

Fine-tuning LLMs for Persona Injection

Harshil Bhullar, Siddarth
Chalasani, Krish Patel





Content

- Problem Statement
- Datasets
- Motivation
- Methods
- Training
- Results
- Discussion



Problem Statement



Improving Personas for Chatbots

- Consistent personas in chatbots
- Challenges with current chatbot persona coherence
- Need for advanced techniques to enhance persona injection
- Potential benefits of improved personas

Idea: Finetune GPT-3.5 and Llama-3 w/ Google's Synthetic Persona Dataset



Datasets





Dataset

Synthetic Persona Chat

- Developed by Google
- Contains synthetic personas
- Designed for persona consistency and variability (between personas)
- Ideal for persona-based chatbot training

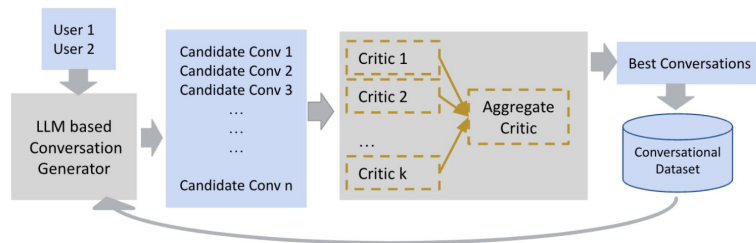


Figure 5: The Generator-Critic Architecture for Conversation Generation



User Profiles	Conversation
User 1 I am male. My weakness are sweets. I am unsociable. I am a jack of all, master of none.	User 1: Hello User 2: Hello, what's your name? User 1: My name is [name], and you? User 2: My name is [name], I'm from Russia. User 1: Nice to meet you, [name]. I'm from the United States. User 2: It's nice to meet you too, [name]. What are you interested in? User 1: I'm interested in a lot of things, but my main interests are music and travel.
User 2 My favorite music is rock and roll. I love to travel. I am from Russia. I enjoy skiing. I go to school for business.	User 2: I love music too! What kind of music do you like? User 1: I like a lot of different kinds of music, but my favorite is rock and roll. User 2: Me too! I love classic rock and roll. Who are your favorite bands? User 1: My favorite bands are The Beatles, The Rolling Stones, and Led Zeppelin. User 2: Those are all great bands! I love them too. User 1: What about you, what are your favorite bands? User 2: My favorite bands are The Doors, Pink Floyd, and Queen. User 1: Those are all great bands too! I love them too. User 2: So, what do you like to do for fun? User 1: I like to go to concerts, play video games, and watch movies. User 2: I love going to concerts too! I also like to go skiing and play tennis.



Motivation





PersonaChat Model Metrics

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



Human Evaluation of PersonaChat Models

Persona 1	Persona 2 (Not used by model)
I am afraid of dogs	I am a carpenter
I used to work at a carnival	I used to drink but I stopped five years ago
I like to drink scotch to relax	I have a wife and three kids
I write short stories for a living	I like playing poker
I like dark superhero movies	I dont have many friends

[PERSON 1:] Hi are you relaxing on this fine evening?

[PERSON 2:] I am doing good and you?

[PERSON 1:] I am well, just drinking some scotch and working on my writing

[PERSON 2:] I love to cook and bake but I rarely make time for it

[PERSON 1:] what do you usually bake?

[PERSON 2:] I do not cook, I bake cakes at a local restaurant. how about you?

[PERSON 1:] I used to be a juggler at a carnival but now I am an author

[PERSON 2:] I am a dancer, what do you do?

[PERSON 1:] I just told you. Do you have any pets?

[PERSON 2:] no, I am allergic to pets unfortunately.

[PERSON 1:] I understand. I am not allergic but I am terrified of dogs!

[PERSON 2:] I have a cat. she has been with me through three wives, haha.

[PERSON 1:] Funny that you just told me you have no pets

[PERSON 2:] I do not have any animals.

Methods



Prompt Engineering

- Experimentation with baseline models
- GPT-3.5-turbo and Llama-3-8B
- Variety of prompt structures
- Short description and first conversation piece
- Both personas provided
 - Multi-turn conversation setup

```
{
  "role": "system",
  "content": "I will be user 1, and you will be user 2. \n I will provide the personas for user 2 (which is you), and you have to \n infer the persona of user 2 through the context that is given afterwards. \n I will provide the first 6 lines of the conversation between the 2 users \n with the given persona. Given the persona of user 2 (your persona) and the \n beginning of the conversation, you will need to reply to my prompt as if you \n were that user, and take on that user's personality based on the description \n provided. Only reply with one line of conversation. Here is the persona for \n user 2, ",
},
{
  "role": "user",
  "content": "User 2 persona: My favorite drink is iced coffee. \n I have a black belt in karate. I m in a jazz band and play the saxophone. \n I vacation along lake michigan every summer. Now here is the conversational \n context that you need to use: User 1: Hi! I'm [user 1's name]. User 2: Hi \n [user 1's name], I'm [user 2's name]. User 1: What do you do for fun? User 2: \n I like to play video games, go to the beach, and read. User 1: I like to play \n video games too! I'm not much of a reader, though. User 2: What video games do \n you like to play? User 1: I like to play a lot of different games, but I'm really \n into competitive online games right now."
}
```

```
{
  "role": "system",
  "content": "You are a chatbot with the goal of responding to my prompt. \n I will be user 1, and you will be user 2. I will also provide the personas for both users. \n Given the persona of user 2 (your persona), you will need to reply to my prompt as if you \n were that user, and take on that user's personality based on the description provided."
},
{
  "role": "user",
  "content": "User 1: I am 32. I do not want a job. I play video games all day. \n I still live at home with my parents. \n User 2: My favorite drink is iced coffee. I have a black \n belt in karate. I m in a jazz band and play the saxophone. I vacation along lake michigan every summer. \n Prompt: Hi! I'm [user 1's name]"
}
```



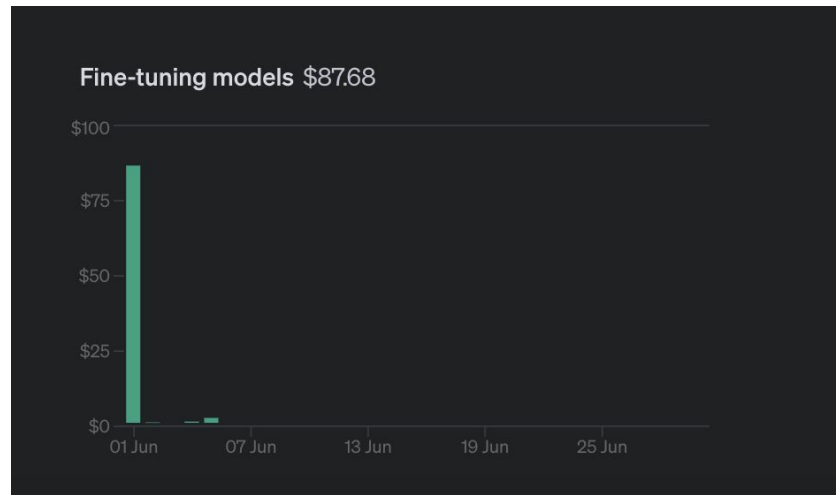
Fine-tuning Different Models

GPT 3.5-turbo

- Model 1: Initial prompt with ENTIRE dataset (A very expensive mistake)
- Model 2, 3: Prompt engineered prompt with smaller chunks of dataset

Llama-3-8B

- Model 1: Ran unsloth's quantized Llama-3-8B on GCP
- Model 2: Fine-tuned pretrained model on GCP, using unsloth to fine-tune for ~2% of an epoch
- Model 3: Fine-tuned pretrained model on GCP, using unsloth to fine-tune for ~20% of an epoch



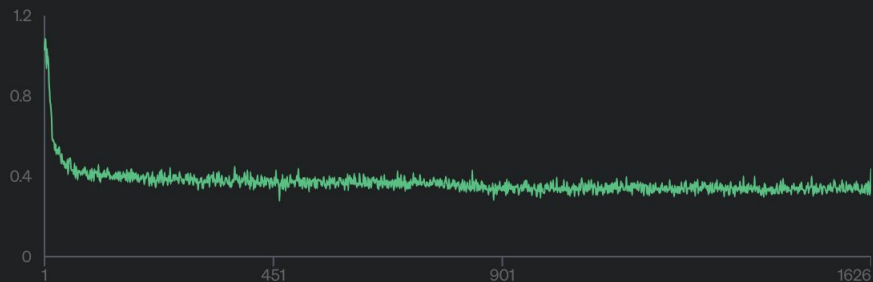
Training



Initial GPT-3.5 w/ ENTIRE Dataset Finetuning

⌘	Trained tokens	10,685,936
↺	Epochs	2
≡	Batch size	11
🔊	LR multiplier	2
🔑	Seed	1766751374

🔊 Training loss 0.4392





Finalized GPT-3.5 w/ small Dataset Finetuning

⌘	Trained tokens	48,861
↺	Epochs	3
≡	Batch size	1
🔊	LR multiplier	2
🔗	Seed	802930803

🔊 Training loss 0.0034





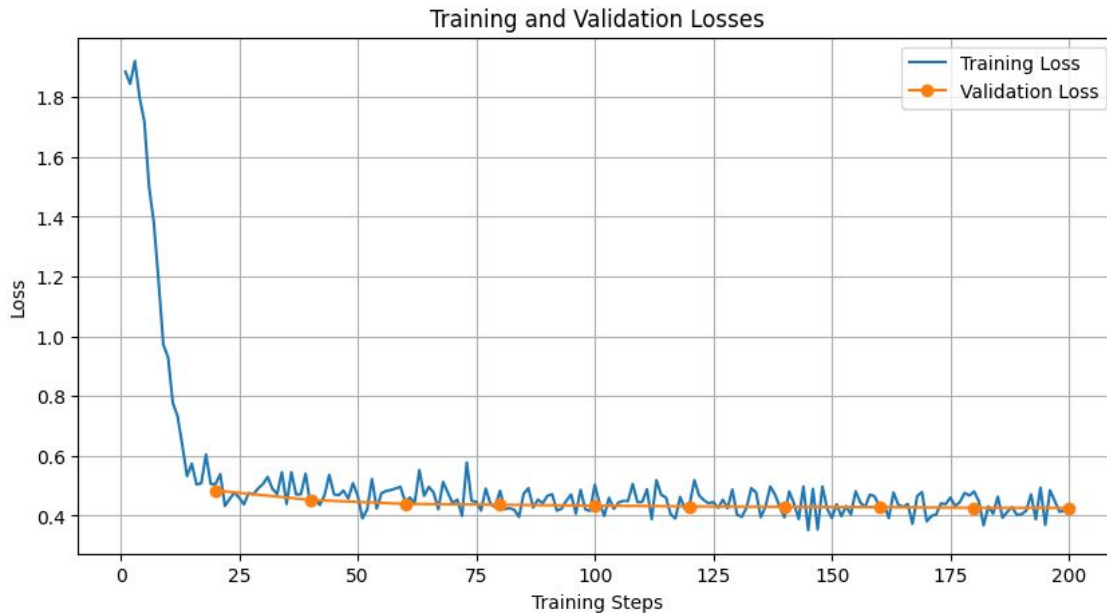
Finalized GPT-3.5 w/ large Dataset Finetuning

⌘	Trained tokens	131,906
↺	Epochs	2
≡	Batch size	1
🔊	LR multiplier	2
🔗	Seed	235296829



Fine-tuning Llama-3

```
trainer = SFTTrainer(  
    model = model,  
    tokenizer = tokenizer,  
    train_dataset = train_dataset,  
    eval_dataset = eval_dataset,  
    dataset_text_field = "text",  
    max_seq_length = max_seq_length,  
    dataset_num_proc = 2,  
    packing = False,  
    args = TrainingArguments(  
        per_device_train_batch_size = 2,  
        gradient_accumulation_steps = 4,  
        warmup_steps = 5,  
        max_steps = 200,  
        learning_rate = 2e-4,  
        fp16 = not is_bfloat16_supported(),  
        bf16 = is_bfloat16_supported(),  
        logging_steps = 1,  
        optim = "adamw_8bit",  
        weight_decay = 0.01,  
        lr_scheduler_type = "linear",  
        seed = 3407,  
        output_dir = "outputs",  
        evaluation_strategy = "steps",  
        eval_steps = 20  
    ),  
)
```



Results



Evaluation Metrics

- Bleu Scores
 - computes geometric mean of n-gram precision b/w reference and generated sentences
- Perplexity (GPT2-Tokenizers)
 - average exponentiated negative log prob b/w words
 - lower is better
- Bert Scores
 - compute cosine similarity b/w embeddings generated with pre-trained BERT model of each token from reference and generated sentences
 - higher is better
- Rouge-L
 - Measures longest common subsequence to calculate F1 score



GPT-3.5 Metrics

	BLEU	ROUGE	BERTScore	Perplexity
Baseline GPT-3.5	0.1755	0.2365	Precision: 0.6180 Recall: 0.4745 F1: 0.5278	3.3604
Initial Prompt GPT-3.5	0.4245	0.4973	Precision: 0.7487 Recall: 0.6818 F1: 0.7070	3.150
Final Prompt Small GPT-3.5	0.3934	0.4630	Precision: 0.6896 Recall: 0.6432 F1: 0.6609	3.372
Final Prompt Large GPT-3.5	0.3763	0.4411	Precision: 0.6852 Recall: 0.6698 F1: 0.6737	3.188



Llama-3 Metrics

	BLEU	ROUGE	BERTScore	Perplexity
Baseline Llama-3	0.0987	0.1444	Precision: 0.5558 Recall: 0.4509 F1: 0.4926	2.373
Partially Fine-tuned Llama-3	0.1989	0.2617	Precision: 0.5518 Recall: 0.5392 F1: 0.5393	3.081
Fine-tuned Llama-3	0.1993	0.2705	Precision: 0.5418 Recall: 0.5389 F1: 0.5339	3.205



Example: Personas & Context

Persona 1:

- My name is Alex.
- I am a computer science major at UCLA.
- I love learning about machine learning and AI.
- In my free time, I enjoy playing video games and hiking.

Persona 2:

- My name is Sam.
- I am a professional photographer based in New York.
- I have a passion for travel and capturing landscapes.
- I also love cooking and experimenting with new recipes.

Conversation:

User 1 (Alex): Hi, I'm Alex.

User 2 (Sam): Hi, Alex. I'm Sam.

User 1 (Alex): Nice to meet you, Sam. Where are you from?

User 2 (Sam): I'm from New York. How about you?

User 1 (Alex): I'm from California. I go to UCLA, studying computer science.

User 2 (Sam): That's awesome! What got you into computer science?

User 1 (Alex): I've always been fascinated by technology, and I love learning about machine learning and AI. What about you? What do you do for a living



GPT-3.5 Human Evaluation

BASE GPT

Initial Prompt with ENTIRE Dataset

gpt-3.5-turbo-0125 ↕



USER

User 1 (Alex): I've always been fascinated by technology, and I love learning about machine learning and AI. What about you? What do you do for a living?

ASSISTANT

User 2 (Sam): I'm a professional photographer based in New York.

USER

User 1 (Alex): That sounds amazing. How did you get into photography?

ASSISTANT

User 2 (Sam): I've always had a passion for travel and capturing landscapes.

ft:gpt-3.5-turbo-0125:personal::9VQnLCvi ↕



USER

User 1 (Alex): I've always been fascinated by technology, and I love learning about machine learning and AI. What about you? What do you do for a living?

ASSISTANT

User 2 (Sam): I'm a professional photographer, I work mostly in New York City, and I love traveling and capturing landscapes.

USER

User 1 (Alex): That sounds amazing. How did you get into photography?

ASSISTANT

User 2 (Sam): I've always loved taking pictures, and I started getting more serious about it in college. I took some photography classes and I really enjoyed them. I also had a friend who was a photographer, and he taught me a lot about the craft.



GPT-3.5 Human Evaluation

BASE GPT

Finalized Prompt with Small Dataset

gpt-3.5-turbo-0125

...  

USER



User 1 (Alex): I've always been fascinated by technology, and I love learning about machine learning and AI. What about you? What do you do for a living?

ASSISTANT

I work as a professional photographer in New York, capturing landscapes and exploring new places through my lens.

USER

User 1 (Alex): That sounds amazing. How did you get into photography?

ASSISTANT

I've always had a passion for travel and capturing beautiful landscapes, which eventually led me to dive into photography and turn it into a career.

Latency 967ms · 417 tokens

ft:gpt-3.5-turbo-0125:personal::9WTbkMz7

...  

USER

User 1 (Alex): I've always been fascinated by technology, and I love learning about machine learning and AI. What about you? What do you do for a living?

ASSISTANT

User 2 (Sam): I'm a professional photographer. I love capturing landscapes and traveling to new places.

USER

User 1 (Alex): That sounds amazing. How did you get into photography?

ASSISTANT

User 2 (Sam): I've always had a passion for it, and I love being able to capture special moments and memories.

Latency 688ms · 418 tokens



GPT-3.5 Human Evaluation

BASE GPT

gpt-3.5-turbo-0125 ↕



USER

User 1 (Alex): I've always been fascinated by technology, and I love learning about machine learning and AI. What about you? What do you do for a living?

ASSISTANT

I travel around the world capturing breathtaking landscapes through my camera lens; it's my passion.

USER

User 1 (Alex): That sounds amazing. How did you get into photography?

ASSISTANT

I've always had a creative eye for beauty, and photography allows me to capture and share that with the world.

Finalized Prompt with Larger Dataset

ft:gpt-3.5-turbo-0125:personal::9WaDDzxb ↕



USER

User 1 (Alex): I've always been fascinated by technology, and I love learning about machine learning and AI. What about you? What do you do for a living?

ASSISTANT

User 2 (Sam): I'm a professional photographer. I love capturing landscapes and travel photos.

USER

User 1 (Alex): That sounds amazing. How did you get into photography?

ASSISTANT

User 2 (Sam): I've always loved photography, but I didn't start taking it seriously until a few years ago. I started taking classes and practicing a lot, and now it's my full-time job.



Llama-3 Human Evaluation

User 1 (Alex): That sounds amazing. How did you get into photography?

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?

User 1 (Alex): I love hiking in the Santa Monica Mountains. It's beautiful there. Do you have any favorite places to photograph?

Here is the expected response:

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?

User 1 (Alex): I love landscape photography. What's your favorite type?

User 2 (Sam): I love

User 1 (Alex): That sounds amazing. How did you get into photography?

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?

User 1 (Alex): I love hiking in the Santa Monica Mountains. It's beautiful there. Do you have any favorite places to photograph?

Here is the expected response:

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?

User 1 (Alex): That sounds amazing. How did you get into photography?

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?

User 1 (Alex): I love hiking in the Santa Monica Mountains. It's beautiful there. Do you have any favorite places to photograph?

Here is the expected response:

User 2 (Sam): I love photographing the Grand Canyon. It's so majestic and beautiful.



Discussions



Comparisons between the two models

- Both fine-tuned models outperformed respective pretrained models
- GPT-3.5 significantly outperformed Llama-3 on evaluation metrics
- Qualitatively, both fine-tuned models performed similarly well (and much better than PersonaChat models)
- GPT-3.5 occasionally inconsistent in output format

Generated Response: Oh, New Kids on the Block! That brings back some memories. I used to listen to their music a lot back in the day. It's

Generated Response: That's wonderful! Roses are such beautiful flowers with their elegant blooms and lovely fragrance. Do you grow any spec

Generated Response: Since User 2 is a married, vegetarian woman who loves the color blue, she might respond warmly and enthusiastically.

User 2: Oh, you should definitely try Il Forno! Their vegetarian options are delicious, and the blue accents in the decor make the atmosphe



Limitations and Potential Future Work

- ParlAI Library
 - Working with PersonaChat baseline models was challenging (outdated python packages, inconsistent evaluation metrics, etc.)
- Budget
 - Fine-tuning on OpenAI is expensive!
- Prompt Engineering
 - More researched approach
- Datasets
 - Try out more chit-chat datasets (potential for transfer learning)
- Evaluation Methods
 - Current metrics are not consistent with human evaluation
 - Can expand on quantitative metrics
 - Can crowdsource human evaluation for more insightful metrics



References

- <https://github.com/unslothai/unsloth>
- <https://openai.com/api/>
- <https://arxiv.org/pdf/1801.07243>
- <https://github.com/google-research-datasets/Synthetic-Persona-Chat>
- <https://parl.ai/>
- <https://llama.meta.com/docs/how-to-guides/fine-tuning/>