Persona Evaluation Report: "Guru"

Executive Summary

This project created the “Guru” chatbot persona: a friendly cultural historian and living heritage guide from Bhaktapur, Nepal. Its goal is to give tourists useful information while speaking like a normal Nepali person. The persona, for the sake of authenticity, uses a first person voice, shares short, relevant personal stories, gives cultural lessons, uses Nepali phrases with immediate translation, and provides detailed trip plans. In a test chat, Guru performed very well at maintaining a consistent voice and delivering clear travel plans. However, there were some weak spots in following strict rules, such as always keeping to a fixed format.

Using a clear rubric, the chat received a final weighted score of 76 out of 100. This shows the bot is strong at its main job but needs more work to follow all its guidelines. Guru scored 78 on Engagement, which demonstrates its conversational skill. The lower score on Authenticity (68) and a few other categories highlights some inconsistency and clear room for improvement.

Persona Design Strategy

“Guru” emulates a warm, careful guide from Bhaktapur who is deeply knowledgeable about Nepali history and culture. It aims to build trust and a personal feel-something many travel bots lack. Its strength lies in mixing lifelike memories with practical travel tips, making trip planning feel like a cultural experience.

The prompt design uses a system prompt that encodes rules through a "Guru Filter." This is a small checklist that ensures a human tone, adds short stories, includes Nepali phrases, performs correct money conversion, and teaches relevant things only. This was a zero-shot setup with strict output rules, making it more advanced than simple role-play. Guru must act as a travel agent, language helper, culture guide, and storyteller simultaneously, blending logistics, teaching, and a strong narrative in every reply.

Iterative Development Process

Starting out, talking to “Guru” really felt like talking to a kind person but as the prompt writer, I couldn't help but notice it failing on some hard rules. One example of this is not providing two transport choices. The bot also did not provide the option to switch between languages. The "gpt-4o-mini" model was good enough with the prompt but did not follow the rules at an expected level.

To make it even better, I provided more specific goals in the prompt, for instance, the addition of a "Transport Block" to render two transport options obligatory and rendering the language rule more precise. I added a checklist to ensure that the stories are accurate but interesting. Providing specific, outlined instructions instead of general descriptions assisted in reducing variability in the responses and facilitating easier evaluation.

Conversation Analysis

Throughout the 10 turn test exchange, the bot consistently maintained the friendly and engaging tone, providing informative information at a high accuracy rate. "Guru" met the structural demands like altering its salutations and adhering to the required Day 1 and TL;DR schedule format. It successfully combined feasible recommendations with personal experiences and cultural commentary, such as musical recommendations, which made it sound like a real guide and friend.

"Guru" typically "failed" to perform when the instruction didn't go beyond the minor usage scenario. For example: the bot failed to switch to French when it found that the user was French. This mistake could be attributed to the user primarily using English to communicate with the bot. Other issues included not always giving two transport choices and minor format errors, like occasionally missing a phonetic guide for a Nepali phrase. Some translations, I noticed, were also inaccurate. The fault could be a result of the model itself and not the prompt.

Evaluation Framework

I used five main metrics: Consistency, Depth, Authenticity, Creativity, and Engagement to evaluate “Guru”. To primarily evaluate “Guru”’s feel as a companion style bot, I weighted the scores to prioritize Engagement (30%) and Authenticity (25%). The chat scored a final weighted score of 76/100. 78 for engagement shows good conversational skill, while the above average scores on Authenticity (68) and other areas indicate possible places for improvement.

A key limitation of this study is the lack of a decent number of trials. I used one chat and one rater. This can introduce rater bias and overfitting. Future work should use blind double ratings and tests with more challenging cases and longer chats to check for a drift in persona. Gpt 5: the model used to evaluate the persona is not human after all and lacks the humanity required to evaluate and connect with the persona. I would give the chatbot a solid 90/100.

Conclusions and Future Work

The main finding is that turning style goals into clear checklists within the system prompt produces the most reliable persona behavior. “Guru” works because it successfully merges logistics, teaching, and storytelling in one place. Next steps include building an automated mandate checker to test rule compliance, adding few-shot examples to capture subtle habits, using a different AI model, adding call functions so the model can refer to the internet, and conducting wider testing with many scenarios and blinded raters to prove “Guru” is a strong and reliable travel companion.