ACE-GraphRAG: Agentic Context Engineering for Hierarchical GraphRAG

Yongfeng Huang[1,*], **Yuren Lai**[1,*], **Ruiying Chen**[2,*]
Haoyu Huang[3], **Mingming Zhao**[4], **James Cheng**[1,†]

[1]CSE, The Chinese University of Hong Kong [2]Wuhan University of Technology
[3]The Hong Kong University of Science and Technology [4]Huawei Noah's Ark Lab
{yfhuang22,jcheng}@cse.cuhk.edu.hk yuren.lai@outlook.com
355227@whut.edu.cn hhuangcp@connect.ust.hk zhaomingming9@huawei.com

Abstract

Hierarchical Graph Retrieval-Augmented Generation (GraphRAG) organizes corpus knowledge at multiple levels of granularity, yet fixed context construction may fail to translate these multi-resolution representations into a context suited to the current query. We identify this mismatch as the *representation–inference gap*. We propose **A**gentic **C**ontext **E**ngineering for Hierarchical GraphRAG (**ACE-GraphRAG**), an inference-time context policy layer that supplements and adapts the initial context for generation. ACE-GraphRAG formulates context construction as a policy over gap-aware refinement, retrieval branches, and task-conditioned adaptation. Parallel Differential Retrieval acquires supplementary evidence from depth-oriented factual and breadth-oriented semantic branches. These evidence increments are consolidated with the initial context while preserving provenance and abstraction levels. FULL-ACE applies the full policy uniformly within each task family, whereas ADAPTIVE-ACE selects task- and topology-specific policies for individual queries. We evaluate ACE-GraphRAG on HotpotQA, 2WikiMultiHopQA, and four UltraDomain subsets across multi-hop QA and query-focused summarization. FULL-ACE outperforms the evaluated RAG and GraphRAG baselines across both task families, while ADAPTIVE-ACE further improves multi-hop QA and is preferred over FULL-ACE on all four UltraDomain subsets. Ablation and topology analyses support treating context construction as a query- and task-dependent inference policy rather than a fixed procedure.

1 Introduction

Retrieval-augmented generation (RAG) enhances large language models (LLMs) with external knowledge, improving their ability to answer knowledge-intensive questions and produce grounded responses (Gao et al., 2023). Conventional RAG systems typically retrieve a small set of semantically similar text chunks, which works well for localized information needs but provides limited support for relational reasoning, multi-hop evidence aggregation, and corpus-level synthesis. Graph-based and hierarchical RAG methods address these limitations by organizing knowledge into entities, relations, communities, and summaries at multiple levels of granularity, making both fine-grained evidence and higher-level semantic evidence available during retrieval (Edge et al., 2024; Gutiérrez et al., 2024; Huang et al., 2025).

However, making evidence available at multiple levels does not ensure that it is assembled into a context suited to the current query. Hierarchical GraphRAG may retrieve entities, relations, source passages, communities, and summaries, yet these heterogeneous representations must still be transformed into a single context for generation. Different queries impose different evidence requirements: a localized entity question may need only a few precise facts, a relational or multi-hop question must preserve intermediate entities and links, and query-focused summarization benefits from broader community- and summary-level evidence. A fixed context construction strategy may omit evidence required by complex queries or introduce unnecessary evidence for simple ones. Effective context construction must therefore determine how the initial context should be supplemented and adapted to the query and downstream task, following the broader view that context can be actively constructed and refined (Zhang et al., 2025). We term this mismatch the ***representation–inference gap*** in hierarchical GraphRAG: the multi-resolution representations built during indexing are not fully trans-

[*]Equal contribution.
[†]Corresponding author.

lated into a context suited to the evidence requirements of the current query.

Prior work has addressed query-dependent evidence requirements primarily through adaptive retrieval. Existing adaptive RAG systems choose among no retrieval, single-step retrieval, and iterative retrieval according to query complexity (Jeong et al., 2024), while recent GraphRAG systems select between local and global retrieval or route between dense and graph retrieval (Zhao et al., 2025; Dong et al., 2026). These methods improve when and where evidence is retrieved, but they do not fully control how the initial context is supplemented and adapted for generation. As a result, the context may still miss an intermediate relation, include evidence at an unsuitable level of abstraction, or retain irrelevant information. Addressing these limitations requires a context policy that jointly controls gap-aware refinement, retrieval branches, and task-conditioned adaptation. Yet such a context policy for an existing hierarchical backbone remains underexplored.

To address this problem, we propose **A**gentic **C**ontext **E**ngineering for Hierarchical GraphRAG (**ACE-GraphRAG**), an inference-time context policy layer that governs how the initial context is supplemented and adapted for generation. ACE-GraphRAG formulates context construction as a policy over gap-aware refinement, retrieval branches, and task-conditioned adaptation. Guided by the selected policy, *Parallel Differential Retrieval* (PDR) acquires complementary depth-oriented factual evidence and breadth-oriented semantic evidence. The resulting evidence increments are consolidated with the initial context while preserving provenance and abstraction levels, after which task-conditioned adaptation produces a generation-ready context. We instantiate the context policy in two settings: FULL-ACE applies the full policy uniformly within each task family, whereas ADAPTIVE-ACE selects task- and topology-specific policies for individual queries.

We evaluate ACE-GraphRAG on multi-hop question answering using HotpotQA and 2WikiMultiHopQA, and on query-focused summarization using four UltraDomain subsets. Under a unified experimental setup, FULL-ACE outperforms the evaluated RAG and GraphRAG baselines across both task families, while ADAPTIVE-ACE further improves multi-hop QA by 4.00 EM points on both datasets and by 3.77/4.42 F1 points on HotpotQA/2WikiMultiHopQA, respectively, and is preferred over FULL-ACE on all four UltraDomain subsets. Ablation and topology analyses further support topology-conditioned policy selection.

Our contributions are threefold:

- We identify the *representation–inference gap* in hierarchical GraphRAG and formulate inference-time context construction as a query- and task-dependent context policy over gap-aware refinement, retrieval branches, and task-conditioned adaptation.
- We propose ACE-GraphRAG, an inference-time context policy layer that realizes this formulation through *Parallel Differential Retrieval*, context consolidation that preserves provenance and abstraction levels, and two context policy settings: FULL-ACE and ADAPTIVE-ACE.
- We evaluate ACE-GraphRAG on multi-hop question answering and query-focused summarization, showing that FULL-ACE outperforms the evaluated baselines, ADAPTIVE-ACE yields further gains, and ablation and topology analyses support topology-conditioned policy selection.

2 Related Work

2.1 Graph-Based and Hierarchical RAG

Graph-based and hierarchical RAG methods organize external knowledge into graph or multi-resolution structures to support relational and multi-granular retrieval. One line of work focuses on constructing such representations: Microsoft GraphRAG builds entity graphs and hierarchical community reports for query-focused summarization (Edge et al., 2024), RAPTOR recursively organizes clustered passages and abstractive summaries across abstraction levels (Sarthi et al., 2024), and HiRAG explicitly integrates hierarchical knowledge into indexing and retrieval (Huang et al., 2025). Another line improves retrieval over these structures: HippoRAG applies Personalized PageRank over a knowledge graph for multi-hop integration (Gutiérrez et al., 2024), while TagRAG uses hierarchical tag chains (Tao et al., 2026), TopoRAG introduces topology-aware search constraints (Wu and Luo, 2026), and CatRAG adapts graph traversal according to query-dependent edge relevance (Lau et al., 2026). ACE-GraphRAG instead focuses

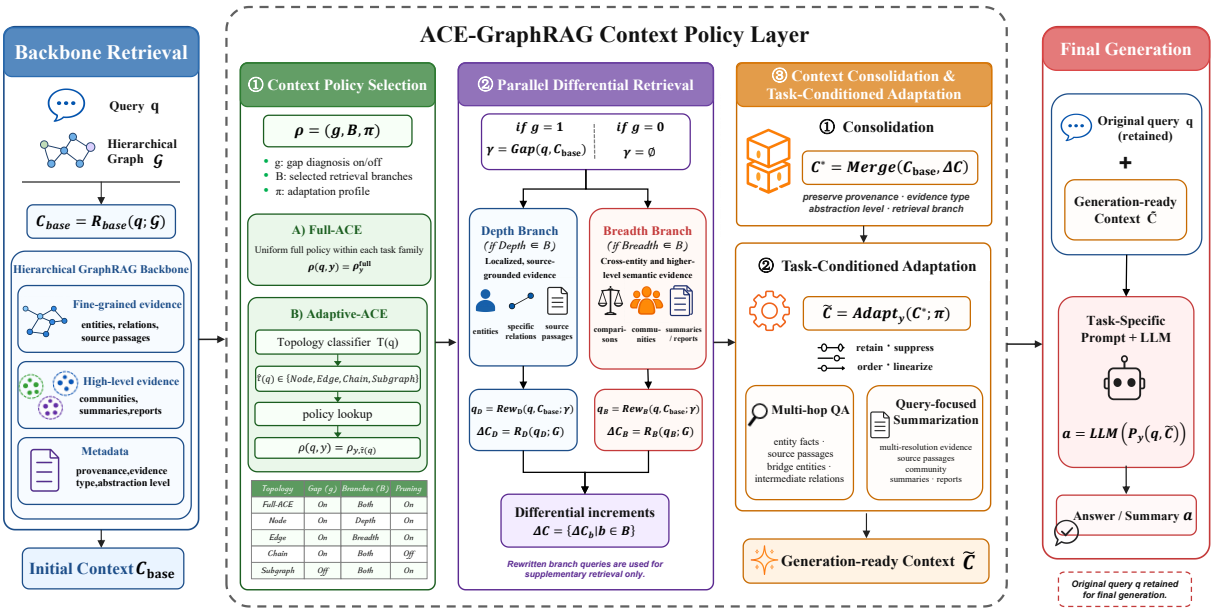

Figure 1: Overview of ACE-GraphRAG. Given a query q and task family y, the hierarchical GraphRAG backbone first retrieves an initial multi-resolution context C_{base}. ACE-GraphRAG then selects a context policy, acquires supplementary evidence from one or both policy-selected retrieval branches through Parallel Differential Retrieval, and consolidates the resulting evidence increments with the initial context while preserving provenance and abstraction levels. Task-conditioned adaptation subsequently produces the generation-ready context $\tilde{C}$.

on inference-time context construction over an existing hierarchical backbone, controlling how the initial context is supplemented, consolidated, and adapted for the current query and task.

2.2 Adaptive and Iterative Retrieval

Adaptive RAG adjusts retrieval behavior to query-specific needs. Adaptive-RAG selects among no retrieval, single-step retrieval, and iterative retrieval based on question complexity (Jeong et al., 2024). KiRAG identifies and retrieves missing knowledge during multi-hop reasoning (Fang et al., 2025), while E^2GraphRAG switches between local and global graph retrieval (Zhao et al., 2025). EA-GraphRAG routes queries between dense RAG and GraphRAG according to structural complexity (Dong et al., 2026), and HiGraAgent coordinates graph retrieval and reasoning through a dual-agent framework (Luu et al., 2026). ACE-GraphRAG extends this adaptive view beyond retrieval selection to inference-time context construction, jointly controlling gap-aware refinement, retrieval branches, and task-conditioned adaptation.

3 Preliminaries

3.1 Hierarchical GraphRAG Backbone

Given a corpus $\mathcal{D}$, a hierarchical GraphRAG backbone organizes corpus knowledge into a multi-resolution structure

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{L}), \tag{1}$$

where $\mathcal{V}$ and $\mathcal{E}$ denote the graph nodes and edges, respectively, and $\mathcal{L} = \{\mathcal{L}_0, \ldots, \mathcal{L}_K\}$ denotes the abstraction levels. Lower levels contain fine-grained evidence such as entities, relations, and source passages, whereas higher levels contain more abstract semantic evidence such as communities, summaries, or reports.

Given a query q, let $\mathcal{R}_{\text{base}}$ denote the initial retrieval operator of the hierarchical GraphRAG backbone. It retrieves evidence from $\mathcal{G}$ to form the initial context

$$C_{\text{base}} = \mathcal{R}_{\text{base}}(q; \mathcal{G}), \tag{2}$$

where C_{base} consists of heterogeneous evidence items associated with their provenance, evidence type, and abstraction level. Although C_{base} makes multi-resolution evidence available, it may not yet match the evidence requirements of the current query and is therefore treated as the starting point for inference-time context construction.

3.2 Task Formulation

Let $\mathcal{Y}$ denote the set of downstream task families, and let $y \in \mathcal{Y}$ denote a downstream task family. Given q and the initial context C_{base}, with access

to the hierarchical knowledge structure $\mathcal{G}$ for supplementary retrieval, ACE-GraphRAG aims to construct a generation-ready context $\tilde{C}$ by supplementing and adapting C_{base} according to the query and downstream task. This context construction is performed at inference time without modifying the backbone index or its initial retrieval procedure.

The final output is generated as

$$a = \text{LLM}\left(P_y(q, \tilde{C})\right), \tag{3}$$

where P_y denotes the task-specific generation prompt for task family y, LLM denotes the language model used for generation, and a denotes the resulting answer or summary.

4 Methodology

4.1 Overview

ACE-GraphRAG introduces an inference-time context policy layer between the initial retrieval of a hierarchical GraphRAG backbone and final generation. Given a query q, task family y, and initial context C_{base}, the overall pipeline consists of three stages:

1. **Context policy selection.** ACE-GraphRAG selects a context policy that controls gap-aware refinement, retrieval branches, and task-conditioned adaptation for the current query and task.

2. **Parallel Differential Retrieval.** Guided by the selected policy, ACE-GraphRAG constructs branch-specific supplementary queries and retrieves supplementary evidence from one or both policy-selected retrieval branches. When enabled, gap diagnosis explicitly targets evidence requirements unresolved by the initial context.

3. **Context consolidation and task-conditioned adaptation.** The supplementary evidence is consolidated with C_{base} while preserving provenance and abstraction levels, and is then adapted into a generation-ready context $\tilde{C}$.

FULL-ACE applies the full context policy uniformly within each task family, whereas ADAPTIVE-ACE selects a task- and topology-specific policy for each query. Figure 1 illustrates the overall inference pipeline.

4.2 Context Policy Selection

ACE-GraphRAG uses a context policy ρ to govern how the initial context is supplemented and adapted during inference:

$$\rho = (g, \mathcal{B}, \pi), \tag{4}$$

where $g \in \{0, 1\}$ indicates whether gap-aware refinement through evidence gap diagnosis is enabled, and $\varnothing \neq \mathcal{B} \subseteq \{\text{DEPTH}, \text{BREADTH}\}$ specifies the retrieval branches to execute. The DEPTH branch targets localized, source-grounded factual evidence, including entities, specific relations, and source passages. The BREADTH branch targets broader cross-entity and higher-level semantic evidence, including comparative relations, communities, summaries, reports, and semantic dimensions not sufficiently covered by the initial context. Finally, π denotes the adaptation profile governing which evidence categories are retained or suppressed and how the retained evidence is ordered and linearized before generation.

FULL-ACE applies the same full context policy to all queries within a task family. It enables gap-aware refinement, executes both the DEPTH and BREADTH branches, and applies the corresponding task-conditioned adaptation profile.

A fixed context policy may not be equally suitable for queries with different structural evidence requirements. ADAPTIVE-ACE therefore uses a lightweight LLM-based classifier T that predicts the topology of the original query:

$$\begin{aligned} &\hat{\tau}(q) = T(q), \\ &\hat{\tau}(q) \in \{\text{NODE}, \text{EDGE}, \\ &\qquad \text{CHAIN}, \text{SUBGRAPH}\}. \end{aligned} \tag{5}$$

NODE queries primarily require localized evidence about a single entity. EDGE queries require broader cross-entity or comparative evidence. CHAIN queries require intermediate evidence that preserves multi-step dependencies. SUBGRAPH queries require broader evidence covering multiple related entities, communities, or semantic regions.

Table 1 summarizes the concrete policy configurations used by the two inference modes. For ADAPTIVE-ACE, each predicted topology is mapped to a predefined configuration of evidence gap diagnosis, retrieval branches, and semantic pruning, fixed before test-set evaluation.

Under the two inference modes, the context pol-

Policy	Gap Diagn.	Branches	Sem. Pruning
FULL-ACE	On	Both	On
NODE	On	DEPTH	On
EDGE	On	BREADTH	On
CHAIN	On	Both	Off
SUBGRAPH	Off	Both	On

Table 1: Context-policy configurations. "Gap Diagn." and "Sem. Pruning" denote gap diagnosis and semantic pruning, respectively. The rows below the divider are the topology-conditioned policies used by ADAPTIVE-ACE.

icy for query q is selected as

$$\rho(q, y) = \begin{cases} \rho_y^{\text{full}}, & \text{FULL-ACE}, \\ \rho_{y,\hat{\tau}(q)}, & \text{ADAPTIVE-ACE} \end{cases} \tag{6}$$

where ρ_y^{full} denotes the full policy for task family y, and $\rho_{y,\hat{\tau}(q)}$ denotes the task- and topology-specific policy assigned to the predicted topology.

The policy assignments reflect different evidence requirements. NODE queries use the DEPTH branch to acquire localized factual evidence. EDGE queries use the BREADTH branch to obtain broader cross-entity and comparative coverage. CHAIN queries retain evidence from both branches and disable semantic pruning to preserve intermediate dependencies. SUBGRAPH queries use both branches and skip evidence gap diagnosis to support broad multi-resolution coverage.

4.3 Parallel Differential Retrieval

Given the selected policy $\rho = (g, \mathcal{B}, \pi)$, Parallel Differential Retrieval acquires supplementary evidence for the evidence requirements of q that remain unresolved by C_{base}.

When gap-aware refinement is enabled, ACE-GraphRAG produces an evidence gap description:

$$\gamma = \begin{cases} \text{Gap}(q, C_{\text{base}}), & g = 1, \\ \varnothing, & g = 0, \end{cases} \tag{7}$$

where Gap denotes an LLM-based evidence gap diagnosis operator that jointly examines q and C_{base} to identify factual details, entity relations, or semantic dimensions that are absent or insufficiently supported, together with the entities or relations that should guide supplementary retrieval. When $g = 0$, branch-specific query rewriting directly uses q and C_{base}.

For each selected branch $b \in \mathcal{B}$, PDR constructs a branch-specific follow-up query and retrieves the corresponding evidence increment:

$$q_b = \text{Rewrite}_b(q, C_{\text{base}}, \gamma), \Delta C_b = \mathcal{R}_b(q_b; \mathcal{G}), \tag{8}$$

where Rewrite_b denotes an LLM-based branch-specific query rewriting operator, $\mathcal{R}_b$ denotes the corresponding retrieval operator, and ΔC_b denotes the retrieved evidence increment.

The DEPTH query targets missing localized facts and specific relations, while the BREADTH query targets broader cross-entity comparisons, semantic dimensions, and perspectives not sufficiently covered by the initial context. When both branches are selected, their follow-up queries are generated in the same planning step, and the corresponding retrieval operators are executed in parallel to produce complementary evidence increments.

The resulting evidence increments are collected as

$$\Delta C = \{\Delta C_b \mid b \in \mathcal{B}\}. \tag{9}$$

These increments supplement C_{base} and are passed to the context consolidation and task-conditioned adaptation stage.

4.4 Context Consolidation and Task-Conditioned Adaptation

ACE-GraphRAG consolidates the supplementary evidence with the initial context:

$$C^* = \text{Merge}\left(C_{\text{base}}, \Delta C\right). \tag{10}$$

Here, Merge organizes the initial and supplementary evidence into structured blocks while preserving provenance, evidence type, abstraction level, and retrieval branch. The resulting C^* forms a consolidated multi-resolution context for subsequent adaptation.

ACE-GraphRAG then applies task-conditioned adaptation according to task family y and adaptation profile π:

$$\tilde{C} = \text{Adapt}_y\left(C^*; \pi\right). \tag{11}$$

Here, Adapt_y denotes task-conditioned context adaptation governed by the adaptation profile π. It determines which evidence categories are retained or suppressed and how the retained evidence is ordered and linearized into the generation-ready context $\tilde{C}$.

For multi-hop question answering, task-conditioned adaptation prioritizes entity-level facts, relational evidence, and source passages. It preserves bridge entities and intermediate

Method	Multi-Hop QA				UltraDomain Pairwise Preference (%)			
	HotpotQA		2Wiki		Mix	CS	Legal	Agr.
	EM	F1	EM	F1				
NaiveRAG	23.00	38.47	24.00	42.04	89.2	76.4	62.3	81.5
GraphRAG	14.00	25.06	28.00	40.15	93.1	81.0	77.0	84.5
LightRAG	25.00	38.42	28.00	46.08	88.8	79.0	85.0	83.0
FastGraphRAG	27.00	43.22	12.00	12.38	98.5	85.0	81.0	84.0
HippoRAG	26.00	39.60	28.00	44.27	–	–	–	–
KAG	–	–	–	–	96.5	85.0	78.5	84.5
HiRAG	24.00	40.02	24.00	40.47	86.2	81.9	84.4	80.5
FULL-ACE	32.00	44.78	52.00	59.85	50.0	50.0	50.0	50.0
ADAPTIVE-ACE†	**36.00**	**48.55**	**56.00**	**64.27**	51.9	61.5	64.0	64.0

Table 2: Main results on multi-hop QA and UltraDomain. For each baseline row, the UltraDomain columns report the pairwise preference rate of FULL-ACE over the listed method. The † row reports the preference rate of ADAPTIVE-ACE over FULL-ACE. Values above 50 favor the method named first in each comparison, and – denotes unavailable comparisons.

relations required for multi-step reasoning, while suppressing broad report-level evidence when specified by π. For query-focused summarization, the adaptation retains broader multi-resolution evidence, including source passages, community-level summaries, and report-level information, and organizes it to support comprehensive and coherent synthesis.

The rewritten branch queries are used only for supplementary retrieval. Final generation conditions on the original query q and the generation-ready context $\tilde{C}$ through the task-specific generation prompt $P_y(q, \tilde{C})$.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets and Evaluation

We evaluate ACE-GraphRAG on two task families. For multi-hop QA, we sample 100 queries from each of **HotpotQA** (Yang et al., 2018) and **2WikiMultiHopQA** (Ho et al., 2020). HotpotQA requires reasoning over evidence distributed across multiple supporting passages, while 2WikiMultiHopQA emphasizes multi-step relational reasoning across linked entities. For query-focused summarization, we use four subsets of the **UltraDomain** benchmark (Qian et al., 2025): Mix, CS, Legal, and Agriculture. These subsets cover both cross-domain and domain-specific corpora and require systems to synthesize retrieved evidence into focused summaries.

For multi-hop QA, we report Exact Match (EM) and token-level F1 using a shared answer extraction and normalization pipeline. For UltraDomain, following prior long-context RAG studies (Guo et al., 2025), we use an LLM-based pairwise judge to evaluate *Comprehensiveness*, *Empowerment*, *Diversity*, and pairwise preference on the *Overall* criterion. All methods use the same judge, scoring prompt, and pairwise comparison protocol.

5.1.2 Baselines

We compare ACE-GraphRAG with representative conventional, graph-enhanced, and hierarchical RAG methods. NaiveRAG (Gao et al., 2023) serves as a standard passage-retrieval baseline, while LightRAG (Guo et al., 2025) and FastGraphRAG (CircleMind AI, 2024) provide lightweight graph-enhanced alternatives. We further include Microsoft GraphRAG (Edge et al., 2024), KAG (Liang et al., 2024), and HiRAG (Huang et al., 2025) as graph-structured and hierarchical baselines, with HiRAG serving as the primary hierarchical backbone baseline. For multi-hop QA, we additionally evaluate HippoRAG (Gutiérrez et al., 2024), which is designed for graph-based multi-hop evidence integration.

5.1.3 Implementation Details

For UltraDomain, we use DeepSeek-V3 (DeepSeek-AI et al., 2024) for planning and generation, GLM-4-Plus embeddings (Team GLM et al., 2024), and gpt-5.4-nano (OpenAI, 2026) as the evaluation judge. For multi-hop QA, we use GPT-4o-mini (OpenAI, 2024) and nvidia/NV-Embed-v2 (Lee et al., 2024) under HippoRAG-style settings (Gutiérrez et al., 2024). Methods within each task family otherwise share the same generation, decoding, and evaluation

settings.

5.2 Main Results

Table 2 summarizes the overall results on multi-hop QA and query-focused summarization. FULL-ACE achieves 32.00 EM and 44.78 F1 on HotpotQA and 52.00 EM and 59.85 F1 on 2WikiMultiHopQA, outperforming all evaluated baselines on both datasets. It is also preferred over every available baseline across all four UltraDomain subsets. These results show that supplementing and adapting the initial context after backbone retrieval can better translate hierarchical representations into generation-ready contexts across both task families.

ADAPTIVE-ACE further improves performance to 36.00 EM and 48.55 F1 on HotpotQA and 56.00 EM and 64.27 F1 on 2WikiMultiHopQA. Compared with FULL-ACE, it gains 4.00 EM points on both datasets and 3.77/4.42 F1 points on HotpotQA/2WikiMultiHopQA, respectively. The pairwise judge also prefers ADAPTIVE-ACE over FULL-ACE on all four UltraDomain subsets, with larger margins on CS, Agriculture, and Legal. Overall, topology-conditioned policy selection better matches context construction to query-specific evidence requirements than applying the same full policy uniformly to every query.

5.3 Policy and Topology Analysis

5.3.1 Component and Policy Analysis

We examine whether the gains of ACE-GraphRAG arise from individual context-policy operations and whether a single fixed policy is sufficient for all queries. Table 3 reports component ablations in the upper block and policy-selection controls in the lower block. The ablations remove the BREADTH branch, the DEPTH branch, gap diagnosis (*w/o Gap Diagnosis*), or semantic pruning (*w/o Semantic Pruning*) from FULL-ACE. Oracle Best Fixed reports the fixed configuration with the highest test-set F1 on each dataset. Shuffled Router randomly permutes the query-to-policy assignments 1,000 times while preserving the number of queries assigned to each policy, and reports the mean and standard deviation. ADAPTIVE-ACE instead applies the topology-conditioned policy selected for each query.

The component ablations reveal clear cross-dataset trade-offs. Removing the BREADTH branch improves 2WikiMultiHopQA but degrades HotpotQA, whereas removing the DEPTH branch produces the opposite pattern. Gap-aware refinement and semantic pruning also have dataset-dependent effects. These results indicate that no fixed context-policy configuration is uniformly optimal across datasets. ADAPTIVE-ACE nevertheless outperforms both Oracle Best Fixed and Shuffled Router on both datasets, suggesting that its gains come from matching context policies to query topology rather than merely varying policies across queries.

	HotpotQA		2Wiki	
Configuration	**EM**	**F1**	**EM**	**F1**
FULL-ACE	32.00	44.78	52.00	59.85
w/o Breadth	29.00	43.39	53.00	62.55
w/o Depth	34.00	46.96	47.00	53.57
w/o Gap Diagnosis	32.00	47.19	46.00	53.30
w/o Semantic Pruning	28.00	42.48	54.00	61.48
Oracle Best Fixed	32.00	47.19	53.00	62.55
Shuffled Router	29.75±2.07	43.90±1.87	50.13±2.02	57.87±1.95
ADAPTIVE-ACE	**36.00**	**48.55**	**56.00**	**64.27**

Table 3: Multi-hop QA component ablations and policy-selection controls. The upper block removes individual operations from FULL-ACE, while the lower block compares oracle-fixed, shuffled, and topology-conditioned policy selection.

Table 4 evaluates the corresponding component ablations on query-focused summarization. Each entry reports the pairwise preference rate of FULL-ACE over the corresponding variant on the *Overall* criterion, with values above 50 favoring FULL-ACE. The variants remove the BREADTH branch, the DEPTH branch, gap diagnosis (*w/o Gap Diagnosis*), or replace the task-specific generation prompt P_y with a generic prompt (*w/o Task-Specific Prompt*), respectively.

The three retrieval-related ablations show that the contribution of each operation varies across domains, while FULL-ACE remains preferred on average over all three variants. Replacing the task-specific prompt with a generic prompt causes a substantially larger and more consistent degradation: FULL-ACE is preferred by 76.5%–85.0% across the four domains, with an average preference rate of 80.0%. These results suggest that retrieval-side operations provide complementary, domain-dependent benefits, whereas task-specific prompting consistently helps the generator use multi-resolution evidence to produce focused and coherent summaries.

5.3.2 Query Topology and Policy Performance

Figure 2(a) shows that query topology distributions differ substantially across datasets. HotpotQA and Legal are dominated by NODE queries, whereas 2WikiMultiHopQA contains a larger proportion

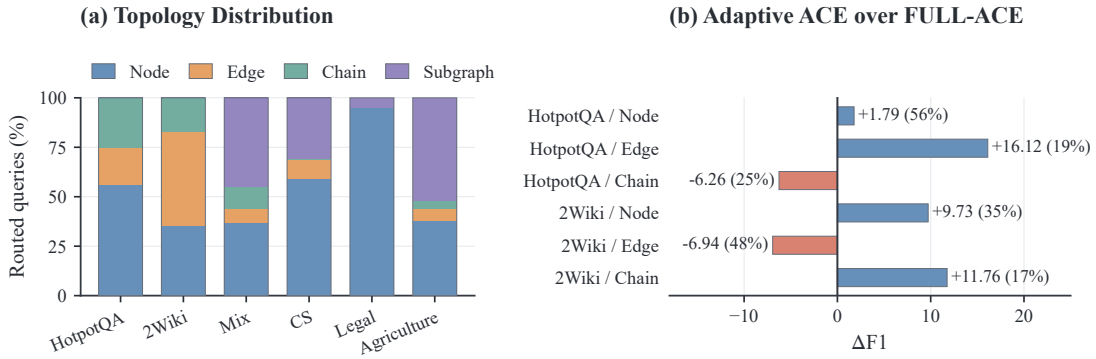

Figure 2: Query topology distributions and topology-conditioned performance. (a) Distributions of predicted query topologies across the multi-hop QA and UltraDomain datasets. (b) F1 changes of ADAPTIVE-ACE relative to FULL-ACE for different query topologies on multi-hop QA.

Variant	Mix	CS	Legal	Agr.	Avg.
w/o Breadth	50.8	54.0	54.5	61.0	55.1
w/o Depth	54.2	60.5	56.5	59.5	57.7
w/o Gap Diagnosis	46.5	60.5	53.8	58.0	54.7
w/o Task-Specific Prompt	85.0	82.0	76.5	76.5	**80.0**

Table 4: UltraDomain component ablations. Entries are pairwise preference rates (%) of FULL-ACE over each variant on the *Overall* criterion.

Context	HotpotQA (%)		2Wiki (%)	
	Cov.	Rec.	Cov.	Rec.
Initial	32.3	–	67.4	–
FULL-ACE	**36.5**	**6.2**	**92.1**	**75.9**
ADAPTIVE-ACE	35.4	4.6	88.8	65.5

Table 5: Gold-answer coverage and recovery on 100 matched queries per benchmark.

of EDGE queries. Mix and Agriculture, by contrast, contain substantial proportions of SUBGRAPH queries. This variation supports selecting context policies at the query level rather than applying a single fixed policy to an entire dataset.

Figure 2(b) reports the F1 change of ADAPTIVE-ACE relative to FULL-ACE within each QA topology. The largest gains occur on HotpotQA EDGE queries and 2WikiMultiHopQA NODE and CHAIN queries. Overall, the topology-dependent performance differences indicate that query topology provides a useful signal for policy selection, while leaving room for further calibration of the topology-to-policy mapping.

5.4 Effectiveness of Context Construction

To verify whether ACE-GraphRAG constructs a generation-ready context that better satisfies query-specific evidence requirements, we replay the same 100 queries per benchmark across the initial, FULL-ACE, and ADAPTIVE-ACE settings. We measure coverage as the proportion of non-yes/no queries whose normalized gold answer appears in the context, and recovery as the proportion of initially uncovered queries for which the constructed context successfully recovers the answer. As shown in Table 5, both FULL-ACE and ADAPTIVE-ACE improve answer coverage and recover missing answer-bearing evidence on both benchmarks, with particularly strong gains on 2WikiMultiHopQA. These results validate that policy-guided supplementary retrieval and subsequent context adaptation effectively transform hierarchical representations into contexts better suited to the current query.

5.5 Transferability and Efficiency

Table 6 evaluates backbone transfer and inference efficiency under matched settings. Panel (a) isolates the effect of adding FULL-ACE to the same GraphRAG Local+Global context, while Panel (b) reports per-query costs on HiRAG.

FULL-ACE transfers effectively to GraphRAG, improving average EM and F1 by 14.50 and 12.69 points, with particularly large gains on 2WikiMultiHopQA and an F1 improvement on HotpotQA. This indicates that the context policy layer is not tied to HiRAG.

Compared with FULL-ACE, ADAPTIVE-ACE reduces context-retrieval calls from 3.00 to 2.19 and the input-token ratio from 2.68× to 2.42×, with similar latency because of the additional routing call. Fixed Multi-Resolution provides a lower-cost reference without gap-aware refinement. Overall, topology-conditioned policy selection reduces token overhead while maintaining comparable in-

(a) Backbone Transfer

Method	HotpotQA		2Wiki		Average	
	EM	**F1**	**EM**	**F1**	**EM**	**F1**
GraphRAG (Local+Global)	**26.00**	38.02	27.00	42.99	26.50	40.51
+ FULL-ACE	25.00	**41.09**	**57.00**	**65.29**	**41.00**	**53.19**
Δ	-1.00	$+3.07$	$+30.00$	$+22.30$	$+14.50$	$+12.69$

(b) Inference Efficiency on HiRAG

Method	Calls		Inference Cost			
	Ctx.	**LLM**	**Input Tok.**	**Output Tok.**	**Time**	**Input Ratio**
HiRAG	1.00	1.00	23,081	202	1.37	1.00×
Fixed Multi-Resolution	3.00	1.00	39,170	94	1.86	1.69×
FULL-ACE	3.00	2.00	62,156	271	2.48	2.68×
ADAPTIVE-ACE	2.19	3.00	56,052	285	2.49	2.42×

Table 6: Backbone transfer and inference efficiency on multi-hop QA. Panel (a) reports the transfer of FULL-ACE to GraphRAG. Panel (b) reports per-query costs on HiRAG, averaged across HotpotQA and 2WikiMultiHopQA. Ctx. and LLM denote context-retrieval and LLM calls, respectively; time is measured in seconds, and input-token ratios are normalized to HiRAG.

ference time.

6 Conclusion

We introduced ACE-GraphRAG, an inference-time context policy layer that refines hierarchical GraphRAG contexts through gap-aware refinement, Parallel Differential Retrieval, and task-conditioned adaptation. FULL-ACE applies a uniform policy, whereas ADAPTIVE-ACE selects task- and topology-specific policies. Results on multi-hop QA and query-focused summarization, supported by ablation, topology, coverage, transfer, and efficiency analyses, show that ACE-GraphRAG better matches hierarchical evidence to each query and task, thereby narrowing the representation–inference gap.

References

CircleMind AI. 2024. FastGraphRAG: Streamlined and promptable graph-based retrieval-augmented generation. GitHub. Accessed: 2026-07-29.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 179 others. 2024. DeepSeek-V3 Technical Report. *arXiv preprint*.

Su Dong, Qinggang Zhang, Yilin Xiao, Shengyuan Chen, Chuang Zhou, and Xiao Huang. 2026. Use graph when it needs: Efficiently and adaptively integrating retrieval-augmented generation with graphs. *Preprint*, arXiv:2602.03578.

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*.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. 2025. KiRAG: Knowledge-driven iterative retriever for enhancing retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18969–18985, Vienna, Austria. Association for Computational Linguistics.

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

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 Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In *Advances in Neural Information Processing Systems 37*, pages 59532–59569, Vancouver, BC, Canada. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Haoyu Huang, Yongfeng Huang, Yang Junjie, Zhenyu Pan, Yongqiang Chen, Kaili Ma, Hongzhi Chen, and James Cheng. 2025. Retrieval-Augmented Generation with Hierarchical Knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6044–6060, Suzhou, China. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

Kwun Hang Lau, Fangyuan Zhang, Boyu Ruan, Yingli Zhou, Qintian Guo, Ruiyuan Zhang, and Xiaofang Zhou. 2026. Breaking the static graph: Context-aware traversal for graph-based RAG. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 5849–5863, San Diego, California, United States. Association for Computational Linguistics.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. *arXiv preprint*.

Lei Liang, Mengshu Sun, Zhengke Gui, Zhongshu Zhu, Zhouyu Jiang, Ling Zhong, Yuan Qu, Peilong Zhao, Zhongpu Bo, Jin Yang, Huaidong Xiong, Lin Yuan, Jun Xu, Zaoyang Wang, Zhiqiang Zhang, Wen Zhang, Huajun Chen, Wenguang Chen, and Jun Zhou. 2024. Kag: Boosting llms in professional domains via knowledge augmented generation. *Preprint*, arXiv:2409.13731.

Hung Luu, Long S. T. Nguyen, Trung Pham, Hieu Pham, and Tho Quan. 2026. HiGraAgent: Dual-agent adaptive reasoning over hierarchical knowledge graph for open domain multi-hop question answering. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1193–1217, Rabat, Morocco. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. OpenAI. Accessed: 2026-07-29.

OpenAI. 2026. GPT-5.4 nano model. OpenAI API Documentation. Accessed: 2026-07-29.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. MemoRAG: Boosting Long Context Processing with Global Memory-Enhanced Retrieval Augmentation. In *Proceedings of the ACM on Web Conference 2025*, pages 2366–2377, Sydney NSW Australia. ACM.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. *arXiv preprint*.

Wenbiao Tao, Xinyuan Li, Yunshi Lan, and Weining Qian. 2026. TagRAG: Tag-guided hierarchical knowledge graph retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 6434–6456, San Diego, California, United States. Association for Computational Linguistics.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, and 39 others. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint*.

Tianhao Wu and Siqiang Luo. 2026. TopoRAG: Graph-based RAG via topology-aware approximate nearest neighbor search. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 34097–34108, San Diego, California, United States. Association for Computational Linguistics.

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

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. 2025. Agentic context engineering: Evolving contexts for self-improving language models. *Preprint*, arXiv:2510.04618.

Yibo Zhao, Jiapeng Zhu, Ye Guo, Kangkang He, and Xiang Li. 2025. E^2GraphRAG: Streamlining graph-based RAG for high efficiency and effectiveness. *Preprint*, arXiv:2505.24226.

A Additional Experimental Results

A.1 Retrieved-Context Evidence Quality Protocol

We evaluate the Initial Context, the evolved context produced by FULL-ACE, and the evolved context produced by ADAPTIVE-ACE using matched replays of the same 100 queries from each benchmark. Answer coverage is defined as the proportion of non-yes/no questions for which the normalized gold answer appears in the retained context; four yes/no questions from HotpotQA and eleven from 2WikiMultiHopQA are therefore excluded. Missing-answer recovery is evaluated only on questions whose gold answer is absent from the Initial Context, and measures the proportion for which the evolved context subsequently recovers the gold answer. Because the replay records do not contain supporting-fact annotations, we do not report supporting-fact recall. The resulting coverage and recovery scores are reported in the main paper.

A.2 UltraDomain Evaluation Dimensions

Table 7 reports the pairwise preference rates of ADAPTIVE-ACE over FULL-ACE across the three UltraDomain evaluation dimensions and the Overall criterion. Values above 50% favor ADAPTIVE-ACE, whereas values below 50% favor FULL-ACE. Overall represents the judge's direct overall preference and is not the arithmetic mean of Comprehensiveness, Empowerment, and Diversity.

Dataset	Comp.	Emp.	Div.	Overall
Mix	49.6	58.1	51.9	51.9
CS	62.0	59.0	53.0	61.5
Legal	65.5	62.5	46.0	64.0
Agriculture	65.5	54.0	53.5	64.0

Table 7: UltraDomain pairwise preference rates (%) of ADAPTIVE-ACE over FULL-ACE across the three evaluation dimensions and the Overall criterion.

B Extended Qualitative Analysis

B.1 Open-Ended Domain Synthesis

We examine the UltraDomain-Mix query *How does the adsorption site affect the state of indeno[1,2-a]fluorene (5)?* The retrieved context contains several potentially relevant factors, including the substrate, charge state, surface defects, and adsorption geometry. To illustrate the role of task-conditioned synthesis, Table 8 presents verbatim response excerpts from FULL-ACE and its variant without the task-specific prompt. The comparison shows how the two responses organize and relate the available factors when explaining the state transition.

The response generated without the task-specific prompt mentions several relevant factors, but their respective roles in the state transition remain only loosely connected. In contrast, the FULL-ACE response more explicitly distinguishes charge-state cycling from the site-dependent mechanism: it states that charge cycling alone is insufficient and relates the transition to adsorption-site geometry and local electrostatic interactions. This example illustrates how task-conditioned synthesis can organize heterogeneous retrieved evidence into a more explicit mechanistic account.

B.2 Recovering Missing Second-Hop Evidence

We examine the 2WikiMultiHopQA query *Who is the spouse of the director of the film My Three Merry Widows?* The initial retrieval identifies Fernando Cortés as the director but does not resolve the spouse relation required by the second hop. Table 9 contrasts the resulting evidence trajectories of the configuration without gap diagnosis and ADAPTIVE-ACE, while Table 10 provides their complete model outputs.

The initial retrieval resolves the first hop by identifying Fernando Cortés as the director, but leaves the spouse relation unsupported. In the configuration without gap diagnosis, no focused retrieval target is formed and the final output remains *Unknown*. In contrast, ADAPTIVE-ACE explicitly represents the unresolved spouse relation, retrieves evidence identifying María del Pilar Cordero as Mapy Cortés, and produces the correct answer. In this example, explicitly representing the unresolved relation guides supplementary retrieval toward the missing second-hop evidence.

C Repeated-Run Robustness

To complement the 100-query evaluation reported in the main paper, we conduct three independent evaluation runs on HotpotQA and 2WikiMultiHopQA, with each run covering 300 questions from each benchmark. Within each run, all methods are evaluated on the same set of questions. Table 11 reports the results of the uniform FULL-ACE configuration and the compared methods. Each run entry gives EM/F1, and the final two columns report the mean and sample standard deviation across

Aspect	w/o Task-Specific Prompt	FULL-ACE
Surface and charge framing	The underlying metal surface also affects the behavior of molecule 5 through its work function; molecule 5 exhibits charge bistability between neutral and anionic states.	However, this charge-state cycling alone does not induce the open/closed-shell switch; it only occurs if the molecule also changes its adsorption site.
Site-specific mechanism	Rather, a change in adsorption site induces sufficient geometric modification of the molecule to cause a corresponding change in its ground-state electronic configuration.	The local electrostatic environment and potential defects at the adsorption site induce sufficient geometric distortion in the physisorbed molecule to alter its fundamental electronic ground state.

Table 8: UltraDomain-Mix synthesis case comparing FULL-ACE with its variant without the task-specific prompt. Entries are verbatim excerpts from the generated responses, with surrounding text omitted.

Stage	w/o Gap Diagnosis	ADAPTIVE-ACE
Initial retrieval	Identifies *My Three Merry Widows* → director → Fernando Cortés.	Identifies the same bridge entity, Fernando Cortés.
Retrieval target	No focused target is formed for the unresolved spouse relation.	Makes the missing relation explicit as Fernando Cortés → spouse → ?.
Supplementary evidence	No evidence resolving the spouse relation is recovered.	Retrieves evidence identifying María del Pilar Cordero, whose stage name is Mapy Cortés.
Final answer	Unknown (EM/F1 = 0/0).	Mapy Cortés (EM/F1 = 1/1).

Table 9: Evidence trajectories for a missing-second-hop case. The configuration without gap diagnosis does not form a focused retrieval target for the unresolved spouse relation, whereas ADAPTIVE-ACE retrieves evidence that resolves it.

the three runs. Because these runs use 300-question evaluation sets rather than the 100-query sample reported in the main paper, their absolute scores should be interpreted within the repeated-run setting and are not directly comparable across the two sample sizes.

Across the three runs, FULL-ACE achieves the highest mean EM and F1 among the evaluated methods on both benchmarks. Its results are also consistent across runs, with standard deviations of 2.01 EM and 1.13 F1 on HotpotQA and 0.51 EM and 0.42 F1 on 2WikiMultiHopQA. These repeated-run results indicate that the improvements of FULL-ACE are not confined to a single evaluation run.

D Overall Inference Procedure

Algorithm 1 summarizes the end-to-end inference procedure of ACE-GraphRAG and instantiates the context-policy formulation introduced in the main paper. Given the initial context retrieved by a hierarchical GraphRAG backbone, ACE-GraphRAG selects either a task-level full policy or a task- and topology-conditioned policy, performs Parallel Differential Retrieval through the enabled branches, consolidates the resulting evidence increments with the initial context, and applies task-conditioned adaptation before final generation.

Algorithm 1 Overall ACE-GraphRAG Inference

Require: Query q, task family y, hierarchical knowledge structure $\mathcal{G}$, inference mode $m \in \{\text{FULL-ACE}, \text{ADAPTIVE-ACE}\}$
Ensure: Final output a

$C_{\text{base}} \leftarrow \mathcal{R}_{\text{base}}(q; \mathcal{G})$
if $m = \text{FULL-ACE}$ **then**
  $\rho \leftarrow \rho_y^{\text{full}}$
else
  $\hat{\tau}(q) \leftarrow \mathcal{T}(q)$
  $\rho \leftarrow \rho_{y,\hat{\tau}(q)}$
end if
$(g, \mathcal{B}, \pi) \leftarrow \rho$
if $g = 1$ **then**
  $\gamma \leftarrow \text{Gap}(q, C_{\text{base}})$
else
  $\gamma \leftarrow \varnothing$
end if
for all $b \in \mathcal{B}$ **in parallel do**
  $q_b \leftarrow \text{Rewrite}_b(q, C_{\text{base}}, \gamma)$
  $\Delta C_b \leftarrow \mathcal{R}_b(q_b; \mathcal{G})$
end for
$\Delta C \leftarrow \{\Delta C_b \mid b \in \mathcal{B}\}$
$C^* \leftarrow \text{Merge}(C_{\text{base}}, \Delta C)$
$\tilde{C} \leftarrow \text{Adapt}_y(C^*; \pi)$
$a \leftarrow \text{LLM}\big(P_y(q, \tilde{C})\big)$
return a

Variant	Complete Model Output
w/o Gap Diagnosis	Reasoning: The director of My Three Merry Widows is Fernando Cortés. The data mentions that Edwin L. Marin was married to actress Ann Morriss. While this relationship does not directly relate to Cortés, there are no specific spousal details stated for him in the provided data. Therefore, the answer regarding the spouse of Fernando Cortés is not explicitly stated. Result: [FINAL_ANSWER] Unknown [/FINAL_ANSWER]
ADAPTIVE-ACE	Reasoning: My Three Merry Widows was directed by Fernando Cortés. The sources indicate that he was married to María del Pilar Cordero, who adopted the stage name Mapy Cortés. Therefore, she is the spouse of the director. Result: [FINAL_ANSWER] Mapy Cortés [/FINAL_ANSWER]

Table 10: Verbatim model outputs for the 2WikiMultiHopQA missing-second-hop case.

Dataset	Method	Run 1	Run 2	Run 3	Mean EM	Mean F1
HotpotQA	**FULL-ACE**	**27.67/39.27**	**29.33/40.19**	**25.33/37.94**	**27.44 ± 2.01**	**39.13 ± 1.13**
	HiRAG	21.33/36.35	20.33/36.15	20.67/35.14	20.78 ± 0.51	35.88 ± 0.65
	GraphRAG	17.67/28.42	19.67/31.05	18.00/30.24	18.45 ± 1.07	29.90 ± 1.35
	HippoRAG	20.33/34.47	21.67/33.97	20.33/30.86	20.78 ± 0.77	33.10 ± 1.96
	LightRAG	23.00/38.17	23.67/37.78	22.67/37.45	23.11 ± 0.51	37.80 ± 0.36
	FastGraphRAG	23.33/37.89	23.33/37.63	22.67/37.35	23.11 ± 0.38	37.62 ± 0.27
	NaiveRAG	19.67/33.41	21.00/34.14	20.67/30.83	20.45 ± 0.69	32.79 ± 1.74
2WikiMultiHopQA	**FULL-ACE**	**58.33/68.69**	**57.67/68.86**	**57.33/68.06**	**57.78 ± 0.51**	**68.54 ± 0.42**
	HiRAG	24.33/48.96	25.00/48.45	31.33/51.88	26.89 ± 3.86	49.76 ± 1.85
	GraphRAG	32.33/49.96	31.33/48.91	30.33/50.36	31.33 ± 1.00	49.74 ± 0.75
	HippoRAG	32.00/50.71	29.67/47.89	28.00/46.92	29.89 ± 2.01	48.51 ± 1.97
	LightRAG	32.69/51.90	30.67/50.59	35.00/54.22	32.79 ± 2.17	52.24 ± 1.84
	FastGraphRAG	20.67/31.75	20.00/30.78	21.22/30.98	20.63 ± 0.61	31.17 ± 0.51
	NaiveRAG	28.70/47.06	25.00/43.81	25.67/44.79	26.46 ± 1.97	45.22 ± 1.67

Table 11: Repeated-run results on HotpotQA and 2WikiMultiHopQA. Each run evaluates 300 questions per benchmark, and each run entry reports EM/F1 (%). The final two columns report the mean and sample standard deviation across three independent runs.

E Prompt Templates

This section reports the inference-time prompts used by ACE-GraphRAG and complements the context-policy formulation in Section 4 and the overall inference pipeline in Figure 1. The templates are presented in execution order. Runtime inputs are replaced with descriptive placeholders, while the routing, retrieval-planning, synthesis, and output-format constraints used during inference are retained.

E.1 Topology Router

For ADAPTIVE-ACE, the topology router implements the query-level context-policy selection described in Section 4. It predicts the logical structure of the original query before supplementary retrieval. The prediction is based on the query alone and is mapped to one of the predefined context-policy configurations summarized in Table 1. Table 12 further specifies the correspondence between the implementation labels used by the router and the topology labels used in the paper. FULL-ACE does not invoke the router and instead applies the task-level full policy directly. Router decoding uses temperature zero and requires a valid JSON object. If parsing fails, generation is retried up to three times; if all attempts fail, `OVERVIEW` is used as the fallback label.

Code label	Topology	Context-policy configuration
`POINT`	Node	Gap diagnosis on; Depth branch only; semantic pruning on.
`CONTRAST`	Edge	Gap diagnosis on; Breadth branch only; semantic pruning on.
`CHAIN`	Chain	Gap diagnosis on; Depth and Breadth branches; semantic pruning off.
`OVERVIEW`	Subgraph	Gap diagnosis off; Depth and Breadth branches; semantic pruning on.

Table 12: Mapping from router implementation labels to the topology labels and context-policy configurations used in the paper.

E.2 Prompt Variables and Output Contracts

Table 13 defines the placeholders used throughout the templates in this section. Together, these interfaces separate the router's context-policy decision,

Placeholder	Role in the prompt
<QUERY>	Original user query retained for final generation.
<ROUND1_CONTEXT>	Initial context returned by the hierarchical GraphRAG backbone and provided to the retrieval planner.
<EVOLVED_CONTEXT>	Generation-ready context produced after evidence consolidation and task-conditioned adaptation.
<RESPONSE_TYPE>	Requested response style or level of detail for open-ended synthesis.
<BRIEF_LOGIC_CHAIN>	Output slot for a concise bridge-entity reasoning trace.
<EXACT_ANSWER>	Output slot for the final entity or short answer span.

Table 13: Placeholders used in the ACE-GraphRAG inference prompts.

the planner's supplementary retrieval actions, and the final generator's use of the evolved context.

E.3 Gap-Aware Differential Retrieval Planner

After initial retrieval, the planner implements the gap-aware refinement and Parallel Differential Retrieval stages described in Section 4. It jointly examines the original query and the initial context and generates two complementary candidate queries: a Depth query targeting unresolved localized evidence and a Breadth query targeting broader perspectives or semantic dimensions not sufficiently represented in the initial context. The selected context policy determines which candidate queries are submitted to their corresponding retrieval branches. When both branches are enabled, their retrieval operations are executed in parallel. The rewritten queries are used only for supplementary retrieval, while the original query is retained for final generation.

If the planner output cannot be parsed, the implementation falls back to one local retrieval using the original query. Under ADAPTIVE-ACE, only the branches enabled by the selected context policy are executed; under FULL-ACE, both retrieval branches are executed. The resulting evidence increments are subsequently consolidated with the initial context before task-conditioned adaptation.

E.4 Task-Conditioned Synthesis

The consolidated context preserves provenance through explicit round, path, goal, and retrieval-mode headers. ACE-GraphRAG then applies separate synthesis instructions for fact-oriented multi-hop question answering and open-ended domain synthesis, corresponding to the task-conditioned adaptation stage in Section 4. In both task families, the original user query is retained for final generation, whereas rewritten branch queries are used only for supplementary retrieval.

E.4.1 Multi-Hop QA

The multi-hop QA prompt is used for HotpotQA and 2WikiMultiHopQA (Yang et al., 2018; Ho et al., 2020). It prioritizes entity-level facts, relations, and source-level evidence. When general reports conflict with more specific evidence, the prompt instructs the generator to prioritize the latter. It also requires the bridge entity to be identified before a concise final answer is produced.

E.4.2 Open-Ended Domain Synthesis

The open-ended synthesis prompt is used for the UltraDomain subsets (Qian et al., 2025). It integrates overview summaries, localized evidence, and alternative perspectives and asks the generator to organize these heterogeneous evidence types into a coherent response. The requested dimensions of comprehensiveness, diversity, and empowerment correspond to the evaluation dimensions adopted in our experimental setup following prior long-context RAG studies (Guo et al., 2025). The prompt further instructs the generator to prioritize more specific evidence when local and high-level descriptions differ.

Topology-routing prompt used by Adaptive-ACE

```
You are the ACE Strategic Router. Analyze the query's LOGICAL TOPOLOGY to dispatch the Context
Management Protocol.

### TOPOLOGY DEFINITIONS:
1. 'POINT' (Fact Lookup): Seeks a specific attribute of 1 entity.
  - Example: "When was Einstein born?"
2. 'CONTRAST' (Comparison): Compares attributes across 2+ entities.
  - Example: "Who is older, A or B?" / "Do they share the same nationality?"
3. 'CHAIN' (Sequential Dependency): Traces a relational chain (A -> B -> C).
  - Example: "Who is the spouse of the father of X?"
4. 'OVERVIEW' (Thematic Synthesis): Seeks broad domain summaries or exploration.
  - Example: "Summarize the history of AI" / "Overview of agriculture policy."

Query: "<QUERY>"

Output ONLY a JSON object:
{"topology": "POINT" | "CONTRAST" |
  "CHAIN" | "OVERVIEW"}
```

Figure 3: Topology-routing prompt used by ADAPTIVE-ACE.

Gap-aware differential-retrieval planning prompt

```
—Role—
You are a Multi-Dimensional Retrieval Planner for a Hierarchical RAG system.

—Goal—
Analyze the Original Query and the Round 1 Context. Your task is to generate TWO distinct,
complementary search queries to maximize Diversity and Comprehensiveness while maintaining
Empowerment.

—Tasks for Queries—
1. [Empowerment/Depth Query]: Focus on the most critical missing factual detail, logical evidence,
or specific data points. Target the "Hard Facts".
2. [Diversity/Breadth Query]: Focus on a completely different angle, alternative perspective, or
related sub-topic that was NOT covered in Round 1. Target the "Missing Dimensions".

—Instructions—
1. Gap Analysis: Briefly state what is missing based on the criteria.
2. Key Entity Extraction: Identify specific entities from the context that need further
exploration.
3. Keyword Optimization: Use dense, noun-heavy keywords. Avoid conversational filler.
4. Differentiation: Ensure Query 1 and Query 2 target different clusters in the Knowledge Graph.

—Output Format (Strict JSON)—
{
  "analysis": "Brief gap analysis",
  "queries": [
    {"query": "Dense keywords for Depth",
      "mode": "hi_local",
      "goal": "empowerment"},
    {"query": "Dense keywords for Breadth",
      "mode": "hi_global",
      "goal": "diversity"}
  ]
}

—Input—
Original Query: <QUERY>
Retrieved Context: <ROUND1_CONTEXT>

—Response (JSON only)—
```

Figure 4: Gap-aware differential-retrieval planning prompt.

Multi-hop QA synthesis prompt

```
—Role—
You are a precision-focused QA expert. Trace the multi-hop reasoning path to find the exact answer.

—Context—
<EVOLVED_CONTEXT>

—Strict Rules—
1. Focus ONLY on specific names, dates, and facts in 'Entities' and 'Sources'.
2. IGNORE general summaries in 'Reports' if they conflict with specific details.
3. Identify the bridge entity that connects the question components.
4. Your final result MUST be a single entity name or a short phrase.

—Output Format—
Reasoning: <BRIEF_LOGIC_CHAIN>
Result: [FINAL_ANSWER] <EXACT_ANSWER> [/FINAL_ANSWER]
```

Figure 5: Multi-hop QA synthesis prompt.

Open-ended domain synthesis prompt

```
—Role—
You are an expert knowledge synthesizer. You are provided with multi-layered context retrieved
through a strategic process (Overview, Deep-dive, and Broad-perspective).

—Goal—
Generate a response that is Comprehensive, Diverse, and Empowerment-oriented.
- Comprehensiveness: Integrate all relevant details from the provided rounds.
- Diversity: Explicitly present different perspectives, background themes, and related entities.
- Empowerment: Use specific evidence and factual details from the deep-dive path to support your
conclusions.

—Instructions—
1. Synthesize, Don't Repeat: Analyze the relationship between the General Overview (Round 1), the
Specific Evidence (Path 1), and the Broad Perspectives (Path 2).
2. Structure: Use markdown sections. Start with a clear executive summary, then elaborate on
specific dimensions, and conclude with actionable insights or implications.
3. Consistency: If Path 1 (Local Facts) clarifies or refines Round 1 (Global Summary), prioritize
the more detailed facts.
4. General Knowledge: Incorporate relevant general knowledge only to connect the provided context,
but do not make up facts.

—Target Response Type—
<RESPONSE_TYPE>

—Multi-Round Context Data—
<EVOLVED_CONTEXT>

—Final Synthesis Task—
Based on the multi-layered information above, provide the most high-quality, professional answer
to the user's question.
```

Figure 6: Open-ended domain synthesis prompt.