics, averaging 2 per session (Spink et al., 2002).

By bridging this gap during pre-training, we aim to enhance the LLM's fundamental multi-turn capabilities, particularly in context memory and understanding, crucial for maintaining coherence and relevance in extended dialogues (Bai et al., 2024). Context memory refers to the LLM's ability to accurately retrieve and utilize prior dialogue information to ensure continuity and relevance. Context understanding, on the other hand, involves resolving anaphora (e.g., identifying pronoun referents) and linking instructions to subsequent inputs across multiple dialogue turns for effective task comprehension. Emphasizing these capabilities during pre-training lays a robust foundation for advanced multi-turn reasoning in fine-tuning.

We propose a conversational data synthesis pipeline that converts web-derived prose into multi-turn, multi-topic, information-seeking dialogues featuring multiple topic switches. Unlike prior approaches that rely heavily on LLMs to generate a large portion of the data (Xu et al., 2024; Nayak et al., 2024; Ge et al., 2024; Long et al., 2024), our pipeline minimizes the use of LLM-generated text, reducing the risk of hallucination in addition to improving the quality and reliability of synthesized dialogues (Liu et al., 2024). Our approach also provides substantial cost benefits, achieving a 70% reduction in generation costs (Appendix A.1), significantly enhancing scalability. In contrast to methods like Dialogue Inpainting (Dai et al., 2022), which produces short, single-topic conversations, our pipeline models the dynamic topical shifts observed in human-LLM interactions, resulting in long, multi-topic conversations.

Applying this pipeline to Wikipedia, we create *DocTalk*, the largest multi-turn conversational

*Work done during internship at Amazon.

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 658–677
Aug 25–27, 2025.

dataset by token count to date, to our knowledge. We continue pre-training Mistral-7B-v0.3 on *DocTalk* and evaluate context memory and understanding using the Conversational Question Answering (CoQA) corpus and a novel LLM-as-a-judge framework. These metrics assess turn-level performance and the ability to incorporate relevant context in extended dialogues. We also employ guardrail metrics to investigate any potential degradation in existing LLM capabilities. Our experiments show that continued pre-training on data mixtures consisting of multi-turn, multi-topic information-seeking conversational data significantly enhances context memory and understanding, without compromising existing capabilities. Our key contributions are as follows:

- We introduce a scalable pipeline for synthesizing conversational data that transforms multiple documents into coherent multi-turn, multi-topic information-seeking dialogues.
- We create *DocTalk*, the largest conversational dataset by token count, by applying our pipeline to Wikipedia documents, demonstrating its effectiveness at scale.
- We provide empirical evidence that our approach enhances LLM context memory and understanding while preserving other model capabilities.

2 Methodology

Our conversational data synthesis pipeline consists of three stages. In *Stage 1*, we construct a document graph connecting multiple related documents and sample a set of documents from it. Leveraging multiple documents, rather than a single source, enables the creation of conversations that span multiple topics. In *Stage 2*, we divide the content of the sampled documents into segments and build a dialogue graph modeling probabilities of switching from one segment to another in natural conversations. We then sample these segments in sequence according to the dialogue graph and consider them as "assistant" utterances. Finally in *Stage 3*, we generate the corresponding user utterances by prompting an off-the-shelf LLM model to provide questions eliciting the assistant's responses at each conversational turn. We posit that pre-training on such structured data improves dialogue capabilities, particularly in aspects of contextual memory and comprehension.

2.1 Stage 1: Document Graph (GDoc)

We aim to systematically obtain and compile a set of n related documents to synthesize structured multi-turn dialogues. This process requires a pool of interconnected documents, where interconnections can be established through various meaningful dimensions of relatedness. Here, we primarily focus on explicit connections through direct references between documents (Frej et al., 2020). Alternative relatedness, such as sequential or hierarchical relationships and semantic topical similarity (Shi et al., 2024), remain as potential directions for future exploration.

GDoc Construction. We construct a weighted directional acyclic graph, $G_{doc} = (V_{doc}, E_{doc})$, where V_{doc} refers to a set of documents as vertices and E_{doc} refers to their connecting edges, i.e., $E_{doc} \subseteq \{(v^i_{doc}, v^j_{doc}) | v^i_{doc}, v^j_{doc} \in V_{doc}, i \neq j\}$. Our document graph construction starts with a random "anchor" document. Then, all related documents are collated to form the first level of connected vertices. Subsequently, from each of the newly formed vertices, related documents will be collated to form the next level of the graph–we stop at depth=3. For assigning edge weights, we adopt a topological approach based on vertex centrality measurements. For edge $e_{i,j}$ connecting v_i to v_j, the edge weight $w_{i,j} = deg^-(v_j)$ where deg^- refers to the out-degree centrality (Xamena et al., 2017). Out-degree centrality measures a document's direct references to others, indicating its prominence within the graph (Yang et al., 2014). Also, high quality and information rich documents typically have broad coverage, referencing numerous subtopics or related articles (Chhabra et al., 2021). This would encourage the synthesis of longer and more informative conversations. Constructing GDoc in this manner enables us to model various potential topical flows from the "anchor" document, representing possible dialogue trajectories in a conversational context. See Figure 1 for an illustration.

GDoc Traversal. We aim to sample a set of n documents–we use $n = 3$. Traversal begins at the anchor document, following a probabilistic random walk, where the next vertex is selected based on the associated edge weights. The traversal stops either when n documents are collected or when a vertex with no outgoing edges is reached, yielding a cohesive and thematically related set of documents–see Appendix A.4.

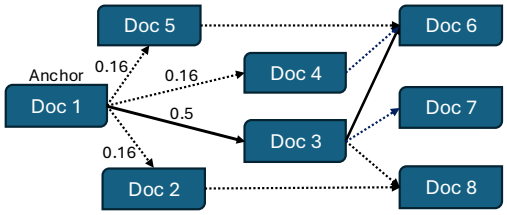

Figure 1: An illustrative example of GDoc constructed in *Stage 1*. The edges (dotted arrows) originating from Doc 1 (the "anchor") are labeled with their respective sampling probabilities–based on the out-degree centrality. The bold arrows indicate the sampled edges.

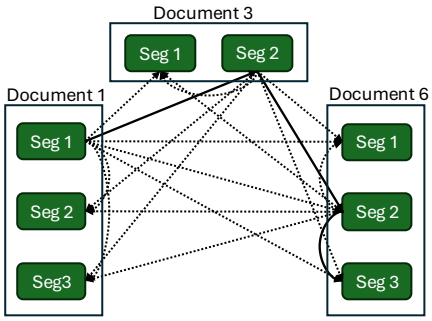

Figure 2: An illustrative example of GDial constructed in *Stage 2* assuming Doc 1, 3 and 6 are sampled during *Stage 1*. Arrows represent edges and bold arrows indicate the order of 4 segments sampled for assistant utterances–i.e., [Doc1, Seg1]→[Doc3, Seg2]→[Doc6, Seg2]→[Doc6, Seg3].

2.2 Stage 2: Dialogue Graph (GDial)

We construct m assistant utterances from the n related documents obtained in Stage 1. We first segment each document and then systematically reorder and interleave these segments to create a natural topic progression. m corresponds to the total number of segments derived from n documents. This interleaving of segments from different but related documents enables the creation of conversations that are both multi-topic and multi-turn. Each utterance contributes novel information while maintaining contextual references to previous utterances derived from the same source document, thereby establishing cross-turn coreference.

GDial Construction. Using m text segments extracted from documents sampled in Stage 1, we construct a separate directed graph $G_{dial} = (V_{dial}, E_{dial})$. We begin by splitting each document into distinct segments, serving as vertices V_{dial}, representing a potential assistant utterance. The graph is fully connected, with edges linking each vertex to every other vertex, regardless of their source documents and with no self-loops, such that $E_{dial} = \{(v^i_{dial}, v^j_{dial}) | v^i_{dial}, v^j_{dial} \in V_{dial}, i \neq j\}$. Each segment is constrained to single use, ensuring that the resulting conversations contain no duplicate assistant utterances. See Figure 2 for an illustration. A critical component in the construction of GDial is the assignment of edge weights that model natural conversational flow. We introduce a novel Conversational Reward (CR) model fine-tuned to assign scores based on the coherence between the assistant's previous utterance(s) and potential subsequent utterances represented by connected vertices. These edges are then sampled based on probabilities proportional to their assigned weights. A higher CR model score indicates that the utterance tied to the connected vertex is better suited as the assistant's next response in a multi-turn, information-seeking conversation. This approach enables us to construct a complete graph, with the CR model guiding its traversal.

Conversational Reward (CR) Model. We conceptualize the CR model as a learned scoring mechanism that quantifies the appropriateness of candidate utterances in the context of preceding assistant turns within multi-turn dialogues. Drawing from established semantic matching objectives, the model is trained to evaluate local coherence and topical progression across turns, thereby assigning scalar values that reflect the conversational flow in information-seeking interactions. We formulate the objective of the CR model as estimating the conditional probability $P(A^c_{t+1}|A_t)$, where A^c_{t+1} is the candidate utterance in turn $t+1$ and A_t is the preceding assistant utterance at turn t. For tractability, we restrict contextual input to one preceding assistant utterance. While this design choice facilitates computational efficiency and interpretability, extending the model to condition on broader dialogue history remains a promising direction for future research.

Our CR model builds upon a pre-trained ranking model, acknowledging relevance as a key determinant of conversational flow. We enhance this foundation through additional training on high-quality conversational data to capture dialogue-specific characteristics. We collect our training data using $4,240$ conversations aggregated from six corpora–for information seeking: HybriDialogue (Nakamura et al., 2022), InScitic (Wu et al., 2023), TopiOCQA (Adlakha et al., 2022), and Wizard of Wikipedia (Dinan et al., 2019); for LLM-

Table 1: CR model performance (Averaged MRR)

Model	LLM-based	Info Seeking
bge-reranker-base	0.199	0.121
Iteration 1 (w. $k = 1$)	0.381	0.182
Iteration 2 (w. $k = 2$)	**0.413**	0.168
Iteration 3 (w. $k = 3$)	0.371	**0.17**

based: ShareGPT*, and UltraChat (Ding et al., 2023). More details are provided in Appendix A.7.

We use *bge-reranker-base* (Xiao et al., 2023) as the starting base model. To prepare training samples, we randomly select a base assistant utterance (excluding the last utterance) from each conversation in the training set, and the next assistant utterance as the "positive" sample y_p. For "negative" sample y_n, we randomly select from a $3*k$ set collected by: *(a)* using the base model, get top-k assistant utterances within the same conversation–as a form of hard negative mining (Robinson et al., 2021); *(b)* using the base model, get top-k utterances from other dialogues; and *(c)* Another k random utterances from other dialogues. We fine-tune the base model iteratively via a pairwise log probability contrastive loss, $-log(R(y_p) - R(y_n))$, where $R(\cdot)$ refers to the CR model output. We update mined hard-negatives iteratively with three iterations of the base CR model, and train each iteration for two epochs–with LR= $3e^{-5}$.

To evaluate the CR model performance, we select 100 conversations from each dataset's test set, with performance measured by Mean Reciprocal Rank (MRR). We train with three iterations, each with a distinct dataset corresponding to k = 1, 2 and 3, respectively. Results are reported in Table 1. We observe that finetuning with k = 2 yielded the highest MRR scores for LLM-based dialogue and performed comparably to the model fine-tuned with k = 3 for information-seeking dialogue. Notably, the fine-tuned models showed significant improvement in the next utterance prediction task compared to the base model. The complexity of this task presents major opportunities, suggesting multiple avenues for future investigation and enhancement.

GDial Traversal After constructing G_{dial}, we sample a set of m assistant utterances by performing a probabilistic random walk. At each node, the next node is sampled with a probability proportional to the edge weights assigned by the CR model. To avoid repetition, each visited vertex is excluded from future selections during the traversal. The traversal ends when every node in GDial has been visited. It should be highlighted that different paths through GDoc and GDial would yield distinct topical trajectories and unique dialogues, effectively capturing the one-to-many nature of conversational dynamics. An algorithm is provided in Appendix A.4.

2.3 Stage 3: User Utterance Generation

We generate user utterances that precede each of the m assistant utterances by prompting an LLM. Essentially, we prompt the Mistral-2-7b-Instruct to generate a question which elicits each assistant utterance in the conversation. Only the user utterances are generated in this process, and no LLM-generated content is included in the assistant utterances. While this approach may slightly impact the naturalness, it significantly minimizes hallucinations in the resulting conversations, which is critical for effective pre-training (Liu et al., 2024). Our simplified approach achieves at least 70% reduction in generation costs, significantly enhancing its potential for large-scale deployment. Further details are provided in Appendix A.1 and A.3.

3 DocTalk

Our pipeline is applied to English Wikipedia documents, resulting in a conversational dataset named *DocTalk*. In *Stage 1*, we utilize the WIKIR toolkit (Frej et al., 2020), providing query documents (i.e. anchor documents) and their relevance labels (qrels) indicating URL references to construct GDoc. We filter out documents with fewer than 10 qrels to exclude shorter or obscure articles, resulting in a final set of 731,511 anchor documents. We sample up to $n = 3$ documents per anchor document, including itself, and limit qrels to 20 per document to form GDoc edges. In *Stage 2*, documents are segmented by paragraph to reduce LLM-generated user utterances and hallucinations while utilizing all text segments. During *Stage 3*, every vertex is visited once during traversal, resulting in a conversation with m turns, where m represents the total number of text segments. Corpus statistics are provided in Table 2. Figure 3 shows the first 5 turns of a *DocTalk* conversation. The first 10 turns and an additional excerpt are in Appendix A.5.

While *DocTalk* conversations are more direct and focused–omitting pleasantries and filler words–they effectively capture topical shifts and coherence necessary for effective human-LLM interac-

*The ShareGPT dataset: https://sharegpt.com/

Table 2: DocTalk statistics summary

Total # of conversations		730,707	
Metric	**Mean**	**Std**	**Median**
# turns per conversation	82.2	53.9	70
Asst Utterance Length (# words)	87.3	63.6	73
User Utterance Length (# words)	26.5	11.3	24
# Doc Shifts per conversation	23.3	15.9	19

Turn #1
USER: Where is the Toronto General Hospital located and what makes it notable?
ASST: The Toronto General Hospital (TGH) is a major teaching hospital in Toronto, Ontario, Canada and the flagship campus of University Health Network (UHN). It is located in the Discovery District of Downtown Toronto…

Turn #2
USER: What specific medical specialties are covered in the final two years of the Doctor of Medicine program at the University of Toronto, and what types of training are provided during these years?
ASST: The Doctor of Medicine program at the University of Toronto is a 4-year MD program with a total enrollment of about 850 students. The first two years are known as the preclerkship curriculum…

Turn #3
USER: Which global university rankings place the school highest for clinical medicine and surgery in the years 2023 and 2020?
ASST: In 2023 the school was ranked 4th in the world for clinical medicine and surgery by "U.S. News & World Report"…

Turn #4
USER: What specific medical procedures and transplants does Toronto General Hospital specialize in and what are some notable achievements in these areas?
ASST: The emergency department now treats 28,065 persons each year, while the hospital also houses the major transplantation service for Ontario…

Turn #5
USER: Which hospitals have merged to form the University Health Network (UHN) and when did these mergers occur?
ASST: A series of mergers over many years has resulted in the UHN in its current form…

Figure 3: First 5 turns of a *DocTalk* conversation based on 3 Wikipedia articles: "Toronto General Hospital" (in orange), "University of Toronto Faculty of Medicine" (in blue), and "United Health Network" (in green).

tions. The primary aim of our conversation synthesis pipeline is to generate a large-scale dataset that fundamentally improves context memory and understanding in multi-turn conversations during pre-training, emphasizing structural aspects such as topic transitions over stylistic naturalness. Although this approach produces conversations that are stylistically less human-like, these aspects can be refined during post-training phases or through fine-tuning. Future work could entail extending our approach to synthesize more natural multi-turn dialogues on a larger scale.

Human Evaluation To ensure the quality of *DocTalk*, we conducted a human evaluation as a form of quality control. We engaged 5 annotators to review 50 randomly selected dialogues from *DocTalk* and assess their quality through a questionnaire. The questionnaire assessed various dimensions of the dialogues, including the contextual relevance of responses (Q1), the specificity and accuracy of user utterances (Q2), the amount of linguistic errors or level of grammatical correctness (Q3), the naturalness and overall resemblance to human-LLM conversation (Q4), the thematic connection between topics (Q5), and the logical consistency of topic transitions (Q6). For further comparison, we also evaluated 50 dialogues from WikiDialog using the same questionnaire (excluding Q5 and Q6 which are specific to DocTalk). See Appendix A.8 for further details.

Table 3: Human evaluation for DocTalk and WikiDialog. Inter-annotator agreements are via Krippendorff's α

Metric	α	**DocTalk**	**WikiDialog**
Q1. Contextual Relevance	0.87	0.94	0.87
Q2. Specificity	0.95	0.96	0.64
Q3. Linguistic Errors	0.94	0.98	0.78
Q4. Naturalness	0.80	0.79	0.79
Q5. Topic Relatedness	0.91	0.95	N/A
Q6. Topic Shift Coherence	0.88	0.83	N/A

Human evaluation results are presented in Table 3. These scores confirm that the assistant utterances in *DocTalk* are highly contextually relevant to the corresponding user utterances, the user utterances exhibit strong precision and specificity, and the dialogues maintain grammatical correctness and naturalness. Additionally, the high scores for Q5 and Q6 indicate that the Wikipedia articles selected in Stage 1 are semantically related and thematically cohesive, and that the topic shifts introduced in Stage 2 support logical and coherent multi-topic conversations. WikiDialog, sharing a common basis in Wikipedia articles, showed comparable levels of naturalness. However, *DocTalk* excelled in contextual relevance, specificity, and grammatical accuracy, highlighting the effectiveness of our pipeline.

4 Experiments

We utilize the annealing data quality assessment approach widely used in other studies (Blakeney et al.; Grattafiori et al., 2024). We continue pre-train the Mistral-7B-v0.3[†] model for 8,000 steps (30B Tokens), applying a learning rate schedule that annealed from 3e-5 to 3e-6. We keep some learning rate bandwidth for instruction fine-tuning using ShareGPT data to prepare the model for benchmarking. For all our experiments, we construct a dataset comprising 25% conversational data, 5% Project Gutenberg books[‡], and 70% RedPajama v1[§] Web content. Following this data composition,

[†]https://huggingface.co/mistralai/Mistral-7B-v0.3

[‡]Books data in the RedPajama dataset was replaced with Project Gutenberg content due to copyright concerns.

[§]Content from sites listed in the U.S. Notorious Markets for Counterfeiting and Piracy report were excluded.

we prepare four variations to study our proposed solution in practice.

DocTalk: We use the *DocTalk* as described in Section 3. Along our main treatment experiment, we also prepare two ablation studies to examine DocTalk's constituent components: 1) We only applying *Stage 1* and *3* of the pipeline (w/o *Stage 2*) to investigate the impact of GDial. In other words, we continuing pre-training on a conversational dataset curated solely with GDoc, preserving the document graph's original traversal sequence without segment reordering. 2) We only apply *Stage 1* of the pipeline (w/o *Stage 2 & 3*). Likewise, this baseline uses Wikipedia articles ordered solely by GDoc, excluding all user utterances.

DocTalk*: We continue pre-training on a *DocTalk* variant restricted to the first 30 turns (roughly one-third of each conversation) to investigate if initial conversational turns are of higher quality than later ones. To match token counts across dataset configurations, we triplicate the resulting dataset.

Plain Wiki: Continue pre-training on Wikipedia articles without any additional treatment. Neither GDoc nor GDial are used.

WikiDialog: Continue pre-training on WikiDialog (triplicated to balance token count).

4.1 Evaluation

4.1.1 Context Memory & Understanding

Existing popular multi-turn benchmarks, such as MTBench (Zheng et al., 2023) and WildBench (Lin et al., 2024), evaluate high-level conversational reasoning. While these benchmarks overlap with our evaluation objectives, they do not explicitly assess context memory and understanding–MTBench is restricted to two turns, whereas WildBench features test samples of up to four turns, with only 20% of dialogues exceeding two. We conduct a targeted evaluation by leveraging CoQA and devising a novel LLM-as-a-judge metric.

CoQA Similar to Huber et al. (2024), we investigate the turn-level performance of the baselines by leveraging the test set provided in the CoQA corpus (Adlakha et al., 2022). Each sample consists of a text passage and a series of questions and answers that form a dialogue. At each turn, the LLM generates a response based on the passage and dialogue history. From the second turn onward, queries depend on both the passage and preceding dialogue, requiring the LLM to track context and resolve pronoun references (e.g., "Who is she?" or "How many were there?"). We report turn-level and overall F1, precision, and recall, computed via word overlap with reference responses.

LLM-as-a-Judge For our LLM-as-a-judge metric, we curtate a 73-sample multi-turn conversational dataset centered on shopping interactions. Each sample includes a conversation, a shopping-specific system prompt, and 1–3 manually labeled user intents that the response must address. Generating a response that addresses these intents would demonstrate an LLM's ability to resolve anaphora and understand the relationship between instructions and inputs. Hence, improvements on this benchmark would reflect gains in context memory and understanding.

Following the system prompt's instruction, each response contains an introduction paragraph and a bullet-point paragraph. We measure how effectively the response addresses the user's intents via 4 metrics derived via an LLM-as-judge approach: $Intro$ for coverage in the introductory paragraph, $BulletPoint$ for coverage in bullet points, $Loose$ for coverage by either the introduction or bullet points, and $Strict$ for coverage by both. See Appendix A.2 for a detailed explanation.

4.1.2 Guardrail Metrics

To ensure that pre-training on *DocTalk* does not degrade general LLM capabilities, standard benchmarks were used to evaluate knowledge, reasoning, and long-context capabilities of the base model (i.e., before IFT). For general knowledge, 5-shot MMLU(Hendrycks et al., 2021a) is used. Commonsense reasoning is assessed using 25-shot ARC (Clark et al., 2018), 5-shot WinoGrande(Wino.G)(Sakaguchi et al., 2019), 0-shot PIQA(Bisk et al., 2019), and 10-shot HellaSwag(H.Swag)(Zellers et al., 2019); logic reasoning relies on 4-shot MATH(Hendrycks et al., 2021b), 3-shot BBH(Suzgun et al., 2022), and 5-shot GSM8k(Cobbe et al., 2021); fluency is evaluated with WikiText2(Merity et al., 2016); and long-context performance with MuSiQue(Trivedi et al., 2022) and 2WikiMultiHopQA(2WikiQA)(Ho et al., 2020).

5 Results & Discussion

Context Memory & Understanding CoQA turn-level and overall F1, precision, and recall scores are shown in Figure 4 and Table 4, respectively. Samples are provided in Appendix 11. Overall, the baseline pre-trained on **DocTalk*** and **DocTalk**

Table 4: CoQA evaluation results

Model	F1	Precision	Recall	Ave # words
DocTalk	0.38	0.36	0.6	9.52
- w/o Stage 2 & 3	0.29	0.27	**0.61**	13.97
- w/o Stage 2	0.27	0.25	0.52	12.06
DocTalk*	**0.40**	**0.37**	**0.61**	9.07
Plain Wiki	0.34	0.31	0.6	12.58
WikiDialog	0.36	0.32	0.59	11.02

Table 5: LLM-as-a-judge evaluation results

	Intro	Bullet Point	Loose	Strict
DocTalk	**0.424**	0.507	**0.542**	**0.331**
- w/o Stage 2 & 3	0.288	0.452	0.381	0.271
- w/o Stage 2	0.263	0.522	0.441	0.254
DocTalk*	0.195	**0.523**	0.525	0.186
Plain Wiki	0.305	0.518	0.424	0.254
WikiDialog	0.271	0.482	0.441	0.263

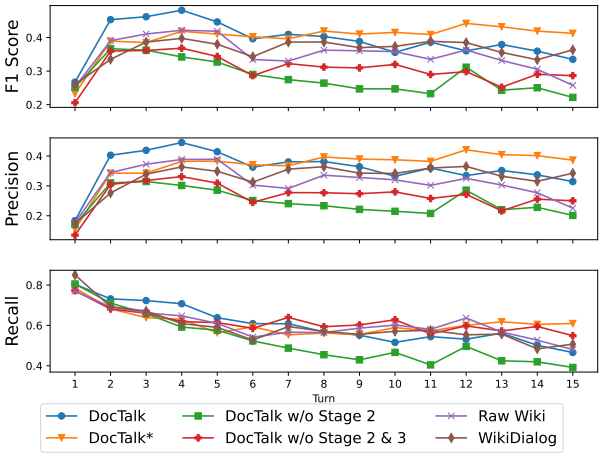

Figure 4: Turn-level F1, precision, and recall plots.

outperformed all other baselines in terms of F1 and precision. Qualitatively, we observe that LLMs pre-trained on **Plain Wiki** or **WikiDialog** are more likely to generate erroneous responses, including incorrect or illogical answers, particularly in conversations with a larger number of turns. In contrast, LLMs pre-trained on **DocTalk** or **DocTalk*** demonstrate a significantly higher rate of accurate answers. It should also be highlighted that the responses generated by baselines pre-trained on **DocTalk*** and **DocTalk** demonstrated gains on F1 and precision despite being shorter in length.

Furthermore, the baseline pre-trained on **DocTalk*** consistently outperformed all other baselines in precision and F1 score after the 8th turn, even surpassing the baseline pre-trained on **DocTalk**. We hypothesize that **DocTalk** dialogues become more abrupt and erratic in later stages as they exhaust all vertices of the dialogue graph and every segment of the Wikipedia document, leading to increasingly dissonant text in subsequent turns. By extracting the first 30 turns, we capture the dialogue's more coherent early sections with logical topic transitions. Future work could explore optimizing performance by adjusting the number of turns extracted. Regarding recall, most baselines showed similar performance, likely because LLMs tend to generate longer responses with greater overlap, boosting recall but reducing precision. The relatively poor performance of baselines pre-trained on **DocTalk** w/o Stage 2 and **DocTalk** w/o Stage 2 & 3 underscores the importance of the dialogue graph and CR model for context memory and understanding.

Additionally, according to the LLM-as-a-judge metrics in Table 5, pre-training on **DocTalk** and **DocTalk*** generally yield the best performance. The performance boost from incorporating the dialogue graph highlights its significance. We notice that the baseline pretrained on **DocTalk** achieves far greater Intro Coverage relative to all other baselines, including **DocTalk***, which also results in a higher Overall strict coverage score. We hypothesize that longer pre-training dialogues, with more context understanding demonstrations across multiple turns, contribute to this improvement.

Guardrail Evaluation Table 6 presents the guardrail benchmarking scores, and Figure 5 depicts the scores during continued pre-training (before IFT). Overall, except for WikiText2 and MMLU, we observed no significant decline in base-model performance following continued pre-training on *DocTalk*. The implemented baselines demonstrate comparable scores across general knowledge and reasoning tasks, showing steady improvement as the number of pre-training steps increases. For WikiText2, the perplexity reflects how much the curated dataset differs from the original Wikipedia. The baseline trained on **Raw Wiki** achieved the lowest perplexity, while **WikiDialog** attained the highest. This is expected since WikiText2 is a language modeling benchmark based on original Wikipedia articles; thus, the more the pre-training dataset deviates from Wikipedia, the poorer the performance.

We note that WikiDialog demonstrates superiority in commonsense reasoning benchmarks, notably in ARC scores, compared to *DocTalk*. This improvement can be attributed to WikiDialog's sentence-level granularity in question-answer pairs, which better align with the structure of reasoning

Table 6: Guardrail benchmarking results for the latest checkpoint.

Model	Knowledge	Commonsense Reasoning				Logical Reasoning			Fluency	Long Context	
	MMLU	ARC	Wino.G	H.Swag	PIQA	MATH	BBH	GSM8k	WikiText	MuSiQue	2WikiQA
DocTalk	0.53	0.45	**0.71**	0.77	**0.80**	**0.07**	0.49	0.20	6.04	**0.12**	**0.28**
- w/o Stage 2 3	**0.55**	0.44	**0.71**	0.77	0.78	0.06	**0.50**	0.20	4.98	0.07	0.18
- w/o Stage 2	**0.55**	0.46	0.68	0.77	**0.80**	0.05	0.46	0.19	5.74	0.05	0.20
DocTalk*	0.52	0.46	**0.71**	0.78	0.79	0.06	0.46	0.12	6.45	0.06	0.19
Plain Wiki	**0.55**	0.45	0.70	**0.80**	0.79	0.05	0.47	0.21	**4.96**	0.08	0.25
WikiDialog	0.51	**0.48**	**0.71**	0.78	**0.80**	**0.07**	0.49	**0.22**	7.36	0.11	0.21

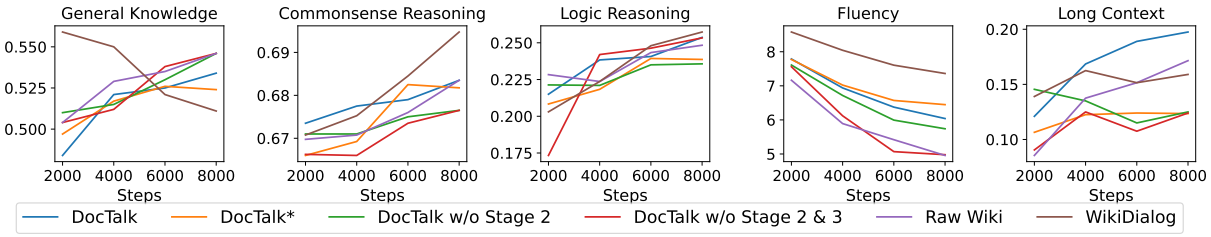

Figure 5: Average general knowledge, reasoning, fluency, and long context scores evaluated every 2000 steps.

benchmarks. However, we observe a degradation in MMLU performance for **WikiDialog** as training steps increase. This decline may be attributed to the large number of user utterances generated via Dialogue Inpainting, showing adverse impact.

The baseline pre-trained on **DocTalk** showed significant performance gains on MuSiQue and 2WikiHotPotQA. We hypothesize that the long multi-turn conversations in *DocTalk* serves to enhance the LLM's long-context capabilities. In contrast, baselines pre-trained on **DocTalk** w/o Stage 2 or w/o Stage 2 & 3 applied performed worse than the baselines pre-trained on **Raw Wiki**, emphasizing the dialogue graph's effectiveness in enhancing long-context performance.

6 Related Work

LLM Pre-training Pre-training is crucial stage where LLMs acquire foundational knowledge, typically through unsupervised learning with language modeling or masked language modeling objectives on large-scale web-crawled text (Radford et al., 2019). Typically, data selection–encompassing filtering, deduplication, and mixing–is performed to curate a high-quality pre-training dataset (Albalak et al., 2024). Sophisticated curation methods have been proposed to align pre-training data with downstream tasks. Previous work has explored incorporating task-specific instructions and responses (Cheng et al., 2024), merging related documents to facilitate cross-document reasoning (Shi et al., 2024), introducing soft prompts (Gu et al., 2022), and using regex patterns to format reading comprehension data (Jiang et al., 2024).

Information Seeking Dialogue Synthesis There has been significant research devoted to the synthesis of information seeking dialogue. A common approach involves iteratively generating document-grounded conversations by extracting relevant information and crafting coherent responses to questions posed by human annotators (Kim et al., 2022; Hwang et al., 2023; Wu et al., 2023). This approach, while effective, relies heavily on resource-intensive human annotation. To mitigate this, alternatives like AutoConv (Li et al., 2023), Dialogue Inpainting (Dai et al., 2022) and SOLID(Askari et al., 2025) leverage LLMs instead. AutoConv fine-tunes LLMs on human dialogues using iterative prompting to generate user and system responses, Dialogue Inpainting inserts conversational utterances between document sentences, and SOLID uses a self-seeding, multi-intent self-instructing approach. In addition to the broad information seeking paradigm, prompt-based LLM approaches have also been used to synthesize dialogues for specific use-cases such as education (Wang et al., 2024) and emotional support (Seo and Lee, 2024).

7 Conclusion

We present a conversational data synthesis pipeline that transforms web-crawled text into multi-turn, multi-topic dialogues, enhancing LLMs' context memory and understanding during pre-training. Using this pipeline, we curate *DocTalk*, the largest conversational dataset to date. Future enhancements could refine the CR model by scoring tran-

sitions with full dialogue context and extend the pipeline to general texts via semantic matching for document relatedness. Future investigations could explore using these dialogues for grounded synthetic conversation generation, potentially enhancing conversational naturalness in data subsets for mid-training or post-training. Additionally, the *DocTalk* corpus mainly consists of straightforward question-and-answer exchanges. Future work could explore incorporating more natural conversational dynamics, such as clarifications, misconceptions, requests for further details, or supporting evidence, to enhance the naturalness of the dialogues. It would be interesting to examine how these enhancements affect multi-turn conversational abilities.

8 Limitations

Similar to other projects that involve the use of synthetic data, there is an inherent risk of hallucination. This risk can have adverse effects on the performance and reliability of the LLM. For our pipeline, we have taken steps to mitigate this risk by only synthesizing the user utterances, not the assistant utterances. By doing so, we aim to reduce the likelihood of hallucinations negatively influencing the LLM's behavior. However, despite these precautions, it is important to recognize that the possibility of hallucination cannot be entirely eliminated. We plan to release the data only to be used for research purpose, and derivatives of data should not be used outside of research contexts.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching. *Preprint*, arXiv:2110.00768.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Preprint*, arXiv:2402.16827.

Arian Askari, Roxana Petcu, Chuan Meng, Mohammad Aliannejadi, Amin Abolghasemi, Evangelos Kanoulas, and Suzan Verberne. 2025. SOLID: Self-seeding and multi-intent self-instructing LLMs for generating intent-aware information-seeking dialogs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6375–6395, Albuquerque, New Mexico. Association for Computational Linguistics.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 7421–7454. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *Preprint*, arXiv:1911.11641.

Cody Blakeney, Mansheej Paul, Brett W Larsen, Sean Owen, and Jonathan Frankle. Does your data spark joy? performance gains from domain upsampling at the end of training, 2024. *URL https://arxiv.org/abs/2406.03476*.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024. Instruction pre-training: Language models are supervised multitask learners. *Preprint*, arXiv:2406.14491.

Anamika Chhabra, Shubham Srivastava, S. R. S. Iyengar, and Poonam Saini. 2021. Structural analysis of wikigraph to investigate quality grades of wikipedia articles. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 584–590, New York, NY, USA. Association for Computing Machinery.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. *Preprint*, arXiv:2205.09073.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *Preprint*, arXiv:1811.01241.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *Preprint*, arXiv:2305.14233.

Jibril Frej, Didier Schwab, and Jean-Pierre Chevallet. 2020. WIKIR: A python toolkit for building a large-scale Wikipedia-based English information retrieval dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1926–1933, Marseille, France. European Language Resources Association.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Patrick Huber, Arash Einolghozati, Rylan Conway, Kanika Narang, Matt Smith, Waqar Nayyar, Adithya Sagar, Ahmed Aly, and Akshat Shrivastava. 2024. Codi: Conversational distillation for grounded question answering. *Preprint*, arXiv:2408.11219.

Yerin Hwang, Yongil Kim, Hyunkyung Bae, Hwanhee Lee, Jeesoo Bang, and Kyomin Jung. 2023. Dialogizer: Context-aware conversational-QA dataset generation from textual sources. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8806–8828, Singapore. Association for Computational Linguistics.

Ting Jiang, Shaohan Huang, Shengyue Luo, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. Improving domain adaptation through extended-text reading comprehension. *Preprint*, arXiv:2401.07284.

Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022. Generating information-seeking conversations from unlabeled documents. *Preprint*, arXiv:2205.12609.

Siheng Li, Cheng Yang, Yichun Yin, Xinyu Zhu, Zesen Cheng, Lifeng Shang, Xin Jiang, Qun Liu, and Yujiu Yang. 2023. AutoConv: Automatically generating information-seeking conversations with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1751–1762, Toronto, Canada. Association for Computational Linguistics.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2024. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. *Preprint*, arXiv:2406.04770.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data. *Preprint*, arXiv:2404.07503.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022. HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492, Dublin, Ireland. Association for Computational Linguistics.

Nihal V. Nayak, Yiyang Nan, Avi Trost, and Stephen H. Bach. 2024. Learning to generate instruction tuning datasets for zero-shot task adaptation. *Preprint*, arXiv:2402.18334.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *International Conference on Learning Representations (ICLR)*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Seungyeon Seo and Gary Geunbae Lee. 2024. DiagESC: Dialogue synthesis for integrating depression diagnosis into emotional support conversation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 686–698, Kyoto, Japan. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2024. In-context pretraining: Language modeling beyond document boundaries.

Amanda Spink, Hüseyin Özmutlu, and Seda Özmutlu. 2002. Multitasking information seeking and searching processes. *JASIST*, 53:639–652.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024. Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9707–9731, Bangkok, Thailand. Association for Computational Linguistics.

Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. InSCIt: Information-seeking conversations with mixed-initiative interactions. *Transactions of the Association for Computational Linguistics*, 11:453–468.

Eduardo Xamena, Nélida Beatriz Brignole, and Ana Gabriela Maguitman. 2017. A structural analysis of topic ontologies. *Information Sciences*, 421:15–29.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *Preprint*, arXiv:2406.08464.

Yang Yang, Yuxiao Dong, and Nitesh V. Chawla. 2014. Predicting node degree centrality with the node prominence profile. *Scientific Reports*, 4(1).

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Appendix

A.1 Cost Advantages

DocTalk consists of approximately 8 billion tokens. If every token in the corpus were synthetically generated using GPT-4's pricing (2.50 USD per 1 million input tokens and 10.00 USD per 1 million output tokens), the cost of generating *DocTalk* would amount to 80,000 USD. This estimate excludes the cost of input tokens, which would vary depending on the complexity of the prompts used. By leveraging our pipeline, only the user queries are generated synthetically, while the assistant's responses are used to prompt the LLM to generate the user queries. This approach significantly reduces costs. Specifically, generating 500 million tokens for user queries would cost approximately 5,000 USD for output tokens and 18,750 USD for input tokens, totaling 23,750 USD. This represents nearly a 70% reduction in cost compared to the fully synthetic generation approach.

A.2 LLM-as-a-Judge Evaluation

In order to rigorously evaluate how well our model retains contextual information, understands user instructions, and addresses specific intents within a multi-turn conversation, we define a set of metrics designed to assess the model's ability to follow the system prompt, resolve anaphora, and cover the user's requested topics. In particular, we focus on the extent to which both the introduction paragraph and the bullet-point paragraph of the model's response address the core user intents identified in our curated shopping-related dataset. By quantifying the proportion of these required intents satisfied in either or both of these components, we aim to capture the model's degree of context memory and understanding in a manner that is both comprehensive and interpretable. Overall, evaluation results show a 90% consistency rate with human annotations, demonstrating the reliability of our metrics.

A.2.1 Response Structure

Following the system prompt's instruction, each response (denoted as $resp$) produced by the LLM consists of two components:

- An **introduction paragraph** (denoted as $resp\_intro$).
- A **bullet-point paragraph** (denoted as $resp\_bulletpoint$).

A.2.2 Metrics for Measuring Context Memory and Understanding

Let $intent$ denote the set of user intents that must be addressed, and let $|intent|$ represent the total number of these required user intents. We define:

$$A_{intro} = \{\text{user intents addressed by the introduction}\} \tag{1}$$

$$A_{bulletpoint} = \{\text{user intents addressed by the bullet points}\} \tag{2}$$

The function:

$$intent\_coverage(A, intent) = \frac{|A \cap intent|}{|intent|} \tag{3}$$

measures the proportion of required user intents successfully covered by a given set A.

We define four evaluation metrics as follows:

- **Intro**: The $intro$ metric assesses if the response introduction addresses each of the user intents.

$$intro(resp\_intro, requirement) = intent\_coverage(A_{intro}, intent) \tag{4}$$

- **Bullet Point**: The $bulletpoint$ metric assess if the bullet points in the response addresses the user intents.

$$bulletpoint(resp\_bulletpoint, requirement) = intent\_coverage(A_{bulletpoint}, intent). \tag{5}$$

LLM-as-a-Judge Intro Prompt

```
You are a judge who is asked to judge whether the given response mentions, utilizes, or includes the information in
the given information. First, write down step-by-step reasoning inside <reason> XML, and then write down the final
conclusion inside <conclusion> XML. Write down the final conclusion inside <conclusion> XML.
    <response>
    {resp_intro}
    </response>
    <information>
    {user_intent}.
    </information>
The conclusion should either be: <conclusion>Yes, it clearly uses the information.</conclusion> or: <conclusion>No,
it misses some information.</conclusion> Must not use any inference or external knowledge!!
```

LLM-as-a-Judge Bullet Points Prompt

```
You are a judge who is asked to judge whether the given response mentions, utilizes, or includes the information in
the given information. First, write down step-by-step reasoning inside <reason> XML, and then write down the final
conclusion inside <conclusion> XML. Write down the final conclusion inside <conclusion> XML.
    <response>
    {resp_bulletpoints}
    </response>
    <information>
    {user_intent}.
    </information>
The conclusion should either be: <conclusion>Yes, it clearly uses the information.</conclusion> or: <conclusion>No,
 it misses some information.</conclusion> Must not use any inference or external knowledge!!
```

Figure 6: Sample from the LLM-as-a-judge evaluation dataset.

- **Loose**: The *loose* metric checks if *either* the introduction *or* the bullet points address the user intents. In set-theoretic terms, this corresponds to the union of addressed intents:

$$loose(resp, intent) = intent\_coverage(A_{intro} \cup A_{bulletpoint}, intent). \tag{6}$$

- **Strict**: The *strict* metric requires that *both* the introduction *and* the bullet points address the user intents. This corresponds to the intersection of addressed intents:

$$strict(resp, intent) = intent\_coverage(A_{intro} \cap A_{bulletpoint}, intent). \tag{7}$$

A.2.3 LLM-as-a-Judge Approach

To compute these metrics, we adopt an LLM-as-a-judge approach:

- We select 10 samples from the curated 73-sample dataset to elicit responses from a LLM(Claude 3).
- These responses are then manually labeled with binary indicators to specify whether each user intent was addressed or not, forming the sets A_{intro} and $A_{bulletpoint}$ for each response.
- The manually labeled responses serve as a ground truth for fine-tuning the LLM-as-a-judge prompt.

By refining the judge prompt with these 10 labeled samples, we achieve at least 80% recall and precision in evaluating whether the LLM's responses demonstrate context memory and understanding. Thus, the LLM-as-a-judge can reliably estimate A_{intro}, $A_{bulletpoint}$, and consequently $intent\_coverage$. The prompt templates are provided in Fig 6.

The final scores shown in Table 5 are obtained by averaging the scores across all samples. Figure 7 provides a representative sample from the evaluation dataset.

A.3 User Utterance Generation Prompt

The prompt used for user utterance in Stage 3 is provided Figure 8.

A.4 Document & Dialogue Graph Traversal Algorithms

The algorithms for sampling the document graph in Stage 1 and the dialogue graph in Stage 2 is provided in Algorithm 1 and 2, respectively.

USER: I'm looking for a smart thermostat with the newest energy-saving features.

ASST: When shopping for a smart thermostat, look for models with built-in learning capabilities that can automatically adjust temperatures based on your habits, as well as voice control and mobile app access for remote adjustments. Key features to compare include compatibility with your HVAC system, smart home platforms, ease of installation, and energy savings estimates. Here are some common options for you:

PRODUCT-LIST:

- Nest Learning Thermostat: The Nest learns your schedule and adjusts temperatures accordingly to save energy. It has a sleek, minimalist design and is compatible with most HVAC systems.
- Ecobee Smart Thermostat with Voice Control: Ecobee's thermostat works with Alexa and Google Assistant for voice control. It has a built-in sensor to detect occupancy and automatically adjusts temperatures.
- Honeywell Lyric T5 Smart Thermostat: The Lyric T5 connects to Wi-Fi for remote access and has geofencing to adjust temperatures when you're away. It's compatible with Apple.

USER: What is the installation process for a smart thermostat?

ASST: Installing a smart thermostat typically involves removing your existing thermostat, connecting the new thermostat to your HVAC system's wiring, and configuring it through the app or on-screen interface. Here are some key steps:

- Check compatibility - Make sure the smart thermostat is compatible with your HVAC system. Most work with forced air furnaces and central AC, but some have specific requirements.
- Gather tools - You'll need a screwdriver, wire strippers, and possibly a multimeter to check wiring. Some models come with an installation kit.
- Turn off power - Go to your circuit breaker box and turn off the power to your thermostat. You don't want to get shocked.
- Remove old thermostat - Take off the cover and remove the screws holding it in place. Disconnect the wires from the terminals. Make note of the wire labels.
- Install new base - Screw the new thermostat base to the wall in the same place as the old one. Make sure it's level.
- Connect wires - Refer to your smart thermostat.

USER: That looks like something I can do. What are your top-selling ones?

LLM-generated ASST: Here are some of the top-selling smart thermostats known for energy-saving features and ease of use:

PRODUCT-LIST:

- Nest Learning Thermostat
- Ecobee Smart Thermostat
- Honeywell Home Smart Thermostat
- Emerson Sensi Smart Thermostat
- Google Nest Thermostat

User Intent: *1. The customer is looking for a smart thermostat.*

2. The customer is looking for a smart thermostat with the newest energy-saving features.

Figure 7: Sample from the LLM-as-a-judge evaluation dataset.

```
Given an utterance by a
helpful assistant, generate a
user utterance which consists
of a question which elicits
the assistant utterance. The
question should be as specific
as possible.

Assistant Utterance: {utt}
User Utterance:
```

Figure 8: Prompt template for user utterance generation in Stage 3.

Algorithm 1 Document Graph traversal in Stage 1

Require: Directed, weighted document graph $G_{doc} = (V_{doc}, E_{doc})$, number of documents n, start vertex v_{start}, out-degree centrality $deg^{-}(v)$ for all $v \in V_{doc}$
Ensure: A set of n related documents D

$D \leftarrow \{v_{\text{start}}\}$
$v_{\text{current}} \leftarrow v_{\text{start}}$
for $k \leftarrow 2$ to n **do**
  Let $O_{v_{\text{current}}} = \{v_j \mid (v_{\text{current}} \rightarrow v_j) \in E_{doc}\}$
  if $O_{v_{\text{current}}} = \emptyset$ **then**
    break {No outgoing edges; cannot sample further}
  end if
  for all $v_j \in O_{v_{\text{current}}}$ **do**
    $p_j \leftarrow \frac{deg^{-}(v_j)}{\sum_{v_u \in O_{v_{\text{current}}}} deg^{-}(v_u)}$
  end for
  Sample v_{next} from $O_{v_{\text{current}}}$ according to probabilities $\{p_j\}$
  $D \leftarrow D \cup \{v_{\text{next}}\}$
  $v_{\text{current}} \leftarrow v_{\text{next}}$
end for
return D

Algorithm 2 Dialogue Graph traversal in Stage 2

Require: Directed dialogue graph $G_{dial} = (V_{dial}, E_{dial})$, number of utterances m, start vertex v_{start}, CR scores $r_{(u \rightarrow v)}$ for all edges $(u \rightarrow v) \in E_{dial}$
Ensure: A sequence of m Assistant utterances U

$U \leftarrow \{v_{\text{start}}\}$
$v_{\text{current}} \leftarrow v_{\text{start}}$
$V_{\text{visited}} \leftarrow \{v_{\text{start}}\}$
for $t \leftarrow 2$ to m **do**
  Let $O_{v_{\text{current}}} = \{v_j \mid (v_{\text{current}} \rightarrow v_j) \in E_{dial}\}$
  $O_{v_{\text{current}}} \leftarrow O_{v_{\text{current}}} \setminus V_{\text{visited}}$ {Prune previously visited vertices}
  if $O_{v_{\text{current}}} = \emptyset$ **then**
    break {No more valid next utterances}
  end if
  for all $v_j \in O_{v_{\text{current}}}$ **do**
    $r_j \leftarrow r_{(v_{\text{current}} \rightarrow v_j)}$
  end for
  Let $Z = \sum_{v_j \in O_{v_{\text{current}}}} r_j$
  for all $v_j \in O_{v_{\text{current}}}$ **do**
    $p_j \leftarrow \frac{r_j}{Z}$
  end for
  Sample v_{next} from $O_{v_{\text{current}}}$ with probability distribution $\{p_j\}$
  $U \leftarrow U \cup \{v_{\text{next}}\}$
  $V_{\text{visited}} \leftarrow V_{\text{visited}} \cup \{v_{\text{next}}\}$
  $v_{\text{current}} \leftarrow v_{\text{next}}$
end for
return U

Turn #1
USER: Who is Joe Biden? What are some highlights of his political career?
ASST: Joseph Robinette Biden Jr. is an American politician who has been the 46th and current president of the United States…

Turn #2
USER: Tell me more about Joe Biden's father. Did he encounter any financial difficulties??
ASST: Biden's father had been wealthy and the family purchased a home in the affluent Long Island suburb…

Turn #3
USER: What role did Gates' parents play in shaping his childhood, including their career aspirations for him, their influence on his upbringing, and their encouragement of competition?
ASST: When Gates was young his parents wanted him to pursue a career in law…

⋮

Turn #22
USER: When and where did a tragic car accident occur that left Joe Biden's wife and infant daughter dead, while their two sons were seriously injured?
ASST: A few weeks after Biden was elected senator, his wife Neilia and one-year-old daughter…

Turn #23
USER: What health conditions did George H.W. Bush experience after his hip replacement surgeries in 2000 and 2007? Did these conditions affect his mobility? If so, when did he start using a wheelchair?
ASST: In May 1991, "The New York Times" revealed that Bush had developed Graves' disease…

Turn #24
USER: What specific activities did George H.W. Bush engage in after leaving the White House, including where he lived and worked, and any notable publications or speaking engagements?
ASST: After leaving office, Bush and his wife built a retirement house in…

Turn #25
USER: Can you tell me about Bill Gates' early life, including his Boy Scout achievements and education?
ASST: Gates was born in Wichita, Kansas, the son of Isabel V. (née Goss) and Melville…

Turn #26
USER: Which universities have awarded Bill Gates with an Honorary Doctorate?
ASST: Gates was awarded an Honorary Doctorate …

Turn #27
USER: What council did Bill Gates join as a member during the COVID-19 pandemic?
ASST: During the COVID-19 pandemic, Gates participated as a member of the National Advisory Council for the COVID Collaborative…

Turn #28
USER: When did Biden test positive for Covid-19?
ASST: On July 21, 2022, Biden tested positive for COVID-19 with reportedly mild symptoms..

Turn #29
USER: How did Covid-19 impact Biden's inauguration?
ASST: Biden's inauguration was "a muted affair unlike any previous inauguration" due to COVID-19 precautions…

⋮

Figure 9: Excerpt of a *DocTalk* conversation based on 3 Wikipedia articles: 'Joe Biden" (in orange), "George H.W. Bush" (in blue), and "Bill Gates" (in green).

A.5 DocTalk Sample

Samples from *DocTalk* are provided in Figure 9 and Figure 10, where different colors represent utterances sourced from distinct documents.

A.6 CoQA Generation Sample

A sample response from CoQA evaluation is provided in Figure 11. Starting from the 15th turn, we provide the responses generated by each of the implemented baselines.

A.7 Conversational Reward Model

To finetune the CR model, we construct the training dataset using open-source multi-turn dialogue datasets, prioritizing information-seeking corpora like HybriDialogue (Nakamura et al., 2022), InScitic (Wu et al., 2023), and TopiOCQA (Adlakha et al., 2022). However, while semantically relevant, assistant utterances from these datasets are much shorter than the utterances in our dataset, typically only containing the direct answer to the user's question. Additionally, only a small portion of conversations in these datasets contain a sufficient number of dialogue turns. Hence, we supplement the information-seeking datasets with more general dialogue datasets such as ShareGPT[¶] and UltraChat (Ding et al., 2023). Additionally, we also include the Wizard of Wikipedia (WoW) corpus (Dinan et al., 2019), a knowledge-grounded open-domain dialogue dataset, which emphasizes conversational naturalness. The final training set includes 2000 conversations each from ShareGPT and UltraChat, and all suitable conversations from HybriDialogue, InScitic, WoW, and TopiOCQA (2240 in total). For evaluation, we extract 100 conversations from each dataset's test set. We use Mean Reciprocal Rank (MRR) to evaluate our CR model performance. MRR

[¶]The ShareGPT dataset: https://sharegpt.com/

Turn #1
USER: Where is the Toronto General Hospital located and what makes it notable?
ASST: The Toronto General Hospital (TGH) is a major teaching hospital in Toronto, Ontario, Canada and the flagship campus of University Health Network (UHN). It is located in the Discovery District of Downtown Toronto…

Turn #2
USER: What specific medical specialties are covered in the final two years of the Doctor of Medicine program at the University of Toronto, and what types of training are provided during these years?
ASST: The Doctor of Medicine program at the University of Toronto is a 4-year MD program with a total enrollment of about 850 students. The first two years are known as the preclerkship curriculum…

Turn #3
USER: Which global university rankings place the school highest for clinical medicine and surgery in the years 2023 and 2020?
ASST: In 2023 the school was ranked 4th in the world for clinical medicine and surgery by "U.S. News & World Report"…

Turn #4
USER: What specific medical procedures and transplants does Toronto General Hospital specialize in and what are some notable achievements in these areas?
ASST: The emergency department now treats 28,065 persons each year, while the hospital also houses the major transplantation service for Ontario…

Turn #5
USER: Which hospitals have merged to form the University Health Network (UHN) and when did these mergers occur?
ASST: A series of mergers over many years has resulted in the UHN in its current form…

Turn #6
USER: Who is the current patron of the hospital within the Canadian Royal Family?
ASST: Currently, Sophie, Duchess of Edinburgh, as a member of the Canadian Royal Family, is patron of the hospital.

Turn #7
USER: When and where was the York General Hospital first established, and what significant events occurred during its history before it moved to its current location?
ASST: The hospital started as a small shed in the old town and was used as a British Army military hospital during the War of 1812..

Turn #8
USER: Which Canadian hospital network is the largest in terms of research funding and has consistently been ranked as one of the top hospitals in the world?
ASST: University Health Network (UHN) is a public research and teaching hospital network in Toronto, Ontario, Canada. It is affiliated with the University of Toronto…

Turn #9
USER: Which medical school in the Greater Toronto Area has the most hospitals affiliated with it, specializing in tertiary care and research, and includes Biosafety level 3 facilities?
ASST: The Faculty of Medicine is also the only medical school in the Greater Toronto Area and operates a health network that comprises twelve teaching hospitals with significant emphasis on tertiary care, including medical treatment, research and advisory services…

Turn #10
USER: When and why did the university suspend the teaching program of its medical school and transfer duties to proprietary schools? What were the consequences of this decision for students seeking university degrees in medicine?
ASST: The university originally opened its medical school in 1843, providing instruction in medicine and medical sciences. In 1853, it suspended the school's teaching program and transferred teaching duties to the city's three proprietary schools…

⋮

Figure 10: First 10 turns of a *DocTalk* conversation based on 3 Wikipedia articles: "Toronto General Hospital" (in orange), "University of Toronto Faculty of Medicine" (in blue), and "United Health Network" (in green).

Text: Wiltshire (or) is a county in South West England with an area of . It is landlocked and borders the counties of Dorset, Somerset, Hampshire, Gloucestershire, Oxfordshire and Berkshire. The county town was originally Wilton, after which the county is named, but Wiltshire Council is now based in the county town of Trowbridge…
Q1: What is Wiltshire characterised by?
Ans1: its high downland and wide valleys
Q2: Where is it located?
Ans2: South West England
Q3: Is it landlocked?
Ans3: yes

⋮

Q12: Where is that one near?
Ans12: Mere.
Q13: How was the county written in the 9th century?
Ans13: Wiltunscir
Q14: What archaeology is it known for?
Ans14: pre-Roman
Q15: What are the most famous Neolithic sites in the UK?
Ans15: Stonehenge and Avebury
DocTalk: Stonehenge and Avebury are two of the most famous Neolithic sites in the UK.
–w/o Stage 2: stonehenge and avebury
–w/o Stage 2 & 3: Stonehenge and Avebury
DocTalk*: stonehenge and avbury
Plain Wiki: stonehenge and avebury
WikiDialog: A West Saxon nobleman who had seized the throne of Queen Saxburga.

Q16: What battle was fought in 675?
Ans16: The Battle of Bedwyn
DocTalk: The battle of Bedwyn
–w/o Stage 2: The battle of bedwyn
–w/o Stage 2 & 3: The outcome of the battle of Bedwyn was that the West Saxon nobleman Escuin was
DocTalk*: The battle of Bedwyn.
Plain Wiki: A West Saxon nobleman seized the throne of Queen Saxburga
WikiDialog: Wulfhere of Mercia

Q17: Who was it between?
Ans17: Escuin and King Wulfhere
DocTalk: Escuin, a West Saxon nobleman who had seized the throne of Queen Saxburga, and King Wulfhere of Mercia.
–w/o Stage 2: escuin and wulfhere
–w/o Stage 2 & 3: escuin
DocTalk*: The danes.
Plain Wiki: King Wulfhere of Mercia.
WikiDialog: the Danes

Figure 11: Samples of responses generated during the CoQA evaluation. Each example displays the outputs from each baseline model, starting from the 15th turn.

measures the effectiveness of ranking by calculating the average reciprocal rank of the first relevant result across all queries.

A.8 Human Evaluation

For the human evaluation, we engaged 5 native English-speaking annotators. Each annotator was presented with 50 conversations from *DocTalk* and 50 conversations from WikiDialog. The evaluation was conducted in person in two 1.5 hour sessions. The annotators evaluated the dialogues using a detailed questionnaire, provided in Fig. 12, designed to measure various aspects of conversational quality. First, they assessed the contextual relevance of the assistant's responses by identifying irrelevant replies and calculating a relevance score based on the proportion of relevant responses to the total number of responses. Next, they evaluated the specificity and precision of user utterances by determining the proportion of specific responses relative to the total. Grammatical accuracy and phrasing were also examined, with annotators counting user utterances containing errors or awkward phrasing and computing the proportion of grammatically accurate responses relative to the total. In addition, the naturalness and cohesiveness of each conversation were rated on a scale from 1 to 4. For *DocTalk* conversations, annotators examined topical relatedness by referencing Wikipedia article titles, categorizing the strength of thematic connections (on a scale from 1 to 4), and assessed the logical coherence of topic shifts by calculating the proportion of logical shifts out of all transitions. This multifaceted approach provided a comprehensive evaluation of conversational dynamics across both datasets.

Human Evaluation Questionnaire

1. Are the assistant's responses contextually relevant to the preceding user utterance?
Participants will be told to count the number of irrelevant responses. The final score for a given dialogue is calculated as follows: (total number of responses - number of irrelevant responses)/total number of responses.

2. Are the user utterances specific and precise?
Participants will be told to count the number of generic and vague responses. The final score for a given dialogue is calculated as follows: (total number of user utterances - number of generic/imprecise user utterances)/total number of user utterances.

***3.* Do the user utterances contain grammatical errors or awkward phrasing that disrupt understanding?**
Participants will be told to count the number of utterances with grammatical errors and awkward phrasing. The final score for a given dialogue is calculated as follows: (total number of user utterances - number of erroneous/awkward user utterances)/total number of user utterances.

4. Does the conversation feel natural and cohesive, resembling an information seeking conversation between a human and a LLM?
Participants will be instructed to choose the option that best characterizes each dialogue. The final score will be determined by averaging the scores assigned to each option (e.g., option 4 is scored as 4, option 3 as 3, and so on) across all dialogues and then normalizing the result.

- 4: The conversation is cohesive and natural. *The dialogue flows smoothly, with responses aligning well with the questions asked. It feels like a comfortable, intuitive exchange between a human and an LLM, resembling a real-life information-seeking conversation.*
- 3: The conversation is largely natural and cohesive. *The exchange is mostly smooth and logical, with minor hiccups or slight deviations that don't significantly disrupt the flow. It still feels like a productive conversation, though it may lack perfect polish.*
- 2: The conversation is largely unnatural and fragmented. *The dialogue feels disjointed, with responses that don't fully connect to the questions or follow a clear thread. It struggles to maintain a natural tone, making it harder to resemble a typical human-LLM interaction.*
- 1: The conversation feels unnatural and fragmented. *The exchange is highly disjointed and awkward, with little coherence or flow. Responses may seem irrelevant or off-topic, making it difficult to recognise as an information-seeking conversation.*

5. Are the topics covered in this conversation related to one another?
Participants will be told to refer to the wikipedia article titles, and instructed to choose the option that best characterizes each dialogue. The final score will be determined by averaging the scores assigned to each option (e.g., option 4 is scored as 4, option 3 as 3, and so on) across all dialogues and then normalizing the result. Not applicable to WikiDialog.

- 4: All topics are closely related and form a clear thematic connection. *Every topic ties directly into a central theme or purpose, creating a unified and focused conversation. The topics are intentional and logically connected.*
- 3: All topics are somewhat related with some thematic connection. *The topics share a general thread or overlap in intent, though the connections might be looser or less obvious.*
- 2: Some topics are unrelated and thematically inconsistent. *While a few topics might align, others diverge significantly, breaking the thematic flow.*
- 1: All topics are unrelated and thematically inconsistent. *The discussion lacks any unifying theme, with topics jumping between unrelated ideas. The topics appear random and disjointed.*

6. Are the topical shifts logical and coherent within the context of the conversation?
Participants will be told to count the number of illogical/incoherent topic shifts. The final score for a given dialogue is calculated as follows: (total number of topic shifts - number of illogical/incoherent topic shifts)/total number of topic shifts. Not applicable to WikiDialog.

Figure 12: Questionnaire used for human evaluation. The questionnaire assesses dialogue quality through six criteria: Q1. contextual relevance of the assistant's responses, scored by counting irrelevant responses; Q2. specificity and precision of user utterances, measured by tallying generic/vague instances; Q3. grammatical correctness and phrasing clarity, evaluated by counting errors or awkwardness; Q4. naturalness and overall resemblance to human-LLM conversation, rated on a 4-point scale; Q5. thematic consistency of topics, scored on a 4-point scale, with reference to Wikipedia articles (not applicable to WikiDialog); and Q6. logical coherence of topic shifts, calculated by counting illogical transitions.