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`

Context

Recurrent Models for NLP (1)

Recurrent Models for NLP (2)

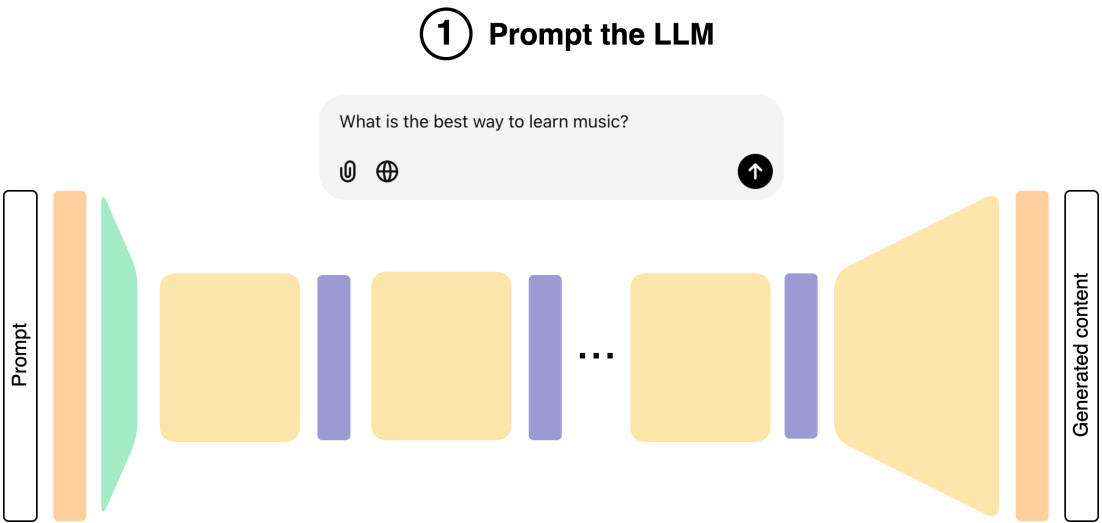
The Transformer Model

Attention Principle

From Transformer-Based Models to Large Language Models (LLMs)

What Use-Cases of LLMs?

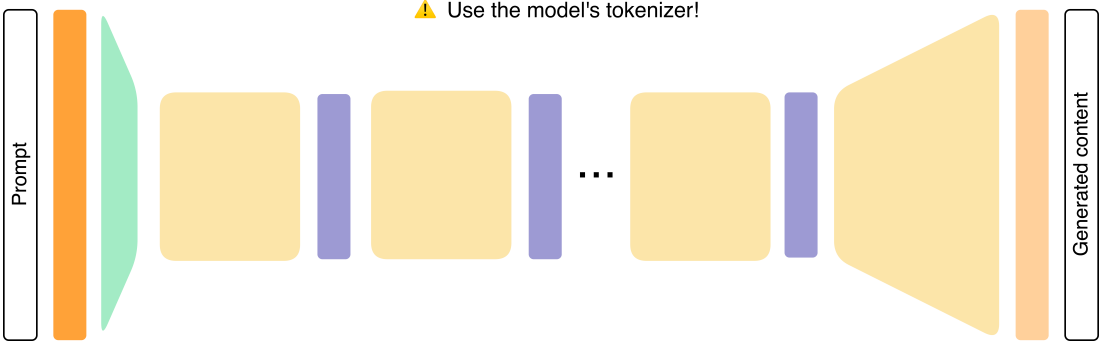
Inside a Decoder-Based LLM



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!



Inside a Decoder-Based LLM

③ Apply an Embedding layer

`Embedding(vocab_size, hidden_size)`

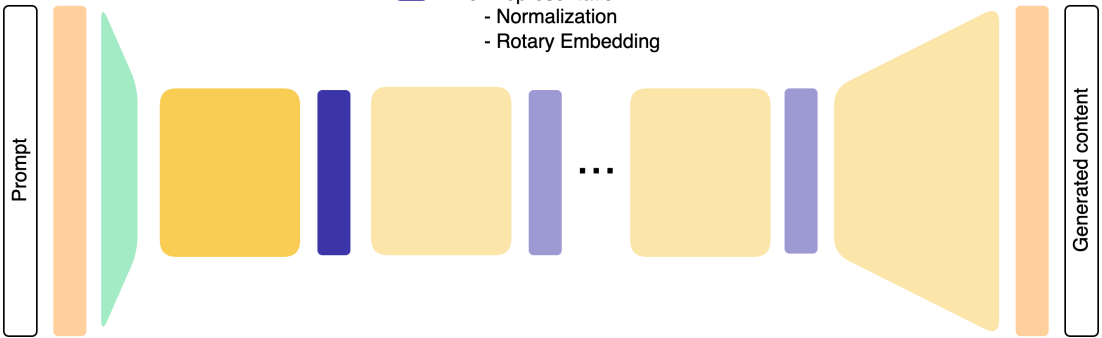
$$i \mapsto (e_1, \ldots, e_{\texttt{hidden\_size}})$$

Inside a Decoder-Based LLM

4 Going Through a Decoder Block

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

Inner Representation
- Normalization
- Rotary Embedding

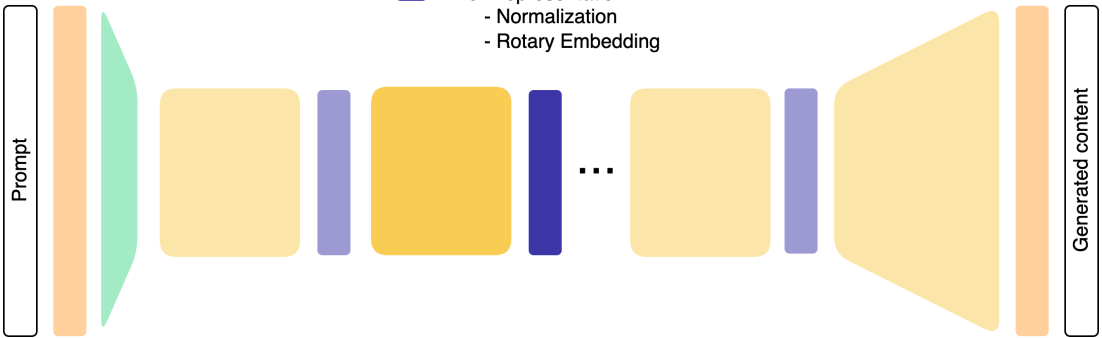


Inside a Decoder-Based LLM

4 Going Through a Decoder Block

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

Inner Representation
- Normalization
- Rotary Embedding

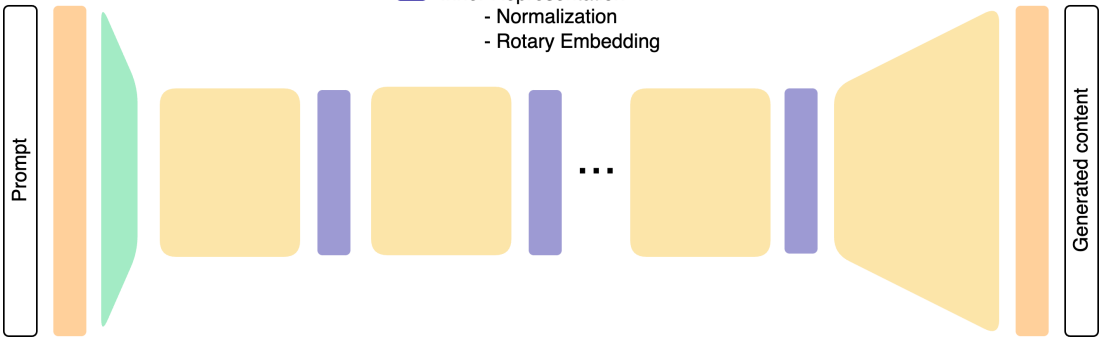


Inside a Decoder-Based LLM

4 Going Through a Decoder Block

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

Inner Representation
- Normalization
- Rotary Embedding

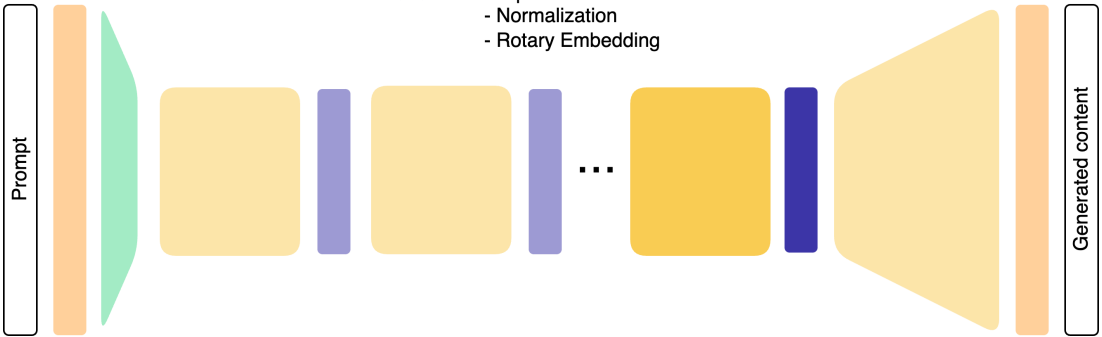


Inside a Decoder-Based LLM

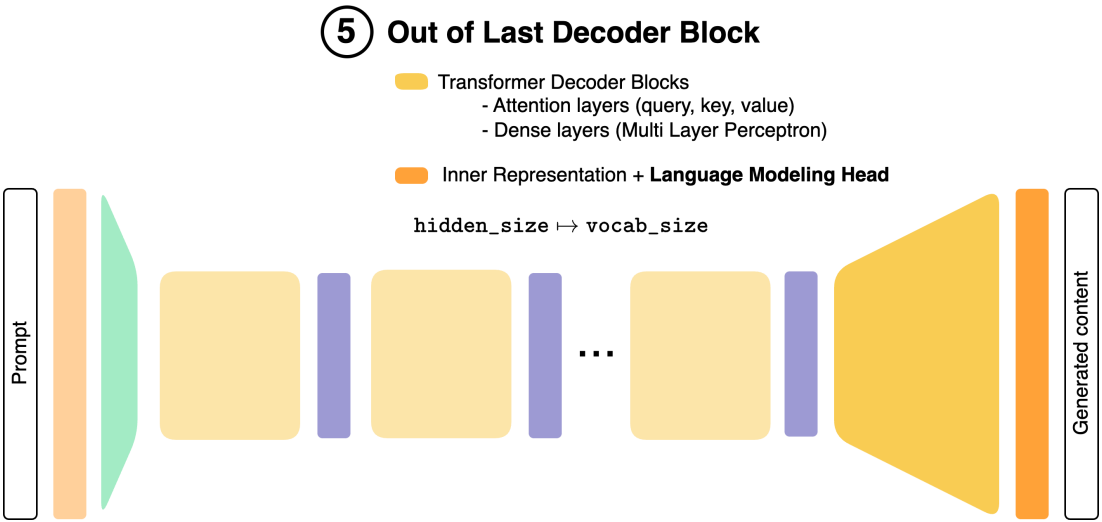
4 Going Through a Decoder Block

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

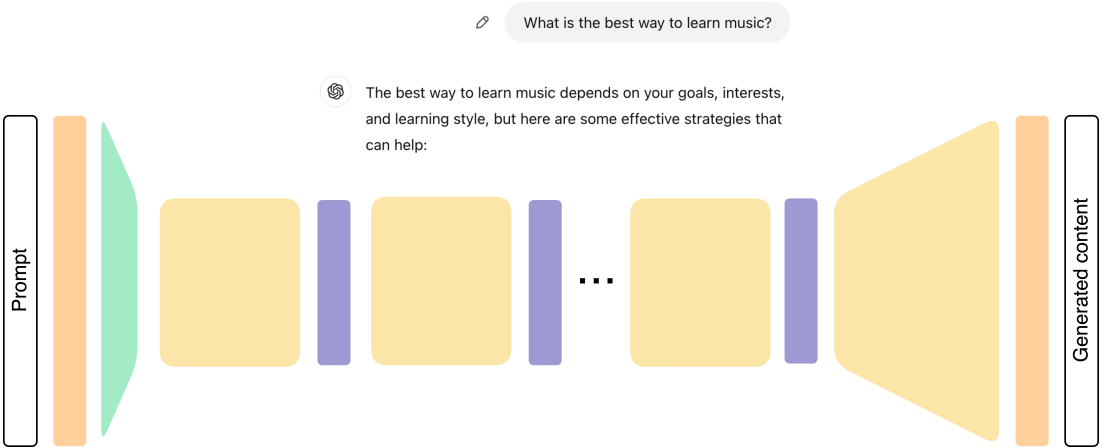
Inner Representation
- Normalization
- Rotary Embedding



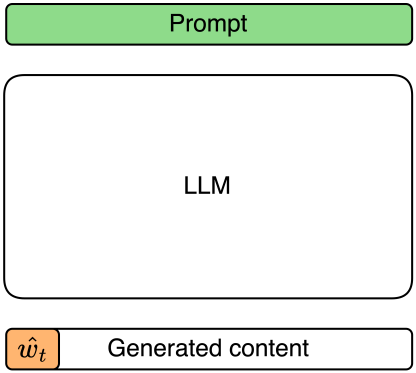
Inside a Decoder-Based LLM



6 Decode Generated Tokens in Natural Language



The Autoregressive Principle

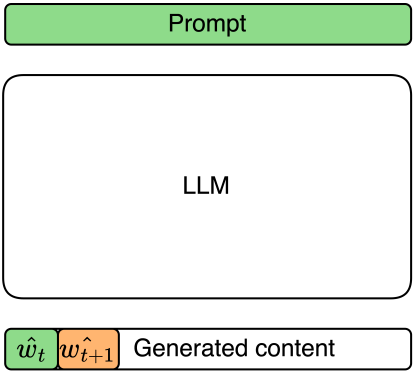


Context

Predicted token

$$\hat{w}_t = \arg\max_{w_i} P(w_i | w_1, w_2, \ldots, w_{t-1})$$

The Autoregressive Principle



Context

Predicted token

$$\hat{w_{t+1}} = \arg\max_{w_i} P(w_i | w_1, w_2, \ldots, w_t)$$