Enhancing LLM Conversational Acuity Using Pragmatic Measures

Andrew Han
Westlake High School
Austin, Texas
andrewhanlt@gmail.com

Nishanth Koushik
Prospect High School
San Jose, California
nkoushik11@gmail.com

Nachiket Bidarkundi
Foothill High School
Pleasanton, California
nachobid08@gmail.com

Michael Naeim Shehata
Miami College High School
Alexandria, Egypt
michaelnaeim01@gmail.com

Vibha Kunchipudi
National Public School HSR
Bengaluru, India
vibessk198@gmail.com

Tunar Mammadli
School No. 42
Baku, Azerbaijan
tunardev@gmail.com

Shourya Mehta
Adlai E. Stevenson High School
Lincolnshire, Illinois
shourya0mehta@gmail.com

Osher Lerner
University of California - Berkeley
Berkeley, California
oshlern@berkeley.edu

***Abstract*—AI-driven communication has the potential to transform society with superhuman capabilities such as real-time multilingual translation, predictive text generation, and personalized content creation. However, current Large Language Models (LLMs) like OpenAI's GPT-3.5 Turbo and 4o-mini often struggle to capture the nuanced writing styles, tones, and behavioral characteristics of individual users. While fine-tuning is a common approach, existing techniques focus primarily on task-specific performance and tend to neglect the integration of personal conversational elements and non-verbal cues. As a result, these methods often fail to preserve the unique conversational tone and vocabulary of individual users. Our work addresses these limitations by analyzing similarity, linguistic, and psychological metrics to fine-tune OpenAI's GPT-3.5 Turbo and 4o mini, improving contextual accuracy, and generating responses that better align with conversational tone and user-specific vocabulary.**

***Index Terms*—Large Language Models, Personalized Content Generation, Generative AI, Human-Computer Interaction**

I. Introduction

In the last several years, Large Language Models (LLMs) have shown extraordinary enhancements in natural language processing, including natural language translation, content generation, and conversational systems [1]. This limitation is particularly evident in key areas such as information retrieval accuracy, linguistic coherence (in line with Grice's maxims), and the overall quality of AI-human interaction [2]. It limits the effectiveness of LLMs in personal contexts, especially when capturing the user's tone, choice of specific words, and the natural flow of a conversation is important to capture.

There is a difficult lack of human-like interaction behavior in these models, such as non-verbal actions and personality profiling. Perhaps one of the most sought-after models currently in the market including; GPT-3.5 Turbo, GPT-4, GPT-4o, LLaMA-2, and Mistral-2 often fail to simulate human personality traits accurately when scored on the IPIP-NEO-120 test, which is designed to measure human-like behavior and interaction [4]. These deficiencies extend to Natural Language Processing (NLP) tasks, where achieving a deep and accurate understanding of text meaning continues to be a significant challenge.

In response to the growing demand for personalized and contextually aware AI, we aim to develop a model that more accurately understands and emulates the unique communication styles of individual users. Unlike existing models, which often lack nuanced personalization, our LLM will analyze and adapt to the characteristics of a user's writing, including vocabulary range, sentence structure, and tone. Furthermore, it will incorporate a psychological framework—the Big Five personality traits (openness to experience, conscientiousness, agreeableness, neuroticism, and extraversion)—to identify and mirror the user's personality traits [4].

Our approach begins with constructing a diverse, real-world dataset of user-generated content, designed to capture a wide range of linguistic behaviors and stylistic nuances. This dataset contains detailed and authentic user inputs to analyze and identify recurring linguistic patterns and infer personality traits based on established psychological frameworks.

To address these challenges, we propose a fine-tuning approach that enhances LLMs by integrating user-specific linguistic and psychological traits. By fine-tuning OpenAI's GPT-3.5 turbo and GPT-4o mini models, we strengthen its ability to better capture and emulate distinctive user-specific linguistic features with greater precision. This focused fine-tuning improves the contextual relevance of the generated responses and ensures they align with the user's unique linguistic style. The outcome represents a significant advance in the personalization of language models, with broad applications in voice assistants and content-generation tools.

The remainder of this paper is structured as follows: Section II discusses relevant literature on recent advancements in LLMs, behavioral psychology, and chat interfaces. Section III outlines our methodology, including data collection, fine-tuning processes, and quantitative evaluations. In Section IV, we present a comprehensive analysis of the improvements achieved. Section V concludes the paper by discussing our findings and their implications. Finally, Section VI outlines potential directions for future research.

II. Related Work

In this section, we review the strengths and limitations of LLMs, the difficulties in capturing nuanced writing styles, and recent progress in fine-tuning techniques aimed at enhancing user-specific outputs.

A. Strengths and Limitations of Large Language Models

Large language models (LLMs), such as GPT-3.5 Turbo and GPT-4o, are distinguished by their extensive training with various datasets and their sophisticated structures, which together give them the ability to understand and produce human-like data with incredible precision and fluency. However, their size and complexity present significant challenges, including computational costs, limitations in understanding nuanced writing styles, and the need for fine-tuning to achieve personalized outputs.

1) OpenAI GPT Models: GPT-3.5 Turbo [12], developed by OpenAI, is known for its consistency and good contextual awareness, as the model is trained on datasets such as Common Crawl and WebText. It performs well in broad applications, such as creative writing and technical documentation, but struggles with domain-specific knowledge and linguistic nuances. Its large size (175 billion parameters) makes real-time use expensive and limits deployment in resource-constrained environments.

GPT-4o [13], also developed by OpenAI, incorporates overparameterization to enhance context and, especially, abstract thinking. The current model's claimed strengths are higher performance on complex tasks yet it has limitations and issues with mimicking the style of certain writers and in managing cultural or emotional contexts.

B. Advancements in Fine-Tuning LLMs for Personalized Outputs

The landscape of AI communication has been significantly influenced by chatbot systems such as Siri, Google Now, and Alexa, which have made people think of AI as a "dialogue system often endowed with human-like behavior" [2]. These systems have motivated the development of AI that can engage in more human-like and contextually aware conversations. After the introduction of ChatGPT, there was a major shift in AI communication, proving that AI now generates original content, rather than relying solely on pre-existing sources [3]. While chatbots and artificial assistants use user interactions via text or speech, their capabilities differ significantly [14]. Chatbots can provide information and execute specific tasks, while artificial assistants leverage complex algorithms in Natural Language Processing (NLP) and Machine Learning (ML) to continuously learn and improve.

To mimic another's behavior, our AI must adjust its communication style while maintaining "social distance." This involves strategies such as convergence, where communicators become more similar, or divergence, where differences are highlighted. Brown and Levinson's politeness theory shows how communication involves the desire for approval (positive face) and autonomy (negative face) [5]. LLMs will improve their conversational abilities and mimic users more effectively by tracking these communication styles.

In addition to these foundational concepts, the integration of pragmatics in AI communication has become a focal point of research. Pragmatics, the study of how context influences the interpretation of meaning in communication, is critical for developing AI that can understand and generate human-like conversation. This involves understanding conversational implicature, where speakers imply and listeners infer meanings that are not explicitly stated. Grice's Maxims, which include principles like quantity, quality, relation, and manner, provide a framework for evaluating the conversational effectiveness of AI. Recent advancements in LLMs, such as GPT-4, have shown that these models can surpass human performance in certain pragmatics tasks, indicating the potential for further development in this area [6]. However, such models still fail to converse well with users, as they often output responses that are oversaturated with information or lack situational understanding.

Incorporating personality traits into LLMs enhances conversational mimicry and user interaction. The Big Five personality traits—conscientiousness, agreeableness, neuroticism, openness to experience, and extraversion—provide a framework for understanding and modeling human behavior. Studies illustrate that LLMs can simulate these traits, making interactions more engaging and personalized [8]. Techniques like zero-shot prompting allow LLMs to exhibit consistent personality traits across interactions, improving their ability to mimic human conversational styles [8].

By integrating principles from psychology, linguistics, and AI research, we can create models capable of understanding and predicting user needs and responses, ultimately enriching the conversational experience.

III. Methodology

In this section, we outline the procedures and techniques employed in our experimentation with Large Language Model (LLM) fine-tuning. This section covers information about our quantitative and qualitative metrics, dataset, experimental environment, and research process.

A. Quantitative Metrics

Tokenization refers to segmenting text into discrete units, such as words or phrases, which can subsequently be analyzed for similarities. Using tokenization similarity to evaluate conversational AI involves assessing how closely the tokens in the AI's response match those in the user's input or context. This approach is useful in gauging the AI's comprehension and coherence. For instance, when an LLM restates part of a user's original prompt in its response, it demonstrates a concrete understanding of the context. This restatement typically ensures that the response is relevant and maintains continuity within the conversation.

Tokenization similarity can be quantified using various metrics such as cosine similarity, Jaccard similarity, and edit distance. These metrics are quantitative methods of comparing the sets of tokens from the user's input and the LLM's output, providing a measure of overlap or distance between them. High token similarity scores suggest that the LLM accurately captures and addresses the key elements of the user's query, indicating strong comprehension and contextual awareness.

Moreover, tokenization similarity is useful for evaluating the consistency and relevance of an LLM's responses over extended conversations. By tracking how the LLM maintains topic coherence and appropriately integrates previous conversational elements, researchers can assess the model's ability to engage in meaningful and contextually appropriate dialogue. This approach can also help identify instances where the LLM might drift off-topic or fail to adequately address the user's input, providing valuable insights for model improvement. While contextual understanding is crucial, a model that overemphasizes token similarity may become overly reliant on restating the user's input, thus limiting the depth and originality of its responses. Overall, using tokenization similarity as a metric for evaluating conversational AI supports the development of more responsive, context-aware, and user-centric conversational agents.

1) Cosine similarity: Cosine similarity is a metric used to measure the similarity between two non-zero vectors in an inner product space, often applied in text analysis to assess how similar two strings of text are. It calculates the cosine of the angle between the two vectors, representing their orientation rather than magnitude. The cosine similarity score ranges from -1 to 1, where a score of 1 indicates that the vectors are identical, 0 implies no similarity, and -1 means they are opposed. Our calculations were built upon the TF-IDF vectorizer from the sklearn library. This vectorizer converts the string input into a matrix of TF-IDF features, which are terms weighted by importance. The term frequency (TF) reflects how often a term appears in the text, while the inverse document frequency (IDF) reduces the weight of terms that appear frequently across both texts. The cosine similarity between these vectors can then be computed from the resultant matrices to determine how similar the texts are.

$$\text{Cosine Similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{1}$$

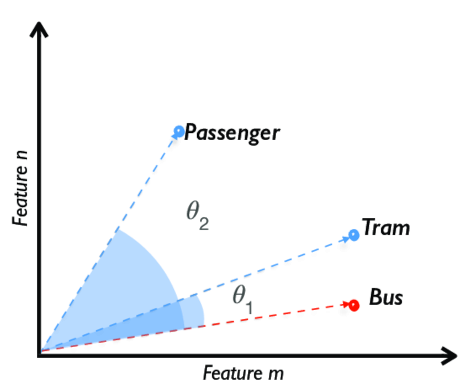

Fig. 1: Visualization of Cosine Similarity [11]

2) Jaccard similarity: Jaccard similarity, also known as the Jaccard index or Jaccard coefficient, is a statistical measure used to evaluate the similarity and diversity of sample sets. It is defined as the size of the intersection divided by the size of the union of the sample sets. Mathematically, the Jaccard similarity between two sets A and B is given by the formula:

$$\text{Jaccard Similarity}(\mathbf{A}, \mathbf{B}) = \frac{|\mathbf{A} \cap \mathbf{B}|}{|\mathbf{A} \cup \mathbf{B}|} \tag{2}$$

where $|A \cap B|$ is the number of elements in the intersection between sets A and B, and $|A \cup B|$ is the number of elements in the union of sets A and B. The Jaccard similarity score ranges from 0 to 1, where 0 indicates no similarity (no common elements), and 1 indicates identical sets (perfect overlap).

In the context of natural language processing and conversational AI, Jaccard similarity can be used to compare the content of two text strings by treating them as sets of tokens, such as words, n-grams, etc. For example, when evaluating the relevance of an AI-generated response to a user's input, Jaccard similarity can measure the overlap in terms of shared tokens. If a user asks, "What are the benefits of regular exercise?" and the LLM responds, "Regular exercise improves cardiovascular health and overall well-being," the sets of tokens from both the input and response can be compared. The Jaccard similarity score in this case would be calculated by identifying the common tokens (e.g., "regular," "exercise") and dividing by the total unique tokens in both texts. A higher Jaccard similarity score indicates that the LLM's response shares a significant portion of tokens with the user's input, suggesting that the response is contextually relevant and aligned with the query. This metric is particularly useful for applications such as text clustering, duplicate detection, and information retrieval, where the goal is to identify and group similar pieces of text.

3) Edit distance: Edit distance, also known as Levenshtein distance, is a metric used to quantify the difference between two strings by counting the minimum number of single-character edits required to transform one string into the other. These edits can include insertions, deletions, or substitutions of characters. The edit distance provides a measure of similarity between the strings: a smaller edit distance indicates greater similarity, while a larger distance suggests more significant differences.

In natural language processing and conversational AI, edit distance is valuable for tasks that involve comparing and evaluating text, such as spell checking, plagiarism detection, and evaluating the quality of generated responses. For instance, when an LLM generates a response to a user's input, the edit distance can be used to measure how much the response deviates from a reference answer or the original input. This metric helps in assessing the precision and relevance of the LLM's output. Consider a scenario where a user asks, "You remove a splinter from your hand with what?" and the LLM responds with, "You remove a splinter from your hand with tweezers." The edit distance between the user's query and the LLM's response is minimal (only some character changes from "what?" to "tweezers."), indicating a high level of similarity and relevance. In another scenario, if the response was "What are the ingredients for a cake?", the edit distance would likely be higher due to multiple word changes, reflecting a shift in focus from the original question. Because of this, edit distance will only be a particularly viable metric for evaluating direct, short-answer responses for clarity and conciseness.

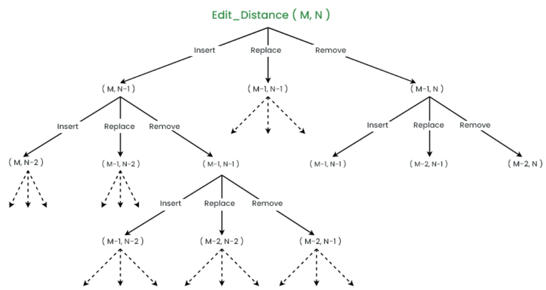

Fig. 2: The decision tree for the Edit Distance [9]

B. *Qualitative Metrics*

We also evaluated LLM's conversational capabilities by employing the Big Five personality traits as qualitative metrics to assess the response quality [8]. The Big Five framework includes openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. These personality traits provide a comprehensive approach to understanding and mimicking human-like conversational nuances. Additionally, we evaluated the responses contextually by considering their vocabulary and tone. By grading LLM responses with these personality traits, we aim to evaluate precisely both the naturalness and the contextual relevance of its interactions.

1) Openness: Openness to experience is characterized by an inclination toward creativity, curiosity, and a wide range of interests. This trait allows us to assess how well the models generate responses that are innovative, insightful, and open to new ideas. High openness in LLM responses is reflected in the model's ability to engage with diverse topics and offer creative solutions to complex questions.

2) Conscientiousness: Conscientiousness describes organization, dependability, and attention to detail. This trait is crucial in evaluating how consistently and reliably the models generate accurate, well-structured, and goal-oriented responses. A highly conscientious LLM response should demonstrate thoroughness, precision, and adherence to structure for what is expected in that conversation.

3) Extraversion: Extraversion is characterized by a tendency to seek out social interactions, excitement, and positive emotional experiences. This trait helps us gauge the LLM's ability to produce engaging, lively, and socially adept responses. High extraversion in LLMs would be evident through enthusiastic and interactive dialogues that effectively engage users and sustain conversational flow.

4) Agreeableness: Agreeableness reflects a propensity for compassion, cooperation, and social harmony. This trait measures the capacity to generate empathetic and friendly responses. In LLM responses, agreeableness is shown by an empathetic and supportive tone, making the interaction feel genuine and pleasant.

5) Neuroticism: Neuroticism relates to emotional stability and the propensity to experience negative emotions. Evaluating this trait helps us understand the model's ability to handle and convey emotional content. Low neuroticism would indicate stability and calmness, whereas higher neuroticism might reflect more reactive or emotionally nuanced responses. For LLMs, lower neuroticism would be preferable, as it would result in more stable, calm, and reassuring interactions rather than overly negative or anxious language. While the model must accurately assess the user's emotions and sentiments, emotional stability in its responses is essential to guarantee consistency and reliability.

6) Tone: The tone of an LLM's response is critical for shaping user perception and engagement. An effective LLM should maintain a tone that is appropriate to the context. The tone can be formal, informal, empathetic, or authoritative, depending on the context that the user has provided to an LLM. By analyzing tone, we evaluate the LLM's ability to align with the user's expectations and the conversational context. A well-calibrated tone enhances the user experience by ensuring that interactions feel natural and respectful.

7) Vocabulary: An LLM with a rich and varied vocabulary is capable of providing accurate answers that are accurate and well suited to the context. We assess the LLM's use of vocabulary by selecting words and phrases that maximize clarity and engage the user. Effective terminology should be flexible, avoid redundancy, and be appropriate and relevant to the user's needs and conversational context.

C. *Dataset*

We created a diverse dataset of 50 prompts and corresponding responses to reflect all Big Five personality traits. These prompts range from simple questions such as "What is the capital of France?" to more complex prompts such as "Explain how you can manage stress during high-pressure situations." For each prompt, we generated a response intended to score highly when evaluated against the Big Five personality traits. The purpose of these responses is to introduce a set of idealized examples for fine-tuning the LLM. For this dataset, we prioritized the variety of prompts and responses over the quantity to effectively assess whether the fine-tuned LLM will exhibit openness, conscientiousness, extraversion, agreeableness, and neuroticism. Furthermore, the small size of this dataset allows for reduced noise with unregulated responses and a lower computation cost. Notably, LLMs can extrapolate from a small set of examples, making this dataset a sufficient proof-of-concept to demonstrate the potential impact of fine-tuning on personality simulation [10]. This property of LLMs ensures that even with limited data, meaningful patterns can be learned, which could be further explored in future studies.

D. Research Process

We began by creating a diverse list of several prompts to be inputted into the original GPT-3.5 Turbo and GPT-4o mini LLMs. In the dataset creation phase, we initially had multiple contributions. To ensure comprehensive coverage, we included three examples per category across a broad range of topics, such as General Knowledge, Science and Technology, History, Literature, Geography, Arts and Culture, Social Issues, Health and Wellness, Education and Career, Current Events, Philosophy, and Ethics. The dataset was randomly split into 80% for training, 10% for validation, and 10% for testing, with each subset distinct and separate. All work was conducted using a Google Colab notebook, utilizing a private API key for access to the OpenAI models. Each LLM's output was evaluated based on the Big Five personality traits, as well as the contextual appropriateness of tone and vocabulary. Through this evaluation, we identified areas for improvement, intending to maximize the models' conversational ability.

After this stage, we created a dataset of 50 diverse prompts, covering categories ranging from science to moral dilemmas. These prompts were also designed to yield LLM responses of ranging lengths. We created an ideal response for each prompt and focused on maintaining conciseness, understanding, and detail. This dataset was later used to fine-tune a GPT-3.5 Turbo and GPT-4o-mini LLM.

After the fine-tuning process, we created a separate test set of three prompts. These prompts were inputted into both the original and the fine-tuned LLMs. From the resulting responses, we compared them against each other to evaluate their conversational ability. These criteria included the Big Five personality traits and the quantitative metrics to determine text similarity.

IV. Results

In this section, we present the results of our study on generating human-like responses using a Large Language Model (LLM). The objective of this study was to evaluate the effectiveness of LLM fine-tuning by assessing their responses across a broad set of real-world prompt examples.

To quantify the differences in conversational performance between the original LLM and the fine-tuned version, we utilized several different metrics: cosine similarity, tokenization similarity, Jaccard similarity, and edit distance. We also qualitatively assessed the LLM's ability to emulate the Big Five personality traits: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism.

TABLE I: Model Comparison

Model	Response	BFPT	Cosine	Jaccard	Edit Dist.
GPT 3.5	Original	0.698	0.620	0.600	300
	Finetuned	0.820	0.675	0.662	250.1
GPT 4o mini	Original	0.667	0.640	0.620	320
	Finetuned	0.780	0.694	0.837	463.17

Comparison of original and fine-tuned model performances across various metrics: Big Five Personality Traits (BFPT), Cosine Similarity, Jaccard Similarity, and Edit Distance.

Our BFPT metric, which is the average of the Big Five Personality Trait scores, shows how our fine-tuning process led to a direct improvement in contextual relevance and overall conversational ability. The model is more enhanced and better able to engage with specific user inputs. For instance, the increase in Cosine and Jaccard similarities shows a better alignment between the model's output and the user's responses. Furthermore, reducing edit distance emphasizes the fine-tuned model's improved accuracy in generating text that closely matches the user's input, minimizing errors and enhancing coherence.

Moreover, the improvement in the Big Five Personality traits shows how it can now generate responses that are more in tune with human-like conversations. This enhancement is not just shown in numbers, as it is now able to mirror user linguistic features and behavioral traits. The model adjusts its language and behavior which makes interactions more natural.

Additionally, the improvements reflect Brown and Levinson's politeness theory, which suggests that the fine-tuned model is having smoother and more respectful communication. The enhanced mimicry and naturalness in the AI's interaction are evident in its ability to deliver responses that are not only technically accurate but also emotionally acceptable. This suggests a more sophisticated understanding of user intent and a greater ability to engage in meaningful, context-aware dialogues.

V. Conclusion

Our research demonstrates that fine-tuning LLMs can lead to significant advancements in their ability to produce responses that are both contextually accurate and stylistically appropriate. By carefully calibrating the model and changing parameters, we improved its understanding of user inputs, resulting in more precise and relevant responses. The refined LLM not only shows enhanced performance in terms of similarity scores but also exhibits a stronger ability to replicate human-like conversational styles. This is particularly evident in the model's alignment with the Big Five personality traits, where the fine-tuning process allowed it to better capture and reflect individual user characteristics. This capacity for nuanced and personalized interaction is crucial for creating more engaging and satisfying user experiences. Overall, our approach to fine-tuning highlights the importance of strategic adjustments in making LLMs more effective and responsive. These enhancements ensure that the model is not just technically proficient but also capable of delivering meaningful and human-like interactions, ultimately making it a more valuable tool for a wide range of applications.

A. Future Work

To build upon our research, we encourage investigating new methods to refine quantitative metrics to evaluate text similarity and relevance. This will contribute to a deeper understanding of language variations and enhance the overall

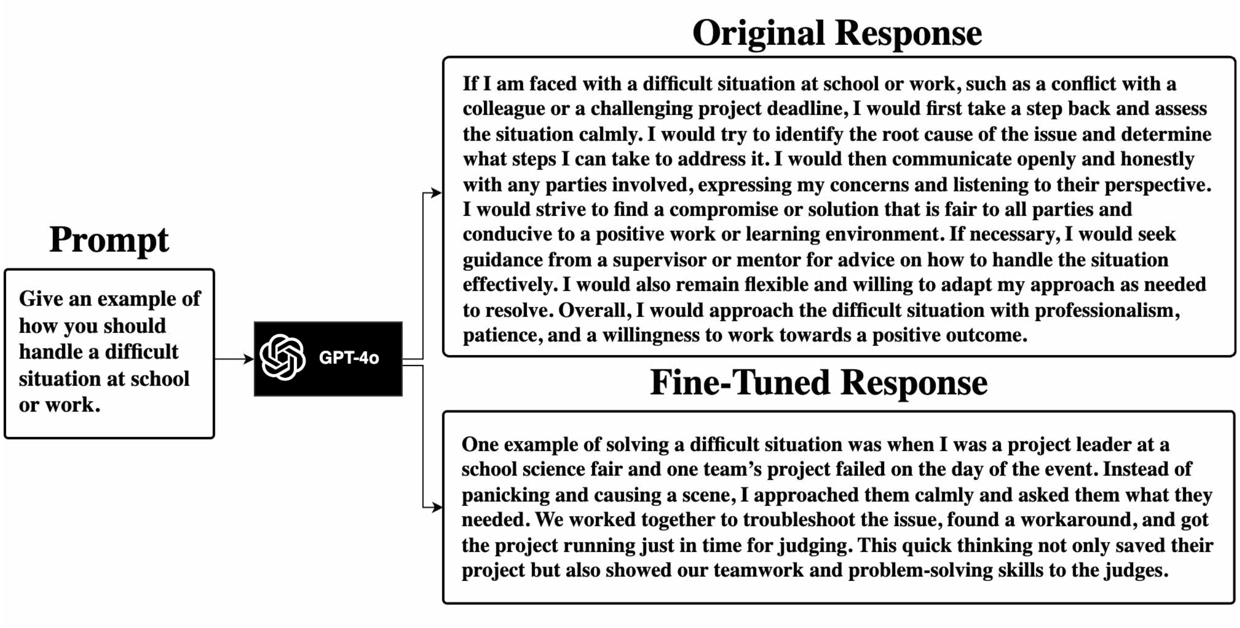

Fig. 3: Comparison between GPT-4's original and fine-tuned responses to a prompt about handling difficult situations at school or work. The fine-tuned response demonstrates improved relevance and specificity.

effectiveness of fine-tuning to help create more conversational LLM responses. Moreover, to increase engagement and satisfaction, a possible implementation is to incorporate reinforcement learning into an active fine-tuning process. This way, the LLM will be able to continually improve over multiple interactions with the user, and further refine the model's conversational ability.

VI. Supplementary Material

We created a dataset of 50 diverse linguistic questions and answers designed to provide a comprehensive evaluation of an LLM's capabilities in conversational acuity. The data along with further details, including data set descriptions, preprocessing code, and experimental results are available at https://github.com/EpicGamer3386/LLM-Conversational-Acuity.

VII. Acknowledgements

We would like to thank Ivan Jose for his contributions in finalizing this paper, and BLAST AI for supporting the project.

References

[1] S. Roller, et al., "Language Models as Knowledge Bases?" *arXiv preprint*, 2020. [Online]. Available: https://arxiv.org/abs/2006.12442.

[2] M. Jadeja and N. Varia, "Perspectives for Evaluating Conversational AI," *arXiv preprint*, 2017. [Online]. Available: https://arxiv.org/abs/1709.04734.

[3] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, H. Albanna, M. A. Albashrawi, A. S. Al-Busaidi, J. Balakrishnan, Y. Barlette, S. Basu, I. Bose, L. Brooks, D. Buhalis, L. Carter, R. Wright, "Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *ACM*, 2023. [Online]. Available: https://dl.acm.org/doi/10.1016/j.ijinfomgt.2023.102642.

[4] A. Sorokovikova, N. Fedorova, S. Rezagholi, and I. P. Yamshchikov, "LLMs Simulate Big Five Personality Traits: Further Evidence," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2402.01765.

[5] H. Giles and T. Ogay, "Communication Accommodation Theory," *PsycNET*, 2007. [Online]. Available: https://psycnet.apa.org/record/2006-21534-016.

[6] L. Bojic, P. Kovacevic, and M. Cabarkapa, "GPT-4 Surpassing Human Performance in Linguistic Pragmatics," *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/pdf/2312.09545.

[7] S. L. Sravanthi, "Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/html/2401.07078v1.

[8] G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu, "Evaluating and Inducing Personality in Pre-trained Language Models," *arXiv preprint*, 2023. [Online]. Available: https://arxiv.org/abs/2206.07550.

[9] Aayushbvc8m, "CSES Solutions – Edit Distance," *GeeksforGeeks*, 2024. [Online]. Available: https://www.geeksforgeeks.org/cses-solutions-edit-distance.

[10] C. Xiao, P. Zhang, X. Han, G. Xiao, Y. Lin, Z. Zhang, Z. Liu, M. Sun, "InfLLM: Training-Free Long-Context Extrapolation for LLMs with an Efficient Context Memory," *arXiv*, 2024. [Online]. Available: https://arxiv.org/html/2402.04617v2.

[11] S. Kalwar, M. Rossi, M. Sadeghi, "Automated Creation of Mappings Between Data Specifications Through Linguistic and Structural Techniques," IEEE, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10077347/.

[12] OpenAI, "GPT-3.5 Turbo," OpenAI, 2023. [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5-turbo.

[13] OpenAI, "GPT-4O," OpenAI, 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/.

[14] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2021, pp. 610-623. [Online]. Available: https://dl.acm.org/doi/10.1145/3442188.3445922.