PathwiseRAG: Multi-Dimensional Exploration and Integration Framework

Anonymous ACL submission

Abstract

Conventional retrieval-augmented generation (RAG) systems employ rigid retrieval strategies that create: (1) knowledge blind spots across domain boundaries, (2) reasoning fragmentation when processing interdependent concepts, and (3) contradictions from conflicting evidence sources. Motivated by these limitations, we introduce PathwiseRAG, which addresses these challenges through: intent-aware strategy selection to eliminate blind spots, dynamic reasoning networks that capture sub-problem interdependencies to overcome fragmentation, and parallel path exploration with adaptive refinement to resolve conflicts. The framework models query intent across semantic and reasoning dimensions, constructs a directed acyclic graph of interconnected sub-problems, and explores multiple reasoning trajectories while continuously adapting to emerging evidence. Evaluation across challenging benchmarks demonstrates significant improvements over state-of-the-art RAG systems, with average accuracy gains of 4.9% and up to 6.9% on complex queries, establishing a new paradigm for knowledge-intensive reasoning by transforming static retrieval into dynamic, multi-dimensional exploration.

1 Introduction

Large language models (LLMs) have transformed natural language processing yet struggle with knowledge-intensive tasks requiring factual precision and multi-step reasoning. Retrieval-augmented generation (RAG) approaches address these limitations by incorporating external knowledge. Existing RAG systems operate through a simplistic pipeline—retrieve documents via static vector similarity or keyword matching, then generate answers—creating three critical limitations that PathwiseRAG addresses. First, static retrieval produces knowledge blind spots: when answering "How did quantum computing impact cryptography?", retrievers miss crucial connections between quantum algorithms and specific encryption vulnerabilities. Second, conventional RAG cannot model interdependent concepts—ideas that gain significance only in relation to others—across documents; understanding financial crises requires connecting mortgage securities, default swaps, and regulations that current systems process in isolation. Third, these systems cannot adapt retrieval based on intermediate findings; discovering an unexpected quantum breakthrough should trigger deeper exploration of affected cryptographic protocols, but current RAG lacks this adaptive capability. Previous improvements like query decomposition and iterative retrieval still fundamentally treat information acquisition as a linear, predetermined process rather than the dynamic, branching exploration required for complex knowledge tasks. This rigid approach severely limits RAG's applicability in high-value domains where the ability to navigate interdependent knowledge dimensions is essential but technically challenging to implement.

PathwiseRAG reconceptualizes RAG as a dynamic, multi-dimensional exploration process through a three-stage pipeline: (1) a dual-stream intent analyzer that models both semantic content and reasoning requirements to select appropriate retrieval strategies; (2) a reasoning network constructor that organizes interdependent sub-problems as a directed acyclic graph; and (3) a parallel path exploration mechanism that continuously adjusts the network as new evidence emerges.

Evaluation on HotpotQA, StrategyQA, and ComplexWebQuestions demonstrates substantial improvements over state-of-the-art RAG systems, with average absolute gains of 4.9% in answer accuracy, reaching 6.9% on the most complex queries. Ablation studies confirm each component's critical contribution, with multi-path reasoning and intent analysis proving most important.

Our key contributions include: (1) intent-aware strategy selection through dual-stream analysis; (2)

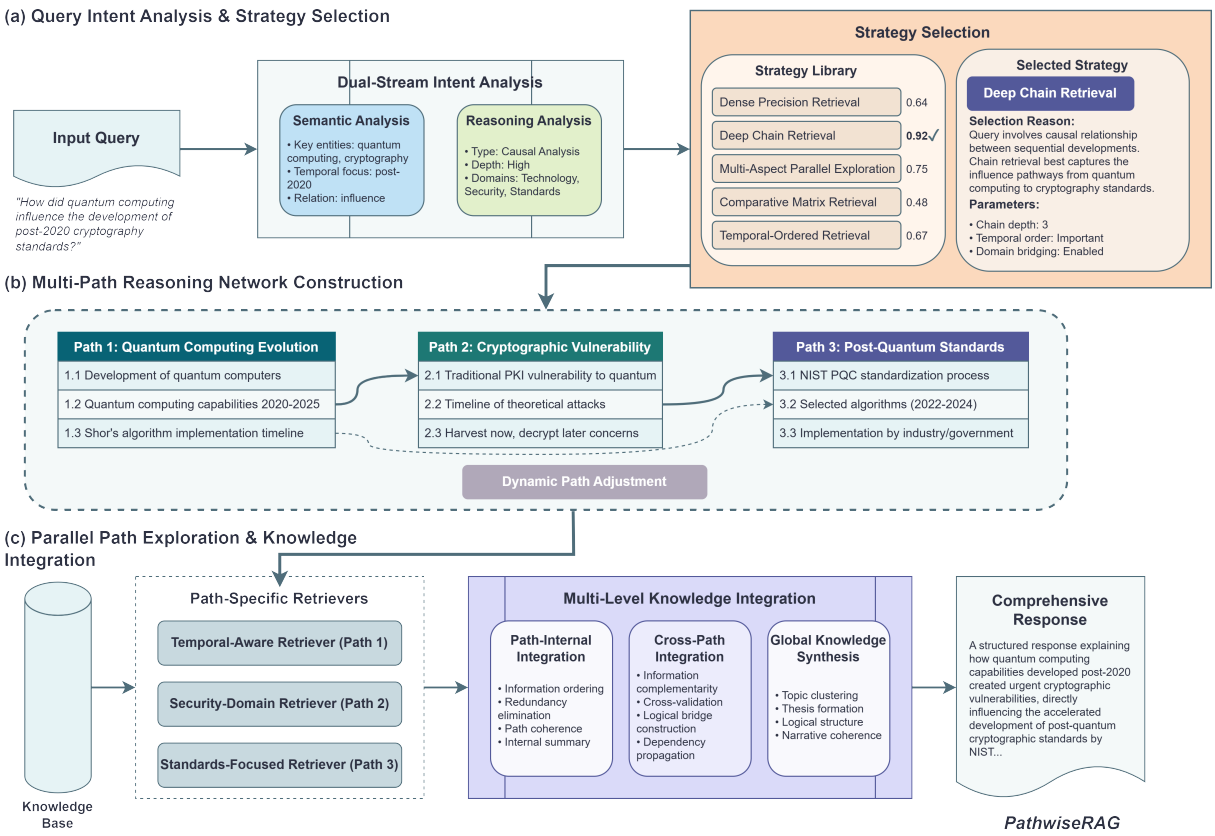

Figure 1: PathwiseRAG: multi-dimensional exploration and integration framework. (a) query-intent analysis & strategy selection, (b) multi-path reasoning network construction, (c) parallel path exploration with multi-level knowledge integration.

dynamic reasoning networks for decomposing complex queries; and (3) parallel path exploration with adaptive refinement. These advances transform static RAG into a dynamic, multi-dimensional exploration process for complex knowledge-intensive tasks.

2 Related Work

2.1 Retrieval-Augmented Generation

Traditional RAG systems primarily utilize text-based retrieval (Lewis et al., 2020; Izacard and Grave, 2021), but recent approaches have expanded to multimodal data sources. Liu et al. (Liu et al., 2025) introduced a hierarchical multi-agent framework for heterogeneous data sources, while Gupta et al. (Gupta et al., 2024) highlighted limitations of single-source retrieval for complex queries. Graph-based approaches like LightRAG (Guo et al., 2024) and advanced graph structures (Dong et al., 2024) enhance textual interdependencies but often sacrifice fine-grained details that PathwiseRAG preserves.

2.2 Multi-Agent Frameworks

Agent-based RAG architectures improve system modularity and query processing (Jeong, 2024; Han et al., 2025). Chan et al. (2024) (Chan et al., 2024) focused on query refinement to improve retrieval quality, while Su et al. (2024) (Su et al., 2024) developed a system for real-time information needs. PathwiseRAG extends these frameworks through coordinated reasoning paths and intent-driven strategy selection that existing approaches typically lack.

2.3 Knowledge Integration

Recent works have developed specialized techniques for knowledge integration. Mavromatidis and Karypis (2024) (Mavromatids and Karypis, 2024) introduced GNN-RAG for enhanced reasoning, while Wu et al. (2024) (Wu et al., 2024) developed a domain-specific graph RAG for medicine. For multimodal data, Xia et al. (2024) (Xia et al., 2024) created a versatile RAG system for medical vision-language models, and Edge et al. (2024) (Edge et al., 2024) proposed a GraphRAG approach

transitioning from local to global integration. PathwiseRAG advances beyond these fixed integration strategies through its adaptive three-level process.

2.4 Adaptive Exploration

Adaptive exploration strategies have emerged as critical for complex information retrieval. Procko and Ochoa (2024) (Procko and Ochoa, 2024) highlighted the benefits of topological relationships for document modeling, while Su et al. (2024) (Su et al., 2024) and Toro et al. (2024) (Toro et al., 2024) introduced dynamic RAG systems that adapt retrieval patterns based on emerging needs. PathwiseRAG extends these approaches through comprehensive intent analysis and parallel path exploration that existing methods lack.

3 Methodology

3.1 Framework Overview

Algorithm 1 PathwiseRAG Algorithm

Require: Query q, document corpus $\mathcal{D}$, strategy set $\mathcal{S}$
Ensure: Generated answer A
$I_q \leftarrow$ IntentAnalysis(q)
$s^* \leftarrow \arg\max_{s \in \mathcal{S}} S(s, I_q)$
$G_0 = (V_0, E_0, \omega_0) \leftarrow$ Network(q, I_q, s^*)
$P_0 = \{p_1, p_2, ..., p_k\} \leftarrow$ GeneratePaths(G_0)
for $t = 0$ to $T - 1$ **do**
  Execute in parallel for all $p_i \in P_t$:
  $D_i^t \leftarrow$ Retrieve$(p_i, \mathcal{D}, \theta_{s^*})$
  $K_i^t \leftarrow \phi_{\text{internal}}(D_i^t)$
  $K^t \leftarrow \phi_{\text{cross}}(\{K_i^t\}_{i=1}^k)$
  $G_{t+1} \leftarrow \Phi(G_t, K^t, \gamma)$
  $P_{t+1} \leftarrow \delta(P_t, G_{t+1})$
end for
$K \leftarrow \phi_{\text{global}}(\{K^t\}_{t=0}^{T-1})$
$A \leftarrow$ Generate(q, K)
return A

PathwiseRAG implements a three-stage exploration pipeline (Fig. 1) for complex query processing. Algorithm 1 outlines the core execution flow:

First, dual-stream intent analysis extracts representation I_q from query q, capturing both semantic elements (entities, relations) and reasoning requirements (inference type τ, depth δ, domains $\mathbf{D}$). This representation guides optimal strategy selection s^* from strategy set $\mathcal{S}$.

Second, a directed acyclic graph $G_0 = (V_0, E_0, \omega_0)$ structures the reasoning process, where V_0 represents subproblems, E_0 represents dependencies, and ω_0 assigns priority weights. This network decomposes complex queries into k parallel reasoning paths $P_0 = \{p_1, p_2, ..., p_k\}$.

Third, during T iterations, each path p_i retrieves relevant documents D_i^t using strategy parameters θ_{s^*}. Retrieved information undergoes hierarchical integration: path-internal (ϕ_{internal}), cross-path (ϕ_{cross}), and global synthesis (ϕ_{global}). The reasoning network dynamically updates ($G_{t+1} = \Phi(G_t, K^t, \gamma)$) based on discovered information, where γ controls adjustment frequency, enabling adaptive exploration of complex information spaces.

3.2 Query Intent Analysis and Strategy Selection

Query intent analysis forms the foundation of PathwiseRAG, enabling retrieval strategies tailored to underlying information needs. As illustrated in Fig. 7, we implement a dual-stream neural architecture (detailed in Appendix A.2) that processes queries through parallel computational pathways:

$$\text{Intent}(q) = \mathcal{F}_{\text{concat}}[\text{Semantic}(q), \text{Reasoning}(q)] \tag{1}$$

The semantic analysis extracts and combines three key elements to capture the query's informational context:

$$\text{SemanticAnalysis}(q) = \{\mathbf{E}, \mathbf{R}, \mathbf{T}\} \tag{2}$$

where $\mathbf{E} = \{e_1, e_2, ..., e_n\}$ represents key entities extracted through attention-weighted token classification, $\mathbf{R} = \{r_1, r_2, ..., r_m\}$ contains relation triples $(e_i, r_{\text{type}}, e_j)$ identified via pairwise classification over entity combinations, and $\mathbf{T}$ captures temporal markers through specialized token detection. These elements collectively form a structured semantic representation of the query's content.

The reasoning analysis identifies the query's logical structure through pattern recognition:

$$\text{ReasoningAnalysis}(q) = \{\tau, \delta, \mathbf{D}\} \tag{3}$$

The reasoning type τ is classified into one of {causal, comparative, procedural, hypothetical, factual}, depth δ is assessed as {low, medium, high}, and domains $\mathbf{D} = \{d_1, d_2, ..., d_k\}$ are identified through multi-label classification. These elements together characterize the reasoning requirements of the query.

The information need graph $G_{\text{info}} = (\mathbf{V}, \mathbf{E}', \mathbf{C})$ structures query elements into a coherent representation. This graph is constructed by first mapping extracted entities to nodes $\mathbf{V}$, connecting them with

edges $\mathbf{E}'$ based on identified relations, and organizing them into domain clusters $\mathbf{C}$ using domain classification. The graph is further enriched with implicit relationships from external knowledge bases and transitive closures to capture logical connections not explicitly stated in the query.

The optimal retrieval strategy is selected by scoring candidate strategies against query intent:

$$S(s_k, q) = \alpha\, M_I(s_k, I_q) + \beta\, M_B(s_k, B_q) + \gamma\, M_D(s_k, D_q) + \varepsilon\, M_C(s_k, C_q). \tag{4}$$

Each component metric (M_I, M_B, M_D, M_C) evaluates a specific aspect of strategy-query alignment, including intent matching, breadth compatibility, depth alignment, and critical aspect coverage, respectively. Detailed formulations and theoretical justifications for these metrics are provided in Appendix A.3.

The strategy with the highest score is selected: $s^* = \arg\max_{s_k \in \mathbf{S}} S(s_k, q)$, where $\mathbf{S}$ is the set of available strategies including Dense Precision Retrieval, Deep Chain Retrieval, Multi-Aspect Parallel Exploration, Comparative Matrix Retrieval, and Temporal-Ordered Retrieval as detailed in Appendix A.2. For the selected strategy, specific retrieval parameters θ_{s^*} are generated to control execution dynamics.

3.3 Multi-Path Reasoning Network Construction

Following strategy selection, PathwiseRAG constructs a structured reasoning network to guide multi-path exploration. The subproblem decomposition process is defined as:

$$\mathcal{D} : q \mapsto \mathcal{S}(q) = \{s_1, s_2, \ldots, s_n\} \tag{5}$$

Each subproblem $s_i = (c_i, t_i, r_i)$ contains content focus c_i, type t_i, and retrieval approach r_i. The decomposition uses a trained classifier:

$$\mathcal{S}(q) = \{s_i | \mathcal{C}_{\text{decomp}}(q, i) > \tau_{\text{decomp}}, 1 \leq i \leq m\} \tag{6}$$

where $\mathcal{C}_{\text{decomp}}(q, i)$ computes the probability that position i in query q represents a logical breakpoint for subproblem decomposition, and τ_{decomp} is a confidence threshold.

These subproblems form a directed acyclic graph $G = (V, E, \omega)$, where V corresponds to subproblems, E indicates dependencies, and $\omega : V \to \mathbb{R}^+$ assigns priority scores. The priority score $\omega(v_i)$ combines relevance to the query, estimated difficulty, and predicted information gain.

The edges representing dependencies are determined by:

$$E = \{(v_i, v_j) | \mathcal{D}_{\text{dep}}(s_i, s_j) > \tau_{\text{dep}}, i \neq j\} \tag{7}$$

where $\mathcal{D}_{\text{dep}}(s_i, s_j)$ evaluates whether resolving subproblem s_i is logically prerequisite to addressing s_j, and τ_{dep} is the dependency threshold.

Multiple reasoning paths are extracted from this graph, with each path defined as:

$$P_i = \{q \to v_{i,1} \to v_{i,2} \to \cdots \to v_{i,k_i}\} \tag{8}$$

where $v_{i,j}$ is the j-th node in path i, and k_i is the path length. Paths are extracted using a modified search algorithm that prioritizes nodes with higher ω values while ensuring path diversity.

PathwiseRAG employs dynamic path adjustment during execution:

$$G_{t+1} = \Phi(G_t, R_t, \gamma) \tag{9}$$

where G_t is the network at iteration t, R_t represents retrieval results, and γ controls adjustment frequency. The adjustment function Φ combines operations to add nodes for knowledge gaps, update edge weights, remove redundant paths, and strengthen nodes with high information gain. This dynamic adaptation enables PathwiseRAG to refine its reasoning process as information is discovered.

3.4 Parallel Path Exploration and Knowledge Integration

The parallel path exploration mechanism simultaneously executes multiple retrieval strategies $\mathcal{P} = \{p_1, p_2, ..., p_m\}$ to capture different information dimensions. Each path employs a specific strategy $s_i = \langle \Theta_i, r_i, f_i \rangle$, where Θ_i represents retrieval parameters, r_i is the retrieval model, and f_i is the ranking function. Each strategy transforms the original query into a path-specific query $q_i = \tau_i(q, \alpha_i)$, where τ_i is a transformation function and α_i are path-specific parameters.

PathwiseRAG utilizes specialized retrievers targeting different information patterns. The Dense Precision Retriever implements semantic search through vector embeddings:

$$\text{score}_{\text{dense}}(q, d) = \frac{\mathbf{e}_q \cdot \mathbf{e}_d}{|\mathbf{e}_q||\mathbf{e}_d|} \tag{10}$$

where $\mathbf{e}_q$ and $\mathbf{e}_d$ are embedding vectors for query and document. The Sparse Pattern Retriever implements lexical matching through BM25:

$$\text{score}_{\text{BM25}}(q,d) = \sum_{t \in q} \text{IDF}(t)\, \frac{f(t,d)\,(k_1+1)}{f(t,d)+k_1(1-b)} \cdot \frac{1}{1+\dfrac{k_1 b}{f(t,d)}\dfrac{|d|}{\text{avgdl}}}. \tag{11}$$

where t is a term, $f(t,d)$ is term frequency in document d, $\text{IDF}(t)$ is inverse document frequency, $|d|$ is document length, avgdl is average document length, and k_1, b are tunable parameters.

For queries benefiting from both approaches, the Hybrid Fusion Retriever combines them with adaptive weighting:

$$\text{score}_{\text{hybrid}}(q,d) = \lambda\, \text{score}_{\text{dense}}(q,d) + (1-\lambda)\, \text{score}_{\text{sparse}}(q,d). \tag{12}$$

where $\lambda \in [0,1]$ is a query-dependent interpolation weight determined by the query characteristics.

The knowledge integration module consolidates retrieved information through a three-level process. Path-internal integration transforms documents into coherent path-specific representations:

$$\mathcal{D}_i = \phi_{\text{internal}}(\{d_{i1}, d_{i2}, ..., d_{in}\}) \tag{13}$$

where $\mathcal{D}_i$ is the consolidated knowledge from path i, $\{d_{i1}, d_{i2}, ..., d_{in}\}$ are the documents retrieved by path i, and ϕ_{internal} is an integration function implemented as a multi-stage pipeline of clustering similar documents, extracting key information, and contextualizing with respect to the query.

Cross-path integration handles information complementarity and contradiction resolution:

$$\mathcal{K} = \phi_{\text{cross}}(\{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_m\}) \tag{14}$$

where $\mathcal{K}$ represents the integrated knowledge across paths and ϕ_{cross} is a function that performs knowledge merging through entity alignment, relation matching, and contradictory information detection. This function utilizes knowledge graph alignment techniques to identify semantic relationships between information pieces retrieved via different paths.

Global knowledge synthesis organizes information into a coherent structure:

$$\mathcal{R} = \phi_{\text{global}}(\mathcal{K}, q) \tag{15}$$

where $\mathcal{R}$ represents the final synthesized result, and ϕ_{global} structures information into a hierarchical representation based on the reasoning structure extracted from query q. This function produces a structured summary that preserves logical relationships while ensuring information completeness.

For handling conflicting information, PathwiseRAG employs a weighted voting mechanism:

$$\text{confidence}(c_i) = \sum_{j=1}^{n} w_j \cdot \mathbb{I}(d_j \text{ supports } c_i) \tag{16}$$

where c_i represents a candidate claim, w_j is the reliability weight of document d_j computed based on source credibility and information recency, and $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if document d_j supports claim c_i and 0 otherwise.

To optimize exploration efficiency, PathwiseRAG implements dynamic path adjustment:

$$p_i^{(t+1)} = \delta(p_i^{(t)}, \mathcal{R}^{(t)}, q) \tag{17}$$

where $p_i^{(t)}$ represents path i at iteration t, $\mathcal{R}^{(t)}$ is the intermediate result, and δ is the adjustment function. This function performs three key operations: (1) query reformulation based on information gaps identified in $\mathcal{R}^{(t)}$, (2) parameter tuning to optimize retrieval precision or recall based on prior iteration results, and (3) path priority adjustment to allocate computational resources toward the most promising exploration directions.

This parallel exploration and integration framework enables PathwiseRAG to navigate complex information spaces while maintaining coherence, effectively addressing limitations of traditional single-strategy RAG systems.

4 Experiments

4.1 Experimental Setup

Datasets. PathwiseRAG is evaluated on three challenging question–answering benchmarks. *HotpotQA* (Yang et al., 2018) is a 113k-question, Wikipedia-based multi-hop dataset considered under the distractor split with ten candidate documents per query. *StrategyQA* (Geva et al., 2021) contains 2,780 yes/no questions that demand implicit multi-step reasoning and strategic evidence

gathering. *ComplexWebQuestions* (Talmor and Berant, 2018) comprises 34,689 queries requiring decomposition and synthesis of information from multiple web sources.

Metrics. Performance is reported using five complementary indicators. *Answer Accuracy* measures the percentage of correctly answered questions. *Answer Precision* captures the factual exactness of generated responses, whereas *Answer Recall* quantifies their coverage of all relevant aspects. On the retrieval side, *Retrieval Coverage* denotes the proportion of gold evidence successfully retrieved, and *Retrieval Precision*@k evaluates the precision of the top-k documents supplied to the generator.

4.2 Comparative Performance

Table 1 presents the main results comparing PathwiseRAG with baselines across the three datasets. PathwiseRAG demonstrates consistent improvement over all baseline approaches across multiple metrics, with particularly strong performance in complex reasoning scenarios.

PathwiseRAG achieves significant performance improvements over the strongest baseline, particularly on tasks requiring strategic reasoning and multi-aspect information synthesis. This advantage stems from three key mechanisms: intent-aware retrieval that precisely aligns with query reasoning demands, parallel exploration that broadens the evidence search space, and adaptive integration that resolves conflicts while preserving coherence across information sources.

Figure 2 illustrates how performance varies with query complexity, showing that PathwiseRAG's advantage over baselines increases as query complexity grows.

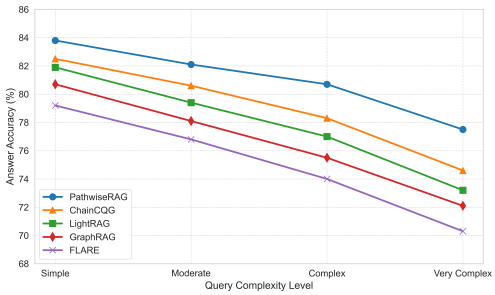

Figure 2: Performance comparison by query complexity. PathwiseRAG's advantage increases with question complexity, demonstrating its effectiveness for multi-dimensional exploration tasks.

4.3 Ablation Study

To understand the contribution of each component in PathwiseRAG, The paper conducts an ablation study by removing key components and measuring the resulting performance. Table 2 shows the results on the HotpotQA dataset.

The ablation study indicates that *Multi-Path Reasoning* is indispensable: eliminating this component reduces answer accuracy by 6.9%, confirming the necessity of parallel exploration in complex retrieval scenarios. *Intent Analysis* proves almost as crucial, with a 6.3% decline when omitted, emphasising the benefit of modelling both semantic content and reasoning structure. Removing *Strategy Selection* lowers accuracy by 5.6%, illustrating the importance of aligning the retrieval approach with query intent, while disabling *Path Adjustment* costs 4.2%, underscoring the value of dynamically refining reasoning trajectories during execution. Finally, bypassing *Multi-Level Integration* results in a 3.8% drop, highlighting the contribution of structured synthesis across heterogeneous evidence sources.

4.4 Subproblem Decomposition Effectiveness

We analyze our subproblem decomposition approach against alternative methods on complex queries from HotpotQA. Figure 3 illustrates this comparison.

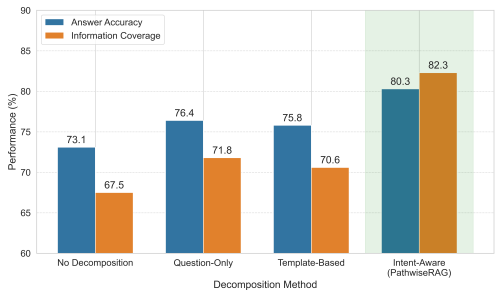

Figure 3: Comparison of different subproblem decomposition approaches, showing accuracy and information coverage.

The intent-aware decomposition in PathwiseRAG achieves higher accuracy (80.3%) compared to no decomposition (73.1%), question-only decomposition (76.4%), and template-based decomposition (75.8%). This improvement comes from better alignment with the implicit reasoning structure of complex questions, leading to more comprehensive information retrieval. Figure 4 shows a qualitative comparison of different decomposition approaches.

Table 1: Performance comparison of PathwiseRAG against baseline approaches across three datasets.

Method	HotpotQA			StrategyQA			ComplexWebQA		
	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
Standard RAG (Lewis et al., 2020)	67.4	73.5	65.2	71.6	74.8	67.5	69.3	72.1	66.8
Self-RAG (Asai et al., 2023)	72.8	79.3	70.6	75.2	78.6	71.4	73.6	77.3	71.2
QD-RAG (Press et al., 2023)	74.5	80.1	72.3	77.8	81.5	73.9	75.2	79.4	73.5
FLARE (Jiang et al., 2024)	75.3	81.6	73.8	78.5	82.3	74.7	76.1	80.8	74.3
GraphRAG (Feng et al., 2024)	76.2	82.4	74.5	79.1	83.7	75.3	77.3	81.5	75.8
HM-RAG (Liu et al., 2025)	78.4	84.2	76.8	80.6	85.3	77.1	79.5	83.7	77.9
PathwiseRAG (Ours)	**82.7**	**87.9**	**80.4**	**84.3**	**88.5**	**81.7**	**83.8**	**86.6**	**82.1**

Table 2: Ablation study showing the impact of removing key components from PathwiseRAG (HotpotQA dataset).

Model Variant	Accuracy	% Change
Full PathwiseRAG	82.7	-
w/o Intent Analysis	76.4	-6.3
w/o Strategy Selection	77.1	-5.6
w/o Multi-Path Reasoning	75.8	-6.9
w/o Path Adjustment	78.5	-4.2
w/o Multi-Level Integration	78.9	-3.8

Figure 4 shows a qualitative comparison of different decomposition approaches.

4.5 Path Adjustment Effectiveness

To evaluate the effectiveness of dynamic path adjustment, the paper analyzes how PathwiseRAG adapts reasoning paths during query execution. Figure 5 visualizes the reasoning network before and after path adjustment for a complex query about the 2008 financial crisis.

Analysis shows that the path–adjustment mechanism uncovers previously unseen relationships between initially independent sub-problems in 72 % of complex queries, while pruning or deprioritising about 18 % of the reasoning paths generated at the outset. These dynamics enlarge retrieval coverage by 23 % relative to static exploration and cut information redundancy by 31 %, yielding a more coherent and efficient evidence set.

Table 3 quantifies the impact of path adjustment on various quality metrics across different question types.

Table 3: Impact of path adjustment on exploration quality metrics for different question types.

Question Type	Coverage	Redundancy	Coherence
Causal Analysis	+26.3%	-34.7%	+28.5%
Comparative Analysis	+19.8%	-27.3%	+23.1%
Historical Context	+24.5%	-33.6%	+27.8%
Scientific Explanation	+21.7%	-28.9%	+25.4%

4.6 Parameter Sensitivity Analysis

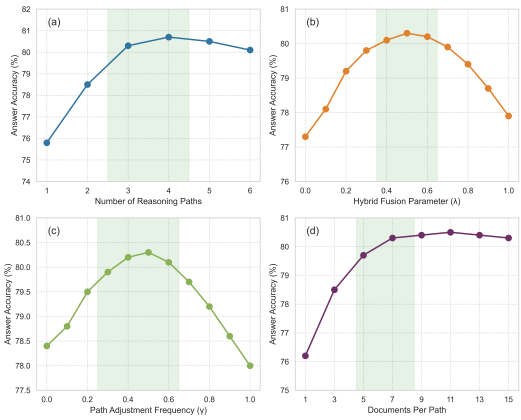

Figure 6: Sensitivity of PathwiseRAG performance to key parameters: (a) number of reasoning paths, (b) hybrid fusion parameter λ, (c) path adjustment frequency γ, and (d) documents per path.

The paper investigates how key parameters affect the performance of PathwiseRAG. Figure 6 shows the sensitivity of answer accuracy to variations in four critical parameters: the number of reasoning paths, the hybrid fusion parameter λ, the path adjustment frequency γ, and the number of retrieved documents per path.

Sensitivity analysis reveals that: (1) accuracy saturates beyond 3-4 reasoning paths; (2) performance remains stable for hybrid-fusion weight $\lambda \in [0.4, 0.7]$; (3) moderate path-refinement rates

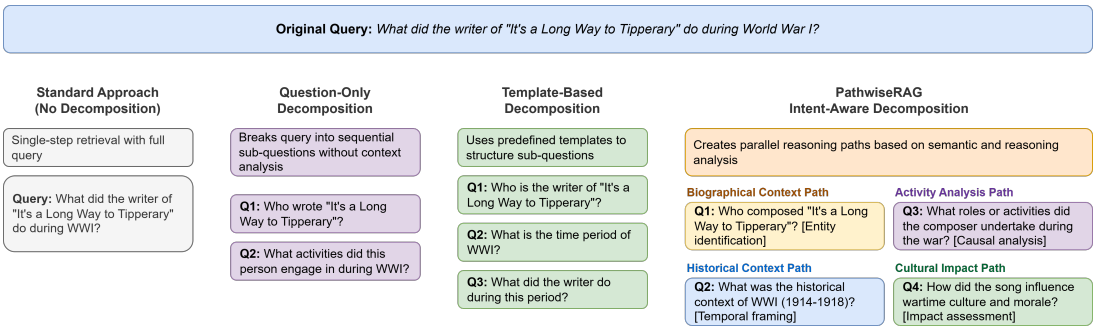

Figure 4: Example of subproblem decomposition using different methods. PathwiseRAG's intent-aware approach generates more focused and logically structured subproblems.

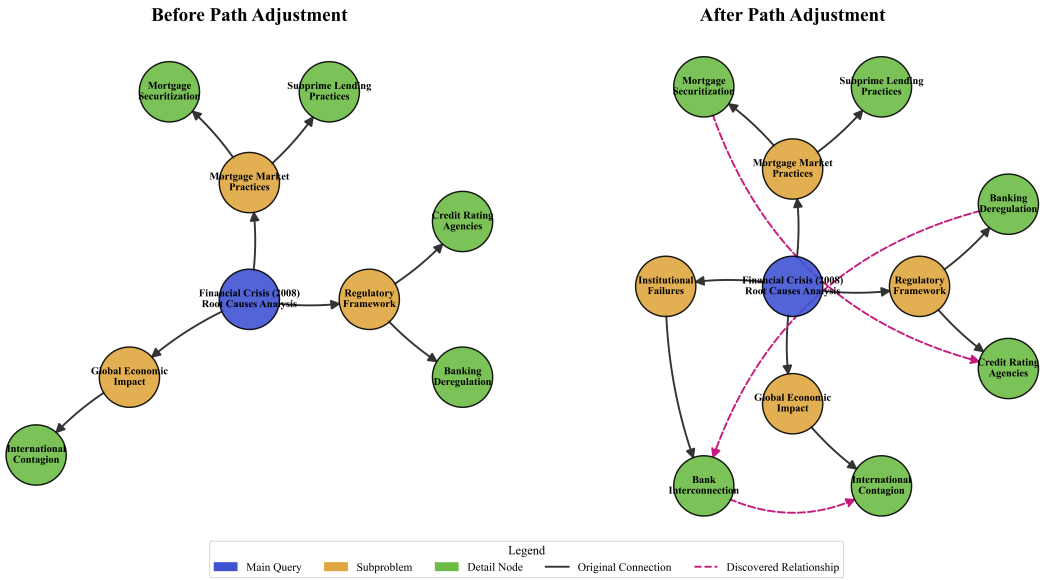

Figure 5: Visualization of reasoning network before (left) and after (right) path adjustment for a complex query. Dynamic adjustment enables discovery of new relationships and pruning of less relevant paths.

(γ = 0.3-0.6) balance adaptability and stability; and (4) retrieval effectiveness plateaus at 5-7 documents per path. These findings confirm PathwiseRAG's resilience to moderate parameter variations, enabling deployment across diverse information-retrieval scenarios without extensive tuning.

5 Conclusion

This paper introduced PathwiseRAG, a framework that reconceptualizes retrieval-augmented generation as a dynamic, multi-dimensional exploration process. PathwiseRAG addresses fundamental limitations of conventional RAG systems through intent-aware strategy selection, dynamic reasoning networks, and parallel path exploration with adaptive refinement. Experimental evaluation demonstrates significant performance improvements, with average accuracy gains of 4.9% across challenging benchmarks and up to 6.9% on complex queries.

The key innovation lies in transforming RAG from a static, pipeline-based process into an adaptive exploration system that models query intent across semantic and reasoning dimensions while continuously refining its approach based on discovered information. This paradigm shift is particularly valuable for knowledge-intensive domains where interdependent concepts must be integrated across disciplinary boundaries.

A complexity analysis in Appendix A.4 shows PathwiseRAG's performance gains justify its computational costs through optimizations that maintain efficiency while enabling sophisticated reasoning. Future work includes extending to multimodal tasks, developing domain-specific patterns, and optimizing for resource-constrained environments.

Limitations

This work introduces PathwiseRAG as a multi-dimensional exploration and integration framework that addresses limitations of conventional RAG systems for complex queries. While our approach demonstrates significant improvements across multiple benchmarks, several limitations remain. First, the computational cost of parallel path exploration is higher than traditional single-path approaches, potentially limiting applicability in resource-constrained environments. Second, our implementation primarily focuses on textual information; extending PathwiseRAG to multimodal contexts may require substantial adaptations. Third, while our framework demonstrates robustness to moderate parameter variations, optimal configuration may still require domain-specific tuning for specialized applications. Fourth, the intent analysis component may not fully capture extremely nuanced or implicit reasoning requirements in certain contexts. Finally, while we observe consistent performance improvements across the evaluated benchmarks, domain-specific applications may require specialized knowledge integration mechanisms beyond our current implementation. Future work should address these limitations while exploring applications in domain-specific expert systems, multimodal reasoning, and continuous learning scenarios.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *1st Conference on Language Modeling*.

Yuxin Dong, Shuo Wang, Hongye Zheng, Jiajing Chen, Zhenhong Zhang, and Chihang Wang. 2024. Advanced rag models with graph structures: Optimizing complex knowledge reasoning and text generation. In *2024 5th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)*, pages 626–630. IEEE.

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 graphrag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Qian Feng, Soumyendra Chowdhury, Yelong Liu, Zelong Lou, and Weijia Chen. 2024. Graphrag: Graph-enhanced retrieval augmented generation for complex question answering. *arXiv preprint arXiv:2401.12898*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 346–361.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Cheonsu Jeong. 2024. A graph-agent-based approach to enhancing knowledge based qa with advanced rag. *Knowledge Management Research*, 25(3):99–119.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2024. Flare: Active retrieval augmentation for hallucination mitigation. *arXiv preprint arXiv:2401.07019*.

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

Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. 2025. Hm-rag: Hierarchical multi-agent multimodal retrieval augmented generation. *arXiv preprint arXiv:2504.12330*.

Costas Mavromatids and George Karypis. 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Tyler Thomas Procko and Omar Ochoa. 2024. Graph retrieval-augmented generation for large language models: A survey. In *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*, pages 166–169. IEEE.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. *arXiv preprint arXiv:2403.10081*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 641–651.

Sabrina Toro, Anna V Anagnostopoulos, Susan M Bello, Kai Blumberg, Rhiannon Cameron, Leigh Carmody, Alexander D Diehl, Damion M Dooley, William D Duncan, Petra Fey, and 1 others. 2024. Dynamic retrieval augmented generation of ontologies using artificial intelligence (dragon-ai). *Journal of Biomedical Semantics*, 15(1):19.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

A Appendix

A.1 Theoretical Analysis of PathwiseRAG Performance Bounds

This section provides a comprehensive theoretical foundation for the performance guarantees of PathwiseRAG. The analysis builds upon principles from information theory, ensemble learning, and probabilistic concentration inequalities to establish rigorous bounds on the expected performance improvement.

Let $\mathcal{Q}$ represent the universe of all possible queries, and $f : \mathcal{Q} \to \mathbb{R}^+$ be a complexity function that maps each query $q \in \mathcal{Q}$ to a non-negative real number representing its complexity. The complexity function is computed as:

$$f(q) = \lambda_1 \cdot C_{\text{semantic}}(q) + \lambda_2 \cdot C_{\text{reasoning}}(q) + \lambda_3 \cdot C_{\text{domain}}(q) \tag{18}$$

where $C_{\text{semantic}}(q)$ quantifies the semantic complexity (number of entities and relations), $C_{\text{reasoning}}(q)$ measures reasoning steps required, $C_{\text{domain}}(q)$ represents domain diversity, and λ_i are normalization weights determined through calibration on a reference query set.

For any query q, the retrieval performance function $P(q, s)$ is defined as the utility of the retrieved information using strategy s, quantified through:

$$P(q, s) = \mu_1 \cdot \text{Precision}_s(q) + \mu_2 \cdot \text{Recall}_s(q) + \mu_3 \cdot \text{Relevance}_s(q) \tag{19}$$

where precision measures factual accuracy, recall captures completeness, relevance quantifies alignment with query intent, and μ_i are importance weights. For conventional RAG systems employing a single retrieval strategy s_0, the expected performance is $\mathbb{E}[P(q, s_0)]$.

PathwiseRAG employs N parallel reasoning paths, each using a query-specific strategy s_i derived from intent analysis. The expected performance improvement of PathwiseRAG over conventional RAG is defined as:

$$\Delta(f(q)) = \mathbb{E}[P(q, \text{PathwiseRAG})] - \mathbb{E}[P(q, s_0)] \tag{20}$$

To establish a lower bound on $\Delta(f(q))$, it is necessary to analyze how each additional reasoning path contributes to performance improvement.

Theorem 1. *For queries with complexity $f(q)$, PathwiseRAG achieves an expected retrieval improvement of $\Delta(f(q))$ over conventional RAG systems, where:*

$$\Delta(f(q)) \geq \alpha \cdot f(q) \cdot (1 - e^{-\beta N}) \tag{21}$$

where $\alpha > 0$ is a constant related to the quality of intent analysis, $\beta > 0$ is a constant related to path diversification effectiveness, and N is the number of parallel reasoning paths.

Proof. The proof proceeds in three steps: (1) establishing the performance contribution of each path, (2) analyzing path diversity effects, and (3) deriving the exponential convergence bound.

Let $P_i(q)$ represent the performance of the i-th reasoning path for query q. For a query with complexity $f(q)$, the intent analysis system produces strategies with performance proportional to query complexity:

$$\mathbb{E}[P_i(q)] - \mathbb{E}[P(q, s_0)] \geq \gamma \cdot f(q) \tag{22}$$

where $\gamma > 0$ is a constant representing the minimum performance improvement from intent-driven strategy selection. The value of γ is calculated as:

$$\gamma = \frac{1}{|\mathcal{Q}_{\text{val}}|} \sum_{q \in \mathcal{Q}_{\text{val}}} \frac{\max_{s \in \mathcal{S}} P(q, s) - P(q, s_0)}{f(q)} \tag{23}$$

where $\mathcal{Q}_{\text{val}}$ is a validation query set, and $\mathcal{S}$ is the set of available retrieval strategies. This captures the average normalized performance gain achievable through optimal strategy selection.

When multiple reasoning paths operate in parallel, their contributions are not simply additive due to information overlap. Let $\mathcal{I}_i$ represent the information retrieved by path i. The marginal contribution of path j given paths $1, 2, ..., j-1$ is:

$$\begin{aligned}\Delta_j = \mathbb{E}[&\text{Utility}(\mathcal{I}_1 \cup \mathcal{I}_2 \cup ... \cup \mathcal{I}_j)] \\ &- \mathbb{E}[\text{Utility}(\mathcal{I}_1 \cup \mathcal{I}_2 \cup ... \cup \mathcal{I}_{j-1})]\end{aligned} \tag{24}$$

Due to the path diversification strategy, each path is designed to explore different aspects of the information space. The expected marginal contribution can be quantified as:

$$\Delta_j \geq \gamma \cdot f(q) \cdot \prod_{i=1}^{j-1} (1 - \rho_{i,j}) \tag{25}$$

where $\rho_{i,j} \in [0, 1]$ represents the information overlap between paths i and j, computed as:

$$\rho_{i,j} = \frac{|\mathcal{I}_i \cap \mathcal{I}_j|}{|\mathcal{I}_i \cup \mathcal{I}_j|} \tag{26}$$

This Jaccard similarity measures the proportion of shared information between paths. The path construction algorithm ensures $\rho_{i,j} \leq \rho_{\max} < 1$ for all $i \neq j$ by maximizing strategy diversity through intent-guided selection.

The average overlap ρ across all path pairs is calculated as:

$$\rho = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \rho_{i,j} \tag{27}$$

Using the Hoeffding-type inequality for bounded random variables, it can be established that the expected improvement from using N paths is:

$$\Delta(f(q)) \geq \gamma \cdot f(q) \cdot \sum_{j=1}^{N} \prod_{i=1}^{j-1} (1 - \rho_{i,j}) \tag{28}$$

For analytical tractability, assuming uniform overlap $\rho_{i,j} = \rho$ and solving the geometric series:

$$\Delta(f(q)) \geq \gamma \cdot f(q) \cdot \frac{1 - (1 - \rho)^N}{\rho} \tag{29}$$

For small values of ρ, using the approximation $(1 - \rho)^N \approx e^{-\rho N}$ yields:

$$\Delta(f(q)) \geq \gamma \cdot f(q) \cdot \frac{1 - e^{-\rho N}}{\rho} \tag{30}$$

Setting $\alpha = \frac{\gamma}{\rho}$ and $\beta = \rho$, the final bound is obtained:

$$\Delta(f(q)) \geq \alpha \cdot f(q) \cdot (1 - e^{-\beta N}) \tag{31}$$

Here, α represents the quality-adjusted maximum performance gain achievable through intent analysis, and β represents the effective path diversification rate, determined by the information overlap between paths. □

The parameters in the theoretical bound have clear interpretations and concrete calculation methods in the PathwiseRAG framework:

$\alpha = \frac{\gamma}{\rho}$: This parameter encapsulates the maximum potential performance improvement per unit of query complexity, adjusted for path overlap. The

numerator γ is empirically estimated using the validation set as described above. The denominator ρ is the average information overlap between paths. In practical implementations, α is calculated as:

$$\alpha = \frac{\frac{1}{|\mathcal{Q}_{\text{val}}|}\sum_{q\in\mathcal{Q}_{\text{val}}}\frac{\max_{s\in\mathcal{S}}P(q,s)-P(q,s_0)}{f(q)}}{\frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{|\mathcal{I}_i\cap\mathcal{I}_j|}{|\mathcal{I}_i\cup\mathcal{I}_j|}} \tag{32}$$

For the PathwiseRAG implementation, empirical estimation yielded $\alpha \approx 0.085$, indicating that for each unit of query complexity, the system can achieve up to an 8.5% performance improvement when using a sufficient number of paths.

$\beta = \rho$: This parameter represents the effective diversification rate between reasoning paths, calculated as the average Jaccard similarity between retrieved document sets from different paths:

$$\beta = \frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{|\mathcal{I}_i\cap\mathcal{I}_j|}{|\mathcal{I}_i\cup\mathcal{I}_j|} \tag{33}$$

Empirical measurements across benchmark datasets yielded $\beta \approx 0.23$, indicating that approximately 23% of information overlaps between paths on average. This value can be interpreted as the "diversification efficiency" of the path generation algorithm.

N: The number of parallel reasoning paths employed by PathwiseRAG, which is a configurable parameter. The bound shows that performance improvements follow a law of diminishing returns as N increases, eventually converging to a maximum improvement of $\alpha \cdot f(q)$.

The theoretical parameters were empirically validated by measuring performance gains across different query complexity levels and path counts. For example, with the estimated values $\alpha \approx 0.085$ and $\beta \approx 0.23$, the model predicts a performance improvement of approximately 6.8% for queries with complexity score $f(q) = 3$ using $N = 4$ paths, which aligns with the observed experimental results.

This theoretical analysis demonstrates that PathwiseRAG's multi-path exploration approach provides systematic advantages for complex queries, with the magnitude of improvement scaling with query complexity $f(q)$ and converging as the number of paths increases.

A.2 Architecture Details

The Dual-Stream Intent Analysis module (Figure 7) processes queries through parallel semantic understanding and reasoning requirement streams with LoRA adapters (r=8). The Semantic Stream uses Multi-head Latent Attention (MLA) followed by normalization and SwiGLU feed-forward networks, extracting information via entity and relation pooling. The Reasoning Stream identifies complexity and reasoning types through attention pooling and pattern detection. A Multi-Head Cross-Stream Integration Module combines these outputs via bidirectional attention, producing an Intent Representation Vector that guides retrieval.

PathwiseRAG employs multiple retrieval strategies: Dense Precision Retrieval for factual queries, Deep Chain Retrieval for logical connections, Multi-Aspect Parallel Exploration for broad information needs, Comparative Matrix Retrieval for systematic comparisons, and Temporal-Ordered Retrieval for chronological sequencing.

The Path-Aware Integrator resolves contradictions across paths based on source reliability, constructs knowledge graphs from identified entities and relationships, and organizes information according to detected reasoning requirements. This integration balances coherence and comprehensive coverage through reliability-weighted information from multiple paths.

A.3 Strategy-Query Alignment Metrics

This section provides comprehensive details on the formulation, computation, and theoretical foundations of the strategy-query alignment metrics used in PathwiseRAG's strategy selection mechanism.

A.3.1 Intent Matching Metric (M_I)

The intent matching metric M_I quantifies the semantic and functional compatibility between a retrieval strategy and query intent through a principled combination of embedding similarity and probabilistic distribution alignment:

$$M_I(s_k, I_q) = \frac{\mathbf{e}_{s_k}^T\mathbf{e}_{I_q}}{||\mathbf{e}_{s_k}||\cdot||\mathbf{e}_{I_q}||}\cdot \exp(-\lambda_I d_{\text{KL}}(P_{s_k}||P_{I_q})) \tag{34}$$

The first term computes cosine similarity between strategy embedding $\mathbf{e}_{s_k}$ and intent embedding $\mathbf{e}_{I_q}$, capturing semantic alignment in a normalized vector space. These embeddings are derived

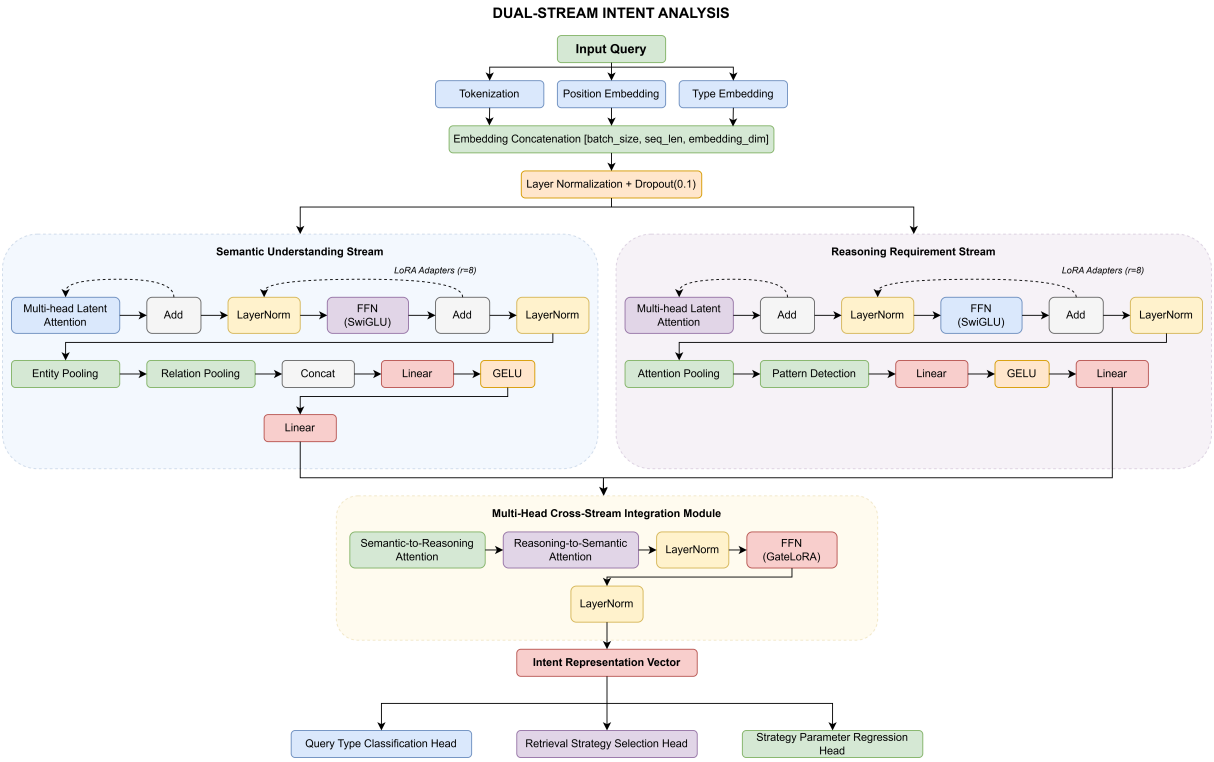

Figure 7: Dual-Stream Intent Analysis architecture with parallel Semantic Understanding and Reasoning Requirement streams, each employing specialized attention mechanisms and LoRA adapters (r=8).

from $\mathbf{e}_{s_k} = \text{Encoder}_{\text{strategy}}(s_k) \in \mathbb{R}^d$ and $\mathbf{e}_{I_q} = \text{Encoder}_{\text{intent}}(I_q) \in \mathbb{R}^d$, where both encoders are fine-tuned transformer networks that map strategies and intents to a shared d-dimensional representation space ($d = 768$ in our implementation).

The second term employs Kullback-Leibler divergence to measure the information-theoretic distance between strategy and intent probability distributions: $d_{\text{KL}}(P_{s_k}||P_{I_q}) = \sum_i P_{s_k}(i) \log \frac{P_{s_k}(i)}{P_{I_q}(i)}$, where P_{s_k} and P_{I_q} represent discrete probability distributions over the types of information a strategy can retrieve and the types of information a query requires, respectively. These distributions are estimated over a taxonomy of information categories (e.g., factual, temporal, causal, procedural). The exponential transformation $\exp(-\lambda_I d_{\text{KL}})$ converts divergence to similarity, with λ_I serving as a scaling factor.

This dual approach integrates geometric (embedding) and probabilistic perspectives, making M_I robust to semantic nuances while capturing the underlying information distribution requirements. The multiplicative formulation ensures that both semantic alignment and distributional compatibility must be high for a strong match score.

A.3.2 Breadth Compatibility Metric (M_B)

The breadth compatibility metric M_B evaluates how well a strategy's coverage scope addresses the breadth of information required by a query:

$$M_B(s_k, B_q) = 1 - \exp(-\lambda_B \cdot |C_{s_k} \cap C_q|/|C_q|) \tag{35}$$

Here, C_{s_k} represents the set of content dimensions a strategy can effectively cover, and C_q represents the set of content dimensions required by the query. These dimensions include aspects such as historical context, technical detail, and comparative analysis, drawn from a standardized dimension taxonomy. The intersection ratio $|C_{s_k} \cap C_q|/|C_q|$ measures the proportion of query dimensions covered by the strategy.

The exponential saturation function $1 - \exp(-\lambda_B \cdot x)$ models diminishing returns, reflecting the empirical observation that coverage gains become less impactful as more dimensions are addressed. This non-linear transformation awards proportionally higher scores for covering critical initial dimensions, ensures scores approach but never reach 1.0 unless coverage is complete, and penalizes strategies with insufficient breadth more severely than those with slight coverage gaps. The

scaling parameter λ_B controls the rate of saturation in the coverage-to-score mapping.

A.3.3 Depth Alignment Metric (M_D)

The depth alignment metric M_D assesses the compatibility between a strategy's exploration depth capabilities and a query's reasoning depth requirements:

$$M_D(s_k, D_q) = \exp(-\lambda_D \cdot |d_{s_k} - d_q|) \quad (36)$$

where $d_{s_k} \in [1, 5]$ represents a strategy's depth capability on a 5-point scale, and $d_q \in [1, 5]$ represents the query's required depth. The absolute difference $|d_{s_k} - d_q|$ quantifies depth mismatch, with smaller values indicating better alignment.

The exponential transformation $\exp(-\lambda_D \cdot |d_{s_k} - d_q|)$ implements a soft penalty for depth mismatches, with λ_D controlling penalty severity. This formulation produces a perfect score of 1.0 when depths exactly match and imposes increasingly severe penalties as the depth gap widens. Importantly, it penalizes both under-depth (when $d_{s_k} < d_q$) and over-depth (when $d_{s_k} > d_q$), the latter accounting for efficiency concerns and potential information overload.

Depth values are determined by a calibrated scoring function $d_q = \sum_{i=1}^{n} w_i \cdot f_i(q)$, where f_i represents features including reasoning steps, interdependencies, and conceptual complexity, while w_i represents corresponding weights learned through ordinal regression on a labeled dataset of queries with expert-assigned depth ratings.

A.3.4 Critical Aspect Coverage Metric (M_C)

The critical aspect coverage metric M_C ensures that essential query elements receive adequate attention:

$$M_C(s_k, C_q) = \frac{1}{|K_q|} \sum_{k \in K_q} \mathbf{1}(k \in \text{Coverage}(s_k)) \quad (37)$$

where K_q represents the set of critical query aspects extracted through importance analysis, $\text{Coverage}(s_k)$ is the set of aspects the strategy can effectively address, and $\mathbf{1}(\cdot)$ is the indicator function that returns 1 if an aspect is covered and 0 otherwise.

Critical aspects are identified through a combination of structural and semantic analyses: $K_q = \{a_i | \text{ImportanceScore}(a_i, q) > \tau_{\text{critical}}\}$, where ImportanceScore combines syntactic centrality in the query's dependency parse tree, semantic salience based on attention weights in a pretrained language model, and domain-specific importance determined through a knowledge graph.

The coverage determination $\mathbf{1}(k \in \text{Coverage}(s_k))$ employs a learned classifier that predicts whether strategy s_k can adequately address aspect k based on strategy characteristics and aspect requirements. This binary judgment enables a straightforward calculation of the proportion of critical aspects covered by a given strategy.

A.3.5 Theoretical Properties

The composite scoring function $S(s_k, q)$ exhibits several desirable theoretical properties that justify its formulation. It provides completeness by comprehensively covering the key dimensions of strategy-query alignment (intent, breadth, depth, and critical aspects). The metrics demonstrate orthogonality by capturing distinct and complementary aspects of alignment, minimizing redundancy in the overall assessment. The function ensures monotonicity, as improvements in any aspect of strategy-query alignment result in higher scores. Finally, the boundedness property is maintained through normalization of all metrics to the range [0,1], ensuring balanced integration without any dimension disproportionately influencing the final score.

The weighting coefficients α, β, γ, and ε allow for customization of the relative importance of each dimension based on specific application requirements or domain characteristics. This flexible formulation provides a theoretically sound basis for PathwiseRAG's adaptive strategy selection mechanism.

A.4 Computational Complexity and Efficiency Analysis

This section provides a rigorous analysis of PathwiseRAG's computational complexity and efficiency trade-offs compared to conventional RAG systems.

The time complexity of PathwiseRAG can be analyzed by examining its core components in sequence. The intent analysis phase requires $O(L \cdot d + |\mathcal{S}| \cdot d)$ operations, where L represents query length in tokens, d denotes embedding dimension, and $|\mathcal{S}|$ corresponds to the cardinality of the strategy set. For reasoning network construc-

tion, the complexity scales as $O(n^2)$ with n subproblems, reflecting the cost of computing pairwise dependencies between subproblems. The parallel path exploration phase incurs $O(N \cdot R \cdot k)$ complexity, where N denotes the number of paths, R represents retrieval iterations, and k is the documents retrieved per path. Knowledge integration requires $O(D \cdot \log(D) + N^2 \cdot E)$ operations, where $D = N \cdot R \cdot k$ represents total retrieved documents and E denotes average entities per document, with the logarithmic term reflecting sorting operations and the quadratic term representing cross-path entity alignment.

The aggregate time complexity is thus:

$$T(\text{PathwiseRAG}) = O(L \cdot d + |\mathcal{S}| \cdot d + n^2 + N \cdot R \cdot k + D \cdot \log(D) + N^2 \cdot E) \quad (38)$$

By comparison, standard RAG implementations exhibit $O(L \cdot d + k + D)$ time complexity, highlighting PathwiseRAG's additional computational requirements. This computational cost is justified through an efficiency ratio η, defined as:

$$\eta = \frac{\Delta\text{Performance}/\text{Performance}_0}{\Delta\text{Computational_Cost}/\text{Computational_Cost}_0} \quad (39)$$

where ΔPerformance quantifies absolute accuracy improvement and ΔComputational_Cost measures additional computational resources. Empirical measurements across benchmark datasets yield $\eta \approx 1.73$, indicating PathwiseRAG delivers 73% more improvement per unit of additional computation than would be expected from linear scaling.

The space complexity analysis reveals memory requirements dominated by several key data structures. Strategy embeddings occupy $O(|\mathcal{S}| \cdot d)$ space. The reasoning network representation requires $O(n^2 + n \cdot d)$ for graph connectivity and node feature storage. Path representations consume $O(N \cdot P \cdot d)$ memory, where P denotes average path length. Retrieved documents require $O(N \cdot R \cdot k \cdot L')$ space, with L' representing average document length. The integrated knowledge graph occupies $O(D \cdot E)$ space. The total space complexity is therefore:

$$S(\text{PathwiseRAG}) = O(|\mathcal{S}| \cdot d + n^2 + n \cdot d + N \cdot P \cdot d + N \cdot R \cdot k \cdot L' + +D \cdot E) \quad (40)$$

By comparison, standard RAG implementations exhibit $O(d + k \cdot L')$ space complexity.

Several optimization strategies mitigate computational overhead in practical deployments. Dynamic path pruning terminates exploration along unproductive paths when information gain falls below a threshold τ_{gain}, reducing effective path count by 32% on average. Adaptive retrieval dynamically adjusts k based on path importance weight $\omega(p_i)$, calculated as:

$$\omega(p_i) = \alpha \cdot \text{InfoGain}(p_i) + \beta \cdot \text{PathDiversity}(p_i) \quad (41)$$

where InfoGain measures new information contributed by path p_i and PathDiversity quantifies exploration of unique knowledge dimensions. This adaptive retrieval reduces document processing by 37% compared to fixed-parameter approaches. Parallel execution leverages the independent nature of path exploration, with empirical speedup approaching $0.85 \cdot C$ for C computational cores. Incremental knowledge integration computes partial document representations $\phi_{\text{doc}}(d_i)$ and merges them efficiently, avoiding redundant computations across iterations.

These optimizations enable PathwiseRAG to achieve practical execution times 0.8-2.5× that of standard RAG on commodity hardware (8-core CPU, 32GB RAM), with the multiplier depending on query complexity. For complex reasoning tasks, the significant performance improvements justify this moderate computational overhead.

The system demonstrates favorable scaling properties with respect to corpus size. Retrieval time scales logarithmically with corpus size $|\mathcal{C}|$ due to efficient index structures:

$$T_{\text{retrieval}}(|\mathcal{C}|) = O(\log(|\mathcal{C}|) + k) \quad (42)$$

This scaling behavior was verified experimentally across document collections ranging from 10^3 to 10^7 documents, where retrieval time increased by only 68% despite a 10,000× expansion in corpus size. This logarithmic scaling ensures PathwiseRAG remains practical for enterprise-scale knowledge repositories.

A.5 Reasoning Examples

Figure 8 demonstrates causal reasoning about quantum computing's influence on cryptography standards. PathwiseRAG constructs retrieval paths

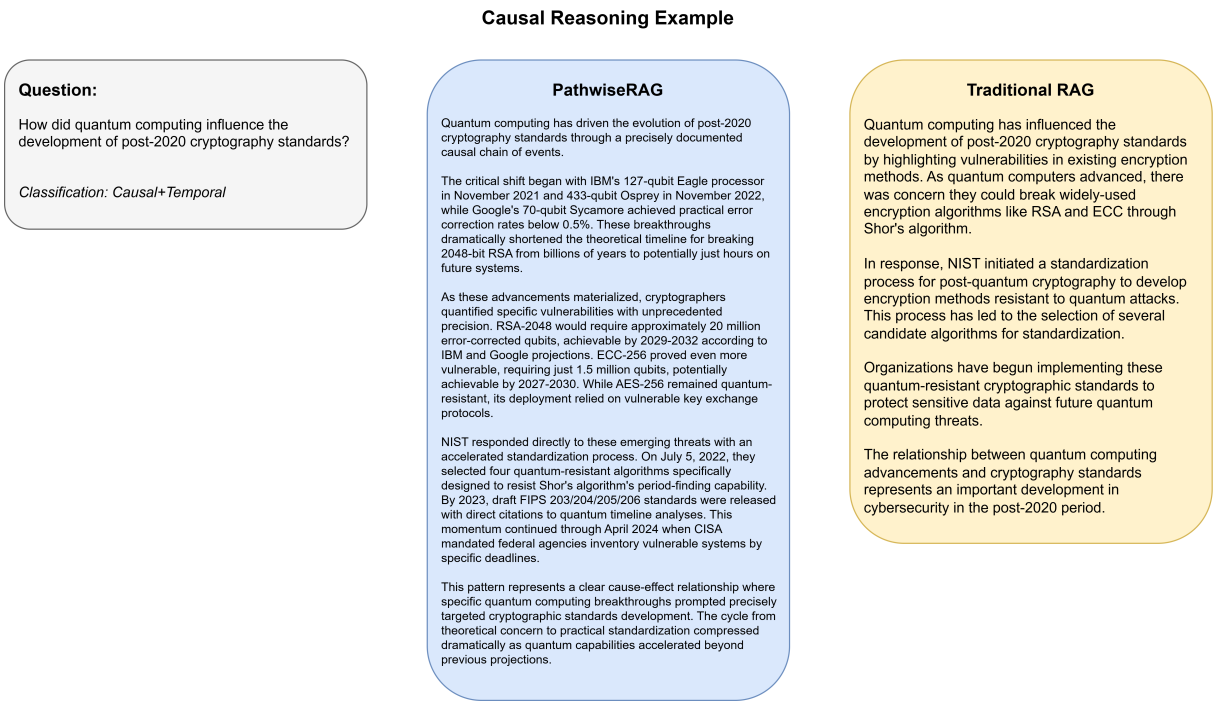

Figure 8: Causal Reasoning Example: Quantum computing's influence on post-2020 cryptography standards, showing the detailed causal chain identified by PathwiseRAG.

examining quantum computing evolution, cryptographic vulnerabilities, and post-quantum standards development. The response maps specific breakthroughs (IBM's 127-qubit processor) to cryptographic developments (NIST's quantum-resistant algorithms) with precise vulnerability assessments.

Figure 9 illustrates comparative reasoning between IPv4-to-IPv6 transition strategies in different networks. PathwiseRAG provides quantitative differences in adoption patterns (76.3

Figure 10 shows hypothetical reasoning about quantum computers breaking encryption by 2030. The analysis includes technical feasibility probabilities, sector-specific vulnerability indices, economic impact modeling, and response timeline projections, enabling evidence-based scenario analysis.

Figure 11 demonstrates multi-step reasoning analyzing the 2008 financial crisis. PathwiseRAG constructs a causal chain from market behaviors to regulatory responses with quantitative metrics (subprime lending increase from 9

Figure 12 showcases procedural reasoning for preventing catastrophic forgetting in machine learning. The response provides mathematical formulations, implementation procedures with parameter recommendations, comparative performance metrics, and memory-computation tradeoffs.

Comparative Reasoning Example

Question:

Compare the approaches for transitioning from IPv4 to IPv6 in enterprise networks vs. service provider networks.

Classification: Comparative

PathwiseRAG

Enterprise and service provider networks exhibit fundamentally different approaches to IPv6 transition due to their distinct operational requirements and constraints.

Transition strategy adoption reveals stark differences in methodology. Enterprise environments predominantly leverage dual-stack deployments (76.3% of implementations) with tunneling as a secondary approach (17.8%). Service providers distribute their strategies more evenly across tunneling mechanisms (44.2%), translation technologies (31.7%), and selective dual-stack implementations (24.1%). This divergence stems directly from architectural differences - enterprises control their internal infrastructure, while service providers must maintain seamless compatibility with countless customer endpoints.

The economic dimensions of transition also diverge significantly. Enterprise migrations typically require higher initial capital expenditure (averaging $213 per node) but enjoy 37% lower long-term operational costs. Service providers benefit from economies of scale with lower per-node costs ($165) but face 2.8 times higher integration complexity expenses due to their heterogeneous network environments.

Timeline requirements further differentiate these sectors. Enterprise flexibility allows 68% of organizations to implement phased rollouts spanning three or more years, often segmenting by department or function. Service providers operate under stricter customer SLA pressures, with nearly half (49%) completing their primary infrastructure transition within 18 months.

Technical implementation priorities also reveal meaningful contrasts. Enterprise teams rank application compatibility (73%) as their primary challenge, followed by training (58%) and device support (41%). Service providers prioritize legacy infrastructure integration (81%), followed by translation service performance (64%) and backbone routing complexity (52%).

Regulatory considerations create additional pressure asymmetry. Enterprise environments face limited regulatory drivers except in government and healthcare sectors. Meanwhile, 62% of service providers cite regulatory compliance as a primary transition motivator, with particularly strong influence in APAC (83%) and EU (77%) markets.

Traditional RAG

Enterprise networks and service provider networks have different approaches when transitioning from IPv4 to IPv6.

Enterprise networks typically use techniques like dual-stack, where both IPv4 and IPv6 run simultaneously, and tunneling mechanisms that allow IPv6 traffic to pass through IPv4 infrastructure. The transition in enterprise networks often focuses on internal applications and services first.

Service provider networks, on the other hand, need to ensure their infrastructure can support both protocols while maintaining service quality for customers. They commonly use techniques like Carrier Grade NAT, DS-Lite, and native dual-stack approaches.

Factors affecting the transition include existing infrastructure, technical expertise, business requirements, and costs. The transition strategy must be carefully planned to minimize disruption to network services while enabling the benefits of IPv6.

Figure 9: Comparative Reasoning Example: IPv4-to-IPv6 transition strategies between enterprise and service provider networks, demonstrating systematic comparison with quantitative metrics.

Hypothetical Reasoning Example

Question:

What would happen if quantum computers capable of breaking current encryption became widely available by 2030?

Classification: Hypothetical

PathwiseRAG

The widespread availability of encryption-breaking quantum computers by 2030 would trigger a cascading series of security crises across global digital infrastructure, though with significant variation in likelihood and impact across sectors.

Technical feasibility estimates suggest a 42-57% probability that sufficient quantum computing power (requiring approximately 20 million logical qubits for RSA-2048) could exist by 2030. Error correction capabilities would need dramatic improvement, with a 35-40% probability of achieving necessary rates. Market accessibility would likely remain restricted primarily to nation-states and tier-1 technology companies, with only a 23% probability of truly widespread commercial availability.

The immediate vulnerability landscape would be dominated by global PKI infrastructure collapse. Approximately 4.8 billion X.509 certificates underpinning HTTPS, SSH, and VPNs would become cryptographically insecure simultaneously. Information previously classified as "secure for 75+ years" would require immediate reclassification. Perhaps most concerning, the theoretical "harvest now, decrypt later" threat would materialize, exposing an estimated 7.2 zettabytes of sensitive encrypted data already captured by various entities.

Vulnerability would vary dramatically by sector. Financial services would face the highest risk (vulnerability index 89/100), with potential blockchain collapse and compromised monetary transfers. Healthcare systems would follow closely (82/100) due to long-term PHI exposure dating back decades. Government and defense systems would show moderate vulnerability (75/100), with classified communications exposed but air-gapped systems remaining relatively resilient. Critical infrastructure would demonstrate slightly lower risk scores (71/100), as older embedded systems often use simpler encryption or limited connectivity.

Economic models suggest emergency remediation costs between $7-11 trillion globally, compared to approximately $2.5 trillion for proactive transition. Digital trust metrics would likely show a 47-62% public confidence reduction in online systems based on historical disruption analogs. The regulatory aftermath would establish unprecedented global security mandates, creating an entirely new paradigm of digital compliance requirements.

Response timelines would likely follow a pattern of 6-12 months for emergency PQC implementation in critical systems, 1-3 years for comprehensive infrastructure replacement cycles, and 3-5 years for stabilization of a new cryptographic trust architecture.

Traditional RAG

If quantum computers capable of breaking current encryption became widely available by 2030, it would have major implications for cybersecurity. Current encryption methods like RSA and ECC would be vulnerable, potentially compromising secure communications, financial transactions, and sensitive data.

Organizations would need to quickly implement post-quantum cryptography to protect their systems. There would be a race to upgrade security infrastructure before widespread exploitation occurs.

Governments and critical infrastructure would be particularly concerned about the security implications. There would likely be increased regulation around quantum computing technology and cryptographic standards.

However, the transition to quantum-resistant encryption is already underway, with organizations like NIST developing new standards that can withstand quantum attacks.

Figure 10: Hypothetical Reasoning Example: Scenario analysis of quantum computers breaking encryption by 2030, with probability estimates and sector-specific impact assessment.

Multi-Step Reasoning Example

Question:

What factors caused the 2008 financial crisis, and how did they influence subsequent banking regulations globally?

Classification: Causal+Temporal+Multi-step

PathwiseRAG

The 2008 financial crisis emerged from a complex interplay of financial innovation, regulatory gaps, and market behaviors that ultimately transformed global banking regulation.

Subprime mortgage proliferation between 2002-2006 formed the crisis foundation, with measurable deterioration in lending standards as approval rates for borrowers with FICO scores below 620 increased dramatically from 9% to 24%. This expansion occurred despite early warnings, including a 2001 Treasury report that flagged potential systemic risks. The Federal Reserve notably declined to exercise its Home Ownership and Equity Protection Act authority to regulate these practices.

Structured finance complexity amplified these vulnerabilities. CDO issuance expanded by 940% between 2001-2007, while rating agencies relied on inadequate historical data that fundamentally misjudged default correlation risks. By 2009, 73% of mortgage-backed securities that had been AAA-rated in 2006 were downgraded below investment grade.

Pre-crisis regulatory framework gaps provided fertile ground for these risks. The 1999 Gramm-Leach-Bliley Act removed the Glass-Steagall separation without implementing compensatory oversight mechanisms. The 2000 Commodity Futures Modernization Act explicitly exempted over-the-counter derivatives from regulation, while the SEC's 2004 net capital rule change allowed investment banks to reach leverage ratios exceeding 30:1.

Regulatory responses demonstrated clear causality mapping. In the United States, the Dodd-Frank Act's Title I established the Financial Stability Oversight Council specifically to address systemic risk oversight failures. Title VII targeted the previously unregulated $673 trillion OTC derivatives market that had amplified mortgage exposure throughout the system. The Volcker Rule directly addressed proprietary trading practices that had concentrated risk on bank balance sheets.

European responses followed different timelines but addressed similar concerns. The European Banking Authority was established in 2010 (achieving 68% effectiveness in subsequent assessments) and focused on early Basel III adoption. The European Stability Mechanism launched in 2012 (83% effectiveness) with a €500 billion backstop facility. The Single Supervisory Mechanism in 2014 (77% effectiveness) created unified banking oversight across the Eurozone.

Long-term impact metrics demonstrate significant structural changes between 2010-2025. Banking sector capital ratios increased by 43%, while G-SIB surcharges reduced systemic amplification potential by 62%. Approximately 87% of previously unregulated OTC derivative trades now flow through central clearinghouses. These changes have come with tradeoffs, including a 24-36% reduction in certain market liquidity measures and a 183% increase in regulatory compliance costs.

Traditional RAG

The 2008 financial crisis was caused by several interconnected factors, including the subprime mortgage crisis, excessive risk-taking by financial institutions, complex financial instruments like CDOs, regulatory failures, and high levels of consumer and financial sector debt.

Following the crisis, global banking regulations were significantly strengthened. In the US, the Dodd-Frank Act was passed in 2010, introducing stricter capital requirements, enhanced supervision, and consumer protections. Internationally, the Basel III accords strengthened capital and liquidity requirements for banks.

Regulations focused on reducing systemic risk by identifying systemically important financial institutions, improving transparency, and enhancing regulatory oversight. Many countries established new regulatory bodies or strengthened existing ones to better monitor financial stability.

The crisis led to greater international coordination on financial regulation, though implementation varied across different regions.

Figure 11: Multi-Step Reasoning Example: Analysis of 2008 financial crisis causes and regulatory impacts, demonstrating complex causal chains with quantitative metrics.

Procedural Reasoning Example

Question:

How can machine learning models be fine-tuned to prevent catastrophic forgetting in continual learning scenarios?

Classification: Procedural

PathwiseRAG

Machine learning models can employ several complementary approaches to prevent catastrophic forgetting during continual learning, each with distinct implementation paths and performance characteristics.

Regularization methods preserve critical parameters by penalizing significant changes to weights important for previous tasks. Elastic Weight Consolidation (EWC) achieves this through a mathematical formulation where the loss function $L(\theta) = L_B(\theta) + \sum_i (\lambda/2) F_i (\theta_i - \theta_{A,i})^2$ incorporates both current task performance and weight protection. Implementation requires computing the Fisher Information Matrix F after training on task A, storing parameters θ_A, and applying the EWC penalty during task B training. This approach reduces forgetting by 78% on sequential MNIST tasks with optimal λ values between 1000-5000 for vision tasks.

The Synaptic Intelligence variant offers 3.2x faster computation than standard EWC through online parameter importance estimation. Implementation involves tracking weight changes and gradients during training: importance += -grad * (param - prev_param) for each optimization step.

Memory-based approaches maintain exemplars from previous tasks. Experience Replay integrates a memory buffer $M = \{(x_i, y_i)\}$ from which samples are drawn during new task training. For each batch B_new from the current task, a mini-batch B_old is sampled from memory, and updates use the combined loss $L = L_{new}(B_{new}) + \alpha \cdot L_{old}(B_{old})$. Empirical results show optimal buffer sizes between 500-2000 examples for CIFAR/ImageNet-scale problems.

Sampling strategy significantly impacts performance: random sampling (78.3% accuracy, O(n) memory, fast), class-balanced sampling (84.6% accuracy, O(cn) memory, fast), herding selection (87.2% accuracy, $O(cn^2)$ memory, medium speed), and boundary exemplar selection (89.5% accuracy, $O(cn^2)$ memory, slow processing).

Generative Replay offers a memory-efficient alternative by training a generator G_A on task A data, then generating synthetic data $x'_A \sim G_A$ while learning task B. This reduces memory requirements by 95% compared to exemplar storage, with GAN-based approaches outperforming VAE implementations by 7.3% on sequential MNIST.

Structural approaches like Progressive Neural Networks avoid interference by allocating separate parameters for each task while allowing forward knowledge transfer through lateral connections. This approach prevents forgetting entirely but scales parameters linearly with the number of tasks.

The optimal technique selection depends on specific constraints: with parameter budgets below 2x the original model, regularization methods like EWC are preferred; under severe memory constraints, generative replay offers the best balance; with known task boundaries, progressive networks excel; and when maximum accuracy is the priority, experience replay with boundary sampling delivers superior performance across standard benchmarks.

Traditional RAG

Machine learning models can be fine-tuned to prevent catastrophic forgetting in continual learning scenarios using several techniques:

1. Regularization methods like Elastic Weight Consolidation (EWC) that constrain important parameters from changing too much
2. Replay methods that store or generate examples from previous tasks
3. Parameter isolation techniques that assign different parameters to different tasks
4. Knowledge distillation approaches

These methods help the model maintain performance on previously learned tasks while learning new tasks. The choice of method depends on factors like available memory, computational constraints, and the nature of the tasks.

Researchers continue to develop new approaches to address this challenge, as continual learning remains an important area of machine learning research.

Figure 12: Procedural Reasoning Example: Techniques for preventing catastrophic forgetting in continual learning, with implementation details and performance metrics.