Quality over Quantity in Supervised Fine Tuning: A Token-Matched Case Study on ALLaM

Manar Alrabie
Umm Al-Qura University
Makkah, Saudi Arabia
s44680509@uqu.edu.sa

Tahani Alsubait
Umm Al-Qura University
Makkah, Saudi Arabia
tmsubait@uqu.edu.sa

Abstract

Large Language Models (LLMs) struggle with low-resource languages like Arabic due to the scarcity of high-quality, culturally grounded instruction data. This paper examines whether curated Arabic instruction data yields better fine-tuning performance than synthetic data when training exposure is held constant via token budget control. ALLaM-7B-Instruct is fine-tuned on two datasets: CIDAR, a human-curated Arabic instruction set, and a token-matched subset of a large synthetic Arabic corpus, across three token budgets (500K, 1M, and 2M tokens). Both models are evaluated on seven public Arabic benchmarks spanning question answering, commonsense reasoning, and cultural alignment. Preliminary results consistently show that curated data outperforms synthetic data across most token budgets and benchmarks, demonstrating that data quality is a more reliable driver of fine-tuning performance than data quantity under token-budget constraints.

1 Introduction

Large Language Models (LLMs) have made significant progress across many languages, yet Arabic continues to lag behind due to the scarcity of high-quality instruction-tuning data. Researchers have increasingly turned to synthetic or machine-translated datasets (Allam et al.), which are easy to scale but often lack linguistic richness, cultural relevance, and may amplify Western-centric biases (Wang et al.). Moreover, excessive reliance on synthetic data risks *model collapse* the gradual erosion of rare linguistic patterns as synthetic content dominates training (Shumailov et al.).

This raises a critical question: **under fixed token constraints, is fine-tuning Arabic LLMs more effective with synthetic quantity or curated quality?** We hypothesize (**H1**) that under a fixed token budget, fine-tuning on human-curated Arabic instruction data yields better Arabic-benchmark performance than fine-tuning on a token-matched synthetic dataset. Despite its practical importance, this question remains unanswered in Arabic NLP (Chouikhi et al., 2024a).

This is addressed by comparing CIDAR (Alyafeai et al., 2024), a human-curated Arabic instruction dataset, against a token-matched synthetic subset across three budgets (500K, 1M, 2M tokens), evaluated on seven Arabic benchmarks.Consistent improvements are observed in our results when curated data is used over synthetic data, and practical guidance is provided for building Arabic LLMs under resource constraints.

Contributions.

- **Empirical comparison:** A controlled study between CIDAR and a token-matched synthetic subset across three token budgets.
- **Comprehensive evaluation:** Assessment on seven Arabic benchmarks spanning question answering, commonsense reasoning, and cultural alignment.
- **Experimental framework:** A token-budget controlled design enabling fair comparison between data quality and quantity.
- **Practical guidance:** Evidence-based recommendations for building Arabic LLMs under limited resource constraints.

2 Related Work

Arabic LLM development has advanced through models such as Jais (Sengupta et al.,

2023), AceGPT (Huang et al., 2024), ALLaM (Bari et al., 2024), and GemmAr (Chouikhi et al., 2024b), most of which rely on synthetic or machine-translated instruction data to compensate for Arabic data scarcity. However, synthetic data carries well-documented risks: cultural bias toward Western norms (Wang et al.; Naous et al.) and model collapse from iterative synthetic training (Shumailov et al.).

On the data efficiency side, Zhou et al. (Zhou et al.) demonstrate that 1,000 curated examples can match GPT-4-level performance, and Kadasi et al. (Kadasi et al., 2025) show that token budget control enables fair data comparisons. Chen et al.(Chen et al., 2024) further showed that filtering instruction data by quality scores improves downstream performance even at reduced scale.

Notably, high-quality synthetic data has proven competitive in some settings. Gunasekar et al.(Gunasekar et al., 2023) showed that textbook-quality synthetic data can train competitive small models, and Mukherjee et al.(Mukherjee et al., 2023) demonstrated that GPT-4-generated explanations improve reasoning through progressive learning. However, these approaches depend on powerful proprietary models for data generation, which may introduce cultural bias in non-English settings such as Arabic (Wang et al.; Naous et al.).

Unlike prior work, data quality versus quantity is directly isolated under fixed token budgets for Arabic instruction tuning.

3 Methodology

3.1 Base Model

ALLaM-7B-Instruct (Bari et al., 2024), a decoder-only Arabic-English LLM, is fine-tuned and kept fixed across all experiments to isolate data quality as the sole variable.

3.2 Datasets

A comparison is conducted between two Arabic instruction-tuning datasets. Only synthetically generated samples from InstAr-500k (291,252 after filtering) are retained to ensure a clean quality contrast against CIDAR.

3.3 Dataset Preprocessing

The preprocessing pipeline is shown in Figure 1. Raw InstAr-500k samples were filtered

Table 1: Dataset Comparison

Property	CIDAR	InstAr-500k
Size	10,000	291,252
Curation	Human	Synthetic
Localization	Manual	Automated
Role	Quality	Quantity

to retain only synthetically generated entries, removing samples shorter than 20 or longer than 6,000 characters. All pairs were formatted using the ALLaM-7B-Instruct instruction template:

`<s>[INST] instruction [/INST] response </s>` (1)

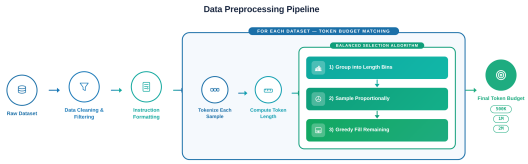

Figure 1: Preprocessing Pipeline

3.4 Token Budget Design

Both datasets are sampled to identical token budgets using a length-balanced sampling algorithm ensuring proportional representation across token-length bins, with a 2% tolerance for practical feasibility.

Table 2: Token Budget Experiments

Experiment	Budget	CIDAR Coverage
1	500K	25%
2	1M	50%
3	2M	100%

3.5 Experimental Design

The overall workflow is illustrated in Figure 2.

3.6 Fine-Tuning Configuration

All models are fine-tuned using QLoRA (Dettmers et al.) in 4-bit NF4 quantization on an NVIDIA A100 (40GB) via Lambda Cloud.

3.7 Evaluation

All six fine-tuned models are evaluated using `lm-evaluation-harness` on seven public Arabic benchmarks: ArabicMMLU, AlGhafa, AraDiCE, ACVA, Arabic Exams, Arab Culture, and OpenAI MMLU (Arabic). To rule out data leakage, normalized 8-gram overlap between each training set and the evaluation

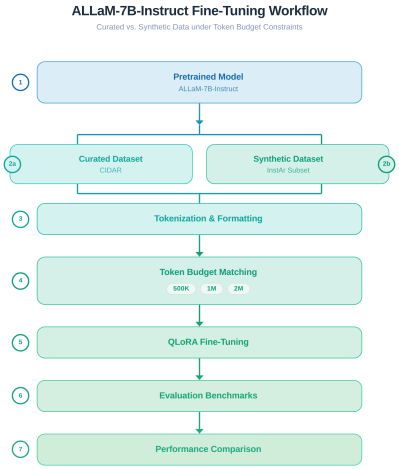

Figure 2: Controlled Experimental Framework

Table 3: Training Configuration

Parameter	Value
LoRA Rank / Alpha	16 / 32
LoRA Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Loss Masking	Instruction tokens only
Optimizer	Paged AdamW (8-bit)
Learning Rate	2×10^{-4}
LR Scheduler	Cosine decay
Warmup Steps	100
Training Epochs	3
Batch Size	64
Gradient Clipping	0.3
Random Seed	42 (fixed across all runs)
Hardware	NVIDIA A100 (40GB)

benchmarks was measured. Overlap was negligible across all benchmarks at most 0.04% for CIDAR and 2.65% for the synthetic subset confirming that observed differences reflect genuine generalization rather than data contamination.

4 Results and Discussion

4.1 Overall Performance

Table 4 presents evaluation results at the 500K token budget. Curated models outperform synthetic models on all seven benchmarks, with the largest gaps on AlGhafa (+5.2pp), ACVA (+2.9pp), and AraDiCE (+2.2pp) all well outside the reported standard errors. Smaller differences such as Arab Culture (+0.001) are within measurement noise. The capability profile in Figure 3 further illustrates the consistent curated advantage across all benchmarks.

Table 4: Evaluation results at 500K token budget with standard error. C = Curated, S = Synthetic. Bold = higher score.

Benchmark	Base	C-500K	S-500K
Arab Culture	0.348	**0.361**±0.008	0.360±0.008
AlGhafa	0.392	**0.594**±0.003	0.542±0.003
AraDiCE†	0.450	**0.581**±0.004	0.559±0.004
ACVA	0.763	**0.770**±0.004	0.741±0.005
Arabic Exams	0.348	**0.508**±0.022	0.497±0.022
ArabicMMLU	0.678	**0.645**±0.004	0.635±0.004
OpenAI MMLU	0.559	**0.412**±0.004	0.372±0.004

†AraDiCE is the mean of Egyptian (0.572/0.549) and Levantine (0.590/0.569) dialect subsets; SE is the pooled standard error of the two subsets.

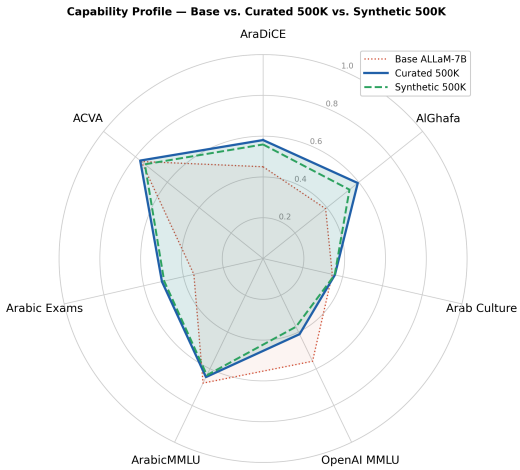

Figure 3: Capability profile comparing base ALLaM-7B, curated (C-500K), and synthetic (S-500K) models across all seven benchmarks.

4.2 Key Findings

Curated data consistently outperforms synthetic. Curated models outperform synthetic models across all seven benchmarks at the 500K token budget. The gaps are most pronounced on AlGhafa (+5.2pp, ±0.003), ACVA (+2.9pp, ±0.004), and AraDiCE (+2.2pp, ±0.004), all well outside the reported standard errors. On OpenAI MMLU the gap reaches +4.0pp (±0.004). These results demonstrate that data quality provides a more reliable supervision signal than data quantity under a fixed token budget.

Cultural benchmarks are most sensitive to data source. ACVA and Arab Culture show the clearest difference between curated and synthetic data. The curated model outperforms the synthetic model on both benchmarks, reflecting the higher cultural fidelity of human-curated instruction examples. Fig-

ure 4 illustrates this concretely: when asked about a traditional Saudi dinner, the curated model correctly identifies Marqouq and Jareesh, while the synthetic model selects toasted sandwiches a Westernised response with no grounding in Saudi food culture, consistent with the cultural bias risks documented by Wang et al. and Naous et al.. Figure 4 illustrates the effect qualitatively: when asked about a traditional Saudi dinner, the curated model correctly identifies *Marqouq* and *Jareesh*, while the synthetic model selects toasted sandwiches.

Cultural Bias: Saudi Dinner Customs

Q	Sa'eed's family eats dinner together. What is served?
A	يتكون العشاء من السلطة و المتبل *yatakawwan al-'asha' min al-salata wal-mutabbal* *(Salad and Mutabbal)*
B	يتكون العشاء من ساندوتشات التوست *yatakawwan al-'asha' min sandawitchat al-tost* *(Toasted sandwiches)*
C	يتكون العشاء من المرقوق و الجريش *yatakawwan al-'asha' min al-marqouq wal-jareesh* *(Marqouq and Jareesh)*
Correct	C
Curated	C
Synthetic	B ×

Figure 4: Synthetic model picks Westernised option over authentic Saudi dish.

Fine-tuning degrades broad knowledge benchmarks. ArabicMMLU and OpenAI MMLU degrade below the base model across all fine-tuned variants, reflecting catastrophic forgetting (McCloskey and Cohen, 1989). This effect is more severe for synthetic models, which lose up to 19.6pp on OpenAI MMLU compared to 14.7pp for curated models, suggesting that the broader topical coverage of curated data partially mitigates knowledge loss.

5 Conclusion

This paper presents a preliminary empirical investigation into whether human-curated instruction data outperforms synthetic data when fine-tuning Arabic LLMs under a controlled token budget. Six variants of ALLaM-7B-Instruct were fine-tuned on CIDAR (curated) and InstAr-500k (synthetic) and evaluated on seven Arabic benchmarks.

Results support H1: curated models outperform synthetic models on all seven benchmarks at the 500K token budget, with the strongest advantages on AlGhafa (+5.2pp), ACVA (+2.9pp), and AraDiCE (+2.2pp). Catastrophic forgetting was observed on knowledge-intensive benchmarks across all variants, but was more severe under synthetic fine-tuning. For resource-constrained Arabic NLP settings, these findings suggest that investing in data quality is more effective than scaling synthetic data pipelines. Future work should explore larger token budgets, open-ended generation evaluation, and hybrid curated-synthetic strategies.

Limitations

This study is limited to a single base model (ALLaM-7B-Instruct) and reports results at the 500K token budget in this preliminary version; full results across all three token budgets are reported in the extended journal version currently under preparation. Findings may not generalize to other Arabic LLMs or larger data scales. Evaluation is restricted to multiple-choice benchmarks, which may not fully capture open-ended generative capabilities.

By construction, human curation and data source are inseparable in this comparison: CIDAR and the synthetic subset differ in origin precisely because of how each is produced, and this difference is the variable under study. We nonetheless acknowledge that the two datasets also differ in topic distribution and formatting; our budget-controlled design isolates the effect of dataset choice as a whole rather than curation alone. Constructing a topic- and format-matched synthetic baseline is left to future work. The proposed link between curated data and reduced catastrophic forgetting via broader topical coverage is observed but not formally quantified. Finally, our contamination analysis relies on n-gram overlap, which detects verbatim but not paraphrased overlap.

References

Abdulrahman Allam, Seif Ahmed, Ali Hamdi, and Khaled Shaban. Scaling arabic medical chatbots using synthetic data: Enhancing genera-

tive AI with synthetic patient records. *Preprint*, arxiv:2509.10108 [cs].

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran A. Q. Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged S. Al-Shaibani. 2024. Cidar: Culturally relevant instruction dataset for arabic. (arXiv:2402.03177). ArXiv:2402.03177.

M. Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan Al-Rashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2024. Allam: Large language models for arabic and english. (arXiv:2407.15390). ArXiv:2407.15390.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. AlpaGasus: Training a better Alpaca with fewer data. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Hasna Chouikhi, Manel Aloui, Cyrine Ben Hammou, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024a. Gemmar: Enhancing llms through arabic instruction-tuning. *Preprint*, arXiv:2407.02147.

Hasna Chouikhi, Manel Aloui, Cyrine Ben Hammou, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024b. Gemmar: Enhancing llms through arabic instruction-tuning. (arXiv:2407.02147). ArXiv:2407.02147.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLORA: efficient finetuning of quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 10088--10115. Curran Associates Inc.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. Acegpt, localizing large language models in arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 81398163, Mexico City, Mexico. Association for Computational Linguistics.

Pritam Kadasi, Abhishek Upperwal, and Mayank SIngh. 2025. Adapt: Learning task mixtures for budget-constrained instruction tuning. *Preprint*, arXiv:2512.04555.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109--165. Academic Press.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv preprint arXiv:2306.02707*.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366--16393. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *Preprint*, arxiv:2305.17493 [cs].

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349--6384. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: less is more for alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 55006--55021. Curran Associates Inc.