Classification with LLMs

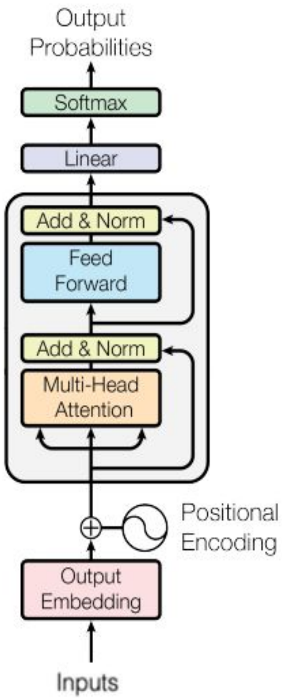
Tanalp Agustoslu
27.01.2025

Structure

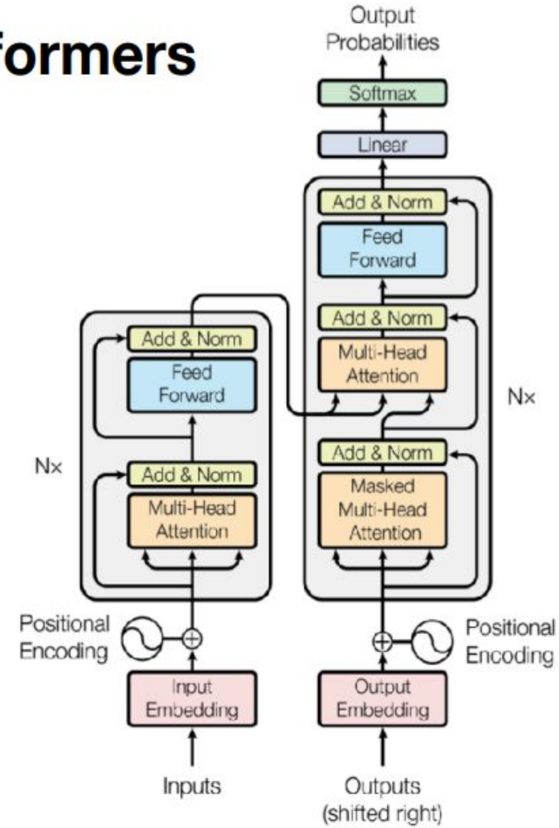
- Types of Transformers
- Evolution of LLMs
- Attention Mechanism, Context Window, Bottleneck of LLMs
- Parameter Efficient Fine-tuning
- Quantization - QLoRA
- Instruction Tuning
- Experiments & Results
- Limitations & Future Work

Transformers

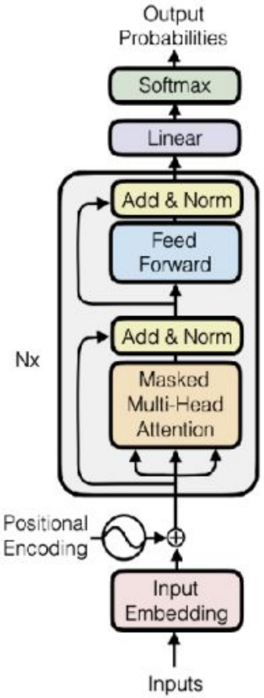
Three Types of Transformers



Encoder-Only Model (e.g. BERT)



Encoder-Decoder Model (Vaswani et al., 2017)

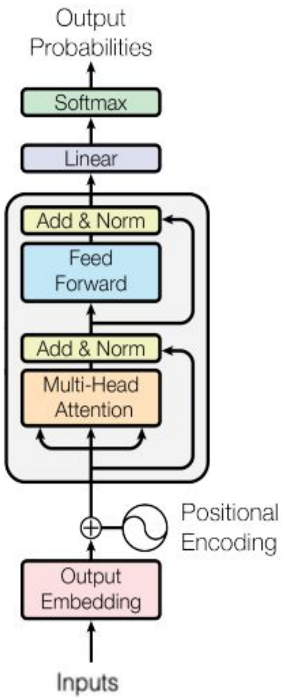


Decoder-Only Model (e.g. GPT, Llama)

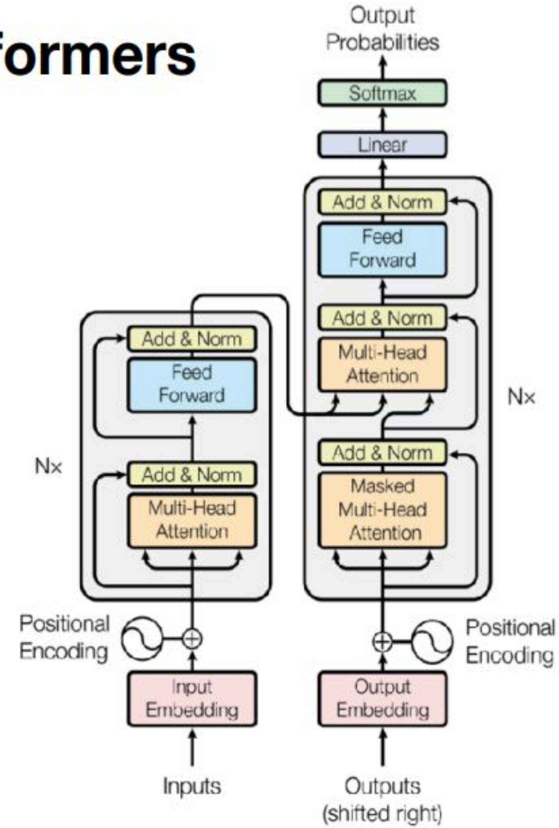
Slide adapted from Plank, Barbara

Transformers

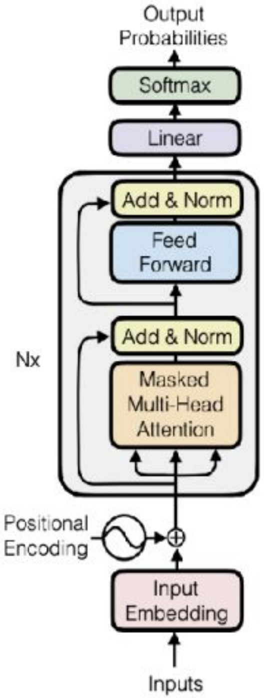
Three Types of Transformers



Encoder-Only Model (e.g. BERT)

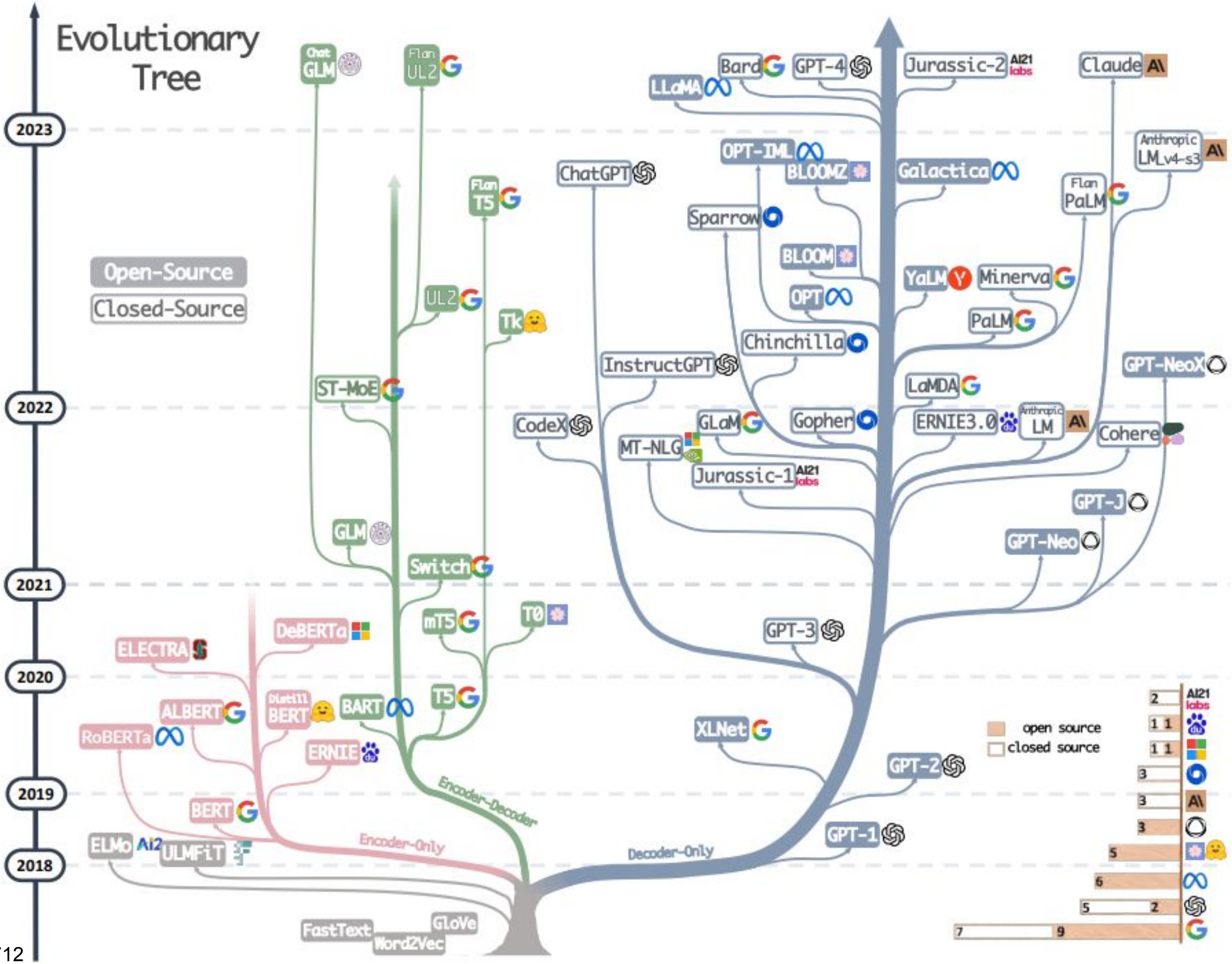


Encoder-Decoder Model (Vaswani et al., 2017)

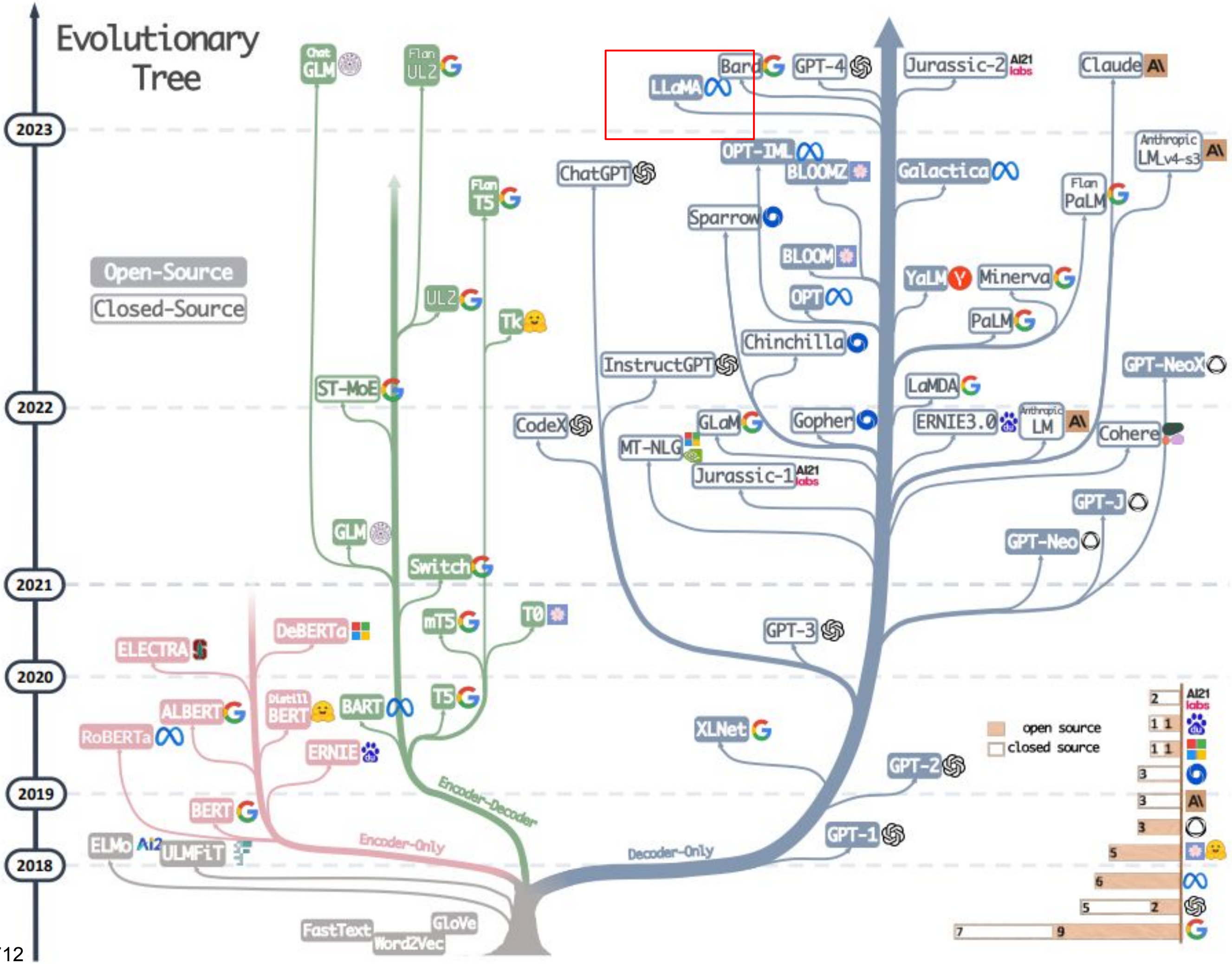


Decoder-Only Model (e.g. GPT, Llama)

Slide adapted from Plank, Barbara

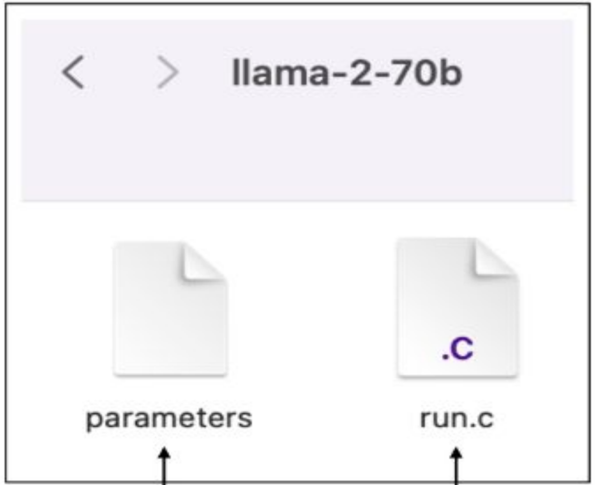


https://arxiv.org/pdf/2304.13712



https://arxiv.org/pdf/2304.13712

Large Language Model (LLM)



140GB

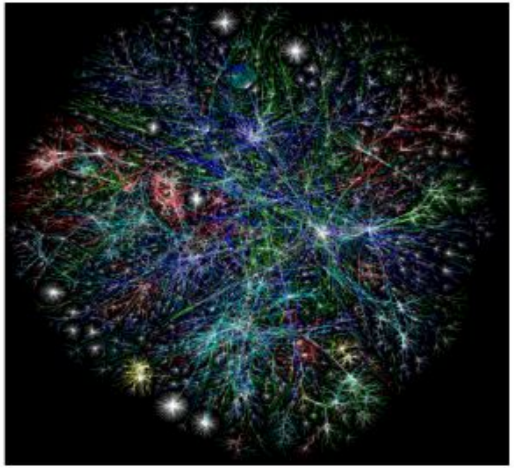
~500 lines
of C code

MacBook

Slide adapted from
Karpathy, Andrej

Training them is more involved.

Think of it like compressing the internet.



Chunk of the internet,
~10TB of text

6,000 GPUs for 12 days, ~$2M
~1e24 FLOPS

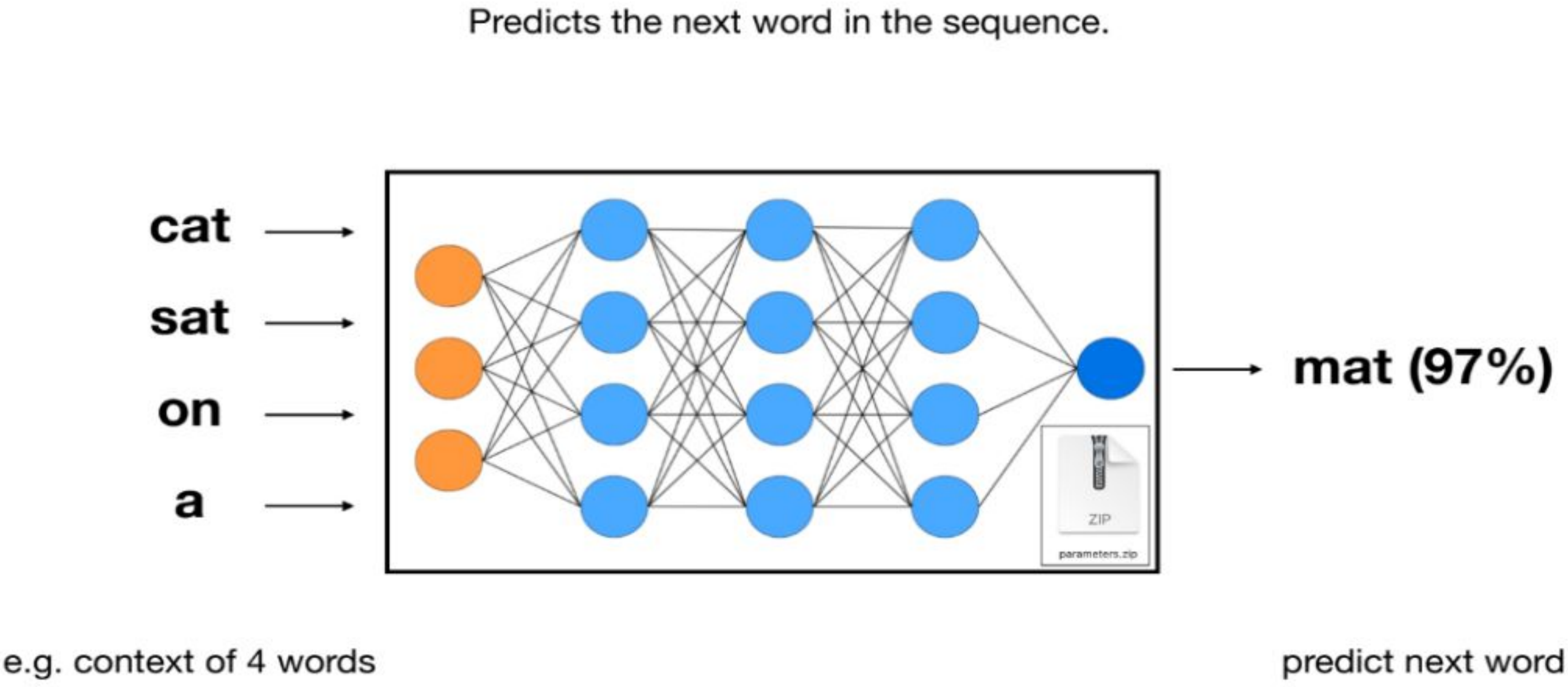
parameters.zip

~140GB file

*numbers for Llama 2 70B

Slide adapted from
Karpathy, Andrej

“You shall know a word by the company it keeps!” - Firth, 1957



Slide adapted from
Karpathy, Andrej

Next word prediction forces the neural network to learn a lot about the world:

Ruth Marianna Handler (née **Mosko**; November 4, 1916 – April 27, 2002) was an American businesswoman and inventor. She is best known for inventing the Barbie doll in 1959,[2] and being co-founder of toy manufacturer Mattel with her husband Elliot, as well as serving as the company's first president from 1945 to 1975.[3]

The Handlers were forced to resign from Mattel in 1975 after the Securities and Exchange Commission investigated the company for falsifying financial documents.[3][4]

Early life [edit]

Ruth Marianna Mosko[5][2][3] was born on November 4, 1916, in Denver, Colorado, to Polish-Jewish immigrants Jacob Moskowicz, a blacksmith, and Ida Moskowicz, née Rubenstein.[6]

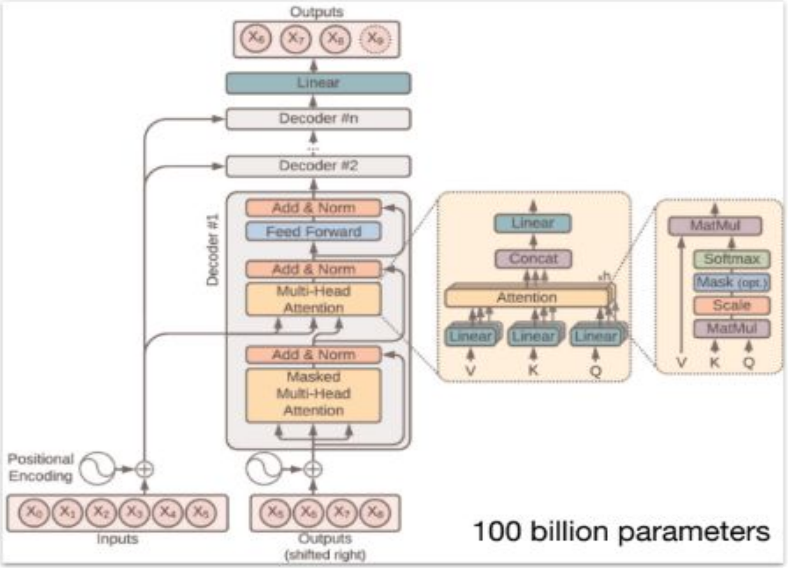
She married her high school boyfriend, Elliot Handler, and moved to Los Angeles in 1938, where she found work at Paramount.[7]

Ruth Handler

Handler in 1961

Born	Ruth Marianna Mosko November 4, 1916 Denver, Colorado, U.S.
Died	April 27, 2002 (aged 85)[1] Los Angeles, California, U.S.

How does it work?



Little is known in full detail…

- Billions of parameters are dispersed through the network
- We know how to iteratively adjust them to make it better at prediction.
- We can measure that this works, but we don't really know how the billions of parameters collaborate to do it.

They build and maintain some kind of knowledge database, but it is a bit strange and imperfect:

Recent viral example: "reversal curse"

Q: "Who is Tom Cruise's mother"?
A: Mary Lee Pfeiffer ✅

Q: "Who is Mary Lee Pfeiffer's son?"
A: I don't know ❌

=> think of LLMs as mostly inscrutable artifacts, develop correspondingly sophisticated evaluations.

Slide adapted from
Karpathy, Andrej

Short distance

One mole of carbon dioxide

Long distance

Harry Potter was a highly unusual boy in many ways. For one thing, he hated the summer holidays more than any other time of year. For another, he really wanted to do his homework but was forced to do it in secret, in the dead of night. And he also happened to be a wizard.

It was nearly midnight, and he was lying on his stomach in bed, the blankets drawn right over his head like a tent, a flashlight in one hand and a large leather-bound book (A History of Magic by Bathilda Bagshot) propped open against the pillow. Harry moved the tip of his eagle-feather quill down the page, frowning as he looked for something that would help him write his essay, "Witch Burning in the Fourteenth Century Was Completely Pointless discuss."

The quill paused at the top of a likely-looking paragraph. Harry Pushed his round glasses up the bridge of his nose, moved his flashlight closer to the book, and read:

Non-magic people (more commonly known as Muggles) were particularly afraid of magic in medieval times, but not very good at recognizing it. On the rare occasion that they did catch a real witch or wizard, burning had no effect whatsoever. The witch or wizard would perform a basic Flame Freezing Charm and then pretend to shriek with pain while enjoying a gentle, tickling sensation. Indeed, Wendelin the Weird enjoyed being burned so much that she allowed herself to be caught no less than fortyseven times in various disguises.

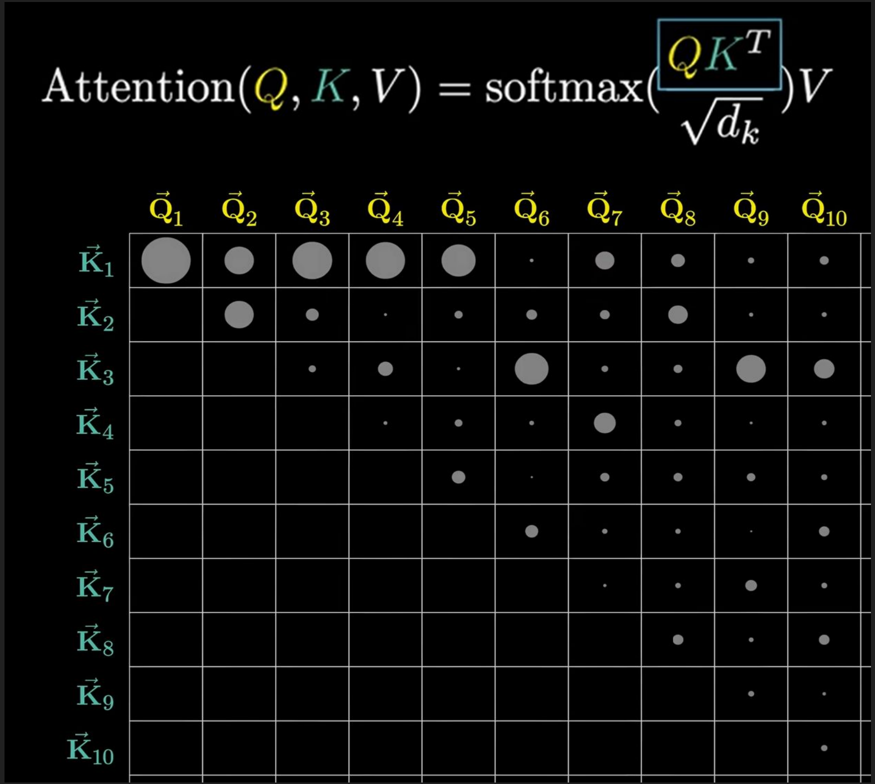
Harry put his quill between his teeth and reached underneath his pillow for his ink bottle and a roll of parchment. Slowly and very carefully he unscrewed the ink bottle, dipped his quill into it, and began to write, pausing every now and then to listen, because if any of the Dursleys heard the scratching of his quill on their way to the bathroom, he'd probably find himself locked in the cupboard under the stairs for the rest of the summer.

The Dursley family of number four, Privet Drive, was the reason that Harry never enjoyed his summer holidays. Uncle Vernon, Aunt Petunia, and their son, Dudley, were Harry's only living relatives. They were Muggles, and they had a very medieval attitude toward magic. Harry's dead parents, who had been a witch and wizard themselves, were never mentioned under the Dursleys' roof For years, Aunt Petunia and Uncle Vernon had hoped that if they kept Harry as downtrodden as possible, they would be able to squash the magic out of him. To their fury, they had been unsuccessful. These days they lived in terror of anyone finding out that Harry had spent most of the last two years at Hogwarts School of Witchcraft and Wizardry. The most they could do, however, was to lock away Harry's spellbooks, wand, cauldron, and broomstick at the start of the summer break, and forbid him to talk to the neighbors.

This separation from his spellbooks had been a real problem for Harry, because his teachers at Hogwarts had given him a lot of holiday work. One of the essays, a particularly nasty one about shrinking potions, was for Harry's least favorite teacher, Professor

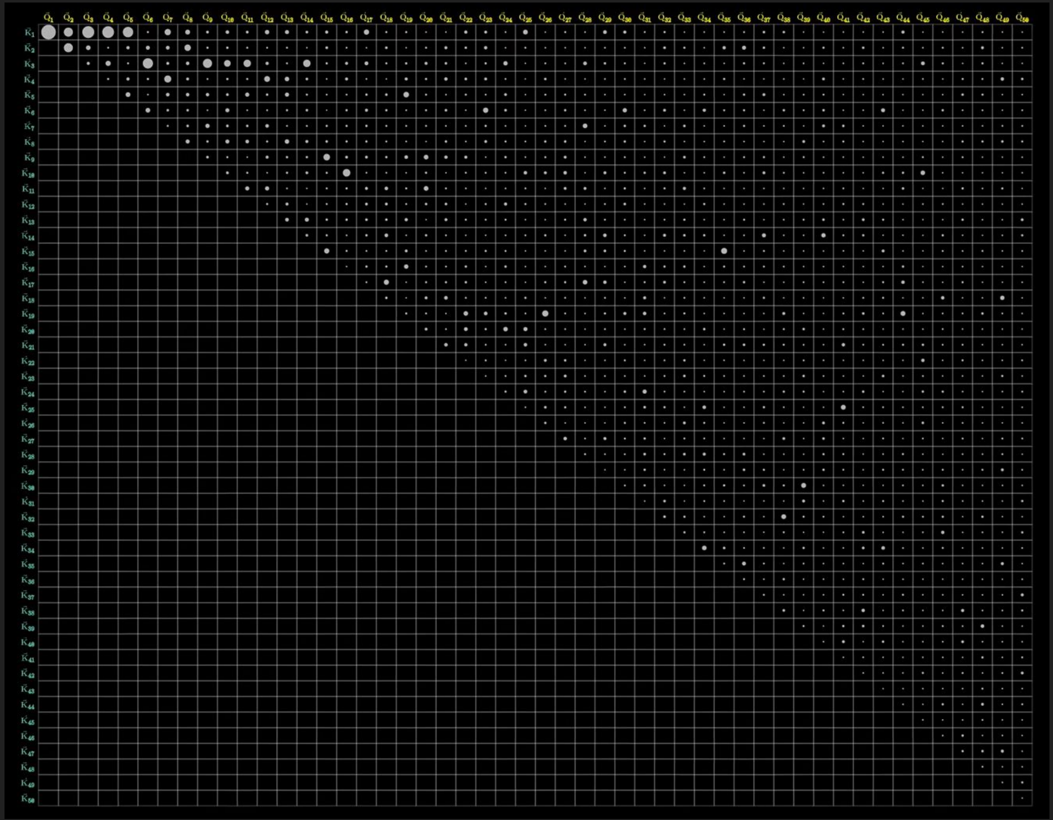
Slide adapted from Sanderson, Grant(3Blue1Brown)

The size of the attention pattern is equal to the half of the square of the context size.

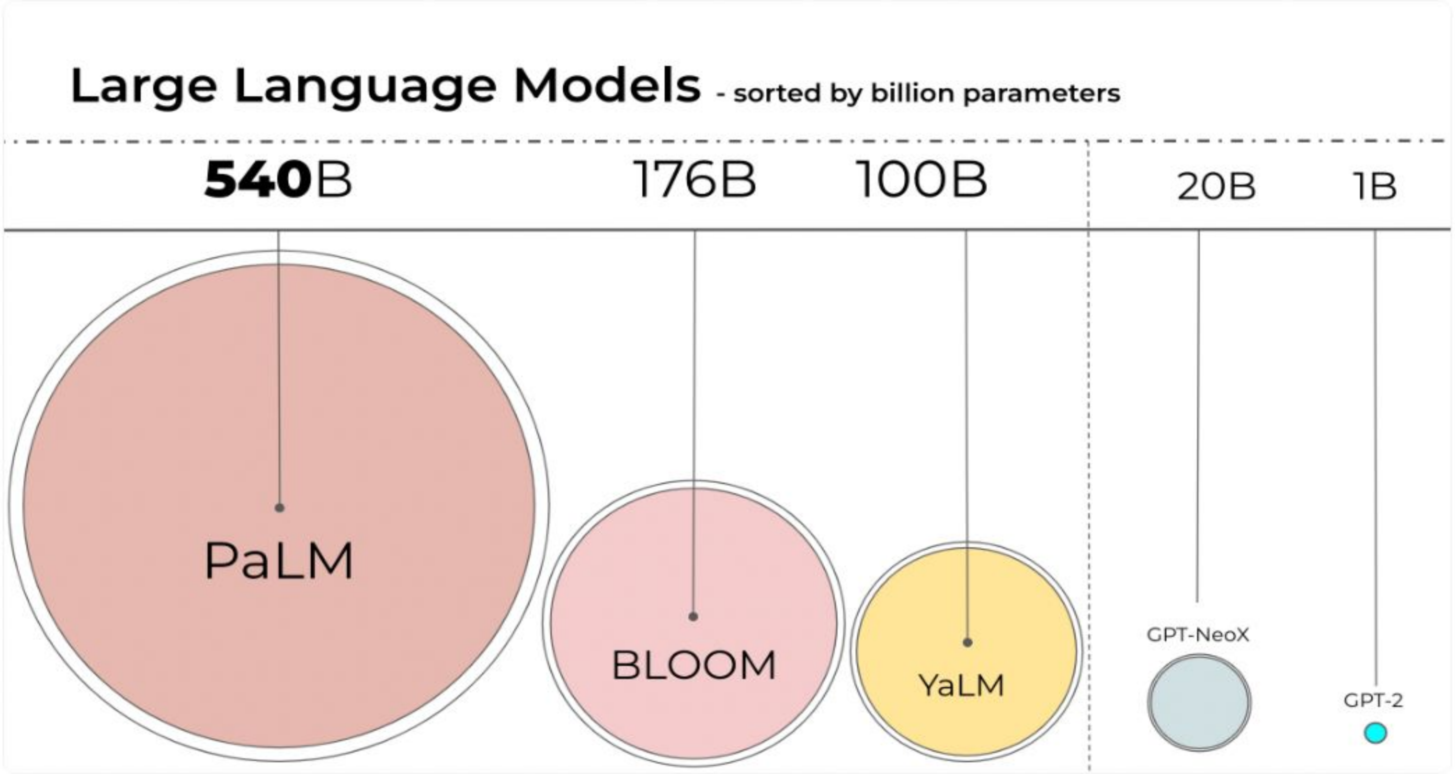


Slide adapted from Sanderson, Grant(3Blue1Brown)

Cost of predicting Professor Snape



Slide adapted from Sanderson, Grant(3Blue1Brown)



Therefore, these models are hard to run on easily accessible devices. For example, just to do inference on BLOOM-176B, you would need to have 8x 80GB A100 GPUs (~$15k each). To fine-tune BLOOM-176B, you'd need 72 of these GPUs! Much larger models, like PaLM would require even more resources.

https://huggingface.co/blog/hf-bitsandbytes-integration

Fortunately we are Gryffindor, so Hermione is in the team. “Capacious Extremis!”

RoFormer: Enhanced Transformer with Rotary Position Embedding

Jianlin Su
Zhuiyi Technology Co., Ltd.
Shenzhen
bojonesu@wezhuiyi.com

Yu Lu
Zhuiyi Technology Co., Ltd.
Shenzhen
julianlu@wezhuiyi.com

Shengfeng Pan
Zhuiyi Technology Co., Ltd.
Shenzhen
nickpan@wezhuiyi.com

Ahmed Murtadha
Zhuiyi Technology Co., Ltd.
Shenzhen
mengjiayi@wezhuiyi.com

Bo Wen
Zhuiyi Technology Co., Ltd.
Shenzhen
brucewen@wezhuiyi.com

Yunfeng Liu
Zhuiyi Technology Co., Ltd.
Shenzhen
glenliu@wezhuiyi.com

LoRA: Low-Rank Adaptation of Large Language Models

Edward Hu* **Yelong Shen*** **Phillip Wallis** **Zeyuan Allen-Zhu**
Yuanzhi Li **Shean Wang** **Lu Wang** **Weizhu Chen**
Microsoft Corporation
{edwardhu, yeshe, phwallis, zeyuana,
yuanzhil, swang, luw, wzchen}@microsoft.com
yuanzhil@andrew.cmu.edu
(Version 2)

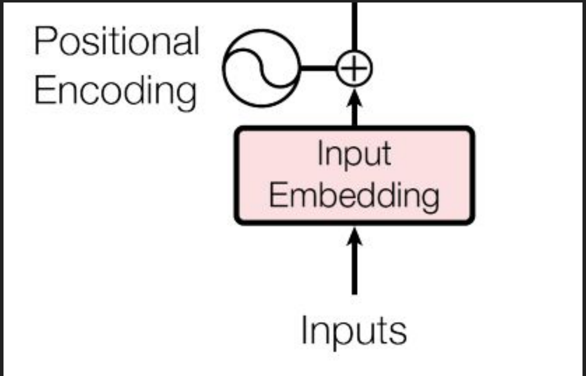
QLoRA: Efficient Finetuning of Quantized LLMs

Tim Dettmers* **Artidoro Pagnoni*** **Ari Holtzman**

Luke Zettlemoyer

University of Washington
{dettmers,artidoro,ahai,lsz}@cs.washington.edu

Positional Encoding



$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

In this paper, we introduce a novel method, namely Rotary Position Embedding(RoPE), to leverage the positional information into the learning process of PLMS. Specifically, RoPE encodes the absolute position with a rotation matrix and meanwhile incorporates the explicit relative position dependency in self-attention formulation. Note that the proposed RoPE is prioritized over the existing methods through valuable properties, including the sequence length flexibility, decaying inter-token dependency with increasing relative distances, and the capability of equipping the linear self-attention with relative position encoding. Experimental results on various long text classification benchmark datasets show that the enhanced transformer with rotary position embedding, namely RoFormer, can give better performance compared to baseline alternatives and thus demonstrates the efficacy of the proposed RoPE.

https://arxiv.org/pdf/1706.03762
https://arxiv.org/pdf/2104.09864

RoPE Scaling

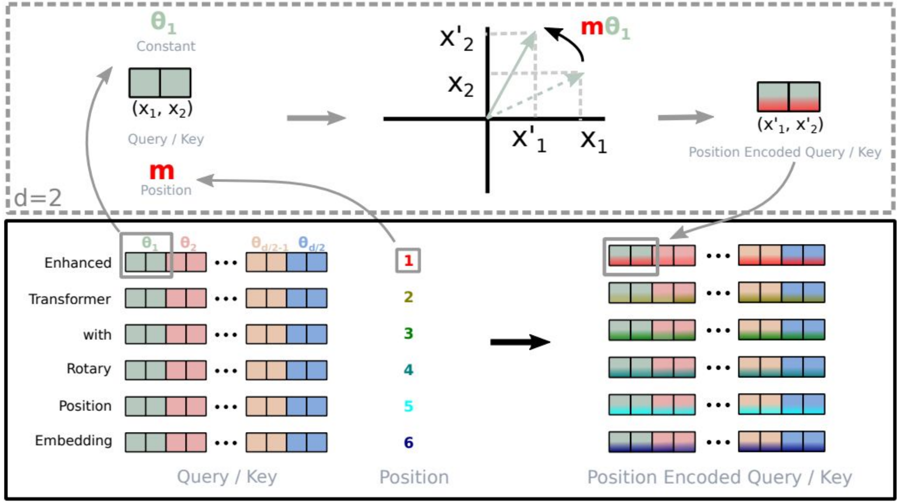
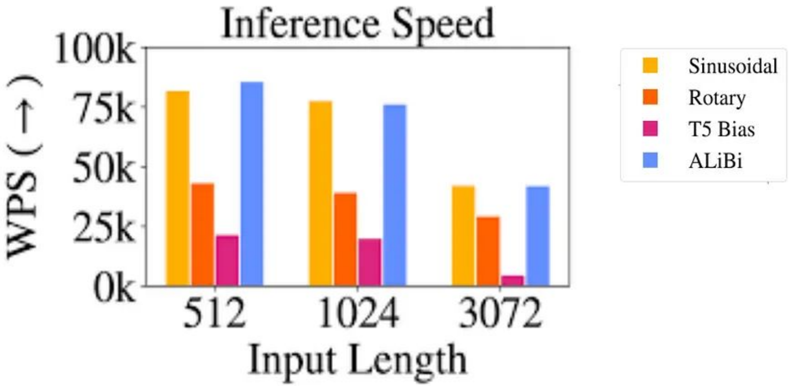


Figure 1: Implementation of Rotary Position Embedding(RoPE).



https://arxiv.org/pdf/2104.09864
https://medium.com/ai-insights-cobet/rotary-positional-embeddings-a-detailed-look-and-comprehensive-understanding-4ff66a874d83

Parameter Efficient Fine Tuning (PEFT) - LoRA

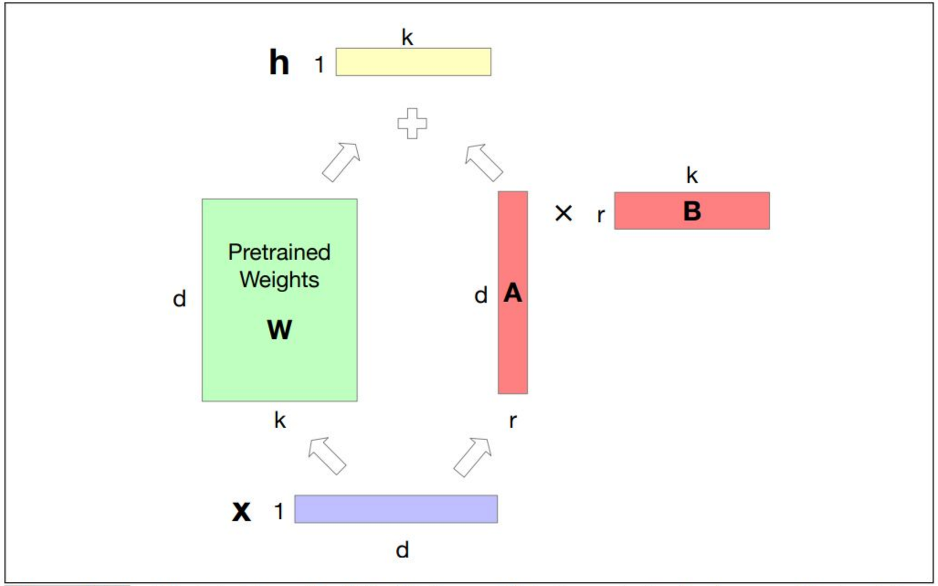


Figure 10.8 The intuition of LoRA. We freeze **W** to its pretrained values, and instead fine-tune by training a pair of matrices **A** and **B**, updating those instead of **W**, and just sum **W** and the updated **AB**.

"We propose Low-Rank Adaptation, or LoRA, which freezes the pretrained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. Compared to GPT-3 175B fine-tuned with Adam, LoRA can reduce the number of trainable parameters by 10,000 times and the GPU memory requirement by 3 times. LoRA performs on-par or better than finetuning in model quality on RoBERTa, DeBERTa, GPT-2, and GPT-3, despite having fewer trainable parameters, a higher training throughput, and, unlike adapters, no additional inference latency. (Hu et al., 2021)."

https://web.stanford.edu/~jurafsky/slp3/10.pdf
https://arxiv.org/pdf/2106.09685

QLoRA

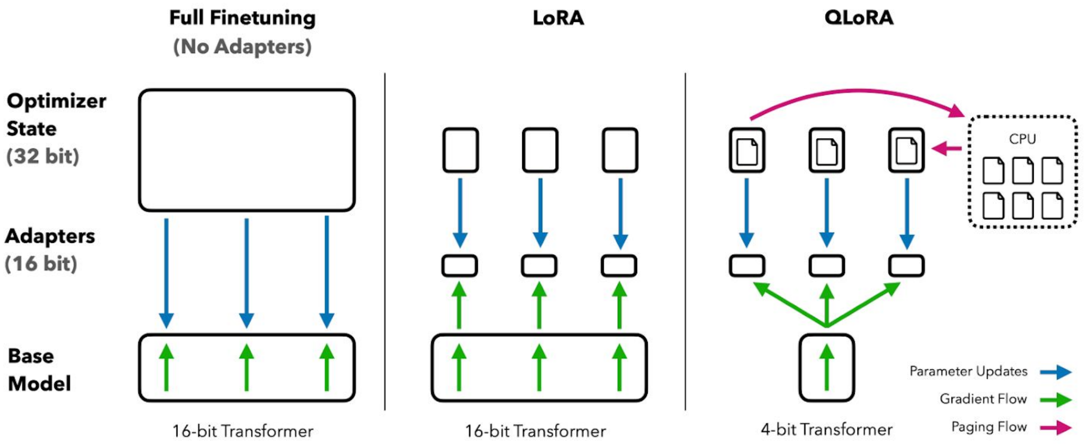


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

https://arxiv.org/pdf/2305.14314

Instruction Tuning with Unsloth

```
alpaca_prompt = """Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

### Instruction:
{}

### Input:
{}

### Response:
{}
"""
```

Instruction Tuning with Unsloth

```
model, tokenizer = FastLanguageModel.from_pretrained(
    model_name="unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit",
    max_seq_length=4600,
    dtype= None,
    load_in_4bit= True,
    token="hf_token",
)

model = FastLanguageModel.get_peft_model(
    model,
    r= 16,
    target_modules=["q_proj", "k_proj", "v_proj", "o_proj",
                    "gate_proj", "up_proj", "down_proj"],
    lora_alpha= 16,
    lora_dropout= 0,
    bias= "none",
    use_gradient_checkpointing= "unsloth",
    random_state= 3407,
    use_rslora= False,
    loftq_config= None,
)
```

Instruction Tuning with Unsloth

```python
trainer = SFTTrainer(
    model=model,
    tokenizer=tokenizer,
    train_dataset= dataset,
    dataset_text_field= "text",
    max_seq_length= 4600,
    dataset_num_proc= 2,
    packing= False,
    args=TrainingArguments(
        per_device_train_batch_size= 2,
        gradient_accumulation_steps= 4,
        warmup_steps= 5,
        max_steps= 200,
        learning_rate= 2e-4,
        fp16= not is_bfloat16_supported(),
        bf16= is_bfloat16_supported(),
        logging_steps= 1,
        optim= "adamw_8bit",
        weight_decay= 0.01,
        lr_scheduler_type= "linear",
        seed= 3407,
        output_dir="output_path",
        report_to="wandb",
    ),
)
```

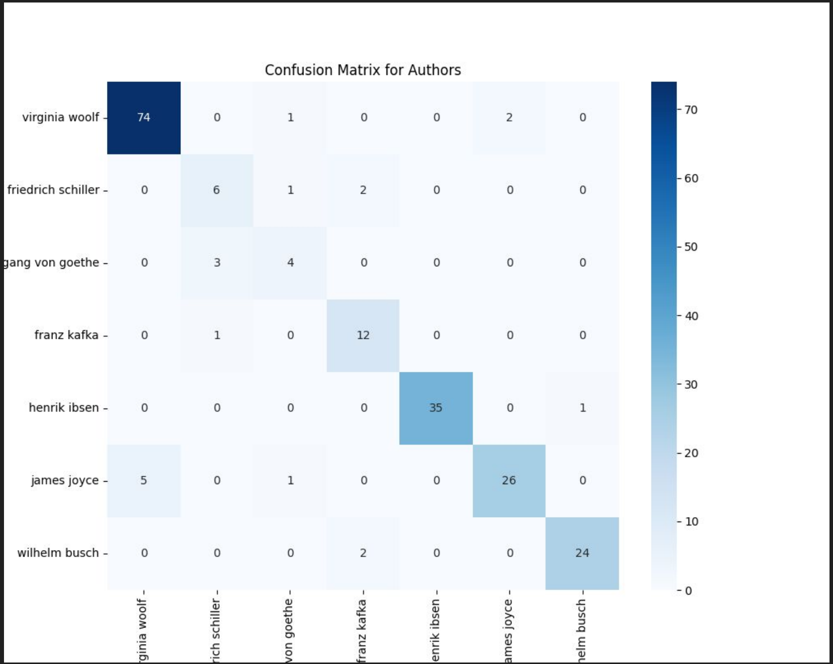
```
Letters Data Train Stats: (500 random instances)
[200, 1.2214837500452995, {"train_runtime": 220.2717,
"train_samples_per_second": 7.264, "train_steps_per_second": 0.908,
"total_flos": 1.5138050631204864e+16, "train_loss": 1.2214837500452995,
"epoch": 3.176}]
```

```
News Data Train Stats: (500 random instances)
[200, 0.9034367097914219, {"train_runtime": 217.9002,
"train_samples_per_second": 7.343, "train_steps_per_second": 0.918,
"total_flos": 8826574976188416.0, "train_loss": 0.9034367097914219, "epoch":
3.176}]
```

```
Sentiment Data Train Stats: (500 random instances)
[400, 1.625782641917467, {"train_runtime": 458.0575,
"train_samples_per_second": 6.986, "train_steps_per_second": 0.873,
"total_flos": 6.570031342829568e+16, "train_loss": 1.625782641917467, "epoch":
6.352}]
```

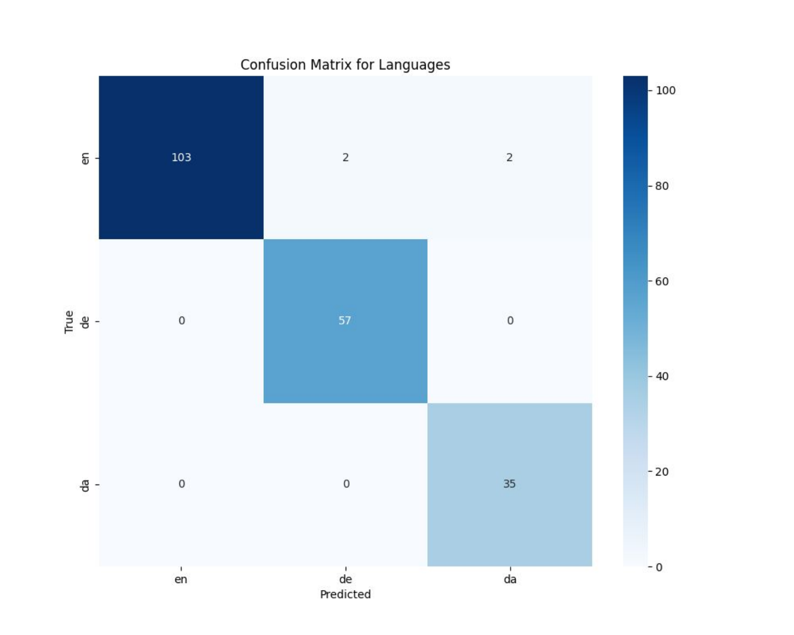
Classification for Authors:

	precision	recall	f1-score	support
franz kafka	0.75	0.92	0.83	13
friedrich schiller	0.60	0.67	0.63	9
henrik ibsen	1.00	0.97	0.99	36
james joyce	0.93	0.81	0.87	32
johann wolfgang von goethe	0.57	0.57	0.57	7
virginia woolf	0.94	0.96	0.95	77
wilhelm busch	0.96	0.92	0.94	26
accuracy			0.91	200
macro avg	0.82	0.83	0.82	200
weighted avg	0.91	0.91	0.91	200

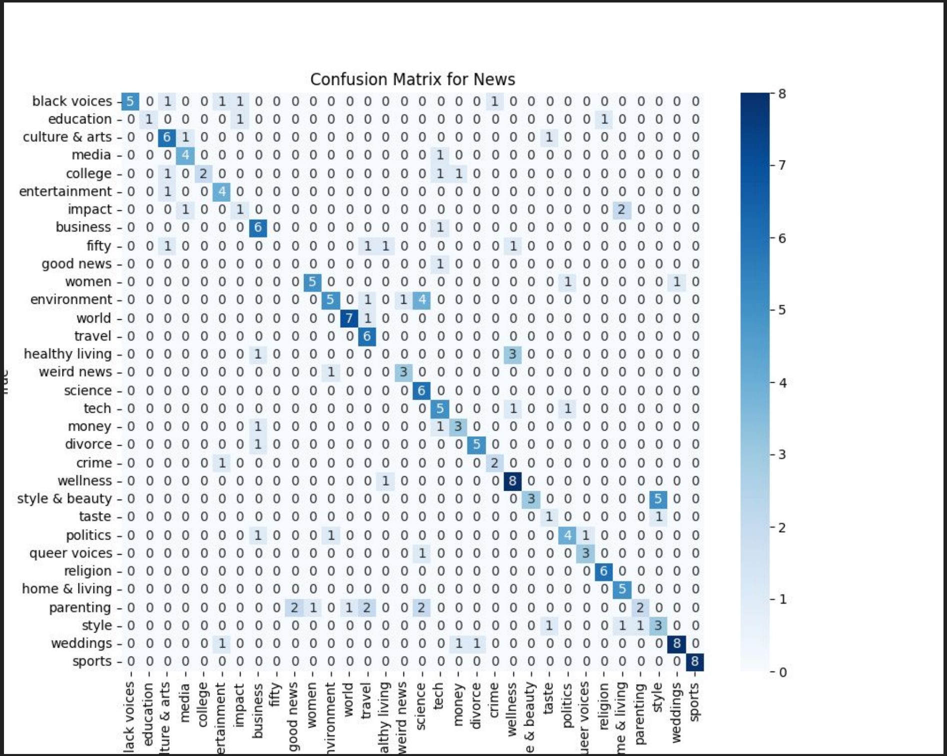


Classification for Languages:

	precision	recall	f1-score	support
da	0.95	1.00	0.97	35
de	0.97	1.00	0.98	57
en	1.00	0.95	0.98	108
unknown	0.00	1.00	0.00	0
accuracy			0.97	200
macro avg	0.73	0.99	0.73	200
weighted avg	0.98	0.97	0.98	200

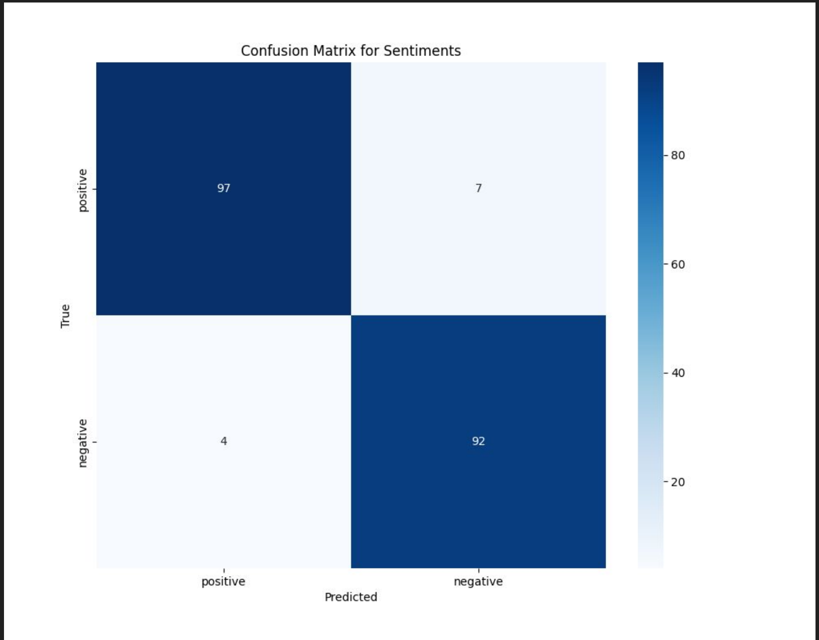


	precision	recall	f1-score	support
black voices	1.00	0.56	0.71	9
business	0.60	0.75	0.67	8
college	1.00	0.40	0.57	5
crime	0.67	0.67	0.67	3
culture & arts	0.60	0.67	0.63	9
divorce	0.83	0.83	0.83	6
education	1.00	0.33	0.50	3
entertainment	0.57	0.57	0.57	7
environment	0.71	0.45	0.56	11
fifty	1.00	0.00	0.00	4
good news	0.00	0.00	0.00	1
healthy living	0.00	0.00	0.00	4
home & living	0.62	1.00	0.77	5
impact	0.33	0.25	0.29	4
media	0.67	0.80	0.73	5
money	0.60	0.60	0.60	5
parenting	0.67	0.20	0.31	10
politics	0.67	0.57	0.62	7
queer voices	0.75	0.75	0.75	4
religion	0.86	1.00	0.92	6
science	0.46	1.00	0.63	6
sports	1.00	1.00	1.00	8
style	0.33	0.50	0.40	6
style & beauty	1.00	0.38	0.55	8
taste	0.33	0.50	0.40	2
tech	0.50	0.71	0.59	7
travel	0.55	1.00	0.71	6
unknown	0.00	1.00	0.00	0
weddings	0.89	0.73	0.80	11
weird news	0.75	0.60	0.67	5
wellness	0.62	0.80	0.70	10
women	0.83	0.71	0.77	7
world	0.88	0.88	0.88	8
accuracy			0.64	200
macro avg	0.65	0.61	0.57	200
weighted avg	0.71	0.64	0.63	200



Classification for Sentiments:

	precision	recall	f1-score	support
negative	0.93	0.96	0.94	96
positive	0.96	0.93	0.95	104
accuracy			0.94	200
macro avg	0.94	0.95	0.94	200
weighted avg	0.95	0.94	0.95	200



Limitations & Future Work

- Hyperparameter Optimization
- Further fine-tuning with the whole dataset
- Prompting Techniques
- Interpretability
- Creating agents to bring the authors back to life :)

Code: 4506 7563

Mentimeter

Thank you for your attention!

References

Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released January 12, 2025. https://web.stanford.edu/~jurafsky/slp3.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. https://arxiv.org/abs/2104.09864

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. https://arxiv.org/abs/2305.14314

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. https://arxiv.org/abs/2106.09685

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. https://arxiv.org/abs/1706.03762