

我国大模型与国外主要差距的报告

2022 年 11 月底，美国 OpenAI 公司发布了对话式大规模自然语言预训练模型 ChatGPT 及体验服务，在自然语言处理领域展现出了惊人的通用智能水平，这项技术变革打开了通向通用 AI 的一扇门，通用人工智能正向我们走来，将会对社会各个相关方面带来深度变革。综合来看，在《新一代人工智能发展规划》指引下，在科技创新 2030——新一代人工智能重大项目持续支持下，近年来我国人工智能技术和产业发展迅速，总体上与美国同属第一发展梯队；但自 2020 年初的新冠疫情以及国际形势的变化，极大的阻碍了我国人工智能领域的研究与国外同行的交流，一定程度上造成了我国与国外的一定差距，并且这种差距有逐渐拉大的趋势，需要引起我们的高度重视。本报告从大模型构建的模型设计、训练数据、训练方法以及性能表现方面，对我国大模型与国外的主要差距进行分析和介绍。

1、模型设计

模型设计方面，国内外的大模型本身在同类基础模型上差异不大，普遍为 Transformer 架构。模型的基座模型设计大体上分为以下三种：

(1) 仅包含解码器 (Decoder-only)，即自回归 (Autoregressive) 模型，代表模型是 GPT[1]。本质上是一个从左到右的语言模型，训练目标是从左到右的文本生成。常用于无条件长文本生成（对

话生成、故事生成等），但缺点是单向注意力机制，不利于 NLU（自然语言理解）任务；

（2）仅包含编码器（**Encoder-only**），即自编码（**Autoencoder**）模型，代表模型是 **BERT**、**ALBERT**、**DeBERTa**[2-4]。自编码模型是通过去噪任务（如利用掩码语言模型）学习双向的上下文编码器，训练目标是对文本进行随机掩码，然后预测被掩码的词。常用于自然语言理解（事实推断、语法分析、分类等），缺点是不能直接用于文本生成；

（3）编码器-解码器（**Encoder-Decoder**），即完整的 **Transformer** 结构，代表模型是 **T5**、**BART**[5,6]。包含一个编码器和一个解码器，接受一段文本，从左到右的生成另一段文本。常用于有条件的生成任务（摘要生成、对话等）。缺点是比 **Encoder-based** 或 **Decoder-based** 模型在同性能下需要更多参数。

以下是对国内外开放的一些模型进行上述架构上的归类介绍：国外：**GPT-3**、**BLOOM**、**OPT**、**LaMDA**、**PALM**[7-11] 以及 **ChatGPT** 等采用 **Decoder** 结构，**Switch Transformer**[12]、**flan-ul2**[13]、**T5** 采用 **Encoder-Decoder** 结构。国内：**GLM**、**PanGu-Alpha**、**Ernie 3.0 Titan**[14-16] 采用 **Decoder** 结构，**PLUG**、**CPM-2**[17] 采用 **Encoder-Decoder** 结构。几乎没有千亿参数量的 **Encoder** 结构，这是因为 **Encoder** 结构的表示能力有限，只能用于特定的任务，对于大模型来说，虽然可以提高特定任务的性能，但一定程度上是“大

材小用”。因此大模型更倾向于使用 **Decoder** 架构，通过自回归预训练，使得其能力更强，生成更丰富，应用更广泛。

机构	模型	参数	语言	时间	架构
百度	Ernie 3.0 Titan	260B	---	2022-01	解码器
清华大学	GLM	130B	中英文	2022-10	解码器
	GLM	10B	中文	2022-09	解码器
浪潮	源 1.0	245B	中文	2021-09	解码器
智源研究院	CPM-2	11B	中文	2021-06	编码器-解码器
	CPM-2	11B	中英文	2021-06	编码器-解码器
	CPM-2	200B	中英文	2021-06	编码器-解码器
鹏城实验室与华为	PanGu-A lpha	13B	中文	2021-05	解码器
	PanGu-A lpha	200B	中文	2021-05	解码器
Meta	OPT	175B	多语言	2022-10	解码器
OpenAI	GPT-3	175B	多语言	2021-10	解码器
Google	LaMDA	137B	多语言	2021-05	解码器
	PALM	540B	多语言	2021-12	解码器

表 1：国内外大模型的训练架构对比

相关文献

- [1] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.
- [2] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.
- [3] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.
- [4] He P, Liu X, Gao J, et al. Deberta: Decoding-enhanced bert with disentangled attention[J]. arXiv preprint arXiv:2006.03654, 2020.
- [5] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551.
- [6] Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[J]. arXiv preprint arXiv:1910.13461, 2019.
- [7] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
- [8] Scao T L, Fan A, Akiki C, et al. Bloom: A 176b-parameter open-access multilingual language model[J]. arXiv preprint arXiv:2211.05100, 2022.
- [9] Zhang S, Roller S, Goyal N, et al. Opt: Open pre-trained transformer language models[J]. arXiv preprint arXiv:2205.01068, 2022.
- [10] Thoppilan R, De Freitas D, Hall J, et al. Lamda: Language models for dialog applications[J]. arXiv preprint arXiv:2201.08239, 2022.
- [11] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint arXiv:2204.02311, 2022.
- [12] Fedus W, Zoph B, Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity[J]. J. Mach. Learn. Res, 2021, 23: 1-40.
- [13] Tay Y, Dehghani M, Tran V Q, et al. Unifying language learning paradigms[J]. arXiv preprint arXiv:2205.05131, 2022.
- [14] Zeng A, Liu X, Du Z, et al. Glm-130b: An open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.

- [15]Zeng W, Ren X, Su T, et al. PanGu- α : Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation[J]. arXiv preprint arXiv:2104.12369, 2021.
- [16]Wang S, Sun Y, Xiang Y, et al. Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation[J]. arXiv preprint arXiv:2112.12731, 2021.
- [17]Zhang Z, Gu Y, Han X, et al. Cpm-2: Large-scale cost-effective pre-trained language models[J]. AI Open, 2021, 2: 216-224.

2、训练数据

相较于国外，我国用于训练大模型的开源数据集更侧重于中文语料，但总体数量较少、丰富性不高且质量较低。下面从大模型常用的三个训练阶段所用到的数据进行阐述：预训练、指令微调和基于人工反馈的强化学习：

在模型预训练方面，我国开源数据集数量较少，且丰富性不高。截至目前，我国开源预训练数据集包括：GLM系列的悟道数据集，其规模为3TB（已开源200GB）；CLUE社区的开源中文数据集CLUE Corpus 2020[1]，其规模为100GB；里屋社区的开源数据集MNBVC，其规模约2.3TB，为互联网收集的中文纯文本语料数据集。

相比较而言，国外开源数据集有数量更多且总体规模更大。例如，PB级的CommonCrawl的网页数据、BigScience开源的1.6TB多语数据集ROOTS[2]、Gao等人构建的825G数据集The Pile[3]。此外，国外开源数据集质量更高且种类丰富：ROOTS既包含网页数据，又收集了GitHub上的代码数据，也从各种下游任务数据集

中收集高质量数据；The Pile 数据集基于学术或专业领域知识源构造，使得该数据的质量很高；还有英文小说数据集 BookCorpus[4]（约 37GB）；百科全书数据源 Wikipedia（20GB 左右）。

	数据集	发布者	规模	特点
国外	BookCorpus	Zhu et al.	Book1: 2.2GB; Book3: 37GB	英文小说数据集
	Wikipedia	维基媒体基金会	100GB	百科全书数据集
	Common Crawl	Common Crawl 团队	PB 级	网页数据，规模大
	ROOTS	BigScience	1.6TB	多元多语数据
	The Pile	Gao et al.	825GB	数据广泛，质量高
国内	悟道	智源研究院	3TB(200GB)	中文数据
	CLUECorpus 2020	CLUE 社区	100GB	中文数据
	MNBVC	里屋社区	2.4TB	中文数据集，互联网数据

表 2：用于大模型预训练的国内外语料对比

在指令微调方面，国外存在开源的指令微调数据集，通过人类标注、模型扩展的方式进行构造，例如包含 1617 种指令的 Super-Natural Instructions[5]（实例数为 5M）和覆盖 52K 种指令的 Self-instruct 数据集[6]（实例数为 82K）。截至目前，国内尚未出现开源的中文指令微调数据集，为此智源社区于 3 月 14 日发布

OpenLabel 数据标注平台，共建 AI 开源数据集，目前开展的第一阶段关注指令任务数据。

截至目前，在基于人工反馈的强化学习方面上国内尚未出现开源的中文人类反馈强化学习数据集，国外的 Anthropic 公司[7]、斯坦福大学、HuggingFace 团队各发布了一个开源人类反馈的数据集，规模为 300K-10M 不等，数据主要来源于和大模型对话和网上论坛。

综上所述，我国预训练数据集较少且数据来源较为单一，缺乏高质量数据集。高质量预训练大模型越来越依赖高质量的指令数据来激活其强大能力，指令数据的缺少将直接导致预训练模型的性能损失。考虑到中文和英文为不同语系，多语言训练和语言的泛化能力无法很好的在中英文之间有效迁移，中文数据集的缺失问题亟待长期布局和解决。

相关文献

- [1] Xu L, Zhang X, Dong Q. CLUECorpus2020: A large-scale Chinese corpus for pre-training language model[J]. arXiv preprint arXiv:2003.01355, 2020.
- [2] Laurençon H, Saulnier L, Wang T, et al. The bigscience roots corpus: A 1.6 tb composite multilingual dataset[J]. arXiv preprint arXiv:2303.03915, 2023.
- [3] Gao L, Biderman S, Black S, et al. The pile: An 800gb dataset of diverse text for language modeling[J]. arXiv preprint arXiv:2101.00027, 2020.
- [4] Zhu Y, Kiros R, Zemel R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books[C]//Proceedings of the IEEE international conference on computer vision. 2015: 19-27.
- [5] Wang Y, Mishra S, Alipoormolabashi P, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 5085-5109.

- [6] Wang Y, Kordi Y, Mishra S, et al. Self-Instruct: Aligning Language Model with Self Generated Instructions[J]. arXiv preprint arXiv:2212.10560, 2022.
- [7] Ganguli D, Lovitt L, Kernion J, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned[J]. arXiv preprint arXiv:2209.07858, 2022.

3、训练方法

训练硬件：国内的盘古预训练模型[1]采用华为自研的昇腾架构的 910 处理器进行训练, GLM[2]采用多个节点的 8 卡 A100 GPU 集群进行训练, ERNIE 3.0[3]使用了 384 块 V100 GPU 集群进行训练。国外的模型[4-9]多采用 A100 GPU 集群进行训练, 包括 GPT-3[4]、BLOOM[5]、OPT[6]、LLaMA[7]。谷歌的 PaLM[8]模型采用自研的 TPU 进行训练。GPT-4[9]并未公开训练硬件方面的细节。

	模型	硬件	训练框架	训练目标
国内	盘古	2048 Ascend 910 处理器	MindSpore	自回归
	GLM	96 A100 (40G * 8)	PipeDream- Flush (Megatron- DeepSpeed)	上下文填 空, 自回归
	悟道 2.0	-	FastMoe	-
	ERNIE 3.0	384 V100 GPU _s	-	句子重排, 句子距离度 量, 自回归
国外	GPT-3	V100	-	自回归
	BLOOM	384 80GB A100	Megatron-D eepSpeed	自回归

	OPT	992 80GB A100 GPUs	Megatron-DeepSpeed	自回归
	LLaMA	2048 80GB A100 GPUs	-	自回归
	PaLM	2*3072 TPU v4	-	自回归
	GPT-4	-	-	-

表 3：国内外大模型的训练细节对比

训练框架：国内的盘古预训练模型采用华为自研的框架 MindSpore 进行分布式训练，为 Ascend AI 处理器提供原生支持以及软硬件协同优化，以提高训练效率。GLM 采用开源训练加速框架 Megatron-DeepSpeed[10]，并使用了 PipeDream-Flush[11]进行流水线优化。悟道 2.0 使用了自研的 FastMoe 技术提升了 MoE 的速度和效率。国外开源的 BLOOM 和 OPT 模型是采用 Megatron-DeepSpeed 框架以加速在英伟达 GPU 上的训练。而其余的模型包括 GPT-4 并未公开训练框架方面的细节。

训练任务：国内的盘古没有公开预训练任务，ERNIE 3.0 使用了句子重排 (sentence reordering)，知识掩码建模 (knowledge masked language modeling)，文档建模 (document language modeling)，句间距离度量 (sentence distance) 等预训练任务来对模型进行预训练，GLM 使用了自回归空白填充作为其主要的预训练目标，掩盖随机的连续文本区间，并使用自回归的方法进行预测。悟道 2.0 是一个 MoE 模型，预训练任务未公开。国外的模型在预训练任务上基本均是基于自回归的预训练方法：预测下一个词的自回归训

练目标。GPT-4 使用了 RLHF（人类反馈强化学习）使得模型在安全性、事实性方面的表现更好。

相关文献

- [1] 曾炜, 苏腾, 王晖等. 鹏程·盘古:大规模自回归 中文预训练语言模型及应用[J]. 中兴通讯技术, 2022, 28: 33-43.
- [2] Zeng A, Liu X, Du Z, et al. Glm-130b: An open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.
- [3] Sun Y, Wang S, Feng S, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation[J]. arXiv preprint arXiv:2107.02137, 2021.
- [4] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
- [5] Scao T L, Fan A, Akiki C, et al. Bloom: A 176b-parameter open-access multilingual language model[J]. arXiv preprint arXiv:2211.05100, 2022.
- [6] Zhang S, Roller S, Goyal N, et al. Opt: Open pre-trained transformer language models[J]. arXiv preprint arXiv:2205.01068, 2022.
- [7] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and Efficient Foundation Language Models[J]. arXiv preprint arXiv:2302.13971, 2023.
- [8] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint arXiv:2204.02311, 2022.
- [9] GPT-4 Technical Report. OpenAI, 2023. Web. 14 March 2023. <https://cdn.openai.com/papers/gpt-4.pdf>.
- [10] Rasley J, Rajbhandari S, Ruwase O, et al. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 3505-3506.
- [11] Narayanan D, Phanishayee A, Shi K, et al. Memory-efficient pipeline-parallel dnn training[C]//International Conference on Machine Learning. PMLR, 2021: 7937-7947.

4、性能表现

国外大语言模型（参数量大于 1000 亿）GPT-4、GPT-3、OPT-175B、PaLM-540B、BLOOM-176B、GLM-130B 等涌现出较强的自然语言理解和推理能力，在多个自然语言处理评测任务上达到优异的性能，甚至超越人类[1]。基于大语言模型 GPT-3，OpenAI 进一步研发的通用对话系统 ChatGPT 在代码生成、多语言翻译、文本内容创作等方面表现极佳，泛化性和通用性表现极好，可用于金融、医疗、教育等领域百余种行业[2]，引发了国内外追逐人工智能技术的浪潮。其中，微软将 ChatGPT 与搜索引擎相结合，能够更准确理解用户的查询指令并可与用户进行多轮交互，进一步提升用户的使用体验和查询信息的准确性。2023 年 3 月 15，OpenAI 发布了超大多模态语言模型 GPT-4[2]，可以同时输入图片和自然语言文本，在各种职业和学术考试上表现和人类水平相当，比如模拟律师考试，GPT-4 取得了前 10%的好成绩，而 GPT-3.5 是倒数 10%。美国高考 SAT 试题测试结果表明，GPT-4 可在阅读写作中拿下 710 高分，数学 700 分（满分 800），在多数方面已经达到了甚至超越了人类，虽然其还有部分出现幻觉和胡说八道的小瑕疵，但这并不影响它带来的惊艳。

紧跟国外大模型技术发展，百度公司发布了大规模中文语言模型 ERNIE Titan 3.0 260B[3]，清华大学科研团队开发了中英双语大模型 GLM-130B[4]，鹏城实验室发布了 2000 亿参数的盘古 α 大模型，且北京智源研究院发布了首个 26 亿参数以中文为核心的大模型。

规模预训练语言模型「悟道·文源」。其中，GLM-130B 在多个中文自然语言处理任务上超越国外大模型(GPT-4 除外)的表现,偏见和生成毒性显著降低，在大规模多样性多任务语言理解 MMLU[5] 上，GLM-130B 小幅超越 GPT-3 0.9%，大幅超过 BLOOM-176B，随着持续训练，水平尚未收敛并将持续提高。其次，GLM-130B 以零样本方式在中文语言理解基准（CLUE[6]）和少样本中文理解基准（FewCLUE[7]）中的 12 个任务（7 个 CLUE 数据集和 5 个 FewCLUE 数据集）上全面超越了 ERNIE Titan 3.0 260B。此外，GLM-130B 在 2 个阅读理解数据集（DRCD[8]和 CMRC2018[9]）上的性能要好至少 260%。鹏城实验室发布的“盘古 α ”，是业界首个全开源 2000 亿参数中文预训练模型。在模型性能方面，鹏城·盘古大模型性能全球领先，16 个下游任务中性能指标优于业界 SOTA 模型，其中零样本学习任务 11 个任务领先、单样本学习任务 12 个任务领先、小样本学习任务 13 个任务领先。在应用方面，鹏城·盘古支持丰富的应用场景，在知识问答、知识检索、知识推理、阅读理解等文本生成领域表现突出。

总体来看，国内大语言模型与国外大语言模型相比，在自然语言文本理解和推理方面差距较小，在部分自然语言处理任务上已超越国外大语言模型。但是，在通用性、鲁棒性和泛化性方面表现差距较大，主要集中在多语言文本理解、代码生成、图像理解、数学符号推理等方面。

模型	发布单位	特点	应用范围	局限性
鹏城·盘古	鹏城实验室	首个拥有全开源 2 000 亿参数的自回归中文预训练语言大模型	文本编辑、编程、翻译、算数	对话能力较弱,代码生成质量较差、准确性低
GLM-130B	清华大学	以较少参数达到了可媲美 GPT-3-175B 优异的效果	文本编辑、编程、翻译、算数	对话能力较弱、不具备多语言理解能力、代码生成能力较弱
混元	腾讯	多模态理解、跨模态理解	自然语言处理、多模态内容理解、文案生成	用户使用量较少
文心一言	百度	生成式搜索、跨模态理解交互	文本生成、加入百度搜索引擎	生成文本较短、事实准确性低
Claude	Anthropic	最大化积极影响、避免提供有害建议、自主选择、加入基于人工智能反馈的强化学习范式	文本编辑更长、更自然,提高事实性和安全性	代码推理能力较弱
Bard	Google	可根据最新事件进行对话、更负责任、与搜索引擎相结合	通过加入 ChromeOS 为搜索引擎、实时外部知识增强	易发生事实性错误、安全性低

GPT-3	Open-AI	大型通用语言模型，可以处理各种语言处理任务	多语言翻译、摘要、代码生成、和文本生成	对话能力较弱，无法连续对话
ChatGPT	Open-AI	支持连续对话、可质疑、主动承认错误、人类反馈的强化学习训练范式	文本编辑、编程、翻译、算数、表格生成	无法进行网页搜索、黑箱问题
GPT-4	Open-AI	超大多模态大模型，输入图像和文本，在各种职业和学术考试上表现和人类水平相当	文本编辑、对话、翻译、算数、编程、问答、文本生成、摘要、多模态内容理解。	在部分场景幻觉和胡说八道问题、模型结构和训练方式未知，实验结果无法考核

表 4：国内外大模型的性能表现对比

相关文献

- [1] Paperno, Denis, et al. "The LAMBADA dataset: Word prediction requiring a broad discourse context." ACL, 2016.
- [2] OpenAI. GPT-4 Technical Report. OpenAI, 2023. Web. 14 March 2023. <https://cdn.openai.com/papers/gpt-4.pdf>.
- [3] Wang S, Sun Y, Xiang Y, et al. Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation[J]. arXiv preprint arXiv:2112.12731, 2021.
- [4] Zeng A, Liu X, Du Z, et al. Glm-130b: An open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.
- [5] Hendrycks, Dan, et al. "Measuring Massive Multitask Language Understanding." ICLR. 2020.

- [6] Xu, Liang, et al. "CLUE: A Chinese Language Understanding Evaluation Benchmark." COLING, 2020.
- [7] Xu, Liang, et al. "Fewclue: A chinese few-shot learning evaluation benchmark." arXiv preprint arXiv:2107.07498 (2021)
- [8] Shao C C, Liu T, Lai Y, et al. DRCD: A Chinese machine reading comprehension dataset[J]. arXiv preprint arXiv:1806.00920, 2018.
- [9] Cui Y, Liu T, Che W, et al. A span-extraction dataset for Chinese machine reading comprehension[J]. arXiv preprint arXiv:1810.07366, 2018.