The Significance of LLM Tokenization
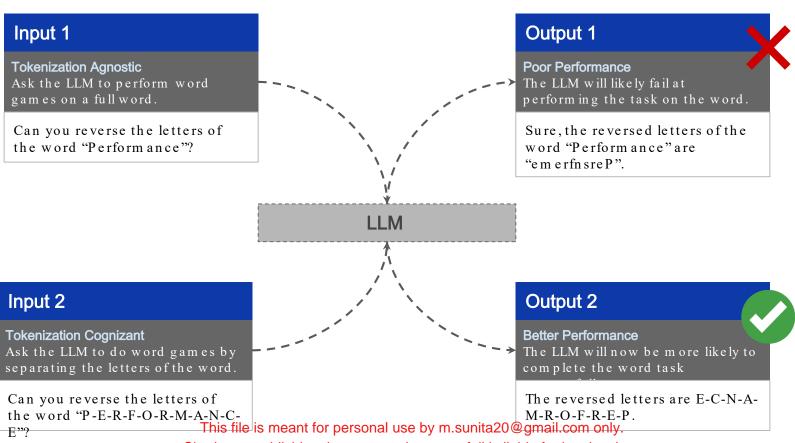
Agenda

In this session, we will discuss:

- Problems caused by Tokenization in LLMs
- Mitigating Tokenization Problems
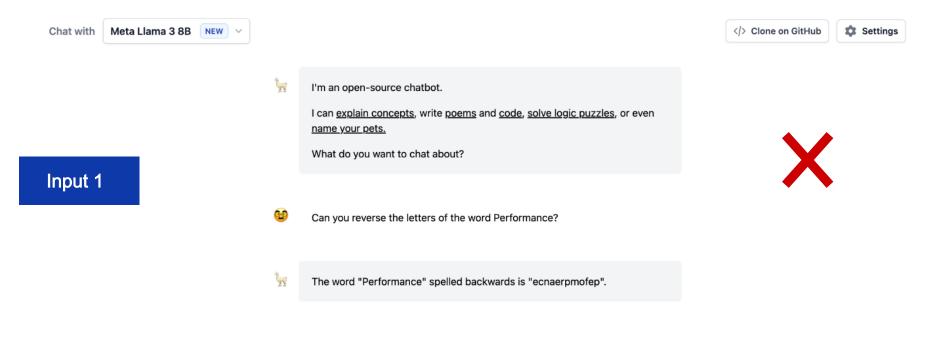
The Significance of Tokenization

Input 1

Tokenization Agnostic
Ask the LLM to perform word games on a full word.

Can you reverse the letters of the word "Performance"?

Output 1

Poor Performance
The LLM will likely fail at performing the task on the word.

Sure, the reversed letters of the word "Performance" are "emerfnsreP".

LLM

Input 2

Tokenization Cognizant
Ask the LLM to do word games by separating the letters of the word.

Can you reverse the letters of the word "P-E-R-F-O-R-M-A-N-C-E"?

Output 2

Better Performance
The LLM will now be more likely to complete the word task

The reversed letters are E-C-N-A-M-R-O-F-R-E-P.

The Significance of Tokenization (Cont.)

Chat with Meta Llama 3 8B NEW

</> Clone on GitHub   ⚙ Settings

🦙 I'm an open-source chatbot.

I can explain concepts, write poems and code, solve logic puzzles, or even name your pets.

What do you want to chat about?

😎 Can you reverse the letters of the word Performance?

🦙 The word "Performance" spelled backwards is "ecnaerpmofep".

Input 1

❌

The Significance of Tokenization (Cont.)

Chat with **Meta Llama 3 8B** NEW ˅

Clone on GitHub   Settings

🦙 I'm an open-source chatbot.

I can <u>explain concepts</u>, write <u>poems</u> and <u>code</u>, <u>solve logic puzzles</u>, or even <u>name your pets.</u>

What do you want to chat about?

Input 2

🤓 Can you reverse the letters of the word P-E-R-F-O-R-M-A-N-C-E ?

🦙 I'd be happy to help you with that.

The reversed letters of the word P-E-R-F-O-R-M-A-N-C-E is E-C-N-A-M-R-O-F-R-E-P.

Summary

Here's a brief recap:

- Tokenization can cause Large Language Models (LLMs) to misinterpret words compared to human reviewers, especially in tasks like Word Games that require alphabet-level tokenization.

- The correct way to resolve these issues is by adding special separator characters between alphabets, ensuring the LLM processes each alphabet as an individual token.