DeepLorIA

Mastering Large Language Models: Efficient Techniques for Fine-Tuning

Barbara Gendron-Audebert, PhD student (MosAIk team)

LORIA, Université de Lorraine, CNRS
DeepLorIA Network

January 15, 2025

About Me

2nd-year PhD student - Knowledge-Enhanced Language Models (Université de Lorraine)

Background & Research

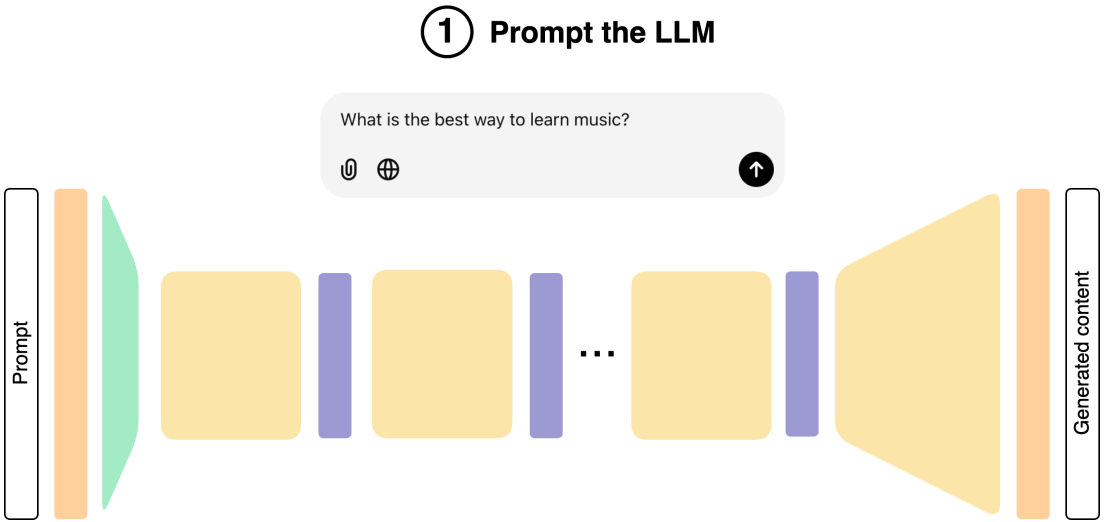
- Maths engineering degree
- Master Thesis: Meta-Learning in Conversational Context
- PhD Topic: Controlled Conversational Models through Conversation-Dedicated Ontology
- *Keywords: Large Language Models (LLMs), Conversational Agents, Ontologies, Fine-Tuning*

More info: `b-gendron.github.io`

Tutorial Objectives

What to be learnt
What to be practiced

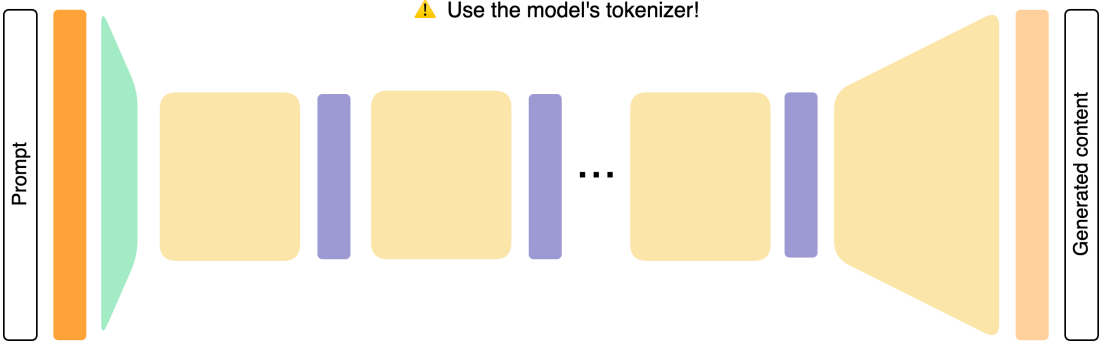
Hello there


1 Prompt the LLM
What is the best way to learn music?
Prompt
Generated content

② Tokenize the Prompt Content

```
"What is the best way to learn music?"
[531, 9, 45, 22, 3316, 2444, 34, 2172, 334]
```

⚠️ Use the model's tokenizer!



③ Apply an Embedding layer

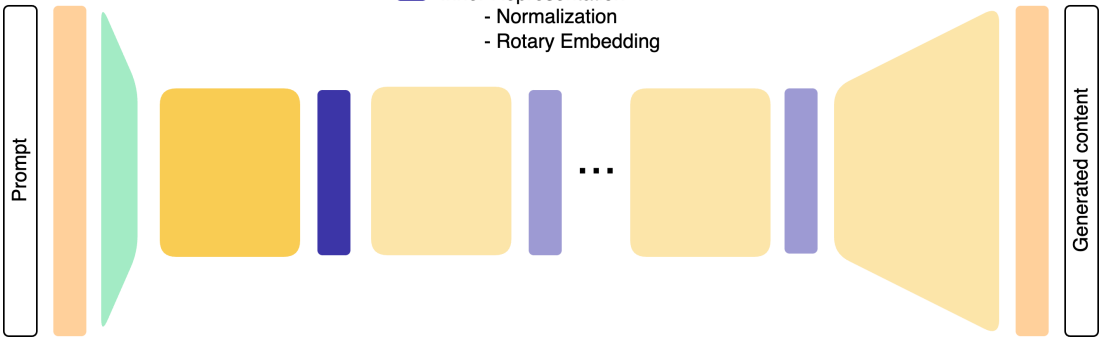
`Embedding(vocab_size, hidden_size)`

$$i \mapsto (e_1, \ldots, e_{\texttt{hidden\_size}})$$

4 Going Through a Decoder Block

Transformer Decoder Blocks
- Attention layers (query, key, value)
- Dense layers (Multi Layer Perceptron)

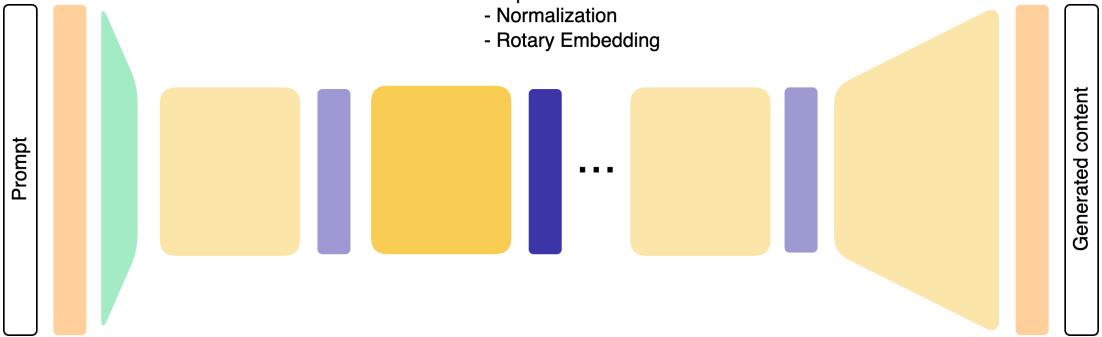
Inner Representation
- Normalization
- Rotary Embedding



4 Going Through a Decoder Block

Transformer Decoder Blocks
- Attention layers (query, key, value)
- Dense layers (Multi Layer Perceptron)

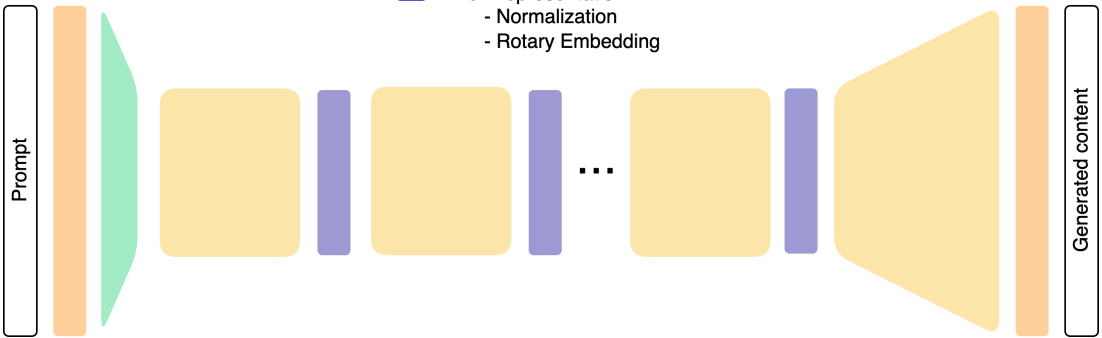
Inner Representation
- Normalization
- Rotary Embedding



4 Going Through a Decoder Block

Transformer Decoder Blocks
- Attention layers (query, key, value)
- Dense layers (Multi Layer Perceptron)

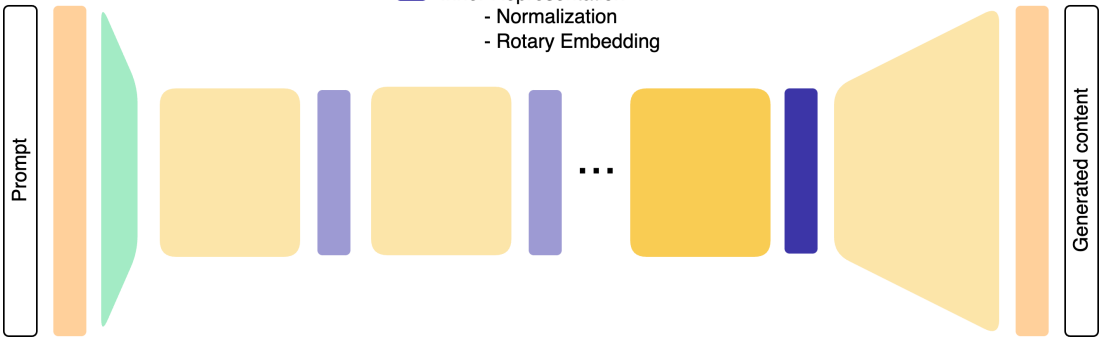
Inner Representation
- Normalization
- Rotary Embedding

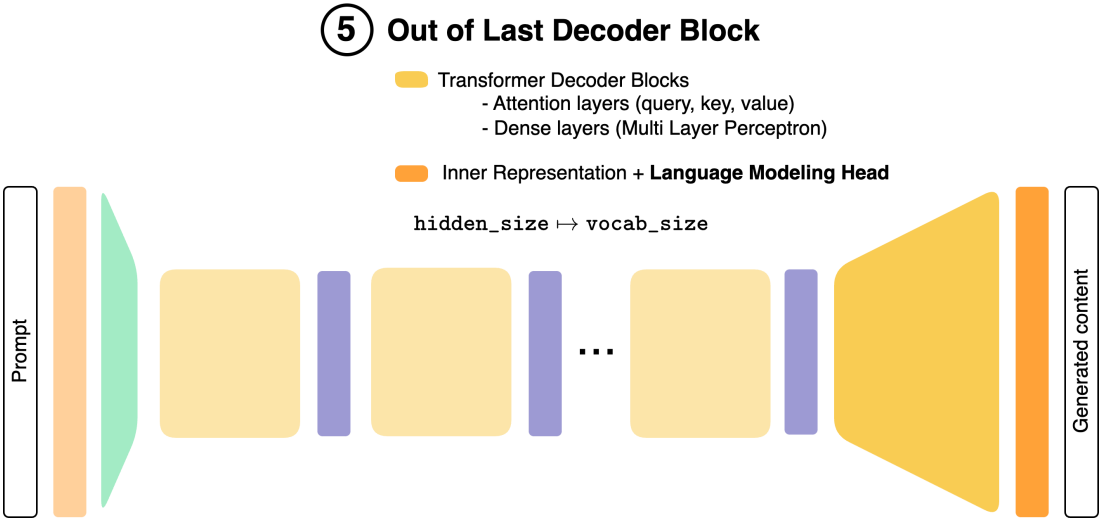


4 Going Through a Decoder Block

Transformer Decoder Blocks
- Attention layers (query, key, value)
- Dense layers (Multi Layer Perceptron)

Inner Representation
- Normalization
- Rotary Embedding




5 Out of Last Decoder Block
Transformer Decoder Blocks
- Attention layers (query, key, value)
- Dense layers (Multi Layer Perceptron)
Inner Representation + Language Modeling Head
hidden_size ↦ vocab_size
Prompt
...
Generated content

6 Decode Generated Tokens in Natural Language