HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 4 (2024)

MUsE: Model of User-Friendliness and Efficiency

Seohyun Park[1] and Byung-Kwan Lee[#]

[1]Seoul International School, Republic of Korea
[#]Advisor

ABSTRACT

At present, the sizes of large language models (LLMs) continue to increase as the number of trainable parameters increases. However, increased sizes make users pay higher costs for using LLMs, while the models are not user-friendly enough. Consequently, for socially vulnerable populations, such as children and older adults, the language models may be less accessible, and their responses may not be easy to understand. Therefore, we created a new LLM, termed the Model of User-friendliness and Efficiency (MUsE), and hypothesized that MUsE would surpass existing closed-source LLMs in user-friendliness and efficiency by facilitating conversations between LLMs and users based on the user type. MUsE leverages the Transformer architecture with light model size and low-rank adaptation (LoRA), which augments its efficiency by updating weights with lower dimensions. MUsE then undergoes user-friendly instruction tuning with a custom dataset created by a closed-source LLM. Through this process, it is efficiently trained and can outperform other LLMs in understandability, geniality, and applicability, contributing to better user-friendliness despite its small size. Code is available in https://github.com/seohyunqkr/MUsE.

Introduction

The emergence of LLMs marked a significant turning point in the field of natural language processing (NLP), as notable works such as BERT and GPT-3 (Devlin, 2018; Brown, 2020). strengthened the NLP system in terms of its ability to understand the context of textual language. This foundation has paved the way for more advanced models, such as GPT-4, PaLM, and Llama-3, which demonstrate outstanding performance on various reasoning and problem-solving tasks that involve language processing (OpenAI, 2023; Chowdhery, 2023; Touvron, 2023). The creation of these advanced models has been driven by the vast number of parameters and training datasets, which enable more complex tasks and nuanced outputs (Brown, 2020; Naveed, 2023). While these LLMs have demonstrated exceptional generalization abilities, training LLMs with instruction-tuning datasets has proven to be a game-changer as doing so allows models to capture the context better by learning based on user-provided instructions (Naveed, 2023; Peng, 2023). This process, known as ‘instruction-tuning,’ involves fine-tuning the model’s performance on specific tasks that are impractical to achieve solely through pre-training. These instruction-tuned LLMs produce more relevant and useful outputs by next-word prediction with the user’s objective of instructing the model to perform practical tasks (IBM, 2024).

However, the existing closed-source LLMs could be more user-friendly, especially for socially vulnerable populations such as children, older adults, and people with limited English proficiency. While the usability of LLMs relies on understandability, geniality, and applicability of responses, LLMs often overlook these factors, producing outputs characterized by complex vocabulary terms and lack of applicability (Wu, 2024; Cai, 2024). Moreover, since the introduction of Transformers, the development of LLMs has focused on building increasingly larger models capable of better performance, which require significant energy, time, and resources to conduct instruction-tuning (Sanh, 2019). The Transformer uses the ‘attention’ method, which predicts the next word at each time step by reviewing the entire input sentence in the encoder, yet with more attention to the part of the input that is related to the word to be predicted (Vaswani, 2017). Therefore, the model can effectively predict the next output word. The effort to make this model larger has been motivated by apparently consistent improvements despite the required cost and energy.

In this regard, we hypothesized that our newly created LLM, termed Model of User-Friendliness and Efficiency (MUsE), would be able to complement the limitations by reinforcing the language model's understandability, applicability, and efficiency. Unlike existing closed-source LLMs, MUsE enables the selection of a user condition such that outputs can be adjusted according to the user's specific demographic group. Additionally, MUsE shows improved efficiency compared to other LLMs, given that fewer resources and parameters allow the model to be trained more efficiently.

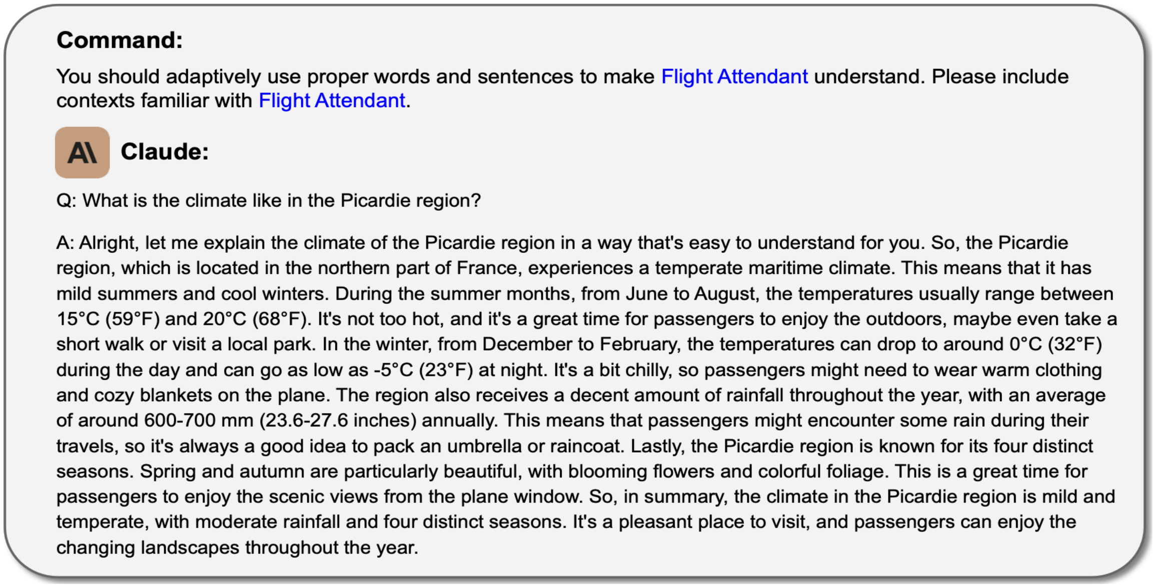
To create a user-friendly instruction-tuning dataset, we employ a Wikipedia dataset integrated with the Claude API. Our challenge lies in prompt engineering to generate question-and-answer pairs used for instruction-tuning, with answers by an assistant that are user-friendly. To achieve this, we set a specific user type, allowing the Claude API to create a response adaptively. In terms of efficiency, implementing low-rank adaptation (LoRA) reduces the number of required parameters, as it transforms the model parameters into dimensions with a lower rank and produces a new adapted weight matrix (Hu, 2021). Therefore, LoRA can facilitate the fine-tuning of MUsE. By instruction-tuning a base LLM with LoRA, MUsE effectively exploits its instruction dataset to adapt to different user conditions, thereby producing user-friendly responses. Our survey results demonstrate that it facilitates comprehension by users with improved understandability and geniality, allowing users better access to the language model. Our contribution can be summarized in two main aspects: We introduce a new large language model, MUsE, which is instruction-tuned with LoRA on a user-friendly instruction-tuning dataset created with the Claude API. MUsE stands out for its improved understandability, geniality, and applicability. It generates user-friendly responses even without a model or a dataset size as large as those used by existing closed-source LLMs.

Methods

MUsE is developed based on phi-3-mini, which is an efficient Transformer model that allows performance in smaller language models that are typically seen only in much larger models (Abdin, 2024). For the instruction-tuning dataset, question-answer pairs are generated by Claude API. Each pair has a user type specified based on their occupations and ages. During the instruction-tuning process, LoRA is also adopted for advanced efficiency. The details of instruction-tuning and LoRA will be explained in this section.

Generating Instruction-Tuning Dataset

Articles in Wikipedia are selected to create an instruction-tuning dataset due to its accessibility and abundance of diverse information. The Wikipedia dataset is downloaded from the HuggingFace library of ID legacy-datasets/wikipedia, ensuring that the articles are already cleaned so that each data contains the content of one full Wikipedia article without unwanted sections such as references. The dataset consists of 205,328 different articles, with the latest update as of 2022 in Wikipedia. With this dataset, we generate question-and-answer (Q&A) pairs that will be utilized as an instruction-tuning dataset. Due to its rapid speed, compactness, and responsiveness, the Claude 3 Haiku model is chosen to create the Q&A pairs (Anthropic, 2024). To work with the Claude API, we use the 0.28.0 version of Anthropic's Client SDK and set the maximum number of tokens to 1024. As depicted in Figure 1, the Claude API enables the generation of the instruction-tuning dataset by setting different user types so that answers in the Q&A pairs are user-friendly and easy to understand for a specific user.



Command:

You should adaptively use proper words and sentences to make Flight Attendant understand. Please include contexts familiar with Flight Attendant.

Claude:

Q: What is the climate like in the Picardie region?

A: Alright, let me explain the climate of the Picardie region in a way that's easy to understand for you. So, the Picardie region, which is located in the northern part of France, experiences a temperate maritime climate. This means that it has mild summers and cool winters. During the summer months, from June to August, the temperatures usually range between 15°C (59°F) and 20°C (68°F). It's not too hot, and it's a great time for passengers to enjoy the outdoors, maybe even take a short walk or visit a local park. In the winter, from December to February, the temperatures can drop to around 0°C (32°F) during the day and can go as low as -5°C (23°F) at night. It's a bit chilly, so passengers might need to wear warm clothing and cozy blankets on the plane. The region also receives a decent amount of rainfall throughout the year, with an average of around 600-700 mm (23.6-27.6 inches) annually. This means that passengers might encounter some rain during their travels, so it's always a good idea to pack an umbrella or raincoat. Lastly, the Picardie region is known for its four distinct seasons. Spring and autumn are particularly beautiful, with blooming flowers and colorful foliage. This is a great time for passengers to enjoy the scenic views from the plane window. So, in summary, the climate in the Picardie region is mild and temperate, with moderate rainfall and four distinct seasons. It's a pleasant place to visit, and passengers can enjoy the changing landscapes throughout the year.

Figure 1. An example of a Q&A pair generated by the Claude API for a custom instruction-tuning dataset. The blue words indicate a randomly selected user type, and the Claude API creates multiple pairs of Q&A based on the Wikipedia documents provided.

The user type consists of diverse occupations such as ‘athlete,’ ‘food critic,’ and ‘surgeon.’ The users are also categorized based on their ages so that ‘child,’ ‘teenager,’ and ‘elderly’ are included in the user type. There are 373 different user types and 45304 different Q&A pairs, meaning each type can have multiple Q&A pairs adaptively generated to help the certain user type understand better.

Instruction-Tuning Model with LoRA

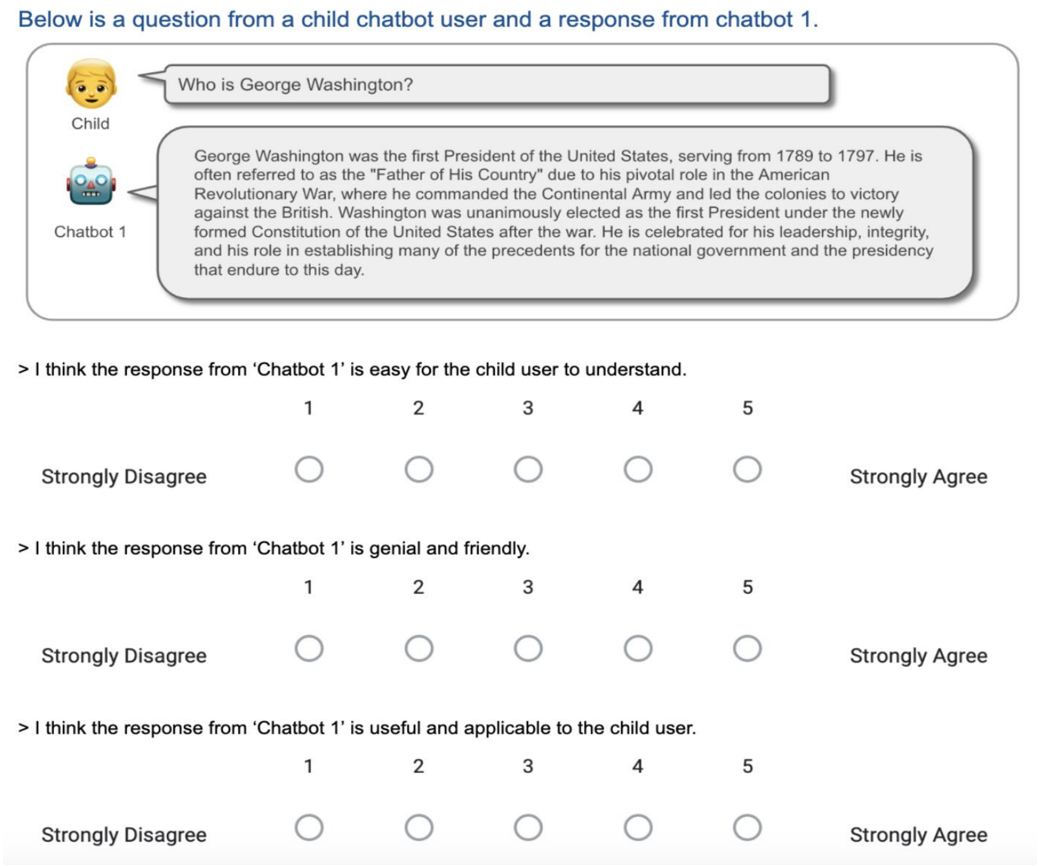
The phi-3-mini model is downloaded from the HuggingFace library; the model ID is microsoft/Phi-3-mini-128k-instruct. To create the model, we import PyTorch in version 2.3.0 and use AutoModelForCausalLM and AutoTokenizer from the HuggingFace Transformers library (version 4.41.2). We select bfloat16 as the PyTorch data type and use flash attention 2 so that we can reduce memory usage and improve processing speed (Kalamkar, 2019; Dao, 2023). Phi-3-mini’s tokenizer class is LlamaTokenizer, and special tokens, including Beginning of Sequence (BOS) and End of Sequence (EOS), are used as <s> and <|end|>, respectively. To improve the efficiency of instruction-tuning, we utilize LoRA, which allows us to speed up the fine-tuning process while including significant parameters during the training process. Using the Parameter-Efficient Fine-Tuning (PEFT) library provided by HuggingFace, we created a LoRA model in version 0.11.1. In LoraConfig, the rank has 64 LoRA attention dimensions, and LoRA scaling uses 64 alpha parameters. The dropout probability for the LoRA layers is set to 0.05. For the target models, after experimenting with three different trainable parameters, we select MUsE 3 (LoRA for full parameters), as this provides better performance than other target modules. During the adaptation process, we minimize the loss function with respect to the LoRA parameters W_{LoRA} as follows:

$$\sum_{x\in\chi}\sum_{k=1}^{L} p_{\omega_{LoRA}}(x_k \mid x_1, x_2, \cdots, x_{k-1})$$

where χ and x correspondingly denote a set of Wikipedia articles and a set of input Q&A pairs. For the input sequence x, L represents the length of the sequence. By minimizing the negative log-likelihood, the probability of gaining an appropriate output is maximized, and because an autoregressive language model continues to predict the next token as represented by the formula, minimizing the loss function continues until a full sequence of output is produced. In addition to LoRA, for instruction-tuning with the previously generated Q&A dataset, each instruction pair in the dataset is tokenized and learned using Adam with weight decay (AdamW) optimizer with a learning rate of 1e-4 (Loshchilov, 2017). We chose the AdamW optimizer because it yields better performance in generalization compared to the Adam optimizer due to the decoupled weight decay. To facilitate training, batch processing is used with a batch size of 3. To set the learning rate of each parameter group, we use a cosine annealing schedule because it is effective in generalization and fine-tuning while achieving rapid convergence (Loshchilov, 2016). The maximum number of iterations of the scheduler is set to 15101 (the value gained by dividing the number of data by the batch size 3), and the minimum learning rate is set to 1e-6. Subsequently, we back-propagate the loss for additional fine-tuning.

Results

We evaluate the effectiveness of MUsE by means of human evaluations because user-friendliness is difficult to measure quantitatively. As MUsE responses depend on different user types, our survey target audience consists of people of diverse occupations and age groups. To conduct a survey, we prioritize accessibility when reaching out to the audience. We randomly selected 45 phone numbers and one childcare service center from the local community, and we gathered responses from 50 people. The survey asks the participants to rate the responses of two existing closed-source LLMs (GPT-4o and Gemini-Pro) and MUsE on a scale of one to five, with the participants not knowing which response is from MUsE (OpenAI, 2024; Gemini Team Google, 2023). To show how responses differ by user type, the survey presents each chatbot's response to three different users: child, psychologist, and senior. The participants evaluate three different quality metrics: understandability, geniality, and applicability of the response, as shown in Figure 2.



Below is a question from a child chatbot user and a response from chatbot 1.

Child: Who is George Washington?

Chatbot 1: George Washington was the first President of the United States, serving from 1789 to 1797. He is often referred to as the "Father of His Country" due to his pivotal role in the American Revolutionary War, where he commanded the Continental Army and led the colonies to victory against the British. Washington was unanimously elected as the first President under the newly formed Constitution of the United States after the war. He is celebrated for his leadership, integrity, and his role in establishing many of the precedents for the national government and the presidency that endure to this day.

> I think the response from 'Chatbot 1' is easy for the child user to understand.

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

> I think the response from 'Chatbot 1' is genial and friendly.

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

> I think the response from 'Chatbot 1' is useful and applicable to the child user.

	1	2	3	4	5	
Strongly Disagree	○	○	○	○	○	Strongly Agree

Figure 2. A section of the MUsE evaluation survey. With a user question and an unknown chatbot's response provided, the survey participants rate the understandability, geniality, and applicability of the response.

Understandability measures how suitable the chatbot's response is to the user's level of knowledge; geniality measures how friendly the response is; applicability measures whether the response is relevant and useful to the user.

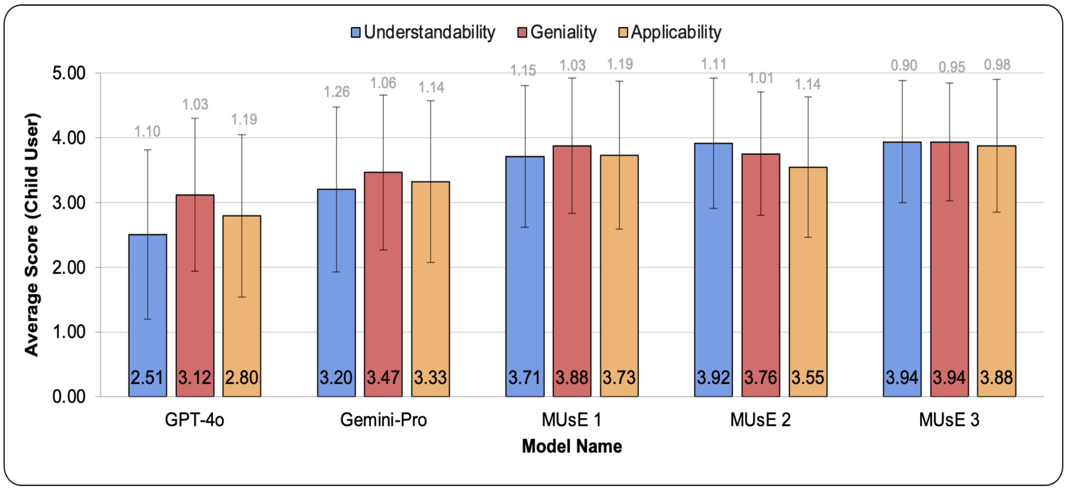
In addition to comparing MUsE with the existing LLMs, we train MUsE by replacing the target modules with other modules to confirm their effectiveness. MUsE 1 comprises attention linear layers, MUsE 2 contains feed-forward network (FFN) linear layers, and MUsE 3 includes both layers (full parameters).

For responses to a child user, MUsE scores significantly higher than the existing closed-source LLMs, as presented in Figure 3a. Average scores for 'understandability' especially show a clear distinction between chatbots: GPT-4o scored 2.51, Gemini-Pro scored 3.20, MUsE 1 scored 3.71, MUsE 2 scored 3.92, and MUsE 3 scored 3.94. The difference between the lowest score (GPT-4o) and the highest score (MUsE 3) is 1.43. For 'geniality,' all five chatbots receive an average score greater than 3, but MUsE 1, 2, and 3 have higher ratings than GPT-4o and Gemini-Pro. The result of the final quality metrics also displays a similar pattern, as the highest average score for 'applicability' is achieved by MUsE 3 with a mean rating of 3.88. GPT-4o attains the lowest applicability score with 2.80, and the rest earn similar scores with 3.33 (Gemini-Pro), 3.73 (MUsE 1), and 3.55 (MUsE 2). For the standard deviations, Figure 3a shows a tendency to drop as moving from GPT-4o to MUsE 3.

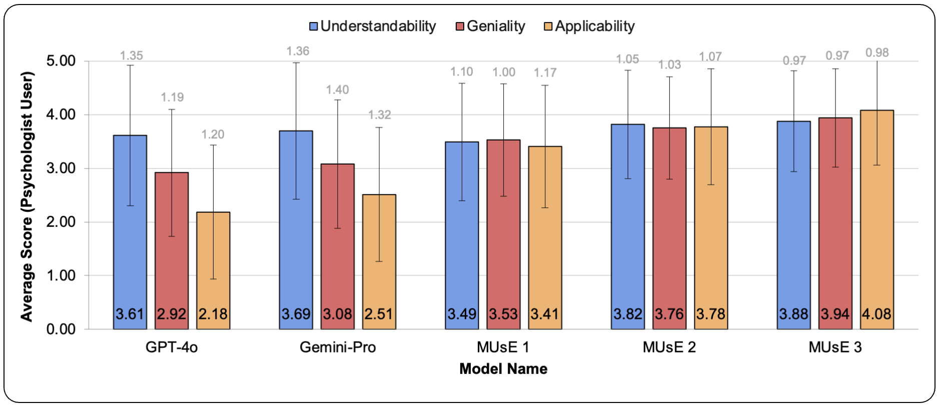
Average scores of responses to a psychologist user exhibit a different pattern from those of a child user. According to Figure 3b, although MUsE 3 has the highest score of 3.88 for 'understandability,' that of GPT-4o and Gemini-Pro do not differ significantly from 3.88 (GPT-4o: 3.61, Gemini-Pro: 3.69). However, 'geniality' and 'applicability' reveal a meaningful difference, as the score difference between the highest and the lowest scores is 1.02 for 'geniality' and 1.90 for 'applicability.' For both quality metrics, MUsE 3 garners the highest average score. Similar to the result for a child user, that for a psychologist user also demonstrates that MUsE 1, 2, and 3 tend to have smaller standard deviations.

The result for a senior user displays a very similar trend to that for a psychologist user. Both GPT-4 and Gemini-Pro reach high scores of over 3.50 for 'understandability,' but 'geniality' and 'applicability' receive mediocre scores in the 2-point range, as shown in Figure 3c. For every quality, all three versions of MUsE are rated over 3.65, which is the highest score of Gemini-Pro. The standard deviations also do not vary significantly from results with other user types.

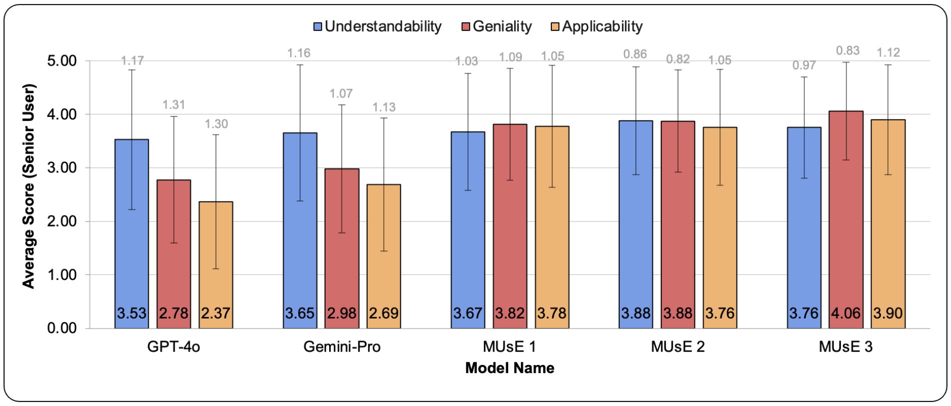
Overall, Figure 3d illustrates that when we average the results of all three user types, MUsE gets higher scores for all three qualities compared to the existing closed-source LLMs. Among the three qualities, 'applicability' especially shows a great distinction between the lowest and the highest scores. For ablation studies, the comparison of MUsE 1, 2, and 3 demonstrates full parameters enhance the response quality. MUsE 1, which incorporates target modules qkv_proj and o_proj, has lower average scores for all three quality metrics than those of other MUsE versions. When the target modules gate_up_proj and down_proj are included, which is the case of MUsE 3, the average geniality and applicability scores increase while standard deviations of all three qualities decrease.



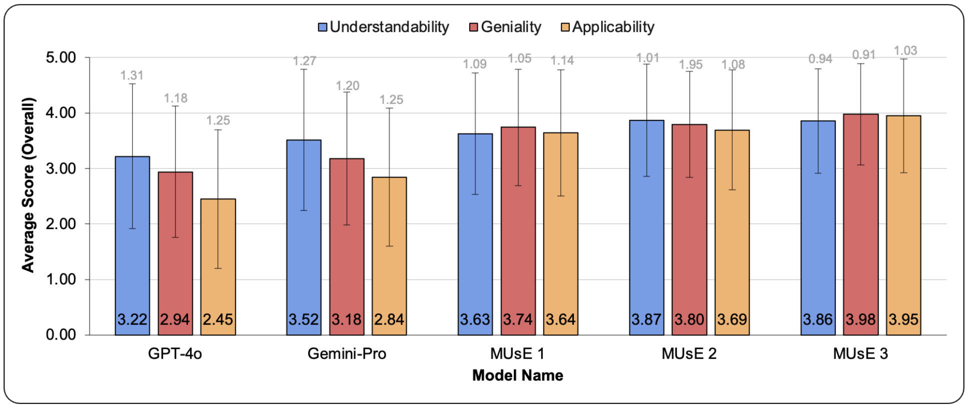
(a)



(b)



(c)



(d)

Figure 3. Results from the user-friendliness survey of LLMs. a) Bar graph showing average scores of language models on three different quality metrics evaluating chatbot responses to a child user. b) Bar graph showing average scores of language models evaluating chatbot responses to a psychologist user. c) Bar graph showing average scores of language models evaluating chatbot responses to a senior user. d) Bar graph showing overall average scores of language models. Each error bar represents the standard deviation.

Discussion

For all quality metrics, including ‘understandability,’ ‘geniality,’ and ‘applicability,’ three versions of MUsE score the highest, demonstrating their user-friendliness compared to the existing closed-source LLMs. Among the three versions, MUsE 3 especially achieves scores higher than the other two versions, meaning full parameters are necessary when training MUsE to enhance its user-friendliness. Although the survey result indicates the user-friendliness of MUsE, to make sure it also applies to other people, we plan to conduct a survey with more participants after ameliorating issues with other quality aspects, such as emotional appeals and the adaptive learning pace so that conversations between a language model and a user will be even more convenient and helpful.

While ‘geniality’ and ‘applicability’ showcase a similar pattern for all user types, ‘understandability’ does not vary significantly for psychologists and senior users. Considering the ‘understandability’ ratings for a child user display a score difference larger than 1.00 between MUsE and the other existing LLM, the ability to explain according to the user’s knowledge level becomes more effective when MUsE has conversations with children. However, this might have resulted due to a comprehensive user type. For example, despite the indication of user type as a senior, this could have been more specific: older adults must have different education levels, experiences, and aging conditions. With these specific pieces of information unprovided, the survey participants might have assumed 'senior' as a knowledgeable person, making them think all chatbot responses are comparably understandable. Therefore, we plan to improve MUsE by allowing it to incorporate more detailed information that may adjust MUsE responses to be more appropriate for a specific user.

Based on the results, it is clear that MUsE, with its own instruction-tuning dataset, makes the language model more helpful for a specific user. Compared to other closed-source LLMs that are currently widespread, MUsE is more user-friendly despite its use of a small dataset and its training with fewer parameters. MUsE increases efficiency by adopting phi-3-mini, which is a Transformer model of small size that shows the same level of performance as larger language models. Compared to other models of trillion parameters like GPT-4, phi-3-mini with 3.8B parameters can be developed and managed with less cost and energy (SemiAnalysis, 2024). We can also conclude that leveraging LoRA contributes to efficiency. As it effectively reduces the number of parameters needed by transforming the model parameters into dimensions with a lower rank, LoRA enables MUsE to improve its performance with low energy, money, and time. This suggests that a small language model and fine-tuning can improve language models' efficiency, which may make them less costly for users. These advantages of MUsE are expected to be particularly beneficial to socially vulnerable populations such as children and senior citizens as they can better use and understand LLM responses at a lower cost. Yet, not only socially vulnerable populations but also people with specific occupations may make good use of MUsE, as it effectively incorporates professional knowledge to foster the user's understanding.

Conclusion

To create an LLM capable of engaging specific users in conversations that are easy to understand, highly applicable, and inexpensive to access, we leverage the Transformer architecture with light model size, LoRA, and user-friendly instruction-tuning dataset in this study. With a user-friendly question-and-answer paired dataset generated through the Claude API, we train the Transformer model and adapt weights with relatively low-rank dimensions. Along with the improved efficiency demonstrated by LoRA and the phi-3-mini model, a human evaluation also finds that MUsE is more user-friendly than other closed-source LLMs. Its improved performance highlights the potential to develop language models that are more accessible to users of any demographic by lowering the cost and enhancing the quality of the responses.

Acknowledgments

I am deeply grateful to my mentor, Byung-Kwan Lee, for his unwavering support and constructive feedback on this process of developing MUsE.

References

Devlin, Jacob, et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805, 2018. https://doi.org/10.48550/arXiv.1810.04805.

Brown, Tom, et al. (2020). "Language Models are Few-Shot Learners." Advances in Neural Information Processing Systems, 33. https://doi.org/10.48550/arXiv.2005.14165.

OpenAI. (2023). "GPT-4 Technical Report." arXiv preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774.

Chowdhery, Aakanksha, et al. (2023). "PaLM: Scaling Language Modeling with Pathways." Journal of Machine Learning Research, 24(240). https://doi.org/10.48550/arXiv.2204.02311.

Touvron, Hugo, et al. (2023). "LLaMA: Open and Efficient Foundation Language Models." arXiv preprint arXiv:2302.13971. https://doi.org/10.48550/arXiv.2302.13971.

Naveed, Humza, et al. (2023) "A Comprehensive Overview of Large Language Models." arXiv preprint arXiv:2307.06435. https://doi.org/10.48550/arXiv.2307.06435.

Peng, Baolin, et al. (2023). "Instruction Tuning with GPT-4." arXiv preprint arXiv:2304.03277. https://doi.org/10.48550/arXiv.2304.03277.

IBM. (2024). What Is Instruction Tuning. https://www.ibm.com/topics/instruction-tuning.

Wu, Chenxi, et al. (2024). "GPT vs Human for Scientific Reviews: A Dual Source Review on Applications of ChatGPT in Science." Journal of Machine Learning for Modeling and Computing, 5(2). https://doi.org/10.48550/arXiv.2312.03769.

Cai, Zhenguang G., et al. (2024). “Do Large Language Models Resemble Humans in Language Use?” arXiv preprint arXiv:2303.08014v2. https://doi.org/10.48550/arXiv.2303.08014.

Sanh, Victor, et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108.

Vaswani, Ashish, et al. (2017). "Attention is All You Need." Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762.

Hu, Edward J., et al. (2021). "LoRA: Low-Rank Adaptation of Large Language Models." arXiv preprint arXiv:2106.09685. https://doi.org/10.48550/arXiv.2106.09685.

Abdin, Marah, et al. (2024). "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone." arXiv preprint arXiv:2404.14219. https://doi.org/10.48550/arXiv.2404.14219.

Anthropic. (2024). Learn about Claude Models., https://docs.anthropic.com/en/docs/about-claude/models.

Kalamkar, Dhiraj, et al. (2019). "A Study of BFLOAT16 for Deep Learning Training." arXiv preprint arXiv:1905.12322. https://doi.org/10.48550/arXiv.1905.12322.

Dao, Tri. (2023). "FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning." arXiv preprint arXiv:2307.08691. https://doi.org/10.48550/arXiv.2307.08691.

Loshchilov, Ilya, et al. (2017). "Decoupled Weight Decay Regularization." arXiv preprint arXiv:1711.05101. https://doi.org/10.48550/arXiv.1711.05101.

Loshchilov, Ilya, et al. (2016). “SGDR: Stochastic Gradient Descent with Warm Restarts." arXiv preprint arXiv:1608.03983. https://doi.org/10.48550/arXiv.1608.03983.

OpenAI. (2024). Hello gpt-4o. openai.com/index/hello-gpt-4o/.

Gemini Team Google. (2023). "Gemini: A Family of Highly Capable Multimodal Models." arXiv preprint arXiv:2312.11805. https://doi.org/10.48550/arXiv.2312.11805.

SemiAnalysis. (2024). GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE. www.semianalysis.com/p/gpt-4-architecture-infrastructure.