

THE PATH MATTERS: EVALUATING SMALL LANGUAGE MODELS BEYOND ANSWER ACCURACY IN KGQA

Eduin E. Hernandez*, Sergio A. Diaz, Luis F. Garcia, Nurassyl Askar, and Stefano Rini

NYCU, Taiwan

ABSTRACT

Small language models (SLMs) are increasingly paired with knowledge graphs (KGs), yet end-to-end KG question answering conflates graph access, search, navigation, reasoning, and answer generation. This coupling makes it difficult both to determine whether an SLM can faithfully execute the reasoning path implied by a question and to attribute failures to navigation rather than to other stages of the pipeline. We isolate this capability by employing the THESEUS navigation and traceability framework and using frozen, off-the-shelf SLMs as local action policies. At each hop, the environment exposes the legal outgoing graph actions, and the model selects one executable graph action and decides whether to stop, without task-specific parameter updates, model-controlled beam search, or free-form answer generation. This controlled setting allows us to evaluate terminal-answer accuracy with Hits@1 together with path fidelity, using Path Edit Distance (PED) as the primary trajectory metric. Across the KINSHIP and MQUAKE-ST KGQAs, similarly sized local models differ substantially in answer accuracy and path fidelity, with the two metrics sometimes favoring different models. This model-dependent behavior also extends to prompting, as a single demonstrated trajectory can improve or degrade navigation depending on the model. These results motivate evaluating SLM graph reasoning beyond endpoint accuracy alone.

Index Terms— Knowledge graph question answering (KGQA); small language models; graph navigation; multi-hop reasoning; path fidelity.

1. INTRODUCTION

Knowledge graphs (KGs) provide explicit relational structure for multi-hop question answering, and recent language-model systems exploit this structure through graph retrieval, path search, planning, and grounded answer generation [1, 2, 3]. These systems demonstrate the value of graph scaffolding, but

they make it difficult to determine whether a language model can choose the correct next graph action when search and answer synthesis are factored out. This distinction is especially important for small language models (SLMs), where graph exploration has been identified as a major bottleneck inside a Think-on-Graph-style pipeline [4]. Meanwhile, recent work on compact KG agents has focused on improving performance through task-specific training rather than evaluating navigation ability of frozen models [5, 6, 7].

We study this question within the THESEUS formulation of multi-hop KGQA as question-conditioned graph navigation [8], where the executed KG path is evaluated together with the terminal answer. Our focus is a deliberately restricted setting: *Can an off-the-shelf SLM faithfully execute a multi-hop reasoning path when the environment exposes the legal local graph actions and the model is responsible only for selecting the next executable action and deciding whether to stop?* Isolating this capability reveals whether navigation failures arise from the model’s local question-conditioned decisions rather than from retrieval, search, or answer synthesis, and provides a controlled baseline for measuring the value of additional reasoning mechanisms. The controller fixes graph access, executes only legal edges, and uses the terminal entity as the answer, removing model-controlled candidate retrieval, pruning, beam/tree search, and free-form answer generation from the decision loop. Related work has studied search-assisted KG reasoning and broader sequential graph navigation [9, 10, 11], whereas our setting isolates local executable graph-action selection and termination within path-annotated multi-hop KGQA.

Answer accuracy alone does not establish path fidelity: a model may reach a valid answer through a trajectory that differs substantially from the reference reasoning path [2, 12, 3]. We therefore evaluate both ordered trajectory agreement, using Path Edit Distance (PED) over graph edges and order-invariant structural overlap, using $F1_{SG}$ over traversed edges [8]. PED serves as our primary trajectory metric. We compare locally deployable SLMs under matched zero-shot and one-shot conditions on KINSHIP and MQUAKE-ST, finding that endpoint correctness and path fidelity can favor different models and that the effect of a single demonstrated trajectory is strongly model-dependent.

Contribution. We provide a controlled empirical charac-

E.E.H., S.A.D., and S.R. are with National Yang Ming Chiao Tung University (NYCU), Taiwan. L.F.G. and N.A. are independent researchers and performed part of this research while previously with NYCU. *Corresponding author: eduin.ee08@nycu.edu.tw.

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

terization of the local multi-hop graph-navigation ability of frozen, locally deployable SLMs under the THESEUS navigation and traceability framework, separating terminal-answer correctness from executed-path fidelity and examining the effect of a single inference-time demonstration. As a secondary analysis, we use this controlled baseline in a compact ToG-style comparison on MQUAKE-ST to examine how explicit search and answer-generation scaffolding affect terminal accuracy and trajectory fidelity.

2. TASK FORMULATION AND EVALUATION

We build on the THESEUS formulation of multi-hop KGQA as question-conditioned graph navigation [8]. Let a knowledge graph be $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is the entity set, $\mathcal{R}$ is the relation set, and $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is the set of directed triplets. Given a natural-language question q , a known topic entity e_s , and a valid answer set $\mathcal{A}(q) \subseteq \mathcal{E}$, the task is to execute a path

$$P = (e_0 = e_s, r_1, e_1, \dots, r_k, e_k) \quad (1)$$

such that every $(e_{i-1}, r_i, e_i) \in \mathcal{T}$ and the terminal entity $e_k \in \mathcal{A}(q)$. As in THESEUS, the underlying reasoning length is not supplied to the agent. Instead, questions are evaluated under a fixed maximum reasoning horizon N . At step t , the legal outgoing graph-action set is

$$\mathcal{N}(e_t) = \{(r, e') : (e_t, r, e') \in \mathcal{T}\}. \quad (2)$$

For our SLM instantiation, the presented graph actions are assigned local identifiers, and the model outputs a pair (i_t, z_t) comprising a graph-action ID i_t (or null) and a stop flag z_t . If i_t identifies a legal graph action, the controller deterministically executes the corresponding KG transition and appends the traversed triplet to the predicted path $\hat{P}(q)$. If z_t indicates stopping, the episode terminates after the selected transition. When i_t is null, the episode instead terminates at the current entity, and the null choice is valid only when stopping. Otherwise, navigation continues until the maximum reasoning horizon N is reached. The terminal entity is returned as the predicted answer.

Our SLM policy produces a single trajectory rather than a ranked rollout set. We therefore measure terminal-answer correctness with Hits@1, where the terminal entity must belong to $\mathcal{A}(q)$. Because no ranked candidate set is produced, MRR is not defined in our evaluation. For multi answer, reaching any valid answer counts as success. For path fidelity, we follow the THESEUS evaluation protocol [8]. PED is the Levenshtein distance between the ordered predicted and reference edge sequences, while $F1_{SG}$ measures their order-invariant edge overlap. Together, these metrics capture complementary aspects of trajectory fidelity: ordered path agreement and order-invariant edge overlap. PED is our primary trajectory metric, with lower values indicating closer agreement and PED = 0 corresponding to an exact reference-path match. Together

Table 1. Datasets and evaluation settings. $\bar{k}$ denotes the mean reference path length over the evaluation questions.

Dataset	$ \mathcal{E} $	$ \mathcal{R} $	$ \mathcal{T} $	Eval. Q	$\bar{k}$	N
KINSHIP	24	12	112	101	2.624	3
MQUAKE-ST SA	38,516	665	724,141	1,504	2.513	4
MQUAKE-ST MA	38,516	665	724,141	870	2.685	4

with Hits@1, these metrics provide a controlled measure of intrinsic local navigation against which additional search or task-specific adaptation can be evaluated.

3. EXPERIMENTAL SETTING

For the local-navigation experiments, all SLMs are evaluated under the same navigation interface defined in Sec. 2. At each step, the model observes the question q , topic entity e_s , current entity e_t , executed path history, and the presented legal outgoing graph actions. We use tuple graph-action selection over (e_t, r, e') , with all models receiving the same indexed triplet representation of presented actions and the same structured response schema—{action: ID/null, stop: flag}. Models are evaluated off the shelf with frozen parameters. Optional model-specific thinking or reasoning modes are disabled to avoid introducing additional inference-time token and compute budgets as a confound. Each navigation decision uses a single model generation, with no parse-retry or repair calls.

Our primary one-shot condition provides one complete solved trajectory from the training split as an in-context demonstration, showing the navigation state, available actions, and gold decision at each hop. The zero-shot condition removes this demonstration while leaving the remaining navigation interface unchanged. We use deterministic decoding with temperature zero, seed 42, and a maximum response length of 64 tokens. Code and evaluation scripts are publicly available.¹

We evaluate on the navigation-ready KINSHIP and MQUAKE-ST resources from THESEUS [8], with MQUAKE-ST derived from the original MQUAKE-CF resource [13]. Each benchmark provides a fixed KG, explicit topic and valid answer entities, and annotated reasoning paths required for path-level evaluation. Table 1 summarizes the graph characteristics and evaluation subsets used in the present experiments. We use action-display caps of 100 for KINSHIP and 200 for MQUAKE-ST. For MQUAKE-ST, we evaluate two settings: Single Answer (SA) and Multi Answer (MA).

The evaluated model set comprises Gemma 4 E4B [14], Granite 3.3 8B [15], Llama 3.1 8B [16], Ministral 3 8B [17], OLMo-3 7B [18], Phi-4 Mini 3.8B [19], Qwen2.5 7B [20], and Qwen3 8B [21]. Llama 3.1, Ministral 3, Qwen2.5, and OLMo-3 use instruction-tuned local artifacts. Llama 3.1, Ministral 3,

¹https://github.com/HalcyonSolutions/LLM_KGQA

and Qwen2.5 use Q4 artifacts. Structured-output compatibility is an inclusion requirement of the controlled interface rather than an ablation dimension.²

ToG-style search. To quantify the effect of explicit search, we additionally evaluate the three strongest models on MQUAKE-ST SA using a ToG-style scaffold [1]. We preserve the same directed KG and maximum horizon $N = 4$, restricting traversal to legal outgoing edges. The scaffold adds LLM-guided relation and entity pruning, global retention of w candidate paths, accumulated-evidence sufficiency checking, and free-form answer generation. These runs use the original ToG prompt templates rather than the structured zero- and one-shot navigation prompts used above. We evaluate $w \in \{1, 3\}$: $w = 1$ provides a single-retained-path condition closest to our local-navigation setting, while $w = 3$ follows the default ToG search width [1]. We report terminal-answer Hits@1, generated-answer accuracy, PED of the highest-ranked retained path, and mean LLM calls per question (Calls/Q).

4. RESULTS

Structural Calibration. Table 2 reports one-shot navigation performance together with the structural calibration references. The unbiased random walk represents unguided traversal, while the shortest-path oracle is given the valid answer set but not the annotated reasoning path and selects a shortest route to an answer. The oracle is therefore not a bound on path fidelity, since the shortest route to a valid answer need not coincide with the reasoning path implied by the question. Models with lower edit distances and higher overlap scores than the oracle follow the annotated reasoning structure more closely than can be achieved from answer knowledge and shortest-path efficiency alone. Performance approaching the random-walk reference instead indicates weaker question-conditioned navigation.

One-Shot Navigation Performance. Across the evaluated SLMs, terminal-answer accuracy and path fidelity vary substantially despite broadly similar deployment scales. On KINSHIP, Ministral 3 achieves the highest Hits@1, whereas Gemma 4 achieves the strongest path fidelity across the reported metrics. This shows that endpoint accuracy and path fidelity can favor different models. On both MQUAKE-ST settings, however, Ministral 3 achieves the strongest performance on both answer accuracy and path fidelity. Qwen3 and Gemma 4 form the next strongest group. Several of the stronger SLMs also outperform the shortest-path oracle on the path-fidelity metrics, indicating closer agreement with the annotated reference reasoning structure than an answer-informed shortest route through the graph. By contrast, weaker models move toward the structural calibration references, indicating less effective question-conditioned navigation. Calls/Q should likewise be interpreted relative to the reference path length

²DeepSeek-R1 was excluded because it did not reliably satisfy the required structured response interface in our runs.

Table 2. One-shot navigation results on KINSHIP and MQUAKE-ST with structural calibration references reproduced from THESEUS [8]. PED is the primary path-fidelity metric, and Calls/Q is the mean number of LLM invocations per question. $\dagger$ denotes terminal answer Hits@1 undefined for the calibration references, and $-$ denotes not applicable.

Model / Reference	Hits@1 $\uparrow$	PED $\downarrow$	$F1_{SG}$ $\uparrow$	Calls/Q
(a) KINSHIP				
<i>RW-AnsMC</i>	$\dagger$	2.546	0.128	$-$
<i>Shortest Path Oracle</i>	$\dagger$	2.099	0.220	$-$
Gemma 4 E4B	0.911	0.158	0.958	2.663
Granite 3.3 8B	0.554	0.842	0.768	2.861
Llama 3.1 8B	0.436	0.832	0.783	3.000
Ministral 3 8B	0.941	0.208	0.936	2.683
OLMo-3 7B	0.337	1.307	0.638	2.931
Phi-4 Mini 3.8B	0.347	1.446	0.530	2.495
Qwen2.5 7B	0.683	0.614	0.825	2.743
Qwen3 8B	0.822	0.446	0.860	2.594
(b) MQUAKE-ST Single Answer				
<i>RW-AnsMC</i>	$\dagger$	3.381	0.034	$-$
<i>Shortest Path Oracle</i>	$\dagger$	1.670	0.380	$-$
Gemma 4 E4B	0.728	0.680	0.790	2.604
Granite 3.3 8B	0.373	2.001	0.489	3.337
Llama 3.1 8B	0.351	2.227	0.561	3.965
Ministral 3 8B	0.894	0.362	0.883	2.545
OLMo-3 7B	0.173	2.322	0.410	3.349
Phi-4 Mini 3.8B	0.297	1.521	0.513	1.751
Qwen2.5 7B	0.666	0.955	0.716	2.849
Qwen3 8B	0.824	0.520	0.843	2.438
(c) MQUAKE-ST Multi Answer				
<i>RW-AnsMC</i>	$\dagger$	3.406	0.044	$-$
<i>Shortest Path Oracle</i>	$\dagger$	1.843	0.362	$-$
Gemma 4 E4B	0.775	0.723	0.792	2.694
Granite 3.3 8B	0.247	2.231	0.450	3.509
Llama 3.1 8B	0.238	2.266	0.549	3.979
Ministral 3 8B	0.849	0.601	0.812	2.657
OLMo-3 7B	0.118	2.540	0.389	3.599
Phi-4 Mini 3.8B	0.391	1.528	0.545	2.128
Qwen2.5 7B	0.523	1.326	0.615	2.928
Qwen3 8B	0.731	0.751	0.786	2.628

$\bar{k}$ rather than minimized independently, since low values can reflect premature stopping.

Effect of One-Shot Demonstration. Table 3 compares zero-shot and one-shot navigation under the same evaluation interface. The effect of a single demonstrated trajectory is strongly model dependent rather than uniformly beneficial. Gemma 4, Ministral 3, and Qwen3 generally improve both terminal-answer accuracy and path fidelity under one-shot

Table 3. Zero-shot and one-shot navigation results on KINSHIP and MQUAKE-ST. Δ Calls/Q is the one-shot minus zero-shot mean number of LLM invocations per question. Negative values indicate fewer calls under one-shot.

Model	Hits@1 $\uparrow$		PED $\downarrow$		Δ Calls/Q
	0-shot	1-shot	0-shot	1-shot	
(a) KINSHIP					
Gemma 4 E4B	0.653	0.911	0.634	0.158	-0.337
Granite 3.3 8B	0.634	0.554	0.861	0.842	+0.119
Llama 3.1 8B	0.455	0.436	0.713	0.832	+0.000
Ministral 3 8B	0.851	0.941	0.238	0.208	-0.059
OLMo-3 7B	0.436	0.337	0.950	1.307	-0.059
Phi-4 Mini 3.8B	0.396	0.347	1.416	1.446	+0.129
Qwen2.5 7B	0.485	0.683	0.911	0.614	+0.287
Qwen3 8B	0.703	0.822	0.653	0.446	+0.119
(b) MQUAKE-ST Single Answer					
Gemma 4 E4B	0.588	0.728	1.111	0.680	-0.578
Granite 3.3 8B	0.417	0.373	1.623	2.001	+0.725
Llama 3.1 8B	0.257	0.351	2.170	2.227	+0.002
Ministral 3 8B	0.761	0.894	0.827	0.362	+0.037
OLMo-3 7B	0.239	0.173	2.211	2.322	+0.023
Phi-4 Mini 3.8B	0.124	0.297	1.658	1.521	+0.370
Qwen2.5 7B	0.590	0.666	1.205	0.955	+0.265
Qwen3 8B	0.681	0.824	0.769	0.520	+0.162
(c) MQUAKE-ST Multi Answer					
Gemma 4 E4B	0.649	0.775	1.121	0.723	-0.523
Granite 3.3 8B	0.313	0.247	1.772	2.231	+0.793
Llama 3.1 8B	0.236	0.238	2.307	2.266	+0.013
Ministral 3 8B	0.708	0.849	0.907	0.601	-0.037
OLMo-3 7B	0.160	0.118	2.451	2.540	+0.090
Phi-4 Mini 3.8B	0.141	0.391	1.730	1.528	+0.638
Qwen2.5 7B	0.462	0.523	1.534	1.326	+0.216
Qwen3 8B	0.630	0.731	0.924	0.751	+0.199

prompting, whereas Granite 3.3 degrades under the same intervention and OLMo-3 also weakens in several settings. Other models exhibit mixed behavior. In particular, Llama 3.1 can improve terminal-answer accuracy without a corresponding improvement in PED, showing that prompting can affect endpoint success and path agreement differently. Changes in Calls/Q are similarly heterogeneous, with some models producing stronger trajectories using fewer calls and others requiring additional calls after the demonstration. Overall, a single trajectory demonstration can affect terminal-answer accuracy, path fidelity, and stopping behavior differently across models.

Effect of Explicit Search. Table 4 evaluates whether adding ToG-style search changes the behavior observed under local navigation. Relative to Table 2, the $w = 1$ setting yields lower Hits@1 and higher PED for all three models, despite requiring more LLM calls. Notably, Ministral 3 and Gemma 4,

Table 4. ToG-style search on MQUAKE-ST Single Answer using the three strongest local-navigation SLMs. Hits@1 evaluates the terminal entity and PED the trajectory of the highest-ranked retained path. Gen. Acc. evaluates the generated answer.

Model	w	Hits@1 $\uparrow$	Gen. Acc. $\uparrow$	PED $\downarrow$	Calls/Q
Gemma 4 E4B	1	0.289	0.412	1.441	5.381
	3	0.152	0.307	1.931	5.963
Ministral 3 8B	1	0.303	0.483	1.496	4.253
	3	0.250	0.490	1.858	10.100
Qwen3 8B	1	0.547	0.700	1.289	6.295
	3	0.393	0.693	1.684	12.987

which are among the strongest models in the isolated navigation setting, degrade substantially under the search scaffold, while Qwen3 remains comparatively stronger. Increasing the width to $w = 3$ further decreases Hits@1 and worsens PED across all models. Generated-answer accuracy nevertheless exceeds terminal Hits@1 in every setting, showing that answer synthesis can recover correct answers from imperfect search trajectories. Overall, under this matched directed setting, additional ToG-style search does not improve trajectory fidelity, and increasing search width provides no consistent benefit in generated-answer accuracy. Because the scaffold jointly introduces prompting, relation and entity selection, path pruning, sufficiency checking, and answer generation, these results further motivate evaluating such components individually rather than attributing system-level behavior to search alone.

5. CONCLUSION

We evaluated frozen, off-the-shelf Small Language Models (SLM) as local graph-navigation policies under a controlled setting that fixes graph access, restricts execution to presented legal graph actions, and uses the terminal entity as the predicted answer. Across KINSHIP and MQUAKE-ST, similarly sized models exhibit substantial differences in both terminal-answer accuracy and path fidelity. Endpoint accuracy and path fidelity can rank models differently, as observed on KINSHIP, while the strongest model on MQUAKE-ST performs well on both criteria. A single trajectory demonstration can also improve or degrade these measures depending on the model. A complementary ToG-style evaluation shows that adding a multi-stage search-and-generation scaffold does not necessarily preserve the strong local-navigation behavior observed in isolation, while generated answers can partially mask trajectory failures. These findings motivate component-wise evaluation of relation and entity selection, search control, stopping, and answer synthesis, as well as future study of task-specific adaptation.

Acknowledgment

This work is partially funded by the NSTC grant number 113-2923-E-A49-001 and by MARC, the MediaTek Advanced Research Center with grant number 114A540531. OpenAI ChatGPT was used during manuscript preparation for language and clarity editing, wording refinement, and limited assistance with research ideation and code development. All scientific decisions, experimental design, result interpretation, and final manuscript content were reviewed and verified by the authors, who take full responsibility for the work.

Compliance with Ethical Standards

This study did not involve human or animal subjects, and no ethical approval was required.

6. REFERENCES

- [1] Jiashuo Sun et al., “Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
- [2] Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan, “Reasoning on graphs: Faithful and interpretable large language model reasoning,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
- [3] Yuan Sui, Yufei He, Nian Liu, Xiaoxin He, Kun Wang, and Bryan Hooi, “FiDeLiS: Faithful reasoning in large language models for knowledge graph question answering,” in *Findings Assoc. Comput. Linguistics: ACL*, 2025, pp. 8315–8330, Association for Computational Linguistics.
- [4] Yi-Jie Cheng, Oscar Chew, and Yun-Nung Chen, “The role of exploration modules in small language models for knowledge graph question answering,” in *Proc. ACL Student Res. Workshop (ACL SRW)*, 2025, pp. 919–928, Association for Computational Linguistics.
- [5] Jinhao Jiang et al., “KG-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2025, pp. 9505–9523, Association for Computational Linguistics.
- [6] Shuwen Xu et al., “Graphwalker: Agentic knowledge graph question answering via synthetic trajectory curriculum,” in *Proc. Conf. Lang. Model. (COLM)*, 2026.
- [7] Jia Ao Sun et al., “Search-on-graph-r1: Training large language models to search knowledge graphs with reinforcement learning,” *arXiv:2607.18481*, 2026.
- [8] Eduin E. Hernandez, Luis F. Garcia, Nurassyl Askar, Sergio A. Diaz, and Stefano Rini, “Theseus in the graph: Towards traceable multi-hop graph navigation,” *arXiv:2609.14528*, 2026.
- [9] Jia Ao Sun et al., “Search-on-graph: Iterative informed navigation for large language model reasoning on knowledge graphs,” in *Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining (KDD)*, 2026, pp. 4475–4485.
- [10] Taraneh Ghandi, Hamidreza Mahyar, and Shachar Klaiman, “Graphwalk: Enabling reasoning in large language models through tool-based graph navigation,” *arXiv:2604.01610*, 2026.
- [11] Daniele Margiotta, Danilo Croce, and Roberto Basili, “Evaluating large language models on wikipedia graph navigation: Insights from the wikigame,” in *Proc. Ital. Conf. Comput. Linguistics (CLiC-it)*, 2025, pp. 659–669, CEUR Workshop Proceedings.
- [12] Kun Li, Tianhua Zhang, Xixin Wu, Hongyin Luo, James R. Glass, and Helen M. Meng, “Decoding on graphs: Faithful and sound reasoning on knowledge graphs through generation of well-formed chains,” in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2025, pp. 24349–24364, Association for Computational Linguistics.
- [13] Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen, “MQuAKE: Assessing knowledge editing in language models via multi-hop questions,” in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Singapore, 2023, pp. 15686–15702, Association for Computational Linguistics.
- [14] Gemma Team, “Gemma 4 technical report,” *arXiv:2607.02770*, 2026.
- [15] IBM Granite Team, “Granite 3.3 8b instruct,” Hugging Face model card, 2025.
- [16] Aaron Grattafiori et al., “The llama 3 herd of models,” *arXiv:2407.21783*, 2024.
- [17] Alexander H. Liu et al., “Ministral 3,” *arXiv:2601.08584*, 2026.
- [18] Team Olmo et al., “Olmo 3,” *arXiv:2512.13961*, 2025.
- [19] Abdelrahman Abouelenin et al., “Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras,” *arXiv:2503.01743*, 2025.
- [20] An Yang et al., “Qwen2.5 technical report,” *arXiv:2412.15115*, 2024.
- [21] An Yang et al., “Qwen3 technical report,” *arXiv:2505.09388*, 2025.