Evaluating the Impact of LoRA Fine-Tuning on A1–A2 Level English Sentence Generation

Muhammet Emin Aydınalp
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Türkiye
muhammet.aydinalp@izu.edu.tr
ORCID: 0009-0004-9423-0217

Ayşe Berna Altınel Girgin
Department of Computer Engineering
Marmara University
Istanbul, Türkiye
berna.altinel@marmara.edu.tr
ORCID: 0000-0001-5544-0925

Abdullah Sönmez
Department of Computer Engineering
Istanbul Sabahattin Zaim University
Istanbul, Türkiye
abdullah.sonmez@izu.edu.tr
ORCID: 0000-0001-7148-3378

***Abstract*—In language learning, practicing with grammatically correct and level-appropriate sentences for the target vocabulary is a fundamental element of vocabulary acquisition. However, it is not always possible to access a sufficient number of level-appropriate example sentences for the target vocabulary. In this study, we investigate the effect of Low-Rank Adaptation (LoRA) based fine-tuning using synthetic data on the ability of open-source large language models (LLMs) to generate level-appropriate English sentences at CEFR levels A1 and A2. We used the Llama-3.2-1B-Instruct and Llama- 3.1-8B-Instruct models. We fine-tuned the Llama-3.2-1B-Instruct model using the LoRA method with synthetic data systematically generated with GPT-4o-mini. The generated sentences were evaluated by two different LLM evaluators based on four linguistic criteria: word usage, grammatical accuracy, clarity, and level appropriateness. The results show that the fine-tuned model performed better than the base model in both evaluators and, in many cases, outperformed the larger LLaMA 8B model. These findings demonstrate that small-scale open-source language models, when trained with domain-specific and well-structured synthetic data, can deliver significant performance gains in goal-oriented tasks such as language learning.**

***Index Terms*—LoRA fine-tuning, English sentence generation, CEFR A1–A2, large language models, synthetic data**

I. Introduction

The role of language learning in international communication is central, and English is the most widely spoken language worldwide [1], [2]. Vocabulary learning is a crucial aspect of language acquisition; however, when words are learned solely through memorization, their long-term retention is unlikely [3]. Hence, it is crucial to learn vocabulary in meaningful, correct and level-dependent sentences in order to retain it effectively. Since this is a requirement, both learners and instructors often struggle to find grammatically accurate sentence examples that are appropriate for specific proficiency levels [4].

Recent advancements in LLMs have led to transformative educational opportunities, offering new possibilities for material preparation, personalized feedback, and interactive learning environments [5]–[9]. However, existing studies have generally focused on the use of commercial models or on the application of open-source models in their base (non-fine-tuned) form.

The benefits of open-source models include increased data privacy, lower cost, and offline usability [10], [11]. Nevertheless, when these smaller models are used in their base (non-modified) form, they tend to perform worse than larger commercial and open-source systems [12]. To address these shortcomings and preserve the advantages of open architectures, recent work has focused on optimizing smaller models using parameter-efficient methods, such as LoRA [13]. LoRA achieves this efficiency by freezing the original pre-trained weights and adding a small number of trainable, low-rank matrices to approximate the weight updates, thereby significantly decreasing the computation and storage costs during fine-tuning.

In this research, we explore how well small-scale open-source LLMs can be fine-tuned to produce English sentences at the A1 and A2 CEFR levels using a synthetic dataset generated by GPT-4o-mini. We fine-tuned Llama-3.2-1B-Instruct [14] and evaluated its performance against both its non-fine-tuned base version and the larger Llama-3.1-8B-Instruct model.

Our technical analyses and evaluations conducted using two distinct LLM-based evaluators indicate that the fine-tuned 1B model significantly outperforms its base version. Specifically, the fine-tuning process significantly improved text generation performance, reflected in a reduction in perplexity of over 77% across both proficiency levels.

Our study has made the following contributions:

- **Generation of a Special Synthetic Data:** We created a specific dataset of 4,495 sentences on the A1 level and 4,365 sentences on the A2 level with the help of GPT-4o-mini to overcome the lack of level-specific educational data.
- **Evidence of Parameter-Efficient Fine-Tuning Efficiency:** We experimentally demonstrated that fine-tuning based on LoRA on a 1B parameter model causes an improvement in perplexity reduction at foundational level of proficiency of more than 70%.
- **Comparative analysis of Open-source Models:** We gave a comparative analysis that indicated that when given well-structured data, small models can be scalable alternatives to larger models such as Llama-3.1-8B on specific educational tasks.

II. Methodology

In this study, we propose a parameter-efficient fine-tuning approach to specialize small language models for CEFR-based sentence generation. This section details the dataset creation, the models employed, the fine-tuning procedure, and the evaluation pipeline.

A. Dataset Preparation

The task defined for the language models in this study is to generate a batch of 10 distinct sentences corresponding to a provided list of 10 target words. To generate high-quality training material for this purpose, we automated the data creation process by developing a Python script that interacts with the GPT-4o API. We prepared the synthetic dataset as a list in which each training sample consists of an instruction with 10 randomly selected words and a target output of 10 sentences, rather than individual word-sentence pairs, enabling the model to learn batch generation while preserving context across the list. As shown in Table I, we generated 5 different sentences for each target word selected from the Oxford English Dictionary corresponding to the target CEFR level. The dataset contained 899 words and 4,495 sentences at the A1 level, and 873 words and 4,365 sentences at the A2 level. Using these word-sentence pairs, we created 2,000 distinct training samples per level in the list format described earlier. The dataset was split into a 90% training set and a 10% evaluation set.

TABLE I
Dataset Statistics for A1 and A2 Levels

Level	Num. of Words	Sentences/Word	Total Sentences
A1	899	5	4,495
A2	873	5	4,365
Total	**1,772**	**-**	**8,860**

B. Models Used

In order to examine the trade-off between the size of the model and domain-specific fine-tuning, we paid attention to the Llama 3 family of open-source models. Specifically, we compared three configurations:

- **Llama-3.2-1B-Instruct (Base):** A lightweight model optimized for edge devices, used as the baseline to measure the impact of fine-tuning.
- **Llama-3.2-1B-Instruct (Fine-Tuned):** The version of the 1B model fine-tuned using our synthetic dataset.
- **Llama-3.1-8B-Instruct:** A larger, general-purpose open-source model used to benchmark whether a smaller, fine-tuned model can compete with a larger architecture.

C. Fine-Tuning Procedure

We used LoRA [13] to fine-tune the Llama-3.2-1B model. LoRA enables efficient adaptation by freezing pre-trained model weights and injecting trainable low-rank matrices into the Transformer architecture.

The training was conducted using the `SFTTrainer` from the TRL library with Unsloth optimization. The specific hyperparameters and LoRA configuration used in the experiments are detailed in Table II.

TABLE II
LoRA Configuration and Training Hyperparameters

Parameter	Value
Rank (r)	8
LoRA Alpha (α)	32
LoRA Dropout	0.05
Target Modules	q_proj, v_proj
Learning Rate	2×10^{-4}
Num. of Epochs	10
Batch Size	16
Optimizer	AdamW (8-bit)

D. Working Environment

All experiments were conducted on a local workstation. The specific hardware and software configurations used in this study are detailed below:

- **Hardware Infrastructure:** Single NVIDIA GeForce RTX 4090 GPU with 24 GB VRAM
- **Operating System:** Linux (Ubuntu 22.04 LTS).
- **Software Environment:** Python 3.11.5 and PyTorch framework with CUDA support.
- **Development Environment:** Visual Studio Code.
- **Fine-Tuning Framework:** Unsloth library was utilized alongside Hugging Face Transformers for efficient parameter-efficient fine-tuning.
- **Model Execution:** *Text Generation WebUI* was employed specifically to run and test the open-source models.
- **Data Processing & Visualization:** NumPy (v2.3.2) and Pandas (v2.3.1) were used for data manipulation, while Matplotlib (v3.10.5) was used for generating the figures.

E. Evaluation Methodology

To ensure a consistent and objective assessment, we utilized an automated evaluation pipeline driven by two advanced Large Language Models: **ChatGPT-5** [15] and **DeepSeek-Reasoner** [16].

1) Sentence Generation: During the testing phase, evaluation was conducted using randomly selected words from the A1 and A2 lists that were not included in the training data. Each model was prompted to generate sentences for these words in batches of 10, adhering to the same format used during training.

2) Scoring Criteria: We automatically routed word-sentence pairs generated by open-source models to LLM evaluators via API and only asked them to score them numerically on a scale of 1-5. We specified this scoring in the instruction prompt and enabled them to score separately according to the four criteria shown in Table III.

Each generated sentence was evaluated twice to ensure consistency. The final score for each sample was derived by calculating the average of these two independent ratings.

TABLE III
EVALUATION CRITERIA FOR SENTENCE GENERATION

Criterion	Description
Word Usage	Accuracy of the target word's meaning and contextual placement.
Clarity	The sentence is meaningful, unambiguous, and easy to understand.
Grammar	Syntactic correctness and absence of structural errors.
Level Appr.	Suitability of vocabulary and sentence structure for the target CEFR level.

III. RESULTS AND DISCUSSION

This section provides an extensive discussion of the exexperimental findings, which assess the effect of the LoRA-based fine- tuning to the Llama- 3.2-1B model. We test the fine-tuned model against its baseline counterpart as well as the bigger Llama- 3.1-8B model on technical measures and linguistic quality evaluation.

A. Technical Findings of the Fine-Tuning Process

To assess the efficacy of the fine-tuning process, we first used quantitative training metrics. The model converged, with evaluation loss decreasing significantly at both CEFR levels, as shown in Table IV.

TABLE IV
TRAINING AND EVALUATION LOSS METRICS BY LEVEL

Level	Initial Train Loss	Final Train Loss	Best Eval Loss
A1	1.68	0.41	0.57
A2	1.76	0.38	0.51

We also compared the perplexity (PPL) of the base and fine-tuned models to examine the effect of fine-tuning. As shown in Table V, perplexity decreases substantially after fine-tuning, suggesting improved confidence and accuracy in predicting target sequences. A perplexity reduction of approximately 77-79% suggests that the model has successfully captured the sentence structures characteristic of the A1 and A2 proficiency levels.

TABLE V
PERFORMANCE IMPROVEMENT: BASE VS. FINE-TUNED MODEL

Level	Base PPL	FT PPL	Improvement (%)
A1	9.78	2.23	77.2%
A2	10.24	2.14	79.1%

B. Overall Model Performance

The linguistic quality of the generated sentences was evaluated by two LLM-based judges, ChatGPT-5 and DeepSeek-Reasoner. Fig. 1 illustrates the average scores obtained under each evaluator, with the y-axis scaled from 4.0 to 5.2 to highlight performance differentials.

Under the DeepSeek-Reasoner evaluation, the fine-tuned 1B model achieved an average score of 4.89, while the base model obtained 4.40 and the Llama-8B model achieved 4.79. Similarly, under the ChatGPT-5 evaluation, the fine-tuned model obtained a score of 4.81, whereas the base model achieved 4.38 and the Llama-8B model achieved 4.67.

When averaged across evaluators, the fine-tuned 1B model achieved an overall score of **4.85**, surpassing the base model by a substantial margin and outperforming the larger Llama-8B model, which achieved an average score of 4.73. This result demonstrates that a small, specialized model can achieve superior performance compared to a larger general-purpose model in domain-specific tasks.

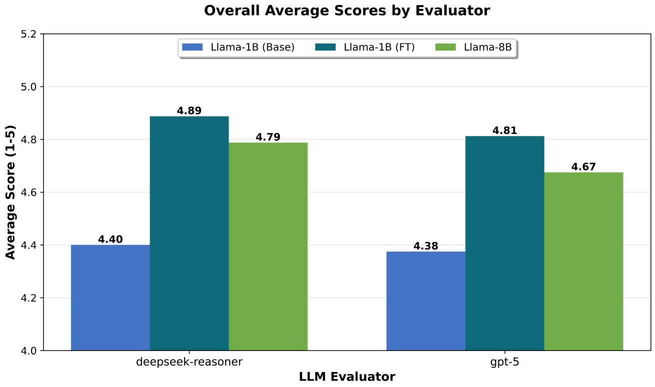

Fig. 1. Overall Average Scores by Evaluator.

C. Performance Across CEFR Levels

When analyzing performance across proficiency levels (Fig. 2), the robustness of the fine-tuned model becomes evident. At the A1 level, Llama-1B FT achieved a near-perfect score of **4.96**, significantly higher than the base model's 4.36. While all models experienced a slight dip at the A2 level due to increased linguistic complexity, the fine-tuned model maintained its lead with a score of **4.74**.

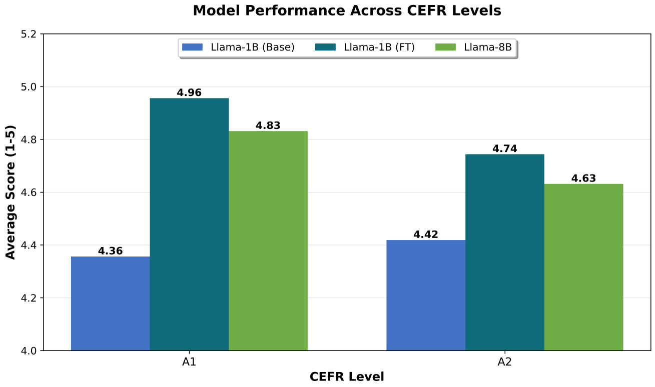

Fig. 2. Model Performance Across CEFR Levels.

D. Linguistic Competence Profile

The radar chart in Fig. 3 provides a multidimensional view of the models' linguistic capabilities. The fine-tuned model shows a balanced and expanded profile compared to the base model.

The most significant gain was observed in **Word Usage** and **Naturalness**. The base model often struggled to place words in contextually appropriate scenarios for beginners. Fine-tuning corrected this, allowing the 1B model to match the semantic precision of the 8B model. For instance, at the A1 level, although the word but was included in the target list, none of the ten sentences generated by the base model contained it. The fine-tuned model, however, correctly used the word in a sentence such as: "I am tired, but I will finish my homework." The fine-tuned model did not exhibit this type of error in any of its generated sentences. This result validates our hypothesis that high-quality synthetic data can effectively bridge the gap between small and large language models for educational content generation.

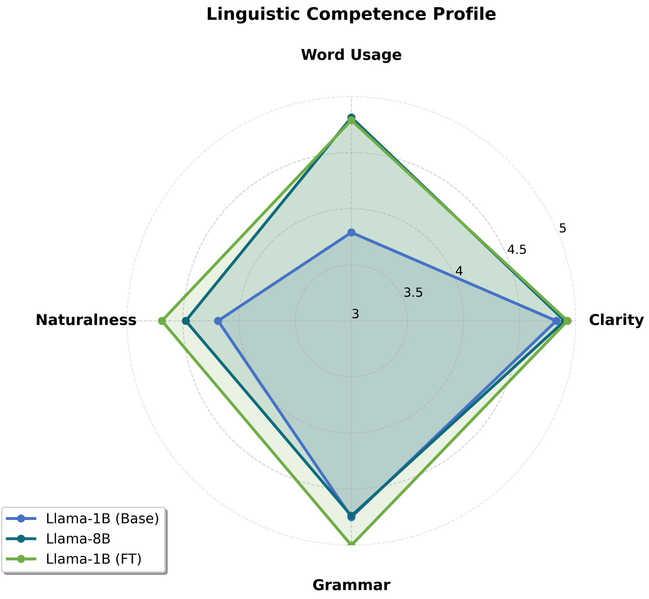

Fig. 3. Linguistic Competence Profile (Radar Chart).

IV. Conclusion

This paper explored the possibility of using small-scale open-source LLM to create educational material of the correct level to English language learners. Using a parameter-efficient fine-tuning method using a synthetic dataset created by GPT-4o-mini, we demonstrated that large model size is not a direct requirement for achieving strong performance in specialized domains. In our experiments, we have found that the fine-tuned Llama- 3.2-1B model performed significantly better than the base version in all of the evaluation metrics with a perplexity drop of almost 80%. More importantly, this smaller 1-billion-parameter model achieved competitive performance relative to the larger, general-purpose Llama-3.1-8B model in sentence generation at the A1 and A2 levels. These findings are significant for the development of educational technologies. We also showed that effective, privacy-preserving, and offline-capable language learning tools can be built on modest hardware without relying on costly external APIs. These models can provide a scalable platform of offering personalized vocabulary practicing to learners across the globe. Future research could extend the scope of this work to higher CEFR levels, allowing the models to be evaluated across a broader range of linguistic complexity.

References

[1] G. Brown, "Understanding spoken language," *TESOL Quarterly*, vol. 12, no. 3, pp. 271–283, 1978.

[2] P. S. Rao, "The role of English as a global language," *Research Journal of English*, vol. 4, no. 1, pp. 65–79, 2019.

[3] N. Lutfiyah, N. Nuraeningsih, and R. Rusiana, "The obstacles in learning vocabulary of EFL students," *Prominent*, vol. 5, no. 2, pp. 114–125, 2022.

[4] U. Y. Elmurodov and S. S. Shorakhmetov, "Essential ways to learn English words," *Academic Research in Educational Sciences*, vol. 2, no. 5, pp. 577–583, 2021.

[5] S. Wang *et al.*, "Large language models for education: A survey and outlook," *arXiv preprint arXiv:2403.18105*, 2024, doi: 10.48550/arXiv.2403.18105.

[6] M. Stahl, L. Biermann, A. Nehring, and H. Wachsmuth, "Exploring LLM prompting strategies for joint essay scoring and feedback generation," in *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, 2024, pp. 283–298.

[7] Y. Wu, "Exploring the influence of large language models (LLMs) on English learners and their teachers," *Journal of Education, Humanities and Social Sciences*, vol. 27, pp. 530–535, 2024, doi: 10.54097/zghke663.

[8] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023, doi: 10.1016/j.lindif.2023.102274.

[9] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?," *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, p. 39, 2019, doi: 10.1186/s41239-019-0171-0.

[10] J. Maharjan *et al.*, "OpenMedLM: Prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models," *Scientific Reports*, vol. 14, no. 1, p. 14156, 2024, doi: 10.1038/s41598-024-64827-6.

[11] W. S. Mathis, S. Zhao, N. Pratt, J. Weleff, and S. De Paoli, "Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods?," *Computer Methods and Programs in Biomedicine*, vol. 255, p. 108356, 2024, doi: 10.1016/j.cmpb.2024.108356.

[12] S. Sandmann, S. Riepenhausen, L. Plagwitz, and J. Varghese, "Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks," *Nature Communications*, vol. 15, no. 1, p. 2050, 2024, doi: 10.1038/s41467-024-46411-8.

[13] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2022, doi: 10.48550/arXiv.2106.09685.

[14] A. Dubey *et al.*, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[15] OpenAI, "GPT-5 technical report," OpenAI Publications, 2025.

[16] DeepSeek AI, "DeepSeek-Reasoner: Model card and technical overview," DeepSeek Research, 2025.