

# TinyGraphEstimator: Adapting Lightweight Language Models for Graph Structure Inference

Michal Podstawski <sup>a</sup>

NASK National Research Institute, Warsaw, Poland  
michal.podstawski@nask.pl

## Keywords:

Graph Reasoning, Structural Inference, Small Language Models, Graph-Theoretic Properties, Efficient Adaptation

## Abstract:

Graphs provide a universal framework for representing complex relational systems, and inferring their structural properties is a core challenge in graph analysis and reasoning. While large language models have recently demonstrated emerging abilities to perform symbolic and numerical reasoning, the potential of smaller, resource-efficient models in this context remains largely unexplored. This paper investigates whether compact transformer-based language models can infer graph-theoretic parameters directly from graph representations. To enable systematic evaluation, we introduce the TinyGraphEstimator dataset - a balanced collection of connected graphs generated from multiple random graph models and annotated with detailed structural metadata. We evaluate several small open models on their ability to predict key graph parameters such as density, clustering, and chromatic number. Furthermore, we apply lightweight fine-tuning using the Low-Rank Adaptation (LoRA) technique, achieving consistent improvements across all evaluated metrics. The results demonstrate that small language models possess non-trivial reasoning capacity over graph-structured data and can be effectively adapted for structural inference tasks through efficient parameter tuning.

## 1 INTRODUCTION

Graphs are a fundamental representation for modeling complex relationships in domains such as social networks, biology, and information systems. Accurately characterizing these networks requires estimating key graph-theoretic properties - such as connectivity, clustering, and chromatic number - that capture their structure and complexity. Traditional graph-learning approaches, including Graph Neural Networks and algorithmic estimators, achieve strong results but depend on task-specific architectures and explicit structural features.

To address this gap, we investigate whether small transformer-based language models can be adapted to estimate structural graph parameters from textual representations of graphs. In our setup, each graph is encoded in a deterministic edge-list format, which serves as a textual

proxy for its topology. The models are trained to predict key graph-theoretic quantities from these structured inputs, enabling us to assess whether compact language models can approximate graph reasoning patterns when fine-tuned on explicit structural data, rather than relying on built-in algorithmic mechanisms or large model capacity.

To this end, we systematically evaluate several small, resource-efficient models on the task and fine-tune them using the Low-Rank Adaptation (LoRA) method. The results show that while these models exhibit limited capability in the zero-shot setting, fine-tuning leads to consistent and substantial improvements across all structural measures. These findings demonstrate that compact language models, when properly adapted, can acquire effective reasoning behavior over graph-structured data while maintaining high computational efficiency.

The main contributions of this paper are as follows:

---

<sup>a</sup>  <https://orcid.org/0000-0003-1222-6894>

- We present a systematic evaluation of small, open language models on the task of inferring graph-theoretic parameters from structured graph representations.
- We demonstrate that LoRA-based fine-tuning yields consistent performance improvements, highlighting the potential of compact models for graph analysis tasks.
- We introduce the TinyGraphEstimator dataset, a balanced and publicly available collection of connected graphs annotated with key structural properties for reasoning evaluation.

## 2 RELATED WORK

The task of predicting graph-theoretic properties has traditionally been addressed using specialized graph learning architectures such as Graph Neural Networks (GNNs) (Kipf and Welling, 2017; Hamilton et al., 2017; Wu et al., 2019). These models operate directly on adjacency structures to learn node embeddings and global representations that capture connectivity and topology. Numerous studies have explored the prediction of structural metrics including clustering coefficients, path lengths, and degree distributions from learned embeddings (You et al., 2018; Dwivedi et al., 2023). However, while GNNs provide strong inductive biases for graph tasks, they often lack general reasoning ability outside the training distribution. Our work diverges from these approaches by treating the inference of graph properties as a language-based reasoning problem, where information is expressed textually and interpreted by small language models.

Recent advances in large language models (LLMs) have demonstrated impressive performance in tasks requiring symbolic reasoning, numerical computation, and pattern generalization (Brown et al., 2020; Bubeck et al., 2023). Studies have shown that transformer-based architectures can internalize relational structures (Wang et al., 2025; Schmitt et al., 2021; Ying et al., 2021) and perform graph-related reasoning when provided with suitable prompts or structured input (Huang et al., 2024; Guo et al., 2023; Guo et al., 2025). Nevertheless, most prior work focuses on large-scale models with hundreds of billions of parameters, whose reasoning capabilities emerge from scale rather than explicit adaptation. In contrast, we investigate whether small language models - under

4B parameters - can infer precise graph-theoretic quantities. This shifts the focus from emergent reasoning in massive models to lightweight, interpretable reasoning in compact architectures.

A growing body of work explores how language models can represent or reason about graphs through textual or hybrid encodings. These approaches encode adjacency information as token sequences, enabling LLMs to reconstruct or classify graph structures (Tang et al., 2024; Chen et al., 2025; Podstawski, 2025). Other works have investigated relational reasoning benchmarks (Hu et al., 2025; Agrawal et al., 2024), showing that LLMs can approximate certain structural statistics with appropriate context. Our approach builds on this foundation but emphasizes the systematic evaluation of structural parameter inference - quantitative graph descriptors such as density, transitivity, and global efficiency - using small models specifically adapted for this purpose.

Fine-tuning language models for specialized reasoning tasks poses significant computational and memory challenges. Parameter-efficient adaptation techniques such as Low-Rank Adaptation (LoRA) (Hu et al., 2022), prefix tuning (Li and Liang, 2021), and adapter-based methods (Houlsby et al., 2019) have emerged as scalable alternatives that modify only a small subset of parameters while preserving base model capabilities. These techniques have been successfully applied across a range of domains including code understanding (Han et al., 2024), reasoning (Bi et al., 2025), and domain adaptation (Lu et al., 2024). In our study, we employ LoRA to specialize small language models for structural graph inference, demonstrating consistent performance improvements across all measured graph parameters. This confirms that lightweight fine-tuning can effectively enhance reasoning abilities even in compact architectures.

In contrast to prior work that primarily investigates either graph representation learning or large-scale emergent reasoning, our research unifies both directions by systematically testing the graph reasoning capacity of small, adaptable language models. By combining structured graph representations with efficient fine-tuning, our approach provides new insights into how compact LMs internalize relational structure and approximate graph-theoretic computations.

### 3 PROPOSED SOLUTION

The proposed approach evaluates and enhances the ability of small language models to infer a wide range of graph-theoretic parameters directly from structured graph representations. Each model operates on undirected, unweighted graphs whose topology is defined entirely by edge connectivity, and is prompted to predict a structured set of key structural properties that collectively describe graph topology and complexity. The inferred parameters include detailed degree-based statistics such as minimum degree, mean degree, maximum degree, and degree standard deviation, along with the global density metric describing overall connectivity. In addition to these fundamental attributes, we evaluate the models' ability to estimate higher-order structural metrics including the total number of triangles, average clustering coefficient, transitivity, average shortest path length, diameter, chromatic number, and global efficiency. Taken together, these parameters capture both local and global aspects of graph organization, providing a comprehensive benchmark for assessing reasoning over graph structures.

In the first stage of our study, we test unmodified (plain) small language models to evaluate their intrinsic ability to perform numerical reasoning and pattern recognition over graph-structured data. This zero-shot evaluation establishes a baseline for each model's inherent capacity to approximate structural graph measures from textual inputs. In the second stage, we apply lightweight fine-tuning using the Low-Rank Adaptation (LoRA) method, which enables efficient parameter adjustment without retraining the full model. This adaptation significantly improves performance across all evaluated parameters, demonstrating that even compact models can be effectively specialized for graph inference through minimal, targeted fine-tuning. The resulting framework provides a scalable and efficient solution for structural property estimation of graphs, highlighting the latent reasoning capabilities of small language models when guided by domain-specific adaptation.

#### 3.1 Models Selection

To investigate the ability of small language models to infer structural properties of graphs, we selected a group of compact yet high-performing models that combine efficient deployment with

strong reasoning capacity. Our goal was to identify models that are powerful enough to capture graph-theoretic patterns while remaining computationally accessible for systematic experimentation. Model selection was guided by comparative performance metrics reported on the LLM-Stats leaderboard (LLM Stats, 2024), focusing on models with fewer than 4B parameters that demonstrate competitive reasoning accuracy across diverse benchmarks.

The final selection includes Qwen-2.5-3B (Yang et al., 2025), Llama-3.2-3B (Grattafiori et al., 2024), and Phi-4-mini (4B) (Abdin et al., 2024), all used in their Instruct variants. These models were chosen because they consistently rank among the strongest sub-4B architectures in reasoning and comprehension tasks while maintaining full open availability for reproducible research. Their extended context capacity allows complete representations of graphs to be processed in a single inference window. Moreover, their compact architecture enables efficient fine-tuning and evaluation on standard hardware, making them ideal candidates for exploring how small-scale language models internalize and generalize graph-structural information.

For reference, we also evaluate two state-of-the-art LLMs - GPT-4.1 (OpenAI, 2024; Achiam et al., 2024) and DeepSeek-V3.2-Exp (Non-Thinking Mode) (DeepSeek, 2025; Liu et al., 2025) - to benchmark our compact models against the current frontier of LLM capability. While these systems exhibit strong general reasoning and broad-domain adaptability, our fine-tuned small models achieve superior accuracy on graph-structural inference, highlighting the efficiency and specialization benefits of lightweight adaptation.

#### 3.2 Dataset Construction

To the best of our knowledge, no existing dataset directly supports the task of mapping raw graph structures to such a wide range of quantitative parameters. Therefore, we constructed a balanced synthetic dataset of connected graphs generated from three canonical random graph models: Erdős-Rényi (ER), Barabási-Albert (BA), and Watts-Strogatz (WS). The ER model produces graphs where each possible edge occurs independently with a fixed probability, capturing purely random connectivity (Erdős and Rényi, 1959). The BA model generates scale-free networks through preferential attachment, empha-

sizing hub formation and heavy-tailed degree distributions (Barabási and Albert, 1999). The WS model interpolates between regular lattices and random graphs, combining high clustering with short path lengths (Watts and Strogatz, 1998).

For each generated instance, the number of nodes  $n$  is uniformly sampled from the range  $[20, 30]$ , and connectivity is enforced by unbiased resampling until a connected instance is obtained, avoiding bias introduced by artificial augmentation. Each model contributes an equal share of graphs, ensuring diversity in structural characteristics such as degree distribution, clustering, and connectivity patterns.

For reproducibility, the following parameter ranges were used during dataset generation. In the ER model, the edge probability  $p$  was sampled from the interval  $[\log(n)/n + 0.01, 0.35]$ . In the BA model, the number of edges to attach from a new node  $m$  was sampled from  $[1, \min(6, n - 1)]$ . In the WS model, the even neighborhood size  $k$  was drawn from  $[2, \min(n - 2, 12)]$  and the rewiring probability  $\beta$  from  $[0.05, 0.35]$ . Each model was sampled uniformly across its respective parameter range, producing connected, non-degenerate graphs suitable for evaluating structural inference capabilities.

Graphs are stored accompanied by metadata that capture key structural properties. Each graph is represented as an edge list, where every line specifies a pair of connected node identifiers. The dataset consists of 1,200 graphs for training and 120 graphs for testing, evenly distributed across the three models. This design provides a well-controlled yet structurally varied benchmark for assessing the generalization and reasoning abilities of lightweight language models in inferring global and local graph parameters.

All generated graphs, along with their metadata and manifest files, constitute the Tiny-GraphEstimator dataset, one of the key contributions of this work. It is publicly accessible<sup>1</sup> to support reproducibility and further research on graph-structured reasoning in language models.

### 3.3 Training Setup

We fine-tune each base model with parameter-efficient supervised instruction tuning using the TRL SFTTrainer (von Werra et al., 2020) and LoRA adapters. Training data consist of paired prompt-completion examples: the prompt includes the normalized graph (edge list), the fixed

target schema, and concise instructions; the completion is the gold JSON with expected fields. To avoid learning spurious prompt tokens and to sharpen output formatting, we use completion-only loss. LoRA adapters are applied to attention and MLP projection modules with rank  $r=32$ ,  $\alpha=32$ , and dropout 0.05. We train with sequence length 2048, batch size 2 and gradient accumulation 8, cosine learning-rate schedule with warmup ratio 0.03, learning rate  $2 \times 10^{-4}$ , and ten epochs over the training set.

**Compute Configuration** All experiments (training and validation) were executed on a single NVIDIA RTX 3090 GPU (Ampere, 24GB VRAM) in a standard Linux x86\_64 environment.

### 3.4 Validation Protocol

We evaluate model outputs with a schema- and tolerance-aware validation pipeline that mirrors the training prompt format while remaining model-agnostic. For each sample, the script loads the graph data and gold metadata and normalizes it to a canonical textual form. A structured prompt then instructs the model to return only a JSON object matching a fixed schema of given fields (density, degree statistics, triangles, clustering, transitivity, path lengths, diameter, chromatic number, global efficiency). Inference is run on a base model or a base model augmented with LoRA adapters (if provided), with JSON-constrained decoding via `lm-format-enforcer` (Gat, 2024) to prevent malformed outputs. The first JSON object in the response is extracted and strictly checked against the schema.

### 3.5 Evaluation Metrics

Performance was evaluated using four standard regression metrics: the normalized root mean squared error based on range ( $\text{NRMSE}_{\text{range}}$ ), the symmetric mean absolute percentage error (sMAPE), the  $R^2$  Accuracy, and the NRMSE Accuracy.

$\text{NRMSE}_{\text{range}}$  represents the root mean squared error normalized by the range of the target variable, providing a scale-independent measure of average deviation between predicted and true values. sMAPE quantifies the mean relative difference between predictions and ground truth, computed symmetrically with respect to both magnitudes to balance over- and underestimation.  $R^2$

<sup>1</sup><https://doi.org/10.5281/zenodo.17387934>

Table 1: Comparison of  $\text{NRMSE}_{\text{range}}$  and  $\text{sMAPE}$  (%) for graph parameter inference. Each model is evaluated in plain (zero-shot) and fine-tuned variants. Lower values indicate smaller relative errors and therefore better predictions. Fine-tuned results are underlined, and the best result for each parameter is shown in bold.

Parameter	$\text{NRMSE}_{\text{range}}$						$\text{sMAPE}$ (%)					
	Llama-3.2-3B		Qwen-2.5-3B		Phi-4-mini (4B)		Llama-3.2-3B		Qwen-2.5-3B		Phi-4-mini (4B)	
	Plain	Fine-tuned	Plain	Fine-tuned	Plain	Fine-tuned	Plain	Fine-tuned	Plain	Fine-tuned	Plain	Fine-tuned
Density	0.902	<u>0.000</u>	0.317	<u>0.005</u>	0.998	<u>0.014</u>	72.764	<u>0.000</u>	56.069	<u>0.201</u>	92.402	<u>0.654</u>
Degree Min	0.349	<u>0.100</u>	0.280	<u>0.090</u>	0.270	<u>0.088</u>	73.697	<u>16.717</u>	46.771	<u>17.376</u>	52.291	<u>17.393</u>
Degree Mean	0.397	<u>0.001</u>	0.311	<u>0.000</u>	0.160	<u>0.004</u>	65.803	<u>0.011</u>	45.800	<u>0.000</u>	23.579	<u>0.217</u>
Degree Max	0.284	<u>0.033</u>	0.250	<u>0.034</u>	0.331	<u>0.037</u>	44.846	<u>3.335</u>	34.764	<u>3.594</u>	60.272	<u>4.559</u>
Degree Std	0.268	<u>0.056</u>	0.293	<u>0.054</u>	0.257	<u>0.053</u>	48.499	<u>9.896</u>	44.476	<u>9.848</u>	45.682	<u>10.266</u>
Triangles Total	0.314	<u>0.030</u>	0.198	<u>0.023</u>	0.291	<u>0.037</u>	163.371	<u>14.027</u>	86.373	<u>9.808</u>	177.232	<u>16.806</u>
Average Clustering	0.594	<u>0.076</u>	0.370	<u>0.056</u>	0.510	<u>0.083</u>	168.042	<u>14.209</u>	105.227	<u>10.034</u>	178.104	<u>15.928</u>
Transitivity	0.619	<u>0.053</u>	0.368	<u>0.042</u>	0.518	<u>0.053</u>	168.327	<u>10.312</u>	102.598	<u>8.151</u>	178.718	<u>11.463</u>
Average Shortest Path Length	0.237	<u>0.048</u>	0.222	<u>0.051</u>	0.345	<u>0.058</u>	26.490	<u>2.720</u>	32.143	<u>2.794</u>	134.941	<u>3.286</u>
Diameter	0.215	<u>0.068</u>	0.236	<u>0.061</u>	0.218	<u>0.074</u>	38.764	<u>7.399</u>	39.536	<u>5.445</u>	46.950	<u>6.565</u>
Chromatic Number	0.224	<u>0.078</u>	0.195	<u>0.076</u>	0.348	<u>0.075</u>	26.137	<u>5.887</u>	21.712	<u>5.298</u>	41.875	<u>5.413</u>
Global Efficiency	0.894	<u>0.017</u>	0.353	<u>0.019</u>	0.994	<u>0.026</u>	140.083	<u>1.290</u>	32.294	<u>1.272</u>	176.948	<u>1.778</u>
Overall (Mean)	0.441	<u>0.047</u>	0.283	<u>0.043</u>	0.437	<u>0.050</u>	86.402	<u>7.150</u>	53.980	<u>6.152</u>	100.750	<u>7.861</u>

Table 2: Comparison of models on  $R^2$  Accuracy (%) and NRMSE Accuracy (%) across graph parameters. Higher values indicate better predictive performance and closer agreement between predicted and true graph-structural properties. Best results for each parameter are shown in bold.

Parameter	$R^2$ Accuracy (%)			NRMSE Accuracy (%)		
	Llama-3.2-3B	Qwen-2.5-3B	Phi-4-mini (4B)	Llama-3.2-3B	Qwen-2.5-3B	Phi-4-mini (4B)
Density	100.000	99.957	99.621	100.000	97.926	93.840
Degree Min	85.213	87.957	88.415	61.546	65.297	65.963
Degree Mean	99.999	100.000	99.974	99.697	100.000	98.394
Degree Max	97.924	97.885	97.415	85.591	85.456	83.921
Degree Std	95.286	95.599	95.775	78.289	79.021	79.445
Triangles Total	98.284	98.996	97.474	86.898	89.981	84.107
Average Clustering	92.539	95.845	90.945	72.686	79.617	69.908
Transitivity	96.465	97.753	96.384	81.198	85.009	80.984
Average Shortest Path Length	94.249	93.636	91.603	76.020	74.773	71.022
Diameter	88.521	90.875	86.226	66.120	69.792	62.886
Chromatic Number	86.986	87.487	87.988	63.926	64.626	65.341
Global Efficiency	99.485	99.317	98.766	92.822	91.736	88.893
Overall (Mean)	94.579	95.442	94.216	80.339	81.936	78.725

Accuracy expresses the proportion of variance in the target data explained by the model, calculated as  $100 \times \max(0, R^2)$ , while NRMSE Accuracy converts normalized RMSE with respect to the target standard deviation into a percentage scale using  $100 \times (1 - \text{NRMSE}_{\text{std}})$ . All accuracy metrics and  $\text{sMAPE}$  are reported in percentage form for comparability across parameters. Lower  $\text{NRMSE}_{\text{range}}$  and  $\text{sMAPE}$  values and higher accuracy scores indicate better predictive performance.

## 4 RESULTS

A comparison of the performance of base and fine-tuned models is presented in Table 1 (accuracy results for untrained models are near zero and therefore omitted). The results show a clear and systematic improvement across all evaluated metrics. Fine-tuning leads to a substantial reduction in normalized errors, with each model achiev-

ing markedly higher precision after adaptation. For  $\text{NRMSE}_{\text{range}}$ , the mean error decreases from 0.441 to 0.047 for Llama-3.2-3B, from 0.283 to 0.043 for Qwen-2.5-3B, and from 0.437 to 0.050 for Phi-4-mini (4B), corresponding to an average improvement of approximately 0.340. A similar trend is observed for  $\text{sMAPE}$ , where mean values drop from 86.4% to 7.2% for Llama, from 54.0% to 6.2% for Qwen, and from 100.8% to 7.9% for Phi - an overall reduction exceeding 70 percentage points. These results demonstrate that LoRA fine-tuning consistently and substantially enhances model precision, reducing both absolute and proportional errors by more than an order of magnitude.

Table 2 summarizes the accuracy of the fine-tuned models. The results remain consistently very high, with  $R^2$  Accuracy exceeding 94% and NRMSE Accuracy surpassing 78% across all models and parameters. Among them, Qwen-2.5-3B achieved the best overall performance, reaching a mean  $R^2$  Accuracy of 95.4% and an NRMSE Ac-

### Comparison of Large and Fine-tuned Small Language Models on Graph-Structural Inference

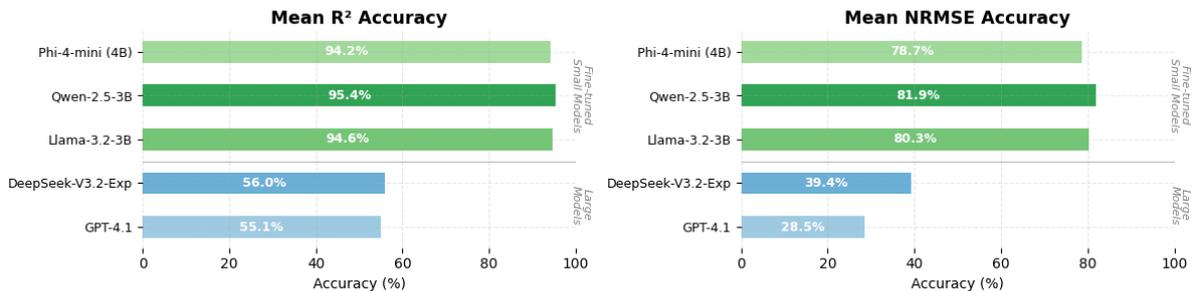


Figure 1: Comparison of Large and Fine-tuned Small Language Models on Graph-Structural Inference. The figure compares average performance across  $R^2$  Accuracy and NRMSE Accuracy. Compact, fine-tuned small models substantially outperform state-of-the-art large models on graph-structural reasoning tasks, demonstrating the effectiveness of efficient, domain-adapted fine-tuning.

curacy of 81.9%. Llama-3.2-3B followed closely with 94.6% and 80.3%, while Phi-4-mini (4B) achieved 94.2% and 78.7%, respectively. Although the differences are modest, Qwen shows a consistent advantage on metrics capturing higher-order structural organization, such as transitivity, clustering, and triangle counts - properties that depend on multi-node relationships and global connectivity. This suggests that Qwen’s architecture or pretraining corpus may better encode relational and hierarchical dependencies in serialized graph input.

In addition to the fine-tuned small models, we evaluated two state-of-the-art proprietary systems to contextualize the relative performance of our compact architectures. As illustrated in Figure 1, the small models substantially outperform these large-scale counterparts across both evaluation metrics. While GPT-4.1 and DeepSeek-V3.2-Exp achieve mean  $R^2$  Accuracies of 55.1% and 56.0%, and Mean NRMSE Accuracies of 28.5% and 39.4%, respectively, the fine-tuned small models reach over 94% and 78% on average for the same metrics. This consistent performance margin underscores the effectiveness of targeted, parameter-efficient adaptation over sheer model scale, demonstrating that compact models can achieve high reasoning precision when optimized for structural inference.

The fine-tuned small models maintain stable accuracy across diverse graph types - from sparse to dense topologies - demonstrating strong generalization and resilience to variations in degree distribution and connectivity. These outcomes highlight the effectiveness of LoRA-based adaptation in enhancing the structural inference capabilities of compact language models.

## 5 DISCUSSION

The results highlight how targeted, parameter-efficient adaptation enables small transformer models to acquire reliable structural inference skills from textual graph representations. Consistent improvements across all evaluated graph parameters confirm that small language models are capable of internalizing both local and global structural dependencies without requiring explicit graph neural architectures.

The ability of compact models to infer structural patterns can be attributed to several complementary mechanisms. First, the self-attention mechanism implicitly constructs relational dependencies between tokens, enabling recovery of adjacency-like patterns from serialized edge lists. Second, weight sharing and low-rank adaptation promote compression of relational rules, supporting the discovery of generalizable structural regularities rather than memorization. Third, pretrained linguistic priors - such as token co-occurrence, dependency tracking, and approximate counting - introduce transferable inductive biases that facilitate reasoning over graph-like inputs. Finally, the hierarchical organization of attention layers allows aggregation from local edge information to global structural descriptors, enabling accurate estimation of higher-order metrics such as clustering, transitivity, and global efficiency. Together, these mechanisms explain how small transformers, despite limited capacity and the absence of explicit graph inductive biases, can learn and generalize structural patterns directly from textual graph representations.

These findings have implications for both theory and methodology. They suggest that struc-

tural reasoning is not solely an emergent property of model scale but can also be encouraged through targeted adaptation, indicating that compact, efficiency-oriented architectures are capable of meaningful reasoning performance when appropriately fine-tuned. Building on this insight, the TinyGraphEstimator framework provides a controlled and reproducible environment for investigating such behaviors within the domain of graph reasoning, offering a principled way to examine how small language models interpret and generalize structural patterns. Although this study focuses on graph-structured data, the same methodological principles could extend to other structured domains in which relational dependencies are fundamental.

Beyond conceptual insights, these results suggest practical opportunities for deploying small language models in resource-constrained or embedded environments. Fine-tuned compact transformers can serve as efficient reasoning components within larger autonomous or decision-support systems, supporting applications such as knowledge graph analysis, network diagnostics, and symbolic planning. By enabling small models to infer and interpret graph-structural properties effectively, TinyGraphEstimator highlights a pathway toward lightweight reasoning agents that combine interpretability, adaptability, and scalability. Overall, the demonstrated predictive performance and generalization ability of fine-tuned small models help bridge the gap between discrete symbolic reasoning and natural language modeling, establishing compact transformers as viable tools for structured inference across both research and applied contexts.

## 6 LIMITATIONS

While this study demonstrates the potential of small language models for graph-structural reasoning, several limitations should be acknowledged. First, although LoRA fine-tuning consistently improves inference accuracy, the models still exhibit challenges in predicting complex or global parameters such as chromatic number and average shortest path length, which require deeper combinatorial reasoning. Second, the evaluation is based on synthetic graphs generated from a selected set of well-established random graph models, which provide controlled variability but may not reflect the full diversity of real-world network structures. Incorporating addi-

tional graph types or domain-specific datasets in future work could further strengthen the assessment of model generalization. Third, the experiments rely on serialized graph formats that, while simple and interpretable, may not capture complex or densely connected structures as effectively as more expressive representations. Future work could explore alternative input forms or hybrid architectures that integrate symbolic and neural reasoning. Despite these limitations, the results highlight promising directions for lightweight, adaptable language models in graph-theoretic inference.

## 7 CONCLUSION

This work explored the capability of small language models to infer a wide range of graph-theoretic parameters directly from structured graph representations. Using the proposed TinyGraphEstimator dataset, we systematically evaluated compact transformer-based models on tasks involving both local and global structural reasoning. The results demonstrate that even models with fewer than 4B parameters possess capacity to approximate structural graph measures when adapted through parameter-efficient fine-tuning using the LoRA method. Across all evaluated metrics, fine-tuned models consistently outperformed their zero-shot baselines, confirming that lightweight adaptation can effectively enhance structured reasoning without the computational overhead of large-scale models.

Overall, this study provides new insight into how small language models can internalize and generalize graph-structural information. By combining efficiency, adaptability, and interpretability, such models offer a promising foundation for scalable graph reasoning and for extending language-based inference to structured, symbolic, and relational domains.

## LLM USAGE STATEMENT

This manuscript acknowledges the use of ChatGPT (OpenAI, 2022), powered by the GPT-5 language model developed by OpenAI, to improve language clarity, refine sentence structure, and enhance overall writing precision.

## REFERENCES

- Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., and Phi (2024). Phi-4 technical report.
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., and OpenAI (2024). Gpt-4 technical report.
- Agrawal, P., Vasania, S., and Tan, C. (2024). Can llms perform structured graph reasoning tasks? In Pattern Recognition: 27th International Conference, ICPR 2024, Kolkata, India, December 1–5, 2024, Proceedings, Part VI, page 287–308, Berlin, Heidelberg. Springer-Verlag.
- Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.
- Bi, J., Wu, Y., Xing, W., and Wei, Z. (2025). Enhancing the reasoning capabilities of small language models via solution guidance fine-tuning. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, Proceedings of the 31st International Conference on Computational Linguistics, pages 9074–9084, Abu Dhabi, UAE. Association for Computational Linguistics.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4.
- Chen, X., Wang, Y., He, J., Du, Y., Hassoun, S., Xu, X., and Liu, L. (2025). Graph generative pre-trained transformer. In Forty-second International Conference on Machine Learning.
- DeepSeek (2025). Deepseek v3.2 experimental model documentation. [https://api-docs.deepseek.com/quick\\_start/pricing#model-details](https://api-docs.deepseek.com/quick_start/pricing#model-details). Accessed: 2025-10-14.
- Dwivedi, V. P., Joshi, C. K., Luu, A. T., Laurent, T., Bengio, Y., and Bresson, X. (2023). Benchmarking graph neural networks. *J. Mach. Learn. Res.*, 24(1).
- Erdős, P. and Rényi, A. (1959). On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290.
- Gat, N. (2024). lm-format-enforcer: Output format enforcement for language models. GitHub repository. <https://github.com/noamgat/lm-format-enforcer>.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., and Llama (2024). The llama 3 herd of models.
- Guo, J., Du, L., Liu, H., Zhou, M., He, X., and Han, S. (2023). Gpt4graph: Can large language models understand graph structured data ? an empirical evaluation and benchmarking.
- Guo, Z., Xia, L., Yu, Y., Wang, Y., Lu, K., Huang, Z., and Huang, C. (2025). Graphedit: Large language models for graph structure learning.
- Hamilton, W. L., Ying, R., and Leskovec, J. (2017). Inductive representation learning on large graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, page 1025–1035, Red Hook, NY, USA. Curran Associates Inc.
- Han, Z., Gao, C., Liu, J., Zhang, J., and Zhang, S. Q. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey.
- Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In International Conference on Machine Learning.
- Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations.
- Hu, Y., Huang, X., Wei, Z., Liu, Y., and Hong, C. (2025). Rethinking and benchmarking large language models for graph reasoning.
- Huang, J., Zhang, X., Mei, Q., and Ma, J. (2024). Can LLMs effectively leverage graph structural information through prompts, and why? Transactions on Machine Learning Research.
- Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations.
- Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4582–4597.
- Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., and DeepSeek-AI (2025). Deepseek-v3 technical report.
- LLM Stats (2024). Llm stats – leaderboard and metrics for large language models. Project homepage: <https://llm-stats.com>, Accessed: 2025-05-14.
- Lu, W., Luu, R. K., and Buehler, M. J. (2024). Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities.

- OpenAI (2022). Chatgpt. <https://chat.openai.com/>. Large language model-based conversational system.
- OpenAI (2024). Gpt-4.1 technical overview. <https://platform.openai.com/docs/models/gpt-4.1>. Accessed: 2025-10-14.
- Podstawski, M. (2025). Applying text embedding models for efficient analysis in labeled property graphs.
- Schmitt, M., Ribeiro, L. F. R., Dufter, P., Gurevych, L., and Schütze, H. (2021). Modeling graph structure via relative position for text generation from knowledge graphs. In Panchenko, A., Malliaros, F. D., Logacheva, V., Jana, A., Ustalov, D., and Jansen, P., editors, Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15), pages 10–21, Mexico City, Mexico. Association for Computational Linguistics.
- Tang, J., Yang, Y., Wei, W., Shi, L., Su, L., Cheng, S., Yin, D., and Huang, C. (2024). Graphgpt: Graph instruction tuning for large language models. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, page 491–500, New York, NY, USA. Association for Computing Machinery.
- von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., and Gallouédec, Q. (2020). Trl: Transformer reinforcement learning. <https://github.com/huggingface/trl>.
- Wang, S., Lin, J., Guo, X., Shun, J., Li, J., and Zhu, Y. (2025). Reasoning of large language models over knowledge graphs with super-relations. In The Thirteenth International Conference on Learning Representations.
- Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442.
- Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2019). A comprehensive survey on graph neural networks. *CoRR*, abs/1901.00596.
- Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., and Qwen (2025). Qwen2.5 technical report.
- Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. (2021). Do transformers really perform bad for graph representation? In Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.
- You, J., Ying, R., Ren, X., Hamilton, W. L., and Leskovec, J. (2018). Graphrnn: Generating realistic graphs with deep auto-regressive models. In Dy, J. G. and Krause, A., editors, ICML, volume 80 of Proceedings of Machine Learning Research, pages 5694–5703. PMLR.