

Prompting Large Language Models

— some first steps —

Peter Steinbach

Helmholtz-Zentrum Dresden-Rossendorf, Department for Information Services and Computing,

HZDR AI Symposium, September 9, 2025

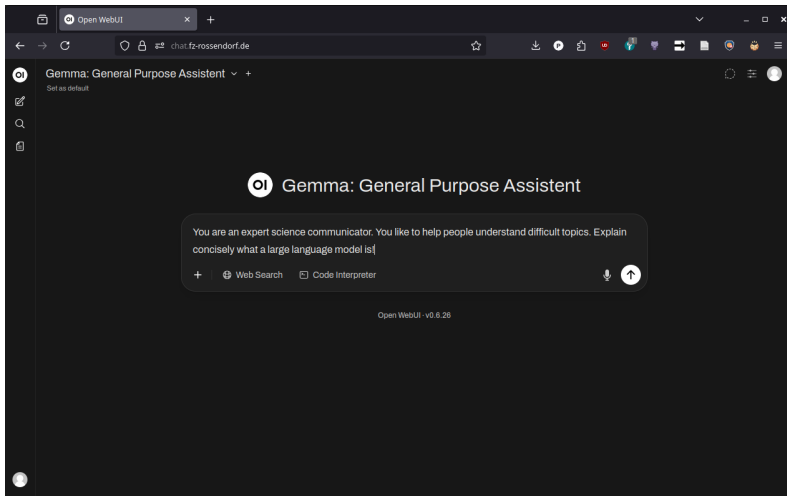
Table of Contents

- 1 Dialogues in LLMs
- 2 The way you ask, the way it replies
- 3 Summary

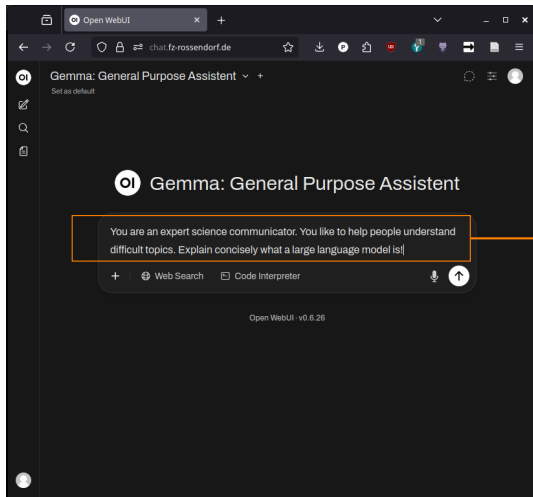
Table of Contents

- 1 Dialogues in LLMs
- 2 The way you ask, the way it replies
- 3 Summary

A chatbot window on <https://chat.fz-rossendorf.de>

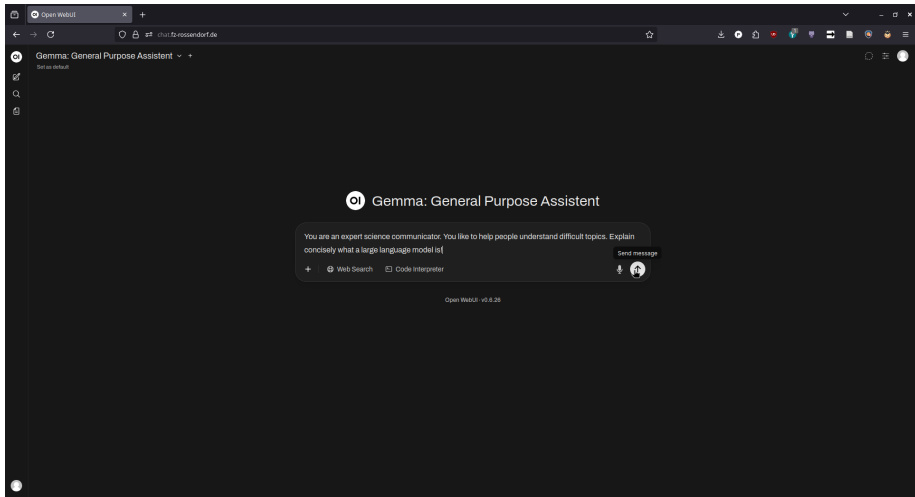


A chatbot window: What happens?

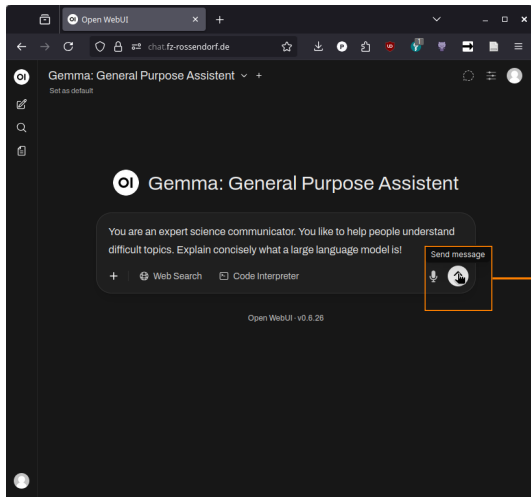


This is a string
(a sequence of characters)
which I just typed in.

A chatbot window: You send something!



Sending input: What happens?



When I click:

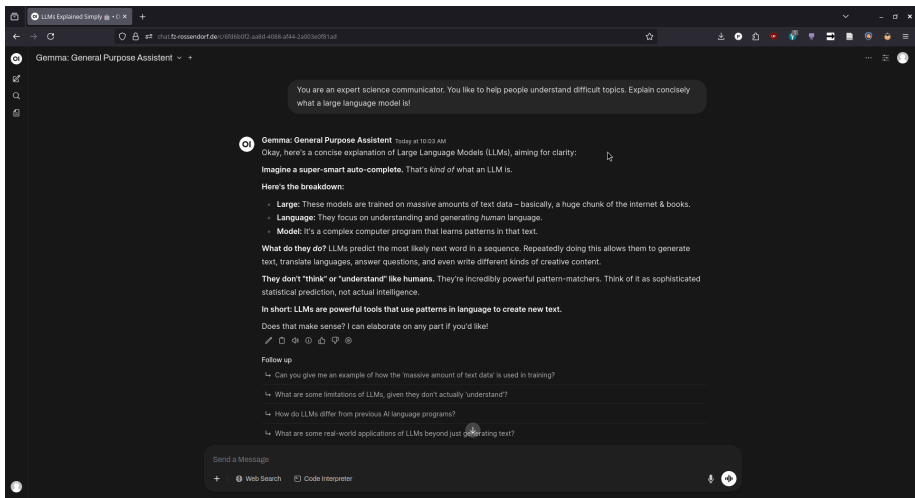
the string will be filtered, checked and submitted to a software on another computer.

This software forwards your string to a **large language model (LLM)**.

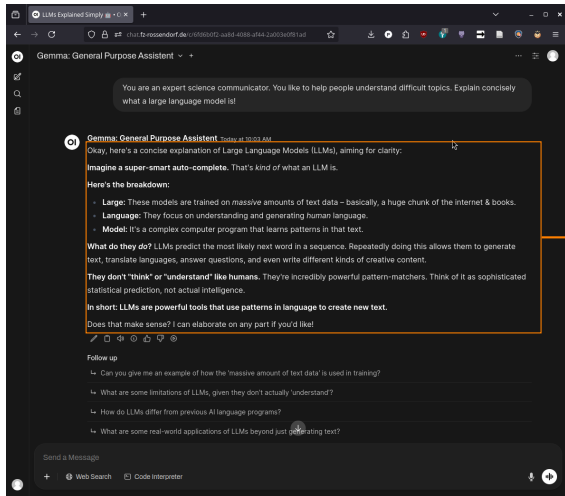
The large language model at work!



A chatbot window: Your results!

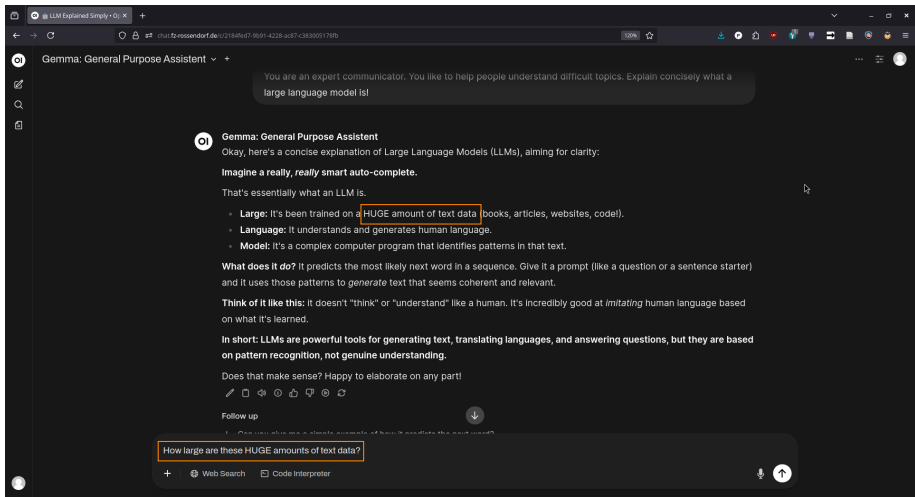


Receiving input



The LLM returns a reply. The web-site software interprets it and shows it in your browser. The text appears formatted.

Starting a Dialogue



It's talking to me?

Gemma: General Purpose Assistant

You're right to ask – "huge" is relative! Let's put some numbers to it. It's genuinely mind-boggling.

Here's a breakdown to give you a sense of scale:

- **Tokens:** LLMs don't actually "read" words, they process pieces of words called "tokens". Roughly, 1 token is about 3/4 of a word.
- **GPT-3 (an earlier, well-known LLM):** Trained on around **175 billion tokens**. That equates to roughly **300 billion words** - enough to fill over 10,000 books!
- **GPT-4 (the model powering ChatGPT Plus):** The exact size isn't publicly disclosed, but estimates suggest it's trained on **over 1.76 trillion tokens** – *more than ten times* the amount of GPT-3. That's like reading the entire internet multiple times over.
- **Google's Gemini:** Similarly, its training dataset is enormous, likely in the trillions of tokens.
- **The Common Crawl:** A major source of data is the "Common Crawl", which is a public archive of web pages. It contains *petabytes* of data (1 petabyte = 1,000 terabytes).

To put it in perspective:

- **All of Wikipedia:** Contains about 18 million articles, representing roughly 8 billion words. An LLM trains on *many, many multiples* of that.
- **Books:** LLMs are trained on massive collections of books - think the entire Library of Congress, and much more.

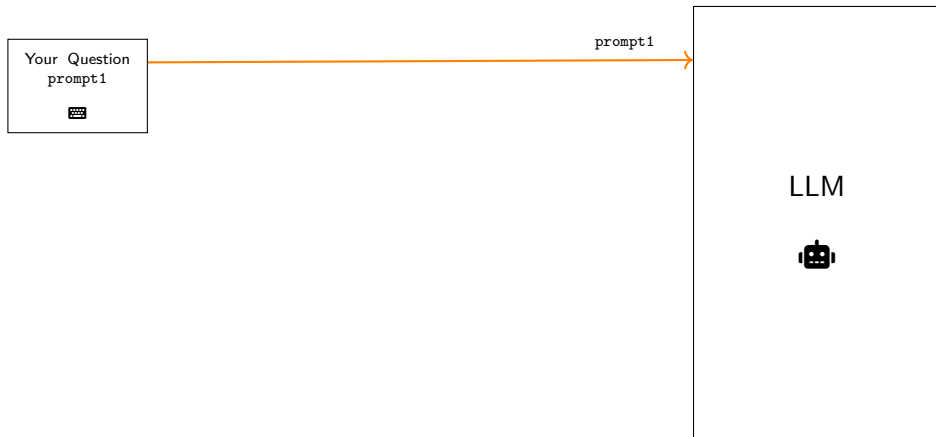
It's not just *how much* data, but also *what kind*: The data includes books, articles, websites, code, and more, giving the LLM a very broad base of knowledge (and biases, which is another topic!).

These numbers are constantly increasing as models are updated and retrained. It's a data arms race!

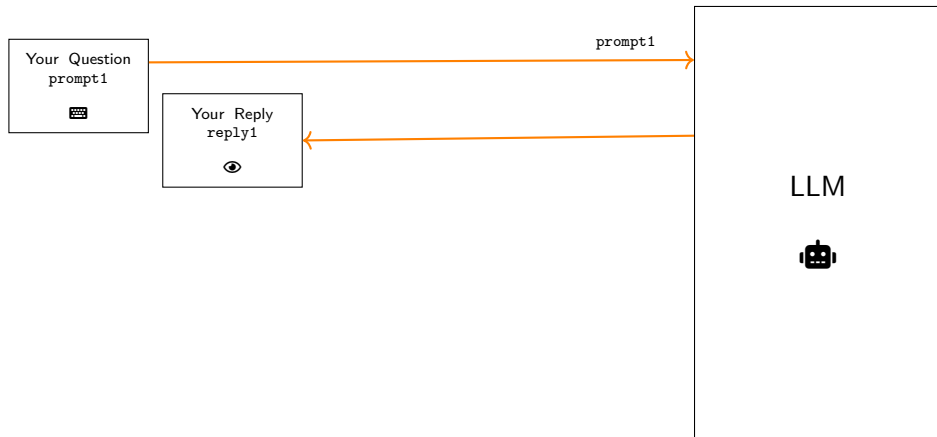
Send a Message

+ Web Search Code Interpreter

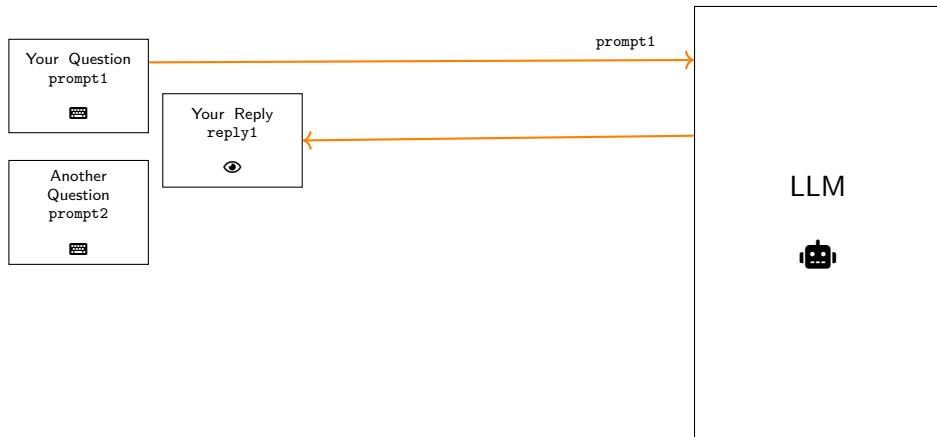
Chatting with a LLM: A view in boxes



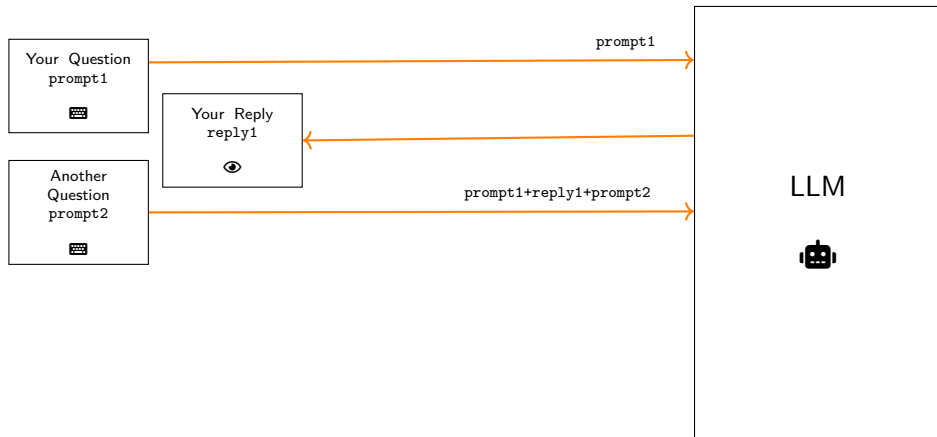
Chatting with a LLM: A view in boxes



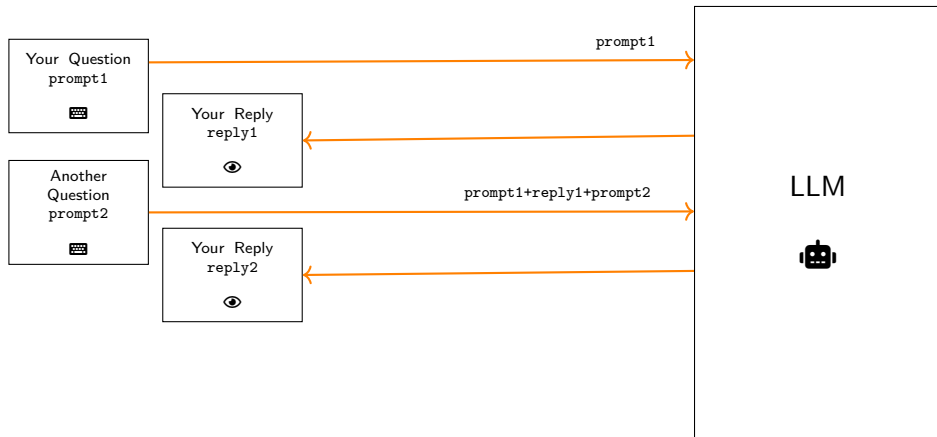
Chatting with a LLM: A view in boxes



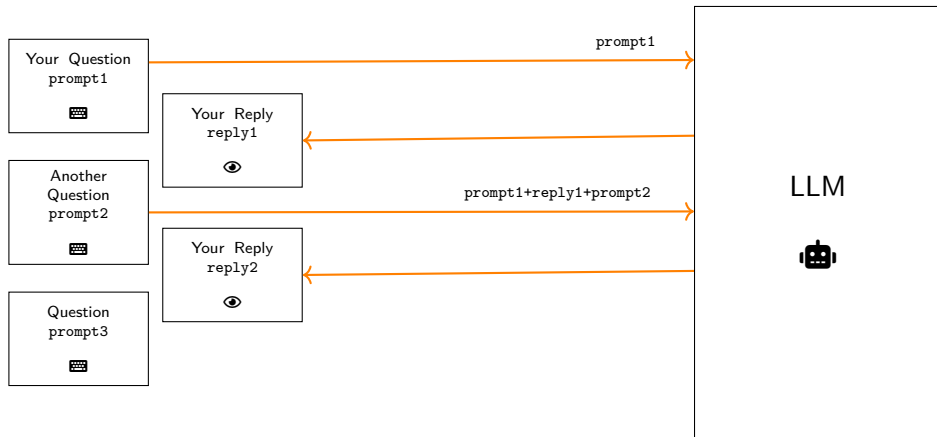
Chatting with a LLM: A view in boxes



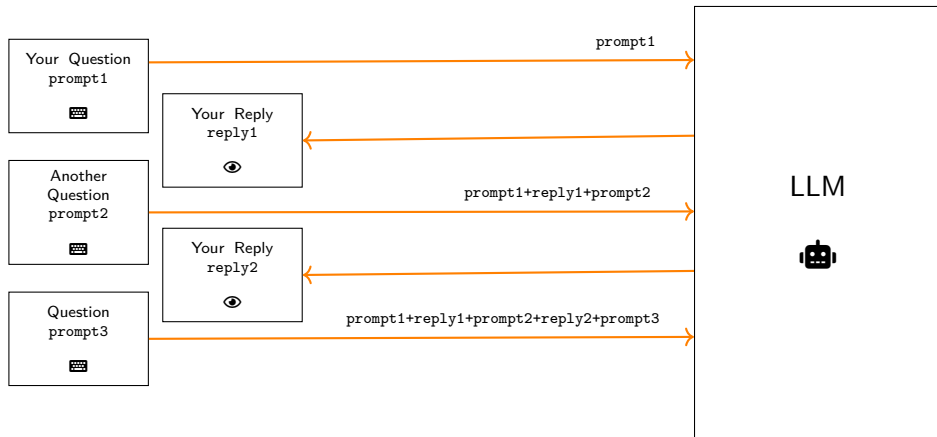
Chatting with a LLM: A view in boxes



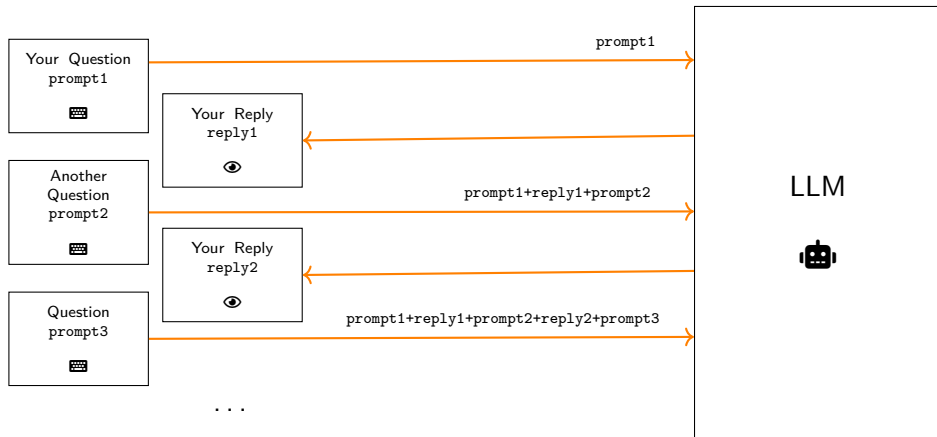
Chatting with a LLM: A view in boxes



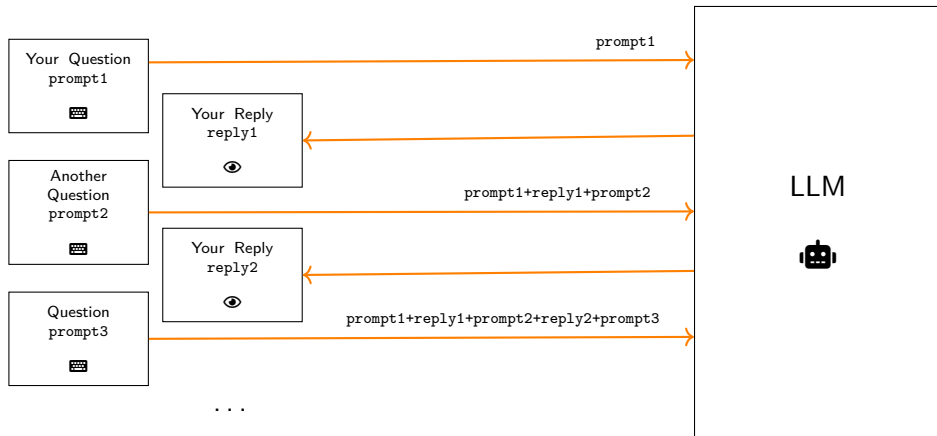
Chatting with a LLM: A view in boxes



Chatting with a LLM: A view in boxes

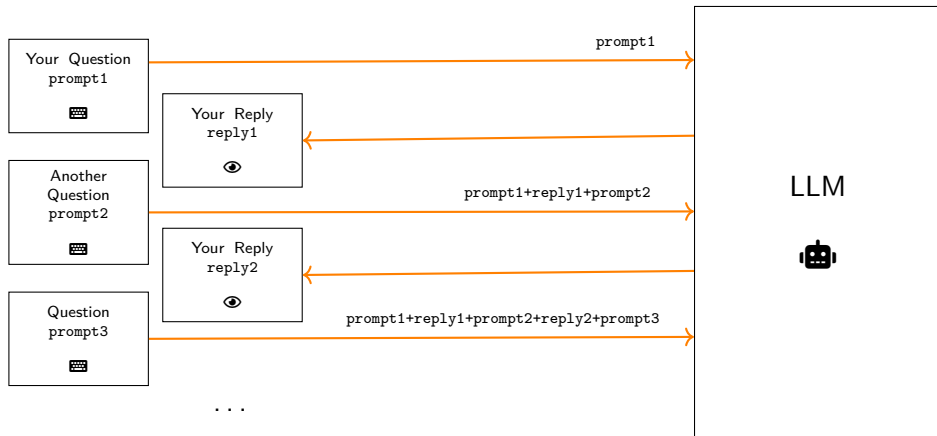


Chatting with a LLM: A view in boxes



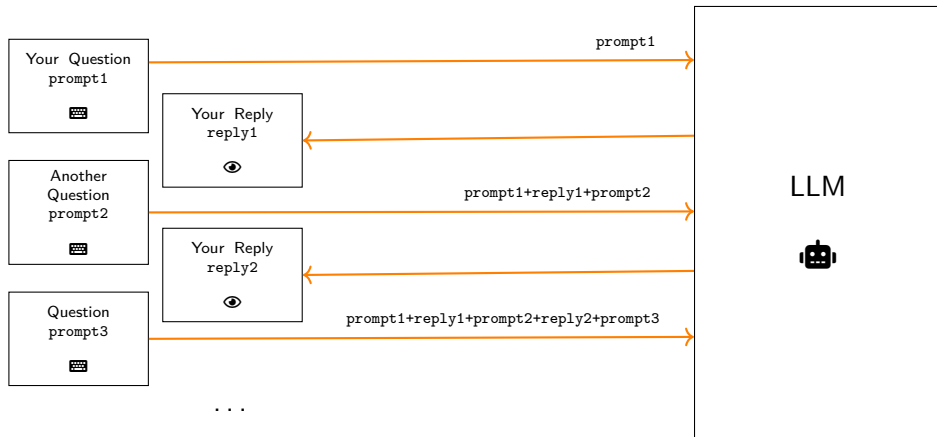
- LLM “remembers” content of a chat dialogue

Chatting with a LLM: A view in boxes



- LLM “remembers” content of a chat dialogue
- mind the **context length**: LLMs can also forget

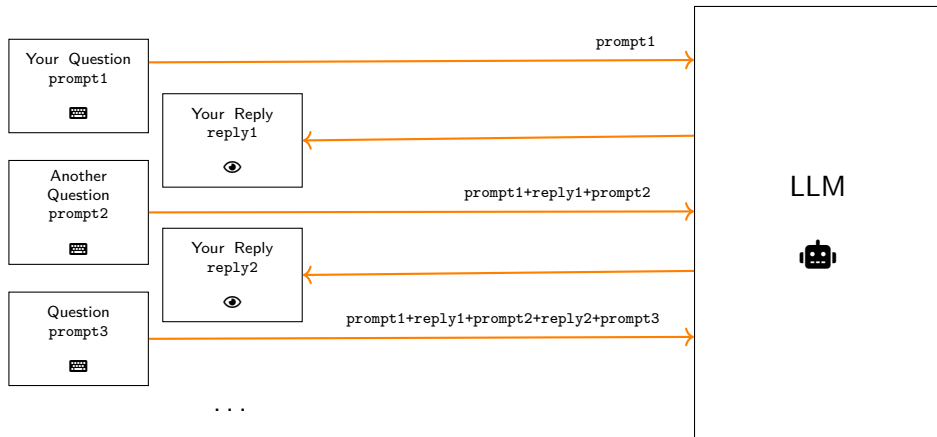
Chatting with a LLM: A view in boxes



- LLM “remembers” content of a chat dialogue
- mind the **context length**: LLMs can also forget

➔ If in doubt, create a new chat!

Chatting with a LLM: A view in boxes



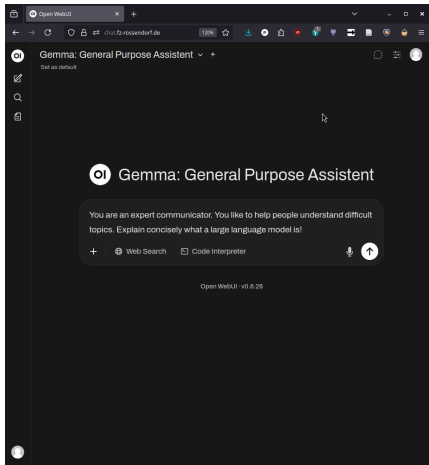
- LLM “remembers” content of a chat dialogue
- mind the **context length**: LLMs can also forget

- ➔ If in doubt, create a new chat!
- ➔ Always double check!

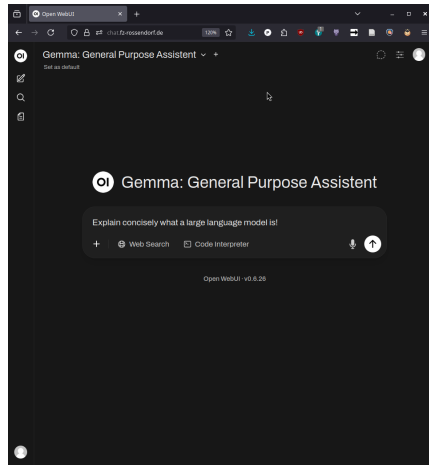
Table of Contents

- 1 Dialogues in LLMs
- 2 The way you ask, the way it replies
- 3 Summary

System Messages: test it yourself!



You are an expert communicator. You like to help people understand difficult topics. Explain concisely what a large language model is!



Explain concisely what a large language model is!

System Messages

- system messages instruct the LLM which role to assume
- they are a relict of instruction tuning
- system messages can be very powerful
- if not changed, system messages influence the entire dialogue

System Messages

- system messages instruct the LLM which role to assume
- they are a relict of instruction tuning
- system messages can be very powerful
- if not changed, system messages influence the entire dialogue

Try another system message:

`"You are a teenage high school student interested in social media. You like to show people that you know stuff. Explain concisely what a large language model is!"`

Table of Contents

- 1 Dialogues in LLMs
- 2 The way you ask, the way it replies
- 3 Summary

Main Takeaways

- instruction tuning has direct consequence on how to interact with LLM

Main Takeaways

- instruction tuning has direct consequence on how to interact with LLM
- a dialogue is constructed by concatenating prompts and replies

Main Takeaways

- instruction tuning has direct consequence on how to interact with LLM
- a dialogue is constructed by concatenating prompts and replies
- a system message can be used to force the LLM into another role

Main Takeaways

- instruction tuning has direct consequence on how to interact with LLM
- a dialogue is constructed by concatenating prompts and replies
- a system message can be used to force the LLM into another role

Main Takeaways

- instruction tuning has direct consequence on how to interact with LLM
- a dialogue is constructed by concatenating prompts and replies
- a system message can be used to force the LLM into another role

Thank you for your attention!

Feel free to ask questions, provide feedback or comments.

Further Reading

- www.promptingguide.ai - compendium of prompting techniques (available in multiple languages)
- [the prompt report](#) - 2024 meta analysis of prompting techniques
- dspy.ai - optimize your prompt programmatically