
MINIRAG: TOWARDS EXTREMELY SIMPLE RETRIEVAL-AUGMENTED GENERATION

Tianyu Fan, Jingyuan Wang, Xubin Ren, Chao Huang*

University of Hong Kong

{tianyufan0504, jameswangjingyuan, xubinrencs, chaohuang75}@gmail.com;

ABSTRACT

The growing demand for efficient and lightweight Retrieval-Augmented Generation (RAG) systems has highlighted significant challenges when deploying Small Language Models (SLMs) in existing RAG frameworks. Current approaches face severe performance degradation due to SLMs' limited semantic understanding and text processing capabilities, creating barriers for widespread adoption in resource-constrained scenarios. To address these fundamental limitations, we present MiniRAG a novel RAG system designed for extreme simplicity and efficiency. MiniRAG introduces two key technical innovations: (1) a semantic-aware heterogeneous graph indexing mechanism that combines text chunks and named entities in a unified structure, reducing reliance on complex semantic understanding, and (2) a lightweight topology-enhanced retrieval approach that leverages graph structures for efficient knowledge discovery without requiring advanced language capabilities. Our extensive experiments demonstrate that MiniRAG achieves comparable performance to LLM-based methods even when using SLMs while requiring only 25% of the storage space. Additionally, we contribute a comprehensive benchmark dataset for evaluating lightweight RAG systems under realistic on-device scenarios with complex queries. We fully open-source our implementation and datasets at: <https://github.com/HKUDS/MiniRAG>.

1 INTRODUCTION

Recent advances in Retrieval-Augmented Generation (RAG) have revolutionized how language models access and utilize external knowledge, demonstrating impressive capabilities across diverse applications from question answering to document synthesis (Fan et al., 2024). While these systems achieve remarkable performance through sophisticated retrieval mechanisms and powerful language models, they predominantly rely on Large Language Models (LLMs) throughout their pipeline - from index construction and knowledge retrieval to final response generation (Gao et al., 2023). This extensive dependence on LLMs introduces substantial computational overhead and resource requirements, creating significant barriers for deployment in resource-constrained scenarios such as edge devices, privacy-sensitive applications, and real-time processing systems (Liu et al., 2024). Despite growing demand for efficient and lightweight language model applications, current RAG frameworks offer limited solutions for maintaining robust performance within these practical constraints, highlighting a critical gap between theoretical capabilities and real-world deployment needs.

The limitations of existing RAG systems become particularly apparent when attempting to utilize Small Language Models (SLMs) for resource-efficient deployment. While these compact models offer significant advantages in terms of computational efficiency and deployment flexibility, they face fundamental challenges in key RAG operations - from semantic understanding to effective information retrieval. Current RAG architectures (*e.g.*, LightRAG Guo et al. (2024) and GraphRAG Edge et al. (2024a)), originally designed to leverage LLMs' sophisticated capabilities, fail to accommodate the inherent constraints of SLMs across multiple critical functions: sophisticated query interpretation, multi-step reasoning, semantic matching between queries and documents, and nuanced information synthesis. This architectural mismatch manifests in two significant ways: either severe perfor-

*Chao Huang is the corresponding author

mance degradation where accuracy drops, or complete system failure where certain advanced RAG frameworks become entirely inoperable when transitioning from LLMs to SLMs.

To address these fundamental challenges, we propose MiniRAG, a novel RAG system that reimagines the information retrieval and generation pipeline with a focus on extreme simplicity and computational efficiency. Our design is motivated by three fundamental observations about Small Language Models (SLMs): (1) while they struggle with sophisticated semantic understanding, they excel at pattern matching and localized text processing; (2) explicit structural information can effectively compensate for limited semantic capabilities; and (3) decomposing complex RAG operations into simpler, well-defined steps can maintain system robustness without requiring advanced reasoning capabilities. These insights lead us to prioritize structural knowledge representation over semantic complexity, marking a significant departure from traditional LLM-centric RAG architectures.

Our design of MiniRAG is motivated by three fundamental observations: (1) while SLMs struggle with semantic understanding, they excel at pattern matching and localized text processing; (2) explicit structural information can compensate for limited semantic capabilities by providing navigational cues for retrieval; and (3) decomposing complex RAG operations into simpler, well-defined steps can maintain system robustness without requiring advanced reasoning capabilities. These insights lead us to prioritize structural knowledge representation over semantic complexity, and to leverage graph-based patterns that naturally align with SLMs’ strengths while circumventing their limitations. This design philosophy enables MiniRAG to achieve efficient and reliable performance even with lightweight models, marking a significant departure from traditional LLM-centric RAG architectures.

Our MiniRAG introduces two key technical innovations that leverage these insights: (1) a semantic-aware heterogeneous graph indexing mechanism that systematically combines text chunks and named entities in a unified structure, reducing reliance on complex semantic understanding, and (2) a lightweight topology-enhanced retrieval approach that utilizes graph structures and heuristic search patterns for efficient knowledge discovery. Through careful design choices and architectural optimization, these components work synergistically to enable robust RAG functionality even with limited model capabilities, fundamentally reimagining how RAG systems can operate within the constraints of SLMs while leveraging their strengths.

Through extensive experimentation across datasets and Small Language Models, we demonstrate MiniRAG’s exceptional performance: compared to existing lightweight RAG systems, MiniRAG achieves 1.3-2.5x higher effectiveness while using only 25% of the storage space. When transitioning from LLMs to SLMs, our system maintains remarkable robustness, with accuracy reduction ranging from merely 0.8% to 20% across different scenarios. Most notably, MiniRAG consistently achieves state-of-the-art performance across all evaluation settings, including tests on two comprehensive datasets with four different SLMs, while maintaining a lightweight footprint suitable for resource-constrained environments such as edge devices and privacy-sensitive applications. To facilitate further research in this direction, we also introduce LiHuaWorld, a comprehensive benchmark dataset specifically designed for evaluating lightweight RAG systems under realistic on-device scenarios such as personal communication and local document retrieval.

2 THE MINIRAG FRAMEWORK

In this section, we present the detailed architecture of our proposed MiniRAG framework. As illustrated in Fig.1, MiniRAG consists of two key components: (1) heterogeneous graph indexing (Sec.2.1), which creates a semantic-aware knowledge representation, and (2) lightweight graph-based knowledge retrieval (Sec.2.2), which enables efficient and accurate information retrieval.

2.1 HETEROGENEOUS GRAPH INDEXING WITH SMALL LANGUAGE MODELS

In resource-constrained RAG systems, Small Language Models (SLMs) introduce significant operational constraints that impact their effectiveness. These limitations primarily manifest in two critical areas: i) reduced capability to extract and understand complex entity relationships and subtle contextual connections from raw text, and ii) diminished capacity to effectively summarize large volumes of text and process retrieved information containing noise and irrelevant content.

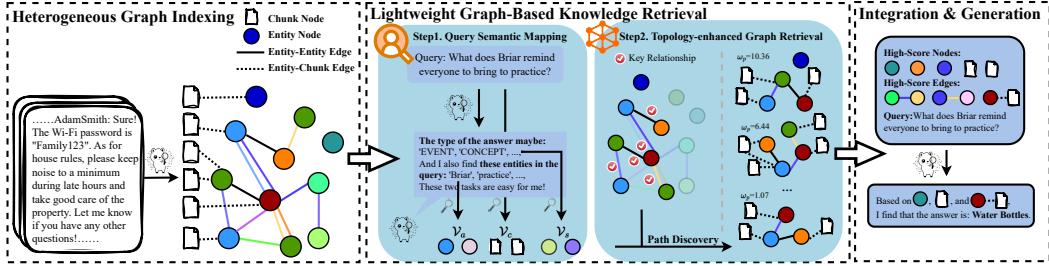


Figure 1: The MiniRAG employs a streamlined workflow built on the key components: heterogeneous graph indexing and lightweight graph-based knowledge retrieval. This architecture addresses the unique challenges faced by on-device RAG systems, optimizing for both efficiency and effectiveness.

As shown in Fig.2, comparing SLM (Phi-3.5-mini (Abdin et al., 2024)) with LLM (gpt-4o-mini (OpenAI, 2023)) reveals these limitations in practice. While both models identify the "HOUSE RULES" entity, the SLM's description lacks specific details and fails to capture the rules and purposes present in the original text (Limitation 1). Furthermore, during the answering phase, SLMs struggle to locate relevant information within extensive contexts, often becoming distracted by irrelevant content - a challenge not faced by LLMs (Limitation 2).

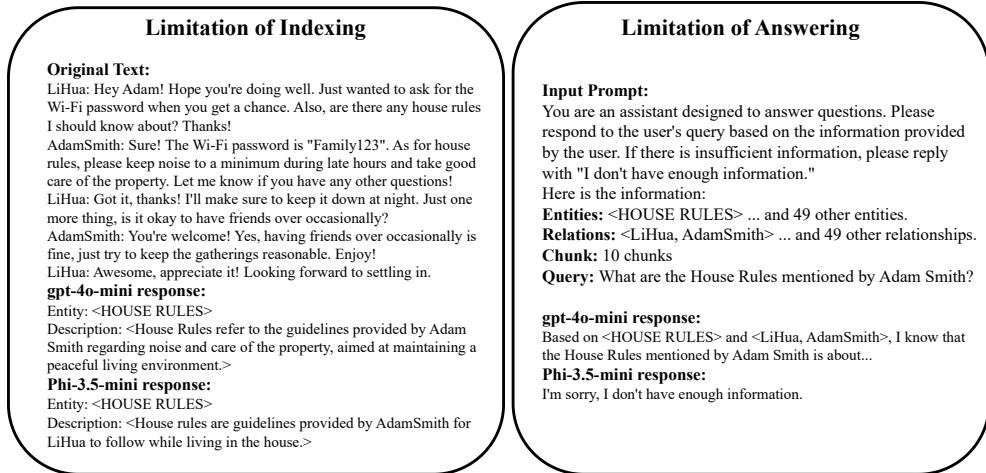


Figure 2: Compared to Large Language Models (LLMs), Small Language Models (SLMs) show significant limitations in both indexing and answering phases. **Left:** SLMs generate notably lower-quality descriptions than LLMs. **Right:** When processing identical inputs, SLMs struggle to locate relevant information in large contexts, while LLMs perform this task effectively.

To address the challenges in resource-constrained RAG systems effectively, our MiniRAG aims to:

- The indexing mechanism should extract the key relationships and contextual connections within the data, despite the small models' reduced entity understanding and summarization capacity.
- The indexing approach should condense the retrieved content to its most query-relevant elements, thereby minimizing potential distractions or misleading information that could impair the small model's capabilities in both summarization and effective denoising of the retrieved content.

To achieve the above goals, we propose a data indexing mechanism that generates a **Semantic-Aware Heterogeneous Graph**. This graph structure systematically incorporates both text chunks and named entities extracted from the raw text, creating a rich semantic network that facilitates precise information retrieval. In the constructed heterogeneous graph, the nodes comprise two primary types:

- **Text Chunk Node \mathcal{V}_c :** Coherent segments of the original text that preserve contextual integrity.

-
- **Entity Node \mathcal{V}_e :** The key semantic elements extracted from chunks include events, locations, temporal references, and domain-specific concepts that serve to anchor semantic understanding.

This dual-node design enables data chunks to directly participate in the retrieval stage, ensuring identification of the most contextually relevant content. This approach effectively mitigates information distortion that could arise from the limited summarization capabilities of the small language model. Within the heterogeneous graph, the connecting edges between nodes fall into two fundamental types:

- **Entity-Entity Connections \mathcal{E}_α :** Links between named entities that capture semantic relationships, hierarchical structures, and temporal or spatial dependencies.
- **Entity-Chunk Connections \mathcal{E}_β :** Bridges between named entities and their corresponding contexts from which the entities are extracted, preserving contextual relevance and semantic coherence.

These connections and inter-dependencies are established through language models' semantic understanding capabilities. For example, when *indexing a document that plans a trip to the 2024 Paris Olympics*, the model might establish entity-entity connections between venue locations (Stade de France), event schedules (swimming finals), transportation options (Metro Line 13), and nearby attractions (Eiffel Tower), while creating entity-chunk connections linking these entities to relevant text segments discussing ticket availability, local accommodation reviews, and optimal travel routes.

To further facilitate the relational semantic understanding, we enhance each edge in the knowledge graph with semantic descriptions generated by language models. These descriptions provide explicit relationship context between connected nodes. Specifically, for each edge $e_\beta \in \mathcal{E}_\beta$ that connects an entity to its corresponding chunk, we employ the language model to generate a description d_{e_β} of the entity as supplementary information for this edge. This description provides detailed content about the entity and reflects the semantic relationship between the extracted entity and the relevant chunk. With the textual description of the entity-chunk edge, it becomes a text-attributed edge $(e_\beta, d_{e_\beta}) \in \mathcal{E}_\beta$. In summary, the indexing process within our MiniRAG framework yields a heterogeneous graph \mathcal{G} that encompasses both entity and chunk nodes with semantic-aware connections as follows:

$$\mathcal{D} = \mathcal{G} = (\{\mathcal{V}_c, \mathcal{V}_e\}, \{\mathcal{E}_\alpha, (e_\beta, d_{e_\beta}) \in \mathcal{E}_\beta\}) \quad (1)$$

2.2 LIGHTWEIGHT GRAPH-BASED KNOWLEDGE RETRIEVAL

In on-device Retrieval Augmented Generation (RAG) systems, the limitations of device computational capabilities and data privacy restrict the use of powerful models, such as large language models and advanced text embedding models, necessitating reliance on smaller alternatives. Consequently, currently used pipelines heavily rely on LLMs for a comprehensive understanding of text semantics when computing embedding similarity for retrieval, facing significant challenges. These smaller models often struggle to capture the precise semantic nuances within lengthy texts, complicating accurate matching. To tackle these challenges, it is essential to: i) **Reduce the complexity of input content** for generation, ensuring that semantic information is clear and concise; ii) **Shorten the length of input content** for smaller language models, facilitating improved comprehension and retrieval accuracy. Additionally, employing effective graph indexing structures can help mitigate performance deficiencies in semantic matching, thereby enhancing the overall retrieval process.

In MiniRAG, we propose a **Graph-based Knowledge Retrieval** mechanism that effectively leverages the semantic-aware heterogeneous graph \mathcal{G} constructed during the indexing phase, in conjunction with lightweight text embeddings, to achieve efficient knowledge retrieval. By employing a graph-based search design, we aim to ease the burden on precise semantic matching with large language models. This approach facilitates the acquisition of rich and accurate textual content at a low computational cost, thereby enhancing the ability of language models to generate precise responses.

2.2.1 QUERY SEMANTIC MAPPING

In the retrieval phase, the primary goal for a user-input query q is to identify elements related to the query (e.g., text chunks) from the constructed index data, thereby aiding the model in generating accurate responses. To achieve this, it is essential to first parse the query and align it with the index data. Some prior RAG methods utilize LLMs to expand or decompose the query into fine-grained queries (Chan et al., 2024; Edge et al., 2024a; Guo et al., 2024), enhancing the match between the

query and the index data. However, this process relies on LLMs to extract high-quality abstract information from the query, which poses challenges for smaller language models. Therefore, in the retrieval process of MiniRAG, we leverage entity extraction—a relatively simple and effective task for small language models—to facilitate the decomposition and mapping of the query q to our graph-based indexing data (i.e., the semantic-aware heterogeneous graph \mathcal{G}).

For a given q , our approach begins with a two-stage entity processing pipeline. First, we employ a small language model to extract relevant entities \mathcal{V}_q from q while simultaneously predicting their potential types (e.g., event, location, person) that may directly contribute to the query’s answer. Following this, we leverage a lightweight sentence embedding model to evaluate semantic similarities across all entity nodes \mathcal{V}_e in the constructed graph $\mathcal{G} = \{\mathcal{V}_c, \mathcal{V}_e\}$, examining various text corpora (i.e., entity names, chunk content) to enable effective node retrieval and grounding.

Query-Driven Reasoning Path Discovery. Within our semantic-aware heterogeneous graph $\mathcal{G} = \mathcal{V}_c, \mathcal{V}_e, \mathcal{M}$ constructs reasoning paths through an intelligent query-guided mechanism. For any input query q , the model identifies relevant text chunks by jointly considering two key aspects: (1) semantic relevance between query and entity nodes, and (2) structural coherence among entity-entity and entity-chunk relationships. This dual-objective optimization framework simultaneously maximizes q - \mathcal{V}_e semantic alignment and preserves $(\mathcal{V}_e\text{-}\mathcal{V}_e)$, $(\mathcal{V}_e\text{-}\mathcal{V}_c)$ relational dependencies, effectively capturing complex reasoning chains within the heterogeneous knowledge graph. The systematic query-relevant reasoning path discovery procedure consists of the following key steps:

- **Initial Entity Identification** ($\hat{\mathcal{V}}_s$): We locate high-confidence starting points by matching query entities with graph nodes, establishing reliable entry points for path exploration.
- **Answer-Aware Entity Selection** ($\hat{\mathcal{V}}_a$): Leveraging predicted answer types, we identify candidate answer nodes from the starting set, enabling type-guided reasoning.
- **Context-Rich Path Formation** ($\hat{\mathcal{V}}_c$): We enrich reasoning paths by incorporating relevant text chunks, creating comprehensive evidence chains that connect query entities to potential answers.

This lightweight framework maintains high efficiency while ensuring semantic accuracy, making it particularly suitable for edge computing scenarios. The subsequent section details our search algorithm for further refining these reasoning paths through importance-based ranking.

2.2.2 TOPOLOGY-ENHANCED GRAPH RETRIEVAL

To address the fundamental limitations of small language models-based methods in knowledge retrieval, we propose a topology-aware retrieval approach that effectively combines semantic and structural information from heterogeneous knowledge graphs. Approaches relying on small language models with limited semantic understanding capabilities, often introduce substantial noise into the retrieval process due to their constrained ability to capture nuanced meanings, contextual variations, and complex entity relationships in real-world knowledge graphs. Our method overcomes these inherent challenges through a carefully designed two-stage process that synergistically leverages both embedding-based similarities and topological structure of the knowledge graph.

The process begins with embedding-based similarity search to identify seed entities $(\hat{\mathcal{V}}_s, \hat{\mathcal{V}}_a)$ through semantic matching, followed by a topology-enhanced discovery phase that leverages the heterogeneous graph structure \mathcal{G} to discover relevant reasoning paths. By integrating entity-specific relevance scores, structural importance metrics, and path connectivity patterns, our approach achieves superior precision in knowledge retrieval while maintaining computational efficiency, ultimately enabling more accurate and interpretable reasoning paths for enhanced generation tasks.

- **Key Relationship Identification:** We first identify high-quality entity-entity connections within graph \mathcal{G} that are relevant to query q through node-edge interactions. In the entity-entity connections \mathcal{E}_{α} , we define an edge as highly relevant if it connects a starting node $\hat{v}_s \in \hat{\mathcal{V}}_s$ to an answer node $\hat{v}_a \in \hat{\mathcal{V}}_a$ along their shortest path. For efficient extraction, we focus on edges proximate to starting or answer nodes. The relevance scoring function $\omega_e(\cdot)$ for each edge $e\alpha \in \mathcal{E}_{\alpha}$ is formally defined as:

$$\omega_e(e) = \sum_{\hat{v}_s \in \hat{\mathcal{V}}_s} \text{count}(\hat{v}_s, \hat{\mathcal{G}}_{e,k}) + \sum_{\hat{v}_a \in \hat{\mathcal{V}}_a} \text{count}(\hat{v}_a, \hat{\mathcal{G}}_{e,k}), \quad (2)$$

where $\hat{\mathcal{G}}_{e,k}$ denotes the k -hop subgraph centered at edge e , encompassing all nodes and edges reachable within k steps from either endpoint. Based on the computed relevance scores ω_e , we construct the key relationships set $\hat{\mathcal{E}}_\alpha$ by selecting the top-ranked edges.

- **Query-Guided Path Discovery:** To systematically discover logically relevant information within our knowledge graph structure, we identify and extract significant paths that serve as meaningful reasoning chains. A reasoning path starts from a carefully selected seed node and progressively advances toward potential answer nodes while maximizing the incorporation of previously extracted key relationships. For each candidate starting node \hat{v}_s in our graph, we comprehensively define the potential reasoning path set $\mathcal{P}\hat{v}_s$ as the collection of all possible acyclic paths of length n originating from \hat{v}_s . For each identified query entity $v_q \in \mathcal{V}_q$, we systematically evaluate these potential paths using a sophisticated entity-conditioned score function $\omega_p(\cdot)$ that quantifies both the overall path importance and query relevance through multiple dimensions:

$$\omega_p(p | v_q) = \omega_v(\hat{v}_s | v_q) \cdot (1 + \sum_{v \in (p \wedge \hat{\mathcal{V}}_a)} \text{count}(v, p) + \sum_{e \in (p \wedge \hat{\mathcal{E}}_\alpha)} \omega_e(e)). \quad (3)$$

The scoring components in our path discovery framework are defined as follows: $\omega_v(\hat{v}_s | v_q)$ measures the semantic similarity between starting node \hat{v}_s and query entity v_q using cosine similarity of their respective embeddings in the vector space, while $\text{count}(v, p)$ serves as a binary indicator function that returns 1 if node v appears in path p and 0 otherwise. For each query entity and starting node pair in the graph, we systematically rank all potential paths according to their computed importance scores and construct the final comprehensive set of reasoning paths \mathcal{P}_q by carefully selecting the top- k highest-scoring paths from each ranking list for subsequent steps.

- **Retrieval of Query-Relevant Text Chunks:** Building upon our indexing structure from Section 2.1, each entity node maintains connections with its source text chunk through entity-chunk inter-dependencies. These text chunks exist as nodes in our indexing graph, connected via text-attributed edges $(e_\beta, d_{e_\beta}) \in \mathcal{E}_\beta$. By traversing these connections, we collect all chunk nodes \mathcal{V}_c^q that are linked to entity nodes present in any reasoning path $p \in \mathcal{P}_q$. **Step 1: Candidate Filtering.** We first systematically filter the candidates to focus on the intersection $\hat{\mathcal{V}}_c \wedge \mathcal{V}_c^q$ to ensure coverage of relevant information. **Step 2: Similarity Computation.** For each candidate chunk in the intersection, we carefully calculate the semantic similarity between the input query and the concatenated content, which combines both the chunk text and its associated edge descriptions. **Step 3: Ranking and Selection.** Finally, we rank these filtered chunks according to their computed relevance scores and select the top-scoring candidates to form the final optimized set $\hat{\mathcal{V}}_c^q$ for subsequent reasoning.
- **Integration for Augmented Generation:** Through our proposed topology-enhanced graph retrieval mechanism and multi-stage filtering process, we efficiently obtain two key components of query-relevant graph knowledge: i) Essential relationships $\hat{\mathcal{E}}_\alpha$ connecting important entities within the knowledge graph, which capture the semantic dependencies and structural patterns; ii) Optimal text chunks $\hat{\mathcal{V}}_c^q$ containing critical contextual information and supporting evidence necessary for accurate answer generation. By systematically integrating these retrieved components with the previously grounded answer nodes $\hat{\mathcal{V}}_a$ through our designed fusion strategy, we construct the comprehensive and well-structured input representation for the final augmented generation process.

3 EVALUATION

Through the novel design of MiniRAG, we enable efficient on-device RAG systems without relying on large models, preserving data privacy while maintaining robust performance. Our evaluation addresses three key research questions (RQs): • **RQ1: Comparative Performance.** How does MiniRAG perform against state-of-the-art alternatives in terms of retrieval accuracy and efficiency? • **RQ2: Component Analysis.** What is the contribution of key components to MiniRAG’s overall effectiveness? • **RQ3: Case Studies.** How effectively does MiniRAG handle complex, multi-step reasoning tasks with small language models, as demonstrated through practical case studies?

3.1 EXPERIMENTAL SETTINGS

Datasets. The evaluation of on-device RAG requires careful consideration of their unique context and practical use cases. While traditional server-side RAG systems are designed to process extensive

documents such as academic papers, technical reports, and comprehensive web content, on-device RAG applications serve fundamentally different purposes aligned with users’ daily device interactions.

Our dataset selection reflects these requirements, focusing on common on-device scenarios including **Instant Messaging** (real-time retrieval from chat histories and personal communications, and emails), **Personal Content** (user-created notes, memos, and calendar entries), and **Local Short Documents** (lightweight text files). This composition aligns with the core strengths of on-device RAG systems - privacy preservation, real-time processing, and efficient personal content management. By focusing on these everyday user interactions rather than complex document processing, our evaluation framework provides a realistic assessment of on-device RAG capabilities in their intended use cases.

Our evaluation employs two datasets (detailed in Appendix Section) that capture essential aspects of real-world, on-device RAG scenarios. The key characteristics of these datasets are as follows:

- **Synthetic Personal Communication Data:** To comprehensively capture real-world personal communications, we leverage GPT-4 to generate a year-long dataset that authentically mirrors the full spectrum of daily life interactions. This extensive dataset encompasses diverse aspects of modern living, including daily necessities (*e.g.*, food, clothing, housing, transportation), social activities and entertainment, work and study-related discussions, personal schedule planning, and shopping decisions. The conversations reflect natural communication patterns across various contexts - from casual chats and task coordination to information sharing and decision making. By utilizing GPT-4’s advanced generation capabilities with effective prompting mechanism, we ensure the dataset maintains realistic temporal coherence and contextual relationships while preserving privacy, making it ideal for evaluating both “Instant Messaging” and “Personal Content” use cases.

- **Short Documents:** We utilize a multi-hop RAG dataset based on contemporary news articles, specifically designed to evaluate the system’s capability in navigating and retrieving information across multiple short documents. This dataset mirrors the real-world scenarios, where users frequently need to retrieve relevant information from various locally stored files. This setup enables comprehensive assessment of both retrieval efficiency and accuracy when handling cross-document information access - for the “Local Short Documents” use case in on-device applications.

Evaluation Protocols and Metrics. We employ two key metrics to assess the quality and reliability of responses generated by various RAG methods. • **Accuracy (*acc*):** Measures the consistency between the RAG system’s response and the expected answer. For instance, given the query “What does Briar remind everyone to bring to practice?” with the expected answer “bottle”, semantically equivalent responses like “water bottle” are considered correct. • **Error Rate (*err*):** Captures instances where the RAG system provides incorrect information without recognizing its mistake. For example, if the system confidently responds with “yoga mat” to the above query, it would count toward the error rate.

Implementation Details. We configure our experimental setup following established practices from prior work (Guo et al., 2024). For text processing, we set the chunk size to 1200 tokens with an overlap of 100 tokens, and utilize nano vector base for vector storage. In our MiniRAG implementation, we configure the top-k retrieval to 5 documents and set the maximum token limit to 6000 tokens.

For the model selection, we employ different efficient configurations for large and small language models. In the advanced LLM setting, we use gpt-4o-mini (OpenAI, 2023) as the language model and text-embedding-3-small as the specialized embedding model. For the lightweight SLM setting, we utilize optimized all-MiniLM-L6-v2 (Reimers & Gurevych, 2019) as the embedding model, paired with various small language models including Phi-3.5-mini-instruct (Abdin et al., 2024), GLM-Edge-1.5B-Chat, Qwen2.5-3B-Instruct (Team, 2024b), and MiniCPM3-4B (Hu et al., 2024).

Baselines. We compare our MiniRAG against several representative RAG systems:

- **NaiveRAG** (Mao et al., 2020) serves as the standard RAG baseline, employing text embedding-based retrieval. It segments documents into chunks stored in a vector database and performs retrieval through direct similarity matching between query and chunk embeddings.
- **GraphRAG** (Edge et al., 2024b) leverages graph-based indexing through language models and the Leiden algorithm for entity clustering. It enhances retrieval by generating community reports and combining local-global information access through a unified retrieval mechanism.

Table 1: Performance evaluation using accuracy (acc) and error (err) rates, measured as percentages (%). Higher accuracy and lower error rates indicate better RAG performance. Results compare various baseline methods against our MiniRAG across multiple datasets. Bold values indicate best performance, while “/” denotes cases where methods failed to generate effective responses.

LiHuaWorld	NaiveRAG		GraphRAG		LightRAG		MiniRAG	
	acc↑	err↓	acc↑	err↓	acc↑	err↓	acc↑	err↓
Phi-3.5-mini-instruct	41.22%	23.20%	/	/	39.81%	25.39%	53.29%	23.35%
GLM-Edge-1.5B-Chat	42.79%	24.76%	/	/	35.74%	25.86%	52.51%	25.71%
Qwen2.5-3B-Instruct	43.73%	24.14%	/	/	39.18%	28.68%	48.75%	26.02%
MiniCPM3-4B	43.42%	17.08%	/	/	35.42%	21.94%	51.25%	21.79%
gpt-4o-mini	46.55%	19.12%	35.27%	37.77%	56.90%	20.85%	54.08%	19.44%

MultiHop-RAG	NaiveRAG		GraphRAG		LightRAG		MiniRAG	
	acc↑	err↓	acc↑	err↓	acc↑	err↓	acc↑	err↓
Phi-3.5-mini-instruct	42.72%	31.34%	/	/	27.03%	11.78%	49.96%	28.44%
GLM-Edge-1.5B-Chat	44.44%	24.26%	/	/	/	/	51.41%	23.44%
Qwen2.5-3B-Instruct	39.48%	31.69%	/	/	21.91%	13.73%	48.55%	33.10%
MiniCPM3-4B	39.24%	31.42%	/	/	19.48%	10.41%	47.77%	26.88%
gpt-4o-mini	53.60%	27.19%	60.92%	16.86%	64.91%	19.37%	68.43%	19.41%

- **LightRAG** (Guo et al., 2024) implements a dual-level retrieval architecture with knowledge graphs. It enhances query understanding by decomposing queries into hierarchical components (low-level details and high-level concepts), enabling more precise document retrieval.

3.2 PERFORMANCE ANALYSIS (RQ1)

• **Performance Degradation in Existing RAG Systems with SLMs**. Current RAG systems face critical challenges when operating with Small Language Models (SLMs), revealing fundamental vulnerabilities in their architectures. Advanced LLM-based RAG methods exhibit severe performance degradation, with LightRAG’s accuracy plummeting from 56.90% to 35.42% during LLM to SLM transition, while GraphRAG experiences complete system failure due to its inability to generate high-quality content. While basic retrieval systems like NaiveRAG show resilience, they suffer from significant limitations, being restricted to basic functionality and lacking advanced reasoning capabilities. This performance analysis highlights a critical challenge: existing advanced systems’ over-reliance on sophisticated language capabilities leads to fundamental operational failures when using simpler models, creating a significant barrier to widespread adoption in resource-constrained environments, where high-end language models may not be available or practical to deploy.

• **MiniRAG’s Unique Advantages**. These innovations enable MiniRAG to maintain strong performance even with simpler language models, making it particularly suitable for resource-constrained environments while preserving the core functionalities of RAG systems.

i) **Semantic-Aware Graph Indexing for Reduced Model Dependency**. MiniRAG fundamentally reimagines the indexing process through a dual-node heterogeneous graph structure. Instead of relying on powerful text generation capabilities, the system focuses on basic entity extraction and heterogeneous relationship mapping. This design combines text chunk nodes for preserving raw contextual information with entity nodes for capturing key semantic elements, creating a robust knowledge representation that remains effective even with limited language model capabilities.

ii) **Topology-Enhanced Retrieval for Balanced Performance**. MiniRAG employs a lightweight graph-based retrieval mechanism that balances multiple information signals through a systematic process. Beginning with query-driven path discovery, the system integrates embedding-based matching with structural graph patterns and entity-specific relevance scores. Through topology-aware search and optimized efficiency, it achieves robust retrieval quality without requiring advanced language understanding, making it particularly effective for on-device deployment.

Table 2: Ablation study results comparing accuracy (acc, \uparrow) and error rate (err, \downarrow) (%) across architectural variants: baseline MiniRAG versus variants with (i) semantic-driven indexing replacement (- \mathcal{I}), (ii) edge information removal (- \mathcal{R}_{edge}), and (iii) chunk nodes removal (- \mathcal{R}_{chunk}). Results validate SLM limitations and the effectiveness of query-guided reasoning path components.

LiHuaWorld	MiniRAG		- \mathcal{I}		- \mathcal{R}_{chunk}		- \mathcal{R}_{edge}	
	$acc \uparrow$	$err \downarrow$	$acc \uparrow$	$err \downarrow$	$acc \uparrow$	$err \downarrow$	$acc \uparrow$	$err \downarrow$
Phi-3.5-mini-instruct	53.29%	23.35%	26.02%	19.12%	48.90%	17.40%	50.47%	15.36%
GLM-Edge-1.5B-Chat	52.51%	25.71%	25.08%	31.50%	46.24%	16.77%	47.81%	20.53%
Qwen2.5-3B-Instruct	48.75%	26.02%	24.14%	15.67%	40.91%	16.14%	48.43%	18.65%
MiniCPM3-4B	51.25%	21.79%	26.18%	15.52%	46.39%	15.83%	48.59%	19.44%

These innovations enable MiniRAG to maintain strong performance with simpler language models, making it ideal for resource-constrained environments while preserving core RAG functionalities.

- **Storage Efficiency While Maintaining Performance.** MiniRAG demonstrates exceptional storage efficiency while preserving high accuracy levels. Empirical evaluations show that MiniRAG achieves competitive accuracy while requiring only 25% of the storage space compared to baselines like LightRAG w/ gpt-4o-mini. This dramatic reduction in storage requirements, coupled with maintained or improved accuracy, represents a significant advancement in efficient RAG system design.

3.3 COMPONENT-WISE ANALYSIS OF MINIRAG (RQ2)

Our ablation study examines the individual contributions of key MiniRAG components through two primary experimental variations, as documented in Table 2. The first variation (- \mathcal{I}) replaces MiniRAG’s heterogeneous graph indexing with a description-based indexing approach that requires comprehensive semantic understanding for generating accurate entity/edge descriptions, similar to methods used in LightRAG and GraphRAG. The second variation (- \mathcal{R}_i) involves selectively deactivating specific modules during graph retrieval. This systematic evaluation framework provides detailed insights into how each component contributes to MiniRAG’s overall performance.

Our experimental results reveal crucial insights into MiniRAG’s architectural effectiveness. • **Validating SLM Limitations.** One key finding emerges when replacing our streamlined indexing method with text semantic-driven indexing techniques (- \mathcal{I}), resulting in substantial performance degradation. This outcome strongly validates our initial hypothesis about Small Language Models (SLMs) and their inherent limitations - specifically their constraints in comprehensive semantic understanding, which impacts both the generation of complex knowledge graphs with entity relationships and the creation of corresponding comprehensive text descriptions. • **Effectiveness of Query-guided Reasoning Path Discovery.** The experiments further demonstrate the critical nature of structural components: the removal of either edge information (- \mathcal{R}_{edge}) or chunk nodes (- \mathcal{R}_{chunk}) significantly impacts system performance. These elements serve dual purposes: they facilitate effective query matching through query-guided reasoning path discovery while simultaneously compensating for the inherent limitations of SLMs during the data indexing phase.

3.4 CASE STUDY ANALYSIS (RQ3)

In this section, we demonstrate MiniRAG’s practical advantages through a case study with LightRAG, focusing on a complex restaurant identification scenario. This study illustrates how our query-guided reasoning approach, combined with heterogeneous graph indexing, effectively handles multi-constraint queries while overcoming the inherent limitations of small language models.

- **Challenge: Complex Query Resolution in Restaurant Identification.** We conducted a comparative case study between MiniRAG and LightRAG using a complex query scenario: “What is the name of the Italian restaurant where Wolfgang and Li Hua are having dinner to celebrate Wolfgang’s

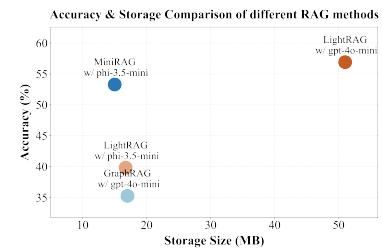


Figure 3: Accuracy vs. Storage Efficiency: Comparative analysis of three RAG systems - MiniRAG, LightRAG, and GraphRAG.

Table 3: Case study comparing LightRAG and MiniRAG on a complex restaurant identification query, demonstrating how query-guided reasoning path discovery effectively addresses small language model (SLM) limitations in multi-constraint information retrieval tasks.

Query: What is the name of the Italian restaurant where Wolfgang and Li Hua are having dinner to celebrate Wolfgang's promotion?
Ground-Truth: Venedia Grancaffe
LightRAG Source: Given the query, I decompose it to low-level information : <"Wolfgang"><"LiHua"><"Restaurant name"> and high-level information : <"Italian restaurant"><"Dinner celebration"><"Promotion">. Using these information, I find these elements: Entities :<"FRIES"><"COLLABORATION"><"WOLFGANGSCHULZ"><"HAILEY'S BAKERY"><"LiHUA"><"JAM SESSION">... Relations :<"ORGANIZATION", "PIZZA ORDER"><"LIHUA", "WOLFGANGSCHULZ"> <"FARRAH", "PIZZA"><"LUNCH", "THANECHAMBERS">... Texts :<Time: 20260318_15:30><Time: 20260508_08:00><Time: 20260514_14:00><Time: 20261103_13:00><Time: 20260326_16:00>...
LightRAG Response: Wolfgang and Li Hua are planning to have dinner at a pasta joint to celebrate Wolfgang's promotion...The specific name of the restaurant wasn't listed in the provided information...
MiniRAG Source: Given the query, I guess the answer type could be: <"SOCIAL INTERACTION"><"LOCATION"> and I found the entities from query are: <"Italian restaurant"><"Wolfgang"><"LiHua"><"dinner"><"promotion">. Using these information, I find these elements: Entities from query : <"ITALIAN PLACE"><"RESTAURANT"><"WOLFGANG"><"WOLFGANG SCHULZ"><"LIHUA"><"DINNER PARTY">... Candidate answer entity : ...<"ITALIAN PLACE"><"GRATITUDE EXPRESSION"><"NEW PIZZA PLACE DOWNTOWN">... Texts : <Time: 20260908_21:00><Time: 20260529_17:00><Time: 20260523_14:00><Time: 20260430_17:00><Time: 20260419_21:00> MiniRAG Response: The name of the Italian restaurant where Wolfgang and Li Hua are having dinner to celebrate Wolfgang's promotion is Venedia Grancaffe . Wolfgang mentioned it in a reminder about the dinner planned for 6 PM. They were excited to celebrate his promotion with a feast, with both looking forward to enjoying some delicious food together....

"promotion?" This query presents multiple challenges, requiring the system to identify specific Italian restaurants from various mentions in online chat data while correlating them with the context of a promotion celebration. LightRAG, despite its capabilities, struggled with this task due to the limitations of its underlying small language model (phi-3.5-mini-instruct). The SLM's constraints in extracting appropriate high-level information, combined with noise in the graph-based index, led to ineffective direct embedding matching and ultimately hindered accurate answer retrieval.

- **MiniRAG's Effective Query-Guided Knowledge Discovery.** MiniRAG successfully resolved this challenge through its query-guided reasoning path discovery mechanism, which enables precise and contextually relevant knowledge retrieval. By leveraging its heterogeneous graph indexing structure, MiniRAG effectively constructs query-relevant knowledge paths, starting with answer type prediction ("Social Interaction" or "Location") and proceeding through targeted entity matching. This structured reasoning approach, combined with strategic decomposition of query elements (focusing on "Italian place" and "restaurant" contexts), allows MiniRAG to navigate the knowledge space efficiently. The synergy between query-guided reasoning and heterogeneous graph indexing enabled MiniRAG to effectively filter through multiple Italian establishments, ultimately identifying "*Venedia Grancaffe*" as the venue specifically connected to the promotion celebration context.

4 RELATED WORKS

Small Language Models (SLMs). The emergence of Small Language Models (SLMs) is driven by the growing need for lightweight, efficient and privacy-preserving AI solutions that can operate

on edge devices, addressing the limitations of large language models in computational resources and deployment costs (Liu et al., 2024; Wang et al., 2024a; Team, 2024b; Abdin et al., 2024; Hu et al., 2024). Recent developments have produced notable models such as MiniCPM3-4B (Hu et al., 2024), phi-3.5-mini (Abdin et al., 2024), Llama-3.2-3B (Grattafiori et al., 2024), Qwen2.5-1.5B (Team, 2024b), gemma-2-2b (Team, 2024a), SmoLLM-1.7B (Allal et al., 2024), and MobiLlama-1B (Thawakar et al., 2024), which have demonstrated impressive performance while maintaining significantly smaller parameter counts. These models excel in inference speed, deployment flexibility, and privacy preservation, making them particularly suitable for resource-constrained environments.

For facilitating visual-language understanding in resource-constrained environments, researchers have developed Multi-modal SLMs by efficiently extending single-modal SLMs with visual capabilities, as exemplified by MiniCPM-V 2.0 (Yao et al., 2024), Qwen2-VL (Wang et al., 2024b), Phi-3-vision (Abdin et al., 2024), and InternVL2-2B (Chen et al., 2024). These multi-modal SLMs have demonstrated remarkable capabilities in combining visual and textual understanding while maintaining the computational efficiency advantages of small-scale models, enabling their deployment in diverse applications from GUI Agent (Lin et al., 2024) to robotics control (Zheng et al., 2024).

While SLMs have demonstrated impressive capabilities in language understanding and multi-modal tasks, the potential of leveraging efficient models for RAG tasks remains largely unexplored. This work fills this gap by introducing a framework that enables SLMs to effectively perform RAG tasks while maintaining their inherent advantages in computational efficiency and deployment flexibility.

Retrieval-Augmented Generation. Retrieval-Augmented Generation (RAG) systems enhance language models’ responses by retrieving relevant knowledge from external databases (Guo et al., 2024; Qian et al., 2024; Gao et al., 2024b). The process consists of three main components: indexing, retrieval, and generation. Given a raw text set, the system first processes it into a database, then retrieves relevant information based on user query q , and finally generates the answer. Two primary approaches have emerged in database construction: (1) chunks-based methods (Mao et al., 2020; Qian et al., 2024), which segment texts into retrievable units, and (2) graph-based methods (Guo et al., 2024; Edge et al., 2024b), which structure information as knowledge graphs.

Recent RAG implementations have evolved along these two paths. Chunks-based methods, exemplified by NaiveRAG (Mao et al., 2020), ChunkRAG (ukhvinder Singh et al., 2024), and RQ-RAG (Chan et al., 2024), focus on optimizing text segmentation and chunk retrieval strategies. Graph-based approaches, such as GraphRAG (Edge et al., 2024b), LightRAG (Guo et al., 2024), and SubgraphRAG (Li et al., 2024), leverage graph structures for better corpus comprehension and efficient retrieval. However, most existing methods require either large context windows (Qian et al., 2024; Li et al., 2024) or strong semantic understanding capabilities (Guo et al., 2024; Edge et al., 2024b), limiting their applicability with small and lightweight language models. This gap motivates the development of more efficient RAG systems suitable for resource-constrained scenarios.

5 CONCLUSION

We present MiniRAG, a novel RAG system designed to address the fundamental limitations of deploying small language models (SLMs) in existing retrieval-augmented generation frameworks. Through its innovative heterogeneous graph indexing and lightweight heuristic retrieval mechanisms, MiniRAG effectively integrates the advantages of both text-based and graph-based RAG approaches while significantly reducing the demands on language model capabilities. Our experimental results demonstrate that MiniRAG can achieve comparable performance to LLM-based methods even when using SLMs. Additionally, to facilitate research in this emerging field, we release the first benchmark dataset specifically designed for evaluating on-device RAG capabilities, featuring realistic personal communication scenarios and multi-constraint queries. These contributions mark an important step toward enabling private, efficient, and effective on-device RAG systems, opening new possibilities for edge device AI applications while preserving user privacy and resource efficiency.

REFERENCES

- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. Smollm - blazingly fast and remarkably powerful, 2024.
- Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610*, 2024.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024a.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2024b. URL <https://arxiv.org/abs/2404.16130>.
- Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 6491–6501, 2024.
- Dawei Gao, Zitao Li, Xuchen Pan, Weirui Kuang, Zhijian Ma, Bingchen Qian, Fei Wei, Wenhao Zhang, Yuexiang Xie, Daoyuan Chen, et al. Agentscope: A flexible yet robust multi-agent platform. *arXiv preprint arXiv:2402.14034*, 2024a.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024b.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. 2024.
- Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.
- Mufei Li, Siqi Miao, and Pan Li. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation, 2024. URL <https://arxiv.org/abs/2410.20724>.
- Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Zechen Bai, Weixian Lei, Lijuan Wang, and Mike Zheng Shou. Showui: One vision-language-action model for generalist gui agent. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024.
- Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, et al. Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. In *International Conference on Machine Learning (ICML)*, 2024.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*, 2020.

OpenAI. Gpt-4 technical report, 2023.

Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery, 2024. URL <https://arxiv.org/abs/2409.05591>.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL <https://arxiv.org/abs/1908.10084>.

Gemma Team. Gemma: Open models based on gemini research and technology, 2024a. URL <https://arxiv.org/abs/2403.08295>.

Qwen Team. Qwen2.5: A party of foundation models, September 2024b. URL <https://qwenlm.github.io/blog/qwen2.5/>.

Omkar Thawakar, Ashmal Vayani, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Michael Felsberg, Timothy Baldwin, Eric P. Xing, and Fahad Shahbaz Khan. Mobillama: Towards accurate and lightweight fully transparent gpt, 2024.

Ishneet S ukhvinder Singh, Ritvik Aggarwal, Ibrahim Allahverdiyev, Muhammad Taha, Aslihan Akalin, Kevin Zhu, and Sean O'Brien. Chunkrag: Novel llm-chunk filtering method for rag systems, 2024. URL <https://arxiv.org/abs/2410.19572>.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhalo Mo, Qiuhalo Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness, 2024a. URL <https://arxiv.org/abs/2411.03350>.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. URL <https://arxiv.org/abs/2408.01800>.

Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies, 2024. URL <https://arxiv.org/abs/2412.10345>.

Appendix

• **Dataset Descriptions.** The rapid growth of mobile computing has led to an unprecedented accumulation of content on personal devices, creating a pressing need for efficient on-device information retrieval and generation systems. Traditional RAG benchmarks, primarily focused on well-structured documents like Wikipedia articles or academic papers, fail to capture the unique characteristics of on-device scenarios. These scenarios present distinct challenges and characteristics:

- (1) **Content Fragmentation and Context-Switching** - Unlike traditional RAG systems that process well-structured documents with clear logical flow (e.g., Wikipedia articles, academic papers), on-device RAG must handle inherently fragmented content that frequently switches between different contexts and conversation threads - a direct reflection of how people naturally communicate and interact across various digital platforms in their daily lives.
- (2) **Temporal Nature and Evolution Patterns** - Unlike traditional RAG's static, historically complete documents, on-device RAG must handle inherently dynamic content that constantly evolves through real-time updates, ongoing conversations, and emerging social interactions. This fundamental difference in temporal dynamics creates unique challenges for on-device RAG systems, which must maintain relevance and accuracy while processing an ever-changing stream of information across various digital platforms and communication channels.
- (3) **Digital-Physical Context Fragmentation**. Digital communication related to personal social activities and events typically presents fragmented and incomplete information, as these interactions span both online and offline contexts. Unlike traditional RAG systems that process self-contained documents, on-device content frequently captures only partial context of real-world events - text conversations might reference in-person meetings, shared experiences, or future plans without full details. This hybrid nature means digital communications often contain implicit references and assumed knowledge that only make sense with additional real-world context, requiring sophisticated context-bridging capabilities to effectively process and understand the information.



Figure A1: LiHuaWorld simulates a digitally interconnected world where AI agents communicate through mobile chat applications. Through the lens of our primary subject, Li Hua, we observe and collect authentic chat interactions within this virtual social ecosystem.

The LiHuaWorld dataset authentically reflects key characteristics of on-device communications, emphasizing digital-physical context fragmentation and temporal evolution patterns. Given the private nature of on-device chat data, we employed a detailed simulation approach to generate

this comprehensive dataset, implementing a year-long life journey across contemporary messaging platforms. The simulation follows our protagonist, Li Hua, through both major life events and daily social interactions that naturally span between digital conversations and physical contexts.

The dataset deliberately incorporates challenging aspects typical of on-device content: conversations span multiple contexts and threads, information develops and updates over time, and messages frequently contain implicit references to offline events. These features reflect the real-world complexity of digital communications, where content evolves dynamically and often requires bridging between online and offline contexts for complete understanding. The temporal nature of interactions is carefully preserved, showing how information and relationships develop over extended periods.

The dataset includes a wide range of scenarios, from social coordination to life transitions and daily activities. For instance, weekend plans might begin with informal group chat coordination and evolve through real-time updates, while housing-related conversations span multiple threads with partial context shared across physical viewings and digital negotiations. Such scenarios demonstrate the dataset's capacity to capture both the breadth and depth of typical on-device communications.

• **Dataset Structure.** The LiHuaWorld timeline begins with Li Hua's relocation to a new city, serving as a strategic starting point that naturally facilitates the simulation of expanding social networks and evolving relationships. Within this narrative framework, the conversation data is systematically organized into two primary categories: one-on-one chats and group chats.

One-on-One Chats

Li Hua, the primary agent in our dataset, engages in conversations with a diverse network of simulated friends. These conversational partners represent varied backgrounds, professions, relationships, and interests, enabling rich and authentic interactions throughout the timeline. Below is a representative one-on-one conversation extracted from LiHuaWorld that illustrates these dynamics:

Time: 20260819_10:00

LiHua: Hey Jake! It's really nice to meet you! I just love the enthusiasm you have for soccer!

JakeWatson: It's really nice to meet you, too, Li Hua! And I like your passion as well!

LiHua: Just wondering if you have some time this weekend to help me out with my dribbling skills? I could really use your expertise!

JakeWatson: Yeah, I'd love to help you out. How about Saturday afternoon? We can hit the field and work on your skills!

LiHua: Saturday afternoon sounds perfect! What time are you thinking?

JakeWatson: How about we meet around 3 PM? Should give us plenty of time to practice! Let me know if that works for you.

LiHua: 3 PM works for me! Looking forward to it. Thanks for helping out!

JakeWatson: No problem at all! It'll be fun, and I'm excited to see you improve. Just bring some water, and we'll have a blast!

LiHua: Definitely! I'll be ready. See you on Saturday!

Group Chat

Beyond one-on-one interactions, LiHuaWorld captures multi-participant conversations in groups. The following excerpt from the “TV Show Lovers” group chat demonstrates these collective interactions:

Time: 20260826_18:00

EmilyBurnett: Hey guys! Let's discuss about the main character Kendall's motivation in the TV series Succession! I think it will be fun to take a close look at him! I will go first!

EmilyBurnett: I think Kendall's motivations are so complex! It's like he's trying to prove himself, not just to his dad but also to the world. Plus, I feel like the pressure of the family legacy weighs heavily on him. What do you all think?

Kieran: I find it interesting how Kendall's struggle for power ties back to his childhood. The constant rivalry and need for approval from Logan really shaped him. Do you think he can ever fully break free from that dynamic?

Tamara: Totally agree! His relationship with Logan is so toxic, but Kendall keeps coming back for more validation. It's like he's in this endless cycle.

...

LiHua: The intensity of Kendall's journey really keeps us on edge! It's hard not to root for him despite everything, especially when you see how hurt he is.

Kieran: I really think it's a mix of both! On one hand, he craves that power and validation, but on the other, he seems desperate to carve out his own identity separate from Logan's shadow. It's such an interesting storyline, watching him fight that internal battle.

...

EmilyBurnett: Absolutely! The suspense makes it so much more thrilling. Plus, with all the character complexities, there's never a dull moment.

- **Event Generation with Human Oversight.** Events serve as conversation catalysts, functioning as carefully crafted scripts that guide character interactions and dialogue topics. While GPT-4-mini occasionally provides creative inspiration, our team primarily authors these events through deliberate human curation to ensure narrative coherence and authenticity. The conversation generation process is powered by **AgentScope** (Gao et al., 2024a), which transforms these event scripts into natural dialogues. Below is a representative excerpt of events from a typical week:

Time	Participants	Case
20260818_10:00	Li Hua and Thane Chambers	Thane Chambers asks Li Hua which character in the game <i>Witcher 3</i> Li Hua likes the best and why.
20260819_10:00	Li Hua and Jake Watson	Li Hua messages Jake Watson asking Jake if he has some time during the weekend to help Li Hua improve his dribbling skills.
20260820_14:00	Li Hua, Emily, and Others in TVfan group	Emily Burnett creates a poll for the group to vote on their favorite HBO series of all time.

Query Set Design. Our query set has two dimensions: event-based content and reasoning complexity. The event-based dimension encompasses six categories (When, Where, Who, What, How, and Yes/No questions), while the reasoning complexity distinguishes between single-hop and multi-hop queries based on required inferential steps. The following examples illustrate these diverse query types:

NO.	Question	Gold Answer	Support Documents	Type
1	What hobbies does Li Hua have?	photography, guitar, fitness, video game, TV show, soccer	NA	What
2	What is the Wi-Fi password at Li Hua's house?	Family123	20260106_09:00	What
3	Who is Li Hua sending a fruit basket to?	Adam	20261027_17:00	Who
4	Did Wolfgang ask Li Hua about watching "Star Wars: A New Hope" after he asked Li Hua about going to see "Overwatch 3"?	Yes	20260121_13:00 <and> 20261009_17:00	YesNo