Knowledge Homophily in Large Language Models

Utkarsh Sahu*
University of Oregon
Eugene, OR, USA
utkarsh@uoregon.edu

Zhisheng Qi*
University of Oregon
Eugene, OR, USA
charq@uoregon.edu

Mahantesh Halappanavar
Pacific Northwest National Laboratory
Richland, WA, USA
hala@pnnl.gov

Nedim Lipka
Adobe Research
San Jose, CA, USA
lipka@adobe.com

Ryan Rossi
Adobe Research
San Jose, CA, USA
ryrossi@adobe.com

Franck Dernoncourt
Adobe Research
San Jose, CA, USA
dernonco@adobe.com

Yu Zhang
Texas A&M University
College Station, TX, USA
yuzhang@tamu.edu

Yao Ma
Rensselaer Polytechnic Institute
Troy, NY, USA
may13@rpi.edu

Yu Wang
University of Oregon
Eugene, OR, USA
yuwang@uoregon.edu

Abstract

Large Language Models (LLMs) have been increasingly studied as neural knowledge bases for supporting knowledge-intensive applications. However, the structural organization of their knowledge remains unexplored. Inspired by cognitive neuroscience, such as semantic clustering and priming, where knowing one fact increases the likelihood of recalling related facts, we investigate an analogous knowledge homophily pattern in LLMs. To this end, we map LLM knowledge into a graph representation through knowledge checking at triplet/entity levels. After that, we analyze the knowledgeability relationship between an entity and its neighbors, discovering that LLMs tend to possess a similar level of knowledge about relevant entities positioned closer in the graph. Motivated by this homophily principle, we propose a Graph Neural Network (GNN) regression model to estimate entity-level knowledgeability scores for triplets by leveraging their neighborhood scores. The predicted knowledgeability enables us to prioritize checking less well-known triplets, thereby maximizing knowledge coverage under the same labeling budget. This not only improves the efficiency of active labeling for fine-tuning to inject knowledge into LLMs but also enhances multi-hop path retrieval in reasoning-intensive question answering. Our code and supplementary is available at https://github.com/utkarshxsahu/kgc.

CCS Concepts

• **Computing methodologies → Natural language processing**.

Keywords

Large Language Model; Knowledge Checking; Homophily

This work is licensed under a Creative Commons Attribution 4.0 International License.
WSDM '26, Boise, ID, USA
© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2292-9/2026/02
https://doi.org/10.1145/3773966.3779394

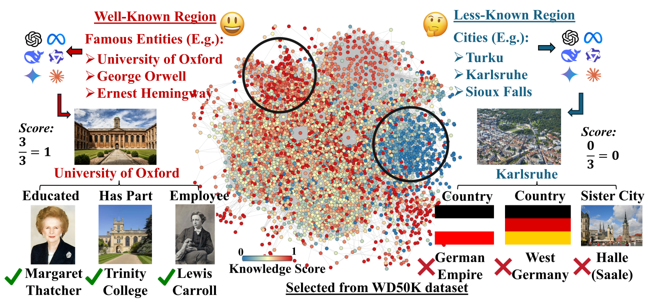

Figure 1: We check whether LLM knows about triple facts and aggregate them to obtain entity knowledgeability scores. The visualized entity-level scores reveal the knowledge homophily, where topologically close entities form distinct high/log-knowledge (red/blue) communities. Graph layout is by ForceAtlas2 [9] to preserve topological proximity.

ACM Reference Format:
Utkarsh Sahu*, Zhisheng Qi*, Mahantesh Halappanavar, Nedim Lipka, Ryan Rossi, Franck Dernoncourt, Yu Zhang, Yao Ma, and Yu Wang. 2026. Knowledge Homophily in Large Language Models. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining (WSDM '26), February 22–26, 2026, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3773966.3779394

1 Introduction

Large Language Models (LLMs) have emerged as powerful neural knowledge bases by encoding vast amounts of world knowledge within their neural parameters [10, 22]. This neural-embedded knowledge enables LLMs to produce contextually relevant and factually rich responses, supporting real-world applications such as fact checking [12] and question answering [11, 33]. To better explore this neural knowledge base, researchers have devised knowledge checking methods to investigate the knowledge patterns of LLMs [1, 38] and leveraged the derived insights to guide knowledge-intensive tasks, including adaptive retrieval [7, 35, 36], knowledge editing [26], and hallucination detection [27].

*Equal contribution and co-first authors.

Despite various knowledge patterns identified previously [10, 22, 37], little attention has been given to whether LLMs' knowledge exhibits structural organization. In fact, in cognitive neuroscience [13], several works have highlighted the semantic clustered patterns of the neural knowledge in human brain networks: (i) semantic clustering in memory recall, where people tend to retrieve related words together (e.g., recalling "dog, cat, horse" in sequence) [2, 17], and (ii) homophily brain networks, where regions with similar functions or inputs are more likely to connect [29]. Analogously, we hypothesize that LLMs also exhibit a similar knowledge homophily pattern, i.e., they tend to possess similar levels of knowledge about conceptually related entities, as illustrated in Figure 1 by checking GPT-3.5's knowledge about triplets from WD50K dataset. Discovering this phenomenon sheds light on how knowledge in LLMs is structurally organized and informs solutions for knowledge-intensive tasks. In particular, estimating a concept's knowledgeability from related concepts helps identify weaker regions, enabling more efficient labeling for knowledge injection and retrieval as shown in Section 4.

Motivated by homophily in other disciplines [15, 32], this paper uncovers this pattern in LLMs and develops graph models to predict knowledgeability. These predictions identify less-known regions to guide efficient fine-tuning and enhance retrieval for multi-hop question answering. Our contributions are:

- **Knowledge Homophily Discovery:** We demonstrate the existence of knowledge homophily in LLMs by measuring knowledge at triplet/entity levels, showing that topologically close entities tend to exhibit similar knowledgeability scores.
- **Knowledge Homophily Application:** We leverage the discovered knowledge homophily to develop a GNN-based estimator that infers the entity knowledgeability, and showcase two applications enhancing knowledge injection efficiency and guiding multi-hop retrieval for question-answering.

2 Related Work

Knowledge Checking for LLMs as Knowledge Bases (KBs). LLMs have evolved into general-purpose agents and neural knowledge bases for knowledge-intensive applications [21, 24]. However, unlike prior knowledge bases with explicit schemas [31], LLM knowledge is implicitly encoded and largely non-interpretable. This lack of transparency motivates the need to verify what LLMs "know" and ensure their reliable use. Existing knowledge checking methods can be categorized into verifying factual accuracy [8, 12], assessing self-awareness [10, 30], and evaluating knowledge coverage and consistency against internal or external sources [14, 16]. While effective, they focus on knowledge content rather than structure.

Structured Understanding of LLMs as Knowledge Bases. Existing structured understandings of LLM knowledge focus on model parameters from two perspectives. The first examines where knowledge is stored, showing that feed-forward layers act as key–value memories for factual knowledge [6], with factual associations often localized and editable in mid-layer "knowledge neurons" [3, 18]. The second investigates how knowledge is structurally organized. [20] evaluates properties such as symmetry, hierarchy, and path-following, revealing failures in complex relational reasoning. Despite exposing implicit structure in LLM knowledge, the role of homophily remains largely unexplored.

Prompt 1: LLM-based Triplet Evaluation

System Message: Evaluate the statement based on your knowledge and respond with `True` or `False`.

Given: Triplet $\mathcal{T} = (sub, \ rel, \ obj)$, **Date D (`Temporal Version`)**

Template: Relation → Pattern (e.g., `son_of` → {SUB} is the son of {OBJ}.)

Procedure:

(1) Retrieve relation-based template for *rel* in triplet $\mathcal{T}$.
(2) Fill {SUB}→*sub*, {OBJ}→*obj* from $\mathcal{T}$ to get statement $\mathcal{S}$.
(3) **Append $\mathcal{S}$ on Date D (`Temporal Version`)**
(4) Prompt **System Msg** + **User Msg:** $\mathcal{S}$ to the LLM.
(5) Record the model output in the format of `True`/`False`

3 Knowledge Homophily Discovery

This section investigates knowledge homophily. We first compute triplet-level knowledgeability and aggregate it into entity-level scores, then assess homophily by measuring knowledgeability differences between neighboring entities in Section 3.2.1 and qualitatively visualizing these scores in Section 3.2.2.

3.1 Knowledgeability Computation

To examine whether LLMs exhibit consistent knowledge about neighboring entities, we first evaluate knowledgeability at the triplet level and then aggregate it to obtain an entity-level score. Given triplets $\mathcal{T} = \{(s_i, d_i, r_i)\}_{i=1}^{|\mathcal{T}|}$ from the knowledge graph, where a source entity s_i is connected to a destination entity d_i via relation r_i, we define the knowledgeability of the LLM on triplet (s_i, d_i, r_i) as $\mathcal{K}(s_i, d_i, r_i)$, reflecting how well the LLM knows about this relational fact. For each entity s_i, we denote its neighbor entity set as $\mathcal{N}(s_i)$, representing the entities adjacent to s_i as either head or tail, and their corresponding neighbor triplet set as $\mathcal{T}(s_i)$. The entity-level knowledgeability of s_i, denoted as $\mathcal{K}(s_i)$, is derived by aggregating knowledgeability scores over its neighboring triplets, capturing how well the LLM knows about entity s_i. Next, we introduce details of calculating triplet and entity knowledgeability.

3.1.1 Calculating Triplet Knowledgeability. Following prior work showing that LLMs are generally well-calibrated in knowing what they know [1, 10, 22], we convert each triplet (s_i, d_i, r_i) into a natural language statement and prompt the LLM to judge whether it recognizes the fact. The model response is recorded as a binary value, with `True`/`False` mapping to 1/0, representing its knowledgeability about the triplet $\mathcal{K}(s_i, d_i, r_i)$. For temporal triplets (s_i, d_i, r_i, t) (e.g., "Donald Trump made a visit to China on 2017-11-08."), we extend the prompt to include the timestamp, allowing us to assess the temporal dimension of LLM knowledgeability. Prompt 1 illustrates the template, with temporal variants highlighted in red.

3.1.2 Calculating Entity Knowledgeability. Given the above triplet knowledgeability, we obtain v_i entity knowledgeability by aggregating scores of all triplets involving v_i [23]:

$$\mathcal{K}(v_i) = |\mathcal{T}(v_i)|^{-1} \sum_{(s,d,r) \in \mathcal{T}(v_i)} \mathcal{K}(s, d, r). \quad (1)$$

Note that the above neighborhood aggregation naturally extends to temporal triplets $(s, d, r, t) \in \mathcal{T}(v_i)$, allowing temporal information to be incorporated into the entity knowledgeability calculation.

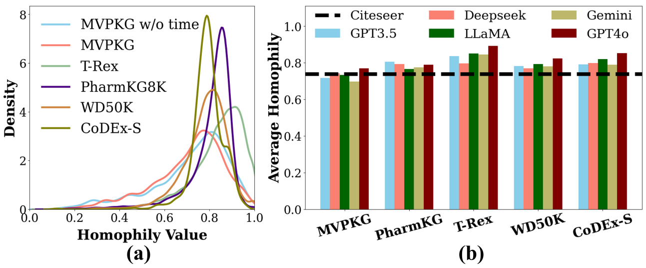

Figure 2: (a): Homophily distribution of node knowledgeability; (b): Average knowledge homophily across datasets/LLMs with black dashed line showing a classic high homophily Citeseer (0.74) dataset for node classification [32].

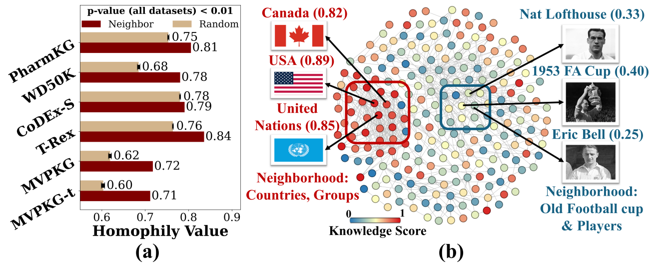

Figure 3: (a) Neighboring nodes possess similar knowledgeability scores to randomly sampled nodes ; (b) Entities with their distinct knowledgeability levels $\mathcal{K}(v)$ indicated by node color (Red = High, Blue = Low).

3.2 Homophily Computation and Analysis

We evaluate whether topologically close entities share similar knowledgeability, i.e., the homophily of entity knowledgeability $\mathcal{H}(v_i)$. Following [15], we compute knowledgeability homophily as one minus the average absolute knowledgeability difference between central node v_i and its neighbors $\mathcal{N}(v_j)$:

$$\mathcal{H}(v_i) = 1 - \frac{1}{|\mathcal{N}(v_i)|} \sum_{v_j \in \mathcal{N}(v_i)} |\mathcal{K}(v_i) - \mathcal{K}(v_j)| \tag{2}$$

where a smaller difference between neighboring entities, $|\mathcal{K}(v_i) - \mathcal{K}(v_j)|$, leads to a higher homophily value $\mathcal{H}(v_i)$. We empirically quantify triplet/entity-level knowledgeability and analyze homophily patterns both quantitatively and qualitatively. We evaluate five representative LLMs, GPT-3.5, 4o, Gemini-2.5 Flash, LLaMA3.3-70B, and DeepSeek-V3, across five knowledge graphs: MVPKG [19], T-Rex [4], PharmKG [39], WD50K [5], and CoDEx-S [25]. T-Rex, WD50K, and CoDEx-S capture general Wikipedia knowledge, while PharmKG8K and MVPKG focus on biomedical and political domains. Graph visualizations in Figures 1 and 3 use ForceAtlas2 [9] to position topologically close nodes visually close, enabling an intuitive assessment of whether they share similar knowledgeability scores.

3.2.1 Quantitative Analysis of Node/Graph Knowledge Homophily. Figure 2(a)/(b) shows node/graph-level homophily across multiple knowledge graphs. In Figure 2(a), node homophily distributions are right-skewed and peak near 0.8, indicating that most entities share similar knowledgeability with their neighbors. This high homophily is known to benefit node-level prediction tasks [40], motivating our use of regression for entity knowledge estimation (Section 4.1). Incorporating temporal information in MVPKG causes a slight left shift. In addition, Figure 2(b) reports average graph homophily, which consistently exceeds that of the Citeseer benchmark [28, 34] across datasets/LLMs, indicating that the observed homophily aligns with the conventional "high-homophily" level [15].

We further compare knowledge homophily to a degree-matched random baseline by replacing each node's true neighborhood $\mathcal{N}(v)$ with a randomly sampled peer group $\widehat{\mathcal{N}}(v)$ of the same size from the graph, and computing homophily relative to this group. True-neighborhood homophily is significantly higher than the random baseline (100 trials per dataset, two-tailed z-test, $p < 0.01$) with tight 99% confidence intervals. *As shown in Figure 3(a), this confirms that knowledge homophily is not a random artifact but an intrinsic structural property of LLMs' knowledge organization.*

3.2.2 Qualitative Analysis of Knowledge Homophily. Figure 3(b) visualizes a T-Rex subgraph colored by knowledgeability $\mathcal{K}(v)$. A geopolitical neighborhood forms a high-knowledge region, while a historical football cluster is similarly coherent but with lower knowledgeability. Despite varying means, the small intra-neighborhood deviations in both groups confirm strong knowledge homophily.

4 Knowledge Homophily Application

After discovering the knowledge homophily, where topologically proximate entities exhibit similar knowledgeability, we apply this insight to two knowledge-intensive tasks: (1) homophily-aware knowledge checking for efficient fine-tuning, and (2) homophily-aware knowledge retrieval for enhanced question answering. We train a GNN-based model to estimate entity-level knowledgeability from neighborhood signals and identify triplets in low-knowledge regions. These triplets are then prioritized for fine-tuning to maximize knowledge injection or for retrieval to complement missing knowledge in answer generation. Both tasks rely on knowledgeability estimation to pinpoint knowledge gaps.

4.1 Homophily-aware Knowledge Estimation

Given that homophily is a sufficient condition for high-utility GNN predictions [15], we design a GNN-based regression model to perform message-passing, aggregate neighboring embeddings, and predict unknown entity scores. Specifically, given a set of entities $\mathcal{V}^{\text{Train}}$ with known knowledgeability (by prompting LLMs), we train a GNN to estimate the knowledgeability of unseen entities. At each layer, the model performs Message Passing (MP) and Feature Transformation (TR), followed by regression:

$$\widehat{\mathcal{K}}_i^l = \text{MP}^l(\widetilde{\mathcal{K}}_j^{l-1} \mid v_j \in \mathcal{N}(v_i) \cup v_i), \quad \widetilde{\mathcal{K}}_i^l = \text{TR}^l(\widehat{\mathcal{K}}_i^l), \tag{3}$$

$$\mathcal{L} = \frac{1}{|\mathcal{V}^{\text{Train}}|} \sum_{v_i \in \mathcal{V}^{\text{Train}}} \left|\widetilde{\mathcal{K}}_i^l - \mathcal{K}_i\right|^2, \tag{4}$$

The initial node feature matrix is $\widetilde{\mathcal{K}}^0 = [\mathcal{X}_1, \ldots, \mathcal{X}_{v_{|\mathcal{V}|}}]^\top$, where each node feature $\mathcal{X}_{v_i}$ is a dense textual embedding from pretrained language models. After training on $\mathcal{V}^{\text{Train}}$, the model is further used to infer the knowledgeability scores for all entities in the knowledge graph, eliminating the need for resource/time-intensive knowledge probing via exhaustive LLM prompting. We utilize the estimated entity knowledgeability scores to guide triplet selection for LLM fine-tuning (Section 4.2) and to guide retrieval for reasoning-intensive multi-hop QA (Section 4.3). Due to space constraints, we summarize the setup, and full details are in Figure 4 of Supplementary.

Table 1: Performance comparison of fine-tuning with triplets selected by knowledgeability estimated by Random, MLP, and GNN. Best result in bold and second-best underlined. L=Llama3-8B, M=Mistral-7B. Selection Quality: percentage of triplets selected for fine-tuning that are unknown to base LLMs. Generalization Gain: percentage of additional 2% evaluation triplets identified by the fine-tuned LLMs. Detailed setting is visualized in Figure 4 in Supplementary.

Task	Method	T-Rex		PharmKG		WD50K		MVPKG		CoDExS		Avg.
		L	M	L	M	L	M	L	M	L	M	
Selection Quality	**Rand**	36.5	44.8	81.9	72.8	41.2	46.8	68.5	66.3	33.8	51.7	54.4
	MLP	**38.4**	48.7	84.8	77.0	44.3	47.8	67.9	68.6	39.1	57.8	57.4
	GNN	37.3	**54.5**	**87.6**	**79.2**	**49.6**	**50.6**	**72.2**	**71.5**	**45.5**	**63.9**	**61.2**
Generalization Gain	**Base**	63.3	64.0	17.8	55.3	54.8	42.9	26.1	52.3	64.9	58.5	49.9
	Rand	86.4	81.9	34.9	41.3	57.8	**56.3**	30.7	65.1	**78.8**	72.1	60.5
	MLP	87.9	90.2	35.8	57.2	56.1	53.2	42.8	74.5	73.7	85.2	65.6
	GNN	**89.1**	**91.9**	**37.0**	**60.7**	**58.8**	55.1	**44.5**	**76.7**	75.6	**88.0**	**67.7**

4.2 Homophily-guided Knowledge Injection

We leverage the homophily to estimate triplet knowledgeability and prioritize selecting less-known triplets for fine-tuning LLMs within a fixed budget, thereby enabling more effective knowledge injection into LLMs. For each dataset, we allocate 4000 triplets as the knowledge-checking budget for selection and fine-tuning, with an additional 2% of all triplets reserved as the test set. Within the 4000 budget, 20% of triplets are sampled as anchor points, for which we directly query the base LLM to obtain ground-truth binary knowledgeability scores (Section 3.1). These anchors provide labeled data to estimate the knowledgeability of their associated entities, which is used to train a GNN model (Eq. (4)) and predict knowledgeability scores for all remaining entities. Based on these predictions, we prioritize triplets with lower-scored entities from the remaining 80% unqueried pool to complete the 4000-triplet set for fine-tuning. We benchmark this triplet selection against two baselines: Random, which uniformly samples triplets, and MLP, which estimates knowledgeability without homophily, eliciting the knowledge homophily contribution to knowledge estimation. We experiment with LLaMA3-8B(L) and Mistral-7B(M).

Table 1 evaluates homophily-guided knowledge injection from two perspectives: selection quality and generalization gain. For selection quality, we assess whether the chosen triplets better capture the knowledge deficiencies of LLMs. Among the 4000 triplets selected for fine-tuning, we compute the percentage that the base LLM does not recognize, following the procedure in Section 3.1. A higher score indicates that more selected triplets are unknown to the LLM and thus more valuable for fine-tuning. Our GNN regressor achieves the highest proportion of unknown triplets, outperforming MLP and Random selection. This demonstrates the advantage of incorporating homophily into GNN design in enabling more effective estimation of ground-truth knowledgeability for knowledge injection. For generalization gain, we test whether fine-tuning on selected triplets improves the knowledgeability over the reserved 2% held-out set. The best performing GNN regressor confirms that its higher selection precision translates into stronger knowledge generalization. This superior generalization gain holds across different evaluation budgets from 1% to 20% in Figure 5 in Supplementary.

Table 2: Multi-hop Question Answering Accuracy by GPT4-as-a-Judge; M=MLP, G=GNN, BS=Beam Search, H = Hop

Dataset	T-Rex		PharmKG		WD50K		MVPKG		CoDExS	
Q-Hop	2-H	3-H	2-H	3-H	2-H	3-H	2-H	3-H	2-H	3-H
Base	30.9	22.6	21.4	16.0	25.1	17.2	24.4	19.1	28.6	20.4
M-BS	33.8	23.1	21.7	16.2	25.8	17.3	24.9	19.4	29.9	20.5
G-BS	**34.2**	**23.7**	**22.2**	**16.6**	**26.0**	**17.5**	**25.4**	**19.5**	**31.1**	**20.9**

4.3 Homophily-guided Knowledge Retrieval

We test whether the estimated knowledgeability can guide entity retrieval to provide better context for question-answering. For each KG, we generate 1000 questions (500 2-hop/500 3-hop). Entity knowledgeability $\mathcal{K}(v)$ is predicted with a GNN regressor trained on 40% of entities labeled by GPT-3.5, excluding entities for generating 1000 evaluation questions. We embed both entities/relations and questions using `all-MiniLM-L6-v2`. Starting with entity linking in the question, we run beam search up to the hop limit and score each neighbor by its knowledgeability $\mathcal{K}(v)$ and semantic similarity $\mathcal{S}(r||d, q)$ to the question q where $r||d$ represents its relation r concatenated with the tail entity d. Baselines are as follows:

- **Baseline (Semantic Beam Search):** It retrieves paths using beam search guided solely by the semantic similarity $\mathcal{S}(r||d, q)$ between the path (relation + tail entity) and the input question.
- **Knowledge-aware Beam Search (BS):** This method adjusts beam search to favor less-known entities. For each expansion, the semantic score $\mathcal{S}$ is penalized by the next entity knowledgeability, $\mathcal{K}(u)$. The final score is $\mathcal{S} \times (1 - \alpha \cdot \mathcal{K}(u))$ with α being weighting factor. Entities with lower knowledgeability receive higher retrieval priority, achieving knowledge-aware search. Beam Search (BS) with GNN/MLP as knowledge estimator are G-BS/M-BS.

Using a GPT-3.5 reader (restricted to retrieved triples) evaluated by GPT-4, we find that M/G-BS consistently outperforms the Baseline, as shown in Table 2. Crucially, G-BS surpasses the homophily-agnostic M-BS, validating the advantage of structural knowledge homophily. G-BS achieves a 4.57% improvement on 2-hop queries (favoring general KGs). While performance declines for all methods on 3-hop queries due to semantic drift, G-BS still secures a 2.62% improvement. These gains remain consistent across training budgets, as shown in Figure 6 in Supplementary.

5 Conclusion

Inspired by the structural knowledge organization in the human brain, we investigate homophily in LLMs' neural knowledge and validate it via correlated knowledgeability scores among neighboring entities in knowledge graphs. Building on this observation, we propose a GNN-based regressor that exploits local neighborhoods to estimate entity-level knowledgeability. We verify its effectiveness in selecting less-known triplets for efficient knowledge injection via fine-tuning, and in improving retrieval for multi-hop question answering. In the future, we plan to explore uncertainty-aware knowledge verification and dynamic homophily modeling to capture how homophily evolves as LLMs acquire new information.

Acknowledgements

This research is supported by the National Science Foundation (NSF) under grant number IIS 2524379.

6 Ethical Considerations

Knowledge homophily, where topologically proximate entities exhibit similar knowledgeability, amplifies the risk of knowledge-extraction attacks. Adversaries can exploit this structure by crafting queries over cohorts of related entities, thereby maximizing unintended information disclosure. This poses privacy and copyright risks, as semantically clustered entities facilitate reconstruction of sensitive facts. To mitigate these threats, several defensive strategies can be employed, including query pre-filtering and sanitization, rate-limiting cohort-based requests, prompt-level heuristics that block verbatim proprietary content, and detector or red-teaming mechanisms for identifying adversarial extraction patterns. When combined with fine-grained access control and differential privacy constraints, these measures can substantially reduce the attack surface introduced by knowledge homophily.

References

[1] Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv* (2022).

[2] Weston A Bousfield and Charles Hill W Sedgewick. 1944. An analysis of sequences of restricted associative responses. *The Journal of General Psychology* (1944).

[3] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint* (2021).

[4] Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[5] Mikhail Galkin, Priyansh Trivedi, Gaurav Maheshwari, Ricardo Usbeck, and Jens Lehmann. 2020. Message passing for hyper-relational knowledge graphs. *arXiv preprint arXiv:2009.10847* (2020).

[6] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint* (2020).

[7] Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A Rossi, Subhabrata Mukherjee, Xianfeng Tang, et al. 2024. Retrieval-augmented generation with graphs (graphrag). *arXiv preprint arXiv:2501.00309* (2024).

[8] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

[9] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* (2014).

[10] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).

[11] Yongjia Lei, Haoyu Han, Ryan A Rossi, Franck Dernoncourt, Nedim Lipka, Mahantesh M Halappanavar, Jiliang Tang, and Yu Wang. 2025. Mixture of Structural-and-Textual Retrieval over Text-rich Graph Knowledge Bases. *arXiv preprint arXiv:2502.20317* (2025).

[12] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).

[13] Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, et al. 2025. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *arXiv preprint arXiv:2504.01990* (2025).

[14] Linhao Luo, Thuy-Trang Vu, Dinh Phung, and Gholamreza Haffari. 2023. Systematic assessment of factual knowledge in large language models. *arXiv preprint arXiv:2310.11638* (2023).

[15] Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. 2021. Is homophily a necessity for graph neural networks? *ICLR* (2021).

[16] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint* (2022).

[17] Jeremy R Manning and Michael J Kahana. 2012. Interpreting semantic clustering effects in free recall. *Memory* (2012).

[18] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems* 35 (2022), 17359–17372.

[19] Xinyi Mou, Zejun Li, Hanjia Lyu, Jiebo Luo, and Zhongyu Wei. 2024. Unifying local and global knowledge: Empowering large language models as political experts with knowledge graphs. In *Proceedings of the ACM Web Conference 2024*.

[20] Vishwas Mruthyunjaya, Pouya Pezeshkpour, Estevam Hruschka, and Nikita Bhutani. 2023. Rethinking language models as symbolic knowledge graphs. *arXiv preprint arXiv:2308.13676* (2023).

[21] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2463–2473.

[22] Pouya Pezeshkpour. 2023. Measuring and modifying factual knowledge in large language models. In *2023 International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 831–838.

[23] Thorsten Rings, Timo Bröhl, and Klaus Lehnertz. 2022. Network structure from a characterization of interactions in complex systems. *Scientific Reports* (2022).

[24] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[25] Tara Safavi and Danai Koutra. 2020. Codex: A comprehensive knowledge graph completion benchmark. *arXiv preprint arXiv:2009.07810* (2020).

[26] Yucheng Shi, Qiaoyu Tan, Xuansheng Wu, Shaochen Zhong, Kaixiong Zhou, and Ninghao Liu. 2024. Retrieval-enhanced knowledge editing in language models for multi-hop question answering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2056–2066.

[27] Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766* (2023).

[28] Zixing Song, Xiangli Yang, Zenglin Xu, and Irwin King. 2022. Graph-based semi-supervised learning: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[29] Olaf Sporns. 2012. The human connectome: a complex network. *Schizophrenia Research* (2012).

[30] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975* (2023).

[31] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* (2014).

[32] Yu Wang and Tyler Derr. 2021. Tree decomposed graph neural network. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 2040–2049.

[33] Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19206–19214.

[34] Yu Wang, Yuying Zhao, Neil Shah, and Tyler Derr. 2022. Imbalanced graph classification via graph-of-graph neural networks. In *Proceedings of the 31st ACM international conference on information & knowledge management*. 2067–2076.

[35] Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li. 2024. Seakr: Self-aware knowledge retrieval for adaptive retrieval augmented generation. *arXiv preprint arXiv:2406.19215* (2024).

[36] Zihan Zhang, Meng Fang, and Ling Chen. 2024. RetrievalQA: Assessing Adaptive Retrieval-Augmented Generation for Short-form Open-Domain Question Answering. In *Findings of the Association for Computational Linguistics ACL 2024*.

[37] Danna Zheng, Mirella Lapata, and Jeff Z Pan. 2024. Large language models as reliable knowledge bases? *arXiv preprint arXiv:2407.13578* (2024).

[38] Shangshang Zheng, He Bai, Yizhe Zhang, Yi Su, Xiaochuan Niu, and Navdeep Jaitly. 2023. KGLens: Towards Efficient and Effective Knowledge Probing of Large Language Models with Knowledge Graphs. *arXiv preprint* (2023).

[39] Shuangjia Zheng, Jiahua Rao, Ying Song, Jixian Zhang, Xianglu Xiao, Evandro Fei Fang, Yuedong Yang, and Zhangming Niu. 2021. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in bioinformatics* (2021).

[40] Jiong Zhu, Ryan A Rossi, Anup Rao, Tung Mai, Nedim Lipka, Nesreen K Ahmed, and Danai Koutra. 2021. Graph neural networks with heterophily. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11168–11176.