VA-BO-INTERN: Adapting Small Language Models to Low-Resource Domains: A Case Study in Hindi Tourism QA

Sandipan Majhi[1,*,†], Paheli Bhattacharya[2]

[1]Indian Institute of Technology Kharagpur, India
[2]Bosch Research and Technology Centre, Bangalore, India

Abstract

Domain-specific question answering in low-resource languages faces two key challenges: scarcity of annotated datasets and limited domain knowledge in general-purpose language models. In this work, we present a multi-stage finetuning strategy to adapt lightweight language models to the Hindi tourism domain by leveraging both original and synthetic training data. Synthetic question–answer pairs are generated using large LLMs (LLaMA-70B, Phi-14B) and used to augment the limited original dataset. We explore several training methodologies and analyze their impact on domain generalization. Our results demonstrate that large models can efficiently generate synthetic data, while small models can effectively adapt to it, offering a scalable pathway for low-resource, domain-specific QA.

Keywords

Question Answering, Synthetic Data Generation, Small Language Model Finetuning, Indic NLP

1. Introduction

Large language models (LLMs) have significantly advanced natural language generation, understanding, and reasoning. Despite their success, adapting these models to domain-specific applications remains challenging due to two main factors: (i) general-purpose LLMs often lack specialized domain knowledge, and (ii) high-quality annotated datasets are scarce and expensive to obtain. The cost and time demands of manual annotation have therefore driven interest in synthetic data as a scalable alternative.

LLMs, owing to their broad pre-training, can act as effective knowledge bases [1, 2] and have been shown to produce high-quality synthetic question–answer (QA) pairs [3, 4, 5]. Synthetic datasets have further demonstrated utility in addressing the limitations of low-resource domains [6, 7], enabling the creation of specialized training resources that would otherwise be infeasible. This has opened a practical avenue for domain adaptation, especially in fields where curated open-source datasets are extremely limited.

At the same time, the emergence of lightweight language models provides new opportunities for efficient domain adaptation. Smaller models are cheaper to finetune, faster at inference, and easier to deploy in resource-constrained environments. While very large LMs are well-suited for generating synthetic training data, compact models are more practical for downstream deployment. Thus, combining synthetic data generation from large models with targeted finetuning of smaller ones represents a promising strategy for building effective, domain-specific QA systems.

In this work, we investigate this paradigm in the context of Hindi tourism, a domain where both language resources and annotated datasets are limited. We generate synthetic QA pairs using large LMs (LLaMA-70B and Phi-14B) and finetune a smaller model (LLaMA-8B) to evaluate its performance. Beyond simple finetuning, we explore mixed-training methodologies that combine synthetic and general-domain data, analyzing their effect on robustness and domain generalization. Our contributions are

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India
*Corresponding author.
†Work done during the internship at Bosch Research and Technology Centre, Bangalore, India
sandipan.majhi.24@kgpian.iitkgp.ac.in (S. Majhi)

threefold: (i) a data augmentation strategy tailored to low-resource, domain-specific QA, (ii) empirical evidence that synthetic data can effectively adapt lightweight models, and (iii) a comparative analysis of training strategies to identify setups that balance efficiency with domain performance.

2. Related Work

Generating **high-quality synthetic question answers** using large language models (LLMs) has been a key focus of recent research [3, 4, 5]. Studies done by Chia et al. [8] and Liu et al. [9] have shown that zero-shot prompting can be highly effective for creating high-quality, structured synthetic data. However, generating synthetic question-answer pairs can sometimes result in unintended redundancy. Studies such as Yadav et al. [10] suggest that exploring different sampling techniques could introduce greater diversity, which may be beneficial for downstream tasks. The primary goal of generating synthetic question-answer pairs is to improve model performance on question-answering tasks. Prior work by Chowdhury and Chadha [11] demonstrated how synthetic data, particularly from "in-the-wild" sources, can lead to performance gains and help achieve natural distribution shifts. Similarly, Kramchaninova and Defauw [12] validated the effectiveness of combining synthetic data with original training data, showing that this approach consistently outperforms models trained exclusively on non-synthetic data, especially on domain-specific test sets. Another study by Harsha et al. [13] on the use of synthetic data within the financial domain confirmed its effectiveness in boosting question-answering performance in specialized fields.

To achieve performance improvements on downstream question-answering tasks, several studies have investigated **different methods for finetuning models** using a combination of synthetic and original training datasets. Namboori et al. [14] proposed a finetuning approach that involves first training on the synthetic data and then on the original training set, arguing that a model should perform better if it is well-conditioned to a high-quality dataset. Conversely, Chada and Natarajan [15] showed performance improvements by finetuning first on the original training data and then on a small, additional amount of synthetic data. Other studies, including [16, 17, 5], also demonstrate that continued finetuning on a small amount of synthetic data can lead to a significant performance uplift in question answering. A study by Gurgurov et al. [18] utilized synthetic data curation by translating English data into other low-resource languages and performing continued pretraining, illustrating how synthetic data can aid in model alignment for new domains.

Researchers have already created several benchmarks based on **synthetic datasets, particularly for low-resource domains**. The Indic-QA Benchmark [19] used synthetic data generation techniques to create question-answering datasets in 11 Indian languages. Similarly, IndiSentiment140 [20] is another such dataset that uses machine translation to generate sentiment analysis datasets across 22 Indian languages. The IndicXTREME Benchmark [21] also leveraged machine translation to create new synthetic datasets from existing English data.

3. Methodology

This study investigates the impact of synthetic data augmentation on the performance of small language models for long-form question-answering task.

Synthetic Data Generation: As shown in Table 1, we utilize contexts from the training set to generate additional question-answer pairs. To achieve this, we employ larger LLMs to create a new corpus of question answer pairs using training set contexts using few shot prompts.

Finetuning SLMs: We train SLMs using the synthetic data and available training data. A comprehensive representation of the workflow has been presented in Figure 1. We follow a mixed-training strategy:

• Baseline Model: We first finetune a small language model on the already available data, to get the baseline finetuned model.

• Continued Finetuning: We use the synthetic data and continue finetuning the baseline finetuned model, to produce a continued finetuned model.

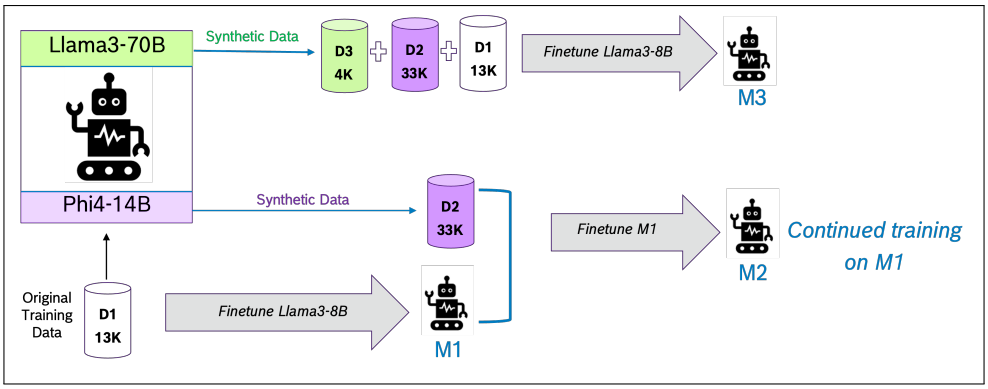

Figure 1: An overview of the two-phased experimental procedure, including synthetic data generation, followed by mixed fine-tuning of a smaller language model on the augmented Hindi-language dataset.

Split	Contexts	QA Pairs	QA/Context	Ques. length	Ans. length
Train	5244	13092	2.50	12.64	16.10
Validation	1134	2798	2.47	12.57	16.07
Test Data-1	1143	2902	2.53	12.68	16.16
Test Data-2	430	1196	2.78	16.40	–
Synthetic Data	5244	37259	5.11	12.16	18.63

Table 1
Key statistics of the VATIKA dataset, including question-answer pairs, average sentence lengths, and the number of contexts in each split. We could not provide the answer length of Test Data-2 as it was held-out.

• Multi-Source Finetuning: In this setting, we use all of the available training data and the augmented synthetic data to finetune the SLM.

4. Dataset

This study utilizes the Varanasi Tourism in Question Answer System (VATIKA) dataset published by Gatla et al. [22], a publicly available resource in Forum for Information Retrieval Evaluation (FIRE) 2025. This Hindi-language dataset consists of instances where each context is paired with one or more related question-answer pairs. The answers were typically long-form and abstractive in nature, utilizing the provided context as their source of information. The originally published dataset had three splits, namely, train, validation and Test Data-1. There is a held-out test set which was provided as a part of the shared task. It only had contexts and questions and not the gold standard answers. We refer this dataset as Test Data-2. To provide a comprehensive overview of the dataset we present its key statistics in Table 1. We see that on average the question and answer lengths in the VATIKA dataset is 13 and 16 words respectively and there are about 2-3 question-answer pairs per contexts. The synthetic data on average produces approximately 3 more QA pairs on average and has similar question and answer length distribution.

5. Experimental Settings

Synthetic Data Generation: For generating synthetic data, LLAMA-3.1-70B[23] and Phi-4-14B[24] were utilized in few shot prompt format. The new corpus of question-answer pairs contained around 4,000 instances generated by LLAMA-3.1-70B and 33,000 instances generated by Phi-4-14B. We use $temperature = 0.7$ and $top - p = 0.9$ for synthetic data generation for both the models.

Model Finetuning: We finetune LLAMA-3.1-8B[23] using the strategies described in Section 3. We experiment with three distinct model configurations as follows. The hyperparameters are in Table 2.
• M1: Baseline Model: LLAMA-3.1-8B was exclusively fine-tuned for 4 epochs on the original 13,092 training instances.
• M2: Continued Fine-Tuning: The 2 epoch trained baseline model (M1) underwent a second phase of fine-tuning for another 2 epochs on a 33,000-instance synthetic dataset generated by Phi-4-14B.
• M3: Multi-Source Fine-Tuning: This model was fine-tuned for 4 epochs on a combined 50,000-instance dataset, which included the original 13,000 training instances along with synthetic data from two distinct large language models: 33,000 instances from Phi-4-14B and 4,000 instances from LLAMA-3.1-70B.

Parameter	Value
Max sequence length	4096
Per-device train batch size	2
Gradient accumulation steps	4
Warmup steps	5
Learning rate scheduler type	"cosine"
Number of epochs	4

Table 2
Fine-Tuning Parameters for LLAMA-3.1-8B.

Evaluation: We report our model's performance using token-based metrics, ROUGE-L[1] and BLEU[2]. In our implementation of the ROUGE-L scores presented in Table 3 we modify of the default tokenization function to incorporate Hindi words and characters. For the semantics-based metric, BERTScore [3] over predicted answers and gold answers on validation and test splits.

6. Results and Analysis

In this section, we first provide our evaluation of the different training strategies on Validation and Test Data-1. Then, we present the organizers' evaluation on Test Data-2.
Validation and Test Data-1: Our experiments presented in Table 3, revealed several key findings regarding model training and distribution robustness. First, as shown in Table 3, the model trained only on the original data (M1[5]) performed best on the development set, while the model with combined original and synthetic data (M3[5]) excelled on Test Data-1. This disparity highlights the models' sensitivity to distribution shifts.

However, the combined multi-source model (M3[5]) underperformed in BLEU-2, which is a precision based metric. A potential reason is that combining multi-source data may introduce conflicting answers for similar questions, leading to ambiguity and performance degradation.
Held-out Test Data-2: Table 4 demonstrates our method's performance on the proprietary and undisclosed Test Data-2 split. Our model configurations has consistent top rankings with M2[5] outperforming other models in QA-F1, a second-place ranking in BLEU-2, a third-place ranking in BLEU-1 and fourth-place ranking in ROUGE-1 and ROUGE-2. The results indicate that supplementing models with a large quantity of high-quality synthetic data can not only improve performance on downstream tasks but also significantly enhance their robustness to unseen data.

A crucial insight comes from the two-stage trained model (M2[5]). Despite its moderate performance on Test Data-1, it achieved superior BLEU-2 and QA-F1 scores on the held-out Test Data-2 (Table 4). This suggests that late exposure to synthetic data is effective for building distribution robustness.
Analysing the Synthetic data: Table 1 outlines the quantitative differences between the original and synthetic datasets, while Table 5 presents a qualitative comparison of the question-answer (QA) pairs

[1]https://huggingface.co/spaces/evaluate-metric/rouge
[2]https://huggingface.co/spaces/evaluate-metric/bleu
[3]https://huggingface.co/spaces/evaluate-metric/bertscore

Model	Settings	Validation Data			Test Data-1		
		Rouge-L	BLEU	BERTScore	Rouge-L	BLEU	BERTScore
M1	Orig 13k	**0.897**	**0.799**	**0.971**	0.917	0.838	0.976
M2	M1 + 33k	0.893	0.790	0.969	0.911	0.828	0.975
M3	All 50k	0.892	0.791	0.968	**0.922**	**0.849**	**0.978**

Table 3
Performance of our model settings on Dev set and Test Data-1 splits of publicly available VATIKA Dataset.

BLEU-1		BLEU-2		ROUGE-1		ROUGE-2		ROUGE-L		QA-F1	
Teams	**Score**	**Teams**	**Score**	**Teams**	**Score**	**Teams**	**Score**	**Teams**	**Score**	**Teams**	**Score**
Scalar1	63.7	Scalar1	41.2	IRel3	0.0824	IRel3	0.0467	IRel3	0.0824	**OUR_M2**	**0.576**
IRel3	61.5	**OUR_M2**	**38.7**	AiNauts1	0.0818	AiNauts1	0.0454	AiNauts1	0.0818	**OUR_M3**	**0.566**
OUR_M3	**60.9**	**OUR_M1**	**38.4**	CSE_SVNIT3	0.0790	CSE_SVNIT3	0.0452	CSE_SVNIT3	0.0790	**OUR_M1**	**0.562**
OUR_M2	**60.1**	IRel3	36.4	**OUR_M2**	**0.0777**	**OUR_M1**	**0.0424**	**OUR_M1**	**0.0777**	IRel3	0.551
NamasteNLP1	59.9	AiNauts2	33.2	**OUR_M1**	**0.0773**	**OUR_M2**	**0.0415**	**OUR_M3**	**0.0773**	Scalar1	0.505
AiNauts1	56.5	NamasteNLP1	31.7	**OUR_M3**	**0.0763**	MUCS1	0.0438	**OUR_M2**	**0.0763**	IRel2	0.461
AiNauts2	55.8	IRel1	30.5	MUCS1	0.0759	**OUR_M3**	**0.0411**	MUCS1	0.0759	AiNauts1	0.453
Scalar-2	52.3	IRel2	28.6	MUCS3	0.0685	MUCS3	0.0411	MUCS3	0.0685	CSE_SVNIT1	0.433
IRel1	51.2	AiNauts1	27.2	NamasteNLP3	0.0685	NamasteNLP3	0.0392	NamasteNLP3	0.0685	IRel1	0.417
OUR_M1	**49.5**	**OUR_M3**	**27.0**	NamasteNLP2	0.0626	NamasteNLP2	0.0360	NamasteNLP2	0.0626	CSE_SVNIT2	0.394

Table 4
Official results from the FIRE VATIKA Competition were tested on a private Test Data-2 split. The different model configurations for our method (VA-BO-INTERN) have been highlighted in bold.

they generated. The outputs from LLAMA-3.1-70B and Phi-4-14B demonstrate considerable overlap, underscoring the critical importance of a data selection stage to filter for quality. A potential direction for future research in developing such quality checks is the joint evaluation of both question and answer within each synthetic pair.

7. Conclusion and Future Work

In this work, we proposed a multi-stage finetuning strategy for lightweight language models in the Hindi tourism domain, leveraging both original and synthetic training data. Models trained with continued finetuning-first on original data, then on synthetic data-consistently outperformed alternative approaches. This staged exposure allows the model to retain grounding in authentic data while benefiting from the scale and diversity of synthetic examples, improving robustness in domain-specific question answering. We also found that indiscriminate or excessive mixing of multi-source synthetic data can degrade performance, highlighting the importance of careful curation and controlled integration in low-resource settings.

As a future work our approach can be extended to other low-resource languages to test its generalizability. Future work also includes systematic evaluation of synthetic data quality, potentially using LLM-based filtering methods. Overall, this study demonstrates that large models can generate synthetic data, but small models can effectively adapt to it, enabling scalable and robust QA systems in low-resource domains.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini-Flash 2.5 in order to rectify: Grammar, spelling check and evaluate microstructure. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Model	Questions and Answers (translated from Hindi for representation)
Context	Bhagirath Kund is 14.1 kilometers away from Pt. Deen Dayal Upadhyaya Railway Station. It can be reached by taxi, cab, or bus. This station was previously known as Mughalsarai and is one of India's major railway junctions. The journey offers views of Varanasi's historic lanes and ghats and provides an experience of its cultural heritage.
Original Training Data	Q1: How far is Bhagirath Kund from Pt. Deen Dayal Upadhyaya Railway Station? A1: Bhagirath Kund is 14.1 kilometers away from the station. Q2: How can one reach Bhagirath Kund from Pt. Deen Dayal Upadhyaya Railway Station? A2: One can reach Bhagirath Kund by taxi, cab, or bus services from the station.
LLAMA-3.1-70B	Q1: What was the previous name of Pt. Deen Dayal Upadhyaya Railway Station? A2: It was earlier known as Mughalsarai. Q1: Is Pt. Deen Dayal Upadhyaya Railway Station an important railway junction? A2: Yes, it is one of India's major railway junctions. Q3: What kind of views are seen on the journey to Bhagirath Kund? A3: The journey offers views of Varanasi's historic lanes and ghats. Q4: What kind of experience do devotees get on this journey? A4: Devotees experience Varanasi's cultural heritage, enhancing the importance of the pilgrimage.
Phi-4-14B	Q1: What was the old name of Pt. Deen Dayal Upadhyaya Railway Station? A1: Its old name was Mughalsarai. Q2: In which category does Pt. Deen Dayal Upadhyaya Railway Station fall? A2: It is one of India's major railway junctions. Q3: What kind of views does the journey to Bhagirath Kund provide? A3: It provides views of Varanasi's historic lanes and ghats. Q4: What kind of experience do devotees encounter on the journey? A4: Devotees experience Varanasi's cultural heritage, enhancing the significance of the site. Q5: Where is Pt. Deen Dayal Upadhyaya Railway Station located? A5: It is located in Varanasi.

Table 5
Examples of synthetic question–answer pairs from LLAMA-3.1-70B and Phi-4-14B, along with training data. The original dataset is in Hindi; the questions and answers are translated into English for representation.

References

[1] K. Sun, Y. Xu, H. Zha, Y. Liu, X. L. Dong, Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs?, in: NAACL, 2024, pp. 311–325.

[2] C. Wang, P. Liu, Y. Zhang, Can generative pre-trained language models serve as knowledge bases for closed-book qa?, in: ACL-IJCNLP, 2021, pp. 3241–3251.

[3] N. Scaria, S. D. Chenna, D. Subramani, How good are Modern LLMs in generating relevant and high-quality questions at different bloom's skill levels for Indian high school social science curriculum?, in: Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), Mexico City, Mexico, 2024, pp. 1–10.

[4] X. Yuan, T. Wang, Y.-H. Wang, E. Fine, R. Abdelghani, H. Sauzéon, P.-Y. Oudeyer, Selecting better samples from pre-trained LLMs: A case study on question generation, in: Findings of the ACL, Toronto, Canada, 2023, pp. 12952–12965.

[5] M. Schmidt, A. Bartezzaghi, N. T. Vu, Prompting-based synthetic data generation for few-shot question answering, in: LREC-COLING, 2024, pp. 13168–13178.

[6] K. Tengler, G. Brandhofer, Exploring the difference and quality of ai-generated versus human-written texts, Discover Education 4 (2025) 113.

[7] H. T. Hakam, R. Prill, L. Korte, B. Lovreković, M. Ostojić, N. Ramadanov, F. Muehlensiepen, Human-written vs ai-generated texts in orthopedic academic literature: Comparative qualitative analysis,

JMIR Formative Research 8 (2024) e52164.
[8] Y. K. Chia, L. Bing, S. Poria, L. Si, RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction, in: Findings of the ACL, 2022, pp. 45–57.
[9] S. Liu, Y. Li, J. Li, S. Yang, Y. Lan, Unleashing the power of large language models in zero-shot relation extraction via self-prompting, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, 2024, pp. 13147–13161.
[10] V. Yadav, H. j. Kwon, V. Srinivasan, H. Jin, Explicit over implict: Explicit diversity conditions for effective question answer generation, in: LREC-COLING 2024, 2024, pp. 6876–6882.
[11] A. Chowdhury, A. Chadha, Generative data augmentation using LLMs improves distributional robustness in question answering, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, 2024, pp. 258–265.
[12] A. Kramchaninova, A. Defauw, Synthetic data generation for multilingual domain-adaptable question answering systems, in: Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, 2022, pp. 151–160.
[13] C. Harsha, K. S. Phogat, S. Dasaratha, S. A. Puranam, S. Ramakrishna, Synthetic data generation using large language models for financial question answering, in: Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal), Abu Dhabi, UAE, 2025, pp. 76–95.
[14] A. Namboori, S. Mangale, A. Rosenbaum, S. Soltan, Gemquad: Generating multilingual question answering datasets from large language models using few shot learning, arXiv e-prints (2024) arXiv–2404.
[15] R. Chada, P. Natarajan, FewshotQA: A simple framework for few-shot learning of question answering tasks using pre-trained text-to-text models, in: EMNLP, 2021, pp. 6081–6090.
[16] X. Chen, J.-Y. Jiang, W.-C. Chang, C.-J. Hsieh, H.-F. Yu, W. Wang, MinPrompt: Graph-based minimal prompt data augmentation for few-shot question answering, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 254–266.
[17] A. Ushio, F. Alva-Manchego, J. Camacho-Collados, An empirical comparison of LM-based question and answer generation methods, in: Findings of the ACL, Toronto, Canada, 2023, pp. 14262–14272.
[18] D. Gurgurov, M. Hartmann, S. Ostermann, Adapting multilingual LLMs to low-resource languages with knowledge graphs via adapters, in: Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024), 2024, pp. 63–74.
[19] A. K. Singh, V. Kumar, R. Murthy, J. Sen, A. Mittal, G. Ramakrishnan, INDIC QA BENCHMARK: A multilingual benchmark to evaluate question answering capability of LLMs for Indic languages, in: Findings of NAACL, Albuquerque, New Mexico, 2025, pp. 7689–7698.
[20] S. Kumar, R. Sanasam, S. Nandi, IndiSentiment140: Sentiment analysis dataset for Indian languages with emphasis on low-resource languages using machine translation, in: NAACL, 2024, pp. 7689–7698.
[21] S. Doddapaneni, R. Aralikatte, G. Ramesh, S. Goyal, M. M. Khapra, A. Kunchukuttan, P. Kumar, Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages, in: ACL, 2023, pp. 12402–12426.
[22] P. Gatla, Anushka, N. Kanwar, G. Sahoo, R. K. Mundotiya, Tourism question answer system in indian language using domain-adapted foundation models, arXiv preprint (2025).
[23] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, CoRR (2024).
[24] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al., Phi-4 technical report, arXiv preprint arXiv:2412.08905 (2024).