2023 年 02 月 23 日

ChatGPT，深度拆解 AI 算力模型

计算机行业

评级及分析师信息

行业评级：推荐

行业走势图

分析师：刘泽晶
邮箱：liuzj1@hx168.com.cn
SAC NO：S1120520020002
联系电话：

ChatGPT 算法的核心壁垒

(1) 庞大的数据训练数据，往往意味着模型精准度的上升；数据量大，往往意味着数据特征维度大，模型的参数越复杂，训练数据维度跟算力指数呈现正相关，算力成本高。

(2) 底层算法 Transformer，相较于传统神经网络综合特征提取能力、远距离特征捕获能力、语义特征提取能力，全部明显增强，正逐步取代 RNN(循环神经网络)。

(3) AI 预训练模型(大模型)，本质是“大算力+强算法”结合的产物，对自然语言理解能力明显上升，谷歌 BERT 模型就是典型跨时代的例子，我们认为其是 AIGC 的初始应用算法。

(4) 多模态数据协同，极大推动 AIGC 的内容多样性与通用性，让 AIGC 不只局限于文本和图像等单个部分，而是多应用相容。

不同类别 AIGC 算法比对

1、ChatGPT ：训练模型为强化学习近端策略优化，可以理解成在“人脑思维”的基础上加入了“人类反馈系统”，是一种奖励模型，拥有 175B 参数，训练数据为语言文本。

2、LaDMA(谷歌 Bard)：参数方面为 137B，奖励模型是人类评分机制，训练数据为对话数据。

3、图神经网络(GNN)作为科学领域预训练模型(大模型)备受瞩目，强大之处在于数据结构，其应用广阔例如推荐系统、药物发现、合成物发现、芯片设计等众多科学前沿领域。

国产 ChatGPT 生态正在形成

百度是少有预训练模型(大模型)语言训练能力的公司，已经经历多次迭代，**参数方面**，模型基于 ERNIE 3.0，拥有千亿级参数。**预训练方面**，具备海量知识沉淀和丰富场景的文心大模型，**跨模态方面**，已有地理-语言、视觉-语言、语音-语言等模型架构，已覆盖众多方向，例如自然语言处理、机器视觉等其他重大任务，此外，根据 IDC 数据，目前已有近百万开发者使用文心大模型，生态正在逐步繁荣，合作厂商覆盖科技、教育、工业、媒体、金融等诸多产业。

投资建议：关注鸿蒙 OS 的生态伙伴

我们认为 AIGC 的出世会产生革命性的影响，同时有望赋能千行百业。我们梳理了三条路径图，积极的推荐以下三条投资主线：

1）具备算力基础的厂商，受益标的为寒武纪、商汤、海光信息、浪潮信息、中科曙光、景嘉微、联想集团、紫光股份、龙芯中科；

2）具备 AI 算法商业落地的厂商，重点推荐科大讯飞、拓尔思，其他受益标的为：汉王科技、海天瑞声、云从科技；

3） AIGC 相关技术储备的应用厂商，受益标的为：百度、同花顺、三六零、金山办公。

风险提示

核心技术水平升级不及预期的风险；AI 伦理风险；政策推进不及预期的风险；中美贸易摩擦升级的风险。

正文目录

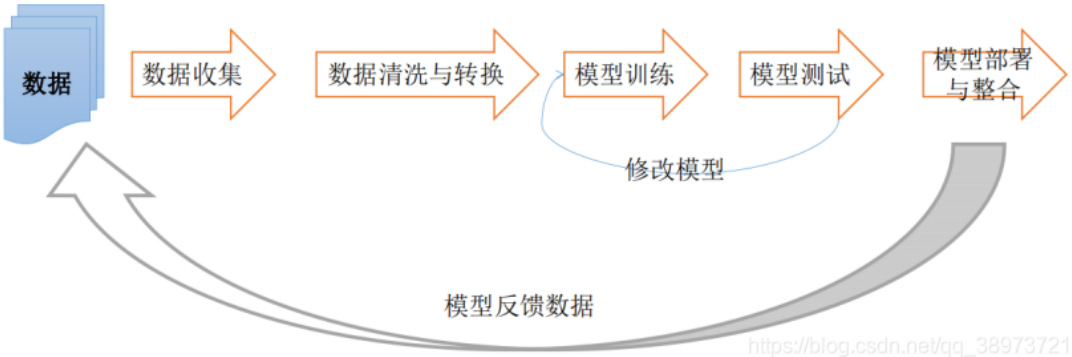
图目录

1.ChatGPT，深度拆解 AI 算力模型

1.1. ChatGPT 算法的核心壁垒

AI 的完整算法生成分为五部分分别是数据收集、数据清洗、模型训练、模型测试、模型部署和反馈。

图表 1 AI 算法的全流程

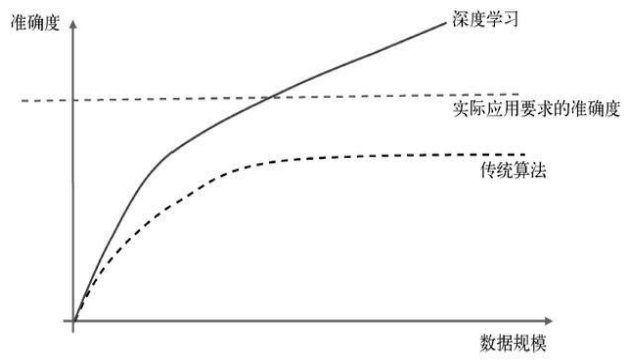


资料来源：CSDN，华西证券研究所

核心壁垒一，庞大的数据训练数据。数据是所有人工智能(或大数据)的“燃料”，根据 appen 的数据，ChatGPT 的前身 GPT-3 就使用了 3,000 亿单词、超过 40T 的大规模、高质量数据进行训练。ChatGPT 在其基础上，加入了人工打标的监督学习，即对话式模型给出结果后，由训练师对结果做出评价并修改结果以更贴切对话内容。

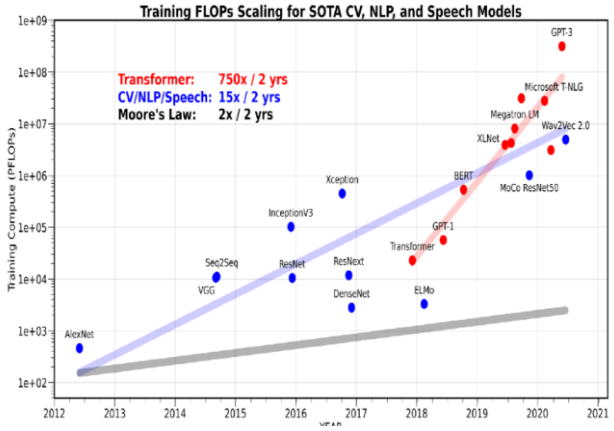
原因，往往愈发庞大的“燃料”意味着模型的精准度的提升，数据量的大小跟深度学习(大数据)的准确度庞大的正相关。此外，**数据量的大小对于运算计算机算力的要求往往呈现指数级别的关系**，这也是强大算法的核心需求。原因是数据清洗和数据标注的核心意义就是将人们理解的非结构化数据转变成计算机可以理解的结构化数据。可以将人工智能的本质理解成矩阵的运算，矩阵的维度往往代表着数据特征的维度，这也是训练神经网络参数的基础，一般情况下，数据维度越多，模型参数量越多，模型越复杂，模型的准确度越高，对算力的指数需求越高。本质是数据维度与算力指数呈现正相关。(不考虑参数堆积、模型过拟合的情况)

图表 2 模型的准确度和数据数量呈现正相关



资料来源：知乎，华西证券研究所

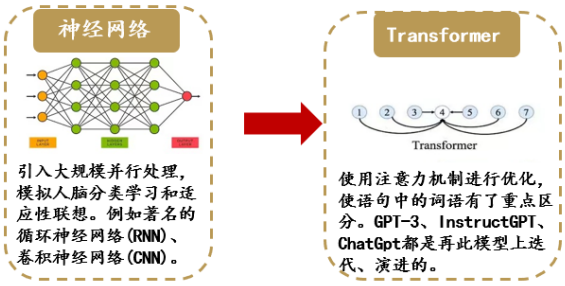
图表 3 AI 需求呈现指数级别的增长



资料来源：腾讯云，华西证券研究所

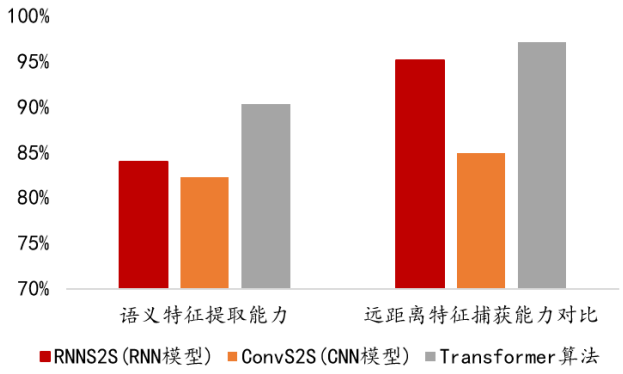
第二，模型训练方面，ChatGPT 强大的底层技术是 Transformer 算法，该算法正逐步取代 RNN(循环神经网络)。Transformer 算法在神经网络中具备跨时代的意义：RNN 和 CNN 已经广泛应用于序列模型、语言建模、机器翻译并取得不错效果，然而在算法上仍有一定限制和不足。Transformer 具备跨时代的意义的原因是算法上添加了注意力机制，这种机制具备突破性的原因在于 1、突破了 RNN 模型不能并行计算的限制；2、相比 CNN 模型，关联所需的操作次数不随距离增长；3、模型解释力度明显加强。从结果上看，根据 CDSN 数据，Transformer 的综合特征提取能力、远距离特征捕获能力、语义特征提取能力，全部明显增强，因此此算法正逐步取代 RNN 算法，也是 ChatGPT 算法的底座。

图表 4 Transformer 算法的前世今生



资料来源：公开资料整理，华西证券研究所

图表 5 Transformer 模型与 RNN、CNN 模型准确度对



资料来源：CSDN，华西证券研究所

第三，模型训练部分，AI 预训练模型(大模型)引发了 AIGC 技术能力的质变。在该模型问世之前，具有使用门槛高、训练成本低、内容生成简单和质量偏低等问题。而在 AIGC 领域，AI 预训练模型拥有巨大参数量模型，AI 预模型可以实现多任务、多语言、多方式等至关重要的作用。

AI 预训练模型的出正是人工智能发展的未来和趋势，AI 预训练模型(大模型)即“大算力+强算法”结合的产物。大模型通常是在大规模无标注数据上进行训练，学习出一种特征和规则。基于大模型进行应用开发时，将大模型进行微调，如在下游特定任务上的小规模有标注数据进行二次训练，或者不进行微调，就可以完成多个应用场景的任务。

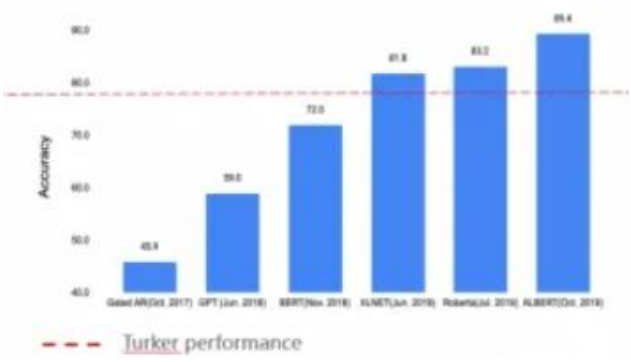
AI 预训练模型的本质是机器对自然语言理解能力的不断提升：其根本原因除 Transformer 算法以外，还有就是参数量的大小，谷歌 BERT 网络模型的提出，使得参数量首次超过 3 亿规模，GPT-3 模型超过百亿。此外，目前较火热 AIGC 的参数量已经超过千亿。此外，**参数量往往是计算空间的复杂程度，模型空间越复杂，往往意味着庞大的计算量，计算量和参数量呈现正比关系**。这也是随着 AI 的功能强大，AI 对算力呈现指数级别根本需求的本质原因。

图表 6 深度学习初期模型越来越大

经典神经网络	AlexNet	VGG16	Inception-V3
模型内存(MB)	＞200	＞500	90-100
参数(百万)	60	138	23.2
计算量(百万)	720	15300	5000

资料来源：博客网，华西证券研究所

图表 7 预模型出现后机器对自然语言的理解不断提升



资料来源：知乎，华西证券研究所

此外，预训练模型(大模型)，按照应用的基本类型分类:可分为 1、自然语言处理(NLP)，例如谷歌的 LaMDA 和 PaLM、OpenAI 的 GPT 系列；2、计算机视觉(CV)，例如微软的 Florence；3、多模态即融合文字、图片、音视频等多种内容形式，例如 OpenAI 的 DALL-E2； 此外，根据不同的领域的应用，可以将预训练模型进一步分类。

图表 8 国外主要 AIGC 预训练模型一览

厂商	预训练模型	应用	参数量	领域
谷歌	BERT	语言理解与生成	4810 亿	NLP
	LaMDA	对话系统		NLP
	PaLM	语言理解与生成、推理、代码生成	5400 亿	NLP
	Imagen	语言理解与图像生成	110 亿	多模态
	Parti	语言理解与图像生成	200 亿	多模态
微软	Florence	视觉识别	6.4 亿	CV
	Turing-NLG	语言理解、生成	170 亿	NLP
Facebook	OPT-175B	语言模型	1750 亿	NLP
	M2M-100	100 种语言互译	150 亿	NLP
Deep Mind	Gato	多面手的智能体	12 亿	多模态
	Gopher	语言理解与生成	2800 亿	NLP
	AlphaCode	代码生成	414 亿	NLP
Open AI	GPT3	语言理解与生成、推理等	1750 亿	NLP
	CLIP&DALL-E	图像生成、跨模态检索	120 亿	多模态
	Codex	代码生成	120 亿	NLP
	ChatGPT	语言理解与生成、推理等		NLP
英伟达	Megatron-Turing NLG	语言理解与生成、推理等	5300 亿	NLP
Stability AI	Stable Diffusion	语言理解与图像生成		多模态

资料来源：腾讯《AIGC 发展报告 2023》，华西证券研究所

谷歌 BERT 作为自然语言处理(NLP)是预训练模型(大模型)的里程碑之作：BERT 模型是谷歌 2018 年发布的的掩码语言模型，当时发布后，在许多自然语言理解任务上取得了最先进的性能，**被当时誉为最先进的神经网络模型。其具有里程碑式结果如下，**机器阅读理解顶级水平测试 SQuAD1.1 中表现出惊人的成绩：全部两个衡量指标上全面超越人类，并且还在 11 种不同 NLP 测试中创出最佳成绩，包括将 GLUE 基准推至 80.4%（绝对改进 7.6%），MultiNLI 准确度达到 86.7% （绝对改进率 5.6%）等。

BERT 取得跨时代的意义是新的预训练模型：在 BERT 模型出世之前，现有的技术已经严重限制了预训练表示的能力，原因是标准语言模型架构是单向的，因此，Bert 采用了 Transformer 技术的双向编码器表示。与最近的其他语言表示模型不同，BERT 旨在通过联合调节所有层中的上下文来预先训练深度双向表示。因此，预训练的 BERT 表示可以通过一个额外的输出层进行微调，适用于广泛任务的最先进模型的构建，比如问答任务和语言推理，无需针对具体任务做大幅架构修改。

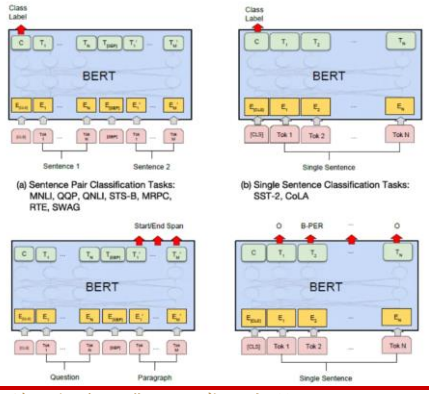
模型的预训练核心机制是其具备里程碑的根本原因：语言建模（15% 的标记被屏蔽，训练目标是在给定上下文的情况下预测原始标记）**和下一句预测**（训练目标是对两个文本跨度进行分类）依次出现在训练语料库中）。因此，BERT 学习了上下文中单词和句子的潜在表示，例如语言推理、文本分类和基于序列到序列的语言生成任务，**此外该阶段的计算成本明显高于微调。我们认为该算法是 AIGC 的初始应用算法。**

图表 9 谷歌 GBRT 取得的能力

Rank	Model	EM	F1
	Human Performance *Stanford University* *(Rajpurkar et al. '16)*	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google A.I.*	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google A.I.*	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

资料来源：知乎，华西证券研究所

图表 10 谷歌 GBRT 预训练架构



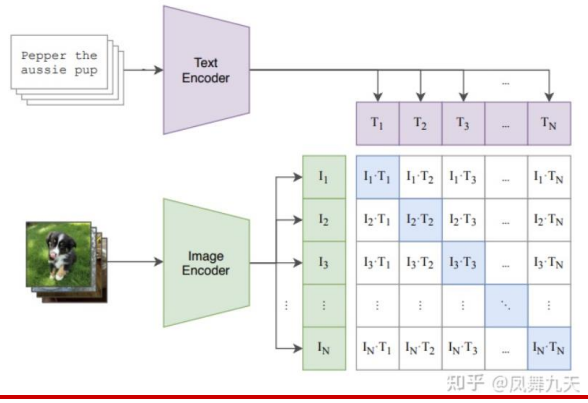
资料来源：稀土掘金，华西证券研究所

第四，模型训练方面，多模态数据协同极大的推动 AIGC 的内容多样性与通用性： 预训练模型更具备通用性、多才艺的根本原因得益于多模型技术(multimodal technology)的使用，即多模态表示图像、声音、语音融合的机器学习。2021 年，OpenAI 团队将跨模态深度学习(CLIP)开源，CLIP 能够将文字和图像进行关联，比如将文字“狗”和图像狗进行关联。CLIP 的优势有两点：

1、同时进行自然语言处理(NLP)和计算机视觉分析(CV)，实现文本和图像的匹配；2、CLIP 模型利用互联网的照片“文本-图像”进行训练，这为后续 AIGC 奠定基础，极大减少数据标注的工作量。

多模态同样具有跨时代的意义： 因此，在多模态技术的支持下，预训练模型已经从早期单一的自然语言处理和机器视觉发展成自动生成图画、图像文字、音视频等多模态、跨模态图型。Dall·E2 就是典型的代表，CLIP 模型让文字和图片两个模态找到能够对话的交界点。

图表 11 CLIP 算法示意图



资料来源：知乎，华西证券研究所

图表 12 Dall·E2 自动生成图画

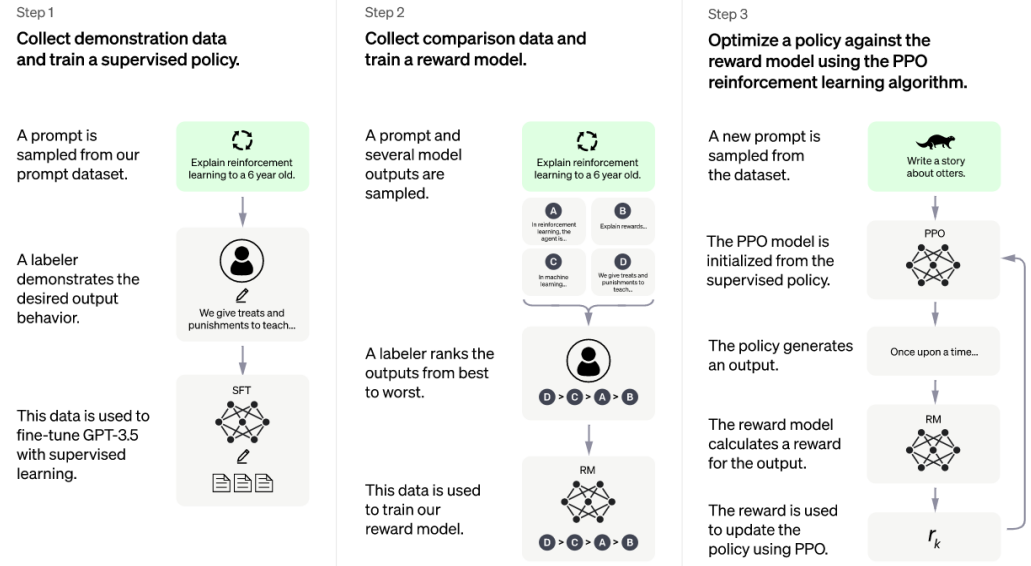
资料来源：OpenAI 官网，华西证券研究所

1.2.不同类别AIGC算法比对

ChatGPT：ChatGPT 基于 GPT-3.5 架构，拥有 175B 个参数。ChatGPT 的训练功能强大的原因就是训练奖励模型数据收集设置略有不同、并加入了强化学习近端策略优化，可以理解成在“人脑思维”的基础上加入了“人类反馈系统”，是一种奖励模型。因此效果更加真实、模型的无害性实现些许提升，编码能力更强。

具体而言：此种强化学习的目的是获得“奖励”，因此 ChatGPT 加入了一个“奖励”模型，每一个问题都生成不同的答案，然后由人类对不同的答案进行排序，排序靠前的回答得分更高，排序较低的回答得分更低。

图表 13 强化学习近端策略优化优化示意图



资料来源：OpenAI 官网，华西证券研究所

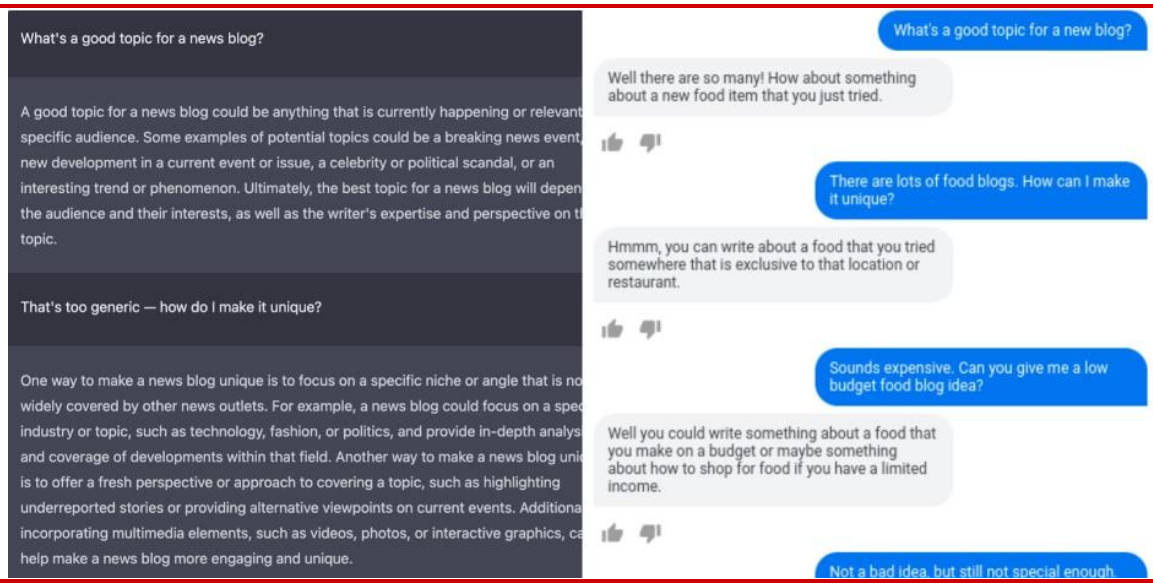
谷歌 LaMDA 是 Google Bard 聊天机器人的程序语言模板：相同点同样是基于 Transformer 的神经语言模型，**不同点，**

1、参数方面由多达 137B 个参数组成，并在 1.56T 的公开可用对话数据和网络文档的单词上进行了预训练。LaMDA 模型具有质量、安全和扎实性三个关键目标，每个目标有各自的衡量指标。

2、奖励模型：LaMDA 的进展是通过收集来自预训练模型、微调模型和人类评分者（即人类生成的反应）对多轮双作者对话的反应来量化的——然后由针对上述定义的指标对一系列问题进行不同的人类评分。 具体行为即对 AI 生成文本进行“点赞”或是“差评”。

3、训练数据：ChatGPT 的训练方式是训练文本，而 LaMDA 的训练方式是训练对话，因此，可以说 GPT-3 专注于生成语言文本，LaMDA 专注于生成对话。

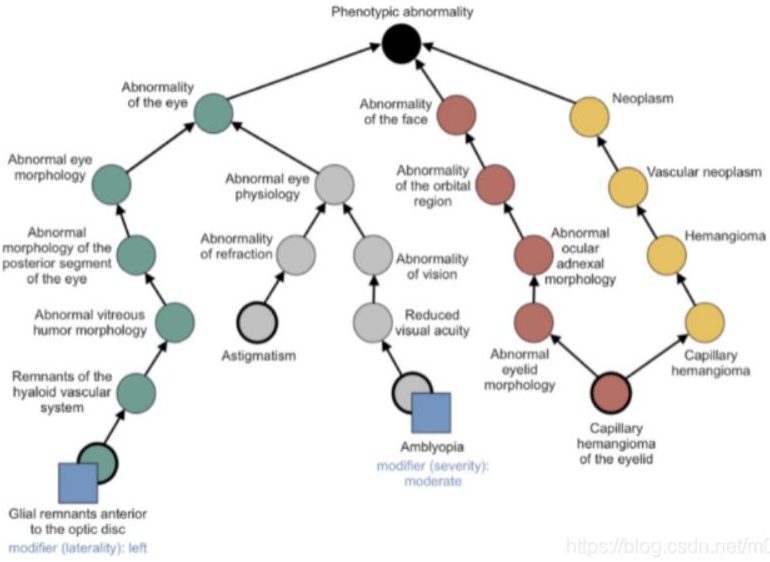
图表 14 ChatGPT 和 LaMDA 的不同(左为 ChatGPT，右为 LaMDA)



资料来源：AI@阅粒，华西证券研究所

此外图神经网络(GNN)作为科学领域预训练模型(大模型)备受瞩目：基本定义，图神经网络（Graph Neural Network，GNN）是指使用神经网络来学习图结构数据，提取和发掘图结构数据中的特征和模式，满足聚类、分类、预测、分割、生成等图学习任务需求的算法总称。**强大之处**，图神经网络相较于普通神经网络最大的特点可以理解成“关系网”，即图神经网络不光可以反映自身的特征，也可以反映邻居结点的特征，换而言之，图结构表示的数据，使得可以进行基于图的解释和推理。

图表 15 图神经网络在电子健康记录建模的应用



资料来源：CSDN，华西证券研究所

图神经网络应用与日俱增，有望成为下一时代的风口浪尖：图形神经网络和相关技术的发展已经具有“脱胎换骨”的意义，例如化学合成、车辆路由、3D 视觉、推荐系统、连续控制、自动驾驶和社交网络分析，目前已经应用在社会结构、