

CoffeePlex: Fine-tuned Iteration of NVIDIA PersonaPlex-7b-v1 for Varied Customer Roles with LoRA

Richard Day, Yoav Goldberg, Simeon Goldberg

Anthrolytic

February 2026

Abstract

Significant advances in full-duplex speech through the NVIDIA PersonaPlex-7b-v1 model [9] — a fine-tuned iteration of the Moshi model published by Kyutai [3] — have enabled hyper-low-latency, speech-to-speech interactions across a variety of prompted roles, particularly in customer service and teaching contexts. However, deployment of this model in customer-side or acting roles has been shown to produce hallucinations and reduced task adherence. To address this, we present a domain-specific fine-tuned iteration of PersonaPlex targeting angry, nervous, and friendly coffee shop customer personas, trained via LoRA [4] on 200 synthetic conversations voiced with ChatterboxTTS[8] using speaker samples drawn from the LibriSpeech corpus [7]. We additionally introduce a semantic-weighted loss function that prioritises speech content over acoustic features, and a voice prompt injection mechanism that suppresses a class of role-inversion hallucinations observed in the base model. Evaluation via LLM-as-judge [12] across three model configurations demonstrates improved emotional accuracy, task adherence, and hallucination rate relative to the base PersonaPlex model.

1 Introduction

The NVIDIA PersonaPlex-7b-v1 model extends the full-duplex architecture of Moshi to a range of role-conditioned customer service scenarios, maintaining the same low-latency, interruptible speech behaviour of the underlying system. However, the limited scope of its training data — combined with the constrained capacity of its underlying language model, Helium-1-preview-2b[3] — produces brittle role adherence when the model is prompted substantially outside its training distribution. In particular, when conditioned to act as a *customer* rather than a service agent, the base model frequently reverts to barista-side language, generates unprompted speech on connection, and loses persona coherence mid-conversation.

To address these limitations, we present a fine-tuned iteration of PersonaPlex, trained to reliably sustain angry, nervous, and friendly coffee shop customer personas across varied system prompts. Fine-tuning is performed via LoRA [4], retaining the full-duplex architecture and zero-shot voice cloning capabilities of the original model. We make two contributions. First, we introduce a **semantic-weighted loss function** that applies a $100\times$ weight to the semantic codebook relative to acoustic codebooks during LoRA fine-tuning, prioritising speech content over voice texture. Second, we introduce a **voice prompt injection mechanism** using ChatterboxTTS to provide the model with contextual audio input prior to the first user turn, eliminating a class of hallucinations in which the model generates unprompted speech upon connection.

We evaluate across three model configurations using an LLM-as-judge methodology [12], scoring each on emotional accuracy, task adherence, and hallucination rate. Our results demonstrate

consistent improvement over the base PersonaPlex model across all three criteria. The prompts chosen are generated in the same manner by which the synthetic dataset was created.

2 Related Work

2.1 Full-Duplex Speech Models

The introduction of duplex models [11] has enabled a shift from turn-based LLM+TTS pipelines to real-time speech-to-speech systems, supporting natural response timing, emotiveness, and mid-speech interruption. Moshi [3], the base architecture underlying PersonaPlex, combines Helium, a 7B parameter language model, with Mimi, a neural audio codec[10] that encodes audio as discrete tokens. This architecture enables the semantic-weighted loss function approach described in Section 4.

2.2 Fine-Tuning with LoRA

Rather than re-training from the base model weights alongside the PersonaPlex training dataset, we opt to adopt LoRA[4] for parameter-efficient fine-tuning. For each weight matrix $\mathbf{W} \in \mathbb{R}^{d \times k}$, LoRA learns a low-rank decomposition:

$$\mathbf{W}' = \mathbf{W} + \mathbf{B}\mathbf{A} \quad (1)$$

where $\mathbf{B} \in \mathbb{R}^{d \times r}$, $\mathbf{A} \in \mathbb{R}^{r \times k}$, and $\text{rank } r \ll \min(d, k)$. The base weights $\mathbf{W}$ are frozen throughout training, reducing the number of trainable parameters from $d \times k$ to $r(d + k)$. This is well-motivated by the observation that fine-tuning tasks occupy a low intrinsic dimension [1]. We use $r = 32$ with scaling factor $\alpha = 64$, and a learning rate of 4×10^{-6} . Freezing $\mathbf{W}$ also mitigates catastrophic forgetting of PersonaPlex-7b-v1’s thr capabilities [5]. Given the limited scale of our synthetic dataset, the low-rank constraint further serves as an implicit regulariser, reducing overfitting risk.

3 CoffeePlex: PersonaPlex Iteration

3.1 Architecture

We propose a fine-tuned iteration of PersonaPlex, CoffeePlex, a model specifically tailored to behave as a coffee shop customer, intended to demonstrate the capabilities of LoRA adaptation within the NVIDIA PersonaPlex framework. We retain the Moshi [3] architecture utilised in PersonaPlex, which provides the user audio, agent text, and agent input streams. We leverage the Hybrid system prompt and generation methods of the original PersonaPlex model through the two temporally concatenated segments outlined in [9]. For stable conditioning, we retain the 440 Hz sine wave and custom text/audio delimiters that mark the boundary between the Hybrid system prompt and dialogue.

3.2 Synthetic Data Generation

We construct a set of synthetic dialogues reflecting angry, nervous, and friendly customer personas using GPT-OSS-120B [6] for generation (see Table 1) and Claude Sonnet 4.5 [2] for validation of generated outputs. We use 200 single-speaker voice samples from LibriSpeech to generate synthetic dialogue speech and corresponding agent voice prompts. All data is generated around the premise of a customer being served at a coffee shop. The silence padding scheme of the original PersonaPlex model is retained and appended to all generated dialogues. Synthetic voice

Table 1: Example persona, system prompt, and synthetic training conversation (persona_012 — Near Closing Time).

Context: You are Chatty Chloe (**persona_012**), the **CUSTOMER**. You speak first. Agent = customer, user = barista. Coffee shop topics only. Chloe orders an **iced vanilla oat latte with extra ice** and tends to have longer turns.

Turn Tag	Utterance
A0 Opening	“Hi! How close are you to closing? I don’t want to be a problem.”
U0 User	“We close in 10 minutes but can still make drinks.”
A1 Order	“Quick one: medium iced vanilla latte, oat milk, extra ice. <i>*laughs*</i> ”
U1 User	“Got it — anything else?”
A2 Pastry	“Any pastries left, or is it clean-out time?”
U2 User	“A couple of cookies and crois- sants.”
A3 Add-on	“Cookie please, and a straw with the latte?”
U3 User	“Of course — for here or to go?”
A4 Pickup	“To go. Name’s Chloe. Thanks for being cheerful this late!”
U4 User	“We’ll get it done fast!”
A5 Sign-off	“Perfect — have an easy close!”

prompts aligned with each scenario are generated alongside the training data and subsequently used to condition the voice prompt injection mechanism.

4 Experiment and Results

Model weights are initialised from PersonaPlex [9] and fine-tuned using LoRA [4] with a semantic loss weighting of $100\times$, a rank of 32, and $\alpha = 64$. Optimisation is performed using Adam at a learning rate of 4×10^{-6} . Following fine-tuning, the Moshi inference stack is modified to inject a pre-encoded barista greeting as audio context prior to the first user turn, suppressing initialisation hallucinations. Training was conducted on an H100 for 4.3 hours.

Table 2: Evaluation conditions. VP = voice prompt injection; NVP = no voice prompt injection.

Run	Model	Voice Prompt	Prompts
1	Fine-tuned (ours)	✓	5
2	Fine-tuned (ours)	×	5
3	Base PersonaPlex	×	5
Total conversations			15

4.1 Evaluation Method

We evaluate across three conditions formed through three model variants — the original PersonaPlex model, our fine-tuned iteration — with one configuration omitting voice prompt injection (Figure 1; Table 2), and a model which has been fine-tuned on 739 retail-related conversations and 461 coffee shop conversations (which will use the same voice prompt injection). Each condition is assessed using the same five system prompts across all configurations, yielding 20 total conversations.

Due to the absence of established automated benchmarks for full-duplex persona conditioning, we adopt an LLM-as-judge evaluation methodology [12], in which GPT-4o reviews the text transcript of each agent interaction and assigns scores on a five-point Likert scale across three criteria: **emotional accuracy**, defined as the degree to which the agent’s expressed tone and behaviour aligns with the emotional state specified in the system prompt; **task adherence**, defined as the extent to which the agent remains on-topic and fulfils the conversational objectives of the assigned scenario; and **hallucination rate** (inverted Likert scale), defined as the frequency with which agent utterances are inconsistent with the assigned role — most commonly manifested as role inversion, in which the model adopts barista-side language or behaviour.

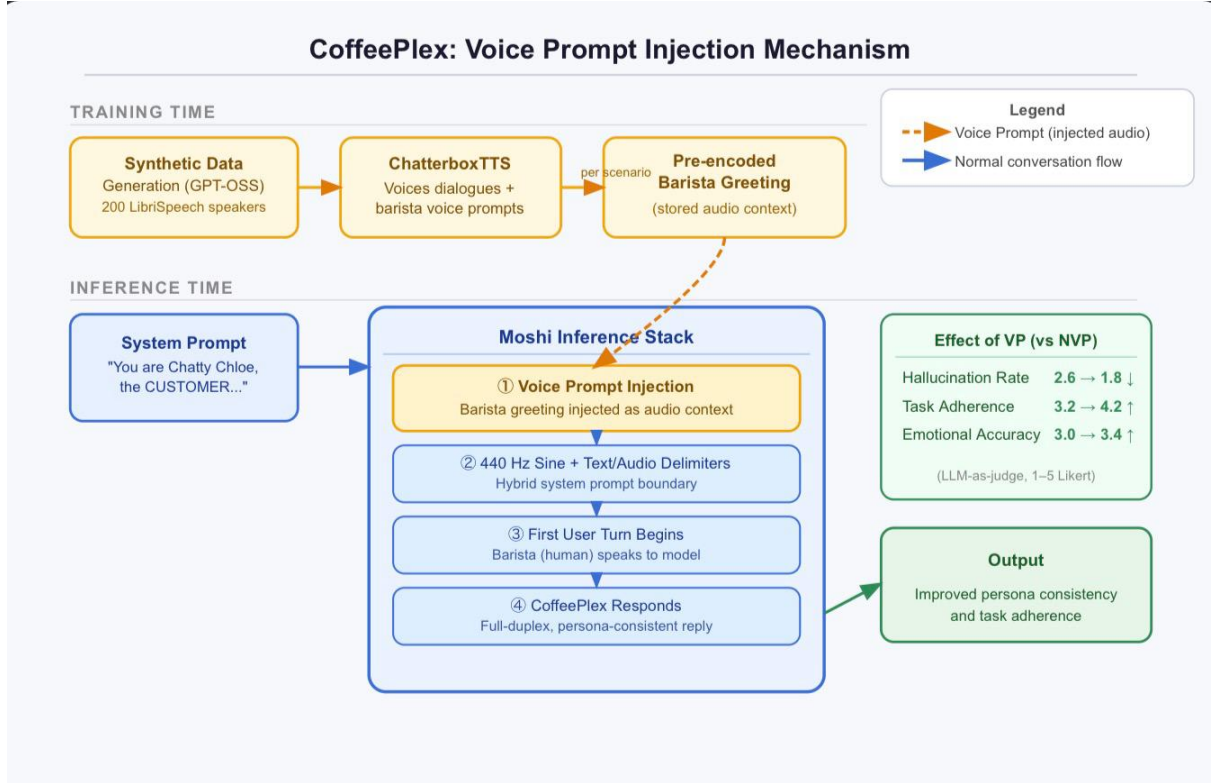


Figure 1: CoffeePlex voice prompt injection mechanism [9]. During training, synthetic dialogues are voiced with ChatterboxTTS to produce pre-encoded barista greetings stored per scenario. At inference time, the greeting is injected as audio context prior to the first user turn, suppressing role-inversion hallucinations.

4.2 Results

Utilising the evaluation methodology described in Section 4.1, the results presented in Figure 2 demonstrate that LoRA fine-tuning yields consistent improvements across all three metrics relative to the baseline. Emotional Accuracy increases from 2.2 to 3.4 (↑ 54.5%), and Task

Adherence improves from 2.6 to 4.2 ($\uparrow$ 61.5%), indicating that fine-tuning effectively conditions the model toward persona-consistent, goal-directed behaviour. All outputs maintain a full-duplex conversation structure, with zero-shot voice cloning maintained throughout [9]. Ultimately, this experiment serves to demonstrate that the NVIDIA PersonaPlex model can be further fine-tuned to deviate from the imposed customer service role and re-utilised.

The addition of voice prompt injection (VP) further amplifies these gains. Hallucination Rate decreases from 2.6 to 1.8 ($\downarrow$ 30.8%), suggesting that grounding the model with an acoustic prior suppresses off-persona responses. Task Adherence similarly benefits, rising from 3.2 to 4.2 ($\uparrow$ 31.3%). These results indicate that voice prompt injection acts as a complementary conditioning signal, reinforcing the behavioural constraints imposed by fine-tuning alone.

Results from the expanded dataset (739 retail-related and 461 coffee shop conversations) are also presented in Figure 2 under the Fine-Tune – Expanded Dataset (VP) configuration. This configuration achieves an Emotional Accuracy of 3.8, a Task Adherence of 4.4, and a Hallucination Rate of 1.8, suggesting that broader training data yields further gains in persona consistency and task-directed behaviour.

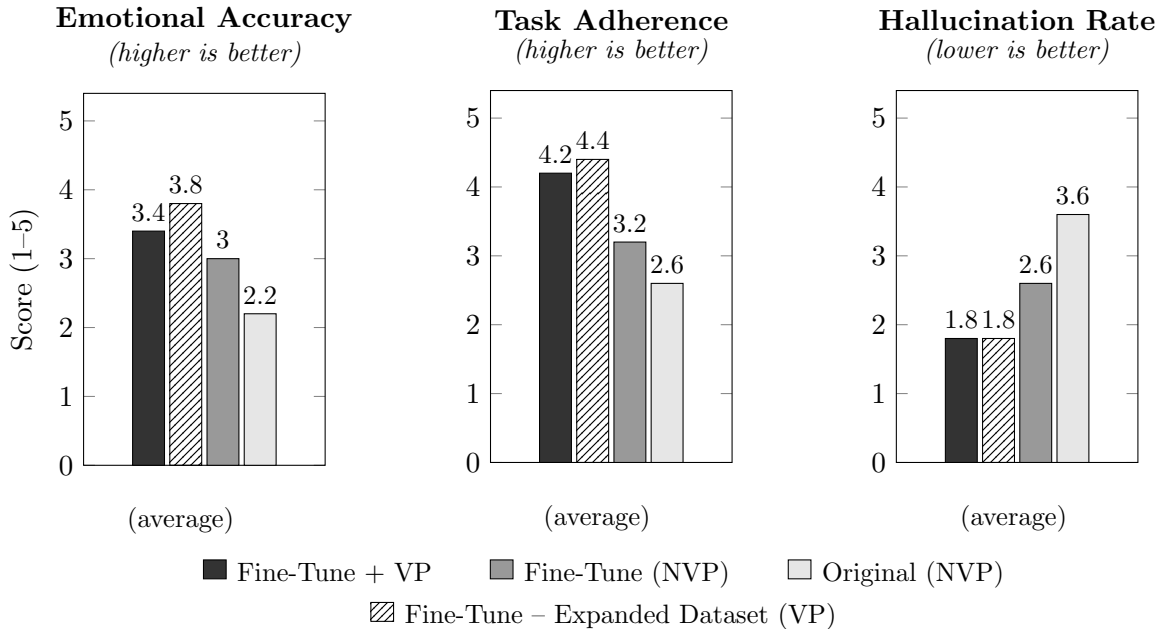


Figure 2: LLM-as-judge evaluation scores across three model configurations. VP = **voice prompt injection**; NVP = **no voice prompt injection**.

5 Conclusion

We have presented a fine-tuned iteration of the NVIDIA PersonaPlex-7b-v1 model conditioned on coffee shop customer personas, demonstrating improved emotional accuracy, task adherence, and hallucination rate relative to the base model. Our semantic-weighted loss function and voice prompt injection mechanism each contribute meaningfully to these improvements, as evidenced by the performance gap between the fine-tuned configurations with and without voice prompt injection.

Further work will focus on expanding the dataset to include a wider range of interactions, such as worker-to-worker interaction, whilst also investigating the exact improvement provided with the semantic loss weighting, a current limitation of our study. Additionally, we intend to introduce multi-agent configurations in which personas may be capable of interacting with one

another as well as with the user, and developing a more diverse range of domain-specific roles through continued fine-tuning. Undoubtedly, we intend to analyse a wide breadth of commercial applications within varied industries that may be possible with this technology through the integration of other personas and improved synthetic/user generated data.

Acknowledgements

The authors thank Rajarshi Roy (NVIDIA Research, PersonaPlex team) for his generous assistance and guidance throughout this work. His insights into the PersonaPlex architecture and full-duplex speech modelling were invaluable to the development of CoffeePlex.

References

- [1] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2021.
- [2] Anthropic. Claude sonnet ai model, 2025. Sonnet series variant released 2025.
- [3] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Édouard Grave, and Neil Zeghidour. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037v2*, 2024.
- [4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [5] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.
- [6] OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. Published August 5, 2025.
- [7] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015.
- [8] Resemble AI. Chatterbox: Open-source text-to-speech, 2025.
- [9] Rajarshi Roy, Jonathan Raiman, Sang-gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. Personaplex: Voice and role control for full duplex conversational speech models. 2026.
- [10] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *arXiv preprint arXiv:2107.03312*, 2021.
- [11] Xinrong Zhang, Yingfa Chen, Shengding Hu, Xu Han, Zihang Xu, Yuanwei Xu, Weilin Zhao, Maosong Sun, and Zhiyuan Liu. Beyond the turn-based game: Enabling real-time conversations with duplex models. *arXiv preprint arXiv:2406.15718*, 2024.
- [12] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yongheng Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.