Semantic Similarity is Not Evidence: Verification Diagnostics for LLM Embeddings

Alex Romanova
Graph AI Studio
McLean, USA
sparkling.dataocean@gmail.com

***Abstract*—This work introduces a diagnostic framework to test when LLM-style semantic similarity reflects relational evidence - and when it does not. LLMs produce strong text embeddings, but semantic similarity is not evidence: when correctness depends on interactions, constraints, or multi-step links, text-only similarity can look convincing while disagreeing with what the data supports. We frame this as an LLM data-quality question and measure structural sensitivity by comparing text-only embeddings with relationship-aware embeddings trained on an evidence graph. Using edge recovery, distributional separation, and top-quantile concentration, results show that text-only similarity is often plausible yet weakly grounded in the observed relation signal, while relationship-aware representations align more reliably with observed structure. Rather than prescribing a specific remedy, our framework makes it measurable when text is enough - and when relationship-aware handling is warranted.**

***Index Terms*—large language models, semantic similarity, structural sensitivity, relational structure, evidence graphs, graph neural networks, link prediction, embedding evaluation, representation auditing, data quality**

I. Introduction

Large Language Models (LLMs) have fundamentally changed how we interact with information. They provide a flexible, unified interface to knowledge: interpreting intent, synthesizing heterogeneous sources, and mapping text into dense representations that capture semantic regularities at unprecedented scale. For many tasks - summarization, semantic search, question answering, and content understanding - this paradigm is not only sufficient but remarkably effective. In a wide range of settings, text itself carries the signal, and LLM-based systems perform exactly as intended.

However, not all problems are linguistic in nature. LLMs are fundamentally sequential models optimized for next-token prediction, which makes them exceptionally strong at capturing semantic regularities in text but not natively compelled to preserve relational structure when correctness depends on it. In many real-world systems, validity depends less on what entities say and more on how they are connected - through interaction patterns, constraints, temporal dependencies, or multi-step relationships. As a result, systems driven primarily by semantic similarity can be plausible yet structurally wrong, confusing plausibility with necessity in domains where relationships, not attributes, determine correctness.

A broader point is that LLM limitations are not unique to our setting, nor are they novel observations. Gentner's structure-mapping theory emphasizes that abstraction and analogy succeed when relational structure is preserved, not when surface attributes match [1]. LeCun likewise argues that next-token learning can perform well on language without enforcing the grounded, hierarchical representations needed for robust reasoning and planning [2], [3]. Together, these perspectives suggest that many errors often labeled "hallucination" stem from limited sensitivity to the relational structure that makes an answer correct.

Crucially, relationships do not always matter: in many settings, text-only representations are sufficient and preferable. The issue is that current evaluations rarely test when relational structure changes what is correct; they mostly reward semantic plausibility. In this paper, we study that gap as a diagnostic question: given an observed relation signal over the same items, to what extent does text-only similarity recover that signal, and how does that compare with a relationship-aware representation trained with access to it? We frame this as an LLM data-quality question - focused on measurement and characterization rather than prescribing specific solutions. To do so, we introduce a diagnostic framework that captures structural sensitivity by isolating cases where semantic similarity and relational structure disagree. The result is a clear signal of when a text-only regime is likely reliable, and when relationship-aware handling is warranted.

II. Related Work

Relationships have been central to information systems well before the recent surge in LLM-centric retrieval. The Semantic Web and knowledge graphs made structure explicit for linking, validation, and provenance, and industry experience highlights the practical challenges of building and maintaining such graphs at scale [4]. Geometric deep learning and graph learning then formalized how relational inductive bias can be learned from data [5]. In the early 2020s, text-first transformers and dense embeddings became dominant because they scale and generalize well, but evaluation shifted toward semantic plausibility. We build on this arc by focusing on *measurement*: quantifying when relational structure changes what is correct versus when text-based similarity is sufficient.

Recent 2026 work reinforces that transformer LLMs, while highly capable, still lack native mechanisms for some system-critical operations. DeepSeek-AI proposes *Engram*, a conditional-memory lookup module that offloads frequent

static patterns from expensive neural computation [6]. In a different setting, Agent World Model (AWM) targets instability in language-only simulation for agent training and proposes scalable, code-driven synthetic environments with consistent state transitions [7]. Both papers respond with architectural or infrastructure remedies; complementarily, we contribute diagnostics for measuring when relational structure changes what is correct and when text-based similarity is likely sufficient.

Graph-based recommender methods learn directly from user–item interaction topology, treating behavior as evidence rather than description [8], [9]. In parallel, work on explanation and faithfulness has shown that fluent rationales can be persuasive while failing to reflect the actual basis of a decision [10], [11]. Standard ranking metrics (e.g., nDCG, MRR) measure utility but provide limited diagnostic insight into what representations capture [12].

Taken together, these lines of work motivate the importance of relational structure, graph-based representation learning, and evaluation beyond surface plausibility. However, they leave open a narrower diagnostic question that is the focus of this paper: when does semantic similarity align with an observed relation signal over the same items, and when does it fail to do so? Our contribution is not a new recommender objective or a general theory of semantic similarity, but a diagnostic framework for comparing text-only and relationship-aware embedding spaces against the same relation signal.

III. Methods

Figure 1 summarizes the diagnostic pipeline. We assume entities with textual descriptions and an evidence graph capturing observed relationships (here, user–item interactions). Because our diagnostics operate on item pairs, we induce an item–item relationship graph from shared-user clicks (a co-click projection) and train the relationship-aware pathway with GNN link prediction on this induced graph. We use this induced graph as an observed relation signal over items, not as a claim about user intent or ideal recommendation behavior. Importantly, link prediction is used only as a training objective: we extract the model's pre-final node embeddings as representations for downstream similarity analysis. We then compare the similarity structures of the text-only and relationship-aware spaces using diagnostic metrics to measure when relational evidence changes which item pairs are supported under the observed relation signal.

A. Data and Evidence Graph Construction

For concreteness, we instantiate the evidence graph using a news-style dataset with (i) item text and (ii) observed user interactions. Each item is associated with a short textual description, which serves as input to the text encoder. Interaction logs record which items are clicked by which users, yielding a natural bipartite user-item evidence graph.

Although our diagnostics operate on item-level embeddings, the bipartite evidence graph is essential because it determines which item-item links are supported by the observed interaction signal. Because our diagnostics are defined over item pairs, we induce an item-item relationship graph from shared-user clicks: two items are treated as linked if they are clicked by the same user, yielding a co-click projection. We use this induced graph as an observed relation signal over items, not as a claim about user intent or ideal recommendation behavior. We retain the underlying bipartite evidence for provenance, while using the induced item-item graph to define positive/negative item pairs and to train relationship-aware representations. This shared evidence foundation ensures that both the text-only and relationship-aware embeddings derive from the same relational signal, allowing the diagnostics to isolate whether similarity reflects semantic plausibility or the observed relation signal in this setting.

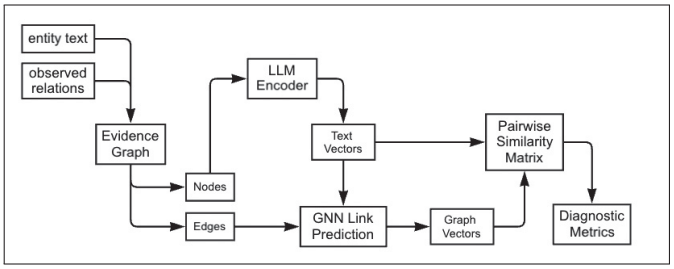

Fig. 1. Diagnostic pipeline. An evidence graph (here, user-item interactions) induces an item-item relationship graph (co-click projection) used to train a relationship-aware pathway via GNN link prediction. We evaluate the resulting text-only and relationship-aware similarity spaces using diagnostic metrics (computed from pre-final embeddings).

B. Text-Only Representation (LLM Baseline)

As a text-only baseline, we encode each news item independently using a pre-trained transformer model from the Hugging Face model repository. Specifically, we use the `all-MiniLM-L6-v2` model, a sentence-transformers architecture designed to map natural language text into a dense semantic vector space. The abstract text associated with each news item is passed through the model to produce a fixed-length embedding in a 384-dimensional space.

This representation captures semantic similarity derived solely from textual content. Each news item is processed independently, without access to user interaction data or relational context. As a result, two items may receive similar embeddings if their abstracts are linguistically or semantically similar, regardless of whether they are clicked by the same users or appear in similar observed interaction contexts.

The resulting text embeddings are used directly for downstream similarity analysis. Pairwise similarities between news items are computed in this embedding space, yielding a text-induced similarity structure that reflects what can be inferred from language alone. This baseline corresponds to a widely adopted modeling paradigm in modern retrieval, recommendation, and semantic matching settings, where large language models are used as standalone semantic encoders. At the same time, by construction, this approach omits the relational evidence encoded in user-item interactions, providing a strong but intentionally incomplete reference point for evaluating the impact of relationships.

C. Relationship-Aware Representation (LLM + GNN)

To incorporate relational evidence, we refine the text-only item embeddings using a graph neural network trained on an induced item-item relationship graph derived from the user-item evidence. We start from the same 384-dimensional text vectors described in Section 3.2, ensuring that any differences in representation arise from introducing relational structure rather than changing textual inputs.

We employ a GraphSAGE-based encoder [8] and train it with a link-prediction objective over the induced co-click graph, where edges represent item-item links supported by the observed interaction signal. Through message passing, each item embedding absorbs neighborhood context, allowing the representation space to reflect relationships supported by the observed relation signal that are weakly expressed or absent in text.

The model is implemented using the Deep Graph Library (DGL) [13], following its standard GraphSAGE link-prediction pipeline. Importantly, link prediction is used only as a training signal: we extract the model's pre-final node embeddings as relationship-aware item vectors and use them for downstream similarity analysis, directly comparable to the text-only baseline. This keeps the comparison focused on representation spaces rather than on link prediction as the primary task outcome.

D. Pairwise Similarity Analysis

To compare text-only and relationship-aware representations, we compute pairwise similarities between news items in each embedding space. This produces two item-item similarity matrices: one from the text-only embeddings and one from the relationship-aware embeddings.

Each similarity matrix defines an implicit neighborhood structure, where an item's nearest neighbors reflect what the representation considers "related." Differences between the two matrices therefore indicate how adding relational information reshapes item neighborhoods beyond what text alone captures.

The analysis is performed over the same item set and the same labeled pair set for both embedding spaces, where positives correspond to induced item–item edges under the observed relation signal and negatives to non-edges. Holding the items, similarity function, and labeled pairs fixed isolates the effect of relational evidence on the induced similarity structure.

E. Diagnostic Metrics

We measure structural sensitivity by asking whether similarity in an embedding space recovers known links under the observed relation signal. Using a fixed set of item pairs labeled as interaction edges (positives) or non-edges (negatives), we compute the same diagnostics on cosine similarity scores for both text-only and relationship-aware embeddings.

We report three complementary metrics:

- **Edge recovery (ROC AUC):** Treat similarity as a score for predicting whether a pair is an interaction edge, and report ROC AUC over the labeled pair set.
- **Distribution separation (Cohen's d):** Compare similarity distributions for edge vs. non-edge pairs and report standardized separation (Cohen's d), alongside simple summaries (mean/median) for interpretability.
- **Top-quantile concentration (Lift@q):** For a quantile q (e.g., top 1%), measure how concentrated edges under the observed relation signal are among the top-q most similar pairs, normalized by the base edge rate (lift).

Together, these diagnostics quantify when semantic similarity aligns with the observed relation signal in this setting and when it does not, in a representation-agnostic way.

IV. Experiments

We evaluate the diagnostic pipeline described in the Methods section (see the overview diagram) on a single day from a news dataset with textual content and observed user interactions. The objective is not to optimize an end-to-end recommender system, but to assess whether incorporating observed relational evidence reshapes similarity structure beyond what is captured by text-only embeddings. All experiments reported in this section use the same representations and diagnostic metrics defined in Sections 3.1-3.5.

A. Dataset and day-level slice

We use the Microsoft News Dataset (MIND), specifically the `MINDsmall` training split, which contains news article metadata and user interaction logs collected from a commercial news recommendation platform [14]. To keep the analysis focused and interpretable, we restrict all experiments to a single day, **2019-11-11**. We use only the two raw tables required by our method: `news.tsv`, which provides article identifiers, categories, and textual content (title + abstract), and `behaviors.tsv`, which records user impressions and click histories. This day-level slice serves as a self-contained snapshot for constructing the observed relation signal and evaluating the resulting representations.

B. Evidence graph construction

From `behaviors.tsv` and `news.tsv` on **2019-11-11**, we build a day-specific bipartite user-item interaction graph that defines the observed relation signal over item pairs. Because our diagnostics are computed on *item pairs*, we induce an item-item *co-click graph* by projecting the bipartite graph onto the item side. For each user, we take the set of items they clicked and add undirected edges between all pairs in that set. The resulting co-click edges define *positive* item pairs under the observed relation signal; pairs of items without a co-click edge are treated as *negative* pairs for evaluation.

We apply a simple day-level filter to ensure users contribute pairwise signal: we keep only users with at least two clicks on that day (users with a single click cannot form co-click pairs). After filtering, the day-level data used in this study is:

- **Users retained:** 10,656

- **Impressions retained:** 22,157
- **Clicks retained:** 37,435
- **Unique news items observed in filtered logs:** 27,945

From this filtered interaction data, the induced co-click graph used for representation learning and diagnostics contains:

- **Graph nodes (news items):** 1,991
- **Graph edges (co-click relationships):** 40,637

Although many items appear in the logs, only the 1,991 items that participate in at least one co-click relationship are included in the induced graph; all representation learning and diagnostics are restricted to this item set.

C. Representations compared

We compare two item representation pathways from Sections 3.2-3.3, each producing a dense vector for every news item. First, we compute **text-only vectors** by encoding each article abstract with the `all-MiniLM-L6-v2` sentence-transformers model. Second, we compute **relationship-aware vectors** using a graph neural network trained in a link prediction setting on the item co-click graph. Although the GNN is trained to predict co-click relationships, we do not use its link predictions directly; instead, we extract the model's **pre-final node embeddings** as item representations, which incorporate both text features and relational context derived from the observed interaction signal.

From both sets of item vectors, we compute cosine similarities for item pairs and evaluate them using the diagnostic metrics in Section 3.5, described next.

D. Reported diagnostics

Edge recovery. Relationship-aware embeddings achieve substantially higher ROC AUC than text-only embeddings (AUC ≈ 0.75 vs. AUC ≈ 0.52), indicating that structural proximity aligns with edges under the observed relation signal (Fig. 2). Text-only similarity remains near chance, consistent with semantic similarity being an incomplete proxy for the observed relation signal in this setting.

Edge vs. non-edge separation. Relationship-aware similarities show clearer distributional separation between edge and non-edge pairs (larger effect size), whereas text-only similarities exhibit weaker separation (Figs. 3-4). This supports the claim that relationship-aware representations encode relational evidence from the observed interaction signal more directly.

Top-quantile concentration. In the highest-similarity region (e.g., top 1%), relationship-aware embeddings yield markedly higher Lift@q, meaning that the most similar pairs are disproportionately likely to be edges under the observed relation signal. Text-only embeddings show substantially lower lift, consistent with high semantic similarity often corresponding to non-edges (Table I).

E. Outlier Analysis

To make the diagnostic disagreement between semantic and relationship-aware similarity concrete, we report representative *outlier pairs* where the two measures strongly diverge. These

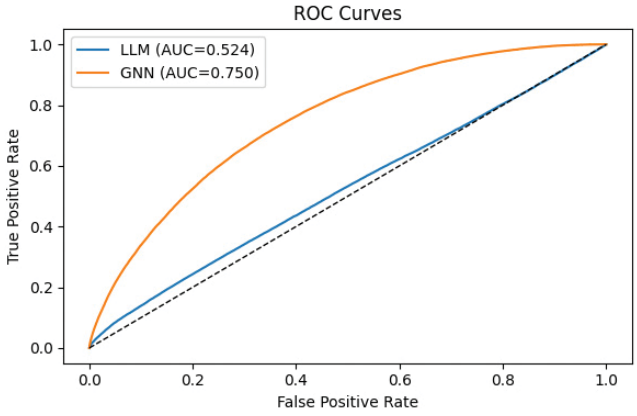

Fig. 2. ROC curves for predicting co-click edges from cosine similarity on 2019-11-11. Relationship-aware embeddings (LLM+GNN) achieve substantially higher AUC than text-only embeddings (LLM).

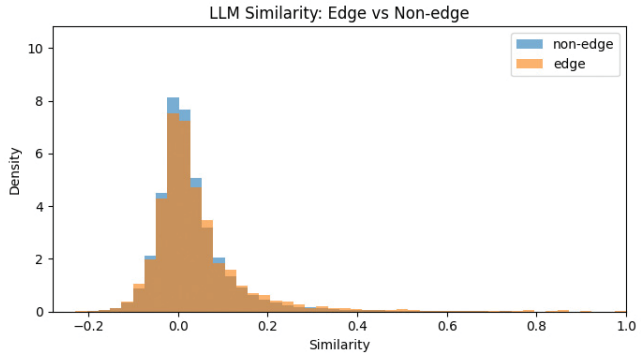

Fig. 3. Text-only similarity distributions for induced edge vs. non-edge item pairs. LLM embeddings show weak separation between pairs linked under the observed relation signal and non-edges.

examples are not an additional metric; they provide qualitative evidence of when similarity is *plausible* in language but not *supported under the observed relation signal*, and vice versa.

We consider two symmetric outlier types using fixed percentile thresholds. First, we identify pairs with **high text similarity but low relationship-aware similarity**: pairs in the top 1% of text-only similarity scores and the bottom 50% of relationship-aware scores Table II. Second, we identify pairs with **high relationship-aware similarity but low text similarity**: pairs in the top 1% of relationship-aware similarity and the bottom 50% of text-only similarity (Table III). To emphasize structural effects rather than trivial lexical overlap, we restrict attention to **cross-category** pairs and rank candidates by the magnitude of the disagreement.

The resulting tables surface two complementary failure modes. Text-high/structure-low pairs illustrate cases where semantic resemblance can be persuasive yet unsupported by the observed relation signal. Structure-high/text-low pairs reveal connections supported by the observed relation signal that are weakly expressed - or absent - in text. Together, these outliers provide interpretable instances of the same distinction quantified by our diagnostics: semantic similarity can look

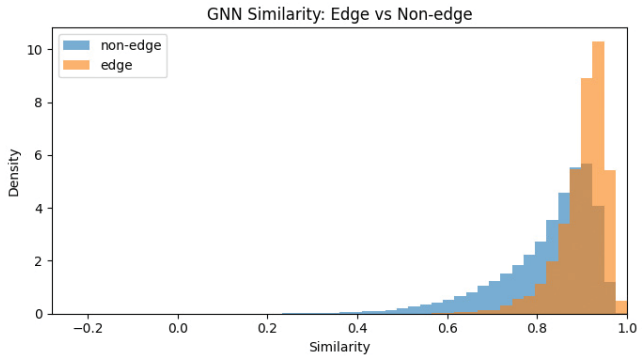

Fig. 4. Relationship-aware similarity distributions for induced edge vs. non-edge item pairs. LLM+GNN embeddings show clearer separation, indicating stronger alignment with the observed relation signal.

TABLE I. TOP-q% CO-CLICK EDGE CONCENTRATION (2019-11-11).

Top fraction q	Edge (LLM)	Edge (GNN)	Lift
0.001 (0.1%)	0.099	0.217	2.19×
0.005 (0.5%)	0.064	0.147	2.30×
0.010 (1%)	0.050	0.125	2.48×
0.050 (5%)	0.034	0.084	2.46×

convincing without evidence, and links under the observed relation signal can be weakly visible in text.

V. CONCLUSION

Our results reinforce a simple but important point: *semantic similarity is not the same as evidence*. Even when LLM embeddings are excellent linguistic representations, they can surface links that feel compelling in language yet lack support from the underlying relational signal. Incorporating interaction evidence through an explicit scaffold yields similarity spaces that more consistently track connections supported by the observed relation signal, providing a practical foundation for explanations that are *grounded* rather than merely fluent.

More broadly, we view this as an *LLM data-quality* question. Instead of asking only ”does it sound related?”, evaluations should also ask ”is it supported by structure?”, and our contribution is to make that distinction measurable with necessity-oriented diagnostics. While we instantiate the evidence graph as a user–item bipartite interaction log, the same metrics apply whenever entities have text and there exists an observed relation signal – whenever relationships can serve as auditable support rather than narrative justification. At the same time, our study is diagnostic rather than task-optimal: the induced co-click graph is useful as an observed relation signal, but it is not equivalent to user intent or recommendation utility.

Looking ahead, this evidence-graph + diagnostics perspective extends naturally beyond our illustrative setting. Any domain where entities have text but correctness depends on how those entities are connected is a candidate - e-commerce, scientific knowledge and discovery, fraud and security, healthcare and biology, software/IT operations, enterprise knowledge management, and beyond. The specific relationships differ across settings, but the underlying question remains: when is language-only similarity sufficient, and when does relational evidence meaningfully change what is correct? A promising direction is to turn that boundary into an operational threshold - so systems can recognize when text is enough, and when relationship-aware handling is warranted, before plausible associations silently become decisions.

TABLE II. CROSS-CATEGORY PAIRS WITH HIGH TEXT SIMILARITY BUT LOW INTERACTION-AWARE SIMILARITY

Item 1 (category: title)	Item 2 (category: title)	LLM	GNN
foodanddrink: The Key to Mastering Goulash, the World’s Most Famous Stew	**sports:** Bud Light Ads Honor Nats Fan	0.685	0.300
video: Tracking freezing temperatures across the U.S.	**weather:** Burst of snow or wintry mix with sleet possible Tuesday; Winter weat...	0.629	0.261
sports: Bucks Five Observations, Including Sloppiness After Time-Outs	**lifestyle:** 30+ Elegant White Rooms To Inspire Your Own Home Decor	0.784	0.468
video: Bison Rolls Around in Fresh Wisconsin Snow	**weather:** Afternoon snow update targets southern Michigan for heaviest snow, s...	0.772	0.492
news: Watch: SpaceX launches 60 Starlink satellites from Cape Canaveral	**video:** SpaceX launches 60 mini satellites into orbit	0.889	0.614

TABLE III. CROSS-CATEGORY PAIRS WITH HIGH INTERACTION-AWARE SIMILARITY BUT LOW TEXT SIMILARITY

Item 1 (category: title)	Item 2 (category: title)	LLM	GNN
video: Underwater camera captures swimming tigers	**weather:** Barham Fire: Crews gain control of Hollywood Hills brush fire after ...	-0.313	0.981
music: Dolly Parton on ’decorating’ her body with hidden tattoos	**tv:** Take That! BIP’s Hannah G. Claps Back After Mispronouncing ’Gnocchi’	-0.284	0.981
movies: Clip - Terminator: Dark Fate	**news:** Iran launches nuclear enrichment at underground Fordow plant, IAEA conf...	-0.289	0.972
finance: Trade War’s Forgotten Farmers Get Crushed in U.S. Cotton Country	**lifestyle:** Why Do Cats Meow?	-0.209	0.981
foodanddrink: These Christmas Dinner Menus Will Create an Unforgettable Meal	**video:** Golden Gums! 200-Year-Old Gold Teeth Made of Ivory & Gold Set to Go fo...	-0.228	0.963

REFERENCES

[1] D. Gentner, “Structure-Mapping: A Theoretical Framework for Analogy,” *Cognitive Science*, vol. 7, no. 2, pp. 155–170, 1983, doi:10.1207/s15516709cog0702_3.

[2] Y. LeCun, “A Path Towards Autonomous Machine Intelligence (Version 0.9.2, 2022-06-27),” OpenReview (working paper), 2022. Available: https://openreview.net/pdf?id=BZ5a1r-kVsf

[3] L. Fridman, “Yann LeCun: Meta AI, Open Source, Limits of LLMs, AGI & the Future of AI,” *Lex Fridman Podcast*, Episode #416, March 7, 2024 (transcript). Available: https://lexfridman.com/yann-lecun-3-transcript/

[4] N. Noy, Y. Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor, “Industry-Scale Knowledge Graphs: Lessons and Challenges,” *ACM Queue*, 2019. Available online: https://queue.acm.org/detail.cfm?id=3332266.

[5] M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges," arXiv preprint arXiv:2104.13478, 2021. Available online: https://arxiv.org/abs/2104.13478.

[6] X. Cheng, W. Zeng, D. Dai, Q. Chen, B. Wang, Z. Xie, K. Huang, X. Yu, Z. Hao, Y. Li, H. Zhang, H. Zhang, D. Zhao, and W. Liang, "Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models," arXiv preprint arXiv:2601.07372, 2026. Available online: https://arxiv.org/abs/2601.07372.

[7] Z. Wang, C. Xu, B. Liu, Y. Wang, S. Han, Z. Yao, H. Yao, and Y. He, "Agent World Model: Infinity Synthetic Environments for Agentic Reinforcement Learning," arXiv preprint arXiv:2602.10090, 2026. Available online: https://arxiv.org/abs/2602.10090.

[8] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive Representation Learning on Large Graphs," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural Graph Collaborative Filtering," *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2019.

[10] A. Jacovi and Y. Goldberg, "Towards Faithfulness in Model Explanation in NLP," *Findings of the Association for Computational Linguistics: EMNLP*, 2020.

[11] Y. Zhang and X. Chen, "Explainable Recommendation: A Survey and New Perspectives," *Foundations and Trends in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2020.

[12] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural Collaborative Filtering," *Proceedings of the 26th International Conference on World Wide Web (WWW)*, 2017.

[13] Deep Graph Library (DGL), "Link Prediction Using Graph Neural Networks," 2018. Available online: https://www.dgl.ai.

[14] F. Wu, Y. Qiao, J. Chen, et al., "MIND: A Large-scale Dataset for News Recommendation," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.