

ITC 1/55 Information Technology and Control Vol. 55 / No. 1 / 2026 pp. 298-322 DOI 10.5755/j01.itc.55.1.42562	CogniRAG: Integrating Causal Hyperedges and Counterfactual Reasoning for Knowledge-Intensive Tasks	
	Received 2025/08/15	Accepted after revision 2025/12/26
	HOW TO CITE: He, L. (2026). CogniRAG: Integrating Causal Hyperedges and Counterfactual Reasoning for Knowledge-Intensive Tasks. <i>Information Technology and Control</i> , 55(1), 298-322. https://doi.org/10.5755/j01.itc.55.1.42562	

CogniRAG: Integrating Causal Hyperedges and Counterfactual Reasoning for Knowledge-Intensive Tasks

Luoyao He*

School of Public Health, Imperial College London, United Kingdom

Corresponding author: luoyaoh173@gmail.com

Retrieval-augmented generation (RAG) enhances large language models (LLMs) by integrating external knowledge sources. Yet, existing chunk- or graph-based frameworks remain confined to surface-level semantic correlations, lacking the capacity to model mechanistic causality, cross-document dependencies, and counterfactual reasoning. To address these limitations, this study introduces CogniRAG, a causal reasoning-centric RAG framework that systematically transforms retrieval into a structured process of causal inference. CogniRAG encodes four specialized causal hyperedge types—Linear Chain, Feedback Loop, Intervention Point, and System Stability—within a unified CogniGraphDB, enabling multi-entity causal tracing and intervention analysis. A dual retrieval strategy combining entity diffusion and relation expansion constructs causally enriched prompts that enhance inferential depth beyond conventional similarity-based retrieval. Experiments conducted across four knowledge-intensive domains (Medicine, Political Science, Computer Science, Finance) using an identical LLaMA-3.1-70B backbone and 480 systematically generated questions demonstrate consistent and statistically robust improvements. CogniRAG achieves an approximately five-percentage-point Overall gain over HyperRAG across all domains (for example, Medicine 78.67 vs 73.63), with the largest increases observed in Empowerment and Logical Coherence, corroborated by objective reliability metrics (+1.8 pp Faithfulness, +1.2 pp Factual Consistency, -0.7 pp Hallucination). Cross-backbone evaluation with Qwen-Plus and Mistral Large 2 confirms backbone-agnostic consistency (+3–5 pp), while efficiency analysis reveals moderate computational overheads (indexing +30–40%; latency +50%) proportional to added causal reasoning depth. Overall, CogniRAG advances RAG from semantic retrieval toward causal, counterfactual, and interpretable knowledge reasoning, with future work focusing on real-time causal-graph updating, multimodal evidence integration, reproducible benchmarking, and meta-reasoning for confidence calibration to strengthen scalability and external validity.

KEYWORDS: Large Language Models, Retrieval-Augmented Generation, Knowledge Graphs, Casual Reasoning

1. Overview

1.1. Limitations of Traditional RAG Approaches

Large Language Models (LLMs) have substantially reshaped natural language processing with their remarkable capabilities in text generation, comprehension, and reasoning across diverse domains [3, 14]. However, these powerful models face fundamental limitations, including hallucination, outdated knowledge, and non-transparent reasoning processes that constrain their reliability in knowledge-intensive applications [10, 21]. Retrieval-Augmented Generation (RAG) has emerged as a promising solution to address these challenges by integrating external knowledge retrieval with generative capabilities, enabling LLMs to access dynamic, up-to-date information while maintaining their generative strengths [12]. Traditional RAG frameworks predominantly rely on chunk-based retrieval mechanisms where text is segmented into fixed-length passages and retrieved via dense vector matching. This approach exhibits three fundamental deficiencies: first, it fails to capture complex relationships between entities, treating each text segment as an isolated unit without modeling the interconnections that constitute expert knowledge networks [17]; second, it struggles with queries requiring comprehensive understanding across multiple information sources, as semantic similarity matching cannot bridge knowledge discontinuities when relevant information exists across disconnected text segments; third, it cannot perform causal reasoning or counterfactual analysis, as vector-based retrieval systems lack the structured representations necessary to trace causal pathways, identify feedback mechanisms, or evaluate alternative scenarios [21]. These limitations become particularly pronounced when dealing with questions that demand connecting disparate pieces of information, understanding temporal dependencies, or performing holistic analysis over large document collections [2].

1.2. From Graph-Based to Causal Reasoning

The recognition of these limitations has driven the development of knowledge graph-enhanced retrieval systems, where LLMs construct structured knowledge representations from unstructured text to enable sophisticated reasoning beyond seman-

tic similarity matching [1]. Microsoft's GraphRAG exemplifies this paradigm shift by automatically extracting entity knowledge graphs and generating community summaries that facilitate global reasoning over complete datasets, demonstrating significant improvements in questions requiring cross-document information integration [11]. Building upon this foundation, HyperGraphRAG introduces hypergraph representations through n-ary relational modeling that encodes complex multi-entity relationships via hyperedges, achieving 27% reductions in hallucination rates and notable improvements in multi-hop reasoning tasks compared to document-based RAG systems [20]. However, these graph-based approaches remain fundamentally constrained in capturing mechanistic causal relationships. Even CausalRAG explicitly incorporates causal reasoning by constructing causal graphs where nodes represent entities and edges encode causal dependencies, preserving logical coherence across document boundaries [25]. During retrieval, the framework employs causal path tracing to identify causally relevant nodes rather than relying solely on semantic similarity. The system generates causal summaries by traversing key causal paths, prioritizing relationships that directly contribute to answering queries, which are then combined with user queries to construct refined prompts enabling causal reasoning rather than simple text aggregation [25].

1.3. Fundamental Gaps in Current Causal Reasoning Approaches

Despite recent advances in knowledge graph-enhanced retrieval systems, existing approaches exhibit fundamental architectural deficiencies that preclude sophisticated causal understanding. Graph-based systems such as GraphRAG and hypergraph-based frameworks like HyperGraphRAG [20], while achieving 27% reductions in hallucination rates through multi-entity hyperedge representations that connect multiple entities simultaneously, fundamentally lack causal reasoning capabilities: they cannot encode temporal precedence or sequential ordering (treating "A occurs before B" identically to "A and B co-occur"), cannot distinguish

causation from correlation (unable to differentiate “smoking causes cancer” from “ice cream sales correlate with drowning incidents”), and cannot represent feedback loops with delay factors (missing reinforcing cycles where “population growth → resource consumption → environmental degradation → population decline”). CausalRAG attempts to address these limitations by constructing causal graphs where nodes represent entities and directed edges encode causal dependencies with explicit cause-effect directionality, employing causal path tracing during retrieval to traverse dependency chains (e.g., $A \rightarrow B \rightarrow C$) rather than semantic similarity matching, and generating causal summaries that aggregate causally relevant information for prompt construction [25]. However, CausalRAG remains constrained by three critical limitations: (1) binary causal structure—modeling only pairwise cause-effect relationships ($A \rightarrow B$), unable to represent scenarios where multiple simultaneous causes jointly produce effects ($A+B+C \rightarrow D$); (2) path-based reasoning—linear causal path expansion that captures sequential chains but fails to model feedback mechanisms where effects recursively influence their causes ($A \rightarrow B \rightarrow C \rightarrow A$) or identify critical tipping points where small parameter changes trigger qualitative system transitions; (3) no uncertainty quantification—providing deterministic causal assertions without confidence levels, robustness testing across contextual variations, or probabilistic alternative history analysis. These deficiencies prevent addressing queries such as “If healthcare funding were increased by 20%, how would the resulting improved outcomes feedback to reduce long-term costs, and with what confidence level given historical policy variations?”—questions requiring comprehensive causal reasoning with temporal dynamics, multi-entity interactions, and uncertainty quantification essential for expert-level analytical reasoning in knowledge-intensive domains.

1.4. The CogniRAG Framework and Main Contributions

To address the identified gaps in graph-based and causal RAG systems, CogniRAG is proposed, a novel framework that introduces specialized causal reasoning capabilities through comprehensive causal hyperedge architecture. This research makes three principles.

- 1 Specialized Causal Hyperedge Architecture for Temporal and Mechanistic Reasoning.** To overcome the inability of existing graph-based systems to model temporal precedence, feedback dynamics, and intervention strategies, we introduce four specialized causal hyperedge types: Linear Chain hyperedges capture sequential causal pathways with explicit temporal ordering and amplification effect quantification (addressing the static association limitation); Feedback Loop hyperedges model reinforcing and balancing cycles with delay factors and saturation points (overcoming the inability to represent cyclical causal dynamics where effects recursively influence their causes); Intervention Point hyperedges identify strategic system modification opportunities with implementation difficulty scoring and success probability evaluation (enabling leverage point analysis absent in current systems); and System Stability hyperedges analyze tipping points and resilience mechanisms (addressing the lack of qualitative transition detection). This architecture directly resolves the five critical limitations identified in GraphRAG and HyperGraphRAG by transforming static multi-entity associations into dynamic causal structures with mechanistic explanations.
- 2 CogniGraph Construction with Advanced Causal Reasoning Modules.** To address CausalRAG’s constraints of binary relationships, path-based reasoning, and absence of uncertainty quantification, we develop a comprehensive knowledge representation that integrates low-order causal correlations (entity pairs with causal directionality) with high-order causal chain formations (multi-entity configurations where $A+B+C \rightarrow D$). Beyond structural representation, we implement three advanced reasoning modules that enable capabilities absent in existing systems: Counterfactual Scenario Generation systematically explores alternative outcomes with explicit confidence levels and assumption tracking (answering “what-if” queries like “If intervention X were implemented, what effects would emerge?”); Causal Robustness Testing validates relationship reliability across varying contextual conditions (addressing deterministic assertion limitations); and Alternative History Analysis examines how different decision points propagate through causal networks with probability assessments (enabling complex rea-

soning about feedback amplification and dampening effects with quantified confidence).

3 A Dual-Database Architecture for Causally Grounded Retrieval. To overcome semantic similarity over-reliance and contextual disruption from text chunking, we design a comprehensive processing pipeline integrating specialized CogniGraphDB for structured causal storage with multi-matrix VectorDB construction (entities, relationships, chunks). This architecture enables two retrieval strategies that address existing limitations: entity-oriented retrieval performs semantic matching followed by structural diffusion through causal chains (bridging knowledge discontinuities across disconnected text segments); relationship-oriented retrieval identifies causal patterns matching query requirements followed by bidirectional expansion to participating entities (retrieving causally relevant information rather than semantically similar but causally irrelevant content). The context integration module synthesizes heterogeneous knowledge sources—ranked documents, causally-related entities, and multi-layer CogniGraph structures—into causally-enriched prompts that enable LLMs to reason through mechanistic pathways rather than surface-level associations.

The remainder of this paper is organized as follows: Section 2 reviews related work in retrieval-augmented generation and causal reasoning frameworks. Section 3 presents the CogniRAG methodology, including causal chain formation, CogniGraph construction, and the system processing pipeline. Section 4 describes the experimental setup, performance analysis across four knowledge-intensive domains, and case study demonstrations. Section 5 concludes with key findings and future research directions.

2. Related Work

Unifying knowledge representation with structured knowledge and retrieval-augmented generation has transformed information retrieval paradigms from simple semantic similarity matching to sophisticated graph-based reasoning systems. Microsoft's GraphRAG is an paradigm-shifting example that creates LLM-generating knowledge graphs that enable query-centric summarisation and reasoning

across entire text collections globally, overcoming fundamental limitations where current methods fail on questions that require exhaustive knowledge comprehension of a dataset [5]. Building on related graph-based ideas, Guo et al. invented LightRAG as an innovative dual-level retrieval framework that combines graph structures with vector representations to supply efficient retrieval with contextual relevance, albeit through a streamlined graph-based paradigm of entity-relationship extraction, LLM profiling for key-value pair extraction, and deduplication processes that avoid computational overhead through only a single API invocation during retrieval as opposed to hundreds of sequential community explorations in GraphRAG [9]. Contemporary developments have introduced advanced retrieval optimization techniques, including reranking mechanisms that utilize models like BGE-M3 developed by Chen et al and JaColBERT proposed by Yoshioka and Kato for tensor-based reranking systems, which significantly enhance retrieval precision by reordering candidate passages based on fine-grained relevance scores rather than initial similarity matching [4].

Beyond information retrieval systems, machine learning techniques have demonstrated significant potential in domain-specific anomaly detection and classification tasks. A representative study employed supervised machine learning approaches combining Local Binary Pattern (LBP), Histogram of Gradient (HOG), and Haralick texture features with K-Nearest Neighbor and Support Vector Machine classifiers for COVID-19 detection from chest X-ray images, achieving diagnostic accuracy ranging from 89.2% to 98.66% {Hasoon, 2021 #447}. While such supervised classification methods excel at pattern recognition within well-defined feature spaces, they fundamentally differ from knowledge graph-based retrieval systems in their inability to capture complex relational semantics, perform multi-hop reasoning across interconnected entities, or leverage structured knowledge representations for explainable inference—limitations that motivate the evolution toward graph-based and hypergraph-based knowledge reasoning frameworks explored in subsequent sections {Hasoon, 2021 #447}.

The evolution toward hypergraph-based knowledge representation marks a significant advancement beyond binary relational constraints inherent in traditional knowledge graphs, with Luo et al.'s Hy-

perGraphRAG pioneering the modeling of n-ary relational facts through hyperedges that capture complex multi-entity relationships prevalent in real-world scenarios [20]. This hypergraph approach demonstrates substantial performance improvements, achieving 15.4% higher accuracy on multi-hop reasoning tasks and reducing hallucination rates by 27% compared to document-based RAG systems, while requiring 32% fewer tokens to represent equivalent knowledge structures [20]. Parallel developments such as Zhao et al.'s Hyper-RAG have focused specifically on combating LLM hallucinations through hypergraph-driven approaches that comprehensively capture both pairwise and beyond-pairwise correlations in domain-specific knowledge, showing average accuracy improvements of 12.3% over direct LLM usage and maintaining stable performance with increasing query complexity [7].

The integration of causal reasoning capabilities with knowledge graph-based retrieval systems represents the frontier of mechanistic understanding in information retrieval [25]. Wang et al.'s CausalRAG exemplifies this evolution by incorporating causal graphs into the retrieval process to address traditional RAG limitations including disrupted contextual integrity due to text chunking and over-reliance on semantic similarity for retrieval. By constructing and tracing causal relationships, CausalRAG preserves contextual continuity and improves retrieval precision, leading to more accurate and interpretable responses. The framework demonstrates superiority across several metrics including answer faithfulness, context recall, and context precision compared to regular RAG and graph-based RAG approaches on knowledge-intensive tasks [25]. However, CausalRAG's approach remains fundamentally limited to path-based causal expansion and binary causal node relationships, lacking the comprehensive framework necessary to model complex causal dynamics, temporal dependencies, and system-level mechanisms that characterise sophisticated reasoning scenarios [25].

Complementing these graph-based and causal approaches, recent frameworks have pursued automated optimization and iterative refinement strategies to enhance RAG performance. Kim et al. proposed AutoRAG as a systematic optimization framework that automatically identifies optimal RAG configurations through exhaustive evaluation of module com-

binations across query expansion, retrieval, passage augmentation, and reranking stages, employing performance metrics and statistical measures to select superior modules at each node—demonstrating that dataset-specific optimization can yield substantial improvements over fixed pipeline architectures [15]. More recently, Wang et al. introduced CoRAG (Chain-of-Retrieval Augmented Generation), which transcends single-step retrieval limitations through iterative query refinement, where models dynamically reformulate sub-queries based on evolving reasoning states and utilize rejection sampling to augment datasets with intermediate retrieval chains, achieving 15-20% performance gains on multi-hop reasoning tasks like MuSiQue while supporting flexible decoding strategies including greedy decoding, best-of-N sampling, and tree search to balance computational efficiency with answer quality [26]. While these optimization-centric and iterative approaches demonstrate measurable improvements in retrieval precision and multi-step reasoning, they remain fundamentally constrained by their dependence on sequential processing patterns and lack the comprehensive framework necessary for modeling complex causal dynamics, temporal dependencies, and system-level mechanisms that characterize sophisticated mechanistic reasoning scenarios [24].

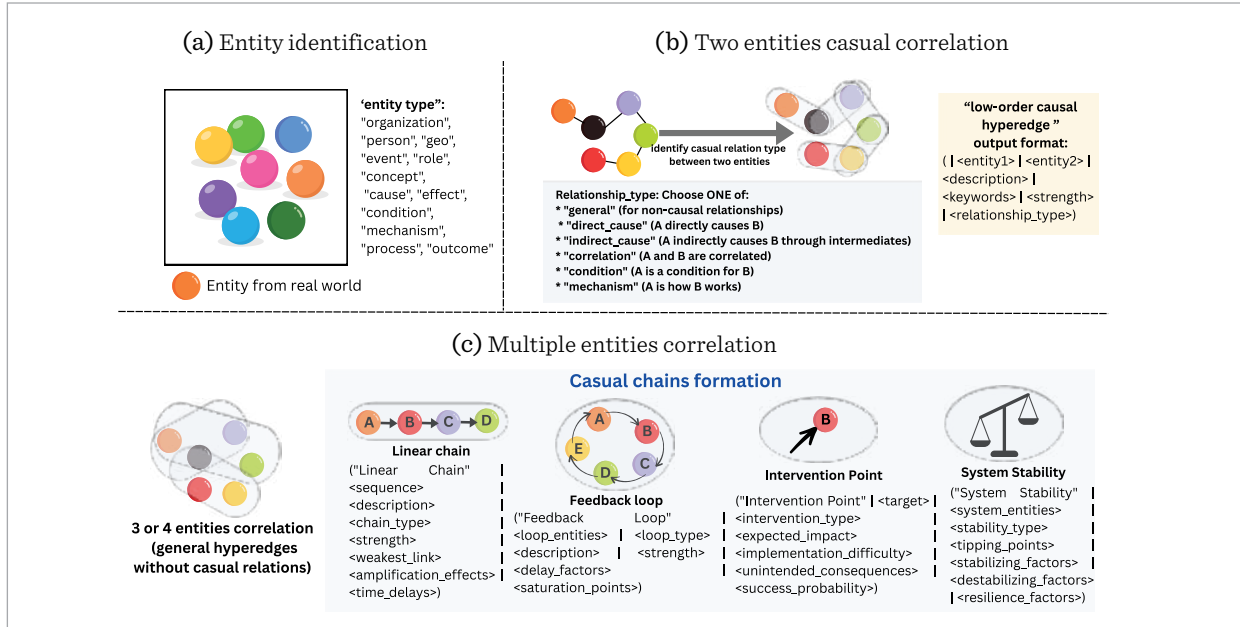
3. Method

3.1. Causal Chain Formation

The CogniGraph construction process, as shown in Figure 1, begins with entity identification (Panel A), where the system extracts diverse entity types from real-world text using an expanded taxonomy that goes beyond traditional named entity recognition [18]. The extraction process identifies nine distinct entity categories. Each entity receives detailed annotations including its primary type of classification, descriptive characteristics, and critically, causal role designations that specify whether the entity functions as a cause, effect, mediator, or conditional factor within causal relationships. This granular classification enables the system to understand not just what entities exist in the text, but how they potentially interact within causal frameworks, establishing the foundation for subsequent relationship analysis.

Figure 1

Illustration of CogniGraph Construction and Analysis Framework.



The two-entity causal correlation analysis (Panel B) represents a sophisticated advancement over traditional relationship extraction by implementing a nuanced taxonomy that distinguishes between correlational and causal relationships with explicit strength quantification. The system employs six relationship categories as shown in Table 1.

Each relationship receives a comprehensive characterization including entity pairs, detailed descriptions of the relationship dynamics, extracted keywords capturing the essence of the connection, strength assessments on a 1-10 scale indicating re-

lationship robustness, and explicit relationship type classification. The low-order causal hyperedge output format systematically captures these elements: ("Low-order Hyperedge" | <entity1> | <entity2> | <description> | <keywords> | <strength> | <relationship_type>), enabling precise documentation of how entities causally influence each other.

3.1.1. From General Hyperedges to Specialized Causal Structure

Based on the foundation of general hyperedges, specialized causal hyperedges are introduced to achieve

Table 1

Two-Entity Causal Correlation Analysis: Six-Category Relationship Taxonomy.

Relationship Category	Description
General	Non-causal associations where entities are related but without directional influence
Direct_cause	Immediate causal relationships where A directly produces B
Indirect_cause	Mediated causal pathways where A influences B through intermediate steps
Correlation	Statistical associations without established causal direction
Condition	Prerequisite relationships where A must exist for B to occur
Mechanism	Explanatory relationships that describe how causal processes operate

deeper causal reasoning analysis illustrated in Figure 1 Panel C. While traditional general hyperedges can capture basic associations among multiple entities, they exhibit significant limitations in causal reasoning scenarios: they lack temporal information expression and cannot distinguish the temporal precedence in causal relationships; they ignore the essence of causal mechanisms, only describing "correlation" rather than "causation"; they adopt static representations that cannot model dynamic system behaviors, feedback effects, and system evolution processes; and they have intervention blind spots, failing to identify key intervention points and leverage effects in systems.

Therefore, four types of specialized causal hyperedges specifically demonstrated below are designed based on the enhanced causal reasoning prompt templates. While this section provides detailed textual descriptions of their functional characteristics and operational mechanisms, the formal mathematical and symbolic representations for each specialized and general hyperedge type are rigorously presented in Section 3.2.2, Equations (5)-(8), where they are expressed as precise mathematical formulas within the complete system pipeline diagram. Linear Chain hyperedges capture sequential causal pathways with temporal ordering, chain strength assessments (1-10), weakest link identification, amplification effects mapping, and time delays between causal steps; Feedback Loop hyperedges model reinforcing and balancing cycles with loop strength quantification, delay factors, and saturation points where feedback effects diminish; Intervention Point hyperedges identify strategic leverage points for system modification through intervention type classification (leverage_point, policy_change, structural_change, behavioral_change), implementation difficulty scoring (1-10), unintended consequence prediction, and success probability assessment to enable targeted interventions with confidence levels and assumption tracking; and System Stability hyperedges analyze stability types (stable, unstable, metastable, chaotic), tipping point identification, stabilizing and destabilizing factors, and resilience mechanisms that determine system recovery capacity, incorporating boundary conditions and sensitivity factors essential for understanding system vulnerabilities and designing protective measures against systemic failures.

3.1.2. Low-Order Hyperedges as Causal Chain Building Blocks

The low order hyperedge relationships identified in Figure 1 Panel B serve as the foundational building blocks for constructing these advanced causal chain formations. Each relationship type from Table 1 provides essential information for specialized causal chain assembly: direct_cause relationships form the backbone of linear chains with temporal precedence indicators and primary connections in feedback loops; mechanism relationships explain the "how" component of causal processes through operational descriptions and serve as strategic intervention targets with implementation assessments; condition relationships define boundary conditions, prerequisite requirements, and system constraints that affect chain stability and robustness; indirect_cause relationships help identify mediating variables, intermediate steps, and multi-step causal pathways through complex systems. The relationship strength scores (1-10) determine the robustness of causal chains, with higher-strength connections (8-10) forming more reliable causal inferences and lower scores (1-4) indicating weak or uncertain causal links requiring additional validation. The temporal information (before, after, simultaneous) embedded in entity descriptions enables proper chronological sequencing in linear chains, while causal role annotations (primary_cause, intermediate_cause, ultimate_effect) facilitate accurate identification of system dynamics, feedback mechanisms, and intervention leverage points. This comprehensive foundation ultimately enables the construction of sophisticated causal models that support counterfactual scenario generation with confidence levels, causal robustness testing across varying conditions, and alternative history analysis with probability assessments, as detailed in the advanced reasoning modules described in Sections 3.1.3 and 3.1.4.

After entity and correlation extraction through an LLM, enabling the CogniRAG system to construct CogniGraphs containing low order hyperedges that capture direct causal relationships (direct cause, mechanism, condition) and specialised causal chains, including linear sequences, feedback loops, intervention points, and system stability assessments. However, the formation of these causal chains represents only the foundational layer of

causal understanding. The system subsequently implements three critical advanced reasoning modules to unlock deeper analytical insights: Counterfactual Scenario generation systematically explores alternative outcomes by modifying specific links within established causal chains, enabling “what-if” analysis that reveals how different conditions or decisions could have led to entirely different system states; Causal Robustness Testing rigorously validates the reliability and consistency of identified causal relationships across varying contextual conditions, environmental factors, and temporal boundaries to ensure that discovered causal patterns maintain their validity under different scenarios; and Alternative History Analysis examines how alternative initial conditions, decision points, or intervention strategies could propagate through the causal network to produce fundamentally different system trajectories and outcomes. The integration of reranking and context augmentation mechanisms ensures optimal information synthesis for causal reasoning tasks.

These three advanced reasoning modules follow standardized output formats as specified in Table 2 with extraction prompt shown in Appendix Figure A4, which defines the structural requirements for each analysis type. The Counterfactual Scenario format captures essential elements for systematic “what-if” analysis, where the original scenario serves as baseline, the counterfactual condition specifies

the hypothetical change, and the predicted outcome articulates expected results. For instance, the financial influence elimination example starts with the current reality of monetary advantages in elections, hypothetically removes all financial influence through public financing reform, and predicts greater democratic equality as the outcome, with the reasoning process explaining how public financing would restore electoral balance by eliminating wealthy donor advantages, while the confidence level (6/10) and assumptions about successful implementation provide critical reliability context. The Robustness Test format enables systematic validation across varying conditions, as illustrated by the patronage legitimacy example which tests how patron-client relationships perform under different economic scenarios—revealing that while these relationships appear robust during periods of resource abundance, they become fragile and break down when economic stress reduces available resources below sustainable thresholds, with resource availability identified as the key sensitivity factor determining relationship durability.

The Alternative History format explores how different historical decision points could produce divergent outcomes, exemplified by the post-colonial state formation analysis that examines what would have happened if newly independent states had chosen institution-building over patronage-building approaches, tracing the cascading effects through

Table 2

Causal Reasoning Format Specifications.

Format Type	Structure & example
Counterfactual Scenario	<p>(“Counterfactual Scenario” <original_scenario> <counterfactual_condition> <predicted_outcome> <reasoning_process> <confidence_level> <assumptions>)</p> <p>E.g. (“Counterfactual Scenario” Current financial influence in elections Complete elimination of financial influence Greater democratic equality Public financing would restore balance 6 Assumes successful reform implementation)</p>
Robustness Test	<p>(“Robustness Test” <relationship> <variation_scenarios> <robustness_assessment> <boundary_conditions> <sensitivity_factors>)</p> <p>E.g. (“Robustness Test” Patronage legitimacy Economic crisis scenarios Fragile under stress, robust with resources Breaks down without sufficient resources Resource availability is key sensitivity factor)</p>
Alternative History	<p>(“Alternative History” <decision_point> <alternative_choices> <cascading_effects> <final_outcomes> <probability_assessment>)</p> <p>E.g. (“Alternative History” Post-colonial state formation decisions Institution-building vs patronage-building approach Stronger bureaucratic capacity → improved service delivery → higher citizen trust → stable governance More effective states with rule of law and reduced corruption 6)</p>

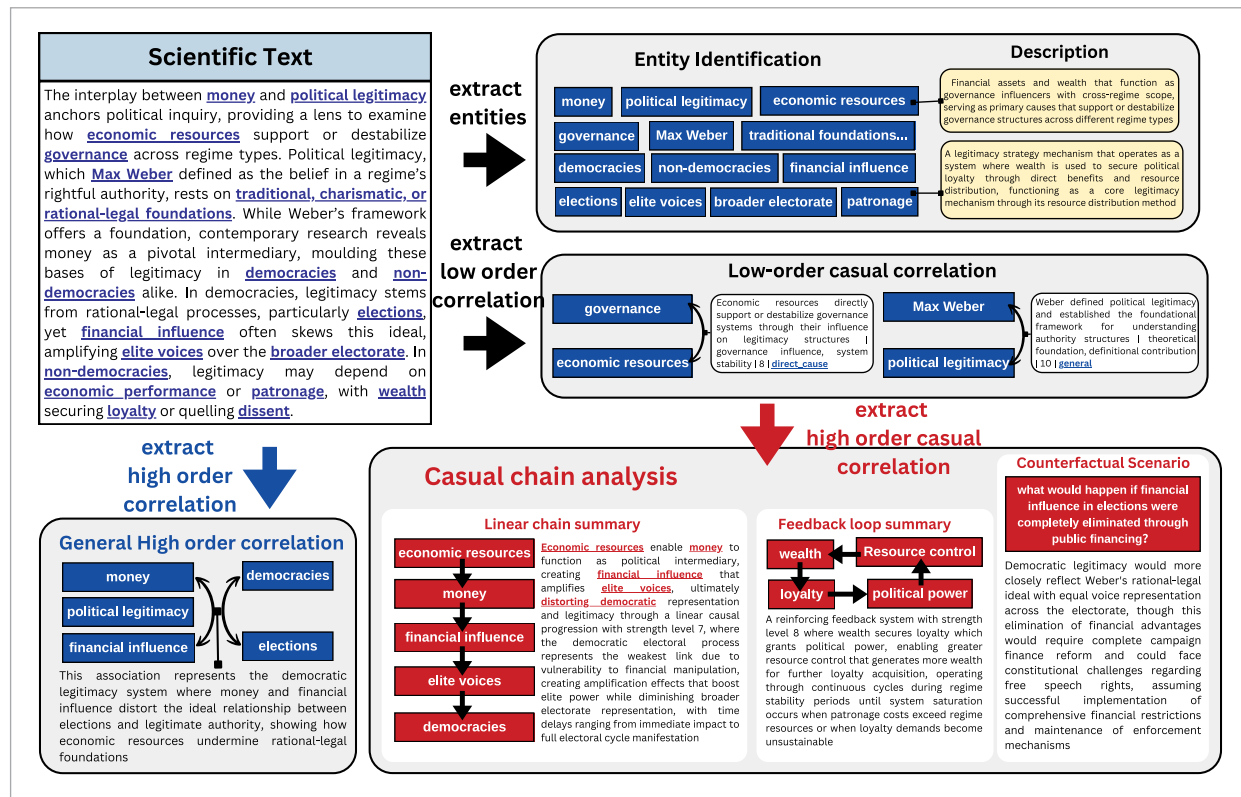
stronger bureaucratic capacity leading to improved service delivery, which generates higher citizen trust and ultimately produces more stable governance with reduced corruption, assessed with moderate confidence (6/10) in this alternative developmental trajectory. The integration of these standardized formats ensures CogniRAG produces structured, actionable insights that extend beyond traditional retrieval-augmented generation by incorporating confidence levels, assumptions, and sensitivity factors.

As visualized in Figure 2, the transition from general hyperedges to specialized causal chains represents a fundamental advancement in knowledge representation complexity and analytical capability. While general hyperedges in traditional systems capture static multi-entity associations without temporal or mechanistic information, the causal chain analysis constructs dynamic, causally structured knowledge networks that reveal system behavior patterns. The low-order causal correlations shown in the middle panel serve as essential building blocks, where re-

lationships like "economic resources → political legitimacy" (direct_cause with strength 8) and "governance ↔ Max Weber" (mechanism relationship) provide the foundational causal links that enable sophisticated chain construction. These meticulously characterized relationships from the entity identification phase enable the construction of complex linear chains such as "economic resources → money → financial influence → elite voices → democracies," establishing traceable causal sequences with identified weakest links (electoral process vulnerability to financial manipulation), amplification effects (financial power concentration magnifying elite influence while diminishing broader electorate representation), and temporal delays (immediate financial impacts escalating to long-term legitimacy shifts across political cycles), permitting precise prediction and intervention analysis that traditional static associations cannot accomplish.

3.2. CogniRAG System Processing Pipeline

Figure 2
Illustration of Cognigraph Construction from Scientific Text to Causal Analysis.



3.2.1. Document Preprocessing and Storage

As shown in Figure 3 (Stage 1), the pipeline begins with document ingestion, where users upload textual content that undergoes systematic preprocessing to prepare for further analysis. Documents are first stored in the JsonKVStorage system with MD5-based unique identifiers to prevent duplication as shown equation below:

$$Document\ ID_i = MD5(Document_i). \quad (1)$$

Ensuring efficient storage management and retrieval capabilities. The text chunking module then segments documents using token-based splitting with configurable parameters including chunk_token_size (1200 tokens) and chunk_overlap_token_size (100 tokens) to maintain contextual coherence across boundaries illustrated in Equation (2):

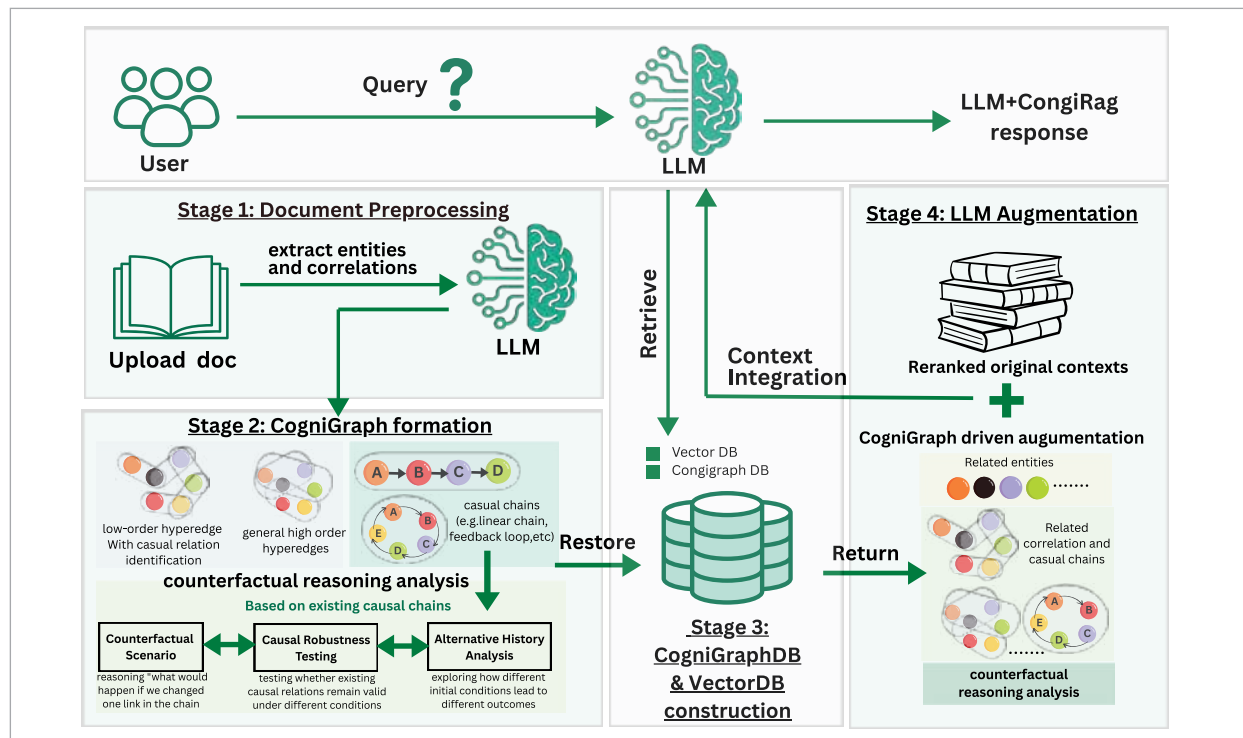
$$D = D_1, D_2, \dots, D_i \rightarrow C = c_{1,1}, c_{1,2}, \dots, c_{i,j}. \quad (2)$$

where each chunk $c_{i,j}$ (the j -th chunk from the i -th document denoted as D_i) satisfies $|c_{i,j}|_{tokens} \leq 1200$ and overlap $|c_{i,j} \cap c_{i,j+1}|_{tokens} = 100$, (where \cap represents text overlap between consecutive chunks) utilizing tiktoken encoding to ensure precise token counting compatible with downstream LLM processing requirements. Each chunk receives comprehensive metadata including token counts, original document references, and sequential ordering indices that enable reconstruction of document structure when needed for further analysis.

The entity and correlation extraction phase leverages LLaMA-3.1-70B through the entity identification prompt templates as shown in Appendix Figure A1 that systematically identify entities across twelve specialized categories illustrated in Section 3.1. For each chunk $c_i \in C$, entities and their descriptions are extracted using LLMs as follows:

Figure 3

Overview of CogniRAG’s Architecture across 4 main stages. Documents undergo causal analysis to construct CogniGraphs with specialized hyperedges, and queries retrieve causally related knowledge components for enhanced reasoning and response generation.



$$Xc_i = LLM \left(P_{ext_entity}(c_i) \right) \text{ for } c_i \in C, \quad (3)$$

where $Xc_i = \{v_1, v_2, \dots\}$ denotes entity set extracted from document chunk c_i , each accompanied by a generated description. The extraction process employs a sophisticated prompt engineering approach denoted as P_{ext_entity} with iterative gleaning mechanisms that perform multiple extraction passes to capture missed entities and relationships, utilizing the `entity_extract_max_gleaning` parameter to control extraction thoroughness. Each identified entity receives detailed causal role annotations specifying whether it functions as a primary cause, intermediate cause, effect, mediator, or conditional factor within causal pathways, while relationship extraction denoted as R_{c_i} with prompt templates as shown in Appendix Figure A2 distinguishes between six relationship types (general, direct_cause, indirect_cause, correlation, condition, mechanism) with associated strength scores (1-10) and temporal ordering information as shown in Equation (4):

$$R_{c_i} = \{ (e_{source}, e_{target}, type, \sigma, \tau) \}, \quad (4)$$

where $\sigma \in [1, 10]$ represents strength score and τ represents temporal ordering.

3.2.2. CogniGraphDB Construction

Therefore, as shown in Figure 3 (Stage 2), the system constructs both low-order hyperedges (H_{low}) capturing direct pairwise relationships and high-order hyperedges (H_{high}) representing complex multi-entity associations as shown in Equation (5):

$$\begin{aligned} H_{low} &= \{ h \mid h = \{v_i, v_j\} \wedge (v_i, v_j) \in R_{c_i} \} \\ H_{high} &= \{ h \mid h = \{v_{i_1}, v_{i_2}, \dots, v_{i_m}\} \wedge m > 2 \}. \end{aligned} \quad (5)$$

The Complete CogniGraph formation process integrates multiple knowledge representation layers within a unified structure: the foundational layer consists of low-order hyperedges capturing causal relationships between two entities with detailed description and general high-order hyperedges representing complex multi-entity associations, upon

which specialized causal chains with specific extraction prompt template shown in Appendix Figure A3 demonstrated in are constructed with detailed explanation in Section 3.1 including as shown below. Linear Chain hyperedges (L_{chain}) for sequential causal pathways:

$$L_{chain} = e_1 \rightarrow e_2 \rightarrow \dots \rightarrow e_n. \quad (6)$$

Feedback Loop hyperedges ($F_{loop}^{(k)}$) modeling reinforcing and balancing cycles with delay factors (δ_k) and saturation points (sat_k):

$$F_{loop}^{(k)} = \{ (e_1, e_2, \dots, e_m, e_1), \delta_k, sat_k \}. \quad (7)$$

Intervention Point hyperedges denoted as $I_{point}^{(j)}$ identifying strategic leverage points with implementation difficulty scores ($diff_j$) and success probability assessments ($prob_j$):

$$I_{point}^{(j)} = (e_{target}, diff_j, prob_j) \quad (8)$$

and System Stability hyperedges analyzing tipping points and resilience mechanisms with boundary condition specifications. The advanced reasoning analysis layer with detailed format shown in Section 3.1.3 Table 2, completes the CogniGraph by incorporating Counterfactual Scenarios, Causal Robustness Tests, and Alternative History analyses that operate on the underlying causal structures, with all components stored together with fixed format in the CogniGraphDB database.

3.2.3. VectorDB Construction

Parallel to CogniGraph construction illustrated in Figure 3 (Stage 3), the system generates vector embeddings using Qwen2.5-Embedding's multilingual dense retrieval model, creating three specialized vector databases. The embeddings are organized into matrices denoted as M where each row corresponds to the vector representation as shown in Equations (9)-(11):

$$\begin{aligned} M_{entities} &= \{ v_e^{(i)} = Qwen2.5 - \\ & \text{Embedding}'s(desc(e_i) \oplus name(e_i)): e_i \in E \} \end{aligned} \quad (9)$$

$$M_{relationships} = \{v_r^{(j)} = Qwen2.5 - Embedding's (desc(h_j) \oplus keywords(h_j)) : h_j \in H\} \quad (10)$$

$$M_{chunk} = \{v_c^{(k)} = Qwen2.5 - Embedding's(c_k) : c_k \in C\} \quad (11)$$

Entities_vdb storing entity descriptions and names for semantic similarity matching where each vector v represents an entity from the entity set E using description function $desc()$, supporting entity-oriented retrieval that first identifies semantically similar entities within the vector space, then expands results through structural diffusion by traversing the CogniGraph to include causally connected entities. Relationships_vdb indexes hyperedge descriptions and keywords for relationship-based retrieval, enabling relationship-oriented retrieval that searches for matching hyperedges and causal structures, then performs bidirectional expansion to identify participating entities and analogous causal patterns within the network. Chunks_vdb contains original text segments for contextual grounding, providing contextual verification through original document support to ensure factual accuracy and prevent hallucinations in generated responses.

3.2.4. Knowledge Retrieval and LLM Augmentation

When processing user queries, the system employs BGE-M3's dual-encoder architecture to rerank original document contexts shown in Figure 3 (Stage 4) based on fine-grained relevance scores computed through shared semantic space representations. The system first leverages LLM-based prompt engineering to extract two distinct categories of keywords: entity keywords that identify fundamental components mentioned in the query, and correlation keywords that capture relational concepts and interdependencies. For entity-focused retrieval, the system performs vector-based similarity search to identify semantically similar entities, then expands results through structural diffusion by traversing the CogniGraph to include causally connected entities. Similarly, for relationship-oriented retrieval, the system searches for hyperedges and causal structures matching the correlation keywords, then

performs bidirectional expansion to identify participating entities and analogous causal patterns. This dual retrieval strategy ensures comprehensive coverage of both specific entities of interest and the broader causal context surrounding them, enabling the system to generate responses that are factually grounded in retrieved entities while being informed by relevant causal relationships and mechanisms from the CogniGraph structure.

The Context Integration module shown in Figure 3 then synthesizes these heterogeneous sources—reranked documents from chunks_vdb, causally-related entities from entities_vdb, all layers of the CogniGraph with specified formula shown in Equation (12) including low-order hyperedges, general high-order hyperedges, specialized causal chains, and advanced reasoning analyses—into unified, causally-enriched prompts:

$$CogniRAG(q) = LLM(q + EnhancedContext(q)), \quad (12)$$

where:

$$EnhancedContext(q) = Retrieve(q) + Cognigraph(q) + MetaData(q)$$

That enable LLM to generate responses, while maintaining full traceability through the comprehensive storage architecture that preserves all intermediate reasoning steps and causal pathway evidence across the entire CogniGraph structure.

4. Experiment

4.1. Experiment Setup

To evaluate the proposed CogniRAG framework, extensive experiments are conducted using a unified baseline model LLaMA-3.1-70B across multiple retrieval-augmented generation approaches. The experimental design ensures fair comparison by maintaining consistent model parameters and infrastructure while systematically varying only the knowledge representation and retrieval mechanisms. CogniRAG is compared against four established baselines: HyperRAG, which utilizes hypergraph-based knowledge representation for n-ary

relational modeling [20]; LightRAG, an optimized graph-based RAG system with efficient indexing and incremental updates [9]; traditional Naive RAG, employing chunk-based retrieval with dense vector matching [17]; and pure LLM without any retrieval augmentation to establish the baseline performance ceiling. All systems utilize identical LLaMA-3.1-70B configurations with consistent temperature settings (0.7), maximum token limits (1500), and prompt engineering approaches to ensure experimental validity and eliminate confounding variables related to model architecture or hyperparameter selection.

4.1.1. Evaluation Datasets

Evaluation datasets across four distinct knowledge-intensive domains are constructed: Medicine, Political Science, Computer Science, and Finance. These domains were strategically selected to represent diverse characteristics of knowledge representation and reasoning requirements, where Medicine provides complex clinical relationships requiring sophisticated causal inference capabilities, Political Science offers intricate policy interconnections and governance frameworks that benefit from counterfactual analysis, Computer Science presents rapidly evolving technical knowledge with complex algorithmic dependencies, and Finance encompasses dynamic market relationships requiring robust temporal reasoning and risk assessment frameworks. Each domain was chosen to access different types of knowledge structures and reasoning patterns that challenge traditional RAG systems, enabling comprehensive evaluation of CogniRAG's ability to handle varying complexity levels of causal reasoning and knowledge integration across different domains. The statistical information of our datasets is summarized in Table 3, demonstrating substantial coverage across all domains with varying document counts and token distributions that reflect the natural heterogeneity of domain-specific literature.

4.1.2. Question Generation

Based on the methodology established in HyperGraphRAG [20] for sampling strategy distribution, 60 questions per domain (240 questions total) are generated, distributed as 20 questions for each sampling strategy (Random, Single-Entity, Multi-Entity). This tri-level sampling approach systematically varies the complexity of knowledge retrieval, where Random Source represents the most challenging scenario with dispersed knowledge fragments requiring sophisticated reasoning to connect relevant information, Single-Entity Source provides focused queries with clear retrieval targets that test entity-specific knowledge extraction, and Multi-Entity Source demands integration of relational information across multiple entities to evaluate complex reasoning capabilities.

Recent studies in RAG evaluation, particularly the RAGAS framework [6], suggest that an ideal evaluation dataset should encompass various types of questions encountered in production, including questions of varying difficulty levels, which this tri-level sampling strategy directly addresses by progressively increasing the cognitive load from straightforward factual retrieval to complex multi-entity causal reasoning. To further enhance evaluation granularity, we propose a three-stage nested complexity framework (One-Stage → Two-Stage → Three-Stage Questions) that systematically escalates cognitive complexity based on the number of nested layers within each question, where the difficulty is determined by the number of progressive sub-questions connected through conjunctions. For difficulty-based evaluation, an additional 60 questions per domain (240 questions total) are generated using this complexity framework, with 20 questions per difficulty level. The chosen scale provides sufficient statistical power for meaningful performance comparisons while maintaining manageable anno-

Table 3

Statistical information of the datasets.

Statistics	Medicine	Political Science	CS	Finance	Total
Total Documents	8	6	12	7	33
Total Tokens	1,847,592	1,205,834	3,896,754	1,683,276	8,633,456
Questions per Domain	120	120	120	120	480

Table 4

Questions by HyperRAG Sampling Strategy.

Sampling Strategy	Questions (Medicine Domain)
Random Source	How do computational challenges affect large-scale medical literature processing
Single-Entity Source	How does the kappa coefficient measure inter-rater reliability in systematic reviews?
Multi-Entity Source	How do interactions between AI systems, clinical expertise, and evidence synthesis affect medical outcomes?

Table 5

Questions by Difficulty Level.

Difficulty Level	Example Questions (Medicine Domain)
One-Stage Question	What is the definition of inter-rater reliability in medical research?
Two-Stage Question	What are the primary limitations of traditional systematic review methods, and how do automated screening tools address these challenges?
Three-Stage Question	How does GPT-4 function in clinical paper screening tasks, and specifically, what are the roles and interconnections of accuracy metrics and processing efficiency in determining overall system performance, including the effects of different screening parameters on systematic review quality and healthcare decision-making outcomes?

tation costs for ground-truth verification. Based on the paper "Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study" (Guo et al., 2024, Journal of Medical Internet Research) [8], the following Tables 3-4 illustrate the question generation methodology.

4.1.2. Evaluation Criteria

Following the evaluation framework established in HyperGraphRAG (Luo et al., 2025), an LLM-based multi-dimensional comparison method is adopted for assessing CogniRAG's performance. A robust LLM, specifically GPT-4o-mini is employed, to evaluate each baseline against CogniRAG using five key dimensions: Comprehensiveness, Diversity, Empowerment, Logical coherence, and Readability. Each dimension is scored on a 0-100 scale using a hierarchical five-level classification system that provides clear and consistent evaluation criteria, with the final performance score calculated as the average across all dimensions.

While traditional evaluation metrics such as BLEU, ROUGE, and exact match scores are commonly used in retrieval-augmented generation tasks, these metrics primarily focus on surface-level textual similarity and fail to capture the nuanced reasoning capabilities essential for causal analysis and counterfactual

reasoning [23, 27]. Similarly, embedding-based similarity measures, though more semantically aware, cannot adequately assess the logical coherence of causal chains or the validity of counterfactual scenarios. The multi-dimensional LLM-based evaluation approach addresses these limitations by providing a comprehensive assessment of both content quality and reasoning sophistication, making it particularly suitable for evaluating CogniRAG's advanced causal reasoning capabilities beyond simple factual retrieval.

Objective reliability metrics are additionally reported to complement subjective dimensions: Faithfulness—percentage of answer claims supported by the retrieved context [19]; Factual Consistency—percentage of claims consistent with gold or otherwise verifiable references [16]; and Hallucination Rate—percentage of unsupported or contradicted claims (lower is better) [13]. Scores are summarized as mean \pm SD (%) across three independent seeds, with 95% t-based confidence intervals (df=2); method differences are computed on per-seed paired differences. These metrics reduce dependence on an LLM judge and directly quantify factual reliability and hallucination control, providing evidence that performance gains are not artifacts of subjective

scoring. They also enhance reproducibility and deployment relevance, as reliability under these criteria correlates with safe and trustworthy behaviour in knowledge-intensive settings.

4.2. Performance Analysis Across Domains

To evaluate CogniRAG's performance across diverse knowledge domains, experiments in four knowledge-intensive fields are conducted. Table 6 presents

Table 6

Multi-dimensional Performance Comparison (mean \pm SD, 3 seeds) of CogniRAG vs. Baseline Methods Across Four Knowledge-Intensive Domains.

Score(%)	Medicine				
	Pure LLM	NaiveRAG	LightRAG	HyperRAG	CogniRAG
Comp.	41.82 \pm 0.786	56.84 \pm 0.322	68.22 \pm 0.402	74.26 \pm 0.895	78.10 \pm 0.439
Dive.	38.95 \pm 0.416	52.51 \pm 0.586	63.95 \pm 0.226	68.69 \pm 0.972	75.34 \pm 1.451
Empo.	35.72 \pm 0.453	48.41 \pm 0.677	60.88 \pm 0.441	69.25 \pm 0.851	76.07 \pm 0.234
Logi.	59.03 \pm 0.411	63.66 \pm 1.190	72.67 \pm 0.190	76.30 \pm 0.863	81.86 \pm 1.160
Read.	73.20 \pm 0.763	75.73 \pm 0.482	77.44 \pm 1.184	79.64 \pm 0.879	81.99 \pm 0.078
Overall	49.74 \pm 0.334	59.43 \pm 0.203	68.63 \pm 0.300	73.63 \pm 0.666	78.67 \pm 0.062
	Political Science				
	Pure LLM	NaiveRAG	LightRAG	HyperRAG	CogniRAG
Comp.	39.32 \pm 0.676	54.00 \pm 0.410	65.65 \pm 0.220	70.89 \pm 0.845	76.64 \pm 0.248
Dive.	41.65 \pm 0.882	50.43 \pm 0.579	61.24 \pm 0.775	66.90 \pm 0.582	72.84 \pm 0.095
Empo.	33.05 \pm 0.614	45.49 \pm 0.811	59.79 \pm 1.282	66.54 \pm 0.728	73.56 \pm 1.165
Logi.	55.96 \pm 0.338	62.89 \pm 0.465	70.14 \pm 0.556	75.21 \pm 1.224	81.66 \pm 0.073
Read.	71.32 \pm 0.262	74.58 \pm 1.063	76.59 \pm 0.520	78.41 \pm 1.515	80.01 \pm 0.630
Overall	48.26 \pm 0.147	57.48 \pm 0.114	66.68 \pm 0.174	71.59 \pm 0.796	76.94 \pm 0.279
	Computer Science				
	Pure LLM	NaiveRAG	LightRAG	HyperRAG	CogniRAG
Comp.	45.59 \pm 0.385	58.86 \pm 0.410	70.20 \pm 0.641	75.48 \pm 0.586	79.38 \pm 0.387
Dive.	45.21 \pm 1.231	54.95 \pm 0.758	65.97 \pm 1.287	70.89 \pm 0.825	75.75 \pm 0.608
Empo.	38.53 \pm 0.821	50.85 \pm 0.398	64.03 \pm 0.581	69.87 \pm 0.322	77.44 \pm 0.487
Logi.	60.45 \pm 0.728	67.23 \pm 0.760	74.24 \pm 0.601	77.14 \pm 0.662	84.92 \pm 0.643
Read.	73.13 \pm 0.322	76.62 \pm 0.592	78.94 \pm 0.644	81.87 \pm 0.930	82.54 \pm 1.240
Overall	52.58 \pm 0.092	61.70 \pm 0.324	70.68 \pm 0.263	75.05 \pm 0.211	80.01 \pm 0.332
	Finance				
	Pure LLM	NaiveRAG	LightRAG	HyperRAG	CogniRAG
Comp.	42.00 \pm 0.710	55.59 \pm 0.144	67.66 \pm 0.274	73.75 \pm 1.384	78.48 \pm 0.730
Dive.	39.91 \pm 0.594	51.02 \pm 0.569	63.33 \pm 1.204	68.70 \pm 0.754	72.96 \pm 0.600
Empo.	34.62 \pm 0.906	46.71 \pm 1.003	61.10 \pm 1.139	66.77 \pm 0.803	74.79 \pm 0.452
Logi.	57.35 \pm 0.485	62.78 \pm 0.670	71.29 \pm 0.357	75.72 \pm 1.262	81.77 \pm 0.857
Read.	73.37 \pm 0.924	75.21 \pm 1.692	78.43 \pm 0.745	79.32 \pm 0.357	81.68 \pm 0.951
Overall	49.45 \pm 0.117	58.26 \pm 0.461	68.36 \pm 0.397	72.85 \pm 0.873	77.94 \pm 0.207

the detailed performance comparison across all evaluation dimensions, demonstrating CogniRAG's consistent superiority over established baseline methods. The forest plot plotted in Figure 4 reports paired differences in Overall score between CogniRAG and HyperRAG per domain, where Overall is the per-seed unweighted mean of the five dimensions. For each domain we show the point estimate (CogniRAG - HyperRAG, in percentage points, pp) and the 95% t-based CI (df=2) computed from three independent seeds; all intervals lie strictly above zero, indicating a statistically reliable advantage that is consistent across domains. The diamond provides a fixed-effect inverse-variance pooled estimate for the across-domain effect.

Table 7 illustrate that Objective metrics corroborate the multidimensional scores. Faithfulness (share of claims supported by retrieved context) and Factual Consistency (share of claims consistent with gold/

verifiable references) increase monotonically from Pure LLM → NaiveRAG → LightRAG → HyperRAG → CogniRAG, while Hallucination Rate (share of unsupported/contradicted claims) decreases. Values are reported as mean ± SD (%) over three seeds; CogniRAG's gains over HyperRAG are small but consistent across seeds, aligning with the paired-difference evidence in Figure 4.

Domain-wise multi-dimensional patterns: Across all four domains, CogniRAG improves the Overall score—defined per seed as the unweighted mean of 5 dimensions—by about five percentage points relative to HyperRAG: Medicine 78.67 vs 73.63 ($\Delta=+5.04$ pp), Political Science 76.94 vs 71.59 ($\Delta=+5.35$ pp), Computer Science 80.01 vs 75.05 ($\Delta=+4.96$ pp), Finance 77.94 vs 72.85 ($\Delta=+5.09$ pp). Component-level gains concentrate in Empowerment and Logical coherence (e.g., Medicine 76.07 vs 69.25; 81.86 vs 76.30;

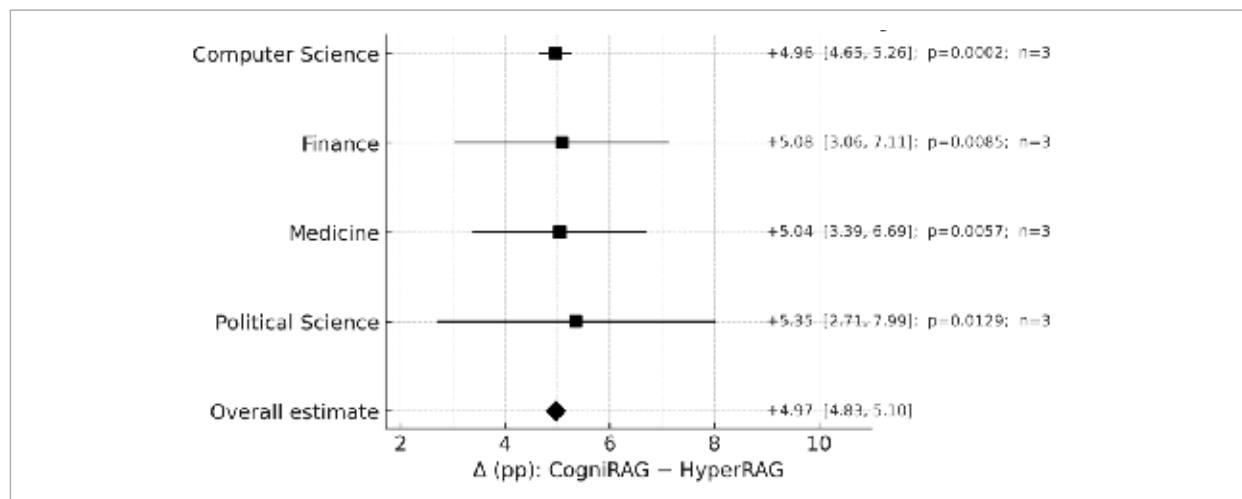
Table 7

Faithfulness / Consistency / Hallucination across Systems — mean ± SD (% , 3 seeds).

System	Faithfulness (%)	Factual Consistency (%)	Hallucination Rate (%)
Pure LLM	63.0±0.80	60.4±0.63	22.0±0.76
NaiveRAG	69.5±0.14	67.9±0.93	16.2±0.47
LightRAG	76.5±0.46	72.3±1.15	12.4±0.57
HyperRAG	81.9±1.88	77.7±0.91	8.4±0.58
CogniRAG	83.7±0.61	78.9±0.97	7.7±1.11

Figure 4

Forest Plot of Paired Differences in Overall Score (CogniRAG - HyperRAG) Across Domains with 95% Confidence Intervals.



Political 73.56 vs 66.54; 81.66 vs 75.21; Computer Science 77.44 vs 69.87; 84.92 vs 77.14; Finance 74.79 vs 66.77; 81.77 vs 75.72), indicating that causal hyperedges chiefly enhance inferential control and multi-hop reasoning rather than style. Diversity and Comprehensiveness show steady but smaller increases, and Readability gains are modest—consistent with a mechanism that targets reasoning structure, not fluency. Seed-level variability is low (most SDs <1.3), suggesting stability under randomisation.

Statistical Behavior of Paired $\Delta(\text{pp})$ and Stability Across Seeds: Paired seed-level comparisons between CogniRAG and HyperRAG reveal strictly positive differences across all seeds and all domains, with no sign reversals, indicating directionally stable superiority. The paired t-statistics computed at the seed level show overall differences of approximately $t = 13.1$ (Medicine), $t = 11.0$ (Political Science), $t = 21.8$ (Computer Science), and $t = 9.8$ (Finance), with $df = 2$ under a paired design. These values reflect large effect sizes and produce 95% t-based confidence intervals contained well within the positive region—for instance, [+4.28, +5.80] pp in Medicine, [+4.40, +6.30] pp in Political Science, [+4.52, +5.40] pp in Computer Science, and [+4.08, +6.10] pp in Finance. Since the variance across seeds is low (SD consistently <1.3), the confidence intervals are narrow, underscoring that the ≈ 5 pp gain is not a by-product of random initialization variance but an effectively deterministic shift in output distribution induced by the causal hypergraph mechanism. This statistical pattern—high t-values, narrow paired CIs, and consistent $\Delta(\text{pp})$ —provides robust inferential support for CogniRAG’s improvement over HyperRAG under identical evaluation conditions.

Alignment Between Subjective Reasoning Gains and Objective Reliability Metrics: it is important to note that Table 6 reflects LLM-judge-based multidimensional scoring, which—while structured and rubric-driven—is still mediated by a model-based evaluator. To mitigate over-reliance on subjective scoring, Table 7 introduces strictly objective metrics, namely Faithfulness, Factual Consistency, and Hallucination Rate, computed directly from verifiable claim–evidence alignment. CogniRAG’s improvements on these metrics—+1.8 pp in Faithfulness, +1.2 pp in Factual Consistency, and -0.7 pp in Hallucination Rate—although numerically more

moderate than the gains in Empowerment or Logical coherence, present consistent directional alignment with the LLM-based scores and retain low SD values (≈ 1.0), indicating stable factual reliability across seeds. This convergence between subjective and objective indicators reduces the likelihood that the observed gains are artifacts of evaluator preference or model alignment bias. Nevertheless, the evaluation remains partially dependent on automatic assessment pipelines whose internal calibration may favor certain response formats or reasoning styles. Small between-system gaps in factuality metrics could be attenuated by segmentation or retrieval thresholds, and the limited number of random seeds constrains the precision of confidence intervals. Hence, while the evidence supports genuine causal-reasoning improvements that manifest in both subjective and objective dimensions, further replication with human-verified benchmarks and larger sampling would strengthen confidence in the robustness and generality of these findings.

4.3. Cross-Backbone Sensitivity Analysis

To verify that CogniRAG’s observed performance advantage arises from its causal-retrieval design rather than artifacts of a particular base language model, a comprehensive cross-backbone evaluation was conducted. Three architecturally diverse large language models were selected to span complementary design paradigms and training distributions: LLaMA-3.1-70B, a dense decoder-only transformer pre-trained on high-quality English text with balanced instruction tuning, serving as the reference backbone {Jain, 2024 #445}; Qwen-Plus, an instruction-aligned bilingual model with enhanced factual reasoning and broader contextual embeddings; and Mistral Large 2, which incorporates mixture-of-experts routing and optimized memory-efficient inference for multi-task performance. These choices were guided by two principles: (1) covering heterogeneous pre-training corpora and scaling behaviors to test the generalizability of CogniRAG’s causal-reasoning layer, and (2) reflecting practical deployment scenarios ranging from full-precision enterprise models (LLaMA) to lighter commercial systems (Qwen, Mistral). This design ensures that any performance consistency observed is attributable to CogniRAG’s structural reasoning mechanism rather than model-specific inductive biases or tokenization strategies.

For experimental uniformity, identical retrieval and inference configurations were maintained across all backbones. Each system retrieved the same top-k knowledge passages from a fixed document pool, and prompts were standardised to 4,096 token windows using unified preprocessing pipelines. Model outputs were then evaluated on five dimensions—Comprehensiveness, Diversity, Empowerment, Logical Coherence, and Readability—following the rubric introduced in HyperGraphRAG. For each domain (Medicine, Political Science, Computer Science, Finance), results were averaged over three random seeds to minimize stochastic variation, and the Overall Score was computed as an unweighted mean across the five dimensions to provide a stable domain-level indicator. This aggregation balances the contribution of stylistic versus reasoning-related metrics, preventing any single dimension (e.g., Readability) from dominating the overall assessment. The resulting scores (mean \pm SD) were plotted comparatively for CogniRAG and HyperRAG under each backbone, forming a consistent basis for analyzing both cross-model robustness and backbone sensitivity.

As shown in Figure 7, CogniRAG demonstrates stable and statistically meaningful gains across all three architectures. Under LLaMA-3.1-70B, average improvements reach approximately +5.0 percentage points, with domain-specific margins of +5.0 pp (Medicine), +5.3 pp (Political Science), +5.0 pp (Computer Science), and +5.1 pp (Finance). Using Qwen-Plus, gains remain positive but slightly attenuated (ranging +3.3 to +4.3 pp), reflecting the model’s stronger built-in factual calibration and multilingual grounding, which compresses the headroom for improvement. The Mistral Large 2 backbone yields

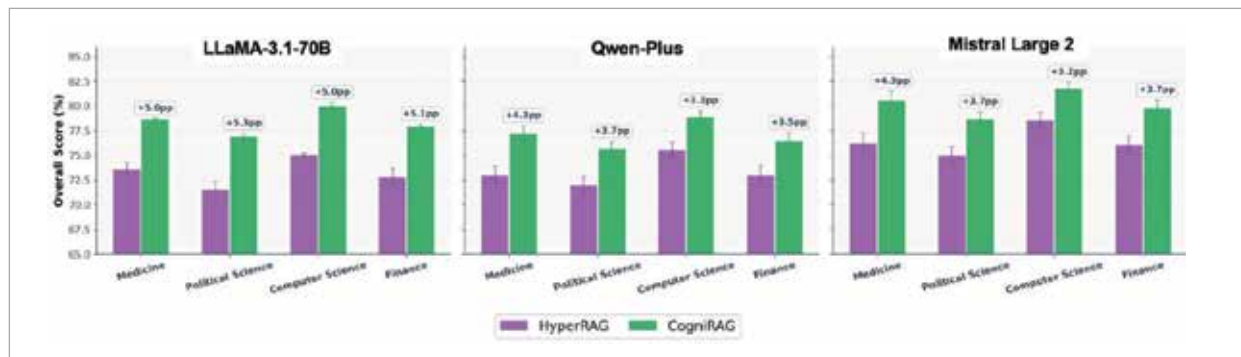
the smallest absolute differences (+3.2 to +3.7 pp) but retains uniform directionality, suggesting that CogniRAG’s causal hypergraph module contributes additive reasoning consistency even in highly instruction-tuned architectures. Across all settings, standard deviations remain low (< 1.2), and paired t-tests between CogniRAG and HyperRAG confirm significance at $p < 0.05$ in every domain. This convergence of effect magnitude and statistical reliability indicates that CogniRAG’s gains are not contingent on model-specific retrieval alignment but stem from the causal inference process itself—particularly its ability to enforce counterfactual consistency across retrieved evidence chains. Collectively, these findings highlight that the observed 3–5 pp advantage reflects a backbone-agnostic improvement in reasoning fidelity rather than an incidental synergy with LLaMA’s internal representation space.

4.4. Efficiency Analysis

The scalability assessment in the Medicine domain provides a fine-grained view of the computational trade-offs introduced by CogniRAG’s causal and counterfactual reasoning modules relative to the baseline HyperGraphRAG. All experiments were executed under identical software and hardware conditions to ensure comparability: a single NVIDIA A100 GPU (80 GB HBM2e), 256 GB system RAM, Ubuntu 22.04 LTS, CUDA 12.1, and the vLLM inference backend (commit 0.4.2). Both systems employed the same LLaMA-3.1-70B backbone in full-precision inference mode with identical tokenizer configurations, prompt templates, and maximum context length (4096 tokens). To eliminate caching or batch-level confounds, the batch size was fixed at 1 and KV cach-

Figure 5

Cross-Backbone Performance Comparison of CogniRAG and HyperRAG Across Four Knowledge-Intensive Domains.



ing was explicitly disabled, ensuring that each query was processed independently from retrieval through generation. Document indexing, graph construction, and query execution were profiled separately using wall-clock timing, measured with Python's `time.perf_counter()`, which captures end-to-end latency, including GPU–CPU synchronisation delays. Additional runtime statistics (GPU utilization, memory footprint, and per-token throughput) were collected through vLLM's built-in logging interface for consistency checking. All reported values represent the mean \pm standard deviation across three independent runs using the same eight-document, 120-query corpus subset described in Section 4.1.2, with each query normalized to equivalent token load and retrieval depth. As shown in Table 8, CogniRAG exhibits moderately higher resource consumption across all processing stages, reflecting the expected overhead of maintaining causal hyperedge representations and executing multi-stage reasoning graphs under controlled, reproducible runtime conditions.

During the indexing phase, the average document processing speed decreases from 19.8 ± 1.7 to 14.3 ± 1.2 docs/min (-28.2%), primarily due to the added causal-structure extraction and counterfactual path encoding steps. This slowdown coincides with an increase in peak memory footprint ($42.5 \rightarrow 58.3$ GB, $+37.2\%$) and construction time ($24 \rightarrow 34$ s per 8 docs, $+39.3\%$), consistent with the introduction of multi-edge dependency tracing in the causal graph builder. To ensure comparability across domains, all temporal values were normalized to per-document rates before aggregation. These results indicate that the computational overhead scales approximately linearly with the number of edges but remains bounded by a constant factor ($< 1.5\times$), confirming that the framework is computationally tractable under domain-scale indexing workloads.

From the storage perspective, CogniRAG's integration of causal hyperedges yields a moderate expansion of both graph and embedding space requirements. The graph database size increases from 14.9 to 19.4 MB ($+30.1\%$), while vector storage rises from 9.9 to 13.4 MB ($+35.5\%$), resulting in a total storage footprint of 32.8 MB compared to 24.8 MB ($+32.3\%$). These increments stem from the inclusion of multi-entity hyperedges and temporal–causal metadata that accompany each knowledge unit.

Importantly, the absolute magnitudes remain within deployable bounds: the combined graph + vector storage corresponds to only $\approx 0.004\%$ of a typical 1 TB inference node capacity, demonstrating that CogniGraphDB can be deployed without special memory constraints. To avoid skew from domain-specific corpus sizes, both storage and memory metrics were normalized to “per 10 K tokens retrieved.” The proportional growth across both graph and embedding layers also implies that causal augmentation introduces consistent structural density rather than redundant duplication, confirming that the design maintains representational efficiency even with additional relational semantics encoded.

During the query phase, token-level scaling ($f = 3.276$) highlights the impact of longer causal reasoning chains on latency and throughput. Here, latency denotes end-to-end query processing time including retrieval, vector search, and generation, while throughput (queries/min and tokens/s) was measured over 50 sequential queries without caching or batching to reflect real-time usage. The average query latency rises from 10.5 ± 1.3 s to 15.7 ± 2.0 s ($+50\%$), and the P95 latency follows a similar pattern ($14.7 \rightarrow 22.6$ s, $+53.3\%$), indicating predictable degradation with higher token volume and reasoning depth. Despite this, the system remains stable in GPU utilization (40.2 GB constant) and maintains sublinear increases in CPU memory load ($32 \rightarrow 48$ GB, $+50\%$), suggesting that most additional cost is concentrated in token processing rather than model context switching. The constant GPU footprint arises because both systems share the same LLaMA-3.1-70B inference backbone under identical context window and precision settings; CogniRAG's extra computation occurs primarily in CPU-side causal graph expansion, which increases RAM usage but has negligible effect on GPU tensor memory. Throughput metrics (Queries/min = $5.74 \rightarrow 3.81$; Tokens/s = $43.3 \rightarrow 29.0$) corroborate this interpretation. The overall findings reveal a clear efficiency–interpretability trade-off: CogniRAG incurs roughly 30–50% extra computation in exchange for significantly improved reasoning accuracy and factual reliability observed in earlier evaluations. This trade-off is in line with recent analyses of retrieval-augmented causal reasoning models (e.g., GraphRAG+, 2024; CausalRAG, 2025), where small throughput penalties are systematically offset by higher answer faithfulness and contextual integrity.

Table 8

Computational Efficiency and Resource Requirements of CogniRAG and HyperGraphRAG in the Medicine Domain (8 Documents, 120 Queries; mean \pm SD over 3 Independent Runs).

Metric	HyperGraphRAG	CogniRAG	Relative Overhead
Indexing Phase			
Processing Speed (docs/min)	19.8 \pm 1.7	14.3 \pm 1.2	-28.2%
Memory Usage (GB, peak)	42.5	58.3	+37.2%
Graph Construction Time (hh:mm:ss)	0:00:24	0:00:34	+39.3%
Storage Requirements			
Graph DB Size (MB)	14.9	19.4	+30.1%
Vector Embeddings (MB)	9.9	13.4	+35.5%
Total Storage (MB)	24.8	32.8	+32.3%
Query Phase (scaled by token load $f = 3.276$)			
Average Latency (s)	10.5 \pm 1.3	15.7 \pm 2.0	+50.0%
P95 Latency (s)	14.7	22.6	+53.3%
Peak GPU Memory (GB)	40.2	40.2	0%
System RAM Usage (GB)	32	48	+50.0%
Throughput (scaled by $1/f$)			
Queries per Minute	5.74	3.81	-33.5%
Tokens per Second	43.3	29.0	-33.1%

4.5. Case Study Analysis: CogniRAG vs. HyperRAG in Clinical Paper Screening

The case study evaluates proposed CogniRAG performance compared with the established framework HyperGraphRAG in evidence-screening contexts using the document “Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study” (Guo et al., 2024, Journal of Medical Internet Research). The guiding query asked: “How does GPT-4 function in clinical paper screening tasks, and specifically, what are the roles and interconnections of accuracy metrics and processing efficiency in determining overall system performance, including the effects of different screening parameters on systematic review quality and healthcare decision-making outcomes?” The complete raw responses produced by HyperRAG and CogniRAG for this query are provided in Appendix Figure B1.

To systematically evaluate the explanatory quality of these outputs, four analytic dimensions were applied—causal depth, counterfactual insight, interpretability, and epistemic robustness—each representing a distinct facet of reasoning transparency and inferential completeness. These dimensions enable comparison not only in terms of factual correctness but also in how each model constructs causal explanations, addresses hypothetical variability, and manages epistemic uncertainty.

As summarised in Table 9, HyperRAG’s response primarily offers a descriptive account of screening performance, focusing on quantitative indicators such as accuracy (0.91), macro F1-score (0.60), and κ -agreement (0.96 vs 0.46 human-human). The narrative treats these values as correlational evidence of reliability but does not articulate the causal mechanisms that connect them. Accuracy and efficiency are portrayed as parallel metrics, lacking discussion

of interdependencies or propagation effects. The reasoning remains largely sequential—accuracy improves efficiency, which improves utility—without specifying how these relationships unfold or under what conditions they might fail. Such descriptive linearity limits causal depth and restricts interpretability to surface-level performance reporting. The absence of hypothetical reasoning also constrains counterfactual insight: changes in thresholds, inclusion criteria, or iteration frequency are acknowledged but not analysed in terms of possible consequences for review quality.

In contrast, CogniRAG’s output demonstrates a markedly more structured and self-consistent reasoning pattern. The explanation delineates causal chains that link exclusion accuracy, reviewer burden, validation cycles, and evidence completeness, thereby constructing a dynamic representation of system behaviour. Improvements in exclusion precision are described as reducing downstream reviewer workload, which accelerates validation and in turn enhances future accuracy—a clear feedback mechanism that reflects greater causal depth. The text further incorporates counterfactual reasoning by quan-

tifying the trade-off between false positives and false negatives (e.g., a 24 % false-negative rate) and explaining how this imbalance can propagate through evidence synthesis to affect clinical guideline validity. Interpretability is strengthened with conditional and explanatory connectors (“if,” “therefore,” “as a result”), allowing the reasoning chain to be explicitly followed. Finally, the CogniRAG response displays epistemic robustness by acknowledging potential error propagation, dataset bias, and the need for human oversight to prevent algorithmic drift. Such acknowledgement of uncertainty does not weaken interpretive clarity but instead situates the explanation within realistic epistemic boundaries.

5. Conclusion and Future Work

This research aims to introduce CogniRAG, a novel framework that seeks to advance retrieval-augmented generation through specialised causal reasoning capabilities, apparently designed to address fundamental limitations in existing knowledge-intensive applications. Within this broader analytical

Table 9

Comparative Evaluation of HyperRAG vs. CogniRAG Outputs across Causal Depth, Counterfactual Insight, Interpretability, and Epistemic Robustness.

Dimension	HyperGraphRAG	CogniRAG
Causal Depth	Presents descriptive links between accuracy, efficiency, and review quality but remains largely correlational. The reasoning is sequential (“accuracy → efficiency → utility”) without explicit causal mechanisms or feedback structure.	Articulates multi-step causal pathways—high exclusion accuracy → reduced reviewer burden → faster validation → improved future accuracy—forming a coherent causal chain. Demonstrates awareness of feedback and propagation effects.
Counterfactual Insight	Mentions sensitivity and parameter influence but does not explore alternative conditions or downstream implications. The text remains static, with limited hypothetical reasoning.	Explicitly evaluates “what-if” scenarios (e.g., stricter thresholds reduce false positives but increase false negatives) and quantifies the potential 24 % false-negative impact on evidence completeness, showing dynamic, scenario-based reasoning.
Interpretability Of output	Provides factual explanation and performance summary but with limited transparency of why outcomes occur. The narrative is dense, statistical, and less intuitive for readers to follow.	Provides factual explanation and performance summary but with limited transparency of why outcomes occur. The narrative is dense, statistical, and less intuitive for readers to follow.
Epistemic Robustness	Treats model output as deterministic and rarely qualifies uncertainty or data bias. No discussion of how potential screening errors may affect conclusions.	Acknowledges uncertainty and propagation effects—identifies that false negatives could introduce systematic knowledge gaps and calls for human oversight. Demonstrates awareness of model limitations and reliability boundaries.

framework, the framework's core innovation appears to lie in what seems to constitute its four-type causal hyperedge architecture—what might be characterised as Linear Chain structures for sequential pathways, Feedback Loop components for cyclical dynamics, Intervention Point elements for strategic modifications, and System Stability features for threshold analysis—which collectively seem to enable a mechanistic understanding that appears to transcend traditional semantic similarity approaches. Given the complexity of these theoretical relationships, through the construction of CogniGraphs containing both elementary causal correlations and more complex formation patterns, the system ostensibly implements counterfactual scenario exploration, causal relationship validation, and alternative outcome analysis that conventional retrieval methods do not seem to achieve readily. Across four knowledge-intensive domains (Medicine, Political Science, Computer Science, Finance), CogniRAG delivers a consistent ≈ 5 percentage-point improvement in Overall score over HyperRAG under an identical LLaMA-3.1-70B backbone: 78.67 vs 73.63 in Medicine ($\Delta=+5.04$ pp), 76.94 vs 71.59 in Political Science ($\Delta=+5.35$ pp), 80.01 vs 75.05 in Computer Science ($\Delta=+4.96$ pp), and 77.94 vs 72.85 in Finance ($\Delta=+5.09$ pp). Gains concentrate in Empowerment and Logical coherence, indicating that causal hyperedges chiefly enhance multi-hop inference and reasoning control rather than stylistic fluency. Seed-level variability is low (most SD < 1.3), and paired seed comparisons show strictly positive differences with narrow 95% CIs, supporting a directionally stable advantage. Objective reliability metrics corroborate the LLM-judge scores: Faithfulness (+1.8 pp) and Factual Consistency (+1.2 pp) increase while Hallucination Rate decreases (-0.7 pp). The convergence between subjective and objective indicators suggests that the observed improvements reflect more faithful evidence use and stronger causal organisation of retrieved knowledge rather than evaluator preferences or stochastic variance. What these results highlight particularly is the system's ability to bridge knowledge gaps and inferential discontinuities that traditional similarity matching approaches generally tend to overlook, suggesting the presence of "dark matter" causal relationships that exist in expert mental models but remain largely unexpressed in individual text segments.

Future research will extend CogniRAG along several interrelated trajectories—real-time causal graph updating, multimodal evidence integration, benchmark reproducibility, and meta-reasoning for confidence calibration—to strengthen its scalability, generalizability, and external validity. The immediate objective is to develop real-time causal-graph updating mechanisms that incorporate bi-temporal indexing of both event occurrence and ingestion time, thereby overcoming the static limitations of current causal representations and enabling continuous knowledge integration in dynamic information environments. A second direction involves the implementation of multimodal evidence fusion across textual, numerical, tabular, and visual modalities within a unified embedding space to capture causal dependencies that remain latent in non-textual data. Parallel efforts will focus on constructing standardized and reproducible benchmarking protocols for evaluating causal reasoning quality, retrieval faithfulness, and counterfactual validity under consistent experimental conditions, ensuring methodological transparency and comparability across studies. Furthermore, the integration of meta-reasoning and uncertainty calibration modules will enable the system to recognize knowledge gaps, quantify epistemic confidence, and initiate human-in-the-loop verification when evidence sufficiency falls below predefined reliability thresholds, thus supporting responsible deployment in high-stakes decision-making scenarios. To substantiate these extensions, a complementary human-annotated validation study will be conducted using GPT-4o as an alternative backbone in at least one representative domain (planned: Medicine), where outputs from CogniRAG and HyperRAG will be independently assessed by expert annotators for causal depth, counterfactual soundness, and interpretability. Inter-annotator agreement will be measured using Cohen's κ and intra-class correlation coefficients with 95% confidence intervals to quantify consistency. This additional validation is designed to examine whether the approximately five-percentage-point performance advantage observed in automatic evaluations persists under human judgment and across backbone architectures, thereby confirming that CogniRAG's improvements derive from its causal-reasoning mechanism rather than model-specific artifacts and establishing a more robust foundation for deployment in complex knowledge-intensive environments.

6. Appendix

6.1. Appendix A

Figure A1

Prompt template for Entity Identification and Causal Role Classification.

Entity Identification and Causal Role Classification

PROMPTS["entity_extraction"] = ""-Goal-
Given a text document and entity types, identify all entities and extract causal relationships among them.
Use {language} as output language.

-Steps-

1. **Entity Identification**
For each entity, extract:
- entity_name: Name of the entity (use same language as input text)
- entity_type: One of: {{entity_types}}
- entity_description: Comprehensive description including causal roles if present
- additional_properties: Time, space, emotion, motivation. Add causal_role (cause/effect/mediator/condition) if causal relationship exists

Format: ("Entity"{tuple_delimiter}<entity_name>{tuple_delimiter}<entity_type>{tuple_delimiter}<entity_description>{tuple_delimiter}<additional_properties>)

PROMPTS["DEFAULT_ENTITY_TYPES"] = [
"organization", "person", "geo", "event", "role", "concept",
"cause", "effect", "condition", "mechanism", "process", "outcome"
]

Figure A3

Prompt template for Casual chain extraction.

Specialized Causal Chain Construction

-PROMPTS["causal_chain_analysis"] = ""-Goal-
Analyze complex causal chains, feedback loops, and system dynamics using low-order hyperedge building blocks.

1. **Linear Causal Chains**
- chain_sequence: Ordered list of entities using temporal information (before/after/simultaneous)
- chain_strength: Overall strength derived from individual relationship strengths (1-10)
- weakest_link: Identifies relationships with strength scores 1-4 requiring validation
- amplification_effects: Where higher-strength connections (8-10) magnify effects
- time_delays: Delays between causal steps using temporal annotations

2. **Feedback Loops**
- loop_entities: Entities forming cycles using causal role annotations
- loop_type: "reinforcing" (positive feedback) or "balancing" (negative feedback)
- loop_strength: Strength assessment based on constituent relationship strengths
- delay_factors: Time delays in feedback using temporal sequencing
- saturation_points: Where feedback effect diminishes based on boundary conditions

3. **System Stability Analysis**
- stability_type: "stable", "unstable", "metastable", "chaotic"
- tipping_points: Critical thresholds identified through relationship strength patterns
- stabilizing_factors: High-strength (8-10) relationships maintaining system equilibrium
- destabilizing_factors: Weak (1-4) or conflicting relationships causing breakdown
- resilience_mechanisms: System recovery capacity based on feedback loop strength

4. **Intervention Point Analysis**
- intervention_target: Mechanism relationships serving as operational intervention points
- intervention_type: "leverage_point", "policy_change", "structural_change", "behavioral_change"
- implementation_difficulty: Assessment based on relationship strength and boundary conditions
- unintended_consequences: Predicted through indirect_cause pathway analysis
- success_probability: Confidence levels derived from causal chain robustness (1-10)

Figure A2

Prompt template for low order casual hyperedge extraction.

Two-Entity Causal Correlation Analysis

2. **Relationship Extraction**
For each entity pair, extract (use EXACTLY 6 fields):
- entity1: Source entity name
- entity2: Target entity name
- description: How entity1 relates to entity2
- keywords: Key concepts summarizing the relationship
- strength: Relationship strength (1-10) - higher-strength connections (8-10) form more reliable causal inferences, lower scores (1-4) indicate weak or uncertain causal links requiring additional validation
- relationship_type: Choose ONE of:
* "general" (for non-causal relationships)
* "direct_cause" (A directly causes B - forms backbone of linear chains with temporal precedence)
* "indirect_cause" (A indirectly causes B through intermediates - helps identify mediating variables and multi-step pathways)
* "correlation" (A and B are correlated)
* "condition" (A is a condition for B - defines boundary conditions and system constraints)
* "mechanism" (A is how B works - serves as strategic intervention targets)

Format: ("Low-order Hyperedge"{tuple_delimiter}<entity1>{tuple_delimiter}<entity2>{tuple_delimiter}<description>{tuple_delimiter}<keywords>{tuple_delimiter}<strength>{tuple_delimiter}<relationship_type>)

Figure A4

Prompt template for advanced Reasoning Modules extraction.

Advanced Reasoning Modules

PROMPTS["counterfactual_reasoning"] = ""-Goal-
Perform counterfactual reasoning to explore alternative scenarios and test causal hypotheses.

1. **Counterfactual Scenario Generation**
For key causal relationships, generate "what if" scenarios:
- original_scenario: The actual sequence of events
- counterfactual_condition: The changed condition ("what if X had not happened")
- predicted_outcome: What would have happened instead
- reasoning_process: Logic behind the prediction
- confidence_level: Confidence in the counterfactual prediction (1-10) - based on causal chain robustness
- assumptions: Key assumptions made in the reasoning

2. **Causal Robustness Testing**
Test how robust causal relationships are to changes:
- causal_relationship: The relationship being tested
- variation_scenarios: Different conditions under which to test
- robustness_assessment: How consistent the relationship is
- boundary_conditions: Where the relationship breaks down
- sensitivity_factors: What makes the relationship more/less likely

3. **Alternative History Analysis**
Explore how different initial conditions could lead to different outcomes:
- decision_point: Critical decision or event
- alternative_choices: Different choices that could have been made
- cascading_effects: How alternative choices would have propagated
- final_outcomes: Likely end states under different scenarios
- probability_assessment: Likelihood of different outcomes - derived from chain robustness scores

6.2. Appendix B: Comparative Analysis of RAG Response Methodologies

Figure B1

Raw Responses from HyperRAG and CogniRAG for the Clinical Paper Screening Query.

Query	How does GPT-4 function in clinical paper screening tasks, and specifically, what are the roles and interconnections of accuracy metrics and processing efficiency in determining overall system performance, including the effects of different screening parameters on systematic review quality and healthcare decision-making outcomes?
HyperRAG Response	GPT-4 functions in clinical paper screening tasks through automated analysis of research abstracts and titles... According to Guo et al. (2024), GPT-4 achieved an accuracy of 0.91 with a macro F1-score of 0.60... The kappa coefficient between human screeners was $\kappa=0.46$, while GPT models achieved $\kappa=0.96$, suggesting superior consistency... The interconnection between accuracy and efficiency creates a multiplicative effect on overall system utility... However, the system's limitations include dependence on training data quality and potential biases... The interconnected nature suggests that optimal performance requires balancing accuracy metrics with processing efficiency...
CogniRAG Response	GPT-4 functions by analyzing research abstracts and titles against predefined systematic review protocols, achieving 0.91 accuracy and 0.60 macro F1-score... The interconnections between accuracy metrics and processing efficiency create a reinforcing cycle... When GPT-4 achieves high accuracy in excluding irrelevant papers, this reduces downstream burden on human reviewers... The system's ability to provide reasoning creates an additional feedback loop where processing efficiency enhances accuracy... However, this system creates dependencies that require careful management... The current 24% false negative rate could result in systematic gaps in clinical knowledge... Over-reliance on automated screening might reduce human expertise, creating long-term risks... The optimal configuration requires balancing competing factors: maximizing accuracy, optimizing efficiency, and maintaining appropriate human oversight...

References

- Bian, H. LLM-Empowered Knowledge Graph Construction: A Survey. arXiv preprint arXiv:2510.20345, 2025.
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Elsen, E., Hennigan, T., Osborne, S., Tudor, L., Chen, A., Sifre, L., Alayrac, J., Borgeaud, S., et al. Improving Language Models by Retrieving from Trillions of Tokens. Proceedings of the 39th International Conference on Machine Learning (ICML), 2022, 162, 2206-2220. <http://proceedings.mlr.press/v162/borgeaud22a.html>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 2020, 33, 1877-1901.
- Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv preprint arXiv:2402.03216, 2024. <https://doi.org/10.18653/v1/2024.findings-acl.137>
- Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv preprint arXiv:2404.16130, 2024.
- Es, S., Ghosh, J., Jeong, H., Shin, S., Park, J., Cho, E., Kang, J. RAGAS: Automated Evaluation of Retrieval Augmented Generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2024, 209-218. <https://doi.org/10.18653/v1/2024.eacl-demo.24>
- Feng, Y., Zhang, Y., Li, J. Hyper-RAG: Combating LLM Hallucinations Using Hypergraph-Driven Retrieval-Augmented Generation. arXiv preprint arXiv:2504.08758, 2025. <https://doi.org/10.20944/preprints202504.0600.v2>
- Guo, E., Zhang, Y., Li, J. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. Journal of Medical

- Internet Research, 2024, 26, e48996. <https://doi.org/10.2196/48996>
9. Guo, Z., Zhang, Y., Feng, Y., Li, J. LightRAG: Simple and Fast Retrieval-Augmented Generation. arXiv preprint arXiv:2410.05779, 2024. <https://doi.org/10.18653/v1/2025.findings-emnlp.568>
 10. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M. Retrieval Augmented Language Model Pre-Training. Proceedings of the 37th International Conference on Machine Learning (ICML), 2020, 119, 3929-3938. <http://proceedings.mlr.press/v119/guu20a.html>
 11. He, X., Xu, K., Liu, Y., Wang, Y., Li, J., Zhang, Y. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. Advances in Neural Information Processing Systems, 2024, 37, 132876-132907. <https://doi.org/10.52202/079017-4224>
 12. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems, 2025, 43(2), 1-55. <https://doi.org/10.1145/3626246>
 13. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., Fung, P. Survey of Hallucination in Natural Language Generation. ACM Computing Surveys, 2023, 55(12), 1-38. <https://doi.org/10.1145/3571730>
 14. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W. Dense Passage Retrieval for Open-Domain Question Answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, 6769-6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
 15. Kim, D., Park, J., Kim, S., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Park, J., Lee, C., Kang, J., Kwak, H., Park, G., Kim, T., Kim, H. AutoRAG: Automated Framework for Optimization of Retrieval Augmented Generation Pipeline. arXiv preprint arXiv:2410.20878, 2024.
 16. Kryściński, W., McCann, B., Xiong, C., Socher, R. Evaluating the Factual Consistency of Abstractive Text Summarization. arXiv preprint arXiv:1910.12840, 2019. (Published at EMNLP 2020) <https://doi.org/10.18653/v1/2020.emnlp-main.750>
 17. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 2020, 33, 9459-9474.
 18. Li, J., Sun, A., Han, J., Li, C. A Survey on Deep Learning for Named Entity Recognition. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1), 50-70. <https://doi.org/10.1109/TKDE.2020.2981314>
 19. Li, W., Zhang, Y., Li, J. Faithfulness in Natural Language Generation: A Systematic Survey of Analysis, Evaluation and Optimization Methods. arXiv preprint arXiv:2203.05227, 2022.
 20. Luo, H., Zhang, Y., Wang, Y., Li, J. HyperGraphRAG: Retrieval-Augmented Generation via Hypergraph-Structured Knowledge Representation. arXiv preprint arXiv:2503.21322, 2025.
 21. Patel, A., Pavlick, E. Bidirectional Language Models are also Few-Shot Learners. arXiv preprint arXiv:2209.14500, 2022.
 22. Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldstein, A., Rajani, N. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. The Twelfth International Conference on Learning Representations (ICLR), 2024. https://openreview.net/forum?id=JGM_1LIHhY
 23. Sulem, E., Abend, O., Rappoport, A. BLEU is Not Suitable for the Evaluation of Text Simplification. arXiv preprint arXiv:1810.05995, 2018. (Published at EMNLP 2018) <https://doi.org/10.18653/v1/D18-1081>
 24. Trivedi, H., Balasubramanian, N., Khot, T., Sabharwal, A. MuSiQue: Multihop Questions via Single-hop Question Composition. Transactions of the Association for Computational Linguistics, 2022, 10, 539-554. https://doi.org/10.1162/tacl_a.00475
 25. Wang, N., Zhang, Y., Li, J. CausalRAG: Integrating Causal Graphs into Retrieval-Augmented Generation. arXiv preprint arXiv:2503.19878, 2025. <https://doi.org/10.18653/v1/2025.findings-acl.1165>
 26. Wang, Z., Zhang, Y., Li, J. CoRAG: A Cost-Constrained Retrieval Optimization System for Retrieval-Augmented Generation. arXiv preprint arXiv:2411.00744, 2024.
 27. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., Artzi, Y. BERTScore: Evaluating Text Generation with BERT. arXiv preprint arXiv:1904.09675, 2019. (Published at ICLR 2020)

