



澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

A Comprehensive Overview of Large Language Models


XuTaosi
taosixu@umac.mo
2024.8.14



Subjects: **Computation and Language (cs.CL)**

Cite as: [arXiv:2307.06435](#) **[cs.CL]**

(or [arXiv:2307.06435v9](#) **[cs.CL]** for this version)

<https://doi.org/10.48550/arXiv.2307.06435> 

Submission history

From: Humza Naveed [[view email](#)]

[\[v1\]](#) Wed, 12 Jul 2023 20:01:52 UTC (542 KB)

[\[v2\]](#) Fri, 18 Aug 2023 13:53:06 UTC (734 KB)

[\[v3\]](#) Wed, 13 Sep 2023 12:13:45 UTC (1,354 KB)

[\[v4\]](#) Thu, 5 Oct 2023 10:29:02 UTC (1,369 KB)

[\[v5\]](#) Thu, 2 Nov 2023 07:59:50 UTC (1,438 KB)

[\[v6\]](#) Thu, 23 Nov 2023 19:23:19 UTC (1,440 KB)

[\[v7\]](#) Wed, 27 Dec 2023 10:15:51 UTC (1,481 KB)

[\[v8\]](#) Tue, 20 Feb 2024 07:19:41 UTC (1,533 KB)

[\[v9\]](#) Tue, 9 Apr 2024 21:38:33 UTC (1,691 KB)

Introduction

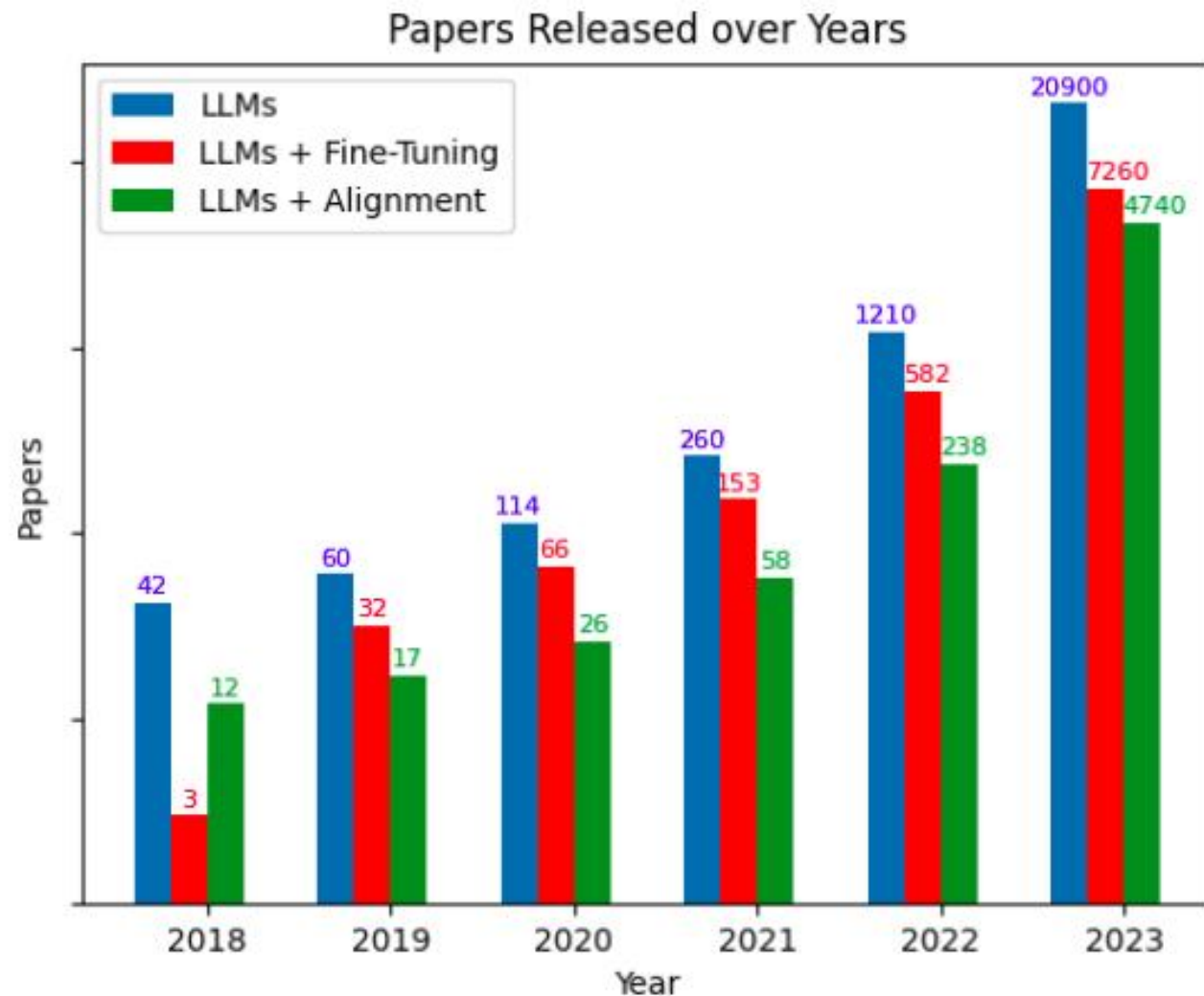


Figure 1: The trend of papers released over years containing keywords "Large Language Model", "Large Language Model + Fine-Tuning", and "Large Language Model + Alignment".

Introduction

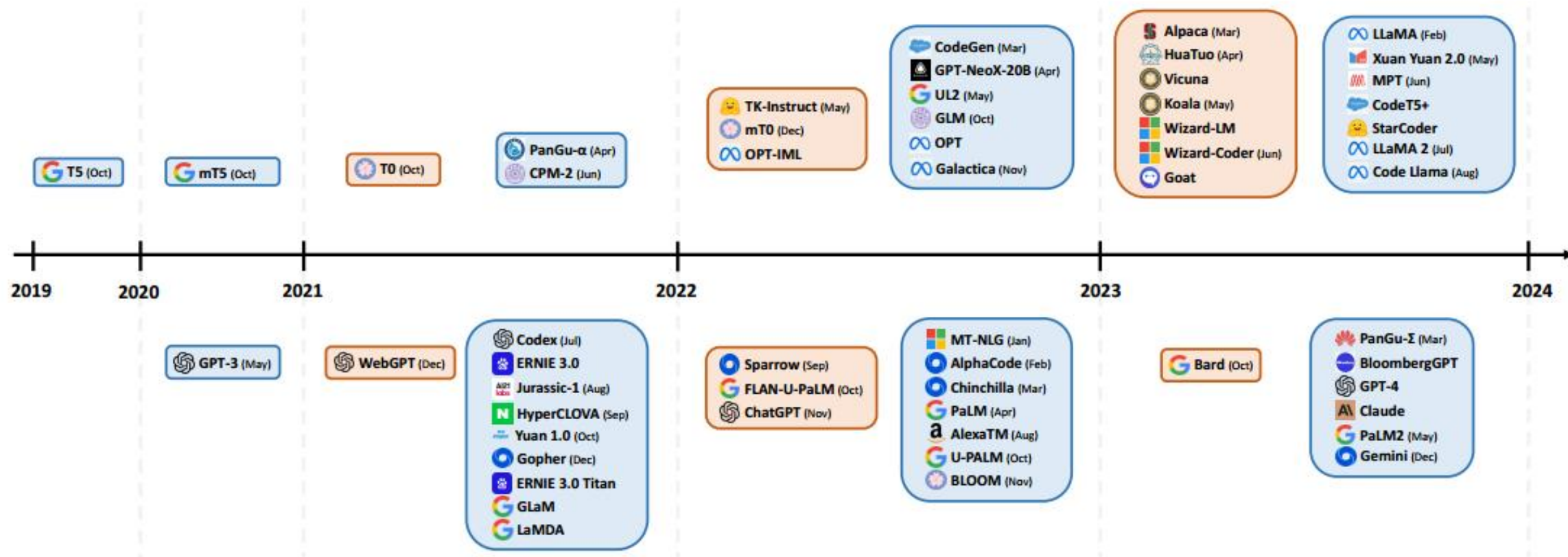
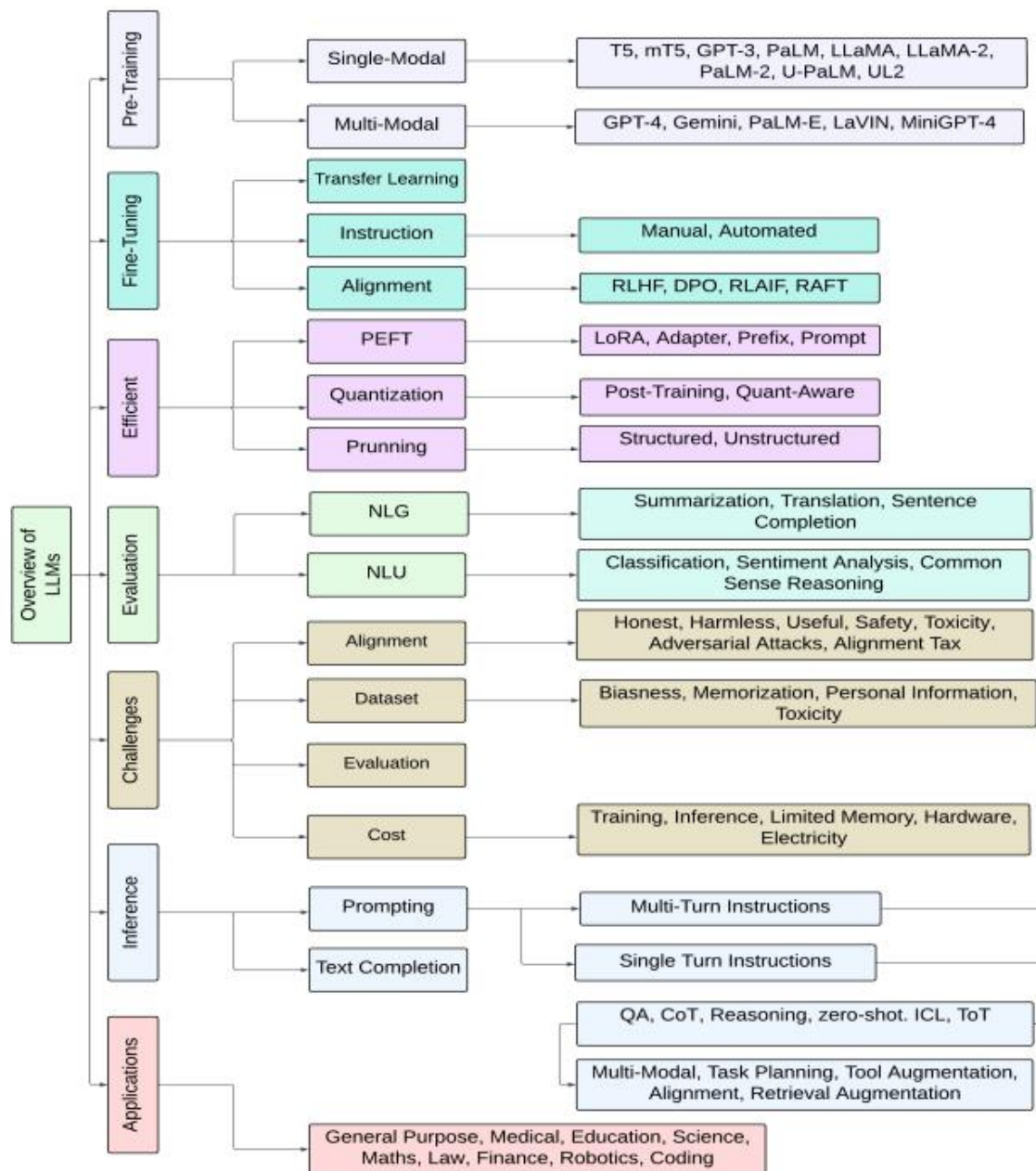


Figure 2: Chronological display of LLM releases: blue cards represent 'pre-trained' models, while orange cards correspond to 'instruction-tuned' models. Models on the upper half signify open-source availability, whereas those on the bottom half are closed-source. The chart illustrates the increasing trend towards instruction-tuned models and open-source models, highlighting the evolving landscape and trends in natural language processing research.

Introduction



自然语言理解(NLU)
自然语言生成(NLG)

Pre-training to utilization

“RL”代表强化学习

“RM”代表奖励建模

“RLHF”代表人类反馈
的强化学习。

- 1、预训练阶段；
- 2、指令调优；
- 3、奖励建模；
- 4、提示与响应；
- 5、强化学习更新；
- 6、人类标注与反馈；
- 7、对齐调优；
- 8、微调与优化

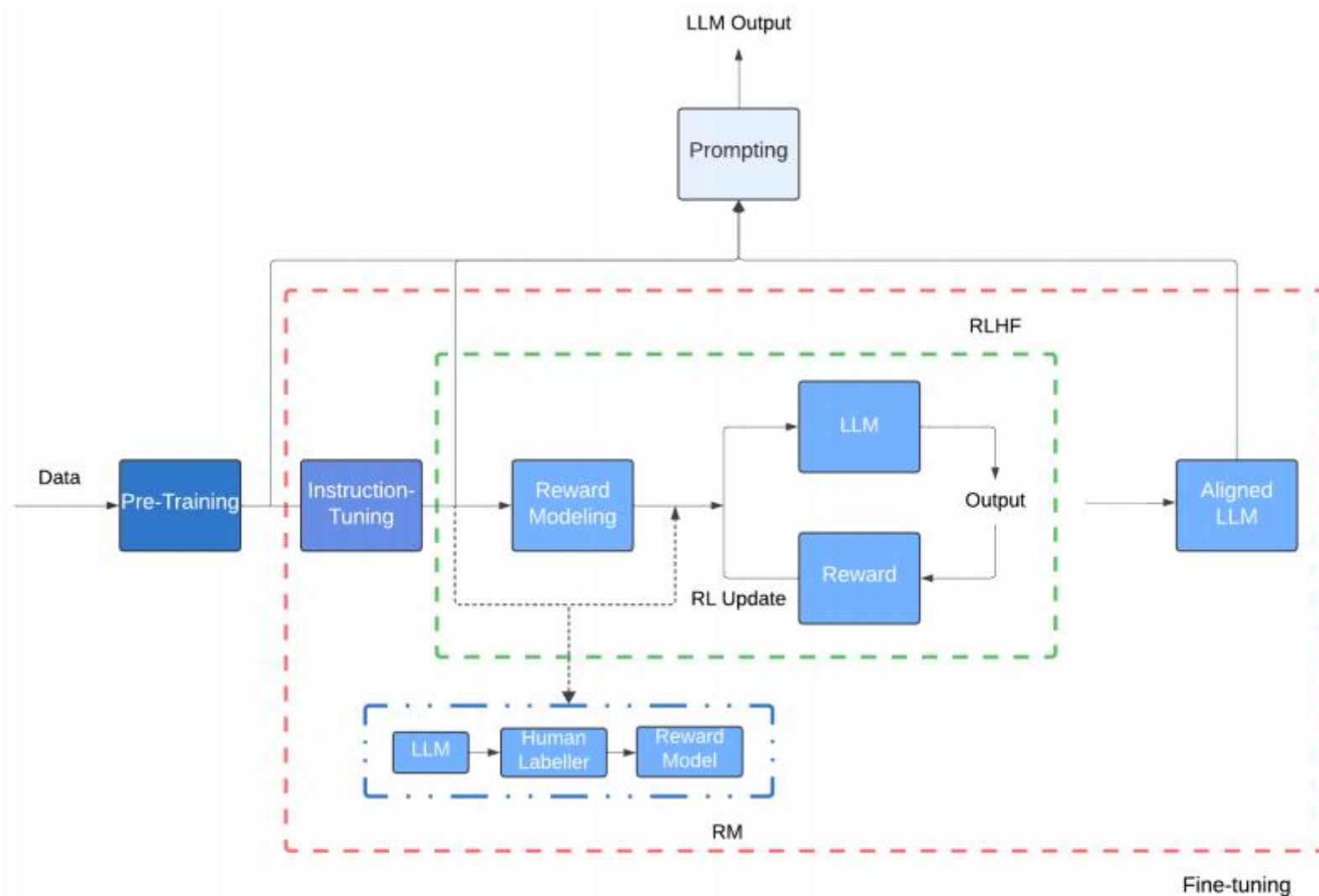


Figure 6: A basic flow diagram depicting various stages of LLMs from pre-training to prompting/utilization. Prompting LLMs to generate responses is possible at different training stages like pre-training, instruction-tuning, or alignment tuning. “RL” stands for reinforcement learning, “RM” represents reward-modeling, and “RLHF” represents reinforcement learning with human feedback.

Pre-trained LLMs

Table 1: Noteworthy findings and insights of *pre-trained* Large Language Models.

Models	Findings & Insights
T5	<ul style="list-style-type: none">• Encoder and decoder with shared parameters perform equivalently when parameters are not shared• Fine-tuning model layers (adapter layers) work better than the conventional way of training on only classification layers
GPT-3	<ul style="list-style-type: none">• Few-shot performance of LLMs is better than the zero-shot, suggesting that LLMs are meta-learners
mT5	<ul style="list-style-type: none">• Large multi-lingual models perform equivalently to single language models on downstream tasks. However, smaller multi-lingual models perform worse
PanGu- α	<ul style="list-style-type: none">• LLMs have good few shot capabilities
CPM-2	<ul style="list-style-type: none">• Prompt fine-tuning requires updating very few parameters while achieving performance comparable to full model fine-tuning• Prompt fine-tuning takes more time to converge as compared to full model fine-tuning• Inserting prompt tokens in-between sentences can allow the model to understand relations between sentences and long sequences• In an analysis, CPM-2 finds that prompts work as a provider (additional context) and aggregator (aggregate information with the input text) for the model
ERNIE 3.0	<ul style="list-style-type: none">• A modular LLM architecture with a universal representation module and task-specific representation module helps in the finetuning phase• Optimizing the parameters of a task-specific representation network during the fine-tuning phase is an efficient way to take advantage of the powerful pre-trained model

GPT-3: 具有元学习的能力，即能够基于少量或没有针对特定任务的训练数据来执行任务。

Pre-trained LLMs

BloombergGPT	<ul style="list-style-type: none">• Pre-training with general-purpose and task-specific data improves task performance without hurting other model capabilities
XuanYuan 2.0	<ul style="list-style-type: none">• Combining pre-training and fine-tuning stages in single training avoids catastrophic forgetting
CodeT5+	<ul style="list-style-type: none">• Causal LM is crucial for a model's generation capability in encoder-decoder architectures• Multiple training objectives like span corruption, Causal LM, matching, etc complement each other for better performance
StarCoder	<ul style="list-style-type: none">• HHH prompt by Anthropic allows the model to follow instructions without fine-tuning
LLaMA-2	<ul style="list-style-type: none">• Model trained on unfiltered data is more toxic but may perform better on downstream tasks after fine-tuning• Model trained on unfiltered data requires fewer samples for safety alignment
PaLM-2	<ul style="list-style-type: none">• Data quality is important to train better models• Model and data size should be scaled with 1:1 proportions• Smaller models trained for larger iterations outperform larger models

Fine-tuning LLMs

Table 2: Key insights and findings from the study of *instruction-tuned* Large Language Models.

Models	Findings & Insights
T0	<ul style="list-style-type: none">• Multi-task prompting enables zero-shot generalization and outperforms baselines• Even a single prompt per dataset task is enough to improve performance
WebGPT	<ul style="list-style-type: none">• To aid the model in effectively filtering and utilizing relevant information, human labelers play a crucial role in answering questions regarding the usefulness of the retrieved documents• Interacting a fine-tuned language model with a text-based web-browsing environment can improve end-to-end retrieval and synthesis via imitation learning and reinforcement learning• Generating answers with references can make labelers easily judge the factual accuracy of answers
Tk-INSTRUCT	<ul style="list-style-type: none">• Instruction tuning leads to a stronger generalization of unseen tasks• More tasks improve generalization whereas only increasing task instances does not help• Supervised trained models are better than generalized models• Models pre-trained with instructions and examples perform well for different types of inputs
WizardCoder	<ul style="list-style-type: none">• Fine-tuning with re-written instruction-tuning data into a complex set improves performance
LLaMA-2-Chat	<ul style="list-style-type: none">• Model learns to write safe responses with fine-tuning on safe demonstrations, while additional RLHF step further improves model safety and make it less prone to jailbreak attacks
LIMA	<ul style="list-style-type: none">• Less high quality data is enough for fine-tuned model generalization

Efficient LLMs

- **Parameter Efficient Fine-Tuning**

Adapter Tuning: 在Transformer中添加少量可训练参数，包括降维、非线性和升维步骤。AdaMix(e mixture of adapter)和LoRA(Low-Rank Adaptation) 是其变体，前者通过多模块平均输出减少延迟，后者通过低秩分解避免延迟。

Prompt Tuning: 利用提示 (Prompt) 适应LLM，微调少量参数 (0.001%-3%) 。P-Tuning等通过可学习映射编码连续提示，Progressive Prompts顺序添加提示嵌入以避免遗忘。

Prefix (前缀) Tuning: 在Transformer层上附加可训练前缀向量作为虚拟令牌，自适应前缀调优通过门控机制控制信息。

Bias Tuning: 仅微调偏置项，对小型至中型数据集有效，达到或接近全微调性能。

Efficient LLMs

- **Quantization**

Post-Training Quantization: 无需或仅需极少训练，以不显著牺牲模型性能为代价。方法包括使用全精度处理异常值特征，其他特征采用8位量化；通过知识蒸馏和独立量化缩放因子改善性能；平滑激活以简化量化难度；利用最优脑压缩算法逐层量化并更新权重以补偿量化误差；以及针对易受影响的权重使用更高精度量化等。

Quantization-Aware Training: 通过微调量化模型以补偿性能下降。技术包括使用二进制编码量化并仅微调量化缩放因子；减少全连接层的精度并微调量化缩放参数；从预训练网络生成训练数据，训练量化学学生模型；以及使用LoRA(Low-Rank Adaptation)方法微调4位量化的预训练LLM等。

Efficient LLMs

- **Pruning (剪枝)**

Unstructured Pruning: 移除不重要的权重，不保留结构。利用LLM特性，如小部分隐藏状态高激活。如. Pruning by weights and activations(Wanda)(通过权重和激活进行修剪)基于权重与输入范数乘积剪枝，无需微调节省成本。 Outlier weighed layerwise sparsity (OWL)(离群值加权分层稀疏度)扩展Wanda，实现非均匀层剪枝，考虑各层异常值差异。 Contrastive pruning (CAP)通过训练稀疏模型迭代剪枝，利用对比损失学习。

Structured Pruning: 以组、行、列或矩阵形式移除参数，加速推理利用硬件优势。LLM-Pruner采用三阶段策略，识别并保留重要隐藏状态组，用Low-Rank Adaptation(LoRA)(低秩的适应)微调Sparsity-induced mask learning(SIMPLE)(稀疏性诱导的掩模学习)等方法通过学习掩码剪枝，或移除因子化权重矩阵中的不重要秩1组件。

Model Configurations(配置)

Table 3: Summary of pre-trained LLMs (>10B). Only the LLMs discussed individually in the previous sections are summarized. “Data/Tokens” is the model’s pre-training data, which is either the number of tokens or data size. “Data Cleaning” indicates whether data cleaning is performed or not. This includes heuristics (Heur), deduplication (Dedup), quality filtering (QF), and privacy filtering (PF), “Cost” is the calculated training cost obtained by multiplying the GPUs/TPUs hourly rate with the number of GPUs and the training time. The actual cost may vary due to many reasons such as using in-house GPUs or getting a discounted rate, re-training, number of employees working on the problem, etc. “Training Parallelism” indicates distributed training using data parallelism (D), tensor parallelism (T), pipeline parallelism (P), model parallelism (M), optimizer parallelism (OP), and rematerialization (R), where for “Library” column, “DS” is a short form for Deep Speed. In column “Commercial Use”, we assumed a model is for non-commercial purposes if its license is unavailable.

Models	Publication Venue	License Type	Model Creators	Purpose	No. of Params	Commercial Use	Steps Trained	Data/ Tokens	Data Cleaning	No. of Processing Units	Processing Unit Type	Training Time	Calculated Train. Cost	Training Parallelism	Library
T5 [10]	JMLR’20	Apache-2.0	Google	General	11B	✓	1M	1T	Heur+Dedup	1024	TPU v3	-	-	D+M	Mesh TensorFlow
GPT-3 [6]	NeurIPS’20	-	OpenAI	General	175B	×	-	300B	Dedup+QF	-	V100	-	-	M	-
mt5 [11]	NAACL’21	Apache-2.0	Google	General	13B	✓	1M	1T	-	-	-	-	-	-	-
PanGu- α [108]	arXiv’21	Apache-2.0	Huawei	General	200B	✓	260k	1.1TB	Heur+Dedup	2048	Ascend 910	-	-	D+OP+P+O+R	MindSpore
CPM-2 [12]	AI Open’21	MIT	Tsinghua	General	198B	✓	1M	2.6TB	Dedup	-	-	-	-	D+M	JAXFormer
Codex [131]	arXiv’21	-	OpenAI	Coding	12B	×	-	100B	Heur	-	-	-	-	-	-
ERNIE 3.0 [110]	arXiv’21	-	Baidu	General	10B	×	120k*	375B	Heur+Dedup	384	V100	-	-	M*	PaddlePaddle
Jurassic-1 [112]	White-Paper’21	Apache-2.0	AI21	General	178B	✓	-	300B	-	800	GPU	-	-	D+M+P	Megatron+DS
HyperCLOVA [114]	EMNLP’21	-	Naver	General	82B	×	-	300B	Clf+Dedup+PF	1024	A100	321h	1.32 Mil	M	Megatron
Yuan 1.0 [115]	arXiv’21	Apache-2.0	-	General	245B	✓	26k*	180B	Heur+Clf+Dedup	2128	GPU	-	-	D+T+P	-
Gopher [116]	arXiv’21	-	Google	General	280B	×	-	300B	QF+Dedup	4096	TPU v3	920h	13.19 Mil	D+M	JAX+Haiku
ERNIE 3.0 Titan [35]	arXiv’21	-	Baidu	General	260B	×	-	300B	Heur+Dedup	-	Ascend 910	-	-	D+M+P+D*	PaddlePaddle
GPT-NeoX-20B [118]	BigScience’22	Apache-2.0	EleutherAI	General	20B	✓	150k	825GB	None	96	40G A100	-	-	M	Megatron+DS+PyTorch
OPT [14]	arXiv’22	MIT	Meta	General	175B	✓	150k	180B	Dedup	992	80G A100	-	-	D+T	Megatron
BLOOM [13]	arXiv’22	RAIL-1.0	BigScience	General	176B	✓	-	366B	Dedup+PR	384	80G A100	2520h	3.87 Mil	D+T+P	Megatron+DS
Galactica [138]	arXiv’22	Apache-2.0	Meta	Science	120B	×	225k	106B	Dedup	128	80GB A100	-	-	-	Metaseq
GLaM [91]	ICML’22	-	Google	General	1.2T	×	600k*	600B	Clf	1024	TPU v4	-	-	M	GSPMD
LaMDA [140]	arXiv’22	-	Google	Dialog	137B	×	3M	2.81T	Filtered	1024	TPU v3	1384h	4.96 Mil	D+M	Lingvo
MT-NLG [117]	arXiv’22	Apache-v2.0	MS.+Nvidia	General	530B	×	-	270B	-	4480	80G A100	-	-	D+T+P	Megatron+DS
AlphaCode [132]	Science’22	Apache-v2.0	Google	Coding	41B	✓	205k	967B	Heur+Dedup	-	TPU v4	-	-	M	JAX+Haiku
Chinchilla [96]	arXiv’22	-	Google	General	70B	×	-	1.4T	QF+Dedup	-	TPUv4	-	-	-	JAX+Haiku
PaLM [15]	arXiv’22	-	Google	General	540B	×	255k	780B	Heur	6144	TPU v4	-	-	D+M	JAX+T5X
AlexaTM [122]	arXiv’22	Apache v2.0	Amazon	General	20B	×	500k	1.1T	Filtered	128	A100	2880h	1.47 Mil	M	DS
U-PaLM [124]	arXiv’22	-	Google	General	540B	×	20k	-	-	512	TPU v4	120h	0.25 Mil	-	-
UL2 [125]	ICLR’23	Apache-2.0	Google	General	20B	✓	2M	1T	-	512	TPU v4	-	-	M	JAX+T5X
GLM [33]	ICLR’23	Apache-2.0	Multiple	General	130B	×	-	400B	-	768	40G A100	1440h	3.37 Mil	M	-
CodeGen [130]	ICLR’23	Apache-2.0	Salesforce	Coding	16B	✓	650k	577B	Heur+Dedup	-	TPU v4	-	-	D+M	JAXFormer
LLaMA [127]	arXiv’23	-	Meta	General	65B	×	350k	1.4T	Clf+Heur+Dedup	2048	80G A100	504h	4.12 Mil	D+M	xFormers
PanGu Σ [92]	arXiv’23	-	Huawei	General	1.085T	×	-	329B	-	512	Ascend 910	2400h	-	D+OP+P+O+R	MindSpore
BloombergGPT [141]	arXiv’23	-	Bloomberg	Finance	50B	×	139k	569B	Dedup	512	40G A100	1272h	1.97 Mil	M	PyTorch
Xuan Yuan 2.0 [142]	arXiv’23	RAIL-1.0	Du Xiaoman	Finance	176B	✓	-	366B	Filtered	80GB	A100	-	-	P	DS
CodeT5+ [34]	arXiv’23	BSD-3	Salesforce	Coding	16B	✓	110k	515B	Dedup	16	40G A100	-	-	-	DS
StarCoder [137]	arXiv’23	OpenRAIL-M	BigCode	Coding	15.5B	✓	250k	1T	Dedup+QF+PF	512	80G A100	624h	1.28 Mil	D+T+P	Megatron-LM
LLaMA-2 [21]	arXiv’23	LLaMA-2.0	Meta	General	70B	✓	500k	2T	Minimal Filtering	-	80G A100	1.7Mh	-	-	-
PaLM-2 [123]	arXiv’23	-	Google	General	-	×	-	-	Ddedup+PF+QF	-	-	-	-	-	-

Model Configurations

Table 4: Summary of instruction tuned LLMs (>10B). All abbreviations are the same as Table 3. Entries in “Data/Tokens” starting with “S-” represents the number of training samples.

Models	Publication Venue	License Type	Model Creators	Purpose	No. of Params	Commercial Use	Pre-trained Models	Steps Trained	Data/ Tokens	No. of Processing Units	Processing Unit Type	Train. Time	Calculated Train. Cost	Train. Parallelism	Library
WebGPT [156]	arXiv'21	-	OpenAI	General	175B	×	GPT-3	-	-	-	-	-	-	-	-
T0 [17]	ICLR'22	Apache-2.0	BigScience	General	11B	✓	T5	-	250B	512	TPU v3	270h	0.48 Mil	-	-
Tk-Instruct [18]	EMNLP'22	MIT	AI2+	General	11B	✓	T5	1000	-	256	TPU v3	4h	0.0036 Mil	-	Google T5
OPT-IML [97]	arXiv'22	-	Meta	General	175B	×	OPT	8k	2B	128	40G A100	-	-	D+T	Megatron
Flan-U-PaLM [16]	ICLR'22	Apache-2.0	Google	General	540B	✓	U-PaLM	30k	-	512	TPU v4	-	-	-	JAX+T5X
mT0 [144]	ACL'23	Apache-2.0	HuggingFace+	General	13B	✓	mT5	-	-	-	-	-	-	-	-
Sparrow [157]	arXiv'22	-	Google	Dialog	70B	×	Chinchilla	-	-	64	TPU v3	-	-	M	-
WizardCoder [154]	arXiv'23	Apache-2.0	HK Bapt.	Coding	15B	×	StarCoder	200	S-78k	-	-	-	-	-	-
Alpaca [148]	Github'23	Apache-2.0	Stanford	General	13B	✓	LLaMA	3-Epoch	S-52k	8	80G A100	3h	600	FSDP	PyTorch
Vicuna [149]	Github'23	Apache-2.0	LMSYS	General	13B	✓	LLaMA	3-Epoch	S-125k	-	-	-	-	FSDP	PyTorch
LIMA [175]	arXiv'23	-	Meta+	General	65B	-	LLaMA	15-Epoch	S-1000	-	-	-	-	-	-
Koala [290]	Github'23	Apache-2.0	UC-Berkley	General	13B	×	LLaMA	2-Epoch	S-472k	8	A100	6h	100	-	JAX/FLAX

Model Configurations

Table 5: Architecture details of LLMs. Here, “PE” is the positional embedding, “nL” is the number of layers, “nH” is the number of attention heads, “HS” is the size of hidden states.

Models	Type	Training Objective	Attention	Vocab	Tokenizer	Norm	PE	Activation	Bias	nL	nH	HS
T5 (11B)	Enc-Dec	Span Corruption	Standard	32k	SentencePiece	Pre-RMS	Relative	ReLU	×	24	128	1024
GPT3 (175B)	Causal-Dec	Next Token	Dense+Sparse	-	-	Layer	Learned	GeLU	✓	96	96	12288
mT5 (13B)	Enc-Dec	Span Corruption	Standard	250k	SentencePiece	Pre-RMS	Relative	ReLU	-	-	-	-
PanGu- α (200B)	Causal-Dec	Next Token	Standard	40k	BPE	Layer	-	-	-	64	128	16384
CPM-2 (198B)	Enc-Dec	Span Corruption	Standard	250k	SentencePiece	Pre-RMS	Relative	ReLU	-	24	64	-
Codex (12B)	Causal-Dec	Next Token	Standard	-	BPE+	Pre-Layer	Learned	GeLU	-	96	96	12288
ERNIE 3.0 (10B)	Causal-Dec	Next Token	Standard	-	WordPiece	Post-Layer	Relative	GeLU	-	48	64	4096
Jurassic-1 (178B)	Causal-Dec	Next Token	Standard	256k	SentencePiece*	Pre-Layer	Learned	GeLU	✓	76	96	13824
HyperCLOVA (82B)	Causal-Dec	Next Token	Dense+Sparse	-	BPE*	Pre-Layer	Learned	GeLU	-	64	80	10240
Yuan 1.0 (245B)	Causal-Dec	Next Token	Standard	-	-	-	-	-	-	76	-	16384
Gopher (280B)	Causal-Dec	Next Token	Standard	32k	SentencePiece	Pre-RMS	Relative	GeLU	✓	80	128	16384
ERNIE 3.0 Titan (260B)	Causal-Dec	Next Token	Standard	-	WordPiece	Post-Layer	Relative	GeLU	-	48	192	12288
GPT-NeoX-20B	Causal-Dec	Next Token	Parallel	50k	BPE	Layer	Rotary	GeLU	✓	44	64	-
OPT (175B)	Causal-Dec	Next Token	Standard	-	BPE	-	-	ReLU	✓	96	96	-
BLOOM (176B)	Causal-Dec	Next Token	Standard	250k	BPE	Layer	ALiBi	GeLU	✓	70	112	14336
Galactica (120B)	Causal-Dec	Next Token	Standard	50k	BPE+custom	Layer	Learned	GeLU	×	96	80	10240
GLaM (1.2T)	MoE-Dec	Next Token	Standard	256k	SentencePiece	Layer	Relative	GeLU	✓	64	128	32768
LaMDA (137B)	Causal-Dec	Next Token	Standard	32k	BPE	Layer	Relative	GeLU	-	64	128	8192
MT-NLG (530B)	Causal-Dec	Next Token	Standard	50k	BPE	Pre-Layer	Learned	GeLU	✓	105	128	20480
AlphaCode (41B)	Enc-Dec	Next Token	Multi-query	8k	SentencePiece	-	-	-	-	64	128	6144
Chinchilla (70B)	Causal-Dec	Next Token	Standard	32k	SentencePiece-NFKC	Pre-RMS	Relative	GeLU	✓	80	64	8192
PaLM (540B)	Causal-Dec	Next Token	Parallel+Multi-query	256k	SentencePiece	Layer	RoPE	SwiGLU	×	118	48	18432
AlexaTM (20B)	Enc-Dec	Denosing	Standard	150k	SentencePiece	Pre-Layer	Learned	GeLU	✓	78	32	4096
Sparrow (70B)	Causal-Dec	Pref.&Rule RM	-	32k	SentencePiece-NFKC	Pre-RMS	Relative	GeLU	✓	16*	64	8192
U-PaLM (540B)	Non-Causal-Dec	MoD	Parallel+Multi-query	256k	SentencePiece	Layer	RoPE	SwiGLU	×	118	48	18432
UL2 (20B)	Enc-Dec	MoD	Standard	32k	SentencePiece	-	-	-	-	64	16	4096
GLM (130B)	Non-Causal-Dec	AR Blank Infilling	Standard	130k	SentencePiece	Deep	RoPE	GeGLU	✓	70	96	12288
CodeGen (16B)	Causal-Dec	Next Token	Parallel	-	BPE	Layer	RoPE	-	-	34	24	-
LLaMA (65B)	Causal-Dec	Next Token	Standard	32k	BPE	Pre-RMS	RoPE	SwiGLU	-	80	64	8192
PanGu- Σ (1085B)	Causal-Dec	Next Token	Standard	-	BPE	Fused Layer	-	FastGeLU	-	40	40	5120
BloombergGPT (50B)	Causal-Dec	Next Token	Standard	131k	Unigram	Layer	ALiBi	GeLU	✓	70	40	7680
Xuan Yuan 2.0 (176B)	Causal-Dec	Next Token	Self	250k	BPE	Layer	ALiBi	GeLU	✓	70	112	14336
CodeT5+ (16B)	Enc-Dec	SC+NT+Cont.+ Match	Standard	-	Code-Specific	-	-	-	-	-	-	-
StarCoder (15.5B)	Causal-Dec	FIM	Multi-query	49k	BPE	-	Learned	-	-	40	48	6144
LLaMA (70B)	Causal-Dec	Next Token	Grouped-query	32k	BPE	Pre-RMS	RoPE	SwiGLUE	-	-	-	-
PaLM-2	-	MoD	Parallel	-	-	-	-	-	-	-	-	-

Model Configurations

Table 6: Summary of optimization settings used for pre-trained LLMs. The values for weight decay, gradient clipping, and dropout are 0.1, 1.0, and 0.1, respectively, for most of the LLMs.

Models	Batch Size	Sequence Length	LR	Warmup	LR Decay	Optimizers			Precision			Weight Decay	Grad Clip	Dropout
						AdaFact	Adan	AdamW	FP16	BF16	Mixed			
T5 (11B)	2 ¹¹	512	0.01	×	inverse square root	✓			-	-	-	-	-	✓
GPT3 (175B)	32K	-	6e-5	✓	cosine		✓		✓			✓	✓	-
mT5 (13B)	1024	1024	0.01	-	inverse square root	✓			-	-	-	-	-	✓
PanGu- α (200B)	-	1024	2e-5	-	-	-	-	-	-	✓	-	-	-	-
CPM-2 (198B)	1024	1024	0.001	-	-	✓			-	-	-	-	-	✓
Codex (12B)	-	-	6e-5	✓	cosine		✓		✓			✓	-	-
ERNIE 3.0 (12B)	6144	512	1e-4	✓	linear		✓		-	-	-	✓	-	-
Jurassic-1 (178B)	3.2M	2048	6e-5	✓	cosine		✓		✓			✓	✓	-
HyperCLOVA (82B)	1024	-	6e-5	-	cosine			✓	-	-	-	✓	-	-
Yuan 1.0 (245B)	<10M	2048	1.6e-4	✓	cosine decay to 10%		✓		-	-	-	✓	-	-
Gopher (280B)	3M	2048	4e-5	✓	cosine decay to 10%		✓			✓		-	✓	-
ERNIE 3.0 Titan (260B)	-	512	1e-4	✓	linear		✓		✓			✓	✓	-
GPT-NeoX-20B	1538	2048	0.97e-5	✓	cosine			✓	✓			✓	✓	×
OPT (175B)	2M	2048	1.2e-4	-	linear			✓	✓			✓	✓	✓
BLOOM (176B)	2048	2048	6e-5	✓	cosine		✓			✓		✓	✓	×
Galactica (120B)	2M	2048	7e-6	✓	linear decay to 10%			✓	-	-	-	✓	✓	✓
GLaM (1.2T)	1M	1024	0.01	-	inverse square root	✓			FP32 + ✓			-	✓	×
LaMDA (137B)	256K	-	-	-	-	-	-	-	-	-	-	-	-	-
MT-NLG (530B)	1920	2048	5e-5	✓	cosine decay to 10%		✓			✓		✓	✓	-
AlphaCode (41B)	2048	1536+768	1e-4	✓	cosine decay to 10%			✓		✓		✓	✓	-
Chinchilla (70B)	1.5M	2048	1e-4	✓	cosine decay to 10%			✓		✓		-	-	-
PaLM (540B)	2048	2048	0.01	-	inverse square root	✓			-	-	-	✓	✓	×
AlexaTM (20B)	2M	1024	1e-4	-	linear decay to 5%		✓			✓		✓	-	✓
U-PaLM (540B)	32	2048	1e-4	-	cosine	✓			-	-	-	-	-	-
UL2 (20B)	1024	1024	-	-	inverse square root	-	-	-	-	-	-	×	-	-
GLM (130B)	4224	2048	8e-5	✓	cosine			✓	✓			✓	✓	✓
CodeGen (16B)	2M	2048	5e-5	✓	cosine		✓		-	-	-	✓	✓	-
LLaMA (65B)	4M Tokens	2048	1.5e-4	✓	cosine decay to 10%			✓	-	-	-	✓	✓	-
PanGu- Σ (1.085T)	512	1024	2e-5	✓	-		✓				✓	-	-	-
BloombergGPT (50B)	2048	2048	6e-5	✓	cosine			✓			✓	✓	✓	×
Xuan Yuan 2.0 (176B)	2048	2048	6e-5	✓	cosine		✓		✓			✓	✓	-
CodeT5+ (16B)	2048	1024	2e-4	-	linear			✓			✓	✓	-	-
StarCoder (15.5B)	512	8k	3e-4	✓	cosine		✓			✓		✓	-	-
LLaMA-2 (70B)	4M Tokens	4k	1.5e-4	✓	cosine			✓		✓		✓	✓	-

Model Configurations

Table 7: Summary of optimization settings used for instruction-tuned LLMs. Values for gradient clipping and dropout are the same as the pre-trained models, while no model uses weight decay for instruction tuning.

Models	Batch Size	Sequence Length	LR	Warmup	LR_Decay	Optimizers			Grad Clip	Dropout
						AdaFactor	Adam	AdamW		
WebGPT (175B)	BC:512, RM:32	-	6e-5	-	-		✓		-	-
T0 (11B)	1024	1280	1e-3	-	-	✓			-	✓
Tk-Instruct (11B)	1024	-	1e-5	-	constant	-	-	-	-	-
OPT-IML (175B)	128	2048	5e-5	×	linear		✓		✓	✓
Flan-U-PaLM (540B)	32	-	1e-3	-	constant	✓			-	✓
Sparrow (70B)	RM: 8+16, RL:16	-	2e-6	✓	cosine decay to 10%	✓			✓	×
WizardCoder (15B)	512	2048	2e-5	✓	cosine	-	-	-	-	-
Alpaca (13B)	128	512	1e-5	✓	cosine	-	-	✓	✓	×
Vicuna (13B)	128	-2048	2e-5	✓	cosine			✓	-	×
LIMA (65B)	32	2048	1e-5	×	linear			✓	-	✓

Datasets and Evaluation

Table 8: Details of various well-known pre-training and fine-tuning datasets. Here, alignment means aligning with human preferences.

Dataset	Type	Size/Samples	Tasks	Source	Creation	Comments
C4 [10]	Pretrain	806GB	-	Common Crawl	Automated	A clean, multilingual dataset with billions of tokens
mC4 [11]	Pretrain	38.49TB	-	Common Crawl	Automated	A multilingual extension of the C4 dataset, mC4 identifies over 100 languages using cld3 from 71 monthly web scrapes of Common Crawl.
PILE [291]	Pretrain	825GB	-	Common Crawl, PubMed Central, OpenWebText2, ArXiv, GitHub, Books3, and others	Automated	A massive dataset comprised of 22 constituent sub-datasets
ROOTs [292]	Pretrain	1.61TB	-	498 Hugging Face datasets	Automated	46 natural and 13 programming languages
MassiveText [116]	Pretrain	10.5TB	-	MassiveWeb, Books, News, Wikipedia, Github, C4	Automated	99% of the data is in English
Wikipedia [293]	Pretrain	-	-	Wikipedia	Automated	Dump of wikipedia
RedPajama [294]	Pretrain	5TB	-	CommonCrawl, C4, Wikipedia, Github, Books, StackExchange	Automated	Open-source replica of LLaMA dataset
PushShift.io Reddit	Pretrain	21.1GB	-	Reddit	Automated	Submissions and comments on Reddit from 2005 to 2019
BigPython [130]	Pretrain	5.5TB	Coding	GitHub	Automated	-
Pool of Prompt (P3) [17]	Instructions	12M	62	PromptSource	Manual	A Subset of PromptSource, created from 177 datasets including summarization, QA, classification, etc.
xP3 [144]	Instructions	81M	71	P3+Multilingual datasets	Manual	Extending P3 to total 46 languages
Super-NaturalInstructions (SNI) [18]	Instructions	12.4M	1616	Multiple datasets	Manual	Extending P3 with additional multilingual datasets, total 46 languages
Flan [16]	Instructions	15M	1836	Muffin+T0-SF+NIV2	Manual	Total 60 languages
OPT-IML [97]	Instructions	18.1M	1667	-	Manual	-
Self-Instruct [19]	Instructions	82k	175	-	Automated	Generated 52k instructions with 82k samples from 175 seed tasks using GPT-3
Alpaca [148]	Instructions	52k	-	-	Automated	Employed self-instruct method to generate data from text-davinci-003
Vicuna [149]	Instructions	125k	-	ShareGPT	Automated	Conversations shared by users on ShareGPT using public APIs
LLaMA-GPT-4 [150]	Instructions	52k	-	Alpaca	Automated	Recreated Alpaca dataset with GPT-4 in English and Chinese
Unnatural Instructions [295]	Instructions	68k	-	15-Seeds (SNI)	Automated	-
LIMA [175]	Instructions	1k	-	Multiple datasets	Manual	Carefully created samples to test performance with fine-tuning on less data
Anthropic-HH-RLHF [296]	Alignment	142k	-	-	Manual	
Anthropic-HH-RLHF-2 [168]	Alignment	39k	-	-	Manual	

Datasets and Evaluation

Table 9: Categorized evaluation datasets used in evaluating LLMs.

Type	Datasets/Benchmarks
Multi-Task	MMLU [297], SuperGLUE [2], BIG-bench [298], GLUE [299], BBH [298], CUGE [300], Zero-CLUE [301], FewCLUE [302], Blended Skill Talk [303], HELM [304], KLUE-STS [305]
Language Understanding	CoQA [306], WiC [307], Wikitext103 [308], PG19 [309], LCQMC [310], QQP [311], WinoGender [312], CB [313], FinRE [314], SanWen [315], AFQMC [301], BQ Corpus [316], CNSS [317], CKBQA 13 [318], CLUENER [301], Weibo [319], AQUA [320], OntoNotes [321], HeadQA [322], Twitter Dataset [323]
Story Cloze and Sentence Completion	StoryCloze [324], LAMBADA [325], LCSTS [326], AdGen [327], E2E [328], CHID [329], CHID-FC [302]
Physical Knowledge and World Understanding	PIQA [330], TriviaQA [331], ARC [332], ARC-Easy [332], ARC-Challenge [332], PROST [333], Open-BookQA [334], WebNLG [335], DogWhistle Insider & Outsider [336]
Contextual Language Understanding	RACE [337], RACE-Middle [337], RACE-High [337], QuAC [338], StrategyQA [339], Quiz Bowl [340], cMedQA [341], cMedQA2 [342], MATINF-QA [343]
Commonsense Reasoning	WinoGrande [344], HellaSwag [345], COPA [346], WSC [347], CSQA [348], SIQA [349], C ³ [350], CLUEWSC2020 [301], CLUEWSC [301], CLUEWSC-FC [302], ReCoRD [351]
Reading Comprehension	SQuAD [352], BoolQ [353], SQUADv2 [354], DROP [355], RTE [356], WebQA [357], CMRC2017 [358], CMRC2018 [359], CMRC2019 [360], COTE-BD [361], COTE-DP [361], COTE-MFW [361], MultiRC [362], Natural Questions [363], CNSE [317], DRCD [364], DuReader [365], Dureader _{robust} [366], DuReader-QG [365], SciQ [367], Sogou-log [368], Dureader _{robust} -QG [366], QA4MRE [369], KorQuAD 1.0 [370], CAIL2018-Task1 & Task2 [371]
Mathematical Reasoning	MATH [372], Math23k [373], GSM8K [374], MathQA [375], MGSM [376], MultiArith [377], AS-Div [378], MAWPS [379], SVAMP [380]
Problem Solving	HumanEval [131], DS-1000 [381], MBPP [382], APPS [372], CodeContests [132]
Natural Language Inference & Logical Reasoning	ANLI [383], MNLI-m [384], MNLI-mm [384], QNLI [352], WNLI [347], OCNLI [301], CMNLI [301], ANLI R1 [383], ANLI R2 [383], ANLI R3 [383], HANS [385], OCNLI-FC [302], LogiQA [386], StrategyQA [339]
Cross-Lingual Understanding	MLQA [387], XNLI [388], PAWS-X [389], XSum [390], XCOPA [391], XWinograd [392], TyDiQA-GoldP [393], MLSum [394]
Truthfulness and Fact Checking	TruthfulQA [395], MultiFC [396], Fact Checking on Fever [397]
Biases and Ethics in AI	ETHOS [398], StereoSet [399], BBQ [400], Winobias [401], CrowS-Pairs [402]
Toxicity	RealToxicityPrompts [403], CivilComments toxicity classification [404]
Language Translation	WMT [405], WMT20 [406], WMT20-enzh [406], EPRSTMT [302], CCPM [407]
Scientific Knowledge	AminoProbe [138], BioLAMA [138], Chemical Reactions [138], Galaxy Clusters [138], Mineral Groups [138]
Dialogue	Wizard of Wikipedia [408], Empathetic Dialogues [409], DPC-generated [96] dialogues, ConvAI2 [410], KdConv [411]
Topic Classification	TNEWS-FC [302], YNAT [305], KLUE-TC [305], CSL [301], CSL-FC [302], IFLYTEK [412]

Datasets and Evaluation

Table 10: An illustration of training datasets and evaluation tasks employed by pre-trained LLMs. Here, “QA” is question-answering, “Clf” is classification, “NLI” is natural language inference, “MT” is machine translation, “RC” is reading comprehension, “CR” is commonsense reasoning, “MR” is mathematical reasoning, “Mem.” is memorization.

Models	Training Dataset	Benchmark								Cloze/ Completion	RC	CR	MR	Coding	Truthful/ Bias/ Toxicity/ Mem.
		BIG- bench	MMLU	Super GLUE	QA	Cif	NLI	MT							
T5	C4 [10]			✓	✓		✓	✓	✓	✓	✓	✓			
GPT-3	Common Crawl, WebText, Books Cor- pora, Wikipedia			✓	✓			✓	✓	✓				✓	
mT5	mC4 [11]				✓		✓	✓							
PanGu- α	1.1TB Chinese Text Corpus				✓		✓		✓	✓	✓				
CPM-2	WuDaoCorpus [109]									✓		✓			
Codex	54 million public repositories from Github												✓		
ERNIE-3.0	Chinese text corpora, Baidu Search, Web text, QA-long, QA-short, Poetry and Cou- plet Domain-specific data from medical, law, and financial area Baidu knowledge graph with more than 50 million facts			✓	✓	✓	✓	✓	✓	✓		✓			
Jurassic-1	Wikipedia, OWT, Books, C4, Pile [291], arXiv, GitHub				✓		✓		✓	✓					
HyperCLOVA	Korean blogs, Community sites, News, KiN Korean Wikipedia, Wikipedia (En- glish and Japanese), Modu-Corpus: Mes- senger, News, Spoken and written lan- guage corpus, Web corpus							✓							

Datasets and Evaluation

Table 11: An illustration of training datasets and evaluation benchmarks used in fine-tuned LLMs. “SNI” is a short of Super-NaturalInstructions.

Models	Training Dataset	BIG-bench	MMLU	BBH	RAFT	FLAN	SNI	PromptSource	TyDiQA	HumanEval	MBPP	Truthful/ Bias/ Toxicity
T0	Pool of Prompts	✓										
WebGPT	ELI5 [414], ELI5 fact-check [156], TriviaQA [331], ARC-Challenge [332], ARC-Easy [332], Hand-written data, Demonstrations of humans, Comparisons between model-generated answers											✓
Tk-INSTRUCT	SNI [18]						✓					
mT0	xP3 [144]											
OPT-IML	PromptSource [17], FLAN [16], SNI [415], UnifiedSKG [416], CrossFit [417], ExMix [418], T5 [10], Reasoning		✓	✓	✓	✓	✓	✓				
Flan	Muffin, T0-SF, Niv2, CoT		✓	✓					✓			
WizardCoder	Code Alpaca									✓	✓	

Datasets and Evaluation

Table 12: Performance comparison of top performing LLMs across various NLU and NLG tasks. Here, “N-Shots” indicate the number of example prompts provided to the model during the evaluation, representing its capability in few-shot or zero-shot learning settings, “f” represents the fine-tuned version, and “B” represents the benchmark.

Task	Dataset/Benchmark	Top-1		Top-2		Top-3	
		Model (Size)	Score (N-shots)	Model (Size)	Score (N-shots)	Model (Size)	Score (N-shots)
Multi-Task	BIG-bench (B)	Chinchilla (70B)	65.1 (5-shot)	Gopher (280B)	53.97 (5-shot)	PaLM (540B)	53.7 (5-shot)
	MMLU (B)	GPT-4 (-)	86.4 (5-shot)	Gemini (Ultra)	83.7 (5-shot)	Flan-PaLM-2 _(f) (Large)	81.2 (5-shot)
Language Understanding	SuperGLUE (B)	ERNIE 3.0 (12B)	90.6 (-)	PaLM _(f) (540B)	90.4 (-)	T5 (11B)	88.9 (-)
Story Comprehension and Generation	HellaSwag	GPT-4 (-)	95.3 (10-shot)	Gemini (Ultra)	87.8 (10-shot)	PaLM-2 (Large)	86.8 (one shot)
	StoryCloze	GPT3 (175B)	87.7 (few shot)	PaLM-2 (Large)	87.4 (one shot)	OPT (175B)	79.82 (-)
Physical Knowledge and World Understanding	PIQA	PaLM-2 (Large)	85.0 (one shot)	LLaMa (65B)	82.8 (zero shot)	MT-NLG (530B)	81.99 (zero shot)
	TriviaQA	PaLM-2 (Large)	86.1 (one shot)	LLaMA-2 (70B)	85.0 (one shot)	PaLM (540B)	81.4 (one shot)
Contextual Language Understanding	LAMBADA	PaLM (540B)	89.7 (few shot)	MT-NLG (530B)	87.15 (few shot)	PaLM-2 (Large)	86.9 (one shot)
Commonsense Reasoning	WinoGrande	GPT-4 (-)	87.5 (5-shot)	PaLM-2 (Large)	83.0 (one shot)	PaLM (540B)	81.1 (zero shot)
	SIQA	LLaMA (65B)	52.3 (zero shot)	Chinchilla (70B)	51.3 (zero shot)	Gopher (280B)	50.6 (zero shot)
Reading Comprehension	BoolQ	PaLM _(f) (540B)	92.2 (-)	T5 (11B)	91.2 (-)	PaLM-2 (Large)	90.9 (one shot)
Truthfulness	Truthful-QA	LLaMA (65B)	57 (-)				
Mathematical Reasoning	MATH	Gemini (Ultra)	53.2 (4-shot)	PaLM-2 (Large)	34.3 (4-shot)	LLaMa-2 (65B)	13.5 (4-shot)
	GSM8K	GPT-4 (-)	92.0 (5-shot)	PaLM-2 (Large)	80.7 (8-shot)	U-PaLM (540B)	58.5 (-)
Problem Solving and Logical Reasoning	HumanEval	Gemini _(f) (Ultra)	74.4 (zero shot)	GPT-4 (-)	67.0 (zero shot)	Code Llama (34B)	48.8 (zero shot)

Challenges and Future Directions

- Computational Cost
- Bias and Fairness
- Overfitting
- Economic and Research Inequality
- Limited Knowledge
- Safety and Controllability
- Real-Time Processing
- Multi-Modality
- Privacy Concerns

Thank You!

Avenida da Universidade, Taipa, Macau, China

Tel : (853) 8822 8833 Fax : (853) 8822 8822

Email : info@um.edu.mo Website : www.um.edu.mo