A Graph Perspective to Probe Structural Patterns of Knowledge in Large Language Models

Anonymous ACL submission

Abstract

Large Language Models have been extensively studied as neural knowledge bases for their knowledge access, editability, reasoning, and explainability. However, few works focus on structural patterns of their knowledge. Motivated by this gap, we investigate these structural patterns from a graph perspective. We introduce triplet/entity knowledgeability to quantify the knowledge of LLMs at both the triplet and entity levels, and analyze how it relates to graph structural properties such as node degree. Furthermore, we uncover the knowledge homophily, where topologically close entities exhibit similar levels of knowledgeability, which further motivates us to develop graph machine learning models to estimate entity knowledge based on its local neighbors. This model further enables more valuable knowledge checking by selecting triplets less known to LLMs. Empirical results show that using selected triplets for fine-tuning leads to superior performance. Our code is publicly available here.

1 Introduction

Large Language Models (LLMs) have emerged as powerful knowledge bases by encoding world knowledge within their neural parameters (Kadavath et al., 2022; Pezeshkpour, 2023; Yin et al., 2023). This world knowledge allows LLMs to generate contextually relevant and factually rich responses to natural language prompts that serve real-world applications. To more wisely leverage this capability, researchers have been probing LLMs' knowledge from various aspects (AlKhamissi et al., 2022; Zheng et al., 2023), including consistency, editability, reasoning, and explainability. These probing efforts have inspired adaptive retrieval, LLM unlearning, confidence calibration, and hallucination detection (Si et al., 2023; Farquhar et al., 2024; Ahdritz et al., 2024).

Despite the above progress (Kadavath et al., 2022; Pezeshkpour, 2023; Zheng et al., 2024), few have examined structural patterns of LLMs' knowledge. Inspired by cognitive neuroscience (Liu et al., 2025), which has uncovered structured patterns in human knowledge organization, such as semantic networks that cluster related concepts (Huth et al., 2016; Hoedemaker and Gordon, 2017), specialized

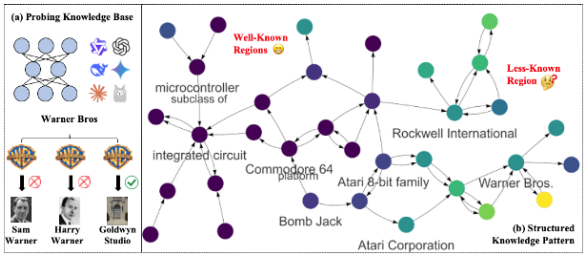


Figure 1: (a) Prompting LLMs to check their knowledge about each triplet and aggregate them to obtain entity knowledgeabilty; (b) These scores are assigned to graph nodes, enabling analysis of structural patterns such as knowledge imbalance (depicted in darker/lighter color), and knowledge homophly where topologically close entities possess similar levels of knowledgeability.

brain regions for specific categories of information (Kanwisher et al., 1997; Binder et al., 2009), and spatial or topographic maps for sensory inputs (Garvert et al., 2017), we hypothesize that similar structured patterns exist within LLMs. Probing these structural patterns provides critical insights into how knowledge is stored, retrieved, and reasoned in LLMs. For example, such understanding could support more flexible retrieval by leveraging structured knowledge organization.

Given the criticality of understanding the structural patterns of knowledge in LLMs and the limited exploration in this field, we take a fresh graph-based perspective to uncover the structural patterns of knowledge encoded in LLMs. Building on these derived structural patterns, we develop graph machine learning models to identify more informative knowledge for fine-tuning LLMs. Our key contributions are as follows:

- **Novel Graph Perspective to Probe Structural Patterns of LLM Knowledge:** We introduce a novel graph-based approach to analyze structural patterns of knowledge in LLMs. Specifically, we define two knowledgeability metrics to quantify LLMs' knowledge at the triplet and entity levels.

- **Discovery of Novel Structural Patterns:** Several novel patterns are revealed, including entity knowledge imbalance, positive correlations between entity degree and knowledgeability, and knowledge homophily, where topologically proximate entities exhibit similar knowledgeability.

- **Graph Learning for Knowledge Prediction and Checking:** We design graph-based regression models to estimate LLM knowledgeability scores for each entity by leveraging its local neighborhood context. These predicted scores are then used to prioritize high-value triplet facts for more effective LLM fine-tuning.

2 Method

Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{R}, \mathcal{F})$ with $\mathcal{V}/\mathcal{R}/\mathcal{F}$ being the set of entities/relations/facts. Each fact is represented as a triplet (v_i, r_{ij}, v_j) with $v_i/v_j \in \mathcal{V}$ being the head/tail entities, and $r_{ij} \in \mathcal{R}$ being their relation. We define the LLMs' knowledgeability for a given triplet (v_i, r_{ij}, v_j)/entity v_i as $\mathcal{K}(v_i, r_{ij}, v_j)/\mathcal{K}(v_i)$, measuring the extent to which the LLM is aware of the triplet fact or entity. Regarding graph structural properties, the degree and clustering coefficient of an entity v_i are denoted as d_{v_i} and c_{v_i}. We define the neighbor entity set $\mathcal{N}(v_i)$ of v_i as the set of entities directly connected to v_i and the neighbor triplet set $\mathcal{T}(v_i)$ of v_i as the set of triplets in which v_i appears as either the head-/tail entity. Next, we introduce knowledgeability measurement at the triplet/entity levels.

2.1 Triplet Knowledgeability

Inspired by prior work (Kadavath et al., 2022; AlKhamissi et al., 2022; Pezeshkpour, 2023), we transform each triplet (v_i, r_{ij}, v_j) into a natural language statement and prompt LLMs to assess whether they recognize the fact. The response of LLMs is recorded as a binary value with `True`/`False` mapped to 1/0, indicating the knowledgeability of LLM about the triplet $\mathcal{K}_{(v_i, r_{ij}, v_j)}$.

To handle temporal triplets with time information (v_i, r_{ij}, v_j, t) (e.g., "Donald Trump made a visit to China on 2017-11-08."), we extend the prompt to explicitly incorporate timestamps, allowing us to consider the temporal impact on LLM knowledgeability. The template of the initial prompt is shown as below with its temporal variation attached in Appendix G:

Prompt 1: LLM-based Triplet Evaluation

System Message: Evaluate the statement based on your knowledge and respond with `True` or `False`.
Given: Triplet $\mathcal{T} = (sub,\ rel,\ obj)$.
Relational Template Map: $T:\ rel \mapsto$ "$\{sub\}\ \ldots\ \{obj\}$".
Procedure:
1. Retrieve relation-based template $t = T(relation)$.
2. Instantiate statement $S = t[\{sub\} \to sub,\ \{obj\} \to obj]$.
3. Prompt **System Msg + User Msg:** S to the LLM.
4. Return "True" or "False."

2.2 Entity Knowledgeability and Homophily

Given the above triplet knowledgeability, we obtain the entity v_i's knowledgeability score by aggregating the knowledgeability of all triplets in which v_i is involved (Jia et al., 2019; Rings et al., 2022):

$$\mathcal{K}(v_i) = |\mathcal{T}(v_i)|^{-1} \sum_{(v_i, r_{ij}, v_j) \in \mathcal{T}(v_i)} \mathcal{K}(v_i, r_{ij}, v_j) \quad (1)$$

Note that the above neighborhood aggregation to obtain the knowledgeability score for each entity also applies to temporal triplets $(v_i, r_{ij}, v_j, t) \in \mathcal{T}(v_i)$, allowing us to account for the temporal impact when assessing an entity's knowledgeability. The change of knowledgeability after incorporating temporal information is shown in Figure 2(a).

Furthermore, we evaluate whether topologically close entities share similar knowledgeability, i.e., the homophily of entity knowledgeability $\mathcal{H}_{v_i}$. Inspired by existing homophily computation (Zhu et al., 2020; Wang and Derr, 2021; Ma et al., 2021), we compute knowledgeability homophily as one minus the average absolute difference in knowledgeability between central node v_i and its neighbors $\mathcal{N}(v_j)$ in the knowledge graph:

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

2.3 Knowledgeability Regression with GNNs.

Given the observed high homophily of entity knowledgeability scores in Figure 2(b), we further design GNN-based graph regression models to approximate the knowledgeability of unknown entities based on known ones. Specifically, given a fixed set of entities $\mathcal{V}^{\text{Train}}$ with known knowledgeability, our goal is to train a GNN model to estimate the entity knowledgeability with unknown scores. We perform message-passing (MP) and feature transformation (TR) followed by regression:

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

$$\mathcal{L} = \frac{1}{|\mathcal{V}^{\text{Train}}|} \sum_{v_i \in \mathcal{V}^{\text{Train}}} \left\| \widetilde{\mathcal{K}}_i^l - \mathcal{K}_i \right\|_2^2, \quad (4)$$

The initial node feature matrix is defined as $\widetilde{\mathcal{K}}^0 = [\mathcal{X}(v_1), \ldots, \mathcal{X}(v_{|\mathcal{V}|})]^\top$, where each node feature $\mathcal{X}(v_i)$ is either a one-hot encoding or a dense text embedding obtained from pretrained language models. By training a regression model on a subset of entities $\mathcal{V}^{\text{Train}}$, we manage to estimate the knowledgeability of all entities without the need for resource and time-intensive knowledge probing via prompting LLMs across the entire entity set.

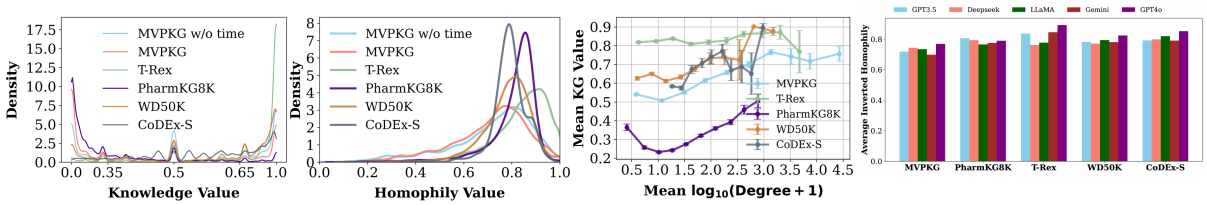


Figure 2: **(a)/(b)**: Distribution of node knowledgeability/homophily for each dataset; **(c)**: Node knowledgeability increases as node degree increases. The results here are based on GPT3.5, and results for other LLMs hold similar observations in Appendix E. **(d)**: Average homophily for all datasets given by different LLMs exceeds 0.6.

3 Experiment

In this section, we quantify triplet/entity knowledgeability, analyze its correlation with structural properties of the underlying graphs, estimate knowledgeability using GNNs, and explore active selection strategies to identify high-valuable triplets for fine-tuning. We evaluate five representative LLMs: commercial ones such as GPT-3.5, 4o, Gemini-2.5 Flash, and two open-source models, LLaMA3.3-70B and DeepSeek-V3. These models are assessed across five knowledge graphs: MVPKG (Mou et al., 2024), T-Rex (Elsahar et al., 2018), PharmKG8K (Zheng et al., 2021), WD50K (Galkin et al., 2020), and CoDEx-S (Safavi and Koutra, 2020). Among them, T-Rex, WD50K, and CoDEx-S represent general factual Wikipedia knowledge, whereas PharmKG8K and MVPKG focus on specialized pharmaceutical and political science. Further details on datasets and experimental configurations are in Appendix D. We now present our key experimental findings.

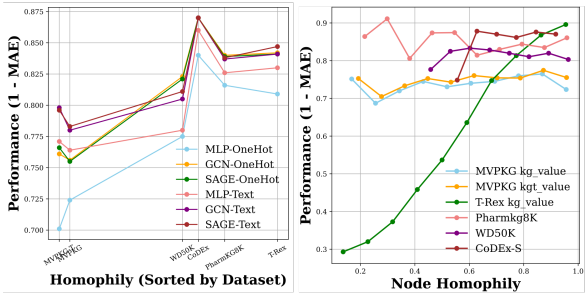
Finding 1 - Figure 2(a) presents the distribution of entity knowledgeability scores across various datasets. The scores exhibit a trimodal pattern with peaks at 0.0, 0.5, 1.0, corresponding to cases where none, some, or all of an entity's triplets are recognized. These patterns exhibit clear domain-specific variation. Specialized datasets such as PharmKG8K and MVPKG are left-skewed, with a dominant peak at 0.0 reflecting LLM's limited knowledge coverage in domains like pharmaceuticals and political science. In contrast, general-purpose datasets like T-Rex and WD50K are right-skewed, with most entities scoring 1.0, indicating substantial knowledge coverage in Wikipedia-based knowledge. Comparing MVPKG and its temporal variant, MVPKG w/o time, we observe an increase in the proportion of entities with zero knowledgeability and a decrease in those scoring 1.0. This indicates challenges of LLMs in understanding time-sensitive knowledge (Yuan et al., 2024).

Finding 2 - Figure 2(b)/(c) presents the node homophily distribution and the average graph homophily across several knowledge graphs. In Figure 2(b), these distributions are all right-skewed, with a peak around 0.8, suggesting that nodes and their neighbors tend to share similar knowledgeability scores. This high homophily property has enhanced graph machine learning in node-level prediction, such as node classification, and inspires our regression to predict entities' knowledge scores in Finding 3. Furthermore, incorporating temporal information into MVPKG results in a slight shift to the left, indicating decreased neighbor score similarity. This shift indicates that the temporal dimension introduces greater complexity and finer knowledgeability distinctions between the nodes and their neighbors. In addition, we compute the average graph homophily by averaging across all nodes and find that it consistently remains above 0.5 across different datasets and LLMs. This exhibits a general tendency for entities to be connected to others with similar knowledgeability scores. This finding reinforces the notion that the LLM's factual recognition is not randomly distributed in the graph but is instead correlated among connected entities.

Finding 3 - Figure 2(d) illustrates the relation between entity degree and knowledgeability. We observe a clear positive correlation, indicating that entities with higher degrees tend to exhibit greater knowledgeability in LLMs. This trend likely arises because high-degree entities are associated with more factual content and appear more frequently in pre-training corpora, increasing their likelihood of being learned during the LLM pre-training process. This observation aligns with findings showing accuracy disparities between popular and less popular entities (Sun et al., 2024). Notably, on the T-Rex dataset, the positive relationship remains but is much less pronounced. This is likely because T-Rex exclusively contains Wikipedia entities, which are generally well represented in LLM training corpora, even for less popular or low-degree entities.

Table 1: Regression of predicting node knowledgeability calculated by (1 - Mean Absolute Error between ground-truth and estimated knowledgeability scores). N/T-X represents the model X with input features being one-hot encoding (N)/textual embedding (T). The best performance is **bolded** and the second best is underlined.

Model	T-Rex	WD50K	Pharm	MVPKG(w/o t)	CoDEx
N-MLP	81%	78%	82%	72% (70%)	84%
N-GCN	84%	**82%**	**84%**	76% (76%)	87%
N-SAGE	84%	82%	84%	76% (77%)	**87%**
T-MLP	83%	78%	83%	76% (77%)	86%
T-GCN	84%	81%	84%	78% (**80%**)	87%
T-SAGE	**84%**	81%	84%	**78%** (79%)	87%



(a) (b)

Figure 3: Relation between regression performance and homophily at **(a)** graph and **(b)** node level.

Finding 4 - Table 1 demonstrates strong regression in predicting node knowledgeability, with absolute errors between 0.15 and 0.25. Comparing models using textual embeddings versus one-hot encodings reveals no consistent performance advantage, indicating that textual similarity between entities does not reliably reflect similarity in knowledgeability. In contrast, GNN-based models consistently outperform their MLP-based counterparts, underscoring the importance of incorporating neighborhood context for knowledgeability prediction. This result aligns with previous findings on the benefits of homophily in relational learning (Ma et al., 2021; Mao et al., 2023). Figure 3(a) visualizes a positive correlation between average regression performance and global graph homophily. However, in Figure 3(b), while this trend holds for T-Rex, WD50K, and CoDEx-S, it is less apparent for PharmKG and MVPKG, suggesting that the effect of homophily may be dataset-dependent.

Application - We demonstrate a practical application of GNN-predicted knowledgeability scores to guide the selection of informative triplets for more effective LLM fine-tuning. Specifically, we fine-tune three LLMs, LLaMA 3 8B, Mistral 7B, and Qwen 2.5 7B, across five datasets using two triplet selection strategies: Random-FT and Graph-FT. Both start by selecting the same initial 20% of triplets for knowledge probing. Random-FT then randomly selects the remaining 80%, while Graph-FT trains a GNN on the initial 20% to estimate entity-level knowledgeability and selects additional triplets involving entities predicted to be less known (i.e., with lower knowledgeability scores). All experiments use identical hyperparameters within each dataset, differing only in the triplet selection strategy.

Table 2: Performance comparison between fine-tuning with random triplet selection (Random-FT) and with knowledgeability-based selection (Graph-FT), where triplets are ranked from high to low based on estimated knowledgeability. The best performance is **bolded** and the second best is underlined.

Dataset	Model	Base	Random-FT	Graph-FT
T-Rex	Llama3 8B	63.25	86.40	**89.05**
	Mistral 7B	63.95	81.85	**91.90**
	Qwen2.5 7B	56.05	**84.80**	83.25
Pharm	Llama3 8B	17.80	34.85	**36.95**
	Mistral 7B	55.30	41.30	**60.70**
	Qwen2.5 7B	39.50	70.20	**74.40**
WD50	Llama3 8B	54.75	57.75	**58.75**
	Mistral 7B	42.87	**56.25**	55.12
	Qwen2.5 7B	49.37	63.00	**64.75**
MVPKG w/o t	Llama3 8B	26.10	30.70	**44.50**
	Mistral 7B	52.30	65.10	**76.70**
	Qwen2.5 7B	37.60	41.30	**65.10**
CoDEx	Llama3 8B	64.87	**78.75**	75.62
	Mistral 7B	58.50	72.12	**88.00**
	Qwen2.5 7B	62.37	67.00	**70.87**
Average Performance		49.64	62.09	**69.04**

In Table 2, both Random-FT and Graph-FT outperform the base models across all datasets. Notably, graph-FT consistently outperforms random-FT, underscoring the benefit of checking triplets with which the model is less familiar rather than redundantly reinforcing known knowledge.

4 Conclusion

This work introduces a novel graph-centric perspective by quantifying LLM knowledge at the triplet/entity levels and examining its relationship with graph structural properties. We uncover key insights, including a strong correlation between node degree and knowledgeability, and a high degree of homophily, where topologically close nodes exhibit similar knowledgeability. These observations motivate the design of a graph machine learning model utilizing neighborhood information to predict entity-level knowledgeability. The predicted scores are then used to actively select more informative triplets for effective fine-tuning LLMs.

5 Limitations

The limitations of this paper are as follows:

- **More applications**: The derived structural patterns are used solely to guide triplet selection for fine-tuning. However, these patterns hold broader potential. For instance, they could inform knowledge graph retrieval by identifying poor knowledge regions and prioritizing retrieving triplets there. Furthermore, this technique can also efficiently identify knowledge deficiency through structural correlations (Song et al., 2025).
- **Limited to knowledge graphs**: The derived structural patterns currently apply only to knowledge graphs with explicitly defined entities and relations. However, real-world networks, such as social or citation networks, are often more complex and rich in textual information. Extending the entity/triplet-level knowledgeability estimation to these text-attributed graphs (Wu et al., 2024) would broaden real-world applications.

References

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L Edelman. 2024. Distinguishing the knowable from the unknowable with language models. *arXiv preprint arXiv:2402.03563*.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Jeffrey R Binder, Rutvik H Desai, William W Graves, and Lisa L Conant. 2009. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, 19(12):2767–2796.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

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)*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

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

Mona M Garvert, Raymond J Dolan, and Timothy EJ Behrens. 2017. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *elife*, 6:e17086.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Qiyuan He, Yizhong Wang, and Wenya Wang. 2024. Can language models act as knowledge bases at scale? *arXiv preprint arXiv:2402.14273*.

Benjamin Heinzerling and Kentaro Inui. 2020. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. *arXiv preprint arXiv:2008.09036*.

Renske S Hoedemaker and Peter C Gordon. 2017. The onset and time course of semantic priming during rapid recognition of visual words. *Journal of Experimental Psychology: Human Perception and Performance*, 43(5):881.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Shengbin Jia, Yang Xiang, Xiaojun Chen, and Kun Wang. 2019. Triple trustworthiness measurement for knowledge graph. In *The World Wide Web Conference*, pages 2865–2871.

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*.

Nancy Kanwisher, Josh McDermott, and Marvin M Chun. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11):4302–4311.

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

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*.

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*.

Yao Ma, Xiaorui Liu, Neil Shah, and Jiliang Tang. 2021. Is homophily a necessity for graph neural networks? *International Conference on Learning Representations*.

Haitao Mao, Zhikai Chen, Wei Jin, Haoyu Han, Yao Ma, Tong Zhao, Neil Shah, and Jiliang Tang. 2023. Demystifying structural disparity in graph neural networks: Can one size fit all? *Advances in neural information processing systems*, 36:37013–37067.

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

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*, pages 2603–2614.

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

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)*, pages 2463–2473.

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

Thorsten Rings, Timo Bröhl, and Klaus Lehnertz. 2022. Network structure from a characterization of interactions in complex systems. *Scientific Reports*, 12(1):11742.

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)*, pages 5418–5426.

Tara Safavi and Danai Koutra. 2020. Codex: A comprehensive knowledge graph completion benchmark. *arXiv preprint arXiv:2009.07810*.

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*.

Linxin Song, Xuwei Ding, Jieyu Zhang, Taiwei Shi, Ryotaro Shimizu, Rahul Gupta, Yang Liu, Jian Kang, and Jieyu Zhao. 2025. Discovering knowledge deficiencies of language models on massive knowledge base. *arXiv preprint arXiv:2503.23361*.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 311–325.

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

Shirley Wu, Shiyu Zhao, Michihiro Yasunaga, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis Ioannidis, Karthik Subbian, James Y Zou, and Jure Leskovec. 2024. Stark: Benchmarking llm retrieval on textual and relational knowledge bases. *Advances in Neural Information Processing Systems*, 37:127129–127153.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 1963–1974.

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

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 arXiv:2312.11539*.

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*.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. 2020. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in neural information processing systems*, 33:7793–7804.

Yanxu Zhu, Jinlin Xiao, Yuhang Wang, and Jitao Sang. 2024. Kg-fpq: Evaluating factuality hallucination in llms with knowledge graph-based false premise questions. *arXiv preprint arXiv:2407.05868*.

A Appendix

B Related Work

B.1 LLM as Knowledge Base (KB)

(Petroni et al., 2019) was among the first works to propose that pretrained LMs encode factual knowledge retrievable via cloze prompts. Subsequent work such as (Roberts et al., 2020) fine-tuned LLM for closed-book QA to match external knowledge systems, (Heinzerling and Inui, 2020) supported LMs as KBs by examining entity representations and paraphrase robustness, and (He et al., 2024) demonstrated that LLMs trained on large-scale data could flexibly retrieve information, further bolstering the concept of LLMs as knowledge bases. This motivates the research on checking knowledge of LLMs as follows.

B.2 Knowledge Checking

To further evaluate this paradigm of LLM as KB, various knowledge checking methods have been developed, such as, factuality testing with TruthfulQA benchmark (Lin et al., 2021), consistency and reliability (Zheng et al., 2024), calibration with self-assessed P(True) and P(I Know) (Kadavath et al., 2022), information-theoretic probing using entropy and KL-divergence (Pezeshkpour, 2023), systematic KG-based evaluation via auto-generated QA from graphs (Luo et al., 2023), and evaluating factuality hallucinations by using false premise questions (Zhu et al., 2024). These approaches look into knowledge and trustworthiness checking but treat the model as a black box, leaving its underlying structural patterns unexplored.

B.3 Topological Understanding of LLM-KB

Some important initial work has looked into local structures of LLMs. (Geva et al., 2020) presented that feed-forward layers act like key–value memories for specific facts. Then (Meng et al., 2022) presented that factual associations are often localized and editable within mid-layer feed-forward modules. (Dai et al., 2021) proposed that factual knowledge is stored in pretrained Transformers in form of knowledge neurons. (Mruthyunjaya et al., 2023) evaluated LLMs on structural indicators such as, symmetry, hierarchy and path among others and show that they often fail on relational tests. These studies demonstrate that some implicit structure exists and yet none characterizes the graph topology or structural patterns of an LLM's knowledge base.

C Dataset Statistics

Our experiments are designed to evaluate and compare the knowledgeability of the LLM across multiple datasets. We illustrate our process on five datasets: **MVPKG** (covering U.S. legislative, election, diplomatic data, etc.), **T-Rex** (containing large-scale high-quality alignments between DBpedia abstracts and Wikidata triples), **PharmKG8K** (biomedical knowledge graph), **WD50K** (dataset derived from Wikidata statements), and **CoDEx-S** (extracted from Wikidata and Wikipedia).

- **MVPKG** (Mou et al., 2024): The MVPKG dataset encompasses U.S. legislative, election, and diplomatic data as well as conceptual knowledge from Wikidata. It originally contains 1,857,410 triplets, 137,117 entities, and 602 relations. Due to scale considerations, we extract the largest strongly connected component, which comprises 255,697 triplets, 9,055 entities, and 602 relations. The MVPKG dataset had a temporal attribute and was evaluated with the temporal component included and excluded. For each triplet, two prompts are generated (with time and without time). Consequently, each entity in MVPKG is assigned two knowledgeability scores corresponding to the two prompt variants for further analysis of the effect of inclusion of temporal information. All other datasets have only one knowledgeability score due to lack of temporal attributes.

- **T-Rex** (Elsahar et al., 2018): The T-Rex dataset is constructed from Wikipedia abstracts aligned with Wikidata entities in English. It contains 6,566,790 unique triplets; the largest connected component comprises 193,781 triplets, 46,891 entities, and 423 relations.

- **PharmKG8K** (Zheng et al., 2021): The PharmKG8K multi-relational, attributed biomedical KG, composed of around 500,000 individual interconnections between genes, drugs, and diseases, with 29 relation types over a vocabulary of around 8000 disambiguated entities. Given the scope of the dataset, we used a strongly connected component of 98,537 edges, 6,877 entities, and 29 relations.

- **WD50k** (Galkin et al., 2020): The WD50K dataset was created using the Wikidata RDF dump of August 2019. It has 233,838 edges and

Table 3: Statistics of the original knowledge graph and the sampled largest connected component.

Dataset	# Nodes		# Triplets		# Avg. Deg		# Avg. CC	
	Original	Sampled	Original	Sampled	Original	Sampled	Original	Sampled
T-Rex	3153568	46891	6566790	193781	4.16	8.26	0.1473	0.5170
WD50K	41334	5140	233838	34208	11.31	13.31	0.0996	0.1332
PharmKG8K	7262	6877	479902	98537	132.16	28.65	0.2512	0.0824
MVPKG	137117	9055	1857410	255697	12.46	28.24	0.0013	0.0140
MVPKG w/o t	137117	9055	1857410	116127	12.46	12.82	0.0013	0.0140
CoDEx-S	2034		36543		35.93		0.0952	

41,334 entities. Since being extracted from Wikidata, there were 14,858 triplets common between the WD50K dataset and the T-Rex largest connected component selected. These were removed to make sure that common triplets were not overshadowing the result comparison between these datasets. Following that, the largest strongly connected component was selected for experimental purposes. This LCC had 34,208 edges, 5,140 entities, and 193 relations.

- **CoDEx-S** (Safavi and Koutra, 2020): CoDEx is a collection of knowledge graph completion datasets extracted from Wikidata/Wikipedia, comprising three subsets of varying sizes. We select CoDEx-S due to its high proportion of triplets involving the “occupation" relation, which poses greater challenges for LLMs, since individuals may hold multiple occupations. CoDEx-S contains 36,543 triplets, 2,034 entities, and 42 relations.

D Experimental Setting

We describe the experimental setup for (1) measuring the triplet and entity knowledgeability, (2) training GNNs to predict knowledgeability scores, and adaptively selecting informative triplets to fine-tune LLMs.

D.1 Measuring Knowledgeability Score

- **Prompt Generation:** Each triplet is converted into a natural language prompt using predefined templates based on the relation type, following (Petroni et al., 2019). These templates were first generated by GPT o-1 mini using the relation and a few of its triplet examples to provide context, and then evaluated to make sure the template made semantic sense. (Luo et al., 2023) used GPT3.5 for generating natural language prompts for triplets, validating that LLMs like GPT3.5 can be used for template or prompt generation. For MVPKG, both time-specific and non-time-specific prompts are created.

- **LLM Evaluation:** The prompts are fed to the LLM, and responses are recorded as binary values (1 for true, 0 for false). This step enables us to quantify the LLM’s internalized knowledge regarding each triplet in a way that’s scalable.

- **Aggregation to Entity-Level Scores:** For every entity, triplet-level scores are aggregated to form the entity-level knowledgeability metric. In MVPKG, separate aggregations are performed for the two prompt types, giving two knowledgeability values for each entity.

D.2 Fine-Tuning: Random VS Graph

The goal of this experiment is to evaluate whether fine-tuning LLMs on entities for which the model has low prior knowledge results in greater performance improvements than fine-tuning on randomly selected entities. We hypothesize that targeting entities about which the model knows less will produce a larger marginal improvement per example than fine-tuning on entities already well encoded in LLM’s internal knowledge inherited during the pre-training phase.

- **Model and Evaluation Set:** To test this, we select three open source models: Llama 3.1 8B, Mistral v0.3 7B, and Qwen 2.5 7B, and constructed an evaluation set for each dataset by randomly sampling a fixed number of triplets. Each triplet is converted into a natural language prompt and is asked to LLM as a True/False evaluation task. Baseline performance is measured by querying each base model on this evaluation set prior to any fine-tuning. The performance metric is the percentage of correct responses by the model on the evaluation set.

- **Fine-Tuning Budget and Initial Query:** We then set a budget that the LLM can be fine-tuned

on, and the size of this budget is adjusted according to the domain and size of the dataset. Twenty percent of the budget is reserved for an initial query set. To set up this initial set, we shuffle the entity list and iterate through it, adding all triples associated with the current entity until the 20% quota is met and if an entity would overshoot the quota, we randomly subsample just enough of its triples to fill the gap.

- **Graph Fine-Tuning:** The triplets in this initial query set are posed to the base model, allowing us to calculate an entity-level knowledgeability score for the selected entities in the initial query. These entity scores are used to train a GraphSAGE model. The model takes text embeddings of entity names generated using the MiniLM-L6-v2 sentence transformer as input to predict knowledgeability scores for all the entities across the dataset. Further, we define an entity's "ignorance" as one minus its predicted knowledgeability. Entities with the highest ignorance are preferred for fine-tuning, and ties are broken first by choosing the entity with the lowest graph degree, to encourage topical diversity, and finally at random. We iteratively add entities and their associated triplets until 80% of the budget is filled. In case an entity's full triplet set would overshoot the remaining slots, we randomly sample within that set to exactly meet the quota. The full Targeted training set thus comprises the initial 20% query triples plus the 80% ignorance-weighted triplets.
- **Random Fine-Tuning:** For the Random Fine-Tuning, we retain the initial 20% query set and additionally randomly sample the remaining 80% of triplets from all unprobed triplets without replacement. This yields a direct random selection comparison to the targeted method.

E Results across different LLMs

E.1 Llama 3.3 70B

See Figure 4 for an overview of model behavior.

- **Knowledgeability Distribution:** Similar to GPT3.5 results, Llama 3.3 70B has a trimodal pattern in the knowledgeability distribution, with domain-specific datasets having higher peaks at 0 while general datasets like T-Rex, which are extracted from Wikipedia, have higher peaks at 1. Peak at 0.5 is largely made of entities with degree 2 where one triplet is evaluated as true while the other one as false.
- **Homophily Distribution:** All datasets have homophily peak at 0.8 and above indicating that nodes and their neighbors tend to share similar knowledgeability scores. We observe overall a higher homophily on the general domain datasets than domain specific ones.
- **Degree vs Knowledgeability:** We observe that all datasets overall have a positive trend between the mean knowledgeability value and mean log degree. Biomedical dataset PharmKG8K has a higher upward trend, while MVPKG has a much shallower trend. This might be attributed to the T-Rex dataset's origin from Wikipedia entities which are well covered by pre-training corpora.

E.2 Deepseek V3

See Figure 5 for an overview of model behavior.

- **Knowledgeability Distribution:** We observe a trimodel pattern with a relatively small peak at 0.5. Entities with a knowledge value of 0 are more common than those with a value of 1, especially in domain-specific datasets. For general datasets like T-Rex, WD50K, and CoDEx-S, a larger proportion of their entities are still recognized by Deepseek, resulting in a higher peak in the number of entities with full knowledgeability.
- **Homophily Distribution:** Homophily for entities across datasets has the highest density at around 0.8, indicating that entities and their neighbors tend to share similar knowledgeability scores. Here, no specific datasets appear to have a clear advantage over others.
- **Degree vs Knowledgeability:** All datasets show a clear positive trend between the degree of entity and their Knowledgeability. T-Rex here has a slightly steeper trend than both GPT 3.5 and Llama 3.3 70 B.

E.3 Gemini 2.5 Flash

See Figure 6 for an overview of model behavior.

- **Knowledgeability Distribution:** The general-domain datasets continue to have a higher proportion of entities with a knowledgeability score of 1, resulting in a right-skewed distribution. In contrast, domain-specific datasets show a higher proportion of entities with a knowledgeability score of 0. A notable improvement of Gemini is that PharmKG8K has a more balanced distribution compared to the other models, like Llama 3.3

70B, GPT3.5, and Deepseek V3. This indicates that it has better knowledge about biomedical-related entities. Although there is still some left skew, it is significantly less pronounced.

- **Homophily Distribution:** Similar to other models, highest homophily density stays around 0.8, suggesting that nodes tend to have similar knowledgeability scores as their neighbors. T-Rex has a homophily to the furthest right, further indicating the nodes have very similar knowledge values to their neighbors.

- **Degree vs Knowledgeability:** A positive trend is observed across all datasets, with each showing an upward-sloping pattern. T-Rex, while following this trend, displays a relatively shallow slope, consistent with the behavior seen in other models, due to it being derived from Wikipedia.

E.4 GPT 4o

See Figure 7 for an overview of model behavior.

- **Knowledgeability Distribution:** GPT-4o demonstrates a higher level of entity knowledgeability across all domains compared to other models. Even in domain-specific datasets like PharmKG8K, GPT-4o recognizes a larger proportion of entities than it does not.

- **Homophily Distribution:** Here, datasets display a high level of homophily, with the highest density peaks being greater than 0.8. Following the pattern across the models, the T-Rex dataset presents the highest homophily among all the other datasets.

- **Degree vs Knowledgeability:** All datasets exhibit an upward trend, suggesting that as the degree associated with an entity increases, so does its knowledgeability.

F KG vs Topology Analysis across models

For each node, we calculate its corresponding graph structural properties and group them based on these properties. For each group, we further calculate the average knowledge and visualize its relation with structural properties. See Figure 8 for GPT-4o; Figure 9 for Llama 3.3 70B; Figure 10 for Gemini 2.5; and Figure 11 for DeepSeek V3 for an overview; Figure 12 for GPT3.5.

- **Degree Centrality:** We observe a general positive upward trend among the mean knowledge value and degree centrality across the models. A high degree node would appear in many facts and would appear in large amount of training corpus. Therefore, if sample those corpus for training the model, that entity would show up at more places and the model would get more examples of the entity, and thus learning about it better.

- **PageRank Centrality:** Here, across the models we observe a positive trend. WD50K displays a large variance towards the top. Since, the bins there contain few entities, variance is presented as large in case of any outlier.

- **Katz Centrality:** We observe a positive trend among 4 out of 5 datasets. WD50K creates a upside down u shape slope with some outliers, along with high variance. This can potentially be attributed to a few entities in the last few bins presenting an increased variance and unexpected behavior.

- **Cluster Centrality:** Across the models we see a positive trend between the mean knowledgeability value and the cluster centrality. This can potentially be caused by the fact that a higher clustering would mean that entity is part of a dense group and would be mentioned over and over whenever the context of that group comes up. However, the rate is less pronounced in some than in others. For example, GPT 4o has a stronger relationship trend than Deepseek V3. T-Rex, for all the models has a very slight but positive trend, mostly staying relatively flat.

- **Closeness Centrality:** Here, the results vary the most. For GPT 4o, almost all datasets have a U-shape, indicating that both peripheral and central nodes get higher knowledgeability values. In contrast, the Llama model has a relatively minor U-shape effect, with some datasets broadly staying flat, and for example, PharmKG8K showing an upward trend.

- **Between Centrality:** Here, datasets with a general domain like WD50K and T-Rex stay relatively flat, whereas domain-specific datasets, such as PharmKG8K, display a strong positive relation, indicating that entities that serve as hubs or bridges tend to have a higher knowledgeability score than nodes on the periphery.

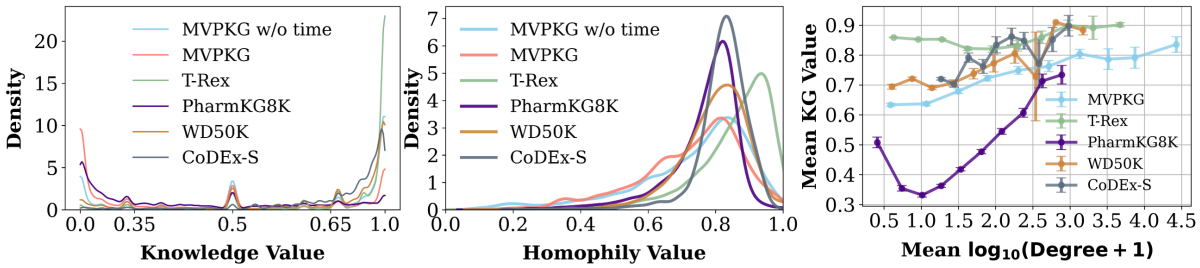


Figure 4: LLaMa **(a)**: Distribution of node knowledgeability for each dataset; **(b)**: Distribution of node homophily for each dataset; **(c)**: Node knowledgeability increases as node degree increases.

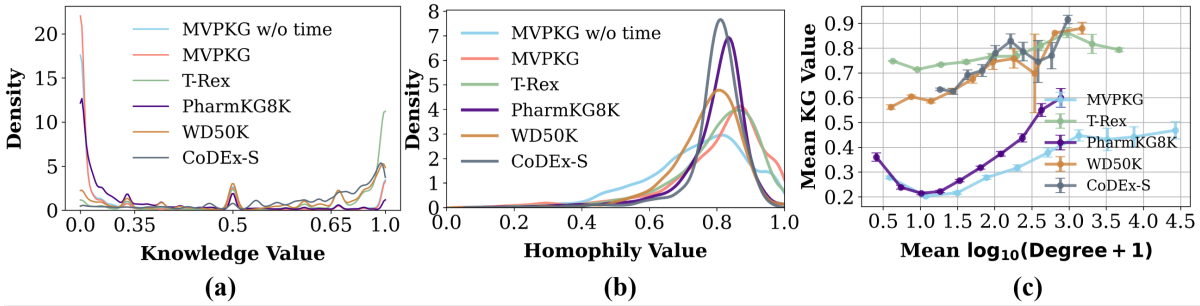


Figure 5: Deepseek **(a)**: Distribution of node knowledgeability for each dataset; **(b)**: Distribution of node homophily for each dataset; **(c)**: Node knowledgeability increases as node degree increases.

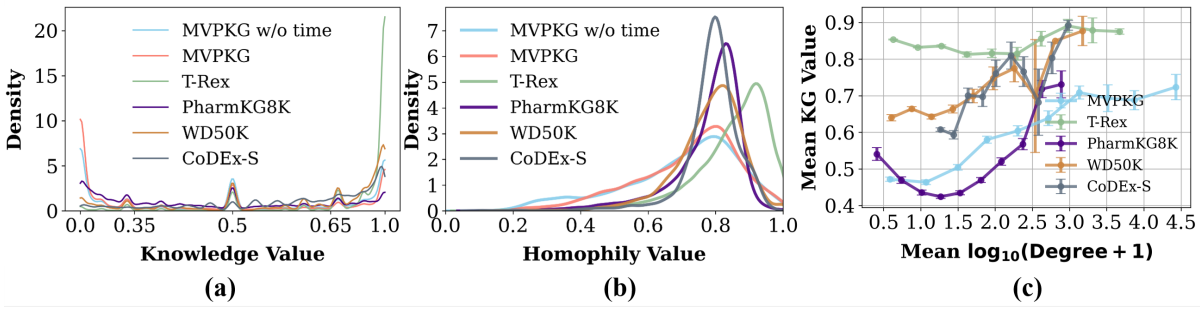


Figure 6: Gemini **(a)**: Distribution of node knowledgeability for each dataset; **(b)**: Distribution of node homophily for each dataset; **(c)**: Node knowledgeability increases as node degree increases.

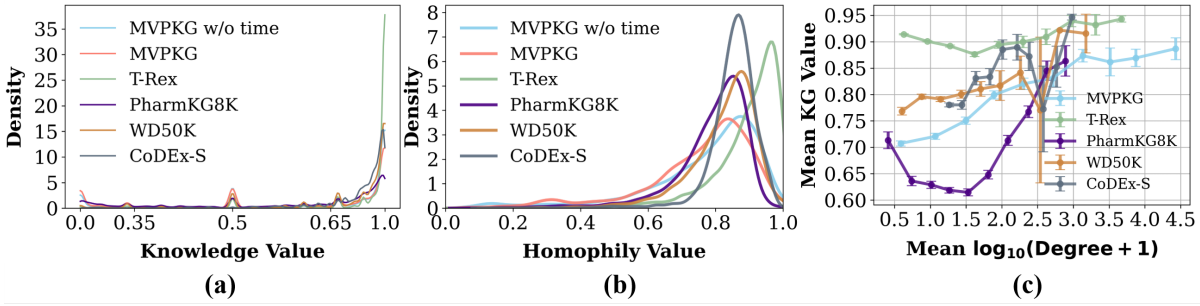


Figure 7: GPT4o **(a)**: Distribution of node knowledgeability for each dataset; **(b)**: Distribution of node homophily for each dataset; **(c)**: Node knowledgeability increases as node degree increases.

We observe that across the models and datasets, some patterns persist. For instance, positive relationship between node degree and their knowledgeability. In addition, pattern of high homophily showcase that nodes and their neighbors tend to have similar knowledgeability.

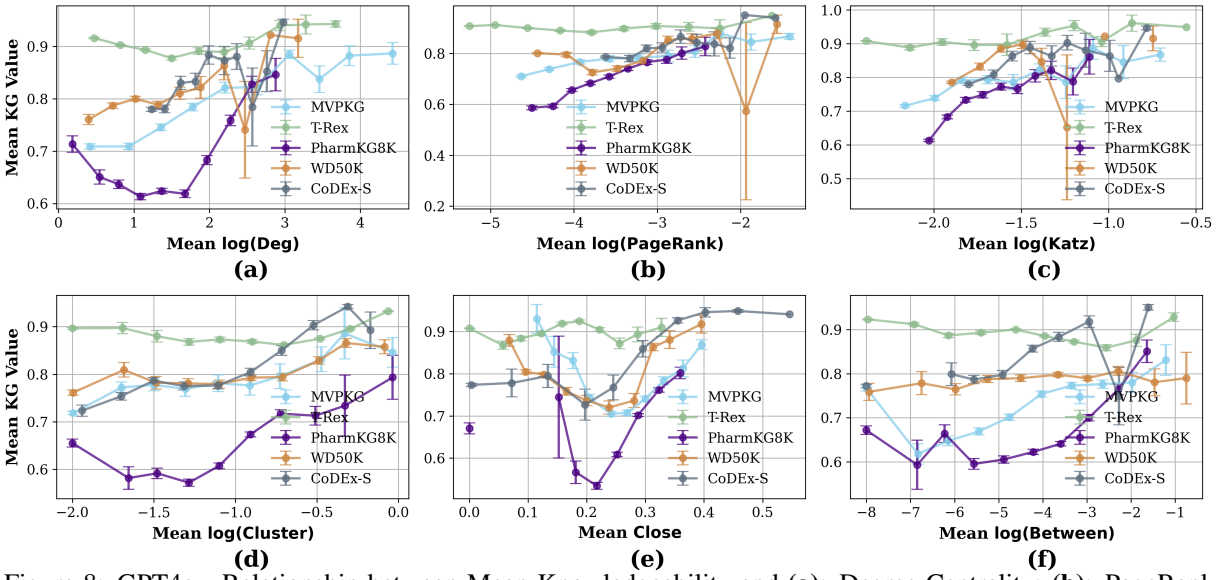


Figure 8: GPT4o - Relationship between Mean Knowledgeability and **(a)**: Degree Centrality; **(b)**: PageRank Centrality; **(c)**: Katz Centrality; **(d)**: Cluster Centrality; **(e)**: Closeness Centrality; **(f)**: Betweeness Centrality.

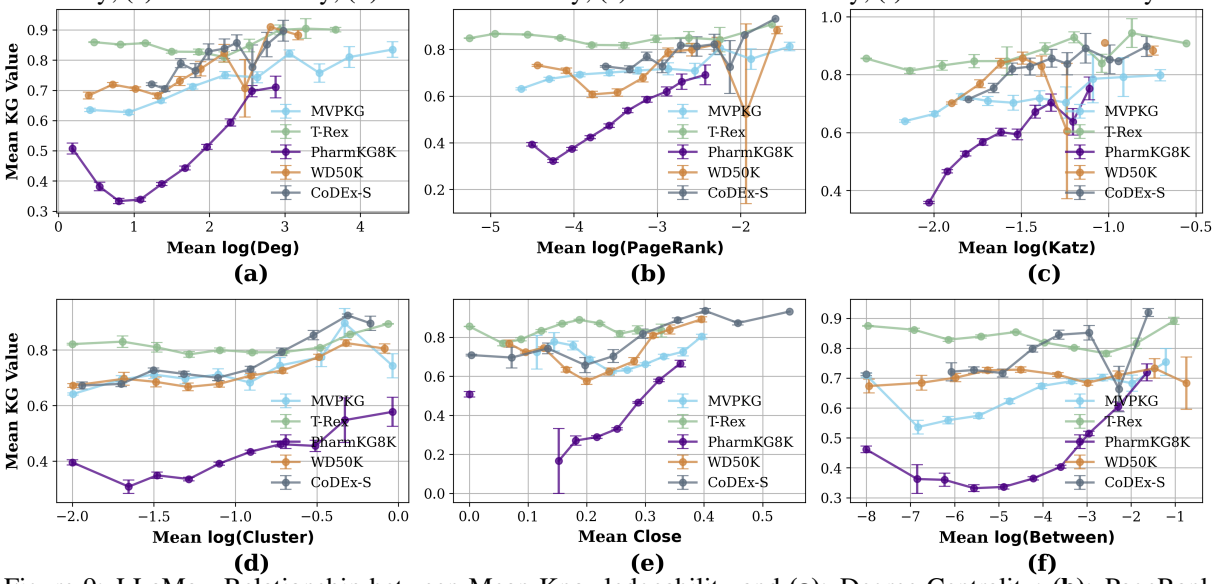


Figure 9: LLaMa - Relationship between Mean Knowledgeability and **(a)**: Degree Centrality; **(b)**: PageRank Centrality; **(c)**: Katz Centrality; **(d)**: Cluster Centrality; **(e)**: Closeness Centrality; **(f)**: Betweeness Centrality.

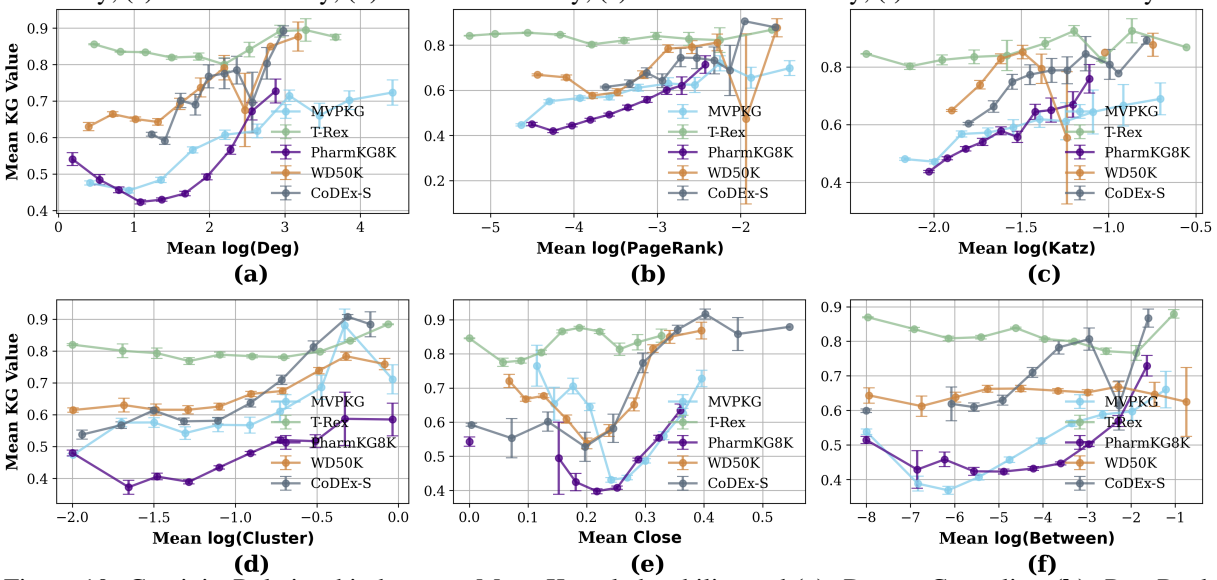


Figure 10: Gemini - Relationship between Mean Knowledgeability and **(a)**: Degree Centrality; **(b)**: PageRank Centrality; **(c)**: Katz Centrality; **(d)**: Cluster Centrality; **(e)**: Closeness Centrality; **(f)**: Betweeness Centrality.

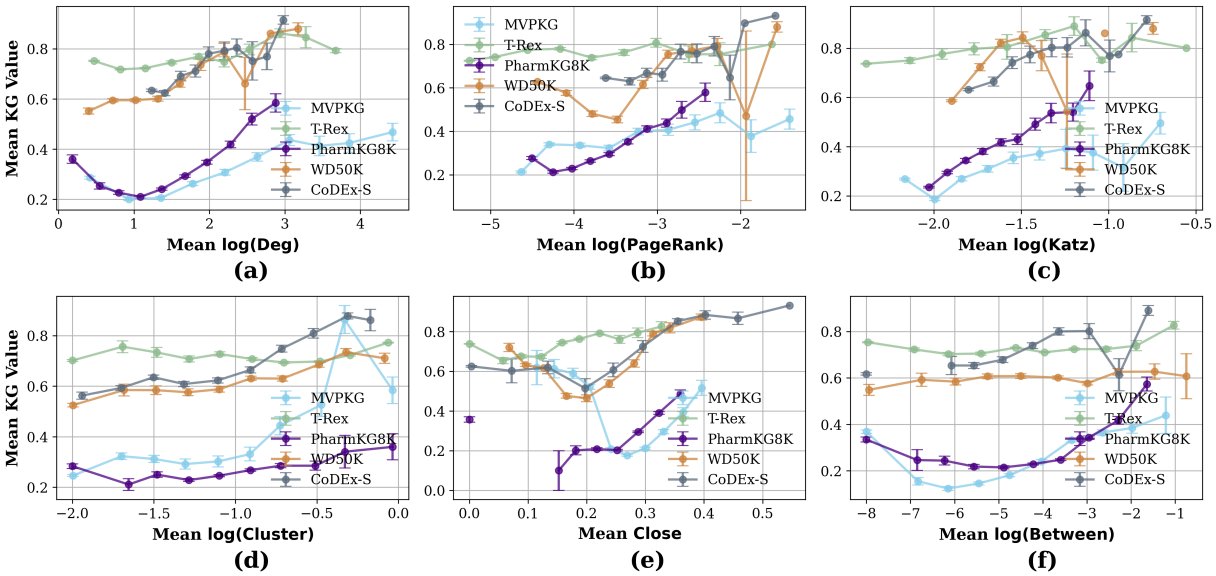


Figure 11: Deepseek - Relationship between Mean Knowledgeability and **(a)**: Degree Centrality; **(b)**: PageRank Centrality; **(c)**: Katz Centrality; **(d)**: Cluster Centrality; **(e)**: Closeness Centrality; **(f)**: Betweeness Centrality.

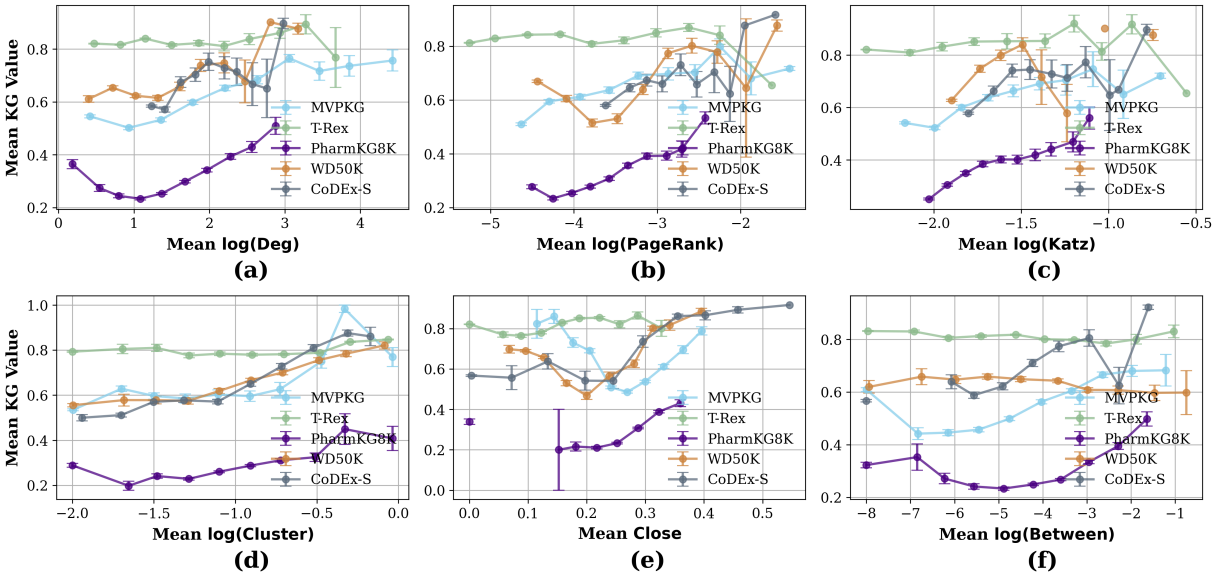


Figure 12: GPT3.5 - Relationship between Mean Knowledgeability and **(a)**: Degree Centrality; **(b)**: PageRank Centrality; **(c)**: Katz Centrality; **(d)**: Cluster Centrality; **(e)**: Closeness Centrality; **(f)**: Betweeness Centrality.

G Temporal LLM-based Triplet Evaluation Prompt

Prompt 2: LLM-based Triplet Evaluation (Temporal Variation)

System Message: Evaluate the statement below; reply only `True` or `False`.
Given: Triplet $\mathcal{T} = (sub,\ rel,\ obj)$, Date D.
Relational Template Map: $T:\ rel \mapsto$ "$\{sub\}\ \ldots\ \{obj\}$".
Procedure:

1. Retrieve template $t = T(rel)$.
2. Instantiate base statement $S_0 = t[\{sub\} \to sub,\ \{obj\} \to obj]$.
3. Append date: $S = S_0$ on D.
4. Send **System Msg + User Msg:** S to LLM.
5. Return "True" or "False."