

Project Final Report: Fine-tuning LLMs for Persona Injection

Harshil Bhullar
hbhullar18@g.ucla.edu

Krish Patel
knpatel@g.ucla.edu

Siddarth Chalasani
siddarth@cs.ucla.edu

Abstract

Recent advancements in large language models (LLMs) have opened new frontiers in natural language processing, particularly in enhancing chatbot interactions through persona injection. This study explores fine-tuning GPT-3.5-turbo and Llama-3-8B on Google’s Synthetic Persona Chat dataset to enhance persona consistency in chatbots. We experimented with various prompt engineering and fine-tuning strategies to effectively inject personas into dialogues. The models were evaluated using BLEU, ROUGE, BERTScore, and perplexity metrics, showing marked improvements over baselines. Fine-tuned models not only enhanced persona coherence but also outperformed traditional models in both quantitative metrics and human evaluations, underscoring the efficacy and potential challenges of persona-driven conversational agents.¹

1 Problem Introduction

In our project, we tackle the task of creating “persona chatbots”, i.e., language models that are capable of carrying a fluent, continued conversation while basing the content of their utterances on a pre-defined persona. Although the problem we attempt to solve is not novel, previous attempts usually either failed to maintain a consistent personality [1] or suffered from long-term memory degeneration [3]. In our project, we survey current literature on persona chatbots, evaluate baseline models currently being employed, and suggest possible improvements to yield improved results.

2 Literature Review and Baseline Model Selection

After reviewing the 3 papers we initially planned on building our project with, we decided to drop the Facebook dataset and Cornell Movie Dialogue,

which will be discussed further below. We are going to continue with Google’s Synthetic Persona Chat Dataset. We now do a short review of each of the papers that we considered for our project.

2.1 Facebook AI’s Persona-Chat Dataset and Baseline Model

The paper “Personalizing Dialogue Agents: I have a dog, do you have pets too?” reintroduces general chit-chat models as a valuable end-application [4]. To aid in the task of creating such models, the authors created the PERSONA-CHAT dataset, which consists of 162,064 utterances between randomly paired crowdworkers who were each assigned one of 1155 personas to base their conversation on. These personas, consisting of at least 5 sentences, were also generated by crowdworkers and assigned randomly. The paired crowdworkers were then asked to chat in a natural, engaging manner. To avoid trivial word overlap in the dataset, each of the 1155 personas was rewritten multiple times by crowdworkers, with explicit word overlap being forbidden.

The paper then evaluates two classes of models: ranking models, which the paper uses as its baseline; and generative models, which the paper uses as its novel approach. The ranking models employed are a greedy IR model and Starship, and the generative model used is based on an LSTM architecture with the hidden encoder state initialized to the embedding of the provided persona. The paper found that, although their generative model did not necessarily outperform the ranking models on their evaluation metrics (perplexity and hits@1), both classes of models were significantly better at the task of persona-based chit-chat when conditioning prediction on their own persona.

Therefore, we consider their ranking models and LSTM-based architecture as our baseline. The pa-

¹Code available at <https://github.com/krish1925/Persona-Chatbot-G28>

per also presents a platform for training and evaluating dialogue models, ParlAI [2], which could aid our experiments. We decide to discard the PERSONA-CHAT dataset in favor of Google’s Synthetic Persona Chat dataset (described in the following section) because of its superior quantity of data at similar quality.

2.2 Google’s Synthetic Persona Chat Dataset

The paper "Faithful Persona-based Conversational Dataset Generation with Large Language Models" introduces a novel method for creating high-quality conversational datasets that are faithful to user personas. The authors propose a structured approach involving three key components: User Generation, User Pairing, and Conversation Generation.

User Generation involves two sub-steps: Persona Expansion and User Profile Construction. The process starts by bootstrapping seed persona attributes using prompts to generate a broader set of persona attributes. These attributes are then used to construct user profiles with the aid of a Natural Language Inference (NLI) model, ensuring the consistency and non-redundancy of the profiles.

User Pairing identifies potential pairs of user profiles for conversation based on their semantic similarity, which is assessed using the BERT model. The goal is to pair profiles that share common persona attributes to potentially enhance the engagement level of the conversations.

Conversation Generation uses a Generator-Critic architecture. The Generator creates initial conversational exchanges between paired profiles, which are then evaluated by the Critic based on predefined quality metrics. This iterative process aims to refine the conversations further, ensuring they meet the desired standards of relevance and persona consistency.

This framework not only allows for the generation of personalized dialogues but also ensures that the conversations are aligned with the constructed personas, thereby avoiding the generation of dialogues that contradict the persona attributes. This means that the dataset created with this paper’s approach is perfect for our use case for creating persona chat bots.

2.3 Cornell Movie Dialogue Dataset Issues

Initially, our team considered using the Cornell Movie Dialogue Dataset, which contains around 9,000 personas or different speakers and over 304,000 conversations. However, while delving deeper into the dataset, several issues emerged that led us to reconsider its usability for our project.

Firstly, the dataset lacked fundamental information about the speakers involved in the conversations. While the gender of the speakers was available, other relevant details such as their interests, backgrounds, or roles in the conversations were absent, and fields such as the movie name and speaker name were irrelevant in this scenario. This lack of speaker information limited our ability to analyze their dialogue and understand the context of the conversation.

Additionally, the dataset only contained dialogues extracted from movies, which presented more challenges. Unlike real-life conversations, movie dialogues are scripted and may not accurately reflect the nuances and dynamics of natural language interactions. Along with that, the context surrounding these dialogues, such as the setting, character relationships, and plot developments, was often missing or ambiguous.

3 Description of Approach

We were inspired by the generative models used in the baseline model’s paper, such as the sequence to sequence model, and we decided we wanted to build off this by fine-tuning on LLMs, specifically GPT-3.5-turbo and Llama-3-8B.

For the GPT-3.5-turbo model, we conducted fine-tuning experiments on three distinct datasets: the entire training dataset, 50 carefully curated samples optimized for recall, and 200 samples with finely-tuned prompts. Subsequently, we compared the performance of these fine-tuned models with the baseline results.

3.1 GPT-3.5-turbo: Fine-Tuned Model on Entire Dataset

Our first attempt at fine-tuning used the entire SCP dataset on GPT-3.5-turbo. Using the OpenAI API, we converted the training data into a JSONL file,

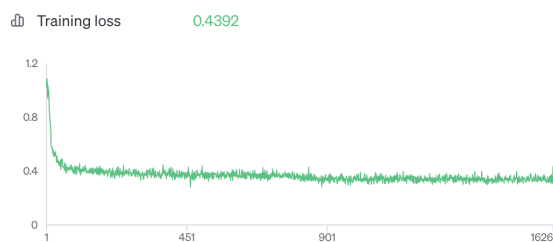


Figure 1: GPT-3.5-turbo fine-tuning loss on entire dataset

where each object had 3 objects: object 1 was the system prompt, which is the 'context' of the task we are explaining to the model. Object 2 was the user prompt, which is where we provide the details for the specific task. In this case, we provided the personas of both user 1 and user 2. Object 3 is the assistant prompt, which is the expected output that we want the model to produce when given the first 2 pieces of the prompt. For this model, the assistant prompt contained the entire conversation that the personas of user 1 and user 2 had. We can see an example of the exact prompt we provided in Figure 2.

This model was a naive attempt at fine-tuning, as it was really more of a brute-force method of providing as much data as possible without really structuring it efficiently. There is no real benefit from providing the entire conversation to the model as an expected output, because in reality we expect the model to respond only once directly to the last conversation piece, until we prompt it again. This method also incurred a large cost to us, since training a model with about 9000 datapoints (each one containing a large number of tokens due to the entire conversation being provided) is very expensive. We can see in the loss curve in Figure 1 that the loss stopped seeing meaningful reduction after only 400 iterations.

After learning from the mistakes of our first fine-tuned GPT model, we decided to curate the dataset and apply our prompt engineering progress to our data. We picked 50 datapoints in the dataset that we felt would best aid our fine-tuning. This was also significant since we wanted to reduce the cost of fine-tuning after the expensive mistake we made earlier.

This fine-tuned model had an improved prompt. The system prompt remained largely the same as our goal was still the same: persona injection. The key change here was that we explained that we

would only be providing the persona of user 2, the user persona that we want the model to embody in its response. The user prompt changed to now provide the persona of user 2, and then also provide the first 6 lines of conversation between user 1 and user 2. This provided an indirect form of in-context learning. Then, we provided the 7th line of the conversation (which would be a response from user 1), and instructed the model to respond to this 7th line as user 2. The assistant prompt also had a big change: we only provided the direct response to the 7th line of the conversation, NOT the entire rest of the conversation.

As we can see from the loss curve in Figure 3, this model performed almost as good as the first model, and it had significantly less data and iterations than the first model. This showed us that we were headed in the right direction, leading us to our 3rd and final GPT model.

3.2 GPT-3.5-turbo: Fine-Tuned Model on 200 Curated Samples

This model only had 1 difference from the previous model: the number of datapoints. At this point, we had reached the end of our prompt engineering and were satisfied with our final prompt. So we decided that the main area for improvement for our model was to increase its fine-tuning by adding 150 more datapoints, for a total of 200 curated samples. As we can observe from our loss curve in Figure 4, this model performed better than the 50 sample fine-tuned GPT model, validating our hypothesis.

3.3 Llama-3-8B

In the case of the Llama-3-8B model, we employed a similar approach. We decided to test the pre-trained model as a baseline against a model fine-tuned on the Synthetic Persona Chat dataset. We initially ran into some issues loading the 'meta-llama/Meta-Llama-3-8B' onto our GPU as it was too large, so we instead used a 4-bit quantized model for improved load time and training efficiency ('unsloth/llama-3-8b-bnb-4bit').

The fine-tuning process for Llama-3 was pretty similar to that of GPT-3.5, but took advantage of the unsloth framework and API. We adapted the processed data from the GPT-3.5 fine-tuning process to pass into Llama-3.

We decided to train for $\sim 20\%$ of an epoch to avoid excessive computation. We noticed that the training loss dropped drastically after the first 20 steps and stayed relatively constant throughout the

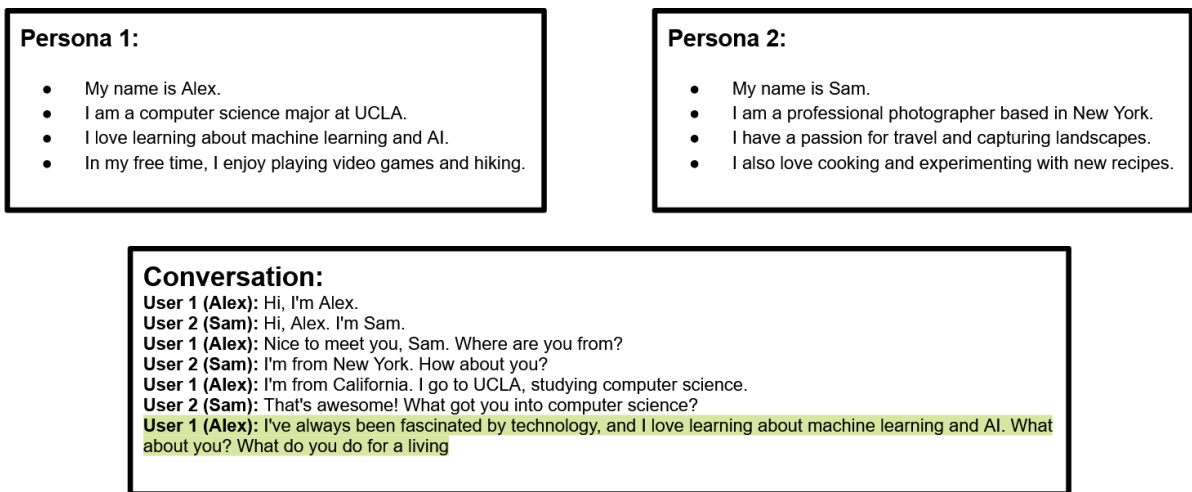


Figure 2: Sample prompt for Entire Dataset GPT

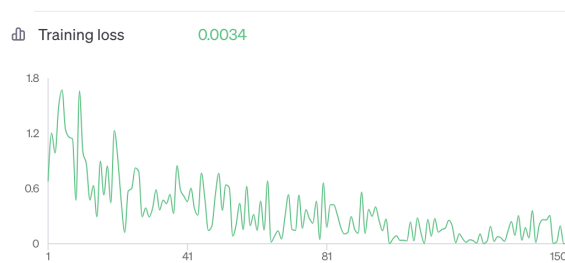


Figure 3: GPT-3.5-turbo fine-tuning loss on 50 curated samples



Figure 4: GPT-3.5-turbo fine-tuning loss on 200 curated samples

remainder of the fine-tuning, so we decided to also compare the model snapshotted after 20 steps to see if the additional training had a noticeable impact.

4 Dataset and Evaluation Framework

4.1 Quantitative Evaluation Metrics

For our evaluation metrics, we decided to use Perplexity, BLEU, and ROUGE to evaluate our fine-tuned model's performance

- **Perplexity:** For calculating perplexity, we evaluated the scores using a GPT-2 model to-

kenizer from the transformers library, which computes the loss for the generated outputs of the fine-tuned/baseline models.

- **BLEU:** We used nltk's BLEU score evaluation metrics, which measure the similarity between the generated text and one or more reference translations, comparing the LLM outputs to the ground truth (reply of the chatbot in the dataset). This operates by comparing n-grams (sequences of n consecutive words) between the generated text and the reference text, calculating a precision score for each n-gram size (usually up to 4-grams), and then combines these scores using a modified brevity penalty to calculate the final BLEU score.
- **BERTScore:** We used BERTScore to evaluate the quality of generated text by leveraging contextual embeddings from BERT (Bidirectional Encoder Representations from Transformers). BERTScore measures the similarity between the generated text and one or more reference translations using pre-trained BERT models. It computes a similarity score by comparing contextual embeddings of tokens between the generated text and the reference text, effectively capturing semantic similarity and providing a robust evaluation of our model's text generation performance.
- **ROUGE or Recall-Oriented Understudy for Gisting Evaluation:** We used nltk's ROUGE score evaluation metrics to measure the overlap between the generated summaries or texts and one or more reference summaries.

ROUGE calculates various measures such as ROUGE-N (which focuses on n-gram overlap), ROUGE-L (which measures the longest common subsequence between the generated and reference texts), and ROUGE-W (which considers weighted versions of ROUGE-N). Similar to the other metrics, a score of 1.0 indicates a perfect match between the two comparison dialogues.

4.2 Qualitative Evaluation Metrics

To complement our quantitative assessments, we employed a qualitative evaluation strategy that involved direct interactions with the fine-tuned models. This approach was aimed at gauging the naturalness and persona consistency of the chatbots in a live conversational context. Each model was assigned a distinct persona, and we engaged in scripted and spontaneous dialogues to explore the depth and adaptability of persona injection. We used personas from the dataset, and then curated our own conversations. Evaluators focused on several key aspects: the relevance and coherency of responses, the ability of the chatbot to maintain persona traits throughout the interaction, and its responsiveness to conversational cues. This method provided insights into the practical efficacy of the models in real-world scenarios, allowing us to assess subtle nuances of dialogue that purely automated metrics might overlook. As we will see in the Results section, while the models that were trained with more data and computational resources may have had better quantitative results, the conversations we had with the other models often outperformed them.

5 Results

Here we talk about the results of our quantitative and qualitative metrics. We mainly observe the evaluation metrics seen in the tables below, and compare some sample conversations between baseline models and our fine-tuned models.

5.1 Quantitative Results

We observe in Table 1 the results of the baseline models. Specifically, we observe how low these values are, relative to the results of our fine-tuned models in Table 2 and Table 3. We knew going into this project that our fine-tuned models would outperform the baseline models, simply because we are leveraging the power of LLMs, whereas

Model	BLEU	P/R/F1	hits@1
IR Baseline	0.2117	0.3028 0.3102 0.3018	0.2114
Starspace	0.1404	0.2306 0.2282 0.2248	0.1400
Seq2Seq (LSTM)	3.822e-11	0.4001 0.0833 0.1355	0.1250

Table 1: Evaluation of PersonaChat baseline models using ParlAI [2] [4]

the baseline models are LSTMS and Transformer variants.

Moving onto the results in Table 2, we observe that as expected, the Full Dataset fine-tuned GPT model performed the best across all metrics. It achieved the highest BLEU score, ROUGE score, precision, recall, and F1 score. It also achieved the lowest perplexity (lower is better). This was expected simply because of the sheer compute resources we gave this model. And while this model achieved the greatest quantitative metrics, the conversations we discuss later will show that is not the full story. Regardless of the qualitative metrics, we still understand that this model is not the most efficient model, as the other models with much less resources had greater potential as seen by their loss curves that we discussed earlier, and their proximity in their quantitative metrics to the Full Dataset Fine-Tuned GPT.

Moving onto the quantitative metrics for Llama-3, seen in Table 3, we first observe that, on the whole, Llama-3 quantitative metrics are lower than those of GPT-3.5 and even some of the PersonaChat baseline models. Furthermore, the partially fine-tuned and fine-tuned model perform better than the baseline but pretty similar to each other on BLEU and ROUGE. However, somewhat surprisingly, the pre-trained baseline has the best BERTScore precision and perplexity scores, and the partially fine-tuned model has the best BERTScore recall and F1 scores. This is probably indicative that the quantitative metrics don't perfectly mirror human judgement, as we see from the qualitative analysis that the fully fine-tuned model significantly outperforms the pre-trained and partially fine-tuned

model in terms of output format consistency and conversational flow.

5.2 Qualitative Results

The finetuned model trained on a more smaller dataset generated smaller but more to the point replies compared to the other gpt-3.5 finetuned models.

Despite the superior performance of the full dataset GPT model in quantitative metrics, the subjective quality of the conversations it generates is not significantly better than those generated by the other models. This highlights the limitations of relying solely on quantitative metrics for evaluating conversation quality. We can observe in the sample conversations 6 7 that all 3 of the GPT models have similar performance, and the detail in their response is pretty close in coherence.

Similarly with Llama-3, we see that the conversations 8 for the 2 partially finetuned models share similar coherence and detail to the GPT models, even though the quantitative metrics are much lower. This also shows that there is a need for better automated quantitative metrics since they do not represent the conversation quality.

5.3 GPT-3.5-turbo vs Llama-3-8B

Consider an example generated by the LLaMA model. Although the LLaMA model did not achieve the highest scores in the quantitative metrics compared to the GPT models, the generated conversations were often more engaging and contextually coherent when evaluated by human users. We observed a more natural flow in dialogues, with the LLaMA model demonstrating a better understanding of conversational nuances. This suggests that while quantitative metrics are important, they do not always capture the full picture of conversational quality and user satisfaction.

All the fine-tuned models outperformed their respective pre-trained versions, with GPT-3.5-turbo significantly surpassing LLaMA-3-8B on evaluation metrics like Perplexity, BLEU, ROUGE, and BERTScore. However, qualitatively, both finetuned models performed similarly well, and much better than the PersonaChat models, indicating that fine-tuning significantly enhances performance. Notably, GPT-3.5-turbo occasionally showed inconsistencies in output format, which did not detract significantly from its overall performance but

highlights areas for potential improvement. Despite the superior quantitative performance of GPT-3.5-turbo, the qualitative evaluation underscores the importance of a balanced approach in evaluating conversational AI models.

6 Limitations and Future Works

This study encountered several limitations that highlight areas for future research. The use of the ParlAI library posed challenges due to outdated Python packages and inconsistent evaluation metrics in the PersonaChat baseline models, impacting our development process. Financial constraints were also notable, as fine-tuning models on OpenAI proved costly, suggesting a need for more budget-friendly solutions. Our approach to prompt engineering, while effective, requires further refinement and a more researched-based methodology. Additionally, exploring a wider range of chat datasets could enhance the transferability and robustness of our findings. In terms of evaluation, the current metrics used do not fully align with human judgment, indicating a potential benefit in expanding both quantitative measures and incorporating crowdsourced human evaluations to gain deeper insights into model performance and user satisfaction.

References

- [1] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany, August 2016. Association for Computational Linguistics.
- [2] Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. ParlAI: A dialog research software platform. In Lucia Specia, Matt Post, and Michael Paul, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [3] Oriol Vinyals and Quoc V. Le. A neural conversational model. *ArXiv*, abs/1506.05869, 2015.
- [4] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of*

Model	BLEU	ROUGE	BERTScore	Perplexity
Pre-Trained GPT-3.5	0.1755	0.2365	P: 0.6180 R: 0.3102 F1: 0.3018	3.360
Fine-Tuned GPT-3.5 with Initial Prompt (Full Dataset)	0.4245	0.4973	P: 0.7487 R: 0.6818 F1: 0.7070	3.150
Fine-Tuned GPT-3.5 with Final Prompt (Small Dataset)	0.3934	0.4630	P: 0.6896 R: 0.6432 F1: 0.6609	3.372
Fine-Tuned GPT-3.5 with Final Prompt (Large Dataset)	0.3763	0.4411	P: 0.6852 R: 0.6698 F1: 0.6737	3.188

Table 2: Evaluation of GPT-3.5-turbo models

Model	BLEU	ROUGE	BERTScore	Perplexity
Pre-Trained Llama-3	0.0987	0.1444	P: 0.5558 R: 0.4509 F1: 0.4926	2.373
Fine-Tuned Llama-3 (~0.02 epochs)	0.1989	0.2617	P: 0.5518 R: 0.5392 F1: 0.5393	3.081
Fine-Tuned Llama-3 (~0.2 epochs)	0.1993	0.2705	P: 0.5418 R: 0.5389 F1: 0.5339	3.205

Table 3: Evaluation of Llama-3-8B models

the Association for Computational Linguistics (Volume 1: Long Papers), pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

A Example Conversations

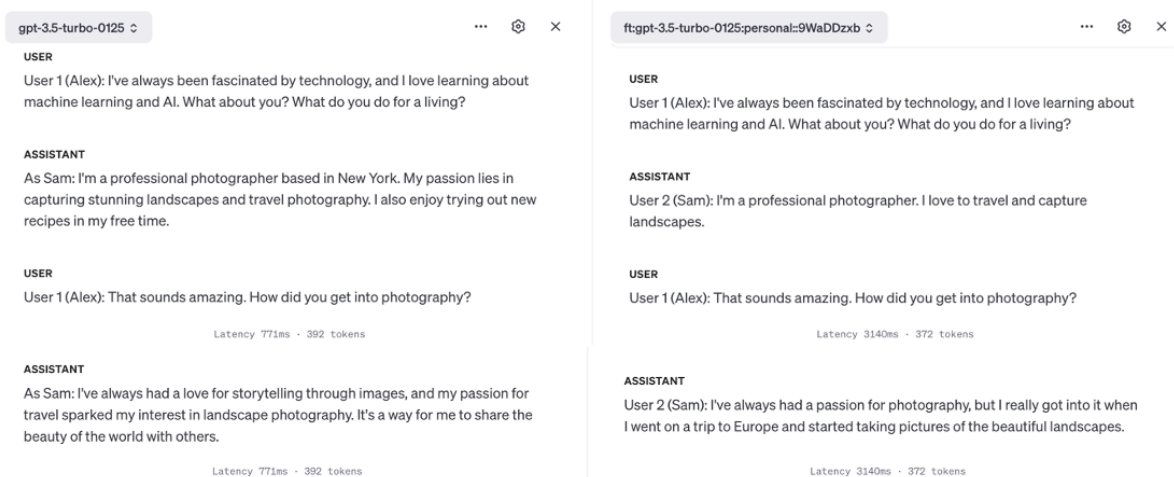


Figure 5: GPT-3.5-turbo: Pre-Trained vs. Fine-Tuned on Entire Dataset

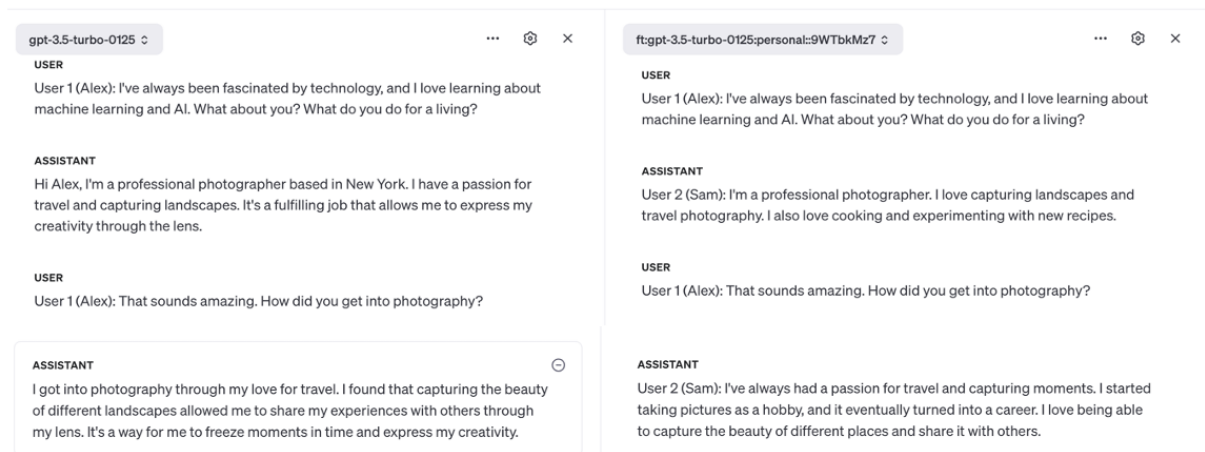


Figure 6: GPT-3.5-turbo: Pre-Trained vs. Fine-Tuned on 50 Curated Samples

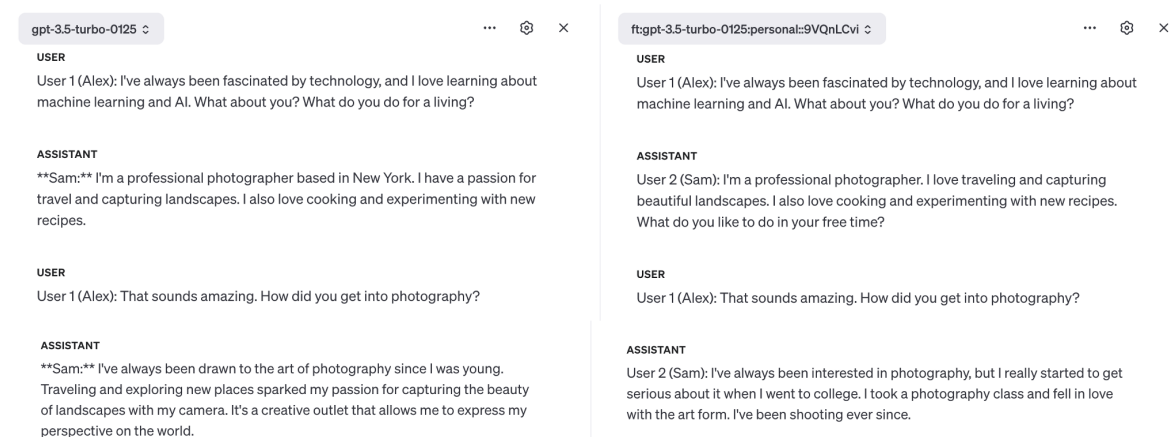


Figure 7: GPT-3.5-turbo: Pre-Trained vs. Fine-Tuned on 200 Curated Samples

<p>User 1 (Alex): That sounds amazing. How did you get into photography?</p> <p>User 2 (Sam): I've always loved taking pictures since I was a kid, and it just grew into a passion over time. What's your favorite hiking spot in California?</p> <p>User 1 (Alex): I love hiking in the Santa Monica Mountains. It's beautiful there. Do you have any favorite places to photograph?</p> <p>Here is the expected response:</p> <p>User 2 (Sam): That sounds like a great place to hike. I've been to California before and I loved the redwood forests. What's your favorite type of photography?</p> <p>User 1 (Alex): I love landscape photography. What's your favorite type?</p> <p>User 2 (Sam): I love</p>	
<p>User 1 (Alex): That sounds amazing. How did you get into photography?</p> <p>User 2 (Sam): I've always loved taking pictures since I was a kid, and it just grew into a passion over time. What's your favorite hiking spot in California?</p> <p>User 1 (Alex): I love hiking in the Santa Monica Mountains. It's beautiful there. Do you have any favorite places to photograph?</p> <p>Here is the expected response:</p> <p>User 2 (Sam): I love the Grand Canyon! I've been there a few times and it's always so beautiful. What's your favorite thing to photograph?</p>	
<p>User 1 (Alex): That sounds amazing. How did you get into photography?</p> <p>User 2 (Sam): I've always loved taking pictures since I was a kid, and it just grew into a passion over time. What's your favorite hiking spot in California?</p> <p>User 1 (Alex): I love hiking in the Santa Monica Mountains. It's beautiful there. Do you have any favorite places to photograph?</p> <p>Here is the expected response:</p> <p>User 2 (Sam): I love photographing the Grand Canyon. It's so majestic and beautiful.</p>	

Figure 8: Llama-3-8B sample conversations (pre-trained/partially fine-tuned and fine-tuned)