Structured Dialogue Refinement: Building Retrieval-Augmented Question Answering on Goal-Oriented Dialogues

Bin Wu [1] * **Sawan Kumar** [2] **Prasetya Ajie Utama** [2] **Mohamed Yahya** [2]

[1] Centre for Artificial Intelligence, University College London
[2] Bloomberg
{bin.wu.23}@ucl.ac.uk
{skumar994, putama, myahya6}@bloomberg.net

Abstract

Retrieval-Augmented Generation (RAG) is widely used for question answering over well-structured document corpora. However, a large amount of real-world problem-solving knowledge is captured in goal-oriented dialogues, where common ground misalignment between users and helpers gives rise to sparse, diffuse, and dynamically refined evidence that challenges standard RAG pipelines. We propose Structured Dialogue Refinement (SDR), a unified framework that adapts dialogue corpora for RAG at both the retrieval and generation stages without altering the underlying pipeline. Specifically, SDR introduces Dual Dialogue Querying for intent-aligned retrieval via issue-centric and solution-centric pseudo-documents, and Graph-Structured Dialogues coupled with a relevance-driven subgraph selection strategy to enable effective utilization of conversational evidence. We further adopt a nugget-based evaluation setup for dialogue-grounded RAG, enabling fine-grained analysis of retrieval coverage and grounded answer generation. Experiments demonstrate that SDR substantially improves both retrieval quality and grounded QA performance under dialogue-specific structural challenges.

1 Introduction

Retrieval-Augmented Generation (RAG) has become a dominant paradigm for knowledge-intensive question answering (QA), grounding large language models (LLMs) in external evidence (Lewis et al., 2020; Fan et al., 2024). While RAG has been extensively studied on *document* corpora, such as Wikipedia (Jin et al., 2025b), curated knowledge bases (Guu et al., 2020), and technical manuals (Xu et al., 2024b), a substantial portion of real-world problem-solving knowledge is instead captured in goal-oriented dialogues. Customer support logs, troubleshooting chats, and help forums

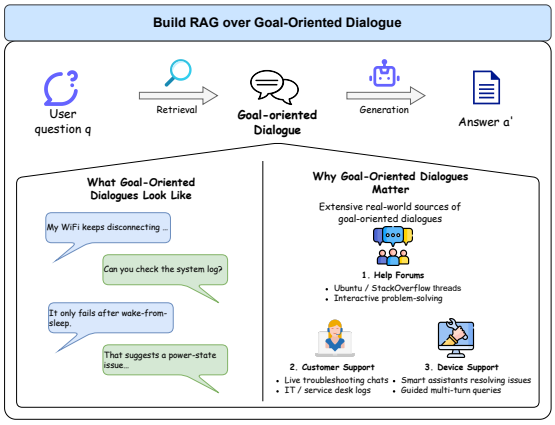

Figure 1: Our task is to build a RAG pipeline over goal-oriented dialogue evidence, an important real-world source of troubleshooting knowledge.

contain large volumes of multi-turn interactions in which users and domain experts collaboratively refine issues and solutions (Figure 1). Utilizing such dialogues for RAG-QA offers an opportunity to extend RAG to a complementary form of knowledge beyond documents, but remains underexplored because dialogue data exhibit structural characteristics that differ from those assumed by existing document-oriented RAG methods.

A central challenge lies in how goal-oriented dialogues express and refine information. Rather than presenting explicit and self-contained explanations, dialogues unfold through interactive problem-solving characterized by **speaker asymmetry** and resulting **common ground misalignment** (Sarkar et al., 2025). Users gradually reveal situational details, while helpers iteratively refine partial solutions. This process gives rise to two structural properties: (1) *information sparsity and diffusion*, where essential issue and solution cues are scattered across turns; and (2) *dialogue dynamism*, where both issue descriptions and solution attempts evolve through refinement, clarification, and cor-

*This work was done during an internship at Bloomberg.

Findings of the Association for Computational Linguistics: ACL 2026, pages 31419–31432

rection. These properties fundamentally challenge standard RAG pipelines. At the retrieval stage, surface-form similarity often fails to capture the underlying intent of dialogues, causing relevant dialogues to be missed. At the generation stage, even correctly retrieved dialogues contain interleaved, incremental, and sometimes abandoned reasoning steps, complicating evidence grounding. Existing RAG methods (Jimenez Gutierrez et al., 2024; Gutiérrez et al., 2025; Jin et al., 2025a), designed for coherent and static documents, are not equipped to handle these structural mismatches.

We address this gap by proposing Structured Dialogue Refinement (SDR), a unified framework that refines dialogue data for retrieval-augmented QA at *both* stages of the RAG pipeline. Our framework alleviates the structural challenges introduced by common ground misalignment while leaving the overall RAG architecture unchanged. Specifically, in the *retrieval stage*, we introduce **Dual Dialogue Querying**, which predicts two complementary sets of latent queries for each dialogue, with one capturing the user's underlying issue and the other representing the helper's underlying solution. These queries form two pseudo-document indices, enabling retrieval that is aligned with the dialogue's intent rather than its surface form. We operationalize this refinement with a dual-indexing and hybrid retrieval mechanism that aggregates evidence across the two perspectives. Second, in the *generation stage*, we introduce **Graph-Structured Dialogues**, which convert original dialogues into directed graphs that expose their internal problem-solution structure. Each dialogue is segmented into issue-solution units connected by refinement or elaboration relations. To ground answers effectively, we further propose a relevance-driven roaming strategy that selects the most pertinent subgraphs while preserving local reasoning flow.

To support rigorous evaluation of RAG-QA on real-world dialogue evidence, we adopt a nugget-based evaluation methodology and apply it to Ubuntu goal-oriented dialogues aligned with AskUbuntu QA pairs. Following prior work on nugget-based assessment (Thakur et al., 2025), we decompose reference answers into minimal factual units (nuggets) and identify dialogue evidence that supports them. This evaluation setup bridges curated QA pairs with raw multi-turn dialogues and enables fine-grained analysis of retrieval coverage and grounded answer generation.

Our contributions are summarized as: (1) **Structured Dialogue Refinement** (SDR), a unified framework for effective retrieval-augmented QA over goal-oriented dialogue corpora without modifying the underlying RAG pipeline. (2) **Dual Dialogue Querying**, a refinement in the retrieval stage that predicts issue- and solution-centric latent queries and enables intent-aligned hybrid retrieval under sparse and diffuse surface forms. (3) **Graph-Structured Dialogues**, a refinement in the generation stage that reorganizes dialogues into refinement-aware issue-solution graphs with a relevance-driven roaming strategy for grounded generation. (4) An evaluation setup for dialogue-grounded RAG that applies nugget-based grounding to analyze retrieval coverage and grounded answer generation, together with experimental analyses that reveal how refinements in the retrieval and generation stage address dialogue-specific challenges.

2 Background & Motivation

2.1 Retrieval-Augmented Question Answering

Retrieval-Augmented Generation (RAG) combines retrieval with LLMs for knowledge-intensive question answering (Lewis et al., 2020; Fan et al., 2024). Given a question q and a corpus $\mathcal{C}$, a retriever returns top-k texts $D = \{d_1, \ldots, d_k\}$, which are aggregated into a representation $H = \Phi(D)$ and used by a generator to produce an answer $\hat{a} = G(q, H)$. This formulation naturally decomposes RAG into two stages: a *retrieval stage* that selects relevant evidence, and a *generation stage* that conditions on this evidence to produce answers.

Most RAG systems are developed for **static, well-structured document corpora**, such as Wikipedia or curated knowledge bases (Karpukhin et al., 2020; Guu et al., 2020). These settings implicitly assume that relevant evidence is explicit, localized, and topically coherent. Such assumptions break down for **dialogue data**, where information is incrementally revealed and distributed across turns, challenging both retrieval and utilization.

2.2 Goal-Oriented Dialogues

Goal-oriented dialogues, such as troubleshooting and customer-service interactions (Young et al., 2013; Eric et al., 2017; Budzianowski et al., 2018), arise from an interactive problem-solving process between a user and a helper. A defining property is **speaker asymmetry**, which leads to **common ground misalignment** (Sarkar et al., 2025): users

possess situational context, while helpers hold domain expertise. As illustrated in Figure 2, this interaction gives rise to two structural properties:

Information sparsity and diffusion. Users disclose problem details incrementally and often incompletely, causing essential issue cues to be *sparse* and scattered across turns. Similarly, solutions emerge as fragmented hints, clarifications, or corrections rather than as a single coherent explanation, making the core problem-solution pair difficult to recover from surface text.

Dialogue dynamism: Issue refinement and solution elaboration. As common ground is negotiated, both issue descriptions and solution attempts evolve. Users *refine* the problem when earlier assumptions fail, while helpers *elaborate* solutions by adding or revising steps. This refinement-driven evolution obscures the underlying problem-solution structure in raw dialogue form.

These challenges have been widely recognized in prior work. Goal-oriented dialogues have been extensively studied for response generation and task completion in customer support, technical forums, and assistants (Young et al., 2013; Budzianowski et al., 2018; Dai et al., 2022). More recent studies examine how LLMs understand, participate in, or reason over such dialogues (Zhang et al., 2023; Yu et al., 2023), highlighting difficulties that arise from incomplete information and common ground misalignment (Sarkar et al., 2025). However, this line of work largely treats dialogues as interaction targets rather than as a retrieval corpus for grounding long-form question answering, leaving open how dialogue structure affects retrieval and utilization in RAG pipelines.

2.3 Challenge of Building RAG on Dialogues

To illustrate these challenges, we analyze AskUbuntu QA pairs with the Ubuntu Dialogue Corpus (Lowe et al., 2015), retrieving the top-20 dialogues using BM25 and evaluating nugget coverage and nugget support in answers generated by `Qwen2.5-72b-instruct` (details in Appendix A).

Failure in the Retrieval Stage. Dialogue cues are sparse, scattered, and often expressed indirectly through iterative refinements, making surface signals poor proxies for underlying intent. As shown in Figure 3(a), BM25 retrieves fewer than half of the nuggets present in the gold dialogues, while BGE also exhibits a substantial performance gap

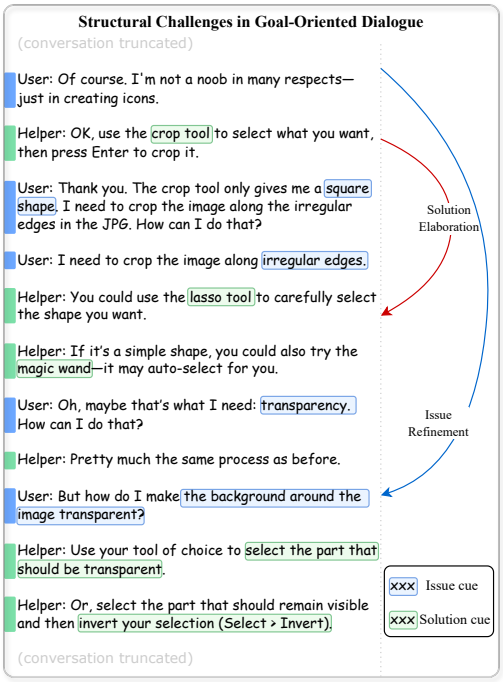

Figure 2: Structural challenges of goal-oriented dialogue. Issue cues (blue) and solution cues (green) appear sparsely and diffusely across turns, and both evolve dynamically through issue refinement (blue arrows) and solution elaboration (red arrows).

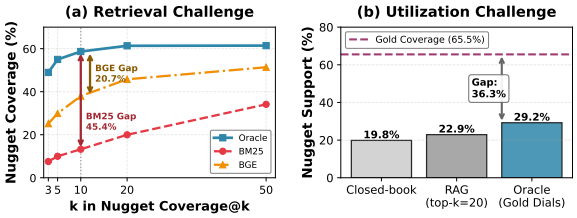

Figure 3: Challenges in the retrieval and generation stage in dialogue-based RAG. (a) Nugget coverage of top-10 dialogues retrieved by question-based retrieval versus gold dialogues. (b) Nugget support in generated answers using retrieved dialogues versus oracle gold dialogues.

at small top-k. These results highlight a significant retrieval bottleneck that persists even before generation.

Failure in the Generation Stage. Even when relevant dialogues are available, interleaved refinements, partial solutions, and conversational noise complicate grounding. As shown in Figure 3(b), oracle gold dialogues improve QA accuracy, yet the model still uses only a fraction of the available nuggets, exposing a fundamental generation gap beyond retrieval.

3 Framework Overview

Goal-oriented dialogues encode problem-solution information in forms fundamentally different from document corpora, creating a structural mismatch with standard RAG pipelines. We address this mismatch with **Structured Dialogue Refinement**, a unified framework that prepares dialogue data for retrieval-augmented QA at two complementary stages (Figure 4).

Specifically, we perform *question-agnostic, offline* refinement of dialogue data using two complementary representations. In the retrieval stage, each dialogue is transformed into a representation that explicitly captures its underlying issue and solution signals, enabling intent-aligned retrieval despite sparse and diffuse surface forms. In the generation stage, dialogues are independently reorganized into structured representations that make their internal problem-solution reasoning flow explicit, reducing ambiguity introduced by conversational dynamics. Together, these two refinements prepare dialogue data for the retrieval and generation stages of RAG while leaving the overall pipeline unchanged.

4 Dialogue Refinement in Retrieval Stage: Dual Dialogue Querying

Standard retrievers rely on stable lexical or semantic cues, yet in goal-oriented dialogues, issue and solution information emerges gradually across turns. Due to *speaker asymmetry* and *common ground misalignment*, user questions are often aligned with issue descriptions, while relevant evidence is expressed through solution-side refinements, making a single surface-form representation insufficient for retrieval. To address this mismatch, we predict two latent query sets for each dialogue, capturing issue- and solution-centric perspectives, and index them as intent-aligned pseudo-documents. At inference time, retrieval is performed independently over the two indices and fused via hybrid scoring, improving robustness under sparse and diffuse surface forms, consistent with prior work on query expansion and multi-view retrieval (Nogueira et al., 2019; Gospodinov et al., 2023).

4.1 Dual Query Construction

For each dialogue X, we predict two sets of representative queries: $Q_\iota(X)$, which approximates the underlying issue, and $Q_\sigma(X)$, which approximates the underlying solution, using an LLM as a query generator during offline preprocessing. These queries abstract the dialogue's core intent from complementary perspectives and are aggregated into two pseudo-documents:

$$\widetilde{X}_\iota(X) = \textsc{Concat}\big(Q_\iota(X)\big), \tag{1}$$

$$\widetilde{X}_\sigma(X) = \textsc{Concat}\big(Q_\sigma(X)\big). \tag{2}$$

The resulting pseudo-documents $\widetilde{X}_\iota(X)$ and $\widetilde{X}_\sigma(X)$ emphasize issue-centric and solution-centric signals, respectively, and are used to build two retrieval indices:

$$\widetilde{\mathcal{C}}_\iota = \{\widetilde{X}_\iota(X) \mid X \in \mathcal{C}\}, \tag{3}$$

$$\widetilde{\mathcal{C}}_\sigma = \{\widetilde{X}_\sigma(X) \mid X \in \mathcal{C}\}, \tag{4}$$

with each entry linked to its source dialogue X.

4.2 Hybrid Retrieval

Given a user question q, we retrieve candidates independently from both indices:

$$D_\iota = \text{Retr}(q, \widetilde{\mathcal{C}}_\iota), \qquad D_\sigma = \text{Retr}(q, \widetilde{\mathcal{C}}_\sigma), \tag{5}$$

map retrieved pseudo-documents back to their dialogues d, and fuse the two relevance scores via

$$s(X) = \alpha\, s_\iota(X) + (1-\alpha)\, s_\sigma(X). \tag{6}$$

The top-k dialogues under $s(X)$ form the final retrieved set D. By combining issue- and solution-centric signals, dual dialogue querying captures complementary intent cues and substantially improves retrieval under dialogue-induced sparsity.

5 Dialogue Refinement in Generation Stage: Graph-Structured Dialogues

Even when relevant dialogues are retrieved, their conversational form complicates grounding. As discussed in Section 2.3, issue refinement by users and solution elaboration by helpers, which are driven by common ground misalignment, lead to *sparsity, diffusion, and dynamism*, making it difficult for LLMs to utilize raw dialogue text. To provide a clearer view in the generation stage, we reorganize dialogues into graph-structured representations that explicitly expose their underlying issue-solution structure.

5.1 Graph Construction

We represent a dialogue using a directed graph, $\mathcal{G} = (V, E)$, where nodes V correspond to coherent issue–solution units and edges E encode refinement or elaboration relations, exposing the dialogue's implicit problem-solution flow.

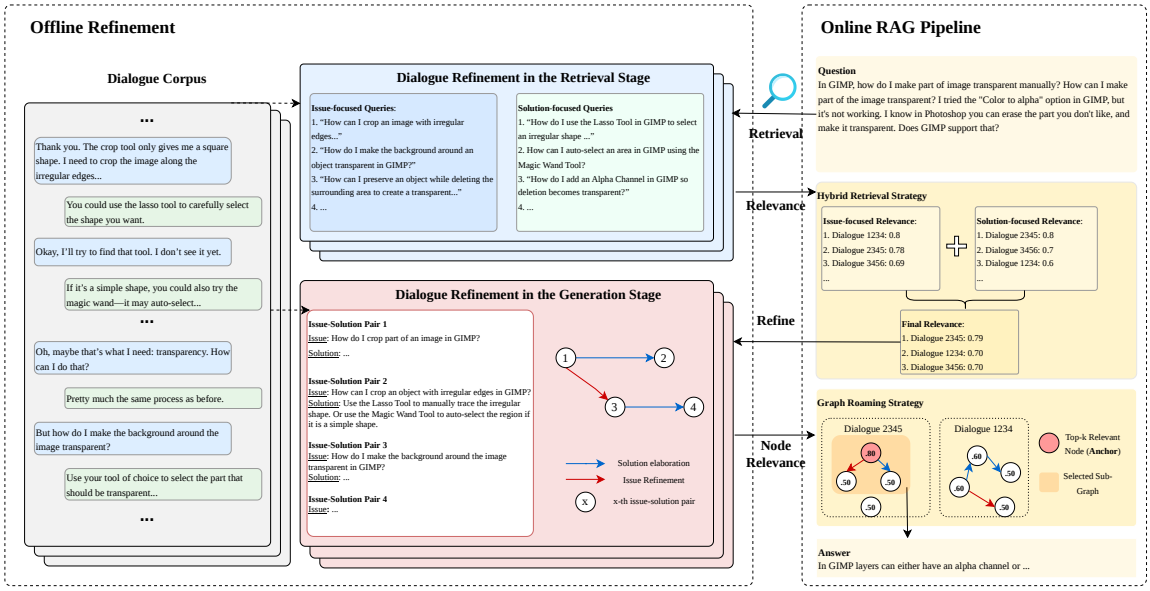

Figure 4: Overview of Structured Dialogue Refinement (SDR). *Retrieval-stage* refinement predicts issue-centric and solution-centric latent queries to build dual indices and enable intent-aligned hybrid retrieval. *Generation-stage* refinement reorganizes retrieved dialogues into graph-structured representations of issue-solution units, and a roaming strategy extracts coherent subgraphs for grounding.

Node extraction. We segment the dialogue into a sequence of issue-solution pairs:

$$V = \{v_1, v_2, \ldots, v_n\}, \qquad v_i = (\iota_i, \sigma_i), \quad (7)$$

where ι_i is a user-side issue expression and σ_i is its associated helper-side solution attempt. Each node condenses a localized interaction segment, abstracting conversational noise while preserving core semantic units. Because node extraction relies on LLM-based segmentation, the resulting issue-solution units may be imperfectly segmented. For example, adjacent units may be merged, or a coherent unit may be split into multiple nodes.

Edge construction. Directed edges connect nodes when one v_i updates another v_j, capturing two common relations in goal-oriented dialogues: (1) **Issue refinement**: ι_j modifies, extends, or corrects ι_i, typically when the user provides new details that narrow or adjust the original problem description. For example, the user may first report "`my WiFi is not working.`" and later clarify "`it only disconnects when waking from sleep.`" (2) **Solution elaboration**: σ_j augments, clarifies, or adjusts σ_i, often triggered when the user seeks clarification about a previously suggested step. For example, the helper may suggest "`check the syslog for errors.`" and the user responds "`I dont know how to open the syslog.`", prompting the helper to elaborate: "`you can open it with 'sudo journalctl -xe'.`" Nodes without any incoming or outgoing refinement or elaboration edges naturally form *isolated units or subgraphs*. These structures arise when the dialogue contains topic shifts, abandoned solution paths, or independent issue-solution segments. Our graph representation preserves these isolated components, which later allows the generation stage to decide whether these components contribute to a given question.

5.2 Graph Utilization

To improve robustness to imperfect segmentation introduced during graph construction, our utilization strategy operates on local graph neighborhoods rather than relying on any single node to capture a complete issue-solution unit perfectly. Given a question q and a retrieved dialogue set $D = \{X_i, ..., X_k\}$, we compute a semantic relevance score for each node v_i in the dialogues of D:

$$r(v_i \mid q) = \text{sim}(q, v_i), \; v_i \in V(X). \quad (8)$$

We select the top-k highest-scoring nodes as anchors. Starting from each anchor node, we perform a light roaming step through adjacent nodes linked by refinement or elaboration edges, forming a small

connected subgraph centered on each anchor:

$$H = \{\mathcal{G}_q^{(1)}, \ldots, \mathcal{G}_q^{(k)}\}. \quad (9)$$

This roaming step allows the model to recover nearby context when a coherent issue-solution unit has been split across nodes, while also avoiding reliance on an entire graph when adjacent units have been overly merged. These subgraphs preserve the dialogue's internal logical progression while filtering out irrelevant conversational turns. Each selected subgraph $\mathcal{G}_q^{(j)}$ is formatted as its node list $V_q^{(j)}$ and edge list $E_q^{(j)}$, explicitly encoding the issue-solution units and their refinement or elaboration relations. The generator then conditions on all selected subgraphs:

$$\hat{a} = G(q, \{\mathcal{G}_q^{(j)}\}_{j=1}^{k}), \quad (10)$$

providing structured grounding that stabilizes reasoning and improves answer quality.

6 Benchmark for RAG-QA on Goal-Oriented Dialogues

Most existing RAG benchmarks are built on document corpora, where information is explicit, topically coherent, and structurally stable. In contrast, evaluating RAG when the underlying knowledge comes from *dialogues* remains underexplored. To support rigorous evaluation in this setting, we adopt a nugget-based evaluation methodology and apply it to RAG-QA over goal-oriented dialogues. Implementation details, statistical information, and evaluation protocols are provided in Appendix C.

Data Sources We collect two complementary resources: the **Ubuntu Dialogue Corpus** (Lowe et al., 2015), which provides large-scale, multi-turn troubleshooting dialogues as raw conversational evidence; and **AskUbuntu QA pairs**, which offer high-quality, community-verified question-answer pairs as evaluation targets. Together, these sources combine realistic but noisy dialogue traces with curated problem-solution outcomes.

Nugget-Based Alignment Because the two corpora originate independently, we align them via nugget-based grounding. Each AskUbuntu answer is decomposed into minimal factual units (nuggets) (Lin and Demner-Fushman, 2005; Pavlu et al., 2012), and dialogues are labeled as relevant if they support these nuggets. Nugget-level grounding enables precise dialogue-QA alignment and fine-grained evaluation beyond surface-form matching. Following prior work (Thakur et al., 2025), we employ LLM-as-a-judge for NLI-based relevance labeling and nugget evaluation.

Evaluation We assess RAG performance by measuring whether the generated answer $\hat{a}$ supports reference nuggets using *All-Strict* nugget accuracy (Pradeep et al., 2025). We also report Recall@k and Nugget Coverage@k to measure dialogue and nugget coverage in the top-k retrieved results.

7 Experiments

We evaluate SDR on three research questions:

RQ1 Does SDR improve end-to-end QA compared to standard RAG and strong refinement baselines?

RQ2 How do refinements in the retrieval and generation stage contribute to performance?

RQ3 Why does hybrid retrieval and graph-based utilization outperform simpler alternatives?

7.1 Experimental Setup

We evaluate SDR and baselines on the Ubuntu-RAG benchmark (Section 6), reporting both end-to-end QA and retrieval performance. For QA, we compare against Naive RAG, and RECOMP-abstractive and RECOMP-extractive (Xu et al., 2024a), which improve utilization via online summarization and re-ranking. All RAG systems use the top-20 retrieved dialogues and are evaluated over five runs. We also include non-RAG references (Close-Book, $\text{Oracle}_{\text{nugget}}$, and Oracle). For retrieval, we compare against original retrieval, Doc2Query (Nogueira et al., 2019), Doc2Query-(Gospodinov et al., 2023), and single-view variants of SDR using only issue- or solution-focused queries.

Model We evaluate two base models, `Qwen2.5-72b-instruct` and `gpt-4o-mini`, and use `BAAI/bge-large-en-v1.5` for dense retrieval. Offline refinement and nugget-based evaluation are performed using `claude-3-7-sonnet`. Additional details are provided in Appendix D.

7.2 Overall Performance (RQ1)

We first evaluate end-to-end dialogue-grounded QA performance. Table 1 compares SDR with Naive RAG and non-RAG reference settings across retrievers and base models. SDR consistently outperforms Naive RAG in all configurations, achieving

Table 1: End-to-end dialogue-grounded QA performance (nugget accuracy). SDR is compared with non-RAG references (Close-Book and Oracle) and Naive RAG. Relative improvements over Naive RAG are shown in green.

	BM25		bge-large-en-v1.5	
	Qwen2.5-72b	gpt-4o-mini	Qwen2.5-72b	gpt-4o-mini
Close-Book	$19.50_{\pm 0.18}$	$22.83_{\pm 0.20}$	$19.50_{\pm 0.18}$	$22.83_{\pm 0.20}$
$\text{Oracle}_{\text{nugget}}$	$29.11_{\pm 0.18}$	$29.77_{\pm 0.22}$	$29.11_{\pm 0.18}$	$29.77_{\pm 0.22}$
Oracle	$29.55_{\pm 0.19}$	$31.82_{\pm 0.26}$	$29.55_{\pm 0.19}$	$31.82_{\pm 0.26}$
Naive RAG	$22.44_{\pm 0.16}$	$26.00_{\pm 0.20}$	$25.37_{\pm 0.24}$	$26.98_{\pm 0.22}$
SDR	$27.37_{\pm 0.21}$ ↑22%	$28.56_{\pm 0.23}$ ↑10%	$28.54_{\pm 0.20}$ ↑12%	$29.12_{\pm 0.18}$ ↑8%

Table 2: Comparison with refinement-based RAG baselines using gpt-4o-mini. Results are reported with and without retrieval enhancement, showing that SDR consistently outperforms RECOMP under both settings.

	BM25	bge-large-en-v1.5
Naive RAG	$\underline{26.00}_{\pm 0.20}$	$26.98_{\pm 0.22}$
w/o retrieval enhancement		
RECOMP-abs.	$22.68_{\pm 0.24}$	$27.10_{\pm 0.19}$
RECOMP-extr.	$25.34_{\pm 0.19}$	$26.26_{\pm 0.24}$
w/ retrieval enhancement		
RECOMP-abs	$24.79_{\pm 0.19}$	$\underline{28.09}_{\pm 0.22}$
RECOMP-extr	$25.34_{\pm 0.19}$	$26.49_{\pm 0.21}$
SDR	$\mathbf{28.56}_{\pm 0.23}$	$\mathbf{29.12}_{\pm 0.18}$

gains of up to 22%, and substantially narrows the gap to oracle performance. These results indicate that structured dialogue refinement is critical for grounding multi-turn, sparsely expressed evidence, beyond what can be achieved by standard retrieval and generation alone.

We next compare SDR with refinement-based RAG baselines using gpt-4o-mini in Table 2. Without retrieval enhancement, RECOMP variants fail to improve over Naive RAG and can even degrade performance under BM25, indicating that refinement in the generation stage alone is insufficient when relevant dialogue evidence is poorly retrieved. With retrieval enhancement, RECOMP benefits from stronger inputs and becomes more competitive, but still consistently underperforms SDR. In contrast, SDR achieves the best performance across retrievers, demonstrating that jointly refining dialogues in both the retrieval and generation stages is critical for robust dialogue-grounded QA.

7.3 Component Contribution Analysis (RQ2)

We analyze how refinement of SDR in the retrieval and generation stage contributes to performance.

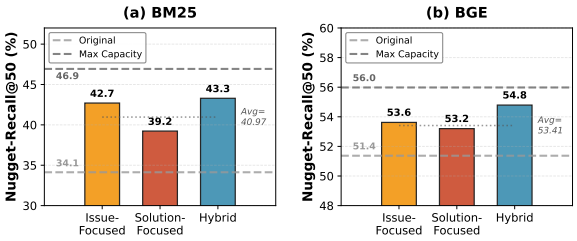

Figure 5: Hybrid retrieval effect. Max capacity denotes nuggets covered by issue- and solution-focused method.

Retrieval stage Table 3 reports retrieval performance under both BM25 and bge-large-en-v1.5. Compared to indexing raw dialogues, query expansion baselines (Doc2Query and Doc2Query-) improve recall and nugget coverage, indicating that retrieval benefits from enriching sparse dialogue cues. However, SDR achieves the strongest overall coverage: hybrid retrieval yields the best R@50 and N@50 for both retrievers, demonstrating that our refinement in the retrieval stage more effectively surfaces nugget-supporting evidence than document-expansion alone.

Generation Stage contribution Table 4 presents a whole-pipeline ablation. Removing refinement in the retrieval stage substantially degrades performance due to missed relevant dialogues under sparsity and diffusion. Removing refinement in the generation stage yields a different failure mode: although relevant dialogues are retrieved, unstructured conversational traces hinder effective grounding. The full model consistently outperforms both variants, confirming that refinements in the retrieval and generation stage address complementary sources of error.

7.4 Design Choice Analysis (RQ3)

Why hybrid retrieval works Figure 5 analyzes issue-only, solution-only, and hybrid retrieval. Issue-focused retrieval captures core problem de-

Table 3: Dialogue retrieval performance. Recall@k (R@k) and Nugget Coverage@k (N@k) are reported for both retrievers. Hybrid retrieval combining issue- and solution-focused queries achieves the strongest overall coverage.

	BM25				bge-large-en-v1.5			
	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50
Original	6.98	17.26	19.97	34.13	34.90	53.96	45.76	51.37
Doc2Query	10.66	22.02	25.83	37.20	39.19	58.18	46.69	52.89
Doc2Query-	10.82	22.11	26.34	37.63	**40.43**	57.64	47.33	53.84
Hybrid (Ours)	**14.33**	**32.96**	**28.28**	**43.29**	39.91	**59.32**	**48.74**	**54.73**

Table 4: Whole-pipeline ablation on a benchmark subset (*bge*). The nugget accuracy of removing refinement in either the retrieval or generation stage is reported.

	Qwen2.5-72b	*gpt-4o-mini*
Naive RAG	21.30	24.67
SDR w/o RetRef	↓ 21.13	↑ 26.52
SDR w/o UtilRef	↓ 20.88	↑ 26.97
SDR	↑ 22.13	↑ 27.43

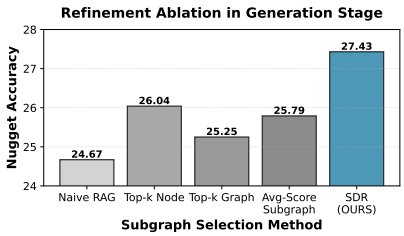

Figure 6: Ablation study of refinement in the generation stage. SDR is compared with reranking on nodes, graphs, and using the average score for identifying the subgraph.

scriptions, while solution-focused queries encode solution cues that are often absent from issue descriptions. Although weak in isolation, solution signals substantially improve coverage when combined with issue signals (i.e., max capacity). Hybrid retrieval consistently approaches the nugget coverage of these two signals, demonstrating that integrating complementary dialogue perspectives is essential for overcoming sparsity and diffusion.

Why roaming-based utilization works Figure 6 compares our roaming-based subgraph selection with simpler strategies. Selecting top-k nodes discards the local reasoning context, while selecting entire graphs often includes irrelevant or abandoned solution paths. Using average relevance improves stability but remains inferior. Roaming-based selection achieves the highest accuracy by extracting minimal, coherent subgraphs that preserve refinement and elaboration structure while suppressing conversational noise.

8 Related Work

Retrieval-Augmented Generation External evidence grounds LLMs to improve factuality and reduce hallucinations (Lewis et al., 2020; Borgeaud et al., 2022; Gao et al., 2023; Tonmoy et al., 2024). Most RAG research and benchmarks focus on document corpora such as encyclopedic passages, curated knowledge bases, or technical manuals (Jin et al., 2025b; Guu et al., 2020; Xu et al., 2024b; Zhang et al., 2024). These corpora assume explicit, coherent, and structurally stable evidence, assumptions often violated by goal-oriented dialogues. While some datasets include dialogue components (Li et al., 2025b,a), they primarily study *conversational RAG for dialogue agents* rather than treating dialogue transcripts as the *retrieval corpus* for long-form, dialogue-grounded QA.

Refinement Methods for RAG A central difficulty in RAG is utilizing retrieved context that is long, noisy, or partially relevant (Cuconasu et al., 2024). Training-based approaches improve robustness via supervised fine-tuning or preference learning (Yoran et al., 2024; Asai et al., 2024; Lin et al., 2024), while training-free approaches refine context via summarization, structured extraction, or retrieval-time pruning/reasoning (Xu et al., 2024a; Jimenez Gutierrez et al., 2024; Wang et al., 2024; Qin et al., 2025). Recent work also reorganizes corpora offline into memory/graph/hierarchical structures (Jimenez Gutierrez et al., 2024; Gutiérrez et al., 2025; Edge et al., 2024; Guo et al., 2025; Jin et al., 2025a). However, these methods often rely on document-style stability (e.g., persistent entities or explicit relations), which is weaker in goal-oriented dialogues; we instead refine dialogues with structure and selection mechanisms tailored to refinement-driven interaction.

Goal-Oriented Dialogues Goal-oriented dialogue research spans customer support, technical forums, and assistants (Young et al., 2013; Eric et al., 2017; Budzianowski et al., 2018; Dinan et al., 2019; Dai et al., 2022; Feng et al., 2021). Many benchmarks are synthetic or semi-synthetic and target response generation or task completion, whereas real troubleshooting dialogues feature incomplete issue descriptions, exploratory solution attempts, and evolving common ground. Recent work studies how LLMs understand and reason in such settings (Zhang et al., 2023; Yu et al., 2023), and Sarkar et al. (2025) identify common ground misalignment as a key challenge. We build on this insight by showing it also induces structural mismatches for RAG, hurting retrieval via sparse/diffused cues and utilization via dynamic refinement, and propose dialogue-specific refinement plus a new dialogue-grounded RAG benchmark.

9 Conclusion

We presented SDR, a unified refinement framework for retrieval-augmented QA over goal-oriented dialogue corpora. By addressing dialogue-specific structural challenges, SDR improves both the retrieval and generation stages without modifying the underlying RAG architecture. Dual Dialogue Querying enables intent-aligned retrieval via complementary issue- and solution-centric views, while Graph-Structured Dialogues organize conversational evidence into coherent reasoning subgraphs. Experiments demonstrate consistent gains in retrieval quality and answer accuracy, underscoring the importance of dialogue-specific refinement for extending RAG beyond document corpora.

Limitations

Our work focuses on structural challenges that broadly characterize goal-oriented dialogues, and the proposed framework is designed to generalize across domains. However, due to the lack of publicly available dialogue-grounded RAG benchmarks, our empirical evaluation is limited to a single benchmark constructed from Ubuntu dialogues and AskUbuntu QA pairs; extending evaluation to additional domains as such benchmarks emerge is a promising direction for future work. In addition, our current formulation assumes a static dialogue corpus, whereas real-world systems often accumulate dialogues continuously. Developing principled methods to incrementally incorporate new dialogues (e.g., updating latent queries or extending graph structures without full reprocessing) would further improve the practicality of SDR in dynamic settings.

Acknowledgments

Bin Wu is supported by the Bloomberg Data Science Ph.D. Fellowship.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *International conference on machine learning*, pages 4558–4586. PMLR.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th annual SIGdial meeting on discourse and dialogue*, pages 37–49.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. Multidoc2dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Mitko Gospodinov, Sean MacAvaney, and Craig Macdonald. 2023. Doc2query–: when less is more. In *European Conference on Information Retrieval*, pages 414–422. Springer.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. LightRAG: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10746–10761, Suzhou, China. Association for Computational Linguistics.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From rag to memory: Non-parametric continual learning for large language models. In *Forty-second International Conference on Machine Learning*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37:59532–59569.

Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, Ye Qi, and Zhicheng Dou. 2025a. Hierarchical document refinement for long-context retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3502–3520.

Jiajie Jin, Yutao Zhu, Zhicheng Dou, Guanting Dong, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao Yang, and Ji-Rong Wen. 2025b. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 737–740.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Haitao Li, Yifan Chen, Hu YiRan, Qingyao Ai, Junjie Chen, Xiaoyu Yang, Jianhui Yang, Yueyue Wu, Zeyang Liu, and Yiqun Liu. 2025a. Lexrag: Benchmarking retrieval-augmented generation in multi-turn legal consultation conversation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3606–3615.

Qiwei Li, Teng Xiao, Zuchao Li, Ping Wang, Mengjia Shen, and Hai Zhao. 2025b. Dialogue-rag: Enhancing retrieval for llms via node-linking utterance rewriting. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24423–24438.

Jimmy Lin and Dina Demner-Fushman. 2005. Automatically evaluating answers to definition questions. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 931–938.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, and 1 others. 2024. Ra-dit: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pages 285–294.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Virgil Pavlu, Shahzad Rajput, Peter B Golbus, and Javed A Aslam. 2012. Ir system evaluation using nugget-based test collections. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 393–402.

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2025. The great nugget recall: Automating fact extraction and rag evaluation with large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 180–190.

Qitao Qin, Yucong Luo, Yihang Lu, Zhibo Chu, Xiaoman Liu, and Xianwei Meng. 2025. Towards adaptive memory-based optimization for enhanced retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7991–8004.

Rupak Sarkar, Neha Srikanth, Taylor Pellegrin, Rachel Rudinger, Claire Bonial, and Philip Resnik. 2025. Understanding common ground misalignment in goal-oriented dialog: A case-study with ubuntu chat logs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3200–3215.

Nandan Thakur, Jimmy Lin, Sam Havens, Michael Carbin, Omar Khattab, and Andrew Drozdov. 2025. Freshstack: Building realistic benchmarks for evaluating retrieval on technical documents. *arXiv preprint arXiv:2504.13128*.

SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Ruobing Wang, Qingfei Zhao, Yukun Yan, Daren Zha, Yuxuan Chen, Shi Yu, Zhenghao Liu, Yixuan Wang, Shuo Wang, Xu Han, and 1 others. 2024. Deepnote: Note-centric deep retrieval-augmented generation. *arXiv preprint arXiv:2410.08821*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024a. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. 2024b. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 2905–2909.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Xiao Yu, Maximillian Chen, and Zhou Yu. 2023. Prompt-based monte-carlo tree search for goal-oriented dialogue policy planning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7101–7125.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6665–6694.

Xuanwang Zhang, Yun-Ze Song, Yidong Wang, Shuyun Tang, Xinfeng Li, Zhengran Zeng, Zhen Wu, Wei Ye, Wenyuan Xu, Yue Zhang, and 1 others. 2024. Raglab: A modular and research-oriented unified framework for retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 408–418.

A Details of Preliminary Study

For our preliminary experiments, we randomly sample 100 question-answer pairs from the benchmark described in Appendix C.

To examine retrieval challenges, we use each question as a query to retrieve the top-20 dialogues and compare their nugget coverage against that of all gold-relevant dialogues. The substantial gap in nugget coverage highlights the difficulty of retrieving relevant evidence from goal-oriented dialogue corpora.

To examine utilization challenges, we evaluate the same 100 QA pairs under three settings: closed-book generation (without external dialogues), Naive RAG, and an oracle setting that provides all gold-relevant dialogues. The improvement of RAG over the closed-book baseline confirms the value of dialogue evidence. The gap between Naive RAG and the oracle setting reflects the impact of retrieval errors, while the remaining gap between oracle performance and the nuggets contained in the relevant dialogues reveals the difficulty LLMs face in effectively utilizing raw dialogue data.

B Details of Method

We present the pseudo code of SDR in Algorithm 1, which summarizes both the offline refinement and online inference stages.

C Details of Benchmark Construction

We provide more details of Ubuntu-RAG.

C.1 Benchmark Construction

We provide additional details for constructing the Ubuntu-RAG benchmark, which aligns Ubuntu goal-oriented dialogues with AskUbuntu QA pairs via nugget-based grounding.

Dialogue Preprocessing. We begin with the Ubuntu Dialogue Corpus (Lowe et al., 2015). Because the raw logs contain user identifiers and cross-user references, we apply two PII-removal steps: (1) all speakers are anonymized using canonical tokens (`user1`, `user2`); (2) mentions of usernames inside utterances—common in channel-style chats—are removed. This yields a clean, speaker-agnostic dialogue corpus suitable for retrieval and structural refinement.

AskUbuntu QA Preprocessing. AskUbuntu answers often include images, scripts, or executable code. To focus our benchmark on text-only scenarios and leave other domains for future work and also consider the time range of the dialogue corpus, we perform the following filtering: (1) *Time-range alignment*: only questions posted within the active period of the Ubuntu Dialogue Corpus are retained, avoiding ambiguity from system-version drift; (2) *Non-text removal*: QA pairs containing image URLs or code blocks are discarded, since multimodal reasoning and code synthesis are beyond the scope of this work; (3) *Quality filtering*: we select the 300 QA pairs with the highest community scores, which serve as reliable gold references due to community validation.

Nugget Generation. Because the dialogue corpus and QA pairs originate from independent sources, aligning them requires bridging a substantial semantic gap. Following recent works on automatically constructing RAG benchmark (Thakur et al., 2025), we decompose each gold answer into a set of minimal factual units ("nuggets"). For each QA pair, an LLM (`claude-3-7-sonnet` in our work) generates a concise list of nuggets that collectively capture the essential information needed to answer the question. Nuggets enable both (i) filtering dialogues that contain necessary evidence and (ii) fine-grained factual evaluation of retrieval and generation.

Dialogue Relevance Labeling. To determine which dialogues truly contain evidence for a given QA pair, we adopt a two-stage procedure. First, because of the scale of the dialogue corpus, we retrieve a candidate pool using BM25 with diverse queries: the question text, the gold answer, and the nugget list. This ensures broad coverage under surface-form mismatch. Second, for each dialogue-nugget pair, we apply LLM-as-a-Judge to perform an NLI-style classification (*Entailment*, *Neutral*, *Contradiction*). We treat only *Entailment* as positive, resulting in high-precision dialogue-nugget relevance labels. This process yields aligned triples of questions, answer nuggets, and grounded dialogues.

C.2 Evaluation

We evaluate systems on two tasks: dialogue retrieval and dialogue-grounded QA generation. Both tasks rely on nugget-level supervision to provide fine-grained factual assessment.

Algorithm 1 Structured Dialogue Refinement (SDR)

Require: Dialogue corpus $\mathcal{C}$; question q; retriever Retr; generator G; fusion weight α; top-k dialogues; top-m anchors; roam radius ρ
Ensure: Generated answer $\hat{a}$

Offline refinement (build once)
for all dialogues $X \in \mathcal{C}$ **do**
  $Q_\iota(X) \leftarrow \text{LLMQueryGen}_\iota(X)$ ▷ issue-focused queries
  $Q_\sigma(X) \leftarrow \text{LLMQueryGen}_\sigma(X)$ ▷ solution-focused queries
  $\widetilde{X}_\iota(X) \leftarrow \textsc{Concat}(Q_\iota(X))$
  $\widetilde{X}_\sigma(X) \leftarrow \textsc{Concat}(Q_\sigma(X))$
  $\mathcal{G}(X) \leftarrow \textsc{BuildGraph}(X)$ ▷ nodes: issue–solution units; edges: refine/elaborate
end for
Build indices $\widetilde{\mathcal{C}}_\iota = \{\widetilde{X}_\iota(X)\}$ and $\widetilde{\mathcal{C}}_\sigma = \{\widetilde{X}_\sigma(X)\}$
Online RAG inference (per query)
$D_\iota \leftarrow \text{Retr}(q, \widetilde{\mathcal{C}}_\iota)$; $D_\sigma \leftarrow \text{Retr}(q, \widetilde{\mathcal{C}}_\sigma)$
for all dialogues X in candidates from $D_\iota \cup D_\sigma$ **do**
  $s(X) \leftarrow \alpha\, s_\iota(X) + (1-\alpha)\, s_\sigma(X)$
end for
$D \leftarrow \textsc{TopK}(\{X\}, s, k)$ ▷ final retrieved dialogues
Graph-structured utilization
$\mathcal{H} \leftarrow \emptyset$
for all $X \in D$ **do**
  $\mathcal{G} \leftarrow \mathcal{G}(X)$
  for all nodes $v \in V(\mathcal{G})$ **do**
    $r(v \mid q) \leftarrow \textsc{sim}(q, v)$
  end for
  $\mathcal{A} \leftarrow \textsc{TopM}(V(\mathcal{G}), r(\cdot \mid q), m)$ ▷ anchor nodes
  for all $v \in \mathcal{A}$ **do**
    $\mathcal{G}_q^{(v)} \leftarrow \textsc{Roam}(\mathcal{G}, v, \rho)$ ▷ expand along refine/elaborate edges
    $\mathcal{H} \leftarrow \mathcal{H} \cup \{\mathcal{G}_q^{(v)}\}$
  end for
end for
$\hat{a} \leftarrow G(q, \mathcal{H})$
return $\hat{a}$

Retrieval Evaluation. Given a question, the system retrieves a ranked list of dialogues. We report two metrics: (1) **Recall@k**: the proportion of gold-relevant dialogues appearing in the top-k. (2) **Nugget Coverage@k**: the fraction of answer nuggets supported by at least one retrieved dialogue. For a nugget set N and top-k dialogues D_k:

$$\text{NC@k} = \frac{|\{\, n \in N \mid \exists d \in D_k \text{ s.t. } d \models n\}|}{|N|}, \tag{11}$$

where $d \models n$ denotes LLM-judged entailment.

QA Evaluation. Given a question and retrieved dialogues, a model generates an answer $\hat{a}$. We again use LLM-as-a-Judge to determine whether each associated nugget in the reference answer is entailed by the generated answer. For nugget set N, we compute:

$$\text{Score} = \frac{1}{|N|} \sum_{n \in N} \mathbf{1}[\hat{a} \models n]\,, \tag{12}$$

which corresponds to the *All-Strict* nugget accuracy (Pradeep et al., 2025). A nugget contributes only when the answer clearly supports it. This evaluation emphasizes factual grounding rather than surface-form similarity.

C.3 Statistical Information

Finally, we obtain the Ubuntu-RAG benchmark. The statistical information of Ubuntu-RAG is

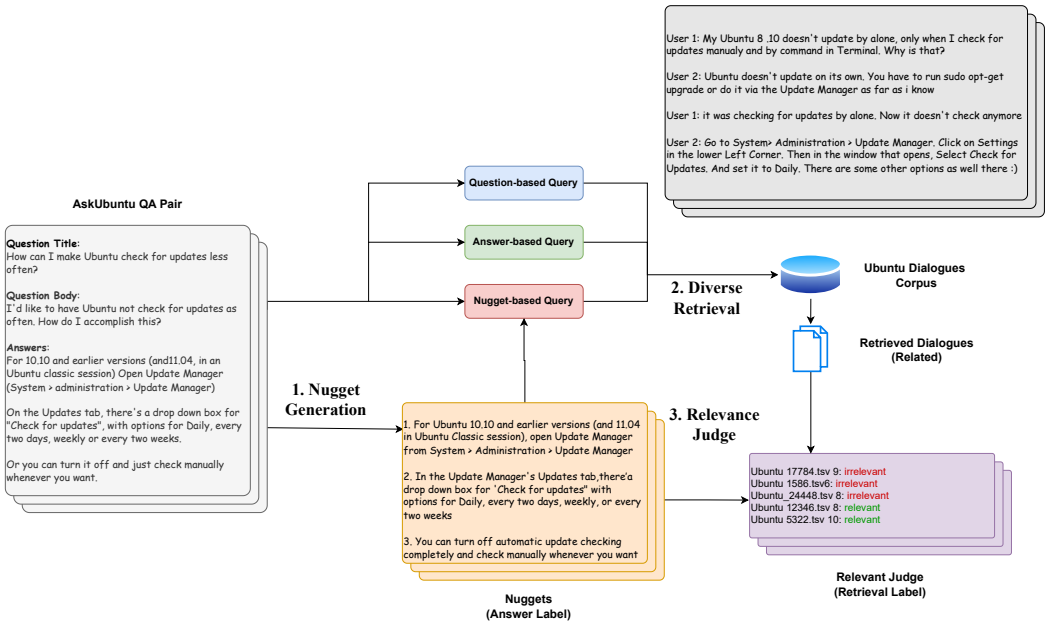

Figure 7: Overview of Benchmark Construction. The employed LLMs first generate the nuggets of the long answers regarding the question. Then the retriever takes the question, answers and nuggets, respectively, to identify the related dialogue from the large Ubuntu Dialogue Corpus. Finally, the LLM-as-a-judge is employed to perform the NLI based on the retrieval dialogues and the associated nuggets to annotate the relevance label.

	Count
# Dialogues	18,000
# QA Pairs	300

Table 5: Statistical Information of Ubuntu-RAG

shown in Table 5.

D Details of Experimental Setup

D.1 Baselines

We evaluate the proposed SDR by comparing with the following baselines:

RAG baselines. For dialogue-grounded QA, we compare with: (1) **Naive RAG**: a direct RAG pipeline with no refinement; (2) **RECOMP-abstractive** (Xu et al., 2024a): summarization-based context compression; (3) **RECOMP-extractive** (Xu et al., 2024a): retrieval-time re-ranking via an encoder. We also include three ablated variants of SDR that replace our graph-structured refinement with simpler structures: (4) **Top-k Node**: a re-ranking method on the captured issue-solution pairs. (5) **Top-k Graph**: a re-ranking method on the captured graphs (a set of issue-solution pairs). (6) **Avg-Score Subgraph**: similar to SDR but using the average score of the whole sub-graph for ranking.

Retrieval baselines. We evaluate the following methods for dialogue retrieval: (1) **Original**: indexing and retrieving from the raw dialogue text; (2) **Doc2Query** (Nogueira et al., 2019): a document-expansion method predicting pseudo-queries for each dialogue; (3) **Doc2Query–** (Gospodinov et al., 2023): an extension that filters predicted queries using relevance scores; as well as two single-view variants of SDR: (4) **issue-focus**: using only predicted issue-centric queries; (5) **solution-focus**: using only solution-centric queries. We focus on domain-agnostic retrieval methods to ensure a fair comparison with our framework, which operates without the dialogue-specific retrievers fine-tuned on the specific dialogue data.

D.2 Others

All RAG models use the top 20 retrieved dialogues. For generations, we run each method 5 times to reduce randomness and set `top_p` to 0.99 and `temperature` to 2e−4 for reproducibility.

For comparisons between SDR and refinement-based baselines (Table 1 and 2), we retrieve 50 dialogues but incorporate only the top 20 into generation, ensuring consistency with the Naive RAG setting.