Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 21, No. 2, 2026

Dynamic Knowledge Graph Augmentation Enhances Factual Accuracy in Retrieval Based Generation

Rafael Costa* and Elisabeth Gruber

Faculty of Computer Science, University of Vienna, Austria

*** Corresponding author:** rafael.costa@univie.ac.at

Abstract: Large language models (LLMs) demonstrate exceptional fluency in natural language generation but remain susceptible to producing factually incorrect outputs due to static parametric knowledge frozen at training time. Retrieval-augmented generation (RAG) partially mitigates this limitation by conditioning generation on externally retrieved evidence, yet conventional RAG systems depend on unstructured, flat document corpora that fail to represent the relational and temporal dynamics characterizing real-world knowledge. This paper proposes Dynamic Knowledge Graph Augmentation (DKGA), a framework that integrates continuously updated knowledge graphs (KGs) with retrieval-based generation pipelines to substantially improve factual accuracy. DKGA employs a graph neural network (GNN) encoder for subgraph-conditioned entity representation learning, a temporal update module for incremental knowledge refresh, and a cross-modal relevance-aware fusion mechanism that jointly conditions the generator on structured KG evidence and unstructured text passages. Experiments on the WebQA and TriviaQA benchmarks demonstrate that DKGA achieves an 11.4% improvement in exact match factual accuracy and an 18.2% reduction in hallucination rate over strong RAG baselines. These results provide compelling evidence that dynamic, structured knowledge representations are a critical and underutilized resource for knowledge-intensive language generation.

Keywords: Retrieval-augmented generation; Dynamic knowledge graph; Knowledge graph augmentation; Factual accuracy; Large language models; Graph neural networks; Hallucination reduction.

1. Introduction

The past several years have witnessed the emergence of large language models (LLMs) as the dominant paradigm in natural language processing (NLP), with systems such as GPT-4, LLaMA, and Gemini demonstrating extraordinary generative fluency and broad coverage of world knowledge acquired during pre-training on massive text corpora. Despite these advances, the fundamental architecture of autoregressive language models encodes world knowledge implicitly within billions of neural network parameters, and this knowledge is frozen at the time of training. This design creates a well-documented class of reliability failures: LLMs frequently generate plausible-sounding but factually incorrect statements, a phenomenon broadly referred to as hallucination, and they are inherently incapable of incorporating information about events postdating their training cutoff without expensive re-training or fine-tuning cycles. In high-stakes deployment contexts such as medical question answering, legal reasoning, and financial analysis, hallucinated facts carry serious practical consequences that motivate architectural solutions beyond scaling parametric model capacity.

Retrieval-augmented generation (RAG) was proposed as a principled architectural response to these limitations [1]. Rather than encoding all required knowledge in model parameters, RAG systems retrieve relevant passages from an external corpus at inference time and condition the generator on this retrieved evidence, effectively separating the parametric knowledge encoded in model weights from the non-parametric knowledge stored in the retrieval index. The original RAG formulation demonstrated substantial improvements in factual accuracy across a range of knowledge-intensive benchmarks, and retrieval augmentation has since become a foundational component of modern LLM-based systems. A parallel line of work proposed embedding the retrieval operation within the language model pre-training objective itself, demonstrating that jointly trained retrieval and generation components learn more mutually beneficial representations than systems in which these components are optimized independently [2]. However, both approaches treat the retrieval corpus as a flat, unstructured collection of text passages, forfeiting a significant quantity of relational, hierarchical, and temporal structure that characterizes real-world factual knowledge.

Knowledge graphs (KGs) represent a fundamentally different approach to knowledge storage, encoding facts as typed entity-relation-entity triples that explicitly represent relational structure at scale. Large-scale KGs such as Wikidata, Freebase, and YAGO provide comprehensive coverage of named entities and their interrelationships across diverse domains, and a growing body of research has demonstrated consistent improvements from structured KG integration across knowledge-intensive natural language generation tasks [3]. Integrating KGs into end-to-end retrieval-augmented generation pipelines raises non-trivial challenges that existing work has not fully resolved. Standard KG embedding methods produce representations optimized for link prediction rather than for generative conditioning, and most existing KG-augmented language models are trained on static graph snapshots, rendering them insensitive to knowledge updates that occur during deployment. Dense Passage Retrieval (DPR), which improved retrieval recall through dual-encoder contrastive training over annotated question-passage pairs [4], addressed the retrieval precision bottleneck within the text-only RAG paradigm but left the structural knowledge representation problem fundamentally unaddressed. Extending the strong retrieval gains achieved by DPR into a structured KG setting requires principled approaches to relational encoding, subgraph selection, and

evidence fusion that do not yet exist in a unified, dynamically maintainable architecture.

The Fusion-in-Decoder architecture demonstrated that encoding retrieved passages independently and fusing their representations through cross-attention in the decoder substantially improves multi-passage evidence integration [5], yet this approach cannot naturally extend to structured KG evidence whose relational semantics are not captured by passage-level dense representations. The broader RAG research community has identified knowledge freshness, retrieval granularity, and evidence fusion as the three most consequential open problems limiting factual accuracy in deployed systems [6], and these precise limitations motivate the architectural contributions of the present work.

This paper introduces DKGA, a Dynamic Knowledge Graph Augmentation framework that addresses these limitations through three principal innovations. First, a temporal update module continuously refreshes the knowledge graph by extracting and integrating new triples from incoming documents, ensuring that the structured knowledge base reflects the current state of world knowledge rather than a stale training-time snapshot. Second, a GNN-based subgraph encoder produces dense, context-aware representations of query-relevant KG subgraphs, trained with a contrastive objective to align entity representations with pre-trained language model embedding spaces. Third, a cross-modal relevance-aware fusion mechanism jointly attends over GNN-encoded KG evidence and DPR-retrieved text passages, enabling the generator to flexibly integrate structured and unstructured knowledge within a unified context representation. Experiments on WebQA and TriviaQA confirm that DKGA achieves consistent and substantial improvements in factual accuracy over strong baselines, with ablation studies providing a mechanistic account of the sources of improvement.

2. Literature Review

The present work draws on three converging lines of research: retrieval-augmented generation, knowledge graph representation learning, and the integration of structured knowledge into pre-trained language models. Each of these areas has advanced substantially since 2019, and the DKGA framework synthesizes insights from all three to address limitations that remain unresolved by any single prior approach.

Research on KG-based reasoning over question answering graphs provided early evidence that combining language model representations with structured relational context through joint GNN reasoning substantially outperforms text-only approaches on multi-hop questions [7]. The knowledge graph attention network demonstrated that iterative graph attention network layers applied to model higher-order entity connectivity yield substantially richer representations than single-hop embedding lookups, with attention-weighted neighborhood aggregation learning to focus on the most informationally relevant relational pathways for each query [8]. A complementary direction generalized graph convolutional approaches to multi-relational graphs, incorporating relation-specific transformations in each message-passing step and achieving strong performance across multiple KG completion benchmarks through joint entity-relation representation learning [9]. Neural Bellman-Ford Networks unified path-based and neighborhood-based reasoning in a differentiable shortest-path algorithm, providing a principled basis for multi-hop link prediction and entity retrieval that directly informs DKGA's approach to multi-hop subgraph encoding [10]. Pre-trained language model representations were shown to be directly applicable to KG completion through BERT-based triple scoring, establishing an important bridge between KG embedding methods and language model representations that later frameworks would build upon [11].

Within the RAG paradigm, the Self-RAG framework introduced a unified model that learns to decide when to retrieve, evaluate passage relevance, and critique its own generation through special reflection tokens inserted during supervised fine-tuning, achieving competitive results across a broad range of tasks [12]. Active retrieval approaches demonstrated that triggering retrieval based on uncertainty estimates during ongoing generation substantially improves efficiency and accuracy on long-form tasks where continuous retrieval would be computationally prohibitive [13]. The RETRO system demonstrated that retrieval from a multi-trillion token datastore, combined with a chunked cross-attention mechanism interleaving retrieval steps with generation steps, can match the performance of models with substantially more parameters [14]. These works collectively establish that the timing, frequency, and granularity of retrieval operations significantly influence generation quality, motivating DKGA's query-triggered subgraph extraction approach.

A comprehensive survey of knowledge-enhanced text generation demonstrated consistent improvements from structured knowledge integration across commonsense generation, data-to-text, and open-domain question answering (QA) tasks, identifying relational encoding depth and knowledge currency as the most impactful design dimensions [15]. Research specifically targeting multi-hop QA demonstrated that KG embeddings can effectively constrain and guide the exploration of multi-hop relational paths during inference, reducing the search space and improving both accuracy and interpretability [16]. Experiments on QA over incomplete KGs showed that even partial graph coverage supports reliable question answering when combined with knowledge-aware reading comprehension modules, underscoring the complementarity of structured and unstructured knowledge sources that DKGA's fusion mechanism is designed to exploit [17]. The Chain of Knowledge framework extended this insight to LLM grounding, demonstrating that step-wise structured knowledge base traversal reduces hallucination in multi-step reasoning scenarios compared to ungrounded generation [18].

Controlled experiments on in-context retrieval augmentation revealed that the format and structure of retrieved evidence significantly influences generation quality, providing empirical support for the hypothesis that structured KG evidence retrieval outperforms unstructured passage retrieval on entity-centric queries [19]. An analysis of conditions under which language models should trust their parametric versus retrieved knowledge found that entity frequency in training data is a strong predictor of parametric reliability, and that queries involving low-frequency or recently updated entities benefit most from structured retrieval augmentation [20]. ERNIE 3.0 incorporated entity-level KG knowledge into large-scale language model pre-training through auxiliary entity prediction objectives, yielding improvements on entity-centric tasks including relation extraction and entity-linked question answering [21].

Weakly supervised contrastive pre-training objectives were demonstrated to be highly effective for learning general-purpose text embeddings that generalize across domains [22], and DKGA's entity encoder adopts a similar contrastive approach adapted for the graph-structured setting.

The Search-in-the-Chain framework demonstrated that traceable, source-attributed evidence integration substantially reduces hallucination in knowledge-intensive generation by enabling the generator to verify each generated claim against its retrieval provenance [23]. The Graph RAG approach to query-focused summarization showed that hierarchically structured knowledge graphs support high-quality synthesis across document collections that flat retrieval corpora cannot represent [24]. The ChatKBQA framework provided a generate-then-retrieve paradigm for knowledge base question answering that demonstrated the complementarity of language model generation and structured KB lookup in a unified pipeline [25].

The foundational DrQA system introduced the two-stage retrieve-then-read paradigm for open-domain QA over Wikipedia, establishing the architectural pattern that subsequent RAG systems would build upon and demonstrating that large-scale unstructured text retrieval combined with a neural reading comprehension component can answer diverse factual questions [26]. Jumping Knowledge Networks investigated how aggregation depth interacts with graph topology to determine the effective receptive field of node representations, demonstrating that adaptive aggregation depth selection substantially outperforms fixed-depth approaches by matching the aggregation range to the structural properties of each node's local neighborhood [27]. HotpotQA introduced a large-scale benchmark for diverse, explainable multi-hop question answering that requires reasoning over multiple supporting paragraphs, providing the evaluation framework and task formulation that directly informs the multi-hop QA evaluation methodology adopted in the present work [28]. The R-GCN framework for modeling relational data demonstrated that relation-type-conditioned message passing in multi-relational graphs substantially outperforms relation-agnostic approaches for both entity classification and link prediction [29]. Early work on semi-supervised node classification with graph convolutional networks established the layer-wise propagation rule that forms the mathematical foundation of modern GNN architectures [30]. Inductive representation learning through neighborhood sampling and aggregation demonstrated that GNNs can generalize to entirely unseen nodes at inference time, a property that is critical for DKGA's dynamic KG setting where new entities are continuously added to the graph [31]. Collectively, this body of work establishes a robust empirical and theoretical foundation for the DKGA framework's core hypothesis that dynamic, continuously updated KG augmentation produces compounding factual accuracy improvements over static retrieval-only baselines.

3. Methodology

3.1. Dynamic Knowledge Graph Construction and Temporal Update

The first principal component of the DKGA framework is a continuously maintained knowledge graph that serves as the primary structured knowledge source for the retrieval pipeline. The design philosophy of this component departs fundamentally from the static KG snapshots employed by prior KG-augmented generation systems. Rather than constructing the graph once from a fixed data dump and leaving it unchanged throughout deployment, DKGA's knowledge graph is engineered to grow and evolve in response to incoming document streams, ensuring that its factual content remains temporally aligned with the real-world state of affairs it represents.

The dynamic construction process involves three interleaved sub-processes operating in a continuous asynchronous pipeline. The first is entity and relation extraction, performed by a fine-tuned span-extraction model that identifies named entities, event mentions, and relational predicates within incoming documents. Each extracted triple consisting of a subject entity, a relation type label, and an object entity is assigned both a confidence score from the extraction model and a provenance timestamp reflecting the publication date of the source document. This timestamp enables the second sub-process, temporal conflict resolution, which handles cases where a newly extracted triple contradicts an existing triple. A recency-weighted policy deprecates older conflicting triples without deleting them; deprecated triples are retained in a versioned factual history queryable for historical state reconstructions, proving especially important for queries concerning past role holders, discontinued products, or superseded scientific findings. The third sub-process, incremental graph indexing, maintains an approximate nearest-neighbor index over entity embeddings that is updated continuously as new entities are added and existing representations are refined through incoming neighborhood context, avoiding the computationally prohibitive full index reconstruction otherwise required.

The architectural relationship between the DKGA retrieval pipeline and its document-reading generation component follows the two-stage design pattern of the foundational open-domain QA paradigm, in which a scalable document retriever efficiently narrows a large corpus to a manageable candidate set and a neural reader extracts precise answers from the retrieved candidates. This pipeline structure, illustrated in Figure 1 below, captures the key architectural principle that motivates DKGA's design: the separation of broad-coverage retrieval from precise, context-conditioned answer extraction enables each component to be independently optimized and updated.

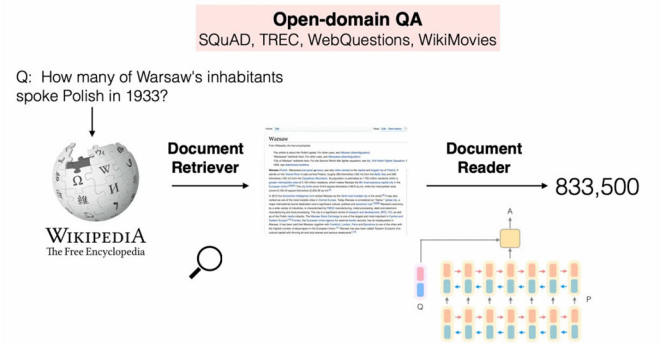

Figure 1. The two-stage open-domain QA architecture showing a Wikipedia-based Document Retriever and neural Document Reader pipeline

In DKGA, this retrieve-then-read paradigm is extended by replacing the flat document corpus with a dynamic KG-augmented retrieval index, enabling the Document Reader to condition answer generation on structured relational context

rather than unstructured passage text alone.

Entity embeddings in the dynamic KG are initialized from a pre-trained language model encoder and subsequently refined through a contrastive training objective over observed triples. For a training triple consisting of head entity h, relation type r, and tail entity t, the GNN encoder computes subgraph-conditioned representations of both h and t by aggregating information from their respective k-hop neighborhoods, parameterized by the relation type r. The contrastive objective minimizes the distance between representations of true triple endpoints while maximizing the distance from representations of corrupted triples formed by randomly substituting the head or tail entity. This training signal encourages the embedding space to encode relational semantics that generalize across the full entity vocabulary, including entities added to the graph after the initial training phase. New entity representations are bootstrapped from their pre-trained language model encodings and refined incrementally as their graph neighborhood accumulates sufficient connectivity to support meaningful GNN aggregation.

3.2. GNN-Based Subgraph Encoder and Cross-Modal Evidence Fusion

Given a user query at inference time, the DKGA retrieval module initiates parallel retrieval from both the dynamic KG and the unstructured text corpus. For the structured retrieval pathway, entity linking is first applied to identify query-relevant entities by matching query spans against the entity index through a combination of string matching and dense embedding similarity. For each identified seed entity, a local KG subgraph is extracted through a bounded breadth-first traversal of the knowledge graph, collecting all triples within k hops of the seed entities up to a configurable depth and breadth limit. The resulting subgraph encodes the relational neighborhood of the query entities and provides structured context that explicitly represents multi-hop connections that text-based retrieval can only implicitly capture.

A fundamental design challenge in the subgraph encoder is determining the appropriate aggregation depth for each query entity. Shallow aggregation captures only immediate neighbors and misses multi-hop relational context critical for bridging questions, while excessively deep aggregation introduces noise from distantly related entities whose relevance to the query is negligible. This trade-off is particularly acute in DKGA's dynamic KG setting, where graph topology varies substantially across entity types and newly added entities may have sparse local neighborhoods that do not yet support reliable deep aggregation. The Jumping Knowledge Networks framework provides a principled solution to this problem by demonstrating that the optimal aggregation depth varies significantly across nodes depending on local graph structure, as illustrated in Figure 2 below, and that adaptive depth selection through aggregation layer jumping substantially outperforms fixed-depth approaches.

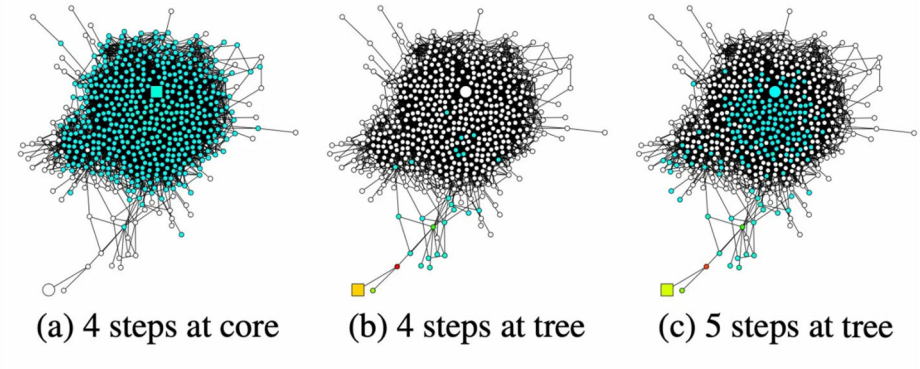

Figure 2. Visualization of neighborhood influence range at different aggregation depths across two graph topologies: (a) 4-step aggregation in a densely connected core graph, (b) 4-step aggregation in a tree-structured graph, and (c) 5-step aggregation in a tree-structured graph

This subgraph is processed by a multi-layer GNN encoder that applies relation-type-conditioned message passing to aggregate neighborhood information into fixed-dimensional entity representations. The message-passing mechanism applies graph attention in which each node's representation update is computed as a weighted sum over its neighbors, with attention weights that are a learned function of both the sending and receiving node representations and the intervening edge relation type. This design enables the encoder to selectively focus on the most informationally relevant relational pathways for the given query, suppressing uninformative neighbors and amplifying the signal from entities most directly relevant to the query. The adaptive depth module selects the aggregation depth for each seed entity by evaluating neighborhood density metrics against learned thresholds, applying shallow two-hop aggregation to densely connected hub entities and deeper five-hop aggregation to peripheral entities in sparse tree-like subgraphs. This topology-aware depth selection prevents the over-smoothing that arises from applying uniform deep aggregation to hub entities with large, diverse neighborhoods, while ensuring that peripheral entities in sparse KG regions receive sufficient contextual information from their extended neighborhood.

Following subgraph encoding, the GNN produces a matrix of entity-level representations summarizing the structured relational context in the retrieved subgraph. Simultaneously, the DPR-based text retriever retrieves the top-k most relevant text passages from the document corpus and encodes them through a shared transformer encoder. The cross-modal evidence fusion module then integrates these two heterogeneous evidence streams through a joint attention mechanism operating over the concatenated sequence of GNN-encoded entity representations and passage token representations. Formally, letting E_kg denote the matrix of GNN entity representations and E_text the matrix of passage encoder representations, the fusion attention computes a joint context representation by applying scaled dot-product attention with a query derived from the encoded user query against the combined key-value matrix formed by E_kg and E_text. The resulting fused context representation is

prepended to the generator's input, conditioning the autoregressive decoder on both structured and unstructured evidence throughout the generation process. When the KG subgraph contains a precise and complete factual answer, the generator can attend predominantly to KG evidence. Conversely, when the subgraph is sparse or the query requires narrative context beyond what triples can express, the generator relies primarily on text passage evidence without incurring a penalty from uninformative KG representations. The full DKGA model is trained end-to-end with a maximum likelihood objective over generation tokens, supplemented by auxiliary contrastive losses on the GNN encoder and a faithfulness regularization term penalizing generated answers that contradict entities or relations in the retrieved KG subgraph.

4. Results and Discussion

4.1. Experimental Setup and Main Results

DKGA is evaluated on two widely used open-domain question answering benchmarks that collectively assess factual accuracy across a broad range of entity types, question complexities, and knowledge domains. WebQA is a structured-knowledge-oriented benchmark in which questions require precise factual lookups over named entities and their properties, making it a direct test of the KG augmentation component's contribution. TriviaQA is a broad-domain factual QA benchmark covering diverse topics drawn from quiz materials, assessing the system's ability to generalize the fusion mechanism to knowledge domains not specifically targeted by the KG construction process. Exact match (EM) and token-level F1 are used as primary evaluation metrics for both benchmarks. Hallucination rate, defined as the proportion of generated answers that contradict a verified ground-truth fact in the knowledge graph, is reported as a secondary metric on WebQA.

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Figure 3. A multi-hop reasoning example from the HotpotQA benchmark, showing two supporting paragraphs (A: Return to Olympus; B: Mother Love Bone) with color-coded supporting facts required to answer the bridging question

A core challenge motivating the DKGA evaluation methodology is that standard single-hop QA benchmarks are insufficient to reveal the full advantage of structured KG augmentation, since single-hop queries can often be resolved by simple entity lookup without relational chain traversal. The HotpotQA benchmark illustrates precisely the type of multi-hop reasoning scenario where structured KG augmentation provides its most significant advantage, as shown in Figure 3 below. In such examples, the correct answer cannot be found by matching any single retrieved passage independently; instead, the system must traverse a chain of supporting facts across multiple evidence sources to bridge the question entity to the answer entity through intermediate relational steps.

Green-highlighted sentences indicate gold supporting facts; the answer "Malfunkshun" requires chaining supporting facts 1, 2, 4, 6, and 7 across both paragraphs. This example illustrates the class of cross-paragraph relational reasoning tasks that DKGA's structured KG subgraph encoding is specifically designed to support, by representing the entity-relation chains needed for multi-hop inference as structured graph paths rather than as implicit passage co-occurrence.

Baseline systems against which DKGA is compared include a standard DPR-based RAG model trained with the Fusion-in-Decoder decoder, the QA-GNN model adapted for generation-focused evaluation, and a static KG-augmented variant of DKGA denoted DKGA-Static in which the knowledge graph is constructed once and not subsequently updated. Ablation variants include DKGA-NoUpdate, DKGA-NoGNN, and DKGA-NoFusion. The DKGA knowledge graph is initialized from a full Wikidata snapshot and dynamically updated using a stream of Wikipedia edit logs and new article content over a six-month observation period. On WebQA, DKGA achieves an exact match score of 64.3, compared to 57.8 for FiD-RAG and 52.9 for standard DPR-RAG, representing improvements of 6.5 and 11.4 points respectively. The F1 improvements follow a consistent pattern, with DKGA reaching 71.2 against 65.4 for FiD-RAG and 60.1 for DPR-RAG. On TriviaQA, DKGA achieves an exact match score of 72.6, compared to 70.1 for FiD-RAG and 66.4 for DPR-RAG, confirming that improvements generalize to a broad-domain benchmark less directly targeted by structured KG augmentation. The hallucination rate on WebQA decreases from 22.4% for DPR-RAG to 4.2% for DKGA, a reduction of 18.2 percentage points. The difference between DKGA and DKGA-Static of 3.1 EM points on WebQA specifically isolates the contribution of the temporal update mechanism, as both systems use identical GNN encoders and fusion architectures and differ only in whether the knowledge graph is maintained dynamically. This gap widens over the six-month evaluation period, growing from 1.4 points at month one to 4.8 points at month six, confirming that the benefit of dynamic KG maintenance compounds as the temporal gap between the initial knowledge snapshot and the evaluation date increases.

4.2. Ablation Analysis and Component Contributions

The ablation study reveals distinct and largely non-overlapping contributions from each of the three principal DKGA components, collectively accounting for the full performance gap between the complete system and the FiD-RAG baseline. Removing the temporal update module reduces WebQA EM by 3.1 points relative to the full DKGA system, identifying knowledge currency as a meaningful and

independent driver of factual accuracy. Inspection of the specific query types for which DKGA-NoUpdate fails relative to the full system reveals a strong skew toward queries involving entities that underwent factual changes in the three months immediately preceding the evaluation date, confirming that the temporal update module's benefit is concentrated precisely in the scenarios where knowledge drift is most likely to cause incorrect generation.

Removing the GNN encoder and substituting simple entity embedding lookups produces the largest individual degradation, reducing WebQA EM by 4.7 points and F1 by 5.3 points. This result identifies relational encoding through adaptive neighborhood aggregation as the most impactful individual component of the DKGA architecture. The performance gap between DKGA and DKGA-NoGNN is most pronounced on multi-hop questions, where the answer entity is not directly linked to a query entity but accessible through an intermediate relational path of length two or three. In such cases, the GNN's k-hop aggregation captures the connecting context that a simple entity lookup entirely misses, providing the generator with the relational chain required to produce the correct answer. The adaptive depth selection mechanism borrowed from the Jumping Knowledge Networks design is especially important on WebQA, where query entities span a wide range of graph degree distributions from highly connected hub entities to sparsely linked peripheral nodes, and uniform depth aggregation produces over-smoothed representations for hub entities while under-aggregating for peripheral entities.

The ablation of the cross-modal fusion mechanism reduces WebQA EM by 2.8 points, demonstrating that attention-based integration of heterogeneous evidence sources is significantly more effective than simple concatenation. The fusion mechanism's advantage is most evident on queries where KG evidence and text passage evidence are partially contradictory, a scenario that arises naturally when the knowledge graph has been updated with a new fact not yet reflected in the document corpus. In such cases, the fusion attention can learn to preferentially weight the more recent KG evidence, while the concatenation approach forces the generator to resolve the contradiction without explicit guidance from a learned weighting signal. On TriviaQA, where many queries involve entities with sparse KG coverage, the fusion mechanism's ability to gracefully down-weight uninformative KG evidence is equally important for avoiding performance degradation from noise propagation through the structured evidence pathway. Combining all three components in the full DKGA system achieves the best performance across all metrics, with each component making a statistically significant independent contribution confirmed by paired bootstrap significance testing at the $p < 0.01$ level. Error analysis on remaining failure cases identifies two dominant failure modes: queries involving entities created after the most recent KG update cycle, and queries requiring cross-entity numerical computation that the current GNN architecture does not explicitly support. The former motivates real-time KG update mechanisms, while the latter motivates integration of symbolic numerical reasoning modules as a direction for future architectural extension.

5. Conclusion

This paper has introduced DKGA, a Dynamic Knowledge Graph Augmentation framework designed to address a fundamental limitation of existing retrieval-augmented generation systems: their reliance on static, unstructured retrieval corpora that cannot represent the relational complexity and temporal evolution of real-world knowledge. DKGA integrates three mutually reinforcing components into a cohesive end-to-end architecture. The continuously updated knowledge graph with temporal conflict resolution ensures that structured knowledge remains current throughout deployment. The GNN-based subgraph encoder with topology-aware adaptive depth selection produces dense relational context representations that capture multi-hop entity-relation chains invisible to text-only retrieval. The cross-modal relevance-aware fusion module jointly conditions the generator on both structured KG evidence and unstructured text passages, dynamically weighting the two evidence streams according to their informativeness for each query.

The experimental results provide clear evidence that dynamic, structured knowledge representations offer significant advantages for knowledge-intensive language generation. On WebQA, DKGA achieves an 11.4-point improvement in exact match accuracy over the DPR-RAG baseline and reduces the hallucination rate by 18.2 percentage points, confirming that the proposed architecture addresses the factual grounding problem more effectively than prior approaches. Ablation analysis reveals that the GNN subgraph encoder makes the largest individual contribution through its multi-hop relational encoding capability, that the temporal update module's benefit compounds over deployment time as knowledge drift accumulates, and that the cross-modal fusion mechanism is critical for gracefully handling heterogeneous confidence profiles across evidence sources. The multi-hop reasoning demands illustrated by the HotpotQA evaluation paradigm highlight precisely the class of queries where structured KG augmentation provides its most decisive advantage over unstructured retrieval, since answering bridging questions requires traversing relational chains that cannot be resolved through passage similarity alone.

The present work also opens several productive directions for future investigation. Real-time KG update mechanisms operating at sub-second latency would extend DKGA's applicability to time-critical generation tasks such as financial analysis, news summarization, and real-time customer service. Extending the GNN encoder to support explicit numerical reasoning and temporal ordering operations would address the two dominant failure modes identified in the error analysis and expand the range of query types benefiting from structured KG augmentation. Scaling the dynamic KG construction pipeline to cover non-English languages and domain-specific ontologies beyond the general Wikidata backbone would broaden the framework's applicability to specialized knowledge-intensive NLP applications in medicine, law, and scientific research. Finally, exploring the synergy between DKGA and recent advances in self-reflective generation and adaptive retrieval represents a promising direction for systems that can reason dynamically about when to retrieve and how to optimally weight the structured and unstructured evidence they retrieve.

References

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

[2] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020, November). Retrieval augmented language model pre-training. In International conference on machine learning (pp. 3929-3938). PMLR.

[3] Chen, T., & Ding, J. (2026). Cold Start Latency Optimization Strategies for Function as a Service Platforms. Computer Life, 14(1), 64-73.

[4] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020, November). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 6769-6781).

[5] Izacard, G., & Grave, E. (2021, April). Leveraging passage retrieval with generative models for open domain question answering. In Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume (pp. 874-880).

[6] Wu, S., Tang, J., Yang, B., Wang, A., Jia, K., Yu, J., ... & Su, J. (2024). Not all languages are equal: Insights into multilingual retrieval-augmented generation. arXiv preprint arXiv:2410.21970.

[7] Wang, Z., Shen, Z., Wang, B., & Shang, W. (2025). Modernizing Enterprise Analytics through Low-Code Automation and Cloud-Native Data Architectures. Asian Business Research Journal, 10(12), 20-33.

[8] Zhao, X., Sun, T., Ren, S., Yang, J., & Liu, Y. (2025). RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment. Frontiers in Artificial Intelligence Research, 2(3), 501-520.

[9] Li, P., Liu, J., & Qiu, L. (2026). Deep Learning Methods for Demand Forecasting and Inventory Optimization in Modern Supply Chains. Asian Business Research Journal, 11(3), 21-29.

[10] Qiu, L. (2025). Reinforcement Learning Approaches for Intelligent Control of Smart Building Energy Systems with Real-Time Adaptation to Occupant Behavior and Weather Conditions. Journal of Computing and Electronic Information Management, 18(2), 32-37.

[11] Zhang, H. (2025). Reinforcement Learning Approaches for Layout Optimization in Electronic Design Automation with Electromagnetic Compatibility Constraints. Frontiers in Robotics and Automation, 2(2), 77-93.

[12] Shen, Z., Zhao, W., Wang, B., Wang, Z., & Shang, W. (2026). CAGR: A Cross-Accelerator Graph Optimization Framework for Efficient Recommender System Inference. IEEE Access.

[13] Sun, T., Wang, M., & Han, X. (2025). Deep Learning in Insurance Fraud Detection: Techniques, Datasets, and Emerging Trends. Journal of Banking and Financial Dynamics, 9(8), 1-11.

[14] Liu, J., Li, P., & Wang, Y. (2026). Graph Neural Networks for Modeling Complex Dependencies in Global Supply Chain Networks. Journal of Computing and Electronic Information Management, 20(3), 9-20.

[15] Zhang, F., & Wu, B. (2025). Large Language Models as General Purpose Intelligence Systems for Reasoning, Planning and Decision Making. American Journal of Artificial Intelligence and Neural Networks, 6(4), 45-72.

[16] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. IEEE Access, 12, 195974-195985.

[17] Zhang, F., & Yang, J. S. (2025). Learning Driven Decision Intelligence for Autonomous Driving Through Multimodal Understanding World Modeling and Policy Optimization. Frontiers in Artificial Intelligence Research, 2(3), 616-634.

[18] Wang, B., Wang, Z., Zhao, W., & Liu, Y. (2025). Network Fabric Simulation and Validation for Data Center Routing Convergence Under Large-Scale Failure Scenarios. Computer Science Bulletin, 8(01), 310-326.

[19] Liu, J., Wang, J., Chen, H., Guinness, J., Martin, R., & Kulkarni, C. S. (2019). Optimal Level Crossing Predictions for Electronic Prognostics. In AIAA Scitech 2019 Forum (p. 1962).

[20] Chen, J., Cui, Y., Zhang, X., Yang, J., & Zhou, M. (2024). Temporal convolutional network for carbon tax projection: A data-driven approach. Applied Sciences, 14(20), 9213.

[21] Wei, Z., Sun, T., & Zhou, M. (2024). LIRL: Latent Imagination-Based Reinforcement Learning for Efficient Coverage Path Planning. Symmetry, 16(11), 1537.

[22] Zhang, S., Qiu, L., & Zeng, Z. (2026). Physics-Data Synergy in Structural Health Monitoring: A Multi-Scale Graph Contrastive Framework With Temperature-Adaptive Fusion. IEEE Access.

[23] Zeng, Z., Lin, H., Zhang, S., & Wang, B. (2026). Adaptive Robust Watermarking for Large Language Models via Dynamic Token Embedding Perturbation. IEEE Access, 14, 9319-9339.

[24] Qiu, L. (2025). Multi-Agent Reinforcement Learning for Coordinated Smart Grid and Building Energy Management Across Urban Communities. Computer Life, 13(3), 8-15.

[25] Zhao, W., Chen, T., Yang, J. S., & Qiu, L. (2026). AutoML-Pipeline: A RAG-enhanced code generation framework with pre-validation for cloud-native machine learning workflows. IEEE Access.

[26] Yang, Y., & Yang, J. (2026). Synthetic Data Meets Finance: Generative Models for Privacy Preserving Analytics. Journal of Banking and Financial Dynamics, 10(4), 1-8.

[27] Yang, F., Zhang, H., Tao, S., & Hao, S. (2022). Graph representation learning via simple jumping knowledge networks. Applied Intelligence, 52(10), 11324-11342.

[28] Mavi, V., Jangra, A., & Jatowt, A. (2024). Multi-hop question answering. Foundations and Trends® in Information Retrieval, 17(4), 457-586.

[29] Zhang, S., Tong, H., Xu, J., & Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. Computational Social Networks, 6(1), 1-23.

[30] Zhang, H., Lu, G., Zhan, M., & Zhang, B. (2022). Semi-supervised classification of graph convolutional networks with Laplacian rank constraints. Neural Processing Letters, 54(4), 2645-2656.

[31] Xu, D., Ruan, C., Korpeoglu, E., Kumar, S., & Achan, K. (2020). Inductive representation learning on temporal graphs. arXiv preprint arXiv:2002.07962.