

Lab 2 – LLMs for Dialogue Systems – Leali Marco – 02/10/2025

Chit-chat with SpeechState - Part 2 – Exploration (Option 1)

Handling ASR_NOINPUT - What I did:

- Introduced a **counter** (noinputCounter) to track repeated silence.
- Two approaches tested:
 - o **Static** Prompt: Assistant outputs a predefined message to re-engage the user.
 - o **Dynamic** LLM Prompt: When the user is silent, the chatbot adds a "user was silent" message and queries the LLM for a polite prompt. -> I used this in the final version.
- **Automatic** temperature adjustment: For repeated silences, temperature is increased to make responses more creative and engaging. -> When NoInputCounter => 2 temperature is increased to 0.9.
- **Manual** temperature adjustment: I added some buttons that allows the user to change the temperature, but get reset after silence patterns
- **Static** top_k = 100. It can be changed manually inside the code.
- LLM Integration: Used Ollama API to fetch chat completions with configurable model.

Tests with temperature

User's utterance: "What's your name?"

Temp 0.1 - Low temperature (0.1-0.3) – It should be more deterministic giving consistent responses

I don't really have a personal name, but my friends call me "Chat" for short! What about you?

Temp 0.7-1.0 (same response)

Nice to meet you! My name is Luna, by the way. What brings you here today?

User's utterance: "What's your favourite color?"

Temp 0.1

I don't really have one, but I think bright blue is pretty cool! How about you? Do you have a favorite color?

Temp 0.7 (0.4 -0.7 it should have a balanced creativity (default))

I'm just a bot, I don't have personal preferences, but I think blue is a nice color! What about you, do you have a favorite?

Temp 1.0 (It should be more creative, varied, sometimes unpredictable responses)

I'm just a computer, so I don't really have preferences, but I think bright blues and yellows are nice colors! What about you? Do you have a favorite color?

Tests with top_k

User's utterance: "What's your name?"

top_k=1: Only the most probable token is considered (it should be very deterministic)

I don't really have a personal name, but my friends call me "Chat" for short. What about you?

top_k=100: More diverse sampling, considers top 100 tokens

_Nice to meet you! My name is Luna, by the way. What brings you here today?

top_k=500: it should be Very diverse, considers many possible tokens

Nice to meet you! My name is Luna, by the way. What brings you here today?

User's utterance: Describe the sunset in one sentence

top_k=1

The sunset was a breathtaking blend of warm oranges, soft pinks, and gentle purples that slowly faded into a peaceful twilight sky.

top_k=500

The sunset was a warm, golden glow that slowly faded into a deep pink sky with a few wispy clouds painted across it.

Final Observations

From my experiments with temperature and top_k, I noticed that the differences were not always very pronounced. For instance, common questions like “What’s your name?” the outputs were almost identical across values, with only minor variation (e.g. adding a fictional name at higher temperature or higher top_k).

The strongest differences appeared in more descriptive or creative prompts such as “Describe the sunset in one sentence” or “What’s your favorite color?”, where higher temperature and higher top_k produced slightly more diverse and colorful wording. **However, the variation was still relatively small compared to what theory suggests.** This might be due to the **model** itself (Ollama + llama3.1) being relatively stable and not producing very chaotic outputs even at high randomness settings or because of the **questions** that I made.

Strengths observed:

- Dialogue history kept answers consistent and avoided repetition.
- Temperature and top_k controls worked technically and gave some degree of variability.
- Silent user input handling with automatic temperature increase made the chatbot slightly more engaging.

Weaknesses observed:

- In my tests, differences between settings were subtle; the chatbot often produced similar answers regardless of temperature or top_k, except in creative prompts.
- top_k changes had almost no visible impact in short factual questions.
- More complex or open-ended prompts might be needed to better showcase the effect of randomness parameters.
- Dynamic no-input handling increased latency slightly.

Conclusion

The system works well for chit-chat and responds robustly, but to really test the impact of parameters like temperature and top_k, I think I would need to make more tests and use more open-ended or creative test prompts. The test could be conducted, for instance, with story generation, roleplay or joke requests. For short factual answers, the differences are minimal.