Comprehensive Comparison of RAG Methods Across Multi-Domain Conversational QA

Klejda Alushi, Jan Strich, Chris Biemann, Martin Semmann
Hub of Computing and Data Science (HCDS)
University of Hamburg, Germany

Correspondence: `{first_name}.{last_name}@uni-hamburg.de`

Abstract

Conversational question answering increasingly relies on retrieval-augmented generation (RAG) to ground large language models (LLMs) in external knowledge. Yet, most existing studies evaluate RAG methods in isolation and primarily focus on single-turn settings. This paper addresses the lack of a systematic comparison of RAG methods for multi-turn conversational QA, where dialogue history, coreference, and shifting user intent substantially complicate retrieval. We present a comprehensive empirical study of vanilla and advanced RAG methods across eight diverse conversational QA datasets spanning multiple domains. Using a unified experimental setup, we evaluate retrieval quality and answer generation using generator and retrieval metrics, and analyze how performance evolves across conversation turns. Our results show that robust yet straightforward methods, such as reranking, hybrid BM25, and HyDE, consistently outperform vanilla RAG. In contrast, several advanced techniques fail to yield gains and can even degrade performance below the No-RAG baseline. We further demonstrate that dataset characteristics and dialogue length strongly influence retrieval effectiveness, explaining why no single RAG strategy dominates across settings. Overall, our findings indicate that effective conversational RAG depends less on method complexity than on alignment between the retrieval strategy and the dataset structure. We publish the code used.[1]

1 Introduction

Conversational search, the task of satisfying information needs through multi-turn dialogue, has gained significant traction due to recent advances in LLMs (Mo et al., 2025). The field is shifting from traditional keyword-based queries to conversational search, characterized by multi-turn natural-language interactions that capture complex and evolving information needs (Mo et al., 2025; Prayitno et al., 2025). To support these interactions, RAG has emerged as the de facto standard (Lewis et al., 2020; Huang and Huang, 2024; Nikishina et al., 2025). By retrieving external evidence from vector databases, RAG mitigates hallucinations and ensures responses are factually grounded and up-to-date (Shuster et al., 2021; Sahoo et al., 2024).

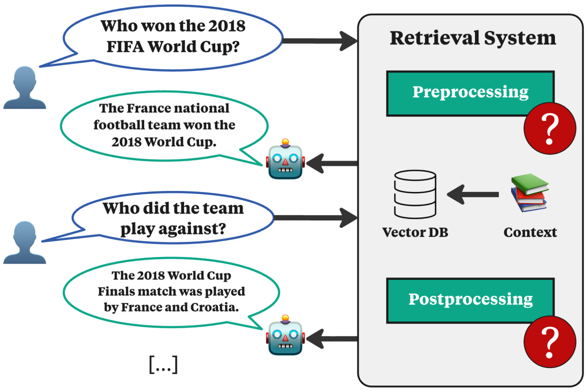
Who won the 2018 FIFA World Cup?
The France national football team won the 2018 World Cup.
Who did the team play against?
The 2018 World Cup Finals match was played by France and Croatia.
[...]
Retrieval System
Preprocessing
?
Vector DB
Context
Postprocessing
?

Figure 1: Conversational Search Problem. One sample from the INSCIT dataset (Wu et al., 2023).

While RAG is well established for single-turn Question Answering (QA), effectively integrating external knowledge into multi-turn conversations introduces significant complexity. In this setting, the system must maintain context across the dialogue history, resolving coreference and handling implicit queries when information is omitted (e.g., ellipsis). Consequently, retrieval effectiveness can vary widely depending on the system's ability to track dialogue history, resolve ambiguity, and adapt to shifting user intent across turns (Saha Roy et al., 2025; Chang et al., 2025; Zhang et al., 2025a). Although the literature reports a rapid evolution of advanced RAG architectures (Gao et al., 2023b; Huang and Huang, 2024) and optimization strategies (Gao et al., 2021, 2023a), these methods are typically evaluated in isolation (Yu et al., 2025).

[1] GitHub Repository

Currently, the field lacks a comprehensive overview of RAG strategies for conversational settings. Existing studies often utilize only vanilla RAG (Liu et al., 2024; Xu et al., 2025), lacking SOTA retrieval metrics, limiting reproducibility, and practical insights for RAG in production systems. Furthermore, the interplay between retrieval performance and the depth of the conversation, specifically how performance degrades or changes as the dialogue progresses, remains underexplored.

To address this gap, we present an empirical analysis of RAG methods for conversational QA. Our contributions are as follows:

- We provide a unified comparison of vanilla RAG and **six advanced RAG methods** under a reliable evaluation on **eight conversational QA datasets.**
- We further analyze the influence of the **position of the conversational turn** on retrieval performance.

2 Related Work

LLMs and Conversational QA. Adapting LLMs for Conversational QA is typically achieved via fine-tuning a pre-trained model or by incorporating external context via RAG (Dhabalia et al., 2025). Instruction tuning, which aligns pre-trained LLMs with conversational instructions, has become a foundational approach (Zhang et al., 2025b). These models have been successfully adapted for diverse domains, ranging from general knowledge (Yang et al., 2018; Joshi et al., 2017) to specialized fields such as medicine (Li et al., 2023; Prayitno et al., 2025) and law (Wu and Ma, 2025). Fine-tuning strategies often involve training on human-rewritten queries (Mo et al., 2023) or converting multi-turn interactions into single-turn problems (Ye et al., 2023).

RAG and Conversational QA. Conversational QA benefits significantly from RAG (Lewis et al., 2020), which integrates additional context by retrieving semantically similar documents from a vector database. Recent work has enhanced this process by incorporating meta-information (Saha Roy et al., 2025) and query rewriting to facilitate accurate generation (Mo et al., 2023). Further advancements include self-check mechanisms, in which the model assesses the correctness of its own answers (Ye et al., 2024), and learning policies that determine when and what to retrieve (Roy et al., 2024). Beyond these conversation-specific adaptations, the broader RAG landscape offers numerous state-of-the-art methodologies (Gao et al., 2023b; Huang and Huang, 2024) that hold potential for this domain. However, existing research often focuses on the end-to-end performance of one method, rather than comparing promising methods from the literature (Gao et al., 2021; Tito et al., 2021; Gao et al., 2023a). Consequently, this work presents a comprehensive comparison of these advanced retrieval strategies within conversational QA.

Conversational QA Datasets. Conversational QA has evolved from single-turn tasks (e.g., SQuAD (Rajpurkar et al., 2016)) to complex multi-turn settings. While datasets such as HotPotQA (Yang et al., 2018) and 2WikiMultiHopQA (Ho et al., 2020) emphasize multi-hop reasoning, contemporary benchmarks have shifted their focus toward retrieval and conversational fidelity. PopQA (Mallen et al., 2023) addresses long-tail knowledge retrieval, while ChatQA (Liu et al., 2024) and ChatQA-2 (Xu et al., 2025) establish a new standard for conversational QA by evaluating models on their ability to reason over retrieved evidence within fluid, multi-turn dialogues. Building on this foundation, we leverage the ChatQA (Liu et al., 2024) dataset to systematically analyze how distinct RAG strategies perform across different knowledge domains.

3 RAG Methods

Without RAG. We establish two reference points: *No RAG*, which sends queries directly to the LLM using only the conversation history, and *Oracle Context*, which provides ground-truth contexts. *No RAG* measures the LLM's internal capabilities through pretraining and sets dataset-specific baselines, whereas *Oracle Context* simulates perfect retrieval to define a ceiling for the generator.

Basic RAG Methods. This category comprises methods that retrieve documents using standard embedding-based retrieval techniques. The *Base RAG* approach follows the original RAG framework (Lewis et al., 2020), in which only the input query is embedded to retrieve the top-k documents, which are then passed unchanged to the generator. As a purely lexical baseline, *Standard BM25* (Robertson and Zaragoza, 2009) ranks documents based on term-frequency and inverse document-frequency statistics, relying on keyword overlap between the query and documents. *Hybrid BM25* (Gao et al., 2021) combines sparse BM25 retrieval with dense vector retrieval, leveraging the complementary strengths of lexical and semantic matching to improve recall and relevance. Finally, the *Reranker* method (Glass et al., 2022) applies a cross-encoder after initial retrieval to reorder documents according to their significance in a shared embedding space.

Advanced RAG Methods. This category encompasses methods that enhance retrieval through either *preprocessing* or *postprocessing* strategies. Preprocessing methods modify the input query to improve retrieval quality. The *HyDE* method (Gao et al., 2023a) generates hypothetical answers for each query and uses them as refined queries to retrieve more relevant documents. In contrast, *Query Rewriting* (Ye et al., 2023) reformulates the original query to better align with the target document distribution. Postprocessing methods operate on retrieved contexts to improve their usefulness for generation. *Summarization* reduces contextual noise by condensing each retrieved document using an LLM, focusing on salient information. *SumContext* applies a similar summarization step while retaining the original full documents for generation, aiming to reduce distractions while preserving content fidelity. Finally, the *HyDE Reranker* performs post-retrieval reranking by leveraging hypothetical answer generation to reorder initially retrieved documents based on semantic alignment.

4 Datasets

For our evaluation, we use ChatRAG-Bench (Liu et al., 2024), a benchmark comprising 10 conversational QA datasets covering diverse topics and formats. We select eight of these subsets for our experiments, as detailed in Table 1. We exclude HybridDial (Nakamura et al., 2022) due to the lack of annotated ground-truth contexts, which renders accurate retriever evaluation infeasible. Additionally, we omit ConvFinQA (Chen et al., 2022) because the answers primarily consist of numerical results derived from simple arithmetic operations, making the F1 score an unreliable metric for performance. The remaining eight subsets contain question-context-answer triples, enabling the independent evaluation of both retriever and generator components. Data preprocessing involved normalizing datasets from Hugging Face to a standardized format: we serialized multi-turn dialogues into a linear text format and removed formatting artifacts from context passages. We describe each dataset in detail below.

Sequential Question Answering (SQA) SQA (Iyyer et al., 2017) is a conversational dataset that focuses on a QA dialogue regarding semi-structured, Wikipedia tables. The aim was to decompose long, complex questions into sequences of small, easy-to-answer sub-questions. WikiTableQuestions (Ho et al., 2020) was used to source the original questions, filtering out those that required arithmetic or could not be answered directly from table cells.

Question Answering in Context (QuAC) The QuAC (Choi et al., 2018) dataset focuses on QA-based conversations between a *teacher* and a *student* regarding a Wikipedia article about an entity. The student knows only the article title, whereas the *teacher* has full access; instead of answering the question in free text, they can only reply with a text excerpt. The interaction continues until one of three outcomes is reached: 12 questions have been answered, two questions remain unanswered, or either the student or the teacher decides to end the dialogue.

Conversations Question Answering (CoQA) Reddy et al. (2019) introduced the CoQA dataset to include diverse data sources, such as children's literature, school exams, and news, as well as casual-sounding speech, in which questions often refer to the dialogue history and answers are direct and without explanation. Each question may have multiple correct answers to account for grammatical or formatting differences. The evaluation is performed between the generated answer and each reference answer, and only the reference with the highest F1 score is selected.

Domain-specific Question Answering (DoQA) DoQA (Campos et al., 2020) is a dataset built upon a continuous dialogue between a *user* and a *domain expert* relating to a specific topic, in this case cooking, travel, or movie forums on Stack Exchange. DoQA aimed for a more natural conversation, based more on follow-up questions than clunky factoids, and, similar to CoQA, there are four correct answers for each question.

Doc2Dial Doc2Dial (Feng et al., 2020), a dataset consisting of dialogues between a *user* and an *agent* on topics related to social welfare in the United

Subset	Source	# Contexts	# QA Pairs	Ctx/Q Ratio	Avg. Tokens		
					Question	Answer	Context
QuAC	Wikipedia	26,315	7,350	3.58	8.81	19.91	511.42
SQA	Wikipedia	185	3,010	0.06	10.69	37.26	453.83
QReCC	Wikipedia	19,275	2,790	6.91	8.12	28.16	505.52
TopiOCQA	Wikipedia	169,231	2,510	67.42	9.12	17.30	97.12
Doc2Dial	Social Welfare	1,238	3,940	0.31	12.52	22.77	350.38
DoQA	StackExchange	395	1,790	0.22	13.16	19.00	145.50
CoQA	Mixed	499	7,980	0.06	7.69	4.43	329.38
INSCIT	Mixed	29,497	502	58.76	12.23	45.32	101.17
Total | W. Avg	Multiple	**246,635**	**29,872**	8.25	9.47	18.82	175.81

Table 1: Summary of the QA datasets, including source, number of contexts and QA pairs, context-to-question ratio, and average token lengths for questions, answers, and contexts. Weighted averages are computed across all subsets, using the number of QA pairs and contexts. Avg. token count based on Llama 3.3 tokenizer.

States as found on *ssa.gov* and *va.gov*. To showcase both dialogue- and document-based contexts, the conversations were sorted into three different categories: *D1*, containing multiple questions relating to the given context, *D2*, in which the conversation revolves around one central inquiry with clarifying questions carried out by the agent, and *D3*, with questions that are irrelevant to the context.

Question Rewriting in Conversational Context (QReCC) The QReCC (Ye et al., 2023) dataset incorporates questions from pre-existing datasets, including QuAC, and includes information sourced from the Common Crawl and web searches. Question rewriting was also used to "fix" any inquiries that had references to the conversation history, thereby preserving the natural-sounding sentence structure while removing ambiguity. These methods involve *replacing* pronouns with their explicit referent, *inserting* the referenced entity into the query itself, and *removing* any unnecessary words.

Topic switching in Open-domain Conversational Question Answering (TopiOCQA) TopiOCQA (Adlakha et al., 2022) focuses on topic switching during a free-form conversation between a *questioner* and an *answerer*. The *answerer* is granted full access to links within the relevant Wikipedia article. In contrast, the *questioner* may only view the metadata and adjust their inquiries accordingly. The questions were divided into general open-ended questions, inquiries about specific entities, and requests for further details. The answers were unrestricted and free-form, facilitating topic switches every 3-4 questions and enabling handling of changing contexts and long-term reasoning.

Information-Seeking Conversations with Mixed-Initiative Interactions (INSCIT) Wu et al. (2023) proposed INSCIT, an information-seeking dataset that would use a variety of interactions between a *user* and an *agent* to challenge hard-to-answer questions regarding Wikipedia passages. It aimed to provide answer structures, categorizing them into: *Direct answers*, with the *agent* providing what they believe is the correct answer, *Relevant answers*, which inform the user of relevant information regarding the query, *Clarifications*, which prompt the user for further information, and *No information*, if not relevant answer is found.

4.1 Dataset Statistics

In total, our analysis encompasses over **29,000** QA pairs and more than **245,000** distinct contexts across multiple domains. As shown in Table 1, the datasets exhibit significant structural variation. The ratio of context tokens to query tokens varies considerably across the different subsets, whereas the average question length remains relatively balanced (7–12 tokens). In contrast, answer and context lengths exhibit substantial variance: average answer lengths range from 4 tokens (CoQA) to 45 tokens (INSCIT), whereas average context lengths range from approximately 100 to 500 tokens. Table 1 presents the answer lengths, which vary considerably, ranging from the more concise answers in the CoQA dataset to the lengthier explanations in INSCIT, which were already accounted for in the system prompts. While longer contexts risk containing more distracting content and reducing generator performance, shorter contexts can be disadvantageous if they do not provide sufficient con-

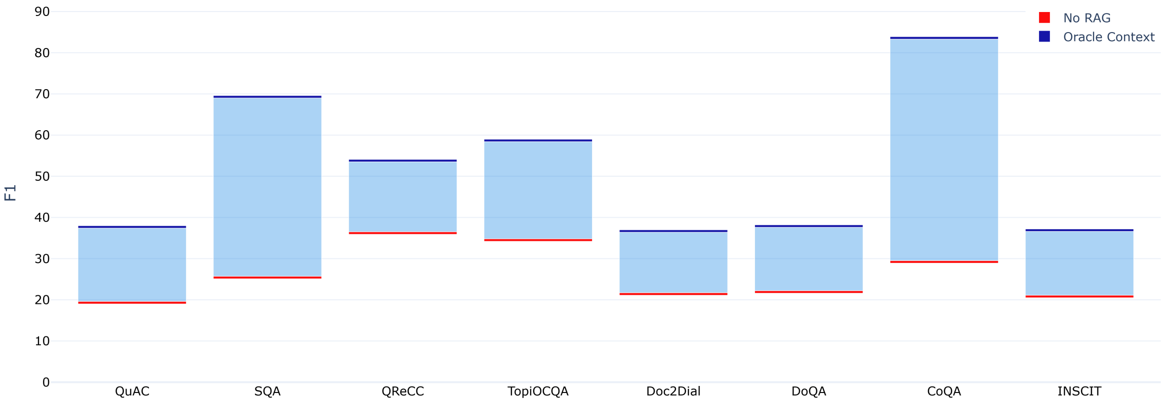

Figure 2: Theoretical minimum and maximum ranges of F1 that can be achieved with the LLM. Minimum is achieved by *No RAG* method, which retrieves no contexts, whereas the maximum is achieved by the *Oracle Context* method, which directly uses the gold label context.

text. A larger number of total contexts, as in the QuAC, INSCIT, and TopiOCQA datasets, could complicate or delay context retrieval.

4.2 Dataset Analysis

To assess whether the datasets are suitable for analysis and whether they theoretically benefit from RAG, we conducted a pre-study in which we queried each dataset with the model using no information and with all information about the query. We evaluate two control settings to disentangle pre-trained knowledge from the effect of context: *No RAG*, which queries the LLM without external context, and *Oracle Context*, which provides only the ground-truth context and serves as an upper bound, as shown in Figure 2. We define the *Oracle Context* setup as the upper bound of the model, given the generator's ability to answer the question using the golden context. In contrast, the lower bound is the LLM's performance without any context, independent of the pre-trained model's knowledge.

For most datasets, the ceiling F1 remains below 40%, with overall values around 15–20%, indicating room for improvements through retriever methods alongside additional performance bottlenecks. CoQA and SQA exhibit wider F1 ranges, facilitating more precise comparisons between retrieval methods. Furthermore, except for QReCC and TopiOCQA, the LLM displays comparable internal knowledge across datasets. Because LLMs are trained on public data, the *No RAG* setting provides a valuable proxy for assessing overlap between the dataset and pre-training or fine-tuning data.

5 Evaluation and Results

This section outlines the experiments we conducted to investigate the effect of RAG methods on multi-domain conversational QA. Firstly, we describe the experimental setups, prompt design, and evaluation metrics used in the assessment in Sections 5.1 to 5.3. This is followed by discussing the results for the retriever and generator, and followed by the analysis of the relation of retriever and generator performance in Sections 5.4 to 5.6. We further analyze the effect of the conversation turn and discuss the results in Sections 5.7 and 5.8.

5.1 Experimental Setup

We evaluated all advanced RAG methods across all eight datasets using the EncouRAGe (Strich et al., 2025) library with the Llama 3 8B Instruct model (Grattafiori et al., 2024), selected for its strong language understanding and manageable computational requirements. Additionally, the results of Gemma 3 27b (Kamath et al., 2025) were added in the camera-ready version in Appendix B and align with all results. Key parameters were set to ensure reproducibility and efficiency: temperature = 0, maximum output length = 1000 tokens, and context length = 40,000 tokens. We conducted multiple runs for each method but observed only negligible differences across runs, so we reported results for only one run per method and dataset. Inference was performed on an NVIDIA RTX A6000 GPU with 48 GB of memory. EncouRAGe (Strich et al., 2025) facilitated dataset and RAG method management, integrating vLLM

RAG Method	QuAC		SQA		QReCC		TopiOCQA		Doc2Dial		DoQA		CoQA		INSCIT	
	Wikipedia								Social Welfare		StackExchange		Mixed			
	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1
No RAG	-	19.0	-	25.1	-	35.9	-	34.2	-	21.1	-	21.6	-	28.9	-	20.5
Oracle Context	100	38.0	100	69.6	100	54.1	100	59.0	100	37.0	100	38.2	100	83.9	100	37.2
Vanilla RAG	35.3	25.4	66.1	43.8	36.6	36.5	8.7	33.9	49.0	26.3	92.0	36.2	72.5	58.3	8.0	19.2
Hybrid BM25	**44.6**	27.5	53.8	45.6	37.4	37.4	9.4	34.0	51.0	27.7	84.8	**36.9**	76.6	67.3	8.1	19.1
Reranker	41.6	**29.2**	**71.0**	**51.3**	35.3	36.0	9.8	34.2	54.0	**28.7**	**93.1**	36.7	78.9	**74.7**	9.5	19.1
Query Rewriting	35.3	20.8	66.1	38.1	36.4	32.5	8.7	26.9	49.0	21.6	92.0	27.6	72.4	47.5	8.0	16.7
HyDE	42.0	26.2	65.3	38.0	**49.0**	**42.2**	**25.1**	**43.5**	**57.5**	**28.7**	90.9	35.4	**86.6**	71.3	**25.2**	**25.9**
HyDE + Reranker	30.5	25.4	61.6	38.4	37.5	37.0	14.6	37.6	48.0	26.4	85.2	34.7	58.2	53.0	13.9	22.0
Summarization	38.6	24.3	57.8	34.9	38.3	39.2	6.1	27.9	44.3	25.6	84.5	29.4	67.6	48.6	7.4	18.4
SumContext	37.7	26.3	57.7	35.1	40.0	37.1	6.7	27.5	44.9	26.7	85.3	35.4	68.0	62.7	7.4	18.5

Table 2: Overall performance (MRR@5 and F1) of RAG methods on all eight conversational QA datasets. MRR@5 is used for retrieval performance and F1 for the generator. **Bold** values indicate the maximum for each column.

(Kwon et al., 2023) for efficient batched inference and MLflow for experiment tracking. External contexts were stored in the Chroma vector database (Chroma Team, 2025) using Sentence Transformers `all-MiniLM-L6-v2` (Reimers and Gurevych, 2020) embeddings and Cosine Similarity for semantic relevance, ensuring efficient retrieval and evaluation across all methods and datasets.

5.2 Prompt Design

Our prompt strategy relies on a consistent zero-shot template, following ChatQA (Liu et al., 2024). The general system prompt instructs the model to prioritize retrieved context and dialogue history, strictly avoiding reliance on internal knowledge to reduce hallucinations. Variations in the prompt were limited to formatting constraints (e.g., extraction vs. generation) to match specific dataset targets; the full set of dataset-specific prompt templates is provided in Appendix A.

5.3 Evaluation Metrics

For generators, we considered using F1, with F1 adopted as the primary metric to align with prior ChatRAG-Bench studies (Liu et al., 2024). We selected the F1 Score from the SQuAD paper (Rajpurkar et al., 2016) to balance token-level precision and recall, and, for datasets with multiple valid answers, we used the maximum score across references. Retriever performance was assessed using Recall@k, indicating whether the ground-truth context appears in the top-k retrievals, and Mean Reciprocal Rank (MRR) (Kantor and Voorhees, 2000) for $k = 5$, which emphasizes correct contexts ranked higher. These metrics together capture both the accuracy and ranking quality of retrieval, facilitating fair comparison across RAG methods.

5.4 Retrieval Results

Table 2 presents the results of the generator and retriever for each of the RAG methods. We also report Recall@*1* and Recall@*5* for each approach in Appendix C. In terms of MRR@*5* performance, for all datasets except for SQA (Iyyer et al., 2017) and DoQA (Campos et al., 2020), the combination of sparse and dense encoders in *Hybrid BM25* leads to better results than *Vanilla RAG*. This effect is also visible for the *Reranker*.

Among the advanced RAG methods, *HyDE* ranks clearly ahead, achieving the best performance on five of eight datasets. It is important to highlight that for INSCIT, *HyDE* triples the performance of *Vanilla RAG*. The two summarization methods yielded poor retrievals, suggesting that summarization may remove crucial contextual information. Dataset-wise, the MRR@*5* values were highest in the DoQA and CoQA datasets and were significantly lower in INSCIT and TopiOCQA, which are analyzed further in Section 5.7.

5.5 Generator Results

Table 2 shows that *Hybrid BM25* consistently achieves slightly higher F1 scores than *Vanilla RAG* across all datasets. Consistent with the retriever results, *HyDE* emerges as the strongest approach, attaining the highest performance on four of the eight datasets and highlighting its effectiveness for conversational QA. This advantage is particularly evident on TopiOCQA, where *HyDE* improves the F1 score by 9.6% over *Vanilla RAG*. In contrast, the performance of *Query Rewriting* is highly dataset-dependent and falls substantially below *No RAG* on the INSCIT, QReCC, and TopiOCQA datasets, where MRR scores are also below average. This may point to differences in conversational struc-

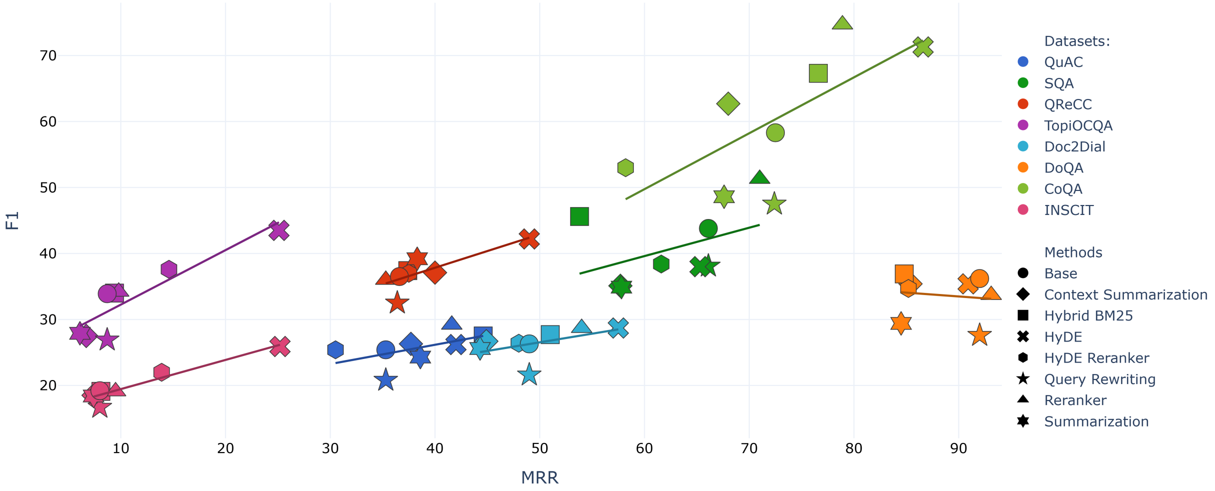

Figure 3: Relationship between retriever (MRR@5) and generator (F1) performance for each dataset and method.

ture, such as larger topic shifts or longer dependency chains, that render these datasets less effective for this method. Finally, for summarization methods, incorporating the original conversational context yields only marginal improvements except for DoQA and CoQA.

5.6 Relationship of Retriever and Generator

Figure 3 illustrates the relationship between retrieval and generation, with a fitted linear regression line summarizing the overall performance trend. Overall, F1 and MRR are positively correlated, indicating that stronger retrieval generally leads to higher answer quality, particularly on INSCIT, TopiOCQA, CoQA, and SQA, where *HyDE* and *HyDE Ranker* follow this trend. For CoQA and SQA, *Reranker* appears to perform best, yielding the highest overall results on these datasets.

In contrast, QuAC and Doc2Dial show only marginal differences across methods, and the correlation is weaker. This disconnect is most pronounced for DoQA, where MRR exceeds 90% but F1 remains below 40%, highlighting that strong retrieval does not necessarily translate into strong generation.

Furthermore, Spearman's ρ in Table 3 illustrates that most datasets have a relatively high correlation between the F1 and MRR values. The only exceptions being SQA and DoQA, which show weak or even negative correlations, suggesting that the two metrics capture different aspects of method performance.

These differences can be attributed to dataset specific problems, particularly relating to variations in answer formats. Answer conciseness is associated with overall high F1 scores, particularly when comparing CoQA's short responses against QReCC and INSCIT's longer, more in-depth answers, which are more challenging to match.

Subset	ρ
QuAC	0.620
SQA	0.383
QReCC	0.857
TopiOCQA	0.874
Doc2Dial	0.687
DoQA	-0.157
CoQA	0.762
INSCIT	0.788

Table 3: Illustration of the Spearman's rank correlation coefficient (ρ), calculated for the F1 and MRR values for each dataset.

The answer content can also heavily influence the performance, such as the social welfare answers of Doc2Dial, which rely on specific case details (user eligibility, personal details) and predetermined scripts, with many answers taking the form of follow-up questions requesting further information. Similarly, the forum-based DoQA dataset contains many informal responses in which respondents draw on personal knowledge rather than relying solely on the provided context. Additionally, the sensitive nature of certain DoQA queries leads to the generator refusing to respond, on the grounds that it cannot provide legal advice or discuss topics relating to violence or weaponry.

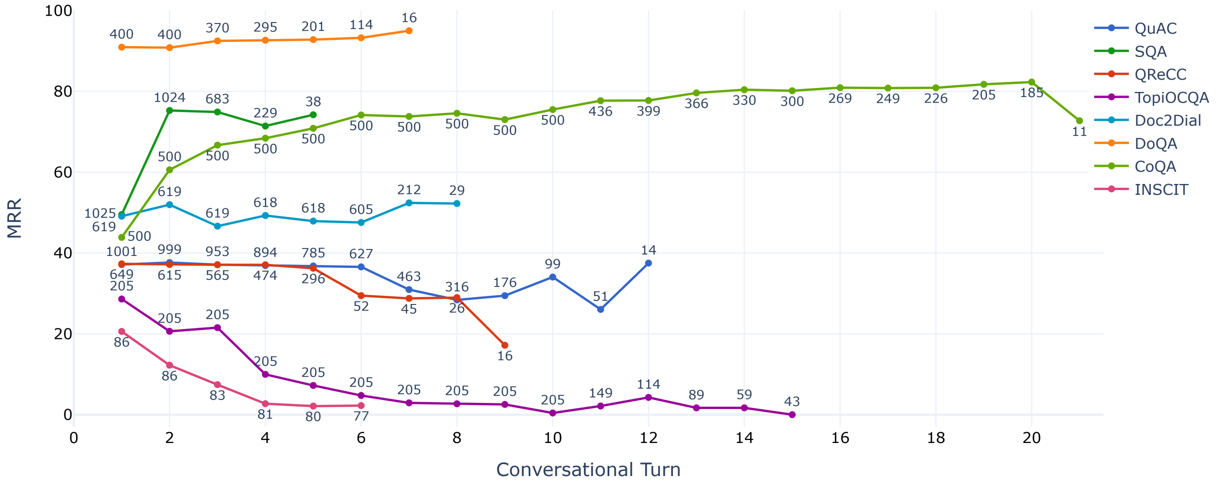

Figure 4: MRR performance across conversational turn for each dataset using *Vanilla RAG*. The annotations indicate the number of samples per turn for each dataset.

5.7 Ablation Study

To further understand the effect of RAG on conversational QA, we examine retrieval performance across conversational turns in Figure 4. We compute MRR and F1 Score by turn position to identify trends in retrieval performance and present the F1 results in Appendix D. We find mixed results for all eight datasets. It is expected that INSCIT and TopiOCQA exhibit low performance that steadily decreases with the number of turns, primarily due to the high Ctx/Q ratio, as shown in Table 1. This generally leads to low retrieval performance, and both datasets are designed to support topic and interaction entity switching.

In contrast, CoQA and SQA benefit from more context and improve performance with each turn, indicating that, when the context is consistent, more information leads to better retrieval performance. We found in addition that for QReCC, QuAC, DoQA, and Doc2Dial, all datasets seem to show no difference regarding the number of conversational turns til turn 5. There, we find that QReCC and QuAC show performance decreases that warrant further investigation in future work.

Overall, the results suggested that when the context is consistent, retrieving it from the entire conversation is beneficial; however, when questions are topic-switching or when other entities are interacting, this approach can decrease retrieval performance. We therefore recommend that calculating the similarity between queries and conversational histories can benefit retrieval performance.

5.8 Discussion

This section examines whether conclusive findings were obtained and, if so, how they could inform future studies in Conversational QA or RAG. We showed in our preliminary analysis that adding ground-truth context boosted performance by 15–50%, providing a ceiling for RAG methods and a baseline for weaker ones. About half of the RAG methods performed on par with *No RAG*, showing that inefficient pipelines either retrieve irrelevant contexts or fail to rank the ground-truth context highly enough.

As observed in the experiments, the results varied considerably across datasets and RAG methods, making it difficult to draw a unified conclusion. For F1, the top three methods were *Reranker* (Glass et al., 2022), *Hybrid BM25* (Gao et al., 2021), and *HyDE* (Gao et al., 2023a), respectively, indicating that *Vanilla RAG* (Lewis et al., 2020) was clearly outperformed across all datasets. Therefore, in one respect, the two advanced methods are recommended as superior alternatives in terms of performance.

In contrast, it is essential to examine how the advanced RAG methods affect computational complexity and runtime overhead relative to *Vanilla RAG*. To calculate relevance scores, *Reranker* retrieves a larger number of candidate contexts, which are then passed through a cross-encoder. On the other hand, *Hybrid BM25* adds a sparse retrieval function to the existing dense retrieval. Overall, although both methods are computationally simple, the experiment runs indicate that the double-retrieval process and score calculation of

Hybrid BM25 result in slightly longer runtimes. While *Vanilla RAG* does not achieve the highest overall performance, its high F1 scores and straightforward implementation make it a worthwhile baseline. Since the advanced methods do not deviate substantially from the baseline computationally, they also provide a valuable reference for estimating the potential performance of other advanced RAG methods.

When examining the impact of dataset characteristics, the preliminary analysis revealed a substantial difference in the model's internal knowledge of specific dataset topics or formats. This is evident in the F1 range difference between Doc2Dial (Feng et al., 2020), centered on specialized, open-ended questions in comparison to the factoid-style questions of CoQA (Reddy et al., 2019). The total number of contexts significantly affected the retriever's performance by reducing the likelihood of finding the correct context.

6 Conclusion

In this paper, we presented a systematic empirical study of vanilla and advanced RAG methods across eight multi-turn conversational QA datasets. Our evaluation, which accounted for both retriever effectiveness and generator quality, revealed that robust yet straightforward techniques, such as *Reranker* and *Hybrid BM25*, consistently outperform *Vanilla RAG* across all evaluated domains. Among the advanced techniques studied, the HyDE method proved the most effective for enhancing retrieval performance, whereas the *Reranker* approach was most successful at improving final answer quality. For future research, the results highlight that performance improvements can be achieved without inflating the computational complexity, emphasizing the need to prioritize retrieval strategies over resource-intensive scaling.

Furthermore, our analysis of conversational turns demonstrated that the impact of dialogue depth varies substantially across datasets, reflecting their distinct structures. While some settings benefit from the accumulation of consistent dialogue history, others suffer from performance degradation as the conversation progresses, particularly when handling topic switching or shifts in user intent. These results suggest that the effectiveness of conversational RAG is determined less by the inherent complexity of the retrieval method than by the strategic alignment between the retrieval strategy and the dataset's specific structural characteristics. We conclude that future advances in the field should prioritize this alignment to ensure that external knowledge is integrated accurately and efficiently into multi-turn dialogues.

Limitations

Methodological and Dataset Heterogeneity. This study was hindered by the wide variety of RAG methods and datasets, which required extensive, dataset-specific preprocessing due to differing prompt formats, answer structures, and context representations. This heterogeneity increased experimental complexity and limited the depth of analysis across all method–dataset combinations. Future work could mitigate this issue by focusing on a smaller, more representative subset of methods and datasets, especially given the redundancy among many RAG enhancements.

Retrieval Challenges from Large and Fragmented Contexts. Another limitation arises from the large number of contexts per query, which makes retrieving the ground-truth passage difficult, particularly for TopiOCQA, QuAC, and INSCIT. This issue could be addressed by grouping contexts during pre-processing using metadata such as titles or by adopting iterative retrieval or generation strategies that increase the likelihood of retrieving relevant evidence and producing accurate answers.

Acknowledgement

This work is supported by the Genial4KMU project, Universität Hamburg, funded by BMBF (grant no. 16IS24044B).

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain Conversational Question Answering with Topic Switching. *Transactions of the Association for Computational Linguistics*.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - Accessing Domain-Specific FAQs via Conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314, Online. Association for Computational Linguistics.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu

Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. 2025. MAIN-RAG: Multi-Agent Filtering Retrieval-Augmented Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2607–2622, Vienna, Austria. Association for Computational Linguistics.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Chroma Team. 2025. Chroma: Open-Source Search And Retrieval Database For AI Applications.

Taksh Dhabalia, Samanyu Bhate, Kushagra Singh, Brandon Cerejo, and Dhananjay Bhagat. 2025. A Comparative Study of RAG and Fine-Tuned Transformer Models for Domain-Specific Chatbots. In *2025 International Conference on Intelligent and Cloud Computing (ICoICC)*, pages 1–6, Bhubaneswar, India.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. Doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021. Complement Lexical Retrieval Model With Semantic Residual Embeddings. In *European Conference on Information Retrieval*, pages 146–160. Springer.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023a. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023b. Retrieval-augmented Generation for Large Language Models: A Survey. *Preprint*, arXiv:2312.10997.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

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.

Yizheng Huang and Jimmy Huang. 2024. A Survey on Retrieval-Augmented Text Generation for Large Language Models. *Preprint*, arXiv:2404.10981.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based Neural Structured Learning for Sequential Question Answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, and et al. Mesnard. 2025. Gemma 3 Technical Report. *arXiv preprint*. ArXiv:2503.19786.

Paul Kantor and Ellen Voorhees. 2000. The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text. *Information Retrieval*, 2(2/3):165–176.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pages 611–626, Koblenz, Germany. Association for Computing Machinery.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for

Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, virtual.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Preprint*, arXiv:2303.14070.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024. ChatQA: Surpassing GPT-4 on Conversational QA and RAG. In *Advances in Neural Information Processing Systems*, volume 37, pages 15416–15459. Curran Associates, Inc.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2025. A Survey of Conversational Search. *Preprint*, arXiv:2410.15576.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. ConvGQR: Generative Query Reformulation for Conversational Search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4998–5012, Toronto, Canada. Association for Computational Linguistics.

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.

Irina Nikishina, Özge Sevgili, Mahei Manhai Li, Chris Biemann, and Martin Semmann. 2025. Creating a Taxonomy for Retrieval Augmented Generation Applications. *Preprint*, arXiv:2408.02854.

La Ode Muhammad Yudhy Prayitno, Annisa Nurfadilah, Septiyani Bayu Saudi, Widya Dwi Tsunami, and Adha Mashur Sajiah. 2025. Conversational Agent for Medical Question-Answering Using RAG and LLM. *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, 4(3):1894–1899.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, USA. The Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Nirmal Roy, Leonardo F. R. Ribeiro, Rexhina Blloshmi, and Kevin Small. 2024. Learning When to Retrieve, What to Rewrite, and How to Respond in Conversational QA. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10604–10625, Miami, Florida, USA. Association for Computational Linguistics.

Rishiraj Saha Roy, Joel Schlotthauer, Chris Hinze, Andreas Foltyn, Luzian Hahn, and Fabian Kuech. 2025. Evidence Contextualization and Counterfactual Attribution for Conversational QA over Heterogeneous Data with RAG Systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 1040–1043, Hannover Germany. ACM.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jan Strich, Adeline Scharfenberg, Chris Biemann, and Martin Semmann. 2025. EncouRAGe: Evaluating RAG Local, Fast, and Reliable. *Preprint*, arXiv:2511.04696.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2021. Document Collection Visual Question Answering. In *16th International Conference on Document Analysis and Recognition*, volume 12822 of *Lecture Notes in Computer Science*, pages 778–792, Lausanne, Switzerland. Springer.

Wenshe Wu and Ning Ma. 2025. A Study of Large Language Modeling for Legal Q&A Based on LoRA

Fine-Tuning. In *Proceedings of the 2nd Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, DEAI '25, pages 258–262, New York, NY, USA. Association for Computing Machinery.

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.

Peng Xu, Wei Ping, Xianchao Wu, Chejian Xu, Zihan Liu, Mohammad Shoeybi, and Bryan Catanzaro. 2025. ChatQA 2: Bridging the Gap to Proprietary LLMs in Long Context and RAG Capabilities. *Preprint*, arXiv:2407.14482.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing Conversational Search: Large Language Model-Aided Informative Query Rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5985–6006, Singapore. Association for Computational Linguistics.

Linhao Ye, Zhikai Lei, Jianghao Yin, Qin Chen, Jie Zhou, and Liang He. 2024. Boosting Conversational Question Answering with Fine-Grained Retrieval-Augmentation and Self-Check. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2301–2305, Washington DC USA. Association for Computing Machinery.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. Evaluation of Retrieval-Augmented Generation: A Survey. In *Big Data*, pages 102–120. Springer Nature Singapore.

Feiyuan Zhang, Dezhi Zhu, James Ming, Yilun Jin, Di Chai, Liu Yang, Han Tian, Zhaoxin Fan, and Kai Chen. 2025a. DH-RAG: A Dynamic Historical Context-Powered Retrieval-Augmented Generation Method for Multi-Turn Dialogue. *Preprint*, arXiv:2502.13847.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and Fei Wu. 2025b. Instruction Tuning for Large Language Models: A Survey. *ACM Comput. Surv.*

A Expanded Dataset Prompts

The system prompts are designed to instruct the LLM on how best to answer the question and to emphasize the focus that should be placed on the previous conversation history and contexts provided, and if they do not provide the answer then the LLM should indicate as such, instead of relying on internal information. A segment of the prompt is used to guide the LLM on how the answer should be formatted whether it be multiple sentences or short phrases.

```
CoQA: You are a helpful assistant who will try
to answer the following question to the best of
your abilities. Use only on the given context
and conversation history and do not use any
assumptions or external information. Make the
answers as direct as possible without using any
redundant information and without using full
sentences. Indicate if you cannot find the
answer based on the context.

Doc2Dial: You are a helpful assistant. Answer
the question strictly based on the given
context. Do not use prior knowledge, make
assumptions, or introduce any information not
present in the context. If the answer is
clearly stated, respond in a complete and
concise sentence. If the context does not
provide enough information, respond with a
relevant follow-up question to clarify the
user's intent.

DoQA: You are a helpful assistant trying to
answer the questions to the best of your
abilities. Use only the given context to answer
the question. and do not use any assumptions or
external information. Keep your answer
relevant, direct and in one sentence. Do not
explain the background, context or reasoning
behind the answer. Do not refer to the context
in your response. Indicate if you cannot find
the answer based on the context.

INSCIT: You are a helpful assistant. Answer the
question strictly based on the given context.
Do not use prior knowledge, make assumptions,
or include any information not present in the
context. Do not refer to the context in your
response. If the answer is not available, say
so clearly. Respond in one full and
complete sentence.

QReCC: You are a helpful assistant. Answer the
question strictly based on the given context.
Do not use prior knowledge, make assumptions,
or introduce any information not present in
the context. If the answer is not available,
clearly state that. Respond in a single, clear,
and complete sentence whenever possible.
```

```
INSCIT: You are a helpful assistant. Answer the
question strictly based on the given context.
Do not use prior knowledge, make assumptions,
or include any information not present in the
context. Do not refer to the context in your
response. If the answer is not available, say
so clearly. Respond in one full and
complete sentence.

QReCC: You are a helpful assistant. Answer the
question strictly based on the given context.
Do not use prior knowledge, make assumptions,
or introduce any information not present in
the context. If the answer is not available,
clearly state that. Respond in a single, clear,
and complete sentence whenever possible.
```

```
QuAC: You are a helpful assistant who will try
to answer the following question to the best of
your abilities. Use only the given context and
conversation history and do not use any
assumptions or external information. Keep your
answer short, direct and in one sentence. Do
not explain the background, context or
reasoning behind the answer. Indicate if you
cannot find the answer based on the context.

SQA: You are a helpful assistant. Use only the
given table and conversation history to answer
the question. Do not rely on outside knowledge
or make assumptions. Return the exact answer
from the table. Use brief phrases or values and
no full sentences.

TopiOCQA: You are a helpful assistant who will
try to answer the following question to the best
of your abilities. Use only the given context
and conversation history and do not use any
assumptions or external information. Make the
answers as direct as possible without using any
redundant information and without using full
sentences. Indicate if you cannot find the
answer based on the context.
```

Figure 5: List of the system prompts used for each dataset.

B Performance of Gemma 3 27b

RAG Method	**QuAC**		**SQA**		**QReCC**		**TopiOCQA**		**Doc2Dial**		**DoQA**		**CoQA**		**INSCIT**	
	Wikipedia								Social Welfare		StackExchange		Mixed			
	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1	MRR	F1
No RAG	-	21.8	-	30.1	-	31.5	-	25.2	-	21.1	-	28.5	-	17.2	-	13.8
Oracle Context	100	47.0	100	78.6	100	55.1	100	61.8	100	42.7	100.0	52.9	100	83.2	100	36.6
Vanilla RAG	35.3	30.4	66.1	51.0	36.6	31.5	8.7	25.2	49.0	28.6	**92.0**	50.8	72.5	57.3	8.0	14.7
Hybrid BM25	**44.7**	33.2	54.0	52.8	37.3	32.9	9.3	25.8	51.3	30.7	84.8	51.2	76.4	66.4	8.1	14.7
Reranker	43.7	**36.3**	**72.6**	**58.1**	34.9	29.9	9.7	26.7	55.4	31.3	93.6	**51.0**	**81.3**	**75.5**	9.6	15.3
Query Rewriting	35.4	24.8	66.1	46.7	36.5	33.4	8.7	19.3	49.0	25.5	**92.0**	37.8	72.6	51.6	8.0	14.2
HyDE	31.5	29.2	63.5	56.2	**47.7**	**47.0**	**30.5**	**46.5**	**56.3**	**35.8**	89.0	45.7	68.0	64.7	**29.4**	**26.9**
HyDE + Reranker	24.8	29.0	60.6	58.0	38.3	42.8	16.3	37.5	46.2	33.8	82.2	47.5	46.2	54.1	15.3	20.1
Summarization	37.8	24.8	63.5	55.7	39.4	34.6	7.5	29.1	34.6	23.6	85.3	33.2	70.9	38.9	6.7	16.4
SumContext	37.7	31.6	63.6	57.1	38.8	43.6	7.5	29.6	34.3	31.6	85.8	47.5	71.6	69.2	7.0	17.0

Table 4: Overall performance (MRR@5 and F1) of RAG methods on all eight conversational QA datasets using Gemma 3 27b (Kamath et al., 2025). MRR@5 is used for retrieval performance, and F1 for the generator. **Bold** values indicate the maximum for each column.

C Recall Retrieval Performance

RAG Method	**QuAC**		**SQA**		**QReCC**		**TopiOCQA**		**Doc2Dial**		**DoQA**		**CoQA**		**INSCIT**	
	Wikipedia								Social Welfare		StackExchange		Mixed			
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Vanilla RAG	24.9	52.6	58.2	79.2	25.0	56.0	5.4	14.2	36.0	70.1	88.3	97.2	66.2	81.7	3.8	15.7
Hybrid BM25	28.7	**77.1**	43.2	76.3	25.1	59.9	5.7	16.3	35.1	**78.4**	74.8	**98.4**	62.4	**96.4**	3.8	16.1
Reranker	28.4	63.4	**61.7**	**86.0**	24.6	53.2	6.6	15.4	40.4	75.1	**90.1**	97.1	**72.3**	88.9	6.0	15.5
Query Rewriting	24.9	52.6	58.2	79.2	25.0	56.1	5.4	14.2	36.0	70.1	88.3	97.2	66.3	81.7	3.8	15.7
HyDE	**30.4**	60.7	56.6	78.8	**35.4**	**71.9**	**18.4**	**37.4**	**44.2**	**78.4**	87.7	95.6	83.4	90.5	**16.3**	**40.8**
HyDE + Reranker	17.7	54.7	52.0	77.2	28.0	52.5	10.9	20.6	35.7	67.8	72.5	86.7	52.7	67.4	9.0	21.5
Summarization	27.7	56.6	49.2	72.3	26.8	58.9	4.0	10.2	31.4	65.1	80.6	92.1	58.7	79.9	4.0	13.3
SumContext	26.8	55.4	48.0	73.4	28.2	60.4	4.1	11.4	32.4	65.6	80.2	92.0	60.3	80.0	4.5	13.2

Table 5: Overall performance (R@1 and R@5) of RAG methods on all eight conversational QA datasets. **Bold** values indicate the maximum for each column.

D F1 Performance Across Conversational Turns

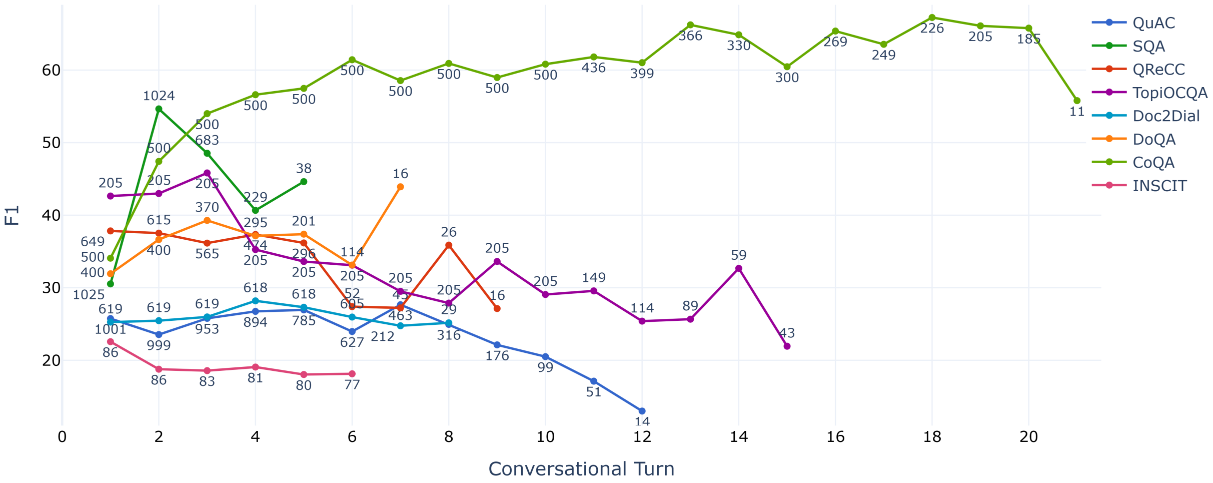

Figure 6: F1 performance across conversational turn for each dataset using *Vanilla RAG*.