

DeepLorIA

Mastering Large Language Models: Efficient Techniques for Fine-Tuning

Barbara Gendron-Audebert, PhD student (MosAlk team)

LORIA, Université de Lorraine, CNRS
DeepLorIA Network

January 15, 2025

About Me

2nd-year PhD student - Knowledge-Enhanced Language Models

Research Focus

- Controlled Conversational Models through Conversation-Dedicated Ontology
- *Keywords: Large Language Models (LLMs), Conversational Agents, Ontologies, Fine-Tuning*

Experience in LLM fine-tuning

- Run pre-defined fine-tuning setups (Causal Language Modeling, Classification,...)
- Develop new fine-tuning pipelines to consider external knowledge
- Focus on textual modality

Experimenting LLM Fine-Tuning

Many successful LLM fine-tunings?

<https://huggingface.co/>

The screenshot shows the Hugging Face Model Hub interface. At the top, there is a search bar and a filter button labeled "finetuned". Below the search bar, the number of models is displayed as "Models 70,963". The main area is a grid of model cards, each containing the model name, author, description, and metrics. The cards are arranged in rows and columns. Some cards have a green checkmark icon, indicating they are finetuned models. The cards include:

- distilbert/distilbert-base-uncased-finetuned-sst-2-...
- FacebookAI/xlm-roberta-large-finetuned-conll03-engl...
- dandelin/vilt-b32-finetuned-vqa
- google/tapas-base-finetuned-wtq
- naver-clova-ix/donut-base-finetuned-cord-v2
- facebook/timesformer-base-finetuned-k400
- EmTpro01/CodeLlama-7b-finetuned-16bit
- VeraSolutions/phi-3.5-mini-finetuned-amp-data-model...
- huuhiu-ai/Llama-3.7-76B-Instruct-ableitized-finetu...
- duyan2003/bartpho-finetuned-qa
- medicalai/MedFound-Llama3-8B-finetuned
- google/bert/bert-large-uncased-whole-word-masking-f...
- Davlan/bert-base-multilingual-cased-finetuned-wolof
- allenai/longformer-large-4096-finetuned-triviaqa
- atharvanundada99/bert-large-question-answering-fine...
- dbmdz/bert-large-cased-finetuned-conll03-english
- google/tapas-large-finetuned-wtq
- henryk/bert-base-multilingual-cased-finetuned-polis...

Experimenting LLM Fine-Tuning

Many successful LLM fine-tunings?
<https://huggingface.co/>

Models 70,963 finetuned

Full-text search N Sort: Trending

- distilbert/distilbert-base-uncased-finetuned-sst-2-
Text Classification - Updated Dec 19, 2023 - 3.6.72M - 467
- dandelin/vilt-b32-finetuned-vqa
Visual Question Answering - Updated Aug 2, 2022 - 3.175K - 397
- naver-clova-ix/donut-base-finetuned-cord-v2
Image-to-Text - Updated Aug 13, 2022 - 3.13.6K - 92
- EmTpro01/CodeLlama-7b-finetuned-16bit
Text Generation - Updated Nov 3, 2024 - 332 - 2
- huhiu-ai/Llama3-3.7-76B-Instruct-ableliterated-finetu...
Text Generation - Updated 8 days ago - 3.189K - 3
- medicalai/MedFound-Llama3-8B-finetuned
Updated 7 days ago - 3.56 - 2
- Davlan/bert-base-multilingual-cased-finetuned-wolof
Fill-Mask - Updated Jun 30, 2021 - 3.15 - 2
- atharvanundada99/bert-large-question-answering-fine...
Question Answering - Updated May 24, 2021 - 3.701 - 15
- google/tapas-large-finetuned-wtq
Table Question Answering - Updated Sep 5, 2023 - 3.114K - 132
- FacebookAI/xlm-roberta-large-finetuned-conll03-engl...
Token Classification - Updated Feb 19, 2024 - 3.1.8M - 159
- facebook/timesformer-base-finetuned-k400
Video Classification - Updated Jan 2, 2023 - 3.46.4K - 28
- VeraSolutions/phi-3.5-mini-finetuned-amp-data-model...
Text Generation - Updated Dec 11, 2024 - 3
- duyan2003/bartpho-finetuned-qa
Text2Text Generation - Updated 15 days ago - 3.33 - 2
- google/bert/bert-large-uncased-whole-word-masking-f...
Question Answering - Updated Feb 19, 2024 - 3.20K - 173
- allenai/longformer-large-4096-finetuned-triviaqa
Question Answering - Updated Oct 4, 2022 - 3.7.56K - 7
- dbmdz/bert-large-cased-finetuned-conll03-english
Token Classification - Updated Sep 7, 2023 - 3.1.2M - 4 - 73
- henryk/bert-base-multilingual-cased-finetuned-polis...
Question Answering - Updated May 19, 2021 - 3.220 - 3

Fine-tuning LLMs in real life?

RuntimeError: probability tensor contains either 'inf', 'nan' or element < 0

"",\n\nAverage Readability Score = 9.45.\n\nPlease let me know if this meets your requirements', "...", "...", "...".\n\nReadability Score = 8.45.\n\nThis text has moderate complexity, making it easy for', "...".\n\nAverage Readability Score = 4.\n\nThis text has short sentences, simple vocabulary words with one', "...".\n\nAverage Readability Score = 14.00.\n\nPlease note that the above text may be difficult for', "...".\n\nAverage Readability Score = 8.45.\n\nThis text has moderate difficulty, making it easily', "..."]
["...", "...", "...", "...", "...", "...", "...", "..."]
["...", "...", "...", "<!", "The new smartphone has many advanced features for improved performance." | Readability Score', "...".\u0000e0067\ufe00d, ✨, 🎉, 99, 🎉, 0, "...", "...".\n\nThis text has an estimated Flesch-Kincaid Grade Level around the range of', "...", "...", "...".\u0000e0067\ufe00e0062\ufe00e0073\ufe00e0063\ufe00e0074\ufe00e00710']

apples fell from trees in autumn.\nPeople enjoy watching movies together.\nApples grow in trees.\nShe enjoys reading books every day.\nApples hang from trees in autumn.\nPeople often use umbrellas when', 'Ice cream melts in hot weather.\nShe enjoys reading books about history.\nApples grow in trees.\nI enjoy reading books about history.\nIce cream tastes delicious.\nShe jumped over the moon today.\nUser Inst', 'I enjoy reading books in my free time.\nShe baked delicious chocolate chip cookies.', 'Apples fall from trees in autumn.\nA bird flew over the mountain peak.'

poch 1: 2%
Sample 0: Yes.,
Sample 1: Yes.,
Sample 2: Today,
Sample 3: In.,
Sample 4: For.,
[1.1630859375, 1.3134765625, 1.140625]

Experimenting LLM Fine-Tuning

Many successful LLM fine-tunings?
<https://huggingface.co/>

<https://huggingface.co/>

Models	70,963	finetuned	Full-text search	11 Sort: Trending
distilbert/distilbert-base-uncased-finetuned-sst-2-english	Text Classification · Updated Dec 19, 2023 · ± 6.72M · ⚡ 667		FacebookAI/xlm-roberta-large-finetuned-conll03-english	Token Classification · Updated Feb 19, 2024 · ± 1.8M · ⚡ 159
dandelin/vilt-b32-finetuned-vqa	Visual Question Answering · Updated Aug 2, 2022 · ± 175K · ⚡ 397		google/tapas-base-finetuned-wtq	Table Question Answering · Updated Jul 14, 2022 · ± 14.9K · ⚡ 207
naver-clova-ix/donut-base-finetuned-cord-v2	Image-to-Text · Updated Aug 13, 2022 · ± 13.5K · ⚡ 92		facebook/timesformer-base-finetuned-k400	Video Classification · Updated Jan 2, 2023 · ± 46.4K · ⚡ 28
EmPro01/CodeLlama-7b-finetuned-16bit	Text Generation · Updated Nov 3, 2024 · ± 332 · ⚡ 2		Verasolutions/phi-3.5-mini-finetuned-amp-data-model	Text Generation · Updated Dec 11, 2024 · ⚡ 3
huwaii-ai/Llama-3.3-78B-Instruct-abiliterated-finetuned	Text Generation · Updated 8 days ago · ± 189K · ⚡ 3		duyan2803/bartpho-finetuned-qa	Text2Text Generation · Updated 15 days ago · ± 33 · ⚡ 2
medicalai/MedFound-Llama3-8B-finetuned	Updated 7 days ago · ± 56 · ⚡ 2		google/bert/bert-large-uncased-whole-word-masking-finetuned	Question Answering · Updated Feb 19, 2024 · ± 20K · ⚡ 173
Davlan/bert-base-multilingual-cased-finetuned-wolof	Fine-Mask · Updated Jun 30, 2021 · ± 15 · ⚡ 2		allennai/longformer-large-4896-finetuned-trivqaqa	Question Answering · Updated Oct 4, 2022 · ± 7.58K · ⚡ 7
atharvaranundada99/bert-large-question-answering-finetuned	Question Answering · Updated May 24, 2021 · ± 701 · ⚡ 15		dbmdz/bert-large-cased-finetuned-conll03-english	Token Classification · Updated Sep 7, 2023 · ± 1.2M · ⚡ 73
google/tapas-large-finetuned-wtq	Table Question Answering · Updated Sep 5, 2023 · ± 114K · ⚡ 132		henryk/bert-base-multilingual-cased-finetuned-polish	Question Answering · Updated May 19, 2021 · ± 220 · ⚡ 3

- Fine-tuning LLMs relies on obscure "magic formulas"

Fine-tuning LLMs in real life?

apples fell from trees in autumn.\nPeople enjoy watching movies together.\n' Apples grow in trees.\n' Apples hang from trees.\nShe enjoys reading books every day.\nPeople often use umbrellas when'.\nIce cream melts in hot weather.\nShe enjoys reading books about history.\nApples grow in trees.\nI enjoy reading books about history.\nbooks taste delicious.\nShe jumped over the moon today.\nUnder Inst',\nI enjoyed reading books in my free time.\nShe baked delicious chocolate chip cookies.\nApples fall from trees in autumn.\nA bird flew over the mountain peak.\n'

```
poch 1: 2% |  
Sample 0: Yes., .....  
Sample 1: Yes., .....  
Sample 2: Today, ..  
Sample 3: In., .....  
Sample 4: For, .....  
[1.1630859375, 1.3134765625, 1.140625]
```

- Fine-tuning LLMs is hard

Experimenting LLM Fine-Tuning

Many successful LLM fine-tunings?
<https://huggingface.co/>

Models	70,963	finetuned	Full-text search	11 Sort: Trending
distilbert/distilbert-base-uncased-finetuned-sst-2-english	Text Classification · Updated Dec 19, 2023 · ± 6.72M · ⚡ 667		FacebookAI/xlm-roberta-large-finetuned-conll03-english	Token Classification · Updated Feb 19, 2024 · ± 1.8M · ⚡ 159
dandelin/vilt-b32-finetuned-vqa	Visual Question Answering · Updated Aug 2, 2022 · ± 175K · ⚡ 397		google/tapas-base-finetuned-wtq	Table Question Answering · Updated Jul 14, 2022 · ± 14.9K · ⚡ 207
naver-clova-ix/donut-base-finetuned-cord-v2	Image-to-Text · Updated Aug 13, 2022 · ± 13.5K · ⚡ 92		facebook/timesformer-base-finetuned-k400	Video Classification · Updated Jan 2, 2023 · ± 46.4K · ⚡ 28
EmPro01/CodeLlama-7b-finetuned-16bit	Text Generation · Updated Nov 3, 2024 · ± 332 · ⚡ 2		Verasolutions/phi-3.5-mini-finetuned-amp-data-model	Text Generation · Updated Dec 11, 2024 · ⚡ 3
huwaii-ai/Llama-3.3-78B-Instruct-abiliterated-finetuned	Text Generation · Updated 8 days ago · ± 189K · ⚡ 3		duyan2803/bartpho-finetuned-qa	Text2Text Generation · Updated 15 days ago · ± 33 · ⚡ 2
medicalai/MedFound-Llama3-8B-finetuned	Updated 7 days ago · ± 56 · ⚡ 2		google/bert/bert-large-uncased-whole-word-masking-finetuned	Question Answering · Updated Feb 19, 2024 · ± 20K · ⚡ 173
Davlan/bert-base-multilingual-cased-finetuned-wolof	Fine-Mask · Updated Jun 30, 2021 · ± 15 · ⚡ 2		allennai/longformer-large-4896-finetuned-trivqaqa	Question Answering · Updated Oct 4, 2022 · ± 7.58K · ⚡ 7
atharvaranundada99/bert-large-question-answering-finetuned	Question Answering · Updated May 24, 2021 · ± 701 · ⚡ 15		dbmdz/bert-large-cased-finetuned-conll03-english	Token Classification · Updated Sep 7, 2023 · ± 1.2M · ⚡ 73
google/tapas-large-finetuned-wtq	Table Question Answering · Updated Sep 5, 2023 · ± 114K · ⚡ 132		henryk/bert-base-multilingual-cased-finetuned-polish	Question Answering · Updated May 19, 2021 · ± 220 · ⚡ 3

X Fine-tuning LLMs relies on obscure "magic formulas"

Fine-tuning LLMs in real life?

apples fell from trees in autumn.\nPeople enjoy watching movies together.\nApples can grow in trees.\nShe enjoys reading books every day.\nApples hang from trees in autumn.\nPeople often use umbrellas when?\nIce cream melts in hot weather.\nShe enjoys reading books about history.\nApples grow in trees.\nI enjoy reading books about history.\nShe jumped over the moon today.\nNumber Inst? I enjoy reading books in my free time.\nShe baked delicious chocolate chip cookies.\nApples fall from trees in autumn.\nA bird flew over the mountain peak.

```
epoch 1: 2%| [.....]
Sample 0: Yes, .....
Sample 1: Yes, .....
Sample 2: Today, .....
Sample 3: In, .....
Sample 4: For, .....
[1.1630859375, 1.3134765625, 1.140625]
```

✗ Fine-tuning LLMs is hard

Deep-Learning-Based Sequence Modeling: Recurrent Models (1)

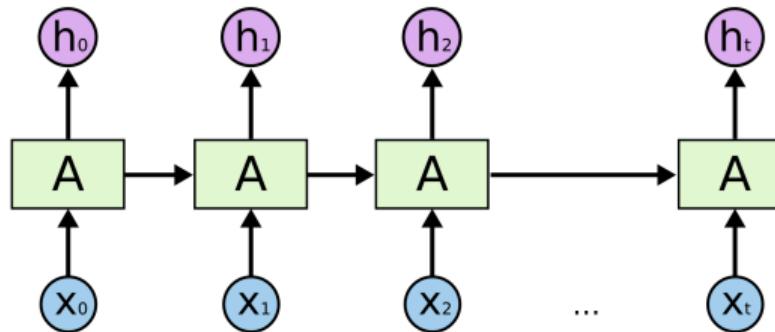


Figure 1 – Illustration of the Recurrent Neural Network (RNN, [18, 10]) architecture¹

- ✓ Keep token order
- ✓ Handle variable-length sequences
- ✓ Parameter sharing across the sequence
- ✗ Exploding and vanishing gradient
- ✗ Long-term dependencies
- ✗ Slow computing, no parallelization

¹ Olah, C. (2015). Understanding LSTM Networks. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Deep-Learning-Based Sequence Modeling: Recurrent Models (2)

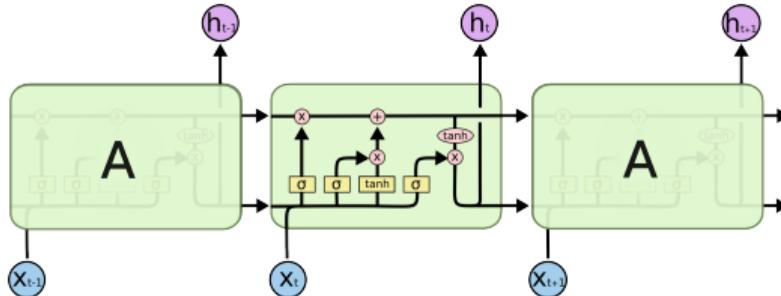


Figure 2 – Illustration of the Long Short Term Memory Neural Network (LSTM, [7]) architecture²

- ✓ Keep token order
- ✓ Handle variable-length sequences
- ✓ Parameter sharing across the sequence
- ✓ No exploding/vanishing gradient
- ✓ Long-term dependencies
- ✗ Slow computing, no parallelization

²Olah, C. (2015). Understanding LSTM Networks. <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

Attention Principle [1] and Transformer Model [20]

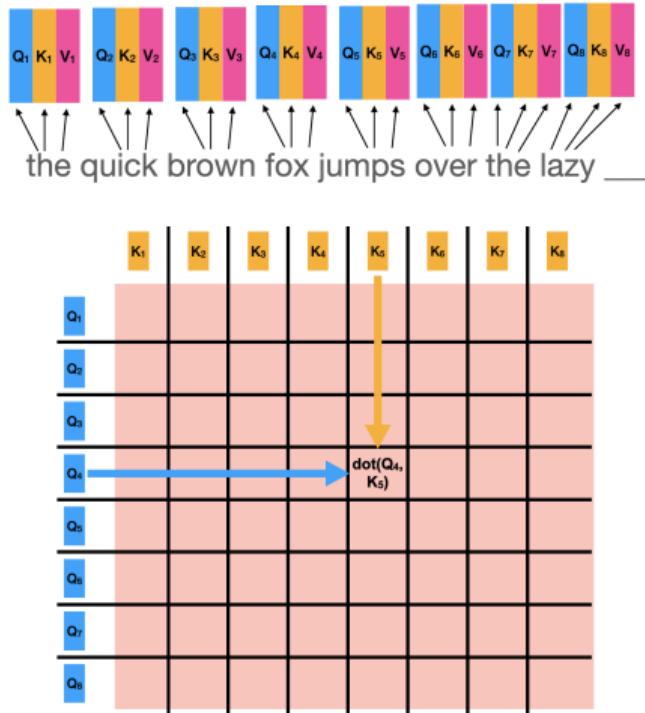


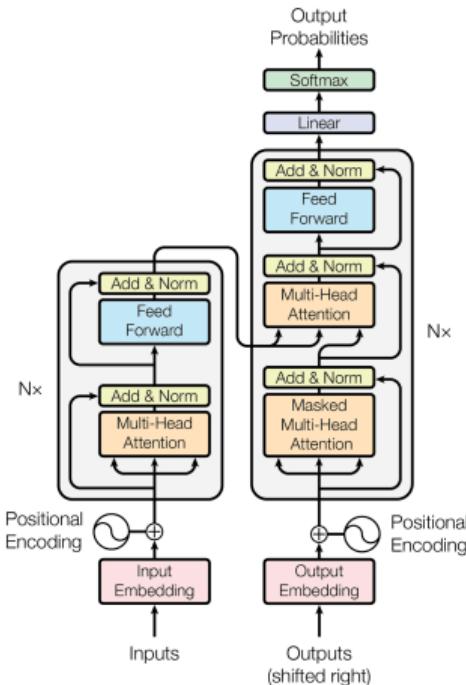
Figure 3 – Attention principle¹

- **Query Q :** current token asking for context
- **Key K :** all tokens defining where to focus
- **Value V :** all tokens information
- d_k : embedding dimension

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V$$

¹<https://learnopencv.com/attention-mechanism-in-transformer-neural-networks/>

Attention Principle [1] and Transformer Model [20]



- **Query** Q : current token asking for context
- **Key** K : all tokens defining where to focus
- **Value** V : all tokens information
- d_k : embedding dimension

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_k}} \right) V$$

Figure 4 – Transformer architecture from the original paper²

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

LLMs scale Transformers by stacking encoders and/or decoders together

- Parallelizable and optimized versions exist (e.g. quantization)
- Enable deeper and broader knowledge representation
- Large context window allows for more accurate generation

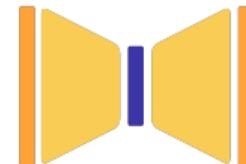


Figure 5 – Full Transformer



Figure 6 – Encoder



Figure 7 – Decoder

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

LLMs scale Transformers by stacking encoders and/or decoders together

- Parallelizable and optimized versions exist (e.g. quantization)
- Enable deeper and broader knowledge representation
- Large context window allows for more accurate generation

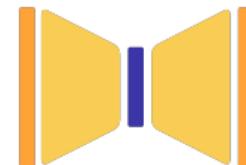


Figure 5 – Full Transformer

Fine-tuning adapts an LLM to a specific task through further parameter updates

- Can be performed with any LLM structure, but:
 - There are *required structures* for some specific tasks
 - There are *preferred models* for some specific tasks



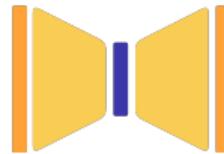
Figure 6 – Encoder



Figure 7 – Decoder

What Use-Cases of LLMs?

Model Structure



Task Examples

- Summarization
- Machine Translation
- Question Answering

Model examples

- BART [11], mBART [12]
- T5 [17], Flan-T5 [3]
- bert2BERT [2]



- Sequence Embedding
- Text Classification
- Regression

- BERT [5], mBERT [15]
- RoBERTa [13]
- DistilBERT [19]



- Text Completion
- Text Generation
- Code Generation

- GPT-3.5, GPT-4o
- Llama-3 [6]
- Qwen2.5 [16]

Inside a Decoder-Only LLM

① Prompt the LLM

What is the best way to learn music?

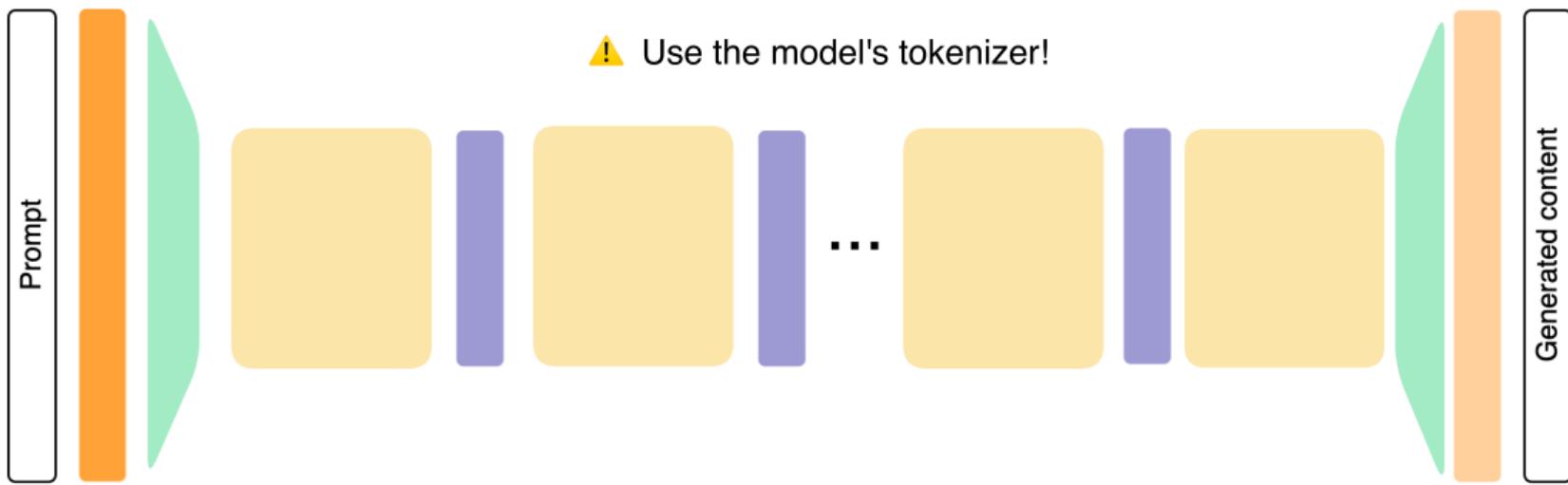


Inside a Decoder-Only LLM

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!



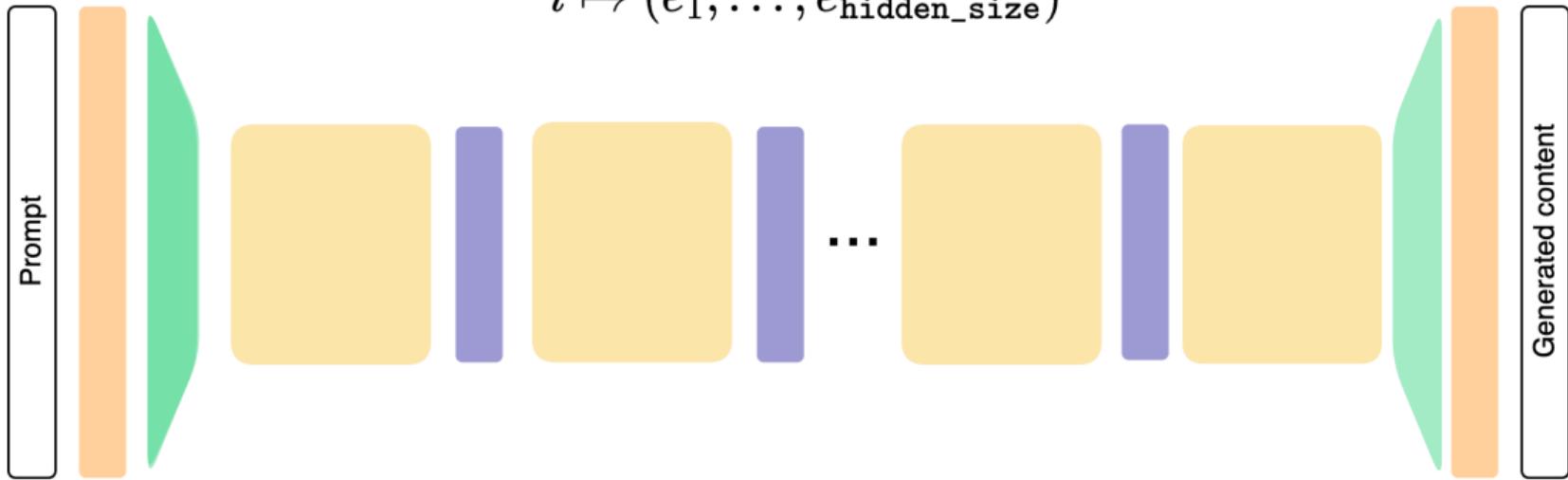
Inside a Decoder-Only LLM

3

Apply an Embedding layer

Embedding (vocab_size, hidden_size)

$$i \mapsto (e_1, \dots, e_{\text{hidden_size}})$$



Inside a Decoder-Only LLM

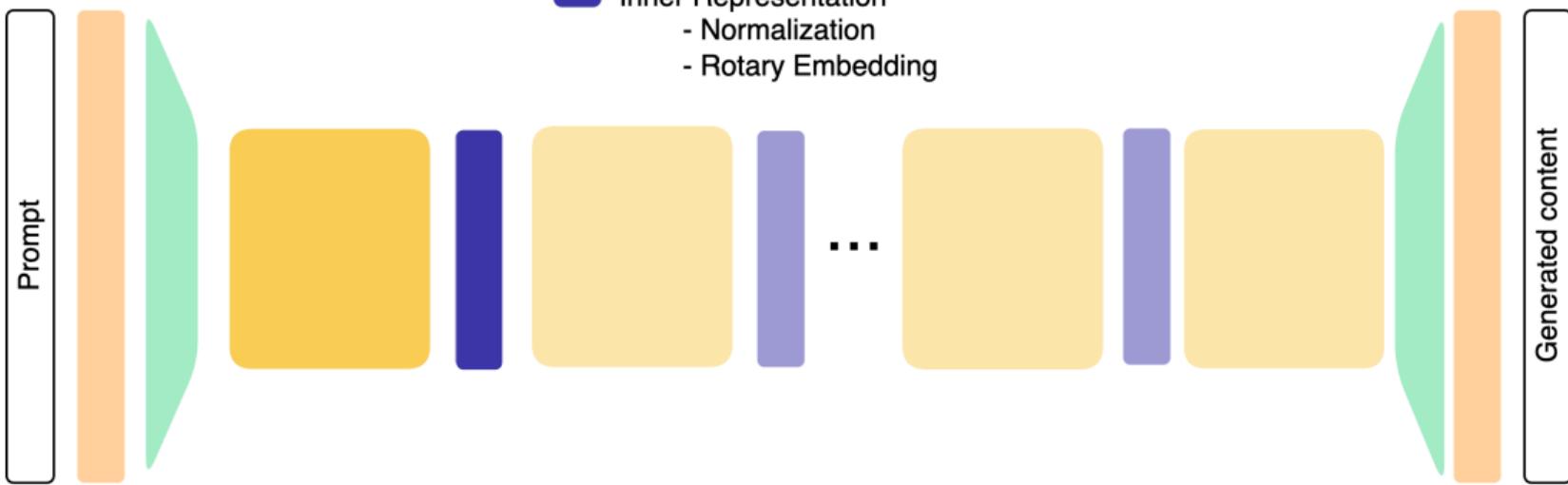
④ Going Through a Decoder Block

Transformer Decoder Blocks

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

Inner Representation

- Normalization
- Rotary Embedding



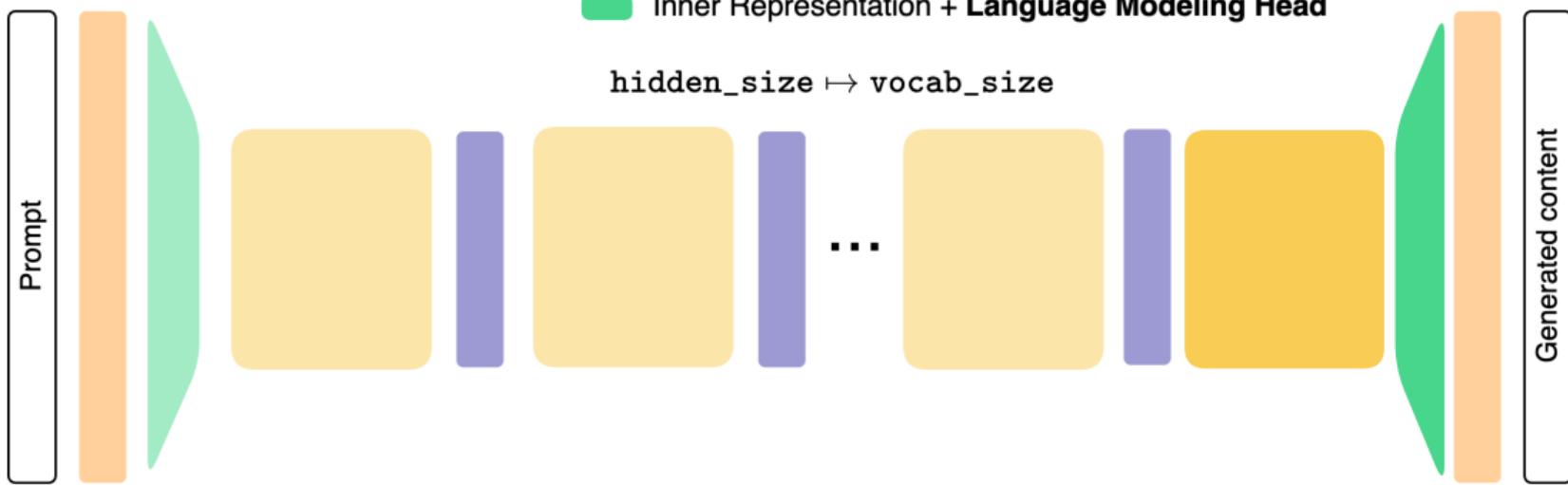
Inside a Decoder-Only LLM

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`



Inside a Decoder-Only LLM

⑥ Decode Generated Tokens in Natural Language



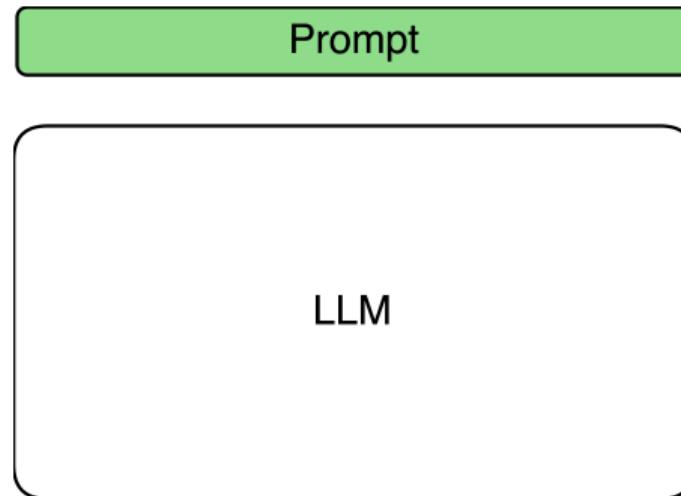
What is the best way to learn music?



The best way to learn music depends on your goals, interests, and learning style, but here are some effective strategies that can help:

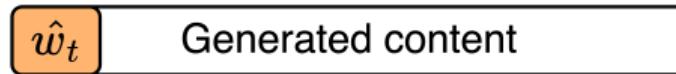


The Autoregressive Principle

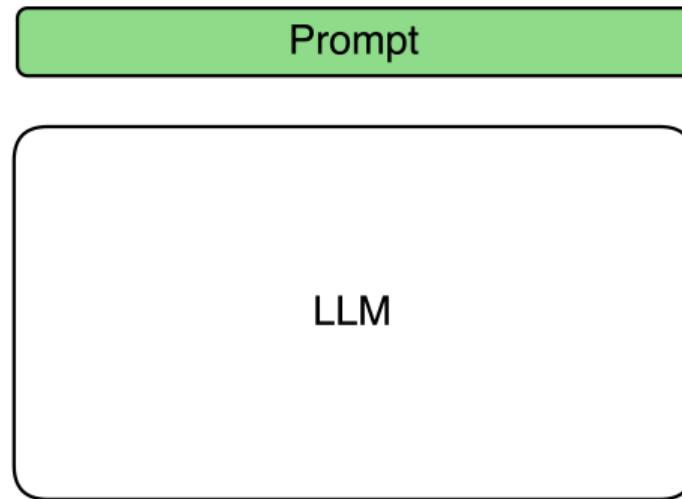


Context
Predicted token

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

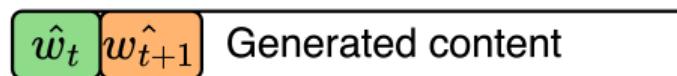


The Autoregressive Principle



Context
Predicted token

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



- **Top p:** adjust the range of tokens to consider regarding their probability.
- **Top k:** choose among the k more likely tokens. Default is 1 (most likely token).
- **Temperature:** control "creativity" by adding random noise to select less likely tokens.

How Large are Large Language Models?

Model	Parameters	Layers	Context Size
Gemini 1.5	200B?	-	10M
GPT-4 turbo	1.8T	120	128k
Claude 2.1	12B	-	200k

Side remark: most LLMs called "open-source" are actually open-weights!

Table: Some *closed-source* model specifications

Model	Parameters	Layers	Context Size
Llama 3.3	70B	80	128k
Phi 4	14B	40	16k
Qwen 2.5	0.5B-72B	24-80	32k - 128k
Mistral-v0.3	7B	32	32k

Table: Some *open-weights* model specifications

How Large are Large Language Models?

Model	Parameters	Layers	Context Size
Gemini 1.5	200B?	-	10M
GPT-4 turbo	1.8T	120	128k
Claude 2.1	12B	-	200k

Table: Some *closed-source* model specifications

Model	Parameters	Layers	Context Size
Llama 3.3	70B	80	128k
Phi 4	14B	40	16k
Qwen 2.5	0.5B-72B	24-80	32k - 128k
Mistral-v0.3	7B	32	32k

Table: Some *open-weights* model specifications

Side remark: most LLMs called "open-source" are actually open-weights!

- Most models involve several gigabytes in RAM GPU to perform inference
- Updating each parameter value during fine-tuning would be too costly

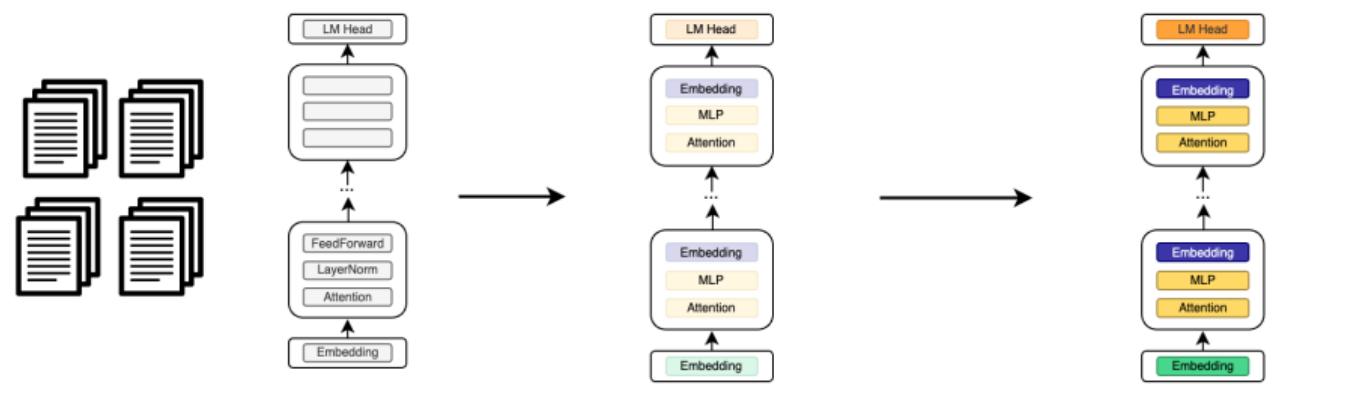
→ fine-tuning should be **efficient**

Fine-Tuning is an *Affinage*: the Cheese Analogy

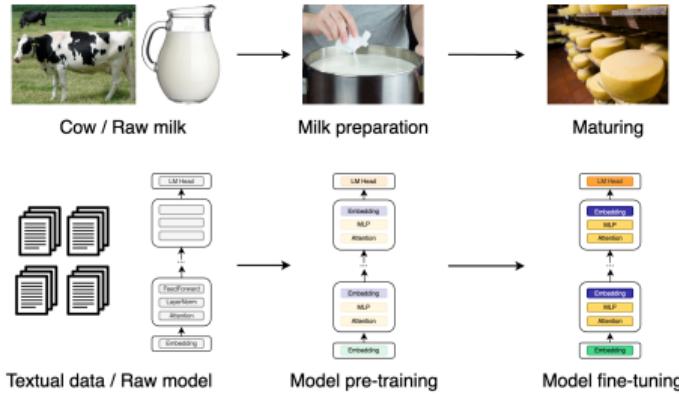
Fine-Tuning is an *Affinage*: the Cheese Analogy



Fine-Tuning is an Affinage: the Cheese Analogy



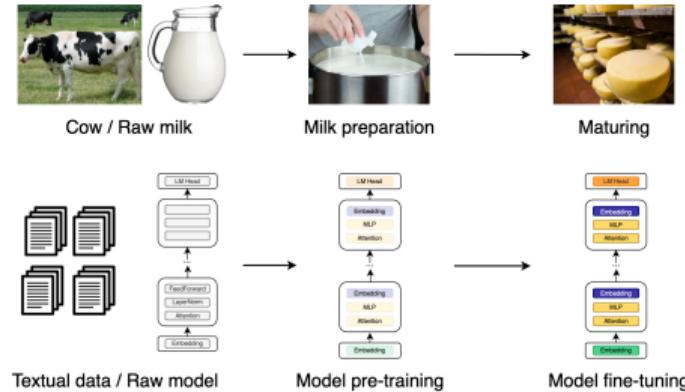
Fine-Tuning is an *Affinage*: the Cheese Analogy



Fine-tuning, as cheese maturing phase, modifies **in place** its given instance.

- Ensure fine-tuning is performed in the right conditions
- Check pre-trained baseline reliability
- Test several (adapted) pre-trained baselines for one fine-tuning experiment

Fine-Tuning is an *Affinage*: the Cheese Analogy



Fine-tuning, as cheese maturing phase, modifies **in place** its given instance.

- Ensure fine-tuning is performed in the right conditions
 - Check pre-trained baseline reliability
 - Test several (adapted) pre-trained baselines for one fine-tuning experiment
- fine-tuning should be **stable** and **consistent** with pre-trained baseline

Efficient Fine-Tuning With Adapters [8]

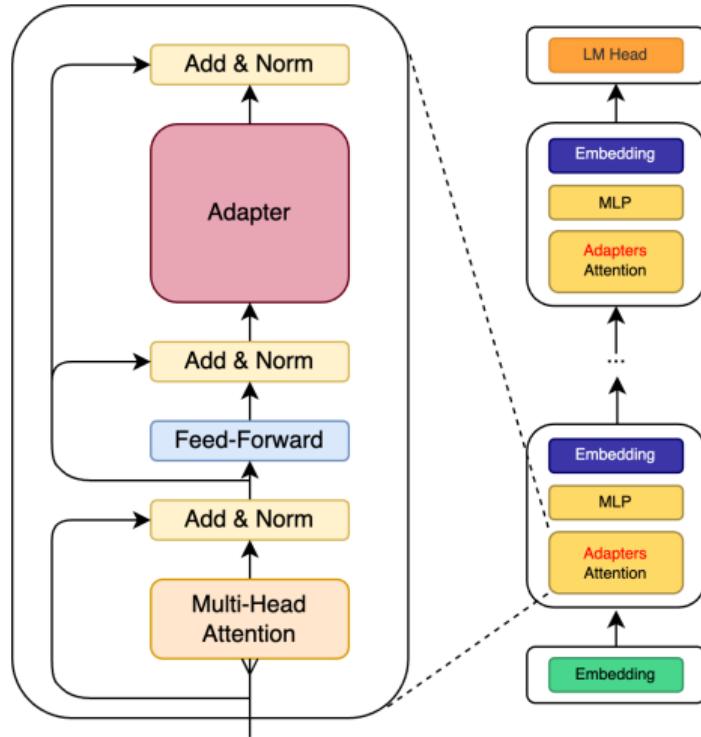


Figure 8 – An attention block with adapters

Adapters layers are inserted to enable parameter-efficient fine-tuning^a

Common usage:

- Freeze all pre-trained model layers
- Insert trainable MLP layers into attention blocks (query, value) and/or model head

^a<https://huggingface.co/PEFT>

^b<https://adapterhub.ml/>

Efficient Fine-Tuning With Adapters [8]

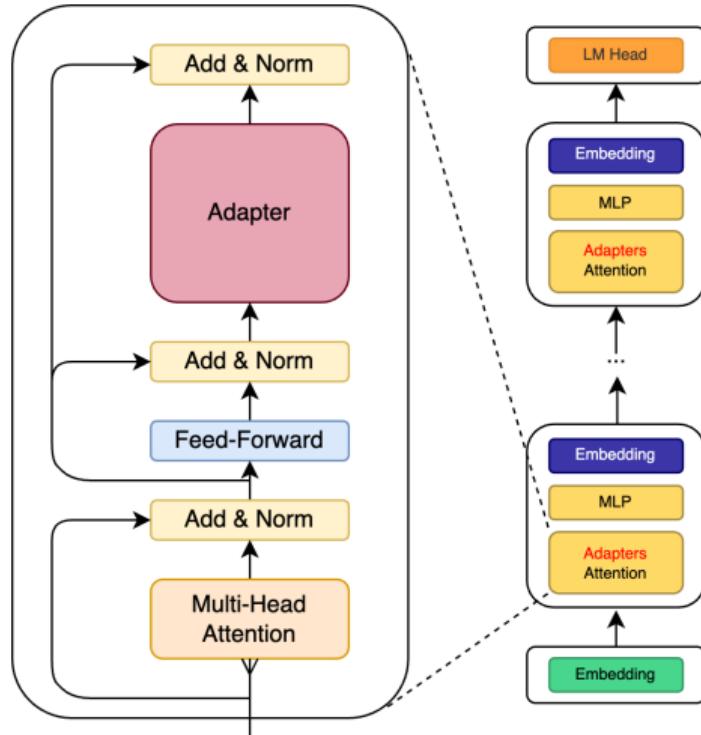


Figure 8 – An attention block with adapters

Adapters layers are inserted to enable parameter-efficient fine-tuning^a

Common usage:

- Freeze all pre-trained model layers
- Insert trainable MLP layers into attention blocks (query, value) and/or model head

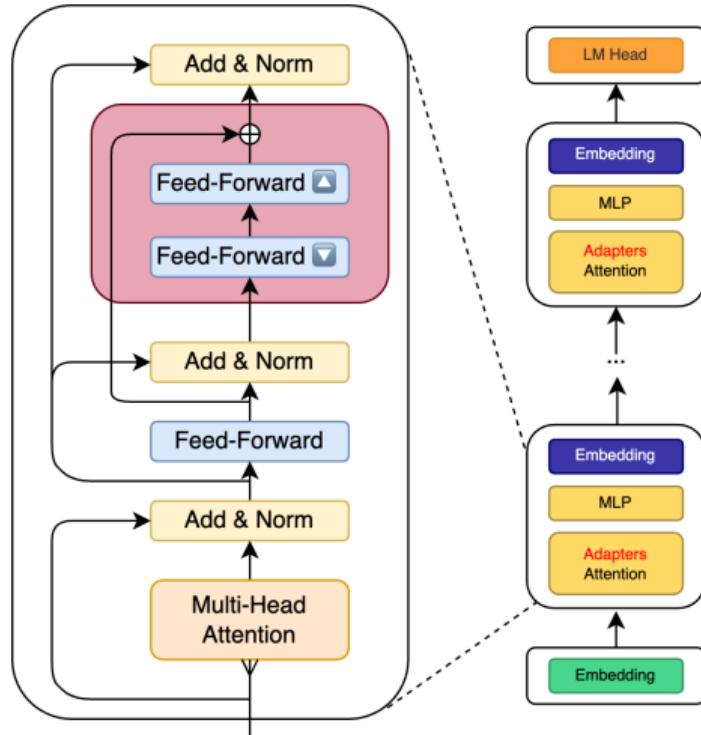
Advantages:

- Fewer param. updates than full fine-tuning
- Helps reduce catastrophic forgetting [14]
- Easy to share fine-tuned models^b

^a<https://huggingface.co/PEFT>

^b<https://adapterhub.ml/>

Efficient Fine-Tuning With LoRA Adapters [9]



LoRA: Low-Rank Adaptation A specific adapter block structure

- Information from the model can be represented (almost) equally well in a lower dimensional space
- The rank r of the lower dim. space should be determined by hyperparameter tuning
- Quantized versions: QLoRA [4]

Figure 9 – An attention block with LoRA adapters

Thanks for your attention!



Figure 10 – Practical session: <https://github.com/B-Gendron/tutorial-deeploria/lab/>

References I

- [1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *ICLR* abs/1409.0473 (2014). URL: <https://api.semanticscholar.org/CorpusID:11212020>.
- [2] Cheng Chen et al. *bert2BERT: Towards Reusable Pretrained Language Models*. 2021. arXiv: 2110.07143 [cs.CL]. URL: <https://arxiv.org/abs/2110.07143>.
- [3] Hyung Won Chung et al. *Scaling Instruction-Finetuned Language Models*. 2022. arXiv: 2210.11416 [cs.LG]. URL: <https://arxiv.org/abs/2210.11416>.
- [4] Tim Dettmers et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. 2023. arXiv: 2305.14314 [cs.LG]. URL: <https://arxiv.org/abs/2305.14314>.
- [5] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *NAACL*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics, 2019.

References II

- [6] Aaron Grattafiori et al. *The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783 [cs.AI]. URL: <https://arxiv.org/abs/2407.21783>.
- [7] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-term Memory”. In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
- [8] Neil Houlsby et al. *Parameter-Efficient Transfer Learning for NLP*. 2019. arXiv: 1902.00751 [cs.LG]. URL: <https://arxiv.org/abs/1902.00751>.
- [9] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: <https://arxiv.org/abs/2106.09685>.
- [10] M I Jordan. *Serial order: a parallel distributed processing approach. Technical report, June 1985-March 1986*. Tech. rep. California Univ., San Diego, La Jolla (USA). Inst. for Cognitive Science, May 1986. URL: <https://www.osti.gov/biblio/6910294>.

References III

- [11] Mike Lewis et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”. In: *ACL*. Ed. by Dan Jurafsky et al. Association for Computational Linguistics, 2020.
- [12] Yinhan Liu et al. *Multilingual Denoising Pre-training for Neural Machine Translation*. 2020. arXiv: 2001.08210 [cs.CL]. URL: <https://arxiv.org/abs/2001.08210>.
- [13] Yinhan Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *ArXiv* (2019).
- [14] Jonas Pfeiffer et al. “AdapterFusion: Non-Destructive Task Composition for Transfer Learning”. In: *CoRR* abs/2005.00247 (2020). arXiv: 2005.00247. URL: <https://arxiv.org/abs/2005.00247>.
- [15] Telmo Pires, Eva Schlinger, and Dan Garrette. “How multilingual is Multilingual BERT?” In: *ArXiv* (2019).

References IV

- [16] Qwen et al. *Qwen2.5 Technical Report*. 2025. arXiv: 2412.15115 [cs.CL]. URL: <https://arxiv.org/abs/2412.15115>.
- [17] Colin Raffel et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *ArXiv* abs/1910.10683 (2019).
- [18] David E. Rumelhart and James L. McClelland. “Learning Internal Representations by Error Propagation”. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. 1987, pp. 318–362.
- [19] Victor Sanh et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: *CoRR* abs/1910.01108 (2019). arXiv: 1910.01108. URL: <http://arxiv.org/abs/1910.01108>.
- [20] Ashish Vaswani et al. “Attention Is All You Need”. In: *CoRR* (2017).