大语言模型技术前沿

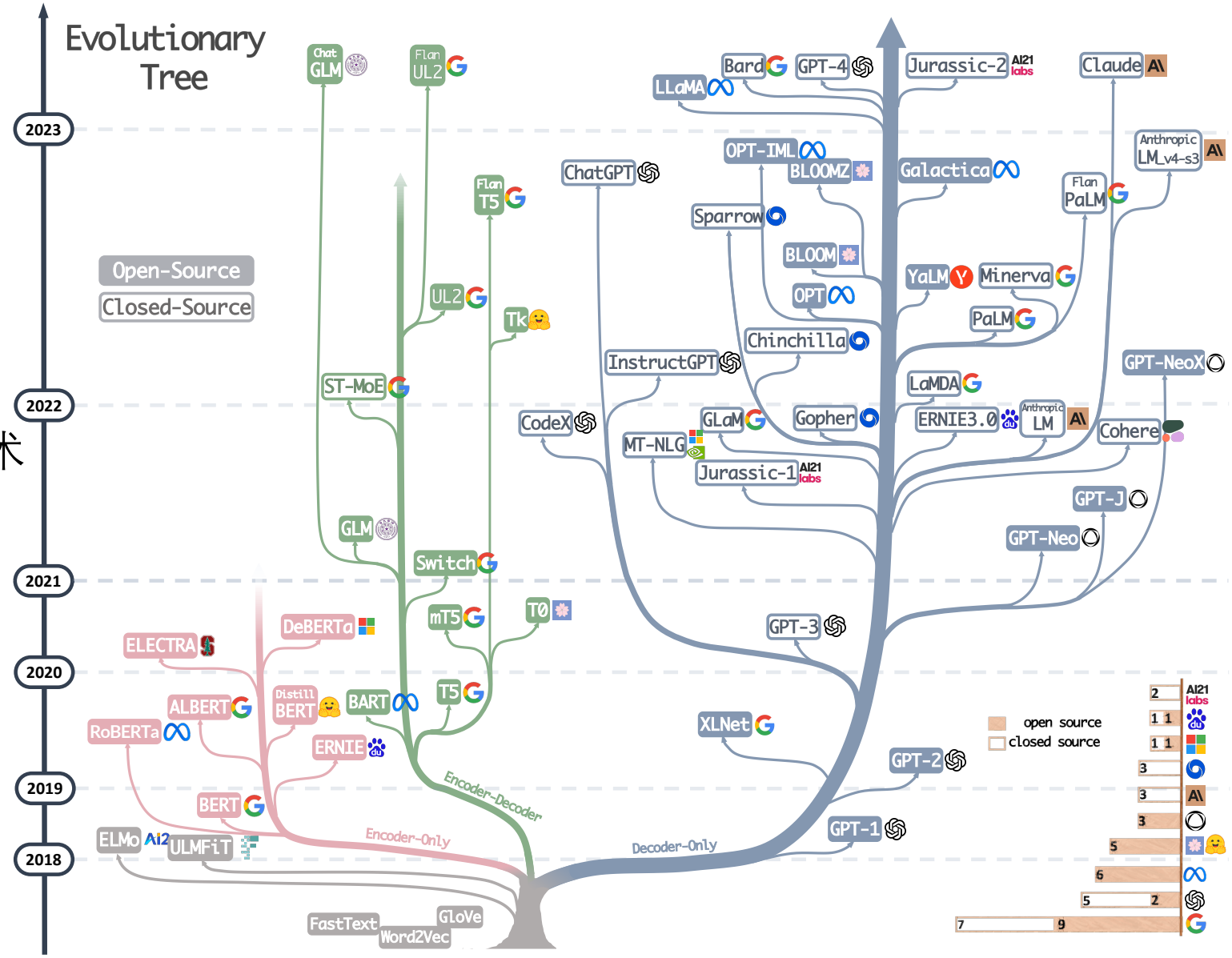
Scott

从PLM到LLM

关键:

- Transformer 架构
- Pretrain-finetune 预训练 + 微调技术
- 自监督(模型规模、数据质量)
- 自监督(模型规模、数据质量) SFT
- Instruction tuning 指令微调、对齐

RLHF



主流开源大语言模型

- LLaMA 1T
- LLaMA 2 2T
- ChatGLM 1T
- ChatGLM 2 1.4T
- Falcon 1.5T
- Bloom 350B
- Baichuan 1.2T
- Yi

模型	训练数据	训练数据量	模型参数量	词表大小
LLaMA	以英语为主的拉丁语系，不包含中日韩文	1T/1.4T tokens	7B、13B、33B、65B	32000
ChatGLM-6B	中英双语，中英文比例为1:1	1T tokens	6B	130528
Bloom	46种自然语言和13种编程语言，包含中文	350B tokens	560M、1.1B、1.7B、3B、7.1B、176B	250880
模型	模型结构	位置编码	激活函数	layer norm
LLaMA	Casual decoder	RoPE	SwiGLU	Pre RMS Norm
ChatGLM-6B	Prefix decoder	RoPE	GeGLU	Post Deep Norm
Bloom	Casual decoder	ALiBi	GeLU	Pre Layer Norm

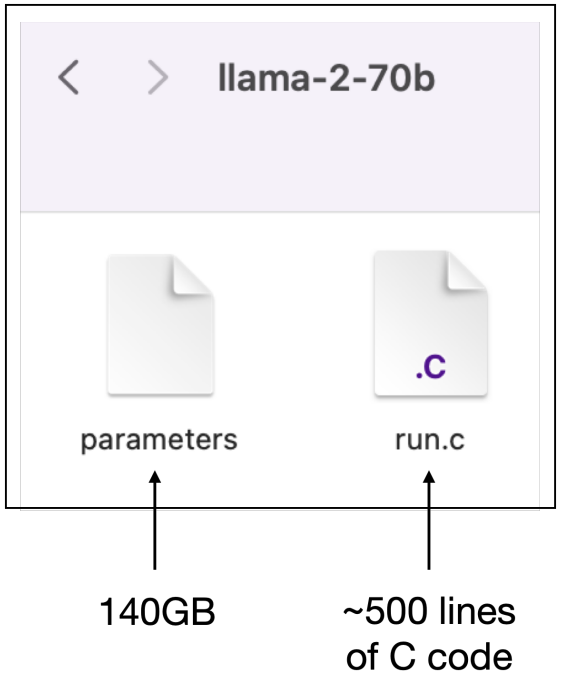
大模型文件

Bert-base-uncased

LICENSE	11.4 kB	
README.md	10.5 kB	
config.json	570 Bytes	
flax_model.msgpack	438 MB	LFS
model.onnx	532 MB	LFS
model.safetensors	440 MB	LFS
pytorch_model.bin	440 MB	LFS
rust_model.ot	534 MB	LFS
tf_model.h5	536 MB	LFS
tokenizer.json	466 kB	
tokenizer_config.json	28 Bytes	
vocab.txt	232 kB	

llama-7b

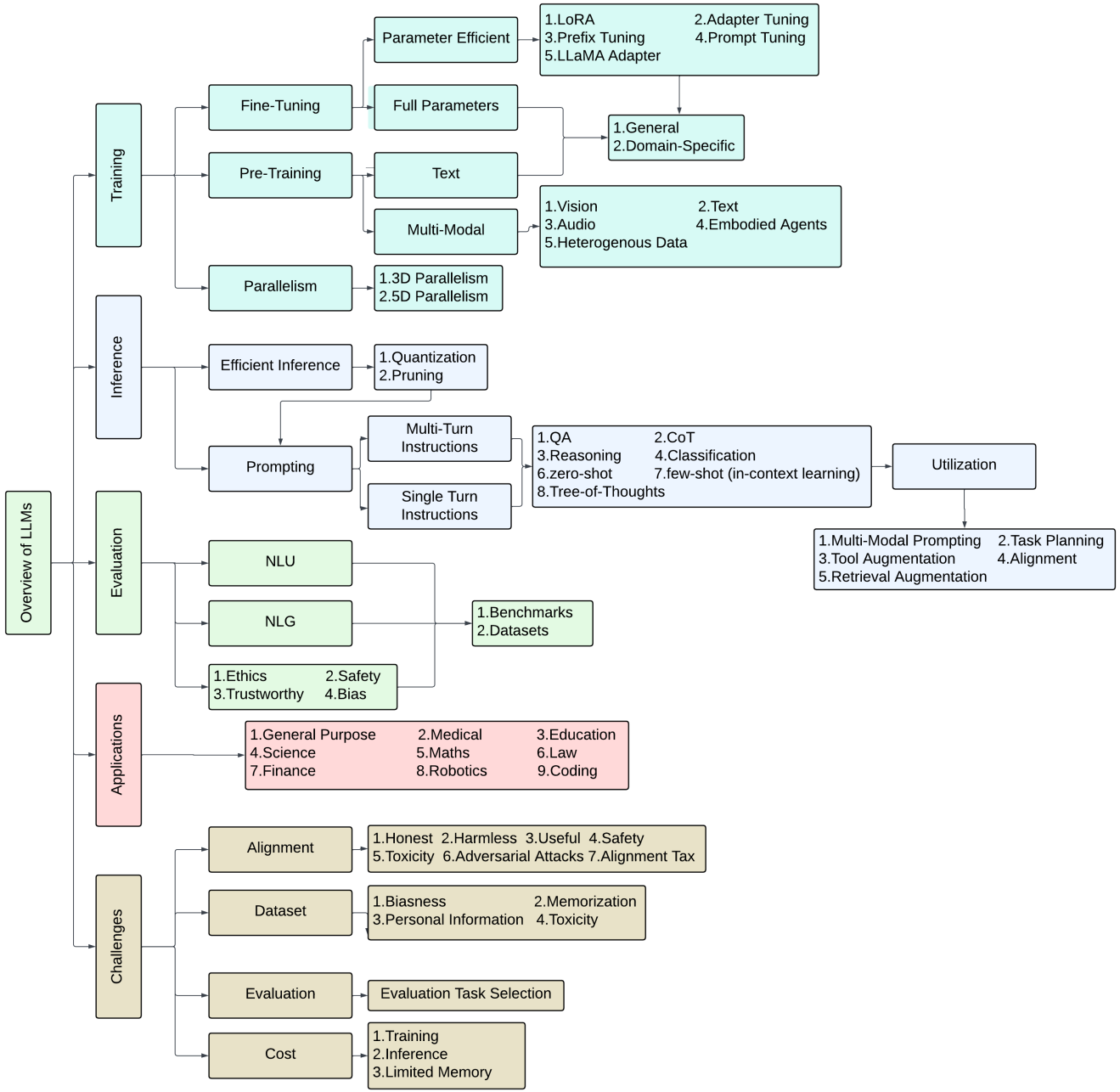
LICENSE	10.6 kB	
README.md	472 Bytes	
config.json	594 Bytes	
generation_config.json	137 Bytes	
model-00001-of-00002.sa…	9.98 GB	LFS
model-00002-of-00002.sa…	3.5 GB	LFS
model.safetensors.index.j…	26.8 kB	
pytorch_model-…	pickle 9.98 GB	LFS
pytorch_model-…	pickle 3.5 GB	LFS
pytorch_model.bin.index.…	26.8 kB	
special_tokens_map.json	411 Bytes	
tokenizer.json	1.84 MB	
tokenizer.model	500 kB	LFS
tokenizer_config.json	700 Bytes	



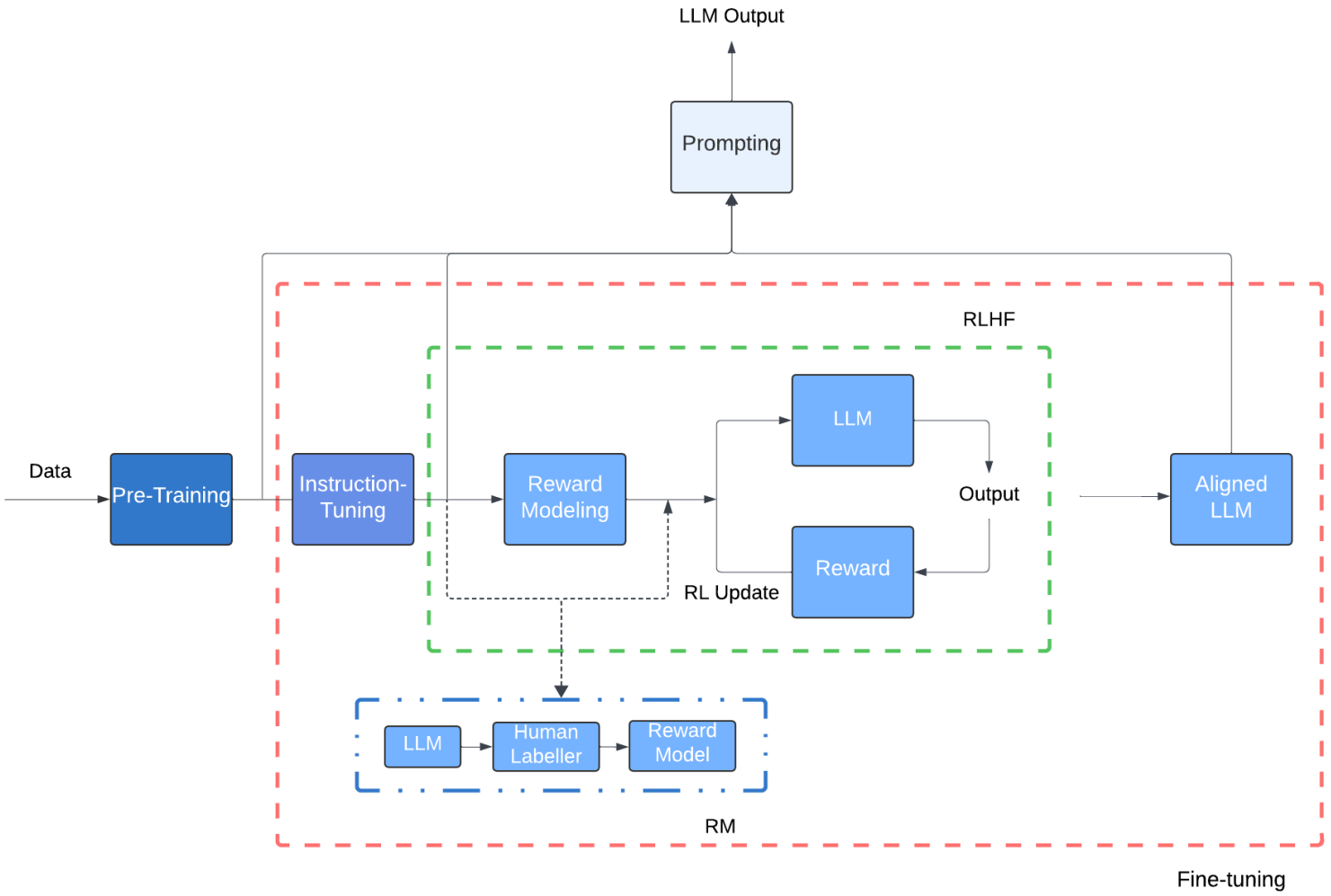
Llamafile
https://github.com/Mozilla-Ocho/llamafile
Just one file!

大模型研究方向

- Training
- Inference
- Evaluation
- Applications
- Challenges



大模型训练流程

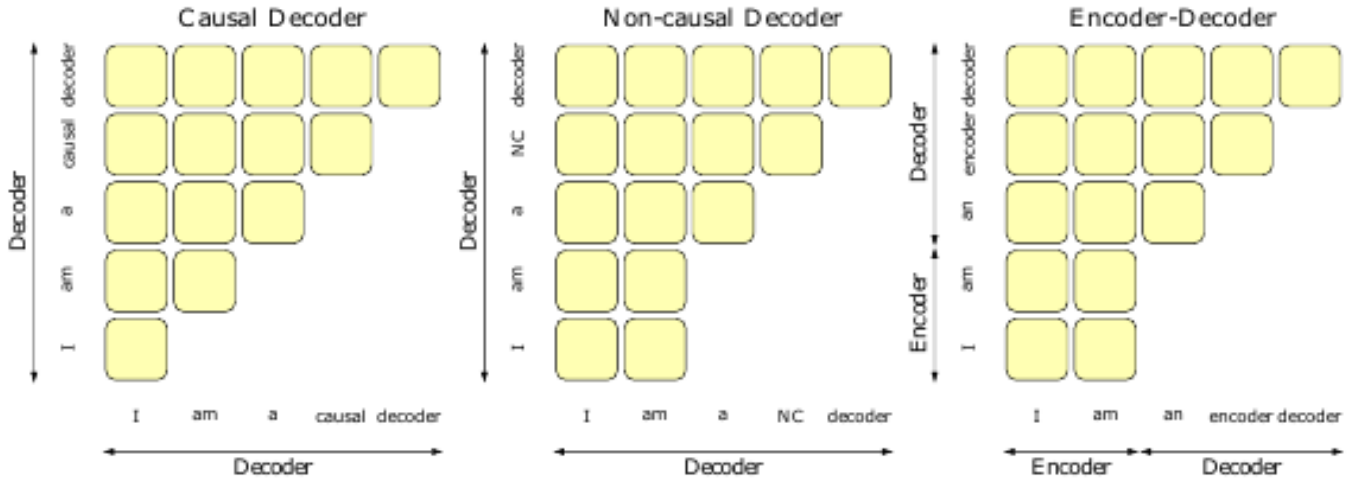

LLM Output
Prompting
RLHF
Data
Pre-Training
Instruction-Tuning
Reward Modeling
LLM
Output
Reward
RL Update
Aligned LLM
LLM
Human Labeller
Reward Model
RM
Fine-tuning

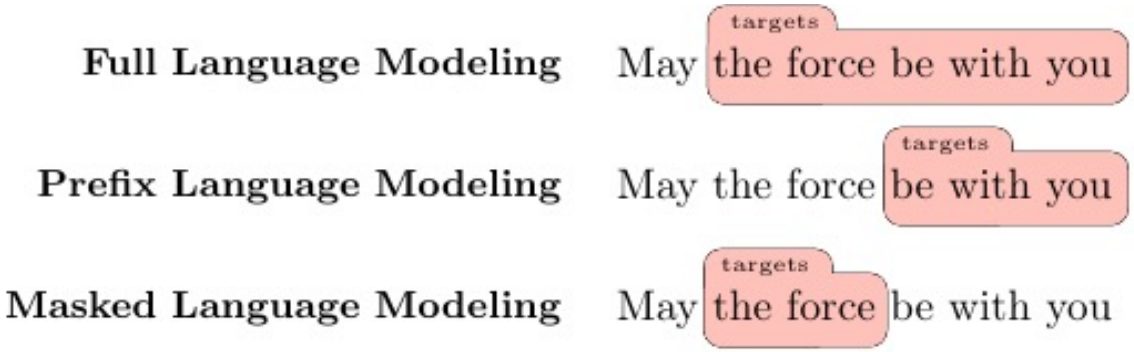
模型架构

Encoder + Decoder
- T5
- Flan-T5
- BART

Causal decoder
- GPT
- LLaMA
- BLOOM
- OPT

Prefix decoder
- GLM
- ChatGLM
- U-PaLM





Tokenization

- WordPiece
- BPE
- UnigramLM
- SentencePiece

模型	词表大小	中文平均token数	英文平均token数	中文处理时间(s)	英文处理时间(s)
LLaMA	32000	1.45	0.25	12.60	19.40
Chinese LLaMA	49953	0.62	0.249	8.65	19.12
ChatGLM-6B	130528	0.55	0.19	15.91	20.84
Bloom	250880	0.53	0.22	9.87	15.60

效率与性能之间的平衡

位置编码

- 绝对位置编码

- 相对位置编码

  ALiBi Attention with Linear Biases

  RoPE 旋转位置编码

注意力机制

- Self-Attention

Intra-attention

- Cross Attention
- Full Attention
- Sparse Attention
- Flash Attention
- Multi-query Attention

激活函数

- ReLU

- GeLU

Gaussian Error Linear Unit

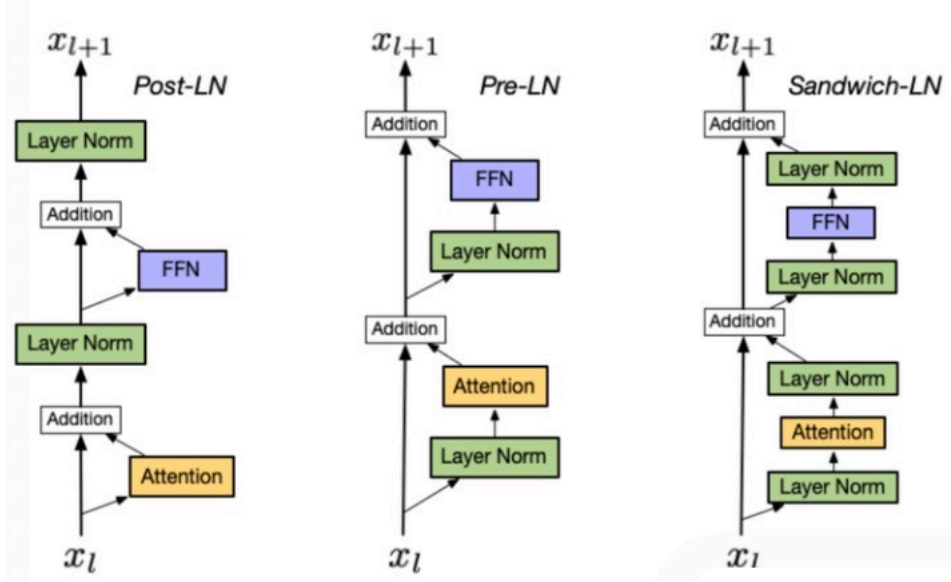
ReLU + dropout + zoneout

- GLU

Gated Linear Unit

层归一化

- LayerNorm
- RMSNorm
- Pre-Norm, Post-Norm, Sandwhich
- DeepNorm



Augmented LLMs

- Retrieval Augmented LLMs
- Tool Augmented LLMs

上下文长度/推理加速

- LLaMA 2048 - 4096 – 32k
- GPT3.5 4096
- GPT4 8192 - 32768 – 128k

Attention Linear Bias (ALiBi)-Better positional encoding
Sparse Attention
Flash Attention
Multi-query Attention
Conditional Computing

量化（Quantization）、剪枝（Pruning）、蒸馏（Distillation）、
参数共享（weight sharing）、矩阵分解（Factorization）

数据集和基线

- MMLU
- SuperGLUE
- BIG-Bench BBH
- GSM8K
- GLUE
- HumanEval
- AGIEval
- GLORE

RAG

- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks https://arxiv.org/abs/2005.11401v4
- Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy https://arxiv.org/pdf/2305.15294v2.pdf
- RETA-LLM: A Retrieval-Augmented Large Language Model Toolkit https://arxiv.org/abs/2306.05212
- Benchmarking Large Language Models in Retrieval-Augmented Generation https://arxiv.org/abs/2309.01431
- Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models https://arxiv.org/abs/2311.09210
- Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection https://openreview.net/forum?id=hSyW5go0v8
- InstructRetro: Instruction Tuning post Retrieval-Augmented Pretraining https://openreview.net/forum?id=4stB7DFLp6
- In-Context Retrieval-Augmented Language Models https://github.com/AI21Labs/in-context-ralm