Overcoming the Impedance Mismatch: A Theoretical Roadmap for Fusing Foundation Models and Knowledge Graphs

Sahil Rajesh Dhayalkar
Arizona State University
sdhayalk@asu.edu

Abstract

Modern artificial intelligence remains fundamentally divided between the continuous, probabilistic spaces of Foundation Models and the discrete, deterministic structures of Knowledge Graphs. While Retrieval-Augmented Generation (RAG) attempts to connect them by serializing graph data into text, we argue this lexical bridging is merely a superficial patch. In this paper, we formalize the underlying structural and geometric friction as the *Impedance Mismatch*. By categorizing current neuro-symbolic integration strategies into a three-tiered hierarchy, we demonstrate that neither surface-level prompt injection nor continuous representation alignment can preserve the strict logical motifs required for reliable multi-hop reasoning. We define the specific mathematical limits, such as the Lexical Bottleneck and Topological Collapse, that show current architectures will eventually hallucinate or conflate semantic nodes. To achieve true semantic fusion, we propose a rigorous theoretical roadmap. We advocate for natively internalizing discrete symbolic structures through Structured Residual Streams, utilizing Vector Symbolic Architectures for latent sub-graph injection, and performing model updates via Orthogonal Subspace Editing. This actionable framework paves the way for models that seamlessly fuse the precision of symbolic logic with the expressivity of parametric memory.

1 Introduction

The architecture of modern artificial intelligence remains fundamentally divided by two distinct paradigms of knowledge representation. On one hand, the subsymbolic paradigm relies on the distributed, continuous representation spaces of Foundation Models, where transformer-based large language models (Vaswani et al., 2017) represent vast amounts of probabilistic world knowledge during pre-training (Brown et al., 2020; Touvron et al., 2023; OpenAI et al., 2024). On the other hand, classical symbolic artificial intelligence utilizes discrete, structured formalisms like Knowledge Graphs to explicitly model declarative knowledge as rigid relational structures (Hogan et al., 2021; Ji et al., 2022). These symbolic frameworks inherently provide the explicit semantics, rigorous compositional structure, and strong mathematical guarantees regarding constraint satisfaction that standard neural architectures natively lack. Bridging this divide is recognized as the next step for Artificial General Intelligence (AGI) (Pan et al., 2024; Luo et al., 2025a).

As foundational models are deployed in high-stakes, knowledge-intensive environments, the need to ground their parametric memory in reliable and up-to-date factual repositories has become critical (Xu et al., 2025; Ma et al., 2025). The prevailing industrial solution is Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Guu et al., 2020; Gao et al., 2024). Current RAG methodologies typically attempt to bridge this gap by serializing knowledge graph subgraphs into natural language strings and injecting them directly into the context window of the model (Edge et al., 2025; Xu et al., 2024). However, we argue that this bridging strategy serves as a superficial patch rather than a mathematical structural solution. Treating the challenge of knowledge integration as mere text retrieval ignores the structural and geometric friction between discrete symbolic edges and continuous parameter spaces (Bian, 2025; Jin et al., 2024).

In this paper, we formalize this structural friction as the *Impedance Mismatch* of neuro-symbolic knowledge integration. Borrowing a foundational concept from object-relational database theory, we define the impedance mismatch as the mathematical degradation that occurs when deterministic graph-structured knowledge bases are artificially mapped into probabilistic self-attention-driven latent spaces (Bian, 2025). Foundational models perceive the world probabilistically through dense vec-

Proceedings of the 4th Workshop on Towards Knowledgeable Foundation Models (KnowFM 2026), pages 78–89

tor similarities, whereas databases and knowledge graphs require strict deterministic algorithmic manipulation. When large language models attempt to process standard knowledge graph structures, they struggle against their own continuous training priors (Jin et al., 2024). This conflict directly results in information loss driven by tokenization mismatches between LLM text encoders and discrete knowledge graph embeddings (Bian, 2025; Pan et al., 2024). Furthermore, converting a rigid relational tuple into a linear sequence of tokens fails to preserve the relational geometry required for multi-hop logical reasoning, directly causing high non-retrieval rates, disconnected subgraphs, and hallucinations (Luo et al., 2025b; Kim et al., 2025; Ma et al., 2025; Edge et al., 2025).

To advance beyond the limitations of text-based retrieval frameworks and achieve true semantic fusion between foundational models and knowledge graphs, we attempt to provide a rigorous theoretical foundation. Our contributions are:

- A Hierarchy of Integration Strategies: We propose a comprehensive hierarchy of integration strategies that categorizes current methodologies from lexical injection to architectural embeddings, highlighting the theoretical capacity limits of each paradigm (Ma et al., 2025; Jin et al., 2024).

- Identification of Core Bottlenecks: We define three bottlenecks preventing true neuro-symbolic fusion, specifically detailing the saturation limits of differentiable logic (van Krieken et al., 2022a), the structural and geometric interference of continuous memory, and the fundamental asymmetry of symbol grounding (Harnad, 1990; Ji et al., 2022).

- A Roadmap for the Knowledge Lifecycle: We chart a theoretical roadmap spanning the complete knowledge lifecycle of emergence, injection, and updating (Dhayalkar, 2025b). We propose mechanisms like latent subgraph injection and orthogonal subspace editing to resolve the impedance mismatch directly within the transformer architecture, paving the way for verifiable compositional generalization (Pan et al., 2024; Luo et al., 2025a).

Hence, we discuss that building knowledgeable foundation models requires moving beyond the assumption that continuous weights can seamlessly absorb discrete facts without explicit, mathematically grounded architectural mediation (Zhu et al., 2025; Pan et al., 2024).

2 The Anatomy of the Impedance Mismatch

To understand why simple text-based retrieval fails to achieve true semantic fusion, we must establish the differences between symbolic graphs and continuous vector spaces. The core technical challenge of integration lies in reconciling the continuous, statistical nature of neural networks with the discrete, logical nature of symbolic systems (d'Avila Garcez et al., 2019; Ji et al., 2022). We categorize this impedance mismatch across three structural dimensions: relational architecture, logical certainty, and memory editability.

2.1 Formalizing the Impedance Mismatch

To ground the impedance mismatch, we must formalize the structural degradation that occurs when mapping discrete relational architectures into continuous latent spaces (Bian, 2025).

Let a Knowledge Graph be defined as a discrete topological space $\mathcal{K} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of entity vertices and $\mathcal{E}$ represents the set of relational edges. This space is equipped with a shortest-path metric $d_{\mathcal{K}}(v_i, v_j)$ that calculates the discrete logical distance between two entities $v_i, v_j \in \mathcal{V}$. Conversely, let the Foundation Model's latent space be a continuous metric space $\mathcal{M} \subseteq \mathbb{R}^h$, where h denotes the dimensionality of the dense vectors, equipped with a geometric distance function $d_{\mathcal{M}}$. Any integration strategy requires a representation mapping function $f : \mathcal{V} \to \mathcal{M}$.

According to the principles of metric embedding theory, mapping an arbitrary discrete graph into a continuous vector space guarantees a strictly positive structural distortion. We formally define the Impedance Mismatch, denoted as $\mathcal{I}$, as the unavoidable mathematical lower bound of this distortion:

$$\mathcal{I} = \inf_{f} \left(\sup_{u \neq v} \frac{d_{\mathcal{M}}(f(u), f(v))}{d_{\mathcal{K}}(u, v)} \times \sup_{u \neq v} \frac{d_{\mathcal{K}}(u, v)}{d_{\mathcal{M}}(f(u), f(v))} \right)$$

where $\inf_f$ denotes the infimum (greatest lower bound) over all possible mapping functions f, and $\sup_{u \neq v}$ denotes the supremum (least upper bound) over all distinct pairs of entities $u, v \in \mathcal{V}$. In a

purely discrete, deterministic system, $\mathcal{I} = 1$, representing perfect structural isometry. However, for dense transformer representations, $\mathcal{I} \gg 1$. This formula shows that continuous spaces cannot faithfully preserve complex graph motifs, such as closed cycles and hierarchical trees, without warping the distances between nodes (Jin et al., 2024). Furthermore, this mismatch manifests as a compounding error during relational composition. In a discrete graph, navigating from a source entity v_1 to a target entity v_3 via sequential relations r_1 and r_2 is a deterministic algebraic composition, yielding an exact target node. In a foundation model, this multi-hop relation is approximated geometrically via sequential self-attention blocks. If $A^{(l)}$ represents the attention matrix at layer l, and L represents the total number of attention layers, the continuous approximation introduces an error term ϵ:

$$\epsilon = \left\| f(v_3) - \prod_{l=1}^{L} A^{(l)} f(v_1) \right\|$$

As the number of logical hops increases, the continuous approximation error ϵ compounds multiplicatively. This formalizes exactly why text-based retrieval frameworks fail at multi-hop logical reasoning (Luo et al., 2025b; Kim et al., 2025): the continuous representation natively lacks the closed algebraic properties required to keep ϵ at zero.

2.2 Structural versus Geometric Relations

In a knowledge graph, knowledge is defined structurally. A relation between a subject entity v_s and an object entity v_o via a predicate r is represented as an explicit, discrete edge $(v_s, r, v_o) \in \mathcal{E}$, where $\mathcal{E}$ is the set of all edges in the graph (Hogan et al., 2021). Retrieving a fact or executing a multi-hop logical query relies on exact graph traversal. The expressive power of such representations depends heavily on the discrete structural motifs used to capture interactions.

Conversely, Foundation Models operate in continuous, high-dimensional vector spaces where internal states are represented by dense tensors (Brown et al., 2020; Touvron et al., 2023). Relations are not explicit edges but are instead approximated geometrically through implicit affine transformations and attention-weighted sums. While a knowledge graph queries adjacency via an indicator function or boolean matrix multiplication, a transformer layer models a relation by computing a soft self-attention distribution (Vaswani et al., 2017):

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V$$

In this geometric space, the relational edge between two concepts is a dense similarity scalar in the attention matrix. This continuous perception struggles to preserve the strict structural constraints required for reliable, multi-step symbolic reasoning (Pan et al., 2024; Jin et al., 2024). When discrete graph architecture is forced into this continuous geometry, the crisp boundaries of symbolic motifs inevitably blur. This geometric blurring directly leads to hallucinated edges, invalid logical hops, and a degradation of verifiable inference (Luo et al., 2025b,a; Edge et al., 2025).

2.3 Certainty versus Probability

The second dimension of the mismatch concerns the truth representation of the encoded knowledge. Knowledge graphs are explicitly built on deterministic logic. An edge either exists or it does not, providing definitive, discrete representations of facts. This structural rigidity makes them suitable for precise querying and explainable, rule-based reasoning (Hogan et al., 2021; Ji et al., 2022).

However, foundational models are fundamentally probabilistic engines trained to minimize cross-entropy loss over token distributions to learn statistical regularities of language (OpenAI et al., 2024). Their internal representation of a fact is inherently statistical and highly contextual. Real-world knowledge is thus modeled not as a binary truth but as a continuous probability density. Merging these two paradigms can cause a structural collapse (Pan et al., 2024). Either the definitive certainty of the knowledge graph must be relaxed into a probabilistic embedding, which mathematically destroys its logical guarantees, or the continuous parameter space of the foundational model must be artificially thresholded to accommodate discrete rules (Luo et al., 2025a; Zhang, 2025). Standard hybrid predictors often assume conditional independence between extracted symbols to bridge this gap. Unfortunately, this assumption limits their ability to model complex interactions and leads to overconfident, miscalibrated predictions (Jin et al., 2024; Luo et al., 2025b).

2.4 The Editability Conflict

Another problem with this impedance mismatch is the difference in how the two systems update

their information. Knowledge graphs are highly dynamic and editable. Updating a fact or correcting an outdated relationship requires a straightforward $O(1)$ operation, executing the direct insertion or deletion of a discrete edge (v_s, r, v_o) (Hogan et al., 2021).

Updating the parametric memory of a foundational model presents a very different theoretical challenge (De Cao et al., 2021; Mitchell et al., 2022). Knowledge in a transformer is heavily interconnected across multiple layers and attention heads via dense vector addition. Modifying a specific fact requires gradient descent or surgical weight perturbations, operations that are inherently unstable for lifelong editing (Meng et al., 2022; Yao et al., 2023). Recent studies in continuous knowledge editing reveal a significant performance decline in both knowledge update efficacy and retention as the number of sequential edits increases (De Cao et al., 2021; Hase et al., 2023). Because the representations are continuous and overlapping, altering the parameters to update one fact often causes degraded interference with adjacent, structurally unrelated knowledge (Meng et al., 2022; Yao et al., 2023; Mitchell et al., 2022). While novel techniques that disentangle and sparsify knowledge representations show promise in alleviating this decline, the fundamental editability conflict remains an unsolved barrier (Pan et al., 2024; Luo et al., 2025a). The distributed nature of the embedding space inherently resists the localized, surgical updates that discrete knowledge graphs effortlessly support.

3 A Hierarchy of Integration Strategies

To analyze neuro-symbolic research, we structure existing literature into a three-tiered maturity model. This hierarchy categorizes integration strategies based on how deeply the discrete knowledge graph penetrates the continuous architecture of the foundational model (Pan et al., 2024; Luo et al., 2025a; Jin et al., 2024). As summarized in Table 1, we can then isolate and expose the specific theoretical limitations inherent to each paradigm.

3.1 Level 1: Lexical and Prompt Injection (Surface-Level)

The most common integration paradigm in industrial and academic settings operates entirely at the surface level. This is mostly realized through Knowledge Graph-Augmented Generation frameworks (Lewis et al., 2020; Gao et al., 2024; Xu et al., 2024; Liu et al., 2025b). In this approach, an external retriever isolates a structurally relevant subgraph, serializes the discrete triples into natural language text, and concatenates this verbalized payload directly into the context window of the foundational model (Lewis et al., 2020; Chen et al., 2025). Recent frameworks have attempted to optimize by retrieving hypothetical reasoning paths to improve evidence selection or by deploying adaptive multi-hop algorithms to reduce the overall token payload (Edge et al., 2025; Liu et al., 2025b).

Critique: While this methodology is accessible and deployable, lexical injection functions as a superficial patch. It inherently suffers from inference latency and remains bottlenecked by context window limitations. Surface-level integration is susceptible to knowledge conflicts, where the model's parametric memory overrides the retrieved context (Luo et al., 2025b; Pan et al., 2024). When the verbalized graph information logically contradicts the pre-trained continuous weights of the foundation model, the architecture frequently discards the prompt in favor of its statistical prior (Mallen et al., 2023; Wang et al., 2025; Luo et al., 2025b). Furthermore, serializing a complex multidimensional graph structure into a flat, linear token stream dismantles the structural motifs required for multi-hop logical deduction (Edge et al., 2025; Bian, 2025).

To formally demonstrate this limitation, we define the mathematical boundary of the Lexical Bottleneck. Let a knowledge subgraph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ possess an average branching factor b and require a logical reasoning depth of k. Let $\mathcal{T}$ represent the token space of a foundational model with a maximum context window length L. Assuming a uniform average branching factor b, the number of distinct reasoning paths of length k diverging from a source entity is b^k. The total number of elements required to fully represent this reasoning subgraph scales geometrically as $\mathcal{O}(b^k)$.

If $c \geq 1$ is the minimum number of tokens required to serialize a single graph element, the minimum token length to represent the subgraph is $c \cdot \mathcal{O}(b^k)$. By the Pigeonhole Principle, if this required length exceeds the fixed capacity L, any deterministic serialization function must truncate information. In classical logic, removing a single premise from a multi-hop chain invalidates the entire deductive path. Consequently, as the reasoning depth k scales, preserving the complete set of relational paths becomes mathematically impossible

without unbounded information loss.

3.2 Level 2: Representation Alignment (Embedding-Level)

To bypass the tokenization bottlenecks of text verbalization, the second tier of integration attempts to align the representations of the knowledge graph and the foundational model within a shared latent mathematical space. Methodologies typically employ Graph Neural Networks or sophisticated translation-based embedding techniques to encode the relational architecture of the discrete graph into dense continuous vectors (Bordes et al., 2013; Kipf and Welling, 2017; Jin et al., 2024; Yasunaga et al., 2021). These graph embeddings are then fused, concatenated, or aligned via multi-task contrastive learning objectives with the native text embeddings of the foundational model during an explicit forward pass or intermediate fine-tuning stage (Liu et al., 2025a; Luo et al., 2025a; Zhang, 2025).

Critique: Embedding-level alignment represents a significant step forward, yet it introduces a representational gap (Pan et al., 2024). Forcing a strict discrete graph into a continuous text embedding space necessitates a mathematical projection that degrades the strict relational properties of the original symbolic graph (Liu et al., 2025a; Bian, 2025). In this paradigm, the continuous vector space acts as a lossy compression algorithm for discrete logic. The system permanently loses the precise relational boundaries inherent to discrete symbols. Hence, while the foundational model gains broad domain awareness, it remains incapable of executing precise algorithmic graph traversals without hallucinating edges or conflating distinct semantic nodes (Luo et al., 2025b; Kiguchi et al., 2025).

To formalize this representational gap, we define the geometric boundary of Topological Collapse as a direct, bounded consequence of the Impedance Mismatch ($\mathcal{I}$) established in Section 2.1. When mapping the discrete metric space of the graph $\mathcal{K} = (\mathcal{V}, \mathcal{E})$ into the continuous latent space $\mathcal{M}$ via an embedding function f, the structural distortion cannot be arbitrarily minimized.

According to Bourgain's Embedding Theorem, embedding a finite metric space of $|\mathcal{V}|$ points into a Euclidean space inherently introduces a minimum structural distortion mathematically bounded by $\Omega(\log |\mathcal{V}|)$. Therefore, we can formally bound the Impedance Mismatch for Level 2 integrations as $\mathcal{I} \geq \Omega(\log |\mathcal{V}|)$. As the size of the ontology grows, this minimum distortion grows logarithmically. Because a perfect, distance-preserving semantic alignment strictly requires $\mathcal{I} = 1$, achieving zero-distortion integration at the embedding level is mathematically impossible. The continuous vector space natively lacks the geometric capacity to preserve the discrete graph structure, unavoidably forcing distinct semantic nodes to overlap and destroying the boundaries required for precise algorithmic traversals.

3.3 Level 3: Architectural Integration (Attention-Level)

The most advanced frontier of current research involves directly modifying the internal computational routing of the transformer architecture to explicitly accommodate graph structures. Rather than treating the knowledge graph as an external text payload or an aligned input vector, these methodologies inject graph priors directly into the message-passing framework or the self-attention calculations of the model (Luo et al., 2025a; Yasunaga et al., 2021). Recent architectural innovations include Graph-Guided Attention modules that non-invasively rewire the native attention matrices of the foundational model based strictly on knowledge graph adjacency (Zhang, 2025; Zhai et al., 2026). Parallel frameworks utilize cross-attention mechanisms to inject semantic graph prompts dynamically across intermediate hidden layers (Hu et al., 2022).

Critique: While architecturally integrated models exhibit state-of-the-art empirical performance on complex reasoning benchmarks (Jin et al., 2024; Yasunaga et al., 2021), they remain theoretically incomplete. They are computationally expensive to scale. They still treat the knowledge graph as an externalized constraint that must be dynamically consulted rather than functioning as an internalized, native knowledge structure. The fundamental mathematical friction remains unresolved because the neural network is still relying on continuous attention weights to approximate discrete logical routing (Pan et al., 2024; Luo et al., 2025a). Until the underlying transformer architecture natively supports discontinuous structural subspaces within its residual stream, true semantic fusion will remain out of reach (Zhai et al., 2026).

To mathematically formalize this architectural limitation, we define the boundary of Attention Approximation Leakage. In a pure symbolic system, logical routing is executed via a discrete adjacency matrix $A \in \{0, 1\}^{n \times n}$. Architecturally integrated

foundational models attempt to approximate this discrete routing using continuous attention matrices $A_{\text{soft}} \in (0,1)^{n \times n}$.

Because the standard attention mechanism relies on the softmax function, it strictly outputs positive probabilities. Approximating a hard, discrete zero (indicating no relationship) requires infinite negative logits, which is impossible in a stable training regime. Therefore, every non-adjacent node contributes a strictly positive residual leakage error $\delta > 0$ during the message-passing calculation. When the model attempts to execute a multi-hop logical query of depth k, the routing calculation approximates $(A_{\text{soft}})^k$. As k increases, the continuous leakage error δ compounds exponentially, leading to severe representation over-smoothing. The precise signal of the true discrete reasoning path is inevitably drowned out by the accumulated noise of the continuous space, proving that approximating discrete routing with continuous attention weights is mathematically unsustainable for deep logical deduction.

4 Core Bottlenecks Preventing True Fusion

To move past the design limits of current integration strategies and achieve true semantic fusion, the community must address three fundamental bottlenecks. These barriers represent incompatibilities between discrete structural constraints and continuous latent spaces.

4.1 Bottleneck A: The Curse of Differentiable Logic

A prevalent method for injecting discrete logic into continuous models utilizes differentiable logic frameworks, which relax Boolean connectives and quantifiers into continuous operators (Rocktäschel and Riedel, 2017; Evans and Grefenstette, 2018; van Krieken et al., 2022b). Soft relaxations algorithmically map strict truth values to the continuous interval $[0, 1]$ via t-norms, s-norms, and fuzzy aggregation operators (van Krieken et al., 2022b; Manhaeve et al., 2018). However, this mapping introduces an optimization bottleneck. The resulting loss landscapes are non-linear and suffer from acute gradient saturation (Giunchiglia et al., 2022; Wang et al., 2019). Once a logical formula is nearly satisfied, the gradients vanish entirely, prematurely halting the optimization process before true semantic alignment is achieved (van Krieken et al., 2022b; Minervini et al., 2019).

Furthermore, soft truth values break classical logical equivalences. In a discrete knowledge graph, De Morgan's laws and contraposition hold absolute certainty. In a relaxed tensor space, these functionally equivalent symbolic rules often yield entirely divergent optimization paths (Giunchiglia et al., 2022; Wang et al., 2019). This inherent conflict makes robust constraint satisfaction mathematically unstable under stochastic gradient descent. Consequently, researchers are forced to choose between Boolean faithfulness and optimization amenability (van Krieken et al., 2022b; d'Avila Garcez et al., 2019).

4.2 Bottleneck B: Structural and Geometric Interference

The second barrier is structural and geometric interference. In a discrete graph, edges provide perfect relational insulation. Editing the relation between a subject node and an object node has no impact on adjacent graph edges. In a continuous representation space, such absolute geometric isolation is mathematically impossible (Meng et al., 2022; Elhage et al., 2021). When discrete symbolic structures are encoded into high-dimensional vectors, they overlap and blend within the same dense space (Elhage et al., 2021).

Updating parametric memory to modify a specific bound relation inherently warps the local geometry of the embedding representation space (Meng et al., 2022; Hase et al., 2023). As the number of overlapping facts in the residual stream increases, theoretical capacity limits are reached, and knowledge extraction operations inevitably suffer from catastrophic crosstalk (Yao et al., 2023; Zhong et al., 2024). Surgically editing a specific semantic relation can inadvertently alter adjacent, structurally unrelated knowledge (Meng et al., 2022; De Cao et al., 2021; Cohen et al., 2023). The fluid nature of the transformer's residual stream lacks the strict orthogonality required to perfectly insulate discrete variables during continuous updates (Wang et al., 2024). This leads to the logical consistency breaking down entirely under minor parameter perturbations (Cohen et al., 2023; Zhong et al., 2024; Hase et al., 2023).

4.3 Bottleneck C: The Symbol Grounding Asymmetry

The final bottleneck centers on the asymmetry in symbol grounding (Harnad, 1990; Ji et al., 2022).

Integration Level	Mechanism	Formal Mathematical Bottleneck	Asymptotic Failure Mode
Level 1: Surface	Lexical Prompt Injection	Lexical Bottleneck: $\mathcal{O}(b^k) > L$	Context truncation; inability to encode exponential path complexity.
Level 2: Embedding	Latent Vector Alignment	Topological Collapse: $D(f) \geq \Omega(\log \lvert\mathcal{V}\rvert)$	Semantic conflation; distortion of discrete relational boundaries.
Level 3: Architecture	Graph-Guided Attention	Approximation Leakage: Compounding softmax error δ in $(A_{\text{soft}})^k$	Representation over-smoothing; discrete signal drowned in continuous noise.

Table 1: A theoretical taxonomy of neuro-symbolic integration strategies, classified by their fundamental mathematical bottlenecks and asymptotic failure modes during multi-hop reasoning.

Knowledge graphs rely on unique entity identifiers to maintain strict referential integrity across diverse contexts (Hogan et al., 2021). On the other hand, foundational models process information through contextualized, distributed sub-word token representations (Brown et al., 2020; OpenAI et al., 2024).

Aligning abstract, immutable symbols with fluid data patterns remains a major theoretical challenge (Pan et al., 2024, 2023). While prior works attempt to bridge this gap using contrastive alignment or dedicated entity embeddings, these methods assume a static mapping that ignores the dynamically overlapping nature of language models (Pan et al., 2024; Luo et al., 2025a; Zhang, 2025). Natively integrating symbolic knowledge requires a mechanism to dynamically instantiate and bind discrete roles to continuous fillers without losing the strict identity of the original symbol (d'Avila Garcez et al., 2019; Smolensky, 1990). Until this structural asymmetry is mathematically resolved, hybrid models will continue to rely on shallow pattern matching rather than exhibiting true, provable compositional generalization (Lake et al., 2016; Bahdanau et al., 2019; Ruis et al., 2020).

5 A Roadmap for the Knowledge Lifecycle

To resolve the bottlenecks in Section 4 and the impedance mismatch, we build upon the framework established by (Dhayalkar, 2025b) to propose an actionable three-stage knowledge lifecycle roadmap that transcends lexical bridging.

5.1 Emergence (Pre-training): Structured Residual Streams

Current pre-training paradigms rely on unconstrained geometric optimization. This reliance directly causes the structural and geometric interference of factual knowledge observed during complex reasoning tasks (Elhage et al., 2021; Bricken et al., 2023). However, recent breakthroughs in Representation Engineering demonstrate that high-level concepts naturally manifest as stable subspace directions or principal-eigenvector backbones within the transformer's residual stream (Zou et al., 2025; Park et al., 2024). Furthermore, models can natively recover spatial separations that directly map to structured human concept categories (Wang et al., 2023; Li et al., 2023).

To formalize this phenomenon, we propose the architectural development of *Structured Residual Streams*. Rather than allowing facts to overlap arbitrarily across the entire embedding latent space, future architectures should introduce explicit graph-theoretic inductive biases during pre-training (Pan et al., 2024; Luo et al., 2025a). By applying regularization penalties that enforce orthogonal subspaces for distinct knowledge domains, discrete relational structures could emerge natively within the continuous weights. This would equip the model with an inherent, mathematically insulated structure, preventing the catastrophic crosstalk that currently degrades multi-hop reasoning (Frady et al., 2020).

5.2 Injection (Inference): Latent Sub-graph Injection via VSAs

The industry standard of text-based retrieval is limited by tokenization bottlenecks and the high influence of the continuous parametric prior (Mallen et al., 2023; Lewis et al., 2020). To bypass this, we must shift from external lexical prompting to *Latent Sub-graph Injection*. We propose utilizing Vector Symbolic Architectures (VSAs) as the mathematical bridge to achieve this integration natively.

VSAs provide a well-defined algebraic framework using operations like binding, bundling, and permutation to represent complex discrete graph data within unified high-dimensional vector spaces (Kanerva, 2009; Kleyko et al., 2022). VSAs retain

fixed-dimensional vectors that align naturally with the native embeddings of the standard transformer architecture (Smolensky, 1990). By encoding a retrieved knowledge graph subgraph directly into a VSA hypervector, researchers can inject explicit role-filler bindings directly into the intermediate attention layers of the foundation model at inference time (Meng et al., 2022; Kanerva, 2009; Dhayalkar, 2025a). This bypasses the superficial text layer and forces the model to condition its generation on strict, mathematically bound relations rather than probabilistic text prompts.

5.3 Updating (Editing): Orthogonal Subspace Editing

The editability conflict requires a new mathematical approach to model updates. Current continuous knowledge editing regimes suffer from a performance decline in knowledge retention as sequential edits increase (Meng et al., 2022; Mitchell et al., 2022; De Cao et al., 2021). While recent methods have advanced the ability to update long-form knowledge using dynamic weight adjustments, they still grapple with coupling of the continuous vector space (Yao et al., 2023; Zhong et al., 2024).

To guarantee localized factual updates without neighborhood interference, we call for the formalization of *Orthogonal Subspace Editing*. Recent dissections of perturbation weights indicate that disentangled and sparsified knowledge representations can alleviate performance degradation during continuous editing (Hase et al., 2023). Building on this insight, we hypothesize that by projecting targeted factual edits strictly along orthogonal feature directions that do not activate unrelated semantic concepts, we can achieve updates that are mathematically equivalent to localized edge-insertion. This theoretical direction would allow foundational models to be patched dynamically and safely, finally bringing the reliable editability of symbolic knowledge bases to neural parameter spaces (Pan et al., 2024; Luo et al., 2025a; Meng et al., 2022).

6 Conclusion

Continuing to treat knowledge graphs merely as external databases or retrieval dictionaries fundamentally limits the evolutionary trajectory of foundation models. Throughout this paper, we have demonstrated that the current industrial standard of text-based retrieval acts only as a superficial patch over a much deeper structural divide. We defined this divide as the Impedance Mismatch, a mathematical friction that occurs when attempting to force rigid, deterministic graph relational structures into fluid, probabilistic embedding spaces.

By categorizing existing integration attempts into a hierarchy of maturity, we revealed that neither lexical prompt injection nor continuous representation alignment can preserve the strict logical motifs required for reliable, multi-hop reasoning. The true barriers to semantic fusion are not engineering hurdles, but rather deep theoretical bottlenecks. The saturation of differentiable logic, the structural and geometric interference of continuous memory, and the fundamental asymmetry of symbol grounding collectively prevent standard transformer architectures from natively internalizing discrete symbolic structures.

To construct truly knowledgeable foundation models, the research community must move beyond the paradigm of lexical bridging. We must confront the fundamental mathematical friction between discrete certainty and continuous probability directly at the architectural level. By pursuing structured residual streams, latent sub-graph injection via vector-symbolic architectures, and orthogonal subspace editing, we can transition from models that mimic factual recall to systems that genuinely harbor structured, editable knowledge. Resolving this impedance mismatch is the necessary next step in the knowledge lifecycle, enabling a future where the precision of symbolic logic and the expressivity of parametric memory are seamlessly and mathematically fused.

Limitations

While this paper establishes a rigorous mathematical foundation for neuro-symbolic integration, it focuses strictly on formal analysis and does not include empirical experiments. Consequently, our proposed frameworks currently serve as theoretical blueprints. Bridging these formalisms, such as Structured Residual Streams and VSA injection into scalable training regimes, represents a natural next step for empirical research. Additionally, because our models assume perfectly deterministic knowledge graphs, future work must explore how these strict geometric constraints adapt to the noise and contradictions inherent in real-world knowledge bases.

References

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. 2019. Systematic generalization: What is required and can it be learned? *Preprint*, arXiv:1811.12889.

Haonan Bian. 2025. Llm-empowered knowledge graph construction: A survey. *Preprint*, arXiv:2510.20345.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jialin Chen, Houyu Zhang, Seongjun Yun, Alejandro Mottini, Rex Ying, Xiang Song, Vassilis N. Ioannidis, Zheng Li, and Qingjun Cui. 2025. Gril: Knowledge graph retrieval-integrated learning with large language models. *Preprint*, arXiv:2509.16502.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *Preprint*, arXiv:2307.12976.

Artur d'Avila Garcez, Marco Gori, Luis C. Lamb, Luciano Serafini, Michael Spranger, and Son N. Tran. 2019. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Preprint*, arXiv:1905.06088.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sahil Rajesh Dhayalkar. 2025a. Attention as binding: A vector-symbolic perspective on transformer reasoning. *Preprint*, arXiv:2512.14709.

Sahil Rajesh Dhayalkar. 2025b. Neuro-symbolic reasoning: A roadmap of unsolved core questions. *TechRxiv*, 2025(1210).

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Richard Evans and Edward Grefenstette. 2018. Learning explanatory rules from noisy data. *J. Artif. Int. Res.*, 61(1):1–64.

E. Paxon Frady, Denis Kleyko, and Friedrich T. Sommer. 2020. Variable binding for sparse distributed representations: Theory and applications. *Preprint*, arXiv:2009.06734.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Eleonora Giunchiglia, Mihaela Catalina Stoian, and Thomas Lukasiewicz. 2022. Deep learning with logical constraints. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, IJCAI-2022, page 5478–5485. International Joint Conferences on Artificial Intelligence Organization.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *Preprint*, arXiv:2002.08909.

Stevan Harnad. 1990. Harnad, s. (1990). the symbol grounding problem. physica d: Nonlinear phenomena, 42(1-3), 335-346. 42:335–346.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *ACM Computing Surveys*, 54(4):1–37.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240, Dublin, Ireland. Association for Computational Linguistics.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. 2022. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514.

Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2024. Large language models on graphs: A comprehensive survey. *Preprint*, arXiv:2312.02783.

Pentti Kanerva. 2009. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1(2):139–159.

Kanan Kiguchi, Yunhao Tu, Katsuhiro Ajito, Fady Alnajjar, and Kazuyuki Murase. 2025. Multi-modal integration analysis of alzheimer's disease using large language models and knowledge graphs. *IEEE Access*, 13:113718–113735.

Soohyeong Kim, Seok Jun Hwang, JungHyoun Kim, Jeonghyeon Park, and Yong Suk Choi. 2025. ReGraphRAG: Reorganizing fragmented knowledge graphs for multi-perspective retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5426–5443, Suzhou, China. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Denis Kleyko, Mike Davies, Edward Paxon Frady, Pentti Kanerva, Spencer J. Kent, Bruno A. Olshausen, Evgeny Osipov, Jan M. Rabaey, Dmitri A. Rachkovskij, Abbas Rahimi, and Friedrich T. Sommer. 2022. Vector symbolic architectures as a computing framework for emerging hardware. *Proceedings of the IEEE*, 110(10):1538–1571.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2016. Building machines that learn and think like people. *Preprint*, arXiv:1604.00289.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*.

Yu Liu, Yanan Cao, Xixun Lin, Yanmin Shang, Shi Wang, and Shirui Pan. 2025a. Enhancing large language model for knowledge graph completion via structure-aware alignment-tuning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20970–20984, Suzhou, China. Association for Computational Linguistics.

Zhaotai Liu, Harald Sack, and Genet Asefa Gesese. 2025b. Hyp-kgrag: Hypothetical path-based knowledge graph retrieval augmented generation with deepseek. In *RAGE-KG 2025: The Second International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs, co-located with ISWC 2025, November 2–6, 2025, Nara, Japan*, volume 4079 of *CEUR Workshop Proceedings*, pages 45 – 55. CEUR-WS.

Linhao Luo, Carl Yang, Evgeny Kharlamov, and Shirui Pan. 2025a. Integrating large language models and knowledge graphs for next-level agi. *Companion Proceedings of the ACM on Web Conference 2025*.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Yuan-Fang Li, Chen Gong, and Shirui Pan. 2025b. Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 41540–41565. PMLR.

Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. 2025. Large language models meet knowledge graphs for question answering: Synthesis and opportunities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24578–24597, Suzhou, China. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. 2018. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Pasquale Minervini, Matko Bošnjak, Tim Rocktäschel, Sebastian Riedel, and Edward Grefenstette. 2019. Differentiable reasoning on large knowledge bases and natural language. *Preprint*, arXiv:1912.10824.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022. Fast model editing at scale. *Preprint*, arXiv:2110.11309.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. Large language models and knowledge graphs: Opportunities and challenges. *Preprint*, arXiv:2308.06374.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. *Preprint*, arXiv:2311.03658.

Tim Rocktäschel and Sebastian Riedel. 2017. End-to-end differentiable proving. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M Lake. 2020. A benchmark for systematic generalization in grounded language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 19861–19872. Curran Associates, Inc.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1):159–216.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Emile van Krieken, Erman Acar, and Frank van Harmelen. 2022a. Analyzing differentiable fuzzy logic operators. *Artificial Intelligence*, 302:103602.

Emile van Krieken, Erman Acar, and Frank van Harmelen. 2022b. Analyzing differentiable fuzzy logic operators. *Artificial Intelligence*, 302:103602.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Po-Wei Wang, Priya L. Donti, Bryan Wilder, and Zico Kolter. 2019. Satnet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver. *Preprint*, arXiv:1905.12149.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024. Knowledge editing for large language models: A survey. *Preprint*, arXiv:2310.16218.

Xi Wang, Procheta Sen, Ruizhe Li, and Emine Yilmaz. 2025. Adaptive retrieval-augmented generation for conversational systems. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 491–503, Albuquerque, New Mexico. Association for Computational Linguistics.

Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. 2023. Concept algebra for (score-based) text-controlled generative models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ran Xu, Patrick Jiang, Linhao Luo, Cao Xiao, Adam Cross, Shirui Pan, Jimeng Sun, and Carl Yang. 2025. A survey on unifying large language models and knowledge graphs for biomedicine and healthcare. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25)*, pages 6195–6205. PMID: 41858611; PMCID: PMC12995553.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. 2024. Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 2905–2909. ACM.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings*

of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 10222–10240, Singapore. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Songlin Zhai, Guilin Qi, Yue Wang, and Yuan Meng. 2026. Knowledge fusion via bidirectional information aggregation. *Preprint*, arXiv:2507.08704.

Qinggang Zhang. 2025. Enhancing large language models with reliable knowledge graphs. *Preprint*, arXiv:2506.13178.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. 2024. Mquake: Assessing knowledge editing in language models via multi-hop questions. *Preprint*, arXiv:2305.14795.

Zhihui Zhu, Yuqi Tang, Qiang Zhang, and Keyan Ding. 2025. Synergizing large language models and knowledge graphs in science: A survey. In *NeurIPS 2025 AI for Science Workshop*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.