

文化和科技融合热点观察

总第 24 期

2023 年 5 月号

本期热点：迈向通用人工智能，大模型拉开新时代序幕

海外视野

OpenAI	10
Google	19

国内风向

百度	29
阿里	32
华为	35
腾讯	39
北京智谱华章科技有限公司	43
北京澜舟科技有限公司	51
名之梦（上海）科技有限公司	61
超对称（北京）科技有限公司	67
清华大学：深言科技、面壁智能、聆心智能	72



关注“文化科技融合汇”
获取产业发展前沿资讯

协会之窗

南京水晶石数字科技有限公司	81
---------------------	----

产业新势力

开心麻花：全景社交解谜互动城《方橙市大侦探》	90
------------------------------	----

主办单位

中国（南京）文化和科技融合成果展览交易会组委会
紫金文化科技融合创新研究院
南京元宇宙产业协会

支持单位

南京市文化投资控股集团有限责任公司
中国人民大学创意产业技术研究院

出品单位

江苏文创文化发展有限公司

本期导读

通用大模型是在大规模数据集上进行预训练的人工智能模型，具有广泛的应用能力和泛化能力。这些模型经过预训练后，可以通过微调或迁移学习等方式，在各种任务和领域中发挥作用。通用大模型是当今人工智能领域的一项重要技术，本期我们着眼于国内外大模型的发展，探讨国内外企业在通用大模型领域取得的成果。

大模型研发热潮可以追溯到近年来人工智能领域的重要突破和进展。随着深度学习技术的不断发展和计算资源的增强，研究人员开始探索如何构建更大、更复杂的神经网络模型，以提高人工智能系统的性能和能力。其中一个重要的里程碑是 2012 年，由 Hinton 等人提出的 AlexNet 模型在 ImageNet 图像分类挑战赛中取得突出的成绩。这个模型引入了较深的卷积神经网络结构和更大规模的参数，彻底改变了计算机视觉领域的格局。AlexNet 的成功引发了对更大模型的探索和研究，开启了大模型热潮的序幕。

另一个重要的里程碑是 2018 年，由 OpenAI 提出的 GPT（Generative Pre-trained Transformer）模型。GPT 是基于 Transformer 架构的预训练模型，在自然语言处理领域取得了巨大成功。通过在大规模语料库上进行预训练，GPT 模型能够生成高质量的文本和进行语言理解任务。GPT 的成功激发了更多研究者对于预训练大模型的兴趣和研究，不断推动了大模型的发展。越来越多的预训练大模型相继问世，如 BERT、XLNet、RoBERTa 等，它们在自然语言处理、机器翻译、问答系统等任务中取得了显著的性能提升。

本期热点观察，我们通过对比研究、对话访谈等方式，深入探索国内通用大模型领域的龙头企业和创业团队。本期我们将为大家梳理大模型相关的基本概念，并详细介绍国内外的大模型开发团队。在海外视野板块将介绍 OpenAI 和谷歌的主要模型。在国内风向板块，我们首先介绍了四家具有全栈能力的四家大厂：百度、腾讯、华为、阿里；其次，我们重点介绍了国内最具实力的三家科创团队：智谱 AI、澜舟科技、名之梦；此外，我们还关注到清华大学在人工智能领域的突出成就和创业企业的孵化能力，并罗列出面壁智能、深言科技、聆心智能等清华系创业团队。

引言

AIGC，全名“AI generated content”，又称生成式 AI，意为人工智能生成内容。AIGC 是人工智能、大数据、云计算、5G 等多个技术领域的整合，是一种跨领域的合作发展模式。

AIGC 的关键在于模型的创造力和生成能力。AI 大模型通过学习大规模数据的模式和语义，具备了较强的生成能力，可以生成逼真、连贯、具有语义理解的内容。这些模型在生成过程中可以根据给定的输入条件、约束或指导，生成与之相符合的内容。

总的来说，AIGC 和 AI 大模型是相互关联的，AI 大模型为生成内容提供了强大的技术支持和能力，而 AIGC 则是 AI 大模型应用的具体体现和产出。

一、AI 大模型的概念

AI 大模型是“人工智能预训练大模型”的简称，是指使用深度学习方法构建的具有大规模参数和复杂结构的模型。它们通过在大规模数据集上进行预训练，并通过微调或其他方式进行优化，以实现在各种任务上的出色表现。这些大模型通常基于深度神经网络，具有多层的神经网络结构，包含数十亿甚至百亿千亿个可训练的参数。它们通过深度学习算法和反向传播来学习数据中的复杂模式和特征表示。

“AI 大模型”结合了预训练和大模型的概念，具有重要的创新意义。

预训练指模型在大规模数据集上进行初始训练，学习通用的语言和语义知识，从而获得丰富的表示能力。在这个预训练阶段，模型通过无监督学习自动捕捉数据中的模式和结构，并形成一种对输入数据的潜在理解。

大模型指的是具有庞大参数量和复杂结构的神经网络模型。这些模型通常拥有数以亿计的参数，通过多层次的神经网络结构实现对输入数据的多层次抽象和理解。大模型具有更强大的学习能力和表达能力，能够更好地捕捉复杂的数据模式和关联。

结合了预训练和大模型的人工智能模式，使得模型能够在大规模数据集上进行预训练，并在后续的微调阶段甚至无需微调即可适应各种应用。这种模式下，模型具有强大的泛化性、通用性和实用性，能够在多个领域实现突破性的性能提升。例如，通过在自然语言处理、计算机视觉和智能语音等领域应用 AI 大模型，可以实现更准确的文本生成、图像识别和语音识别等任务。

二、AI 大模型的特点

参数数量庞大：AI 大模型通常拥有数以亿计的参数，远远超过传统机器学习模型。这使得它们能够更好地学习和表示复杂的数据模式和关联。

复杂的网络结构：AI 大模型往往采用深层次的神经网络结构，包括多个堆叠的隐藏层。这种深层结构可以帮助模型提取更高级别的抽象特征，从而更好地理解和处理输入数据。

泛化能力：AI 大模型能够从大规模数据中学习通用的特征表示，具有很强的泛化能力，可以适应各种不同的任务和数据。

高性能：由于拥有庞大的模型规模和丰富的参数，AI 大模型能够捕捉到数据中的更多细节和复杂性，从而在许多任务上取得卓越的性能。

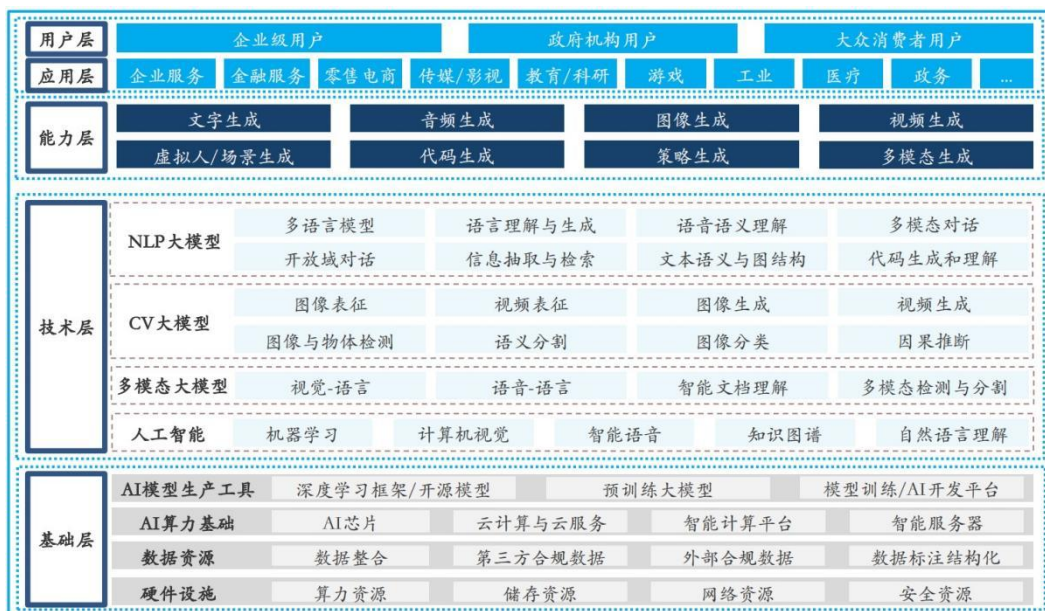
多领域应用：AI 大模型在自然语言处理、计算机视觉、语音识别等领域都取得了显著的成果，并能够解决复杂的问题，如自动翻译、图像分类、语义理解等。

可迁移性：预训练的 AI 大模型可以通过微调或迁移学习的方式应用到新的任务上，从而减少了训练模型所需的数据量和时间，并提供了更快的部署和应用的能力。

自适应性：AI 大模型具有一定的自适应性，能够从新的数据中不断学习和更新模型，以适应不断变化的环境和需求。

三、AI 大模型技术架构

AI 大模型的技术架构通常涉及多个层次可以分为基础层、技术层、能力层、应用层、终端层五大板块，其中核心技术层涵盖 AI 技术群和大模型的融合创新，为各行业深度赋能。



（一）基础层

AI 大模型的基础层涉及硬件基础设施和数据、算力、算法三大核心要素。

随着 AI 大模型规模的不断扩大对计算资源的需求也在增加。因此，高性能的硬件设备、海量场景数据、强大的算力基础和升级迭代的算法模型成为支持 AI 大模型发展的关键。深度学习模型的不断升级和迭代，增强了 AI 算法的学习能力；同时，开源模式将使 AI 大模型成为海量应用、网络和服务的基础。

1、算力

算力指的是计算机硬件资源。包括处理器、内存、存储等。用于处理数据、支持算法的运行。

2、算法

算法指的是人工智能的实现方式。包括机器学习、深度学习、自然语言处理等。用于实现数据分析、训练人工智能模型等。

3、数据

数据指的是训练算法和实现算法所需的信息。包括事实、图像、声音、文本等。用于训练 AI 模型，同时通过数据训练 AI 模型的结果也可以反哺优化提升算法。

总的来说，算力属于基础设施能力，算法是基于基础设施之上运作的工作方法，而数据则是用来指导算法运作的依据。未来随着技术的进步和应用场景的不断拓展，这三个概念将继续发挥重要作用，推动整个行业的创新和发展。

（二）技术层

AI 大模型的技术层主要涉及模型构建。

目前，Transformer 架构在 AI 大模型领域占据主导地位，如 BERT、GPT 系列等。AI 大模型包括 NLP 大模型、CV 大模型、多模态大模型等。这些模型采用预训练和微调的策略，先在大量无标注数据上学习语言或图像的基本表示，然后针对特定任务进行微调。

经过大规模预训练的大模型，能够在各种任务中达到更高的准确性、降低应用的开发门槛、增强模型泛化能力等，是 AI 领域的一项重大进步。大模型最早的关注度源于 NLP 领域，随着多模态能力的演进，CV 领域及多模态通用大模型也逐渐成为市场发展主流。这些大模型通常基于深度学习的神经网络架构，并通过对大规模语料库上进行预训练来学习通用的语言表示和语义理解。

模型	功能	常见案例
自然语言处理模型	处理文本数据，例如文本分类、命名实体识别、情感分析等。	BERT、GPT、Transformer 等。
计算机视觉模型	处理图像和视频数据，例如图像分类、目标检测、图像分割等。	ResNet、YOLO、MaskR-CNN 等。
语音识别模型	将语音转换为文本，用于语音识别和语音转写任务。	DeepSpeech、Wav2Vec 等。
多模态模型	能够同时处理多种不同类型数据，将它们综合起来，得到更全面、准确的分析和结果。	XMC-GAN、SimVLM、ALIGN、PolyVit 等。
跨模态模型	能够有效地在不同模态之间传递和交互信息，它能够理解不同类型的数据，并将它们关联起来，以便更好地理解和分析数据。	Flamingo、Parti、LOLNerf、Phenaki 等。

1、NLP 大模型

NLP（自然语言处理）大模型是指用于处理文本和语言任务的具有大规模参数和计算能力的模型。这些大模型通常基于深度学习的神经网络架构，并通过在大规模语料库上进行预训练来学习通用的语言表示和语义理解。

在 2018 年以来，以 BERT 和 GPT 为代表的语言大模型，弥补了自然语言处理标注数据的缺点，促进了 NLP 技术的发展。从技术的角度，这些大模型的训练，透过事先遮住一些文本片段，让 AI 模型通过自监督学习，通过海量语料库的预训练，逐步掌握上下文语境，把这些被遮住的片段，尽可能合乎逻辑的方式填上去

从 2019 年的谷歌 T5 到 OpenAIGPT 系列，参数量爆炸的大模型不断涌现，尤其在 2022 年 11 月底 ChatGPT 的出现，更是引起了社会各界的广泛关注。3 月以来，国内的百度、阿里等公司，也陆续发布了自己的大模型。

2、CV 大模型

CV（计算机视觉）大模型是指用于处理图像和视觉任务的具有大规模参数和计算能力的模型。这些大模型通常基于深度学习的卷积神经网络（CNN）架构，并通过在大规模图像数据集上进行训练来学习图像的特征表示和语义理解。

相比于 NLP 大模型，目前 CV 大模型的发展，还是存在一些挑战，主要包括四点：

- （1）在 CV 领域可用于训练的有效数据，比 NLP 领域有不少差距。
- （2）CV 领域的学习方法还需要突破。
- （3）不同的视觉应用仍需要依赖于不同的模型，如何建立通用的视觉模型还是未知。
- （4）供训练的图像尺寸越来越大，较小的模型也可能有很大的计算量。

所以，综合来看，CV 大模型短期还很难出现大的爆发。现阶段比较有名的 CV 大模型包括微软的 swin-transformer 系列，以及谷歌的 ViT 系列，Meta 也在近期发布了 SAM 模型，可以识别图像和视频中物体，完成 CV 底层技术突破。国内方面，代表性的是：商汤科技在 4 月初发布的日日新大模型。

3、多模态大模型

多模态大模型是指能够同时处理多种模态（如图像、文本、音频等）数据的具有大规模参数和计算能力的模型。这些模型能够从不同模态的数据中提取特征，并进行跨模态的联合建模和分析，以获取更全面、准确的信息和理解。

多模态大模型的设计和训练通常基于深度学习的方法。模型结构中包含多个分支或子模型，每个分支专门处理一个模态的数据。这些分支在训练过程中可以独立地学习每个模态的特征表示，同时通过共享参数来促进跨模态的交互和融合。通过这种方式，多模态大模型可以在不同模态的数据之间建立关联，并从中提取更丰富的信息和语义。

多模态大模型的应用非常广泛。例如，在图像和文本的多模态场景下，可以使用多模态大模型来进行图像标注、视觉问答、图像生成等任务。在语音和文本的多模态场景下，可以将多模态大模型应用于语音识别、语音生成、语音翻译等任务。多模态大模型的优势在于能够充分利用不同模态之间的互补信息，提供更全面、准确的分析和预测。

尽管多模态大模型面临着挑战，如数据集的融合、模态间的对齐、计算资源的需求等，但它们代表了人工智能领域的前沿研究方向，有望进一步提升多模态数据的处理能力和应用领域的拓展。

（三）能力层、应用层及用户层

在基础层和技术层的支持下，AI 大模型拥有了文字音频、图像、视频、代码、策略、多模态生成能力等，具体应用于金融、电商传媒、教育、游戏、医疗、工业、政务等多个领域，为企业级用户、政府机构用户、大众消费者用户提供产品和服务。

同时，AI 大模型在各个领域都有广泛的应用，以下是一些成功的案例：

领域	案例
自然语言处理	BERT 模型在语言理解任务中表现出色，它能够预训练从而学习到丰富的语义表示，用于文本分类、命名实体识别、句子关系判断等任务。
	GPT 模型，它可以生成连贯的文本，用于自动写作、对话系统等。
计算机视觉	ResNet 是一个深度卷积神经网络，能够有效地解决图像分类、目标检测和图像分割等任务。
	GAN 模型，它可以生成逼真的图像，用于图像合成、风格迁移等。
语音识别	DeepSpeech 是一个基于深度学习的语音识别模型，能够准确地将语音转换为文本。这种技术在语音助手、语音翻译和语音指令识别等应用中得到广泛应用。
推荐系统	YouTube 使用 AI 大模型来推荐用户感兴趣的视频。
	Netflix 利用 AI 大模型为用户推荐适合的电影和电视节目。这些模型通过分析用户的观看历史和行为模式，提供个性化的推荐结果。
医疗诊断	基于 AI 大模型的深度学习算法可以在医学图像中进行病变检测和诊断，辅助医生进行准确的诊断。这种技术在放射影像学、病理学和眼科等领域具有很高的潜力。

除了以上领域，AI 大模型还被应用于金融风控、交通管理、智能客服、虚拟现实等各个行业和领域。随着技术的不断发展和优化，预计 AI 大模型将在更多领域中实现突破和应用。

四、AI 大模型的发展趋势

（一）“通用大模型”“行业小样本”是未来主流 AI 发展范式

AI 大模型是一种新型平台模式，是移动互联网向沉浸互联网发展的中间形态。从平台的逻辑出发，AI 产业链包括上层大模型、中层小模型、下层应用及内容，下层应用及内容包括直接接入大模型或通过小模型接入大模型两种方式。

头部 AIGC 应用产品属于“小模型”+“应用”的技术范式，本身具备较高质量的 AI 能力，若接入匹配的多模态大模型，有望实现能力上的质变突破。一方面上层大模型持续为小模型、应用及内容赋予通用型能力，其兼容性取决于大模型的输出与小模型应用及内容的输入在模态上是否匹配；另一方面小模型、应用及内容持续产生的大数据可不断迭代大模型的通用能力，并让大模型更为适配垂类场景。

1、“行业小样本”的数据集来自小模型或应用及内容：

AI 产业链包括上层大模型、中层小模型、下层应用及内容，包括应用及内容直接接入大模型或通过小模型接入大模型两种方式，即“大模型+应用及内容”或“大模型+小模型+应用或内容”其中具备特定功能的 AIGC 软件产品我们理解为“小模型”+“应用”的技术范式，本身具备较高质量的 AI 能力，若接入匹配的多模态大模型，有望实现能力上的质变突破。

2、“行业小样本”的结合方式包括“能力调用”及“能力训练”两层：

（1）能力调用

是指下游垂类场景直接调用通用大模型的通用能力，并基于垂类场景内产生的特性化数据不断提升调用能力在垂类场景内的适配程度。我们认为现阶段下游应用及内容主要采取此类方式接入大模型能力，此类方式可高效快速调用大模型先进能力，在时间上及成本上具备优势。

（2）能力训练

是指下游垂类场景将通用大模型针对特性化数据集进行再训练，从而形成垂类场景专属大模型。例如彭博让利用自身丰富的金融数据源，基于开源的 GPT-3 框架再训练，开发出了金融专属大模型 BloombergGPT。我们认为未来拥有丰富特性化数据集的下游垂类场景将主要采取此类方式，有助于构筑更强的能力壁垒。

（二）MaaS（Model as a Service）将重构 AI 大模型商业化生态

MaaS 能够降低 AI 应用开发门槛，重构 AI 产业的商业化结构生态，激发新的产业链分工和商业模式不断涌现。MaaS 将可能成为未来大模型的主流商业模式。

1、概念

随着大模型的功能越来越强大，AI 领域逐渐形成一种独特的新商业模式，叫做模型即服务，即 MaaS。

简单来说，MaaS 是一种新的商业模式。作为 AI 时代的基础设施，MaaS 为下游应用提供安全、高效、低成本的模型使用与开发支持。用户可以直接在云端调用、开发与部署模型，而无需投资构建和维护自己的模型所需的基础设施、硬件和专业知识。大模型作为 MaaS 的重要基座，是“大算力+强算法”结合的产物，将成为人工智能未来的发展趋势。大模型以迅速发展态势，实现 AI 从“手工作坊”到“工厂模式”的转变，达到更强的通用性和智能程度，使 AI 更广泛地赋能各行业应用

2、架构

MaaS 模式将由基础层、中间层和应用层三部分组成，这是一种将 AI 模型以服务的形式提供给用户的架构。

（1）基础层

基础层是 MaaS 模式的底层基础设施，包括大规模的计算资源、存储设备和网络基础设施等。这些基础设施为模型的训练和推理提供必要的计算和存储支持。

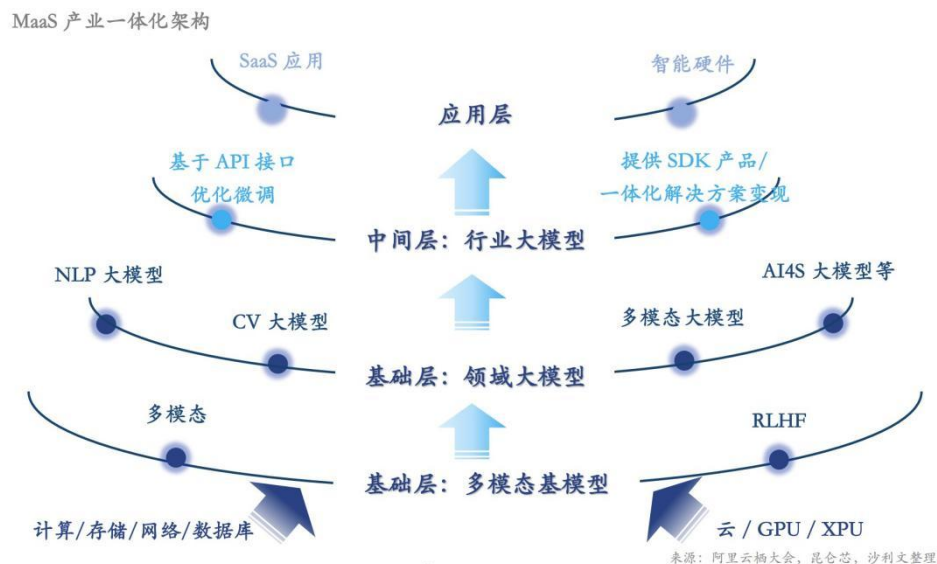
（2）中间层

中间层是 MaaS 模式的核心组成部分，负责管理和部署 AI 模型。它包括模型的训练、优化、部署和监控等功能。中间层提供了一系列的工具和接口，使用户能够方便地使用和管理 AI 模型。

（3）应用层

应用层是 MaaS 模式的最上层，为用户提供基于 AI 模型的具体应用和服务。这可以包括自然语言处理、图像识别、推荐系统等各种应用领域。应用层通过调用中间层提供的接口和服务，实现了将 AI 模型集成到具体应用中的功能。

通过 MaaS 模式，用户无需自己搭建和管理庞大的 AI 模型基础设施，而是可以通过云服务提供商或专门的 AI 平台来获取所需的 AI 模型能力。这种模式能够降低用户的技术门槛和成本，并使得 AI 模型的应用更加灵活和可扩展。同时，MaaS 模式也为模型提供商提供了更好的商业机会，可以将其模型以服务的形式销售给广大用户。



3、商业化

MaaS 模式在 B/G 端和 C 端的商业化落地有所区别。

（1）C 端

C 端用户量巨大，工具使用门槛较低落地以及未来成长速度更快，“应用商店”等创新商业模式不断涌现。通过“对话+插件”形式将大模型单点工具接入厂商自有或第三方应用，将打造新的用户交互界面和入口。目前 Chat-4 已应用于微软必应搜索引擎，ChatGPT 接入第三方网上购物及机票预订等平台、阿里巴巴“通义千问”大模型将连接旗下所有产品。插件形式将覆盖用户生活的方方面面，打造完整的生态系统。而随着 Adept 等无需 App 交互，自动执行操作响应用户需

求的产品出现，未来 MaaS 模式下 C 端商业化落地形式将不断创新，颠覆传统，具有广阔的想象空间。

未来 C 端市场，MaaS 落地的主流商业模式为软件订阅费用，以及第三方 App 的推广和订阅分成费用。基础层大模型厂商可以基于大模型推出类似 ChatGPT 的单点工具，以月度或年度订阅费提供产品功能使用，为用户提供灵活和便捷的购买方式。未来盈利模式也将可能向 AppStore 式靠拢，通过排行榜、广告位等收取费用。

（2）B/G 端

B/G 端，需要针对行业领域和业务场景进行大量工程工作，尤其是传统行业的知识获取和积累需要较长时间，即使是同一个行业下，细分场景的痛点不同，AI 大模型渗透率也有明显差异应用场景碎片化的特点导致低成本、易用、泛化能力较强的能力平台构建需较长周期，但 B/G 端客户付费能力更强，未来盈利空间及成长空间广阔据分析，从美国市场看，目前 ToB 应用的数量大于 ToC 应用，通用工具数量大于具体场景应用数量主要集中于市场销售、客服/CRM/CEM、企业内部生产力工具等。

未来 B/G 端市场，MaaS 落地的主流商业模式将按照数据请求量和实际计算量计算通过对外开放大模型的 API 调用接口，让开发者灵活地使用基模型服务，典型案例是 GPT 基于对外 API 的收费模型。基于此，垂直行业厂商可以提供 SDK 产品或一体化落地解决方案变现。大模型厂商也可以通过推出 TOB/TOG 的单点工具，按文本、图像或语音等不同形式的内容量收费，如 DALL·E 可以根据每张图片的分辨率和请求计算量计费。

海外视野

OpenAI

一、公司发展

OpenAI 是一家人工智能研究实验室和公司，成立于 2015 年，总部位于美国加利福尼亚州的圣弗朗西斯科。OpenAI 的使命是确保人工智能的好处可以普惠全人类，并致力于推动人工智能技术的发展和应用。

OpenAI 的研究和开发涵盖了广泛的领域，包括机器学习、深度学习、自然语言处理、计算机视觉等。公司致力于开发具有强大智能能力的系统，并推动人工智能技术的发展。OpenAI 在人工智能领域的研究成果和项目备受瞩目，其开发的大型预训练语言模型 GPT 系列在自然语言处理和文本生成方面取得了重大突破。

从时间维度上看，OpenAI 的研究方向大致聚焦在三个领域：强化学习领域、自然语言处理领域、多模态模型领域。在强化学习领域，OpenAI 采用通用的强化学习策略来帮助 AI 应对灵活多变的困难问题；在自然语言处理领域，OpenAI 开辟自回归建模路径，大语言模型无需在特定领域微调也有出色的任务表现；在多模态模型领域，OpenAI 围绕 Transformer 主架构不断增强多模态能力。OpenAI 的研究方向始终关注问题的通用性解决方案，而不是聚焦在特定问题之上，与其通用人工智能的研究目标契合。

二、主要模型

AIGC 的发展离不开大模型底层技术的支撑，以下表格所示便是目前 OpenAI 开放的主要模型，此外还有包括 Point-E、Whisper、Jukebox，和 CLIP 在内的开源模型。

模型	发布时间	描述
GPT 系列	2018 年—2023 年	一组可以理解并生成自然语言或代码的模型
DALL·E 系列	2021 年—2022 年	一种可以在给定自然语言提示情况下生成和编辑图像的模型
WhisperBeta	2022 年 9 月 21 日	一种可以将音频转换为文本的模型
Codex (Deprecated)	2021 年 8 月 10 日	一组可以理解和生成代码的模型，包括将自然语言翻译成代码
Moderation	/	一种可以检测文本是否敏感或不安全的微调模型
Embeddings	2022 年 12 月 15 日	一组可以将文本转换为数字形式的模型

当前，随着大语言模型的不断成熟发展，全球多家科技智能公司纷纷推出系

列模型，同时 ChatGPT 版本持续更新受到广泛关注。以下我们将详细介绍 ChatGPT 背后的系列模型及发展过程。

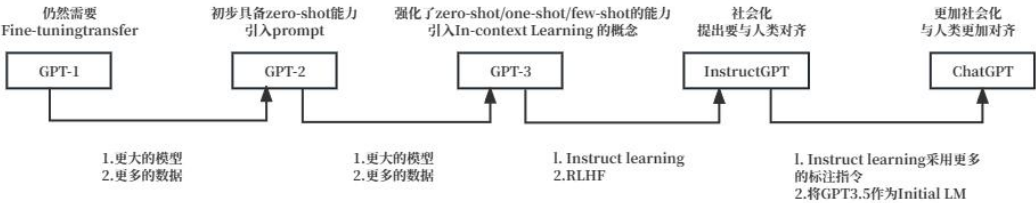
（一）GPT 系列模型

GPT 系列模型包括 GPT-1，GPT-2，GPT-3，GPT-3.5 以及 GPT-4。该系列模型均采用了生成式预训练 Transform 模型¹。GPT 系列模型结构秉承了不断堆叠 Transformer 的思想，通过不断提升训练语料的规模、网络的参数数量来完成 GPT 系列的迭代更新。

GPT，生成式预训练模型，全称 Generative pre-trained transformer model，是一种延伸自转换器架构（Transformer）的自然语言生成模型。

其名称中的“生成式”指的是它可以进行微调以完成各种自然语言处理任务，例如文本生成、代码生成、视频生成、文本问答、图像生成、论文写作、影视创作、科学实验设计等。

其名称中的“预训练”指的是在大型文本语料库上进行的初始训练过程，其中模型学习预测文章中下一个单词，这为模型在具有有限的任务特定数据的下游任务中表现良好提供了坚实的基础。



1、GPT-1、GPT-2、GPT-3

下图是 GPT-1、GPT-2、GPT-3 模型的简单介绍，它们属于无监督的预训练模型，可以发现模型参数量越来越大，能力越来越强。

模型	发布时间	参数量	训练方式	训练数据来源	预训练数据量
GPT-1	2018.6	1.2 亿	Pre-training+Fine-tuning	书籍	4.5GB
GPT-2	2019.2	15 亿	Pre-training	网页	40GB
GPT-3	2020.5	1750 亿	Pre-training	网页、书籍	570GB

（1）GPT-1

GPT-1 模型训练主要包括两个阶段，无监督预训练阶段和有监督微调阶段。

¹Transform 模型，具体介绍详见附录“生成算法”部分。

对于下游任务的训练，GPT-1 往往只需要简单的微调便能取得非常好的效果。但在未经微调的任务上虽然也有一定效果，但是其泛化能力远远低于经过微调的有监督任务，说明了 GPT-1 只是一个简单的领域专家，而非通用的语言学家。

(2) GPT-2

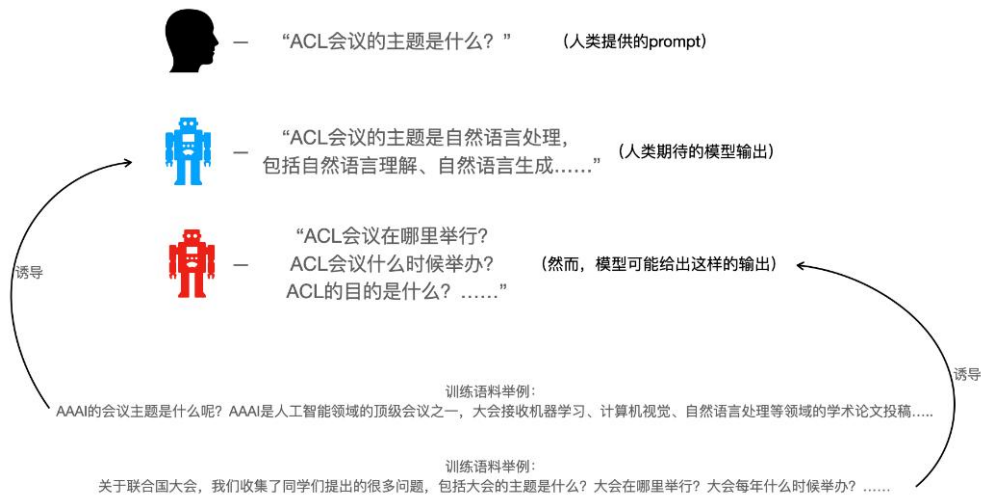
GPT-2 使用更大的模型去学习，还提出了 zero-shot，零样本学习，即将预训练好的模型，直接接入诸多的下游任务，在多个任务下都可以取得很好的效果。GPT-2 模型针对 GPT-1 存在的模型不具有通用性的问题提供了解决方案，核心就是提升模型的容量和数据多样性，让语言模型能够达到解决任何任务的程度。

(3) GPT-3

GPT-3 模型的结构与 GPT-2 的方法相似，在其基础上增大了模型尺寸，进一步增强 GPT-2 的性能。GPT-3 主要分析了各种情境学习能力的差异，可以在许多 NLP 的数据集上都有很好的表现，在各种场景中，可以写文章、写故事，还可以进行多轮对话、写代码、做表格、生成图表等，具有非常可观的商业价值。

2、InstructGPT

在 GPT-3 之后，继续进化出了 InstructGPT，这是为解决 LM（语言模型）有时不能遵循用户意图²而诞生的。主要原因是 LM 的训练目标与其用户没有对齐，LM 并没有对输出是否无用、是否有毒的监督。因此，InstructGPT 要做的就是使模型的输出符合人类的意图。下图所示便是 InstructGPT 社会化、与人类对齐的效果案例。



总的来说，InstructGPT 的优化目标有以下三个（3H）：

(1) Helpful

模型应该帮助用户解决他们的任务。

(2) Honest

²不能遵循用户意图，表示语言模型可能会生成不真实、有毒或对用户毫无帮助的输出。

模型不应该编造信息来误导用户。

(3) Harmless

模型不应对人或环境造成身体、心理或社会伤害。

3、ChatGPT

ChatGPT 是基于 GPT 模型进一步发展而来的人工智能聊天机器人。

2022 年 11 月 30 日 OpenAI 通过 GPT-3.5 系列大型语音模型微调而成的，全新对话式 AI 模型 ChatGPT 正式发布。

2023 年 3 月 15 日 OpenAI 震撼推出了大型多模态模型 GPT-4，不仅能够阅读文字，还能识别图像，并生成文本结果，现已接入 ChatGPT 向 Plus 用户开放。

模型	类型	功能	文字输入长度	考试能力	API 价格
GPT-3.5	自然语言处理模型	文字问答 剧本写作	3000 字	司法考试排名后 10% SAT 数学考试 590 分 生物奥林匹克竞赛排名前 69%	0.002 美元 /1000token
GPT-4	多模态模型	看图作画 文字问答 数据推理 分析图表 角色扮演	25000 字	司法考试排名前 10% SAT 数学考试 700 分 生物奥林匹克竞赛排名前 1%	请求 0.06 /1000token 回复 0.12 /1000token

总的来说，GPT3.5 是一种自然语言处理模型，只能听或看懂语言。而 GPT-4 是一种多模态模型，可以处理多种媒体数据，并且将他们整合到统一的语义空间之中。具体差异如下图所示。

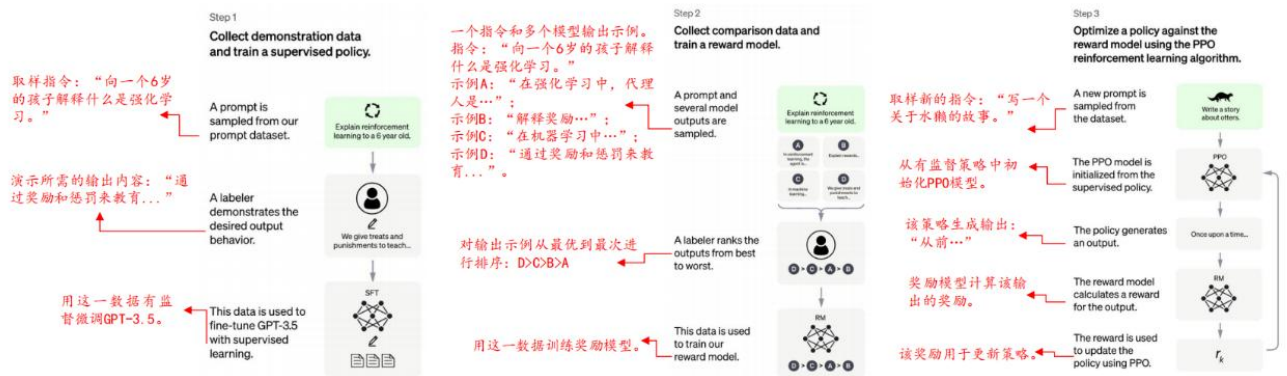
(1) 技术原理

ChatGPT 的训练过程分为三步：

- 1.收集演示数据，使用有监督学习微调 GPT 模型，训练得到有监督策略。
- 2.收集对比数据，并训练奖励模型。
- 3.基于奖励模型，使用 PPO³强化学习算法优化策略。

下图是 ChatGPT 的详细训练过程：

³ PPO 是一种深度强化学习算法，用于训练智能体（agent）在复杂环境中学习和执行任务。



(2) 迭代更新——ChatGPT4

首先，作为多模态大模型，ChatGPT4除了文本能力的增强以外，最大的功能就是可以接受图像作为输入并生成说明、分类和分析结果。具体来说，它能在用户输入散布式文本与图像后，自主生成文本输出（自然语言、代码等）。在包括带有文本和照片的文档、图表与屏幕截图中，ChatGPT4展示了与纯文本输入同样强大的功能。此外，它还可以通过为纯文本语言模型开发的测试时间技术实现进化，能运用少量镜头和思维链提示。

1. 文本理解和图像识别并行

ChatGPT4可以接受文本和图像的提示，允许用户指定任何视觉或语言任务。在一系列域（包括包含文本和片的文档、图表或屏幕截图）上，ChatGPT4表现出与纯文本输入类似的功能。

用户 这张图片有什么不寻常之处？

资料来源：巴诺拉马

GPT-4 这张照片的不同寻常之处在于，一名男子正在行驶中的出租车车顶上的熨衣板上熨烫衣服。

2. 辨别幽默元素

ChatGPT4 还能辨别图像中存在的幽默元素，并做出相关解释。



3. 直接阅读并分析带有图片的论文

给几张论文的照片，ChatGPT4 可以做总结，也可以对用户指定的图片的内容进行展开解释。



其次，是语言能力，在语言测试中 ChatGPT4 完胜其他大模型。

1. 英语测试

ChatGPT4 的准确率为 85.5%，高出第二名 15pct，取得碾压优势除英语外的 26 种语言测试中，ChatGPT4 在 24 种的准确率获胜，包括拉脱维亚语威尔士语、斯瓦希里语等资源匮乏的语言，准确率集中在 70%—85%，虽与其他模型相比优势突出，但未来仍存在较大提升空间。

2.文字输入限制大幅提升

由基于 ChatGPT3.5 的 ChatGPT 的 3000 字限制提升至 ChatGPT4 的 25000 字，用户使用将开拓更多使用场景。

最后，是推理能力，ChatGPT4 已表现出人类水平的性能，较 ChatGPT3.5 更进一步。

1.学术和专业基准测试

在与其他模型的对比中，ChatGPT4 除了在阅读理解和算术方面与最佳外部模型略有差距，其他的学术测试表现均获胜，基于日常的常规推理和小学多选科学题的准确度更是达到了 95%以上，较外部模型高 10pct 以上。

		GPT-4	GPT-3.5	最佳外部语言模型	最佳外部模型
多项选择题 (57个学科)	★	86.4%	70.0%	70.7%	75.2%
基于日常的常规推理	★	95.3%	85.5%	84.2%	85.6%
小学多选科学题	★	96.3%	85.2%	85.2%	86.5%
围绕代词解析的常识推理	★	87.5%	81.6%	85.1%	85.1%
Python代码任务	★	67.0%	48.1%	26.2%	65.8%
阅读理解和算术		80.9%	64.1%	70.8%	★ 88.4%
小学数学题	★	92.0%	57.1%	58.8%	87.3%

2.应试方面展现“恐怖实力”

ChatGPT4 在模拟律师资格考试中进入前 10%，SAT 的读写和数学得分均在 700 分以上，GRE 接近满分，在 GLUE 语言理解基准测试上取得了 90.1 的平均得分，某些性能已经与人类持平甚至超越。

Exam	GPT-4	GPT-4 (no vision)	GPT-3.5
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)	298 / 400 (~90th)	213 / 400 (~10th)
LSAT	163 (~88th)	161 (~83rd)	149 (~40th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)	710 / 800 (~93rd)	670 / 800 (~87th)
SAT Math	700 / 800 (~89th)	690 / 800 (~89th)	590 / 800 (~70th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)	157 / 170 (~62nd)	147 / 170 (~25th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)	165 / 170 (~96th)	154 / 170 (~63rd)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)	4 / 6 (~54th)	4 / 6 (~54th)
USABO Semifinal Exam 2020	87 / 150 (99th - 100th)	87 / 150 (99th - 100th)	43 / 150 (31st - 33rd)
USNCO Local Section Exam 2022	36 / 60	38 / 60	24 / 60
Medical Knowledge Self-Assessment Program	75 %	75 %	53 %
Codeforces Rating	392 (below 5th)	392 (below 5th)	260 (below 5th)
AP Art History	5 (86th - 100th)	5 (86th - 100th)	5 (86th - 100th)
AP Biology	5 (85th - 100th)	5 (85th - 100th)	4 (62nd - 85th)
AP Calculus BC	4 (43rd - 59th)	4 (43rd - 59th)	1 (0th - 7th)
AP Chemistry	4 (71st - 88th)	4 (71st - 88th)	2 (22nd - 46th)
AP English Language and Composition	2 (14th - 44th)	2 (14th - 44th)	2 (14th - 44th)
AP English Literature and Composition	2 (8th - 22nd)	2 (8th - 22nd)	2 (8th - 22nd)
AP Environmental Science	5 (91st - 100th)	5 (91st - 100th)	5 (91st - 100th)
AP Macroeconomics	5 (84th - 100th)	5 (84th - 100th)	2 (33rd - 48th)
AP Microeconomics	5 (82nd - 100th)	4 (60th - 82nd)	4 (60th - 82nd)
AP Physics 2	4 (66th - 84th)	4 (66th - 84th)	3 (30th - 66th)
AP Psychology	5 (83rd - 100th)	5 (83rd - 100th)	5 (83rd - 100th)
AP Statistics	5 (85th - 100th)	5 (85th - 100th)	3 (40th - 63rd)
AP US Government	5 (88th - 100th)	5 (88th - 100th)	4 (77th - 88th)
AP US History	5 (89th - 100th)	4 (74th - 89th)	4 (74th - 89th)
AP World History	4 (65th - 87th)	4 (65th - 87th)	4 (65th - 87th)
AMC 10	30 / 150 (6th - 12th)	36 / 150 (10th - 19th)	36 / 150 (10th - 19th)
AMC 12	60 / 150 (45th - 66th)	48 / 150 (19th - 40th)	30 / 150 (4th - 8th)
Introductory Sommelier (theory knowledge)	92 %	92 %	80 %
Certified Sommelier (theory knowledge)	86 %	86 %	58 %
Advanced Sommelier (theory knowledge)	77 %	77 %	46 %
Leetcode (easy)	31 / 41	31 / 41	12 / 41
Leetcode (medium)	21 / 80	21 / 80	8 / 80
Leetcode (hard)	3 / 45	3 / 45	0 / 45

(3) 应用

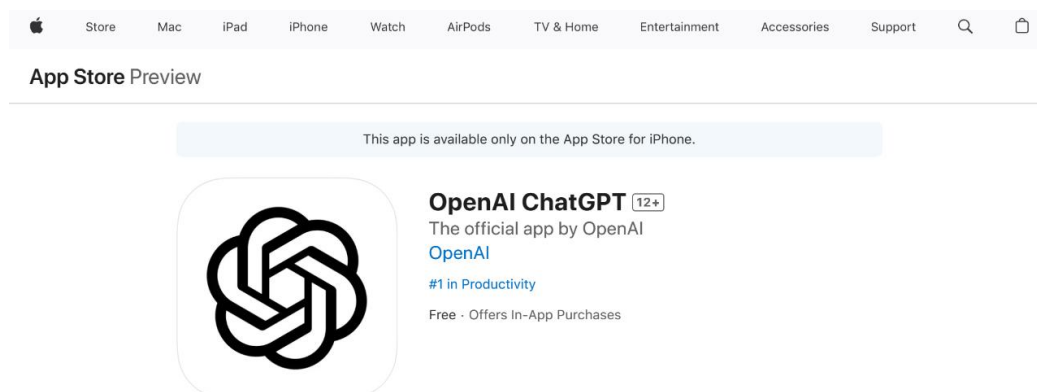
ChatGPT 的应用场景多元，ChatGPT 潜在的市场规模是空前的，能够覆盖各个领域。包括自然语言处理、对话系统等六个方面。借助其强大的算力和算法分析 ChatGPT 有望在教育、科研、游戏、新闻等多个领域发挥作用，并不断拓展其市场潜力。ChatGPT 应用举例如下图所示。



(4) 最新动态

2023 年 5 月 20 日，OpenAI 正式推出了适用于苹果 iOS 系统的 ChatGPT 应用程序。该应用程序免费使用（需要绑定 ChatGPT 账号），并且可以跨设备实现同步历史记录。此外还集成了 OpenAI 的开源语音识别系统 Whisper，支持语音输入。

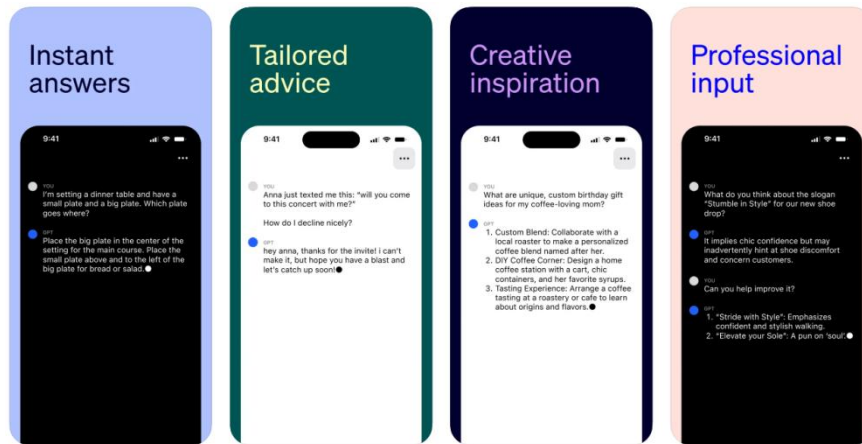
不过目前，ChatGPT iOS 应用程序只在美国上线，未来几周将在其他国家或地区推出。OpenAI 表示，安卓版 ChatGPT App 也将在不久之后到来。OpenAI 将收集更多用户反馈，并持续改进 ChatGPT 的功能和安全性。



具体来说，ChatGPT App 提供的功能包括：

- 1.即时回答：获得精确的信息，而无需通过广告或多个结果进行筛选。
- 2.量身定制的建议：寻求关于烹饪、旅行计划或制作贴心信息的指导。

- 3.创造性灵感：生成礼物创意，起草演示文稿，或写出完美的诗句。
- 4.专业意见：通过想法反馈、笔记总结和技术主题协助来提高生产力。
- 5.学习机会：以你自己的节奏探索新的语言、现代历史和更多。



可以说，通过构建适用于 iOS 系统的 ChatGPT 应用程序，OpenAI 将最先进的研究转化为有用、便捷的工具，让人们更容易获得的同时为他们持续赋能。

Google

一、公司发展

谷歌是一家全球领先的科技公司，成立于 1998 年，总部位于美国加利福尼亚州的山景城。谷歌以其强大的搜索引擎技术而闻名，提供广泛的互联网相关产品和服务。

其核心产品是谷歌搜索引擎，为用户提供准确、全面和实时的搜索结果。除了搜索引擎，谷歌还开发和提供许多其他产品和服务，包括在线广告平台 (Google Ads)、电子邮件服务 (Gmail)、地图导航 (Google Maps)、云计算服务 (Google Cloud Platform)、智能助理 (Google Assistant) 等。

此外，谷歌还在人工智能和机器学习领域进行广泛的研究和开发。谷歌的人工智能项目包括谷歌大脑 (Google Brain) 团队，致力于推动深度学习和神经网络的发展，并开发了一系列重要的大模型和算法，如 BERT 和 Transformer 等。

作为一家全球科技巨头，谷歌在全球范围内拥有广泛的业务和影响力。它的产品和服务被广泛应用于个人用户、企业客户和开发者社区。谷歌的愿景是通过技术创新来改善人们的生活，并致力于推动科技的进步和社会的发展。

二、主要模型

谷歌在大模型领域有多项重要的成果，下表是关于部分模型的简单介绍。

模型	发布时间	描述
Inception	2014 年	一个用于图像分类和识别的深度卷积神经网络
Seq2Seq	2015 年	一个在机器翻译等序列到序列任务上取得了突破性的性能模型
Transformer	2017 年	一个基于自注意力机制的神经网络结构，用于自然语言处理任务的模型
BERT	2018 年	一个通过在大规模文本数据上进行预训练的语言模型
T5	2019 年	一个通用的文本到文本转换模型
MUM	2020 年	一个能够处理文本、图像和语音等多种输入的多模态大模型
LaMDA	2022 年	一个会话式大型自然语言模型
PaLM2	2023 年	一个具有改进的多语言、推理和编码功能的语言模型。

（一）Bert

（1）概念

Bert，于 2018 年提出，全称 Bidirectional Encoder Representations from Transformers，是一种基于 Transformer 架构的预训练语言模型。

Bert 模型的设计主要解决了传统语言模型在处理自然语言时所面临的两个主要问题：上下文理解和双向性。传统的语言模型在处理文本时，通常是从左到右或从右到左进行阅读，这导致了对上下文信息的不完全捕捉。而 Bert 模型引入了 Transformer 的自注意机制，能够同时考虑文本中的左右上下文，从而更好地理解整个句子的语义。

（2）训练过程

Bert 模型的训练过程主要分为两个阶段：预训练和微调。

在预训练阶段，模型使用大规模的未标注文本数据进行训练，通过学习句子级别和单词级别的任务来学习通用的语言表示。具体而言，Bert 模型通过遮盖输入文本中的一些单词，让模型预测被遮盖的单词，以此来学习上下文信息。此外，Bert 还引入了一个额外的任务，即预测句子间的关系，以帮助模型学习句子级别的表示。

在微调阶段，Bert 模型使用有标注的任务特定数据集进行进一步的训练。通过微调，Bert 模型可以根据具体的任务进行适应性调整，如文本分类、命名实体识别、问答系统等。在微调阶段，Bert 模型通常会在任务特定的数据集上进行 fine-tuning，通过在特定任务上的训练，使得模型可以具体地理解和解决该任务。

（3）优点

Bert 模型的优点在于它能够从大规模未标注的数据中学习通用的语言表示，从而在多个自然语言处理任务上都能取得很好的性能。另外，Bert 模型的双向性质使得它能够更好地理解 and 捕捉上下文信息，从而提升了在语义理解和生成任务中的表现。

总的来说，Bert 模型通过引入 Transformer 架构和预训练—微调的方式，在自然语言处理领域取得了重大突破，成为当前最为流行和强大的语言模型之一。

（二）T5

（1）概念

T5，于 2019 年提出，全称 Text-to-Text Transfer Transformer，是一种基于 Transformer 架构的大型预训练语言模型，被广泛应用于自然语言处理（NLP）领域。T5 模型以 Text-to-Text 的方式统一了各种 NLP 任务的表达方式，通过将不同任务转化为文本到文本的转换问题，实现了多任务学习和迁移学习的能力。

在 T5 模型提出之前，NLP 领域的主流模式是先在大量无标签的数据上预训练一个语言模型，然后再在下游具体任务上进行有监督的 **fine-tune**，以此取得还不错的效果。在这个模式下，各大公司争奇斗艳，各个机构奋起直追，今天 A 提出了一种预训练目标函数；明天 B 提出了另一种预训练目标函数，并且收集了更多的训练语料；后天 C 提出了参数更庞大的模型，采用了不同的优化策略；很难分析比较这些工作的有效贡献量。

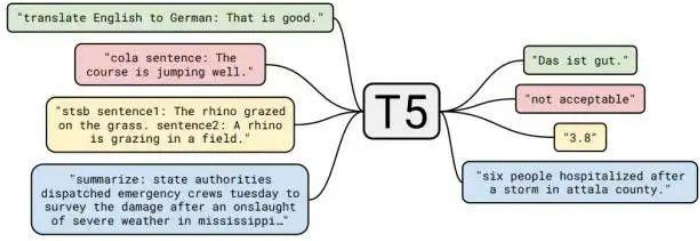


Figure 1: A diagram of our text-to-text framework. Every task we consider—including translation, question answering, and classification—is cast as feeding our model text as input and **training it to generate some target text**. This allows us to **use the same model, loss function, hyperparameters, etc. across our diverse set of tasks**. It also provides a standard testbed for the methods included in our empirical survey. “T5” refers to our model, which we dub the “Text-to-Text Transfer Transformer”.

因此，T5 希望提出一个统一的模型框架，将各种 NLP 任务都视为 **Text-to-Text** 任务，也就是输入为 **Text**，输出也为 **Text** 的任务。由此可以方便地评估在阅读理解、摘要生成、文本分类等一系列 NLP 任务上，不同的模型结构，预训练目标函数，无标签数据集等的影响。

（2）训练过程

T5 模型的训练过程主要分为两个阶段：预训练和微调。

在预训练阶段，有自我监督预训练和多任务预训练两种训练方法。

1. 自监督预训练。通过对预训练任务的细致搜索，最终 T5 模型采用了类似于 BART 模型的文本填充任务进行预训练，如下表所示。与 BART 模型稍有不同，这里对不同位置的文本片段使用不同的掩码标记；同时，在目标端不对原始句子进行完全重构，而是重构丢弃的文本片段，并通过掩码标记指示恢复片段的位置信息。

原文本	Thank you for inviting me to your party last week .
输入序列（随机丢弃 15% 的标记）	Thank you <X> me to your party <Y> week .
目标序列（重构丢弃的标记片段）	<X> for inviting <Y> last <Z>

2. 多任务预训练。除了使用大规模数据进行无监督预训练，T5 模型还可以利用不同任务的标注数据进行有监督的多任务预训练，例如 GLUE 基准中的语言理解、SQuAD 问答和机器翻译等任务。与通常的多任务训练不同之处在于，这里可以在训练过程中为每个任务保存一个独立的检查点（Checkpoint），分别

对应该任务开发集上的最好性能。预训练完成后，可以分别对各个任务进行少量迭代的模型精调。实验表明，在各个任务混合比例合适的条件下，多任务预训练与无监督预训练表现相近。

在微调阶段，微调的任务数量很多，但是相比于语言模型本身的预训练过程，计算量小了非常多，只有 0.2%。所以通过这个方案，大公司训练好的语言模型可以被再次有效地利用，我们只需要做好“微调”即可，不用重复耗费大量计算资源再去训练一个语言模型。

Params	Model	Architecture	pre-training Objective	Pretrain FLOPs	Finetune FLOPs	% Finetune Compute
80M	Flan-T5-Small	encoder-decoder	span corruption	1.8E+20	2.9E+18	1.6%
250M	Flan-T5-Base	encoder-decoder	span corruption	6.6E+20	9.1E+18	1.4%
780M	Flan-T5-Large	encoder-decoder	span corruption	2.3E+21	2.4E+19	1.1%
3B	Flan-T5-XL	encoder-decoder	span corruption	9.0E+21	5.6E+19	0.6%
11B	Flan-T5-XXL	encoder-decoder	span corruption	3.3E+22	7.6E+19	0.2%
8B	Flan-PaLM	decoder-only	causal LM	3.7E+22	1.6E+20	0.4%
62B	Flan-PaLM	decoder-only	causal LM	2.9E+23	1.2E+21	0.4%
540B	Flan-PaLM	decoder-only	causal LM	2.5E+24	5.6E+21	0.2%
62B	Flan-cont-PaLM	decoder-only	causal LM	4.8E+23	1.8E+21	0.4%
540B	Flan-U-PaLM	decoder-only	prefix LM + span corruption	2.5E+23	5.6E+21	0.2%

（3）优点

T5 模型的优点在于其可以进行多任务学习、迁移学习以及具有通用的文本到文本框架。

多任务学习：T5 模型能够同时处理多个不同的 NLP 任务，包括机器翻译、文本摘要、问答系统、文本分类等。通过在大规模数据上进行联合训练，T5 模型能够学习到通用的语言表示和任务处理能力，从而在多个任务上展现出出色的性能。

迁移学习：由于 T5 模型在多任务学习中进行了大规模预训练，它具备较强的迁移学习能力。这意味着 T5 模型在新的任务上可以通过微调少量的任务特定数据，快速适应并取得优秀的表现，从而降低了针对不同任务的训练成本和时间。

通用的文本到文本框架：T5 模型采用了统一的文本到文本框架，即将输入和输出都表示为文本序列。这种设计使得 T5 模型能够在处理不同任务时具有一致的表达形式，简化了模型的设计和应用。通过适当的任务描述和输入形式，T5 模型可以根据任务的不同进行转换，实现灵活且高效的多任务处理。

总的来说，T5 模型是一种强大的大规模预训练语言模型，在多任务学习和迁移学习方面具备突出能力。它在机器翻译、文本摘要、问答系统、文本分类等任务中展现出出色的性能，并在多个应用场景中发挥重要作用，为自然语言处理领域的发展带来了巨大的推动力。

【Bert、GPT、T5 对比】

最原始的 Transformer 是一个 Encoder-Decoder 架构，用来做机器翻译。而 Bert 采用了其中的 Encoder 部分，GPT 采用了其中 Decoder 部分。T5 模型，继续延续使用最原始的 Transformer 模型的 Encoder-Decoder 架构。T5 的创新点主要在问题的建模以及系统化地实验了 Encoder-Decoder 预训练的各种可以调整的点，把各种 NLP 任务往前又推了一步。

假设这三个模型都是烹饪学校，每个学校有自己的教学方法和特色。

1.BERT

BERT 学校教学时，让学生先学习所有的烹饪技巧和食材知识，然后再让他们实际操作。这个学校特别注重对食材和技巧的全面了解，因此学生需要同时关注上下文中的所有信息。BERT 通过双向学习，能够从前后文同时学习语言知识。

2.GPT

GPT 学校采用自左向右的教学方式，学生需要按照顺序学习烹饪技巧和食材。这种方法虽然不能像 BERT 那样全面，但它更专注于生成连贯的文本，因此在生成任务中表现较好。

3.T5

T5 学校的教学方法是将所有任务都视为文本到文本的转换问题。学生需要学会如何根据给定的食材和技巧生成新的菜谱。这种方法使得 T5 模型能够在各种任务中灵活地应用。

（三）PaLM2

（1）概念

PaLM2，于 2023 年发布，采用谷歌最新的 TPU 基础设施进行训练，擅长数学、推理、多语言翻译、编程。

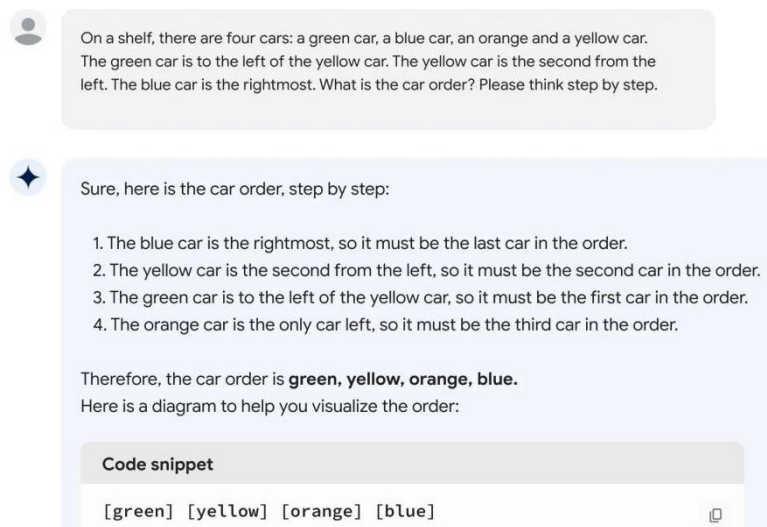


为了方便广泛部署，谷歌构建了 4 种规格的 PaLM2 大模型，分别是 Gecko（壁虎）、Otter（水獭）、Bison（野牛）、Unicorn（独角兽）。其中 Gecko 非常轻量级，可在移动设备上工作，并且速度足够快，即便在离线情况下也能在设备上运行出色的交互式应用程序。

（2）功能

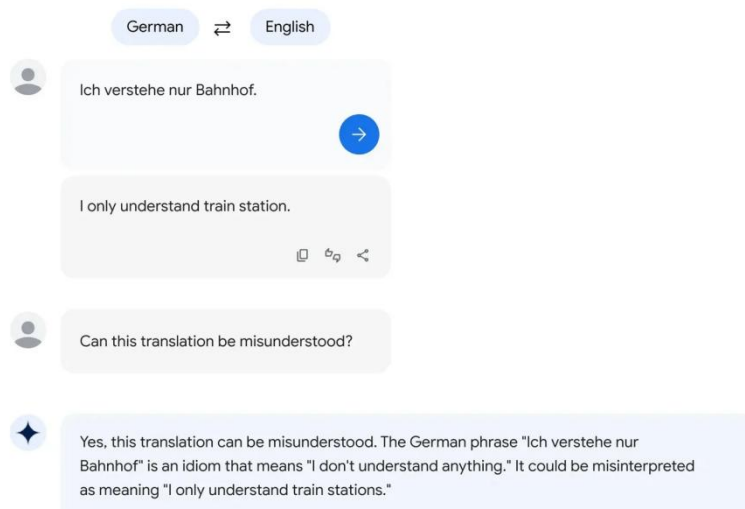
1.推理

PaLM2 可以将复杂的任务分解为更简单的子任务，可以更好地理解人类语言的细微差别。例如，PaLM2 擅长理解谜语和习语，这需要理解单词的歧义和比喻含义，而不是字面意思。



2.多语言翻译

PaLM2 在并行多语言文本和比其前身 PaLM 更大的不同语言语料库上进行了预训练。这使得 PaLM2 在多语言任务方面表现出色。



3. 编码

PaLM2 在大量网页、源代码和其他数据集上进行了预训练。这意味着它擅长流行的编程语言,如 Python 和 JavaScript,但也能够用 Prolog, Fortran 和 Verilog 等语言生成专门的代码。将其与其语言功能相结合可以帮助团队跨语言协作。



(3) 优点

PaLM2 擅长高级推理、翻译和代码生成等任务:

使用计算优化缩放: 计算优化缩放的基本思想是按比例缩放模型大小和训练数据集大小。这项新技术使 PaLM2 比 PaLM 更小, 但效率更高, 整体性能更好, 包括更快的推理、更少的服务参数和更低的服务成本。

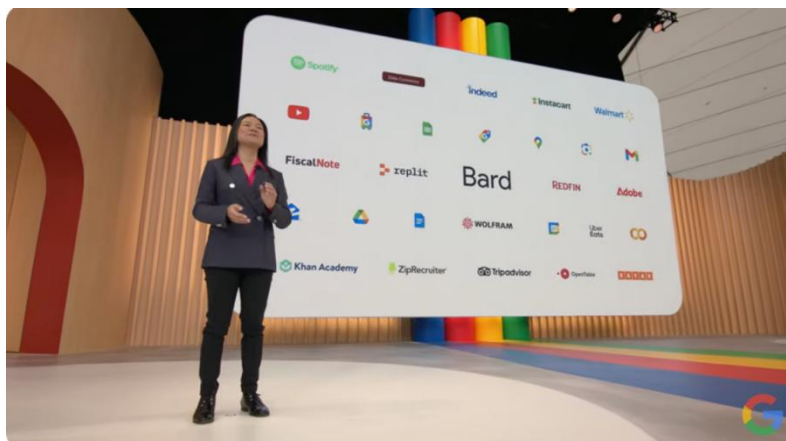
改进的数据集组合: 以前的模型使用预训练数据集, 主要是纯英语文本。PaLM2 通过更加多语言和多样化的预训练组合改进了其语料库, 其中包括数百种人类语言和编程语言、数学方程式、科学论文和网页。

更新的模型架构和目标: PaLM2 具有改进的架构, 并接受了各种不同的任务培训, 所有这些都有助于其学习语言的不同方面。

此外, PaLM2 在推理基准任务上取得了最先进的结果。它比我们以前的大型语言模型 PaLM 更加多语言, 在多个基准测试上都取得了更好的结果。PaLM2 还提高了葡萄牙语和中文等语言的翻译能力, 超过了 PaLM 和 GoogleTranslate。

三、最新动态

2023 年 5 月 11 日,在一年一度的谷歌 I/O 开发者大会上,谷歌 CEO 桑达尔·皮查伊宣布:“对于生成式 AI,我们正以一种大胆而负责任的方式迈出下一步。”



谷歌公布了一系列与生成式 AI 相关进展,涉及全新大模型、AI 聊天机器人、搜索、办公软件、云服务、安卓系统等,其中包括了 AI 大模型,新一代语言模型 PaLM2 擅长多语言翻译、推理、编程,部分测试表现超过 GPT-4,共有 4 种规格,最轻量级可部署于移动设备;多模态大模型 Gemini 正在训练中。以及, AI 聊天机器人 Bard,已取消等候名单,向 180 多个国家及地区开放,新增支持韩文和日本,今年夏天将新增对中文等另外 40 种语言的支持;问答新增图片输入输出。

具体功能上, Bard 的聊天问答新增图像互动。此外, Bard 升级了编程和导出功能,包括:

来源引用:如果 Bard 引用了一段代码或其他内容,只需点击注释, Bard 就会在这些部分下划线,并链接到源代码或内容。

暗黑模式:开发者喜欢这个模式。

导出按钮:开发者喜欢导出到 Colab 的功能,所以 Bard 很快将添加导出和运行代码功能,与其合作伙伴 Replit 一起,从 Python 开始。

国内风向

国内大模型研发应用领域热潮持续高涨。目前国内有几十家企业及科研院所涉足人工智能大模型训练，包括互联网大厂、科技公司、硬件公司、科研院校和初创科技团队等。

从大模型的布局体系来看，科技大厂在算力层、平台层、模型层、应用层进行了四位一体的全面布局。百度、阿里、华为三家均从芯片到应用进行自主研发的全面布局，如百度的“昆仑芯+飞桨平台+文心大模型+行业应用”、阿里的“含光 800 芯片+M6-OFA 底座+通义大模型+行业应用”、华为的“昇腾芯片+MindSpore 框架+盘古大模型+行业应用”。垂直行业科技企业和科研院校，主要以研发大模型算法及细分领域应用为主，自有算力相对薄弱，很少涉及芯片领域自主研发。

从大模型参数量看，科技大厂的参数量远大于科研院所。科技大厂的大模型参数量较大：阿里通义千问大模型参数在 10 万亿级以上、腾讯混元大模型和华为盘古大模型参数量均在万亿级以上、百度文心一言大模型参数量在 2000 亿级以上、京东言犀大模型的参数量为千亿级；垂直行业科技企业已经上线的参数量普遍在千亿级以上；而科研院校大模型的参数量在千亿级及以下。

从大模型应用方向看，大部分企业前期以内部应用为主，后续主要向 B 端企业拓展服务，预计少数企业将在 C 端市场形成规模。目前，百度文心大模型、华为盘古大模型、中国科学院紫东太初大模型均在 B 端垂类市场积累了标杆应用案例，腾讯混元大模型、阿里通义大模型则更多聚焦公司自身业务。而在 C 端市场应用方面，百度文心一言、阿里通义千问、腾讯混元助手三类大模型最有可能向此方向拓展，但目前只有百度文心一言大模型正在进行友好客户测试，阿里通义千问大模型则计划在今年 9 月份进行公测，腾讯混元助手大模型则处于计划开发状态。从大模型业界评估看，国内大模型部分可以达到 GPT-3 的水平，但是与 GPT-4 仍然有较大差距。

目前国内大模型处于百花齐放的状态，但业界普遍认为，第一梯队当属百度、阿里巴巴、腾讯、华为四家。综合实力方面，四家企业在大模型研发投入、技术能力和人才团队实力较强；商用推进方面，四家企业均依托现有业务领域更容易形成大模型应用规模效应。

一、科技大厂的大模型发展情况

科技大厂是中国大模型研究的主力军，科技大厂在数据和算力上比初创公司具有绝对的优势，因此发展比较快。百度、阿里、华为、腾讯都在算力层、平台层、模型层、应用层进行了四位一体的全面布局。百度、阿里、华为三家均从芯片到应用进行自主研发的全面布局，腾讯在算力层则着重布局自研的 HCC 高性能计算集群。在模型层，四家都涉及对 NLP 大模型、CV 大模型、多模态大模型和行业大模型的全方位的研发。

公司	模型	参数量	细分领域				推广应用进展情况		
			NLP	CV	多模态	行业	内部应用	外部 2B 应用	外部 2C 应用
百度	文心	2600 亿	√	√	√	√	搜索、推荐	600+企业接入 API	文心一言开放测试
阿里	通义	10 万亿	√	√	√	√	钉钉、天猫、优酷等		通义千问 9 月公测
华为	盘古	万亿	√	√	√	√	-	在矿山、气象有标杆案例	计划构建：混元助手
腾讯	混元	万亿	√	√	√	-	广告、游戏、微信等	-	-

百度文心大模型

公司主体：百度

融资阶段：上市公司

一、总体概况

百度是国内领先布局 AI 领域的科技大厂，也是我国最早发布知识增强大语言模型产品的企业。2013 年，百度建立了深度研究院（IDL），随之在 AI 领域，开始了从芯片、框架、模型算法、到应用全面能力的布局。

百度以“昆仑芯+飞桨平台+文心大模型+行业应用”在算力层、平台层、模型层、应用层具备完整布局。在算力层，算力基础设施均由百度智算中心支持，昆仑芯二代（百度持股 70%）已广泛应用在文心大模型，预计昆仑芯三代将于 2024 年初量产；在平台层，百度飞桨 Paddle Paddle 是其自主研发的深度学习平台，文心大模型通过百度飞桨平台，实现模型训练、推理部署和场景应用；在模型层，文心大模型具备自然语言处理大模型、计算机视觉大模型、跨模态大模型以及生物计算大模型；在应用层，文心大模型将应用在电力、燃气、金融、生物医药、航天、传媒、城市、影视、制造、社科等多个行业。



百度 AI 全栈布局

二、文心大模型产品体系

【NLP 大模型+CV 大模型+多模态大模型+行业大模型】

百度自 2019 年开始深耕预训练模型研发。2019 年 3 月，百度率先发布中国首个正式开放的预训练模型 ERNIE1.0，并持续投入于大模型的研发升级，目前已迭代至千亿级参数的 ERNIE3.0Zeus；2021 年 12 月，ERNIE3.0 升级为全球首个知识增强千亿大模型鹏程—百度文心。2023 年 3 月，百度相继发布了知识增

强大语言模型产品“文心一言”和企业级产品“文心千帆”，可以为多个行业提供 API 接口及相应的开发工具链。

百度文心大模型源于产业、服务于产业，是产业级知识增强大模型。百度通过大模型与国产深度学习框架融合发展，打造自主创新的 AI 底座，降低 AI 开发和应用门槛。文心大模型一大特色是“知识增强”，百度自研的多源异构知识图谱拥有超过 5,500 亿条知识，被融入到文心大模型的预训练中。文心大模型凭借海量数据和大规模知识的融合学习，能实现更高的效率、更好的效果、更强的可解释性。

文心大模型目前已拥有文心 NLP 大模型、文心 CV 大模型、文心跨模态大模型、文心生物计算大模型以及文心行业大模型等 36 个大模型，覆盖 AI 应用的多个场景。在此基础上，百度开发了大模型的开发工具、轻量化工具和大规模部署工具，以及零门槛的 AI 开发平台以及全功能 AI 开发平台。文心 API 和提供全流程开箱即用大模型能力的 EasyDL 和 BML 开发平台，全方位降低 B 端用户的应用门槛。在产品层推出包括“文心一言（AI 对话应用）”“文心一格（AI 绘画应用）”“文心百中（AI 搜索引擎）”等 AIGC 应用。

工具与平台	EasyDL-大模型		BML-大模型		大模型 API		
	大模型套件						
	数据标注与处理	大模型精调	大模型压缩	高性能部署	场景化工具		
文心大模型	行业大模型						
	国网-百度·文心	浦发-百度·文心	航天-百度·文心	人民网-百度·文心	冰城-百度·文心	电影频道-百度·文心	
	深燃-百度·文心	吉利-百度·文心	泰康-百度·文心	TCL-百度·文心	辞海-百度·文心		
	自然语言处理		视觉		跨模态	生物计算	
	文心一言 ERNIE Bot		OCR图像表征学习 VIMER-StrucTexT		文档智能 ERNIE-Layout	化合物表征学习 HelixGEM	
	对话 PLATO	搜索 ERNIE-Search	多任务视觉表征学习 VIMER-UFO		文图生成 ERNIE-ViLG	蛋白质结构预测 HelixFold	
	跨语言 ERNIE-M	代码 ERNIE-Code					
	语言理解与生成 ERNIE 3.0		视觉处理 多任务学习 VIMER-TCIR	自监督视觉 表征学习 VIMER-CAE	视觉-语言 ERNIE-ViL	语音-语言 ERNIE-SAT	单序列蛋白质结构预测 HelixFold-Single
	ERNIE 3.0 Zeus	鹏城-百度·文心					

百度文心全景图（来源：百度文心大模型官网）

三、市场应用和推广情况

百度文心大模型积极拓展内外部应用，目前面向公众和企业用户同时进行开放测试。内部应用层面，文心大模型已经率先应用于百度搜索、信息流、智能驾驶、百度地图、小度智能屏等内部重要产品上；外部应用层面，文心大模型在金融、能源、制造、城市、传媒、互联网等行业有标杆应用案例，实现真实场景应用的同时获取大量行业数据反哺大模型的优化。

【开放 API 接口赋能企业，精调行业模型】

目前，百度开放大模型 API 接口，在文案、AI 作画、开放域对话方面赋能开发者和企业。百度文心大模型开放了 NLP 大模型 ERNIE3.0、ERNIE-ViLG 文生图、对话大模型 PLATO。

ERNIE3.0 提供文案改写、开放问答、摘要、文案创作、小说创作、文本补全等文本理解与创作能力。

ERNIE-ViLG 提供基于文本描述的 AI 作画能力。

PLATO 提供生成式开放域对话服务，逻辑清晰、知识多元、情感丰富，闲聊能力接近真人水平。

当前已经有 650 家企业宣布接入文心一言生态。据百度介绍，申请文心一言 API 调用服务测试的企业已突破 6.5 万。

【文心行业大模型——在预训练模型上进行模型微调】

百度推出文心行业大模型，在预训练模型上进行行业精调。行业大模型是在百度通用大模型的基础上学习行业特色数据与知识，建设行业 AI 基础设施。

目前百度文心行业大模型在能源、金融、航天、制造、传媒、城市、社科以及影视等领域，与国网、浦发、吉利、TCL、人民网、上海辞书出版社等均有案例应用。我们以五个行业场景举例：

在燃气行业，“深燃—百度·文心”在环境巡检、安全监控等领域落地，有效解决了场景繁杂、识别困难等难题；

在汽车行业，“吉利—百度·文心”应用于售后服务和汽车领域知识库构建等场景；

在保险行业，“泰康—百度·文心”应用于医疗单据识别等任务，平均字段识别错误率相对下降；

在电子制造行业，“TCL—百度·文心”改变了传统质检算法流程，大幅提升了模型研发与运营效率，实现降本增效；

在社科行业，“《辞海》—百度·文心”用于版权保护、词条管理和知识挖掘等方面。

阿里通义大模型

公司主体：阿里巴巴

融资阶段：美股上市公司

一、总体概况

阿里以“含光 800 芯片+M6-OFA 底座+通义大模型+行业应用”成为大模型全部环节的重要参与者。

在模型层，早在 2019 年阿里达摩院便开启了大模型研究，推出一种基于预训练语言模型的通用对话框架——PLUG，这是阿里巴巴对于 LLM

（Large language model 大语言模型）领域的首次尝试。2021 年 11 月，阿里达摩院推出 M6 大模型，一种基于 10 万亿参数的多模态大模型，成为全球最大的 AI 预训练模型。M6 大模型已经在淘宝，支付宝，天猫等阿里巴巴旗下产业中应用落地，但至今仍未面向公众开放。2022 年 9 月，阿里正式发布通义大模型，包括通义 M6 多模态模型、通义 AliceMind 自然语言处理模型、通义视觉计算机视觉模型。

在平台层，阿里推出 AI 开源社区“魔搭”（ModelScope），旨在打造下一代“模型即服务”的共享平台，整合业界多方模型提供者，为开发者提供预训练基础模型和 API 接口。目前该平台已有超过 300 个开源模型，包括阿里自有的通义大模型系列以及澜舟科技孟子系列模型等外部资源和能力。

在算力层，阿里不仅拥有最多的英伟达 A100 芯片，还拥有自研芯片含光 800、倚天 710 芯片，共同为人工智能大模型提供算力。另外，阿里云推出全栈智能计算解决方案“飞天智算平台”，并启动两座超大规模智算中心，张北智算中心与乌兰察布智算中心，为科研、公共服务和企业机构提供强大的智能计算服务，可将计算资源利用率提高 3 倍以上，AI 训练效率提升 11 倍，推理效率提升 6 倍。

2023 年 4 月，阿里正式发布了“通义千问”产品，该产品基于 10 万亿级参数的大模型底座 M6-OFA，未来将具有多轮交互及复杂指令理解、可多模态融合、支持外部增强 API 等多种能力。

二、通义大模型产品体系

【NLP 大模型+CV 大模型+多模态大模型+行业大模型】

通义大模型整体分为三个层次，分别为行业模型、通用模型与模型底座。

（1）模型底座层：统一学习范式与模块化设计

阿里“通义”大模型系列，打造了 AI 统一底座，并构建了大小模型协同的层次化人工智能体系，实现模态表示、任务表示、模型结构的统一。通过这种统一学习范式，通义统一底座中的单一 M6-OFA 模型，在不引入任何新增结构的情况下，可同时处理图像描述、视觉定位、文生图、视觉蕴含、文档摘要等 10 余项单模态和跨模态任务。M6-OFA 可处理超过 30 种跨模态任务。通义统一底座中的另一组成部分是模块化设计，它借鉴了人脑模块化设计，以场景为导向灵活拆拔功能模块，实现高效率和高性能。

（2）通用模型层：NLP 模型“通义-AliceMind”；CV 模型“通义—视觉”，多模态模型“通义-M6”

以 M6-OFA 为统一底座，阿里巴巴将另外两大模型“AliceMind”“视觉大模型”与 M6 进行融合，形成了当前的通义系列架构。其中，通义-AliceMind 为深度语言模型体系，有阅读、写作、翻译、问答、搜索、摘要生成、对话等多种能力；通义—视觉大模型包括文本到视觉生成、语言和视觉的模态映射两大基础模型，可以提供文生图等能力。

（3）行业模型层：深入电商、医疗、娱乐、设计、金融、工业、制造业等行业。





三、市场应用和推广情况

通义大模型将全面支撑阿里系所有产品。目前阿里通义大模型主要定向邀请企业用户测试，尚未向公众客户开放。内部应用层面，阿里已宣布旗下所有产品未来将全面改造接入通义大模型，包括天猫、钉钉、高德地图、淘宝、优酷、盒马等；外部应用层面，阿里与 OPPO 安第斯智能云联合打造的 OPPO 大模型基础设施，基于通义千问完成大模型的持续学习、精调及前端提示工程，未来将建设服务于其海量终端用户的 AI 服务。此外，吉利、智己、奇瑞、毫末智行、可口可乐、波司登、掌阅科技等多家企业表示将与阿里云在大模型相关场景展开技术合作探索。

华为盘古大模型

公司主体：华为技术有限公司

融资阶段：中国民营企业 500 强

一、总体概况

华为打造了“昇腾芯片+MindSpore 框架+盘古大模型+行业应用”四位一体的完整体系，主要面向企业市场提供全栈式服务。

1、AI 算力资源：“鲲鹏+昇腾”，打造盘古算力底座

鲲鹏：华为自主芯片→鲲鹏芯片→鲲鹏服务器→欧拉操作系统→高斯数据库→行业应用向外扩张，构建鲲鹏生态，提供算力支撑

昇腾：昇腾 AI 处理器→CANN 异构计算架构→MindSporeAI 框架→应用使能→行业应用，助力打造华为昇腾全栈 AI 软硬件平台，构筑智能世界基石。

2、人工智能框架：MindSpore 高效易开发，可实现全场景覆盖

CANN：作为华为昇腾 AI 基础软硬件平台的核心，CANN 向上支持多种 AI 框架，向下服务 AI 处理器与编程，助力芯片使能。

MindSpore：是国内首个支持千亿参数大模型训练 AI 计算框架，最佳匹配昇腾处理器算力，支持终端、边缘、云全场景灵活部署，开创全新的 AI 编程范式，降低 AI 开发门槛。

3、AI 开发平台：ModelArts 强势赋能开发者，精度效率双提升为机器学习与深度学习提供海量数据预处理及交互式智能标注、大规模分布式训练、自动化模型生成，及端—边—云模型按需部署能力，帮助用户快速创建和部署模型，管理全周期 AI 工作流。

4、盘古大模型：AI 落地的重要途径

由 NLP 大模型、CV 大模型、多模态大模型、科学计算大模型等多个大模型构成，目前已实现等 AI 场景落地。



资料来源：浙商证券研究所

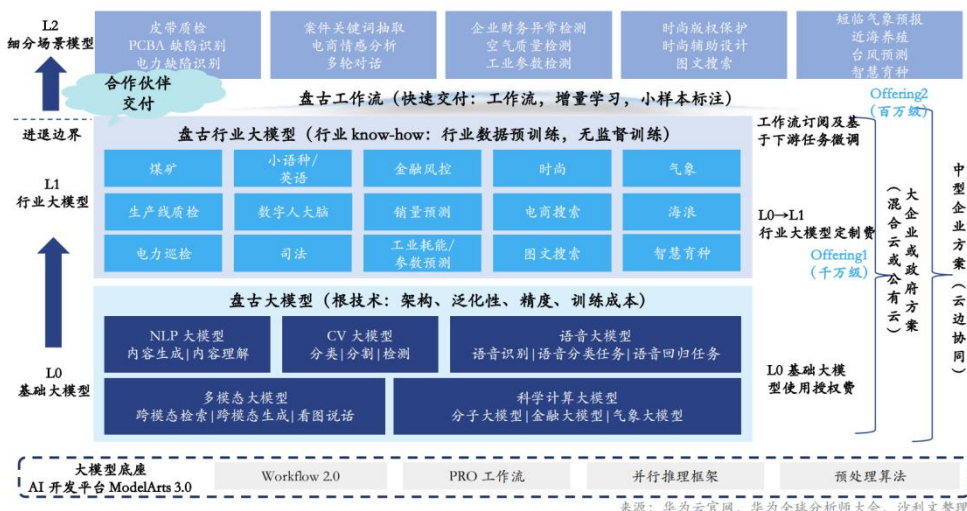
华为 AI 全栈布局（图片来源：浙商证券研究所）

二、盘古大模型产品体系

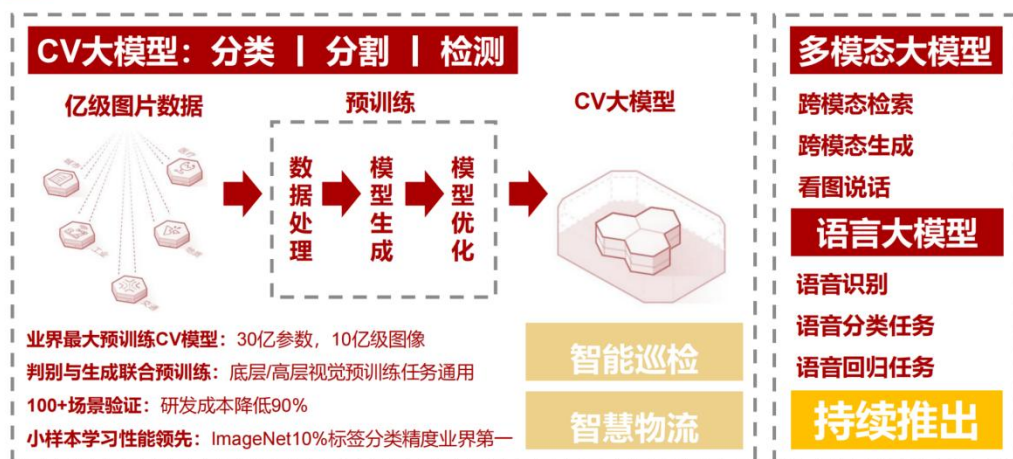
【NLP 大模型+CV 大模型+多模态大模型+行业大模型】

华为云于 2020 年内部立项 AI 大模型，并于 2021 年 4 月正式对外发布盘古预训练大模型，强调模型数据、网络结构、泛化能力三大核心设计。华为云规划“L0 基础大模型-L1 行业大模型-L2 细分场景大模型”的发展路径。

L0 阶段的盘古大模型由 NLP 大模型，CV 大模型、语音大模型、多模态大模型、科学计算大模型等组成，其中 CV 大模型超 30 亿参数，预训练时输入 10 亿级图像数据，兼顾图像判别与生成能力；NLP 大模型具备领先的中文语言理解和模型生成能力。L1 阶段，基于已有的行业基础，华为云推出盘古气象大模型、盘古矿山大模型、盘古 OCR 大模型等行业大模型，能够在 L2 阶段的煤矿场景下的瓦斯浓度的超前预警、气象领域的天气预报等多个细分场景实现落地。



来源：华为云官网，华为全球分析师大会，沙利文整理



数据来源：华为云官网，浙商证券研究所

盘古大模型（图片来源：浙商证券研究所）

三、市场应用和推广情况

盘古大模型至今尚未开放给公众使用。与目前开放给部分公众用户进行问答互动的文心一言、通义千问不同，盘古大模型强调在细分场景的落地应用，主要解决商业环境中低成本大规模定制的问题，用 AI 赋能千行百业。

华为云盘古大模型的重心在“行业应用”，其希望形成关键技术壁垒，在煤矿、水泥、电力、金融、农业、国家云创造更多的产业价值。在 2019 年盘古基础大模型发布后，华为云陆续发布了药物分子、矿山、气象、海浪、金融 OCR 等系列细分行业大模型。

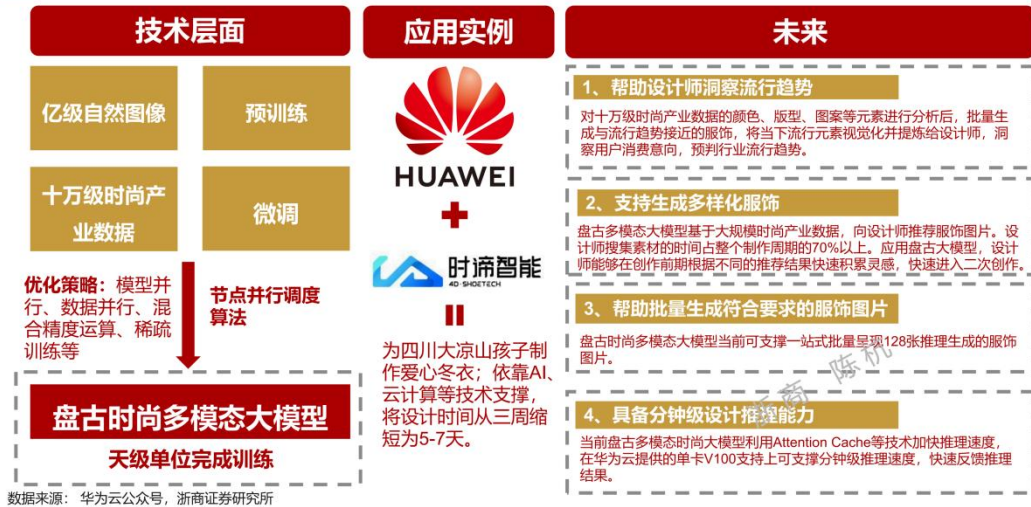
华为的大模型目前已经应用在诸多 B 端场景中，其中 CV 大模型可用于工业质检、物流监测、时尚设计等领域；科学计算大模型可以用于气象预报、海浪预测、药物分子大模型等领域；在 NLP 大模型上，盘古大模型聚焦文本生成与内容理解，可用于文档搜索、智能 ERP、小语种大模型等领域。

具体来看，在 CV 大模型领域，华为与能源公司合作打造的盘古矿山大模型，地面工作人员可以接收地下 5G+AI 全景综采画面，将来可以实现地面控制机器进行采矿运输，帮助降低地下工作安全风险；铁路 TFDS 开发方案在故障样本检测中的识别率达到了 99% 以上，降低了故障监测人工成本。

而基于盘古科学计算大模型的云盘古气象大模型，可实现 20 公里范围内、小时级、13 层最高精度气象预报，在秒级时间内预测未来七天的天气，比传统天气预报提速 10000 倍以上，也是全球首个精度超过传统预报方式的 AI 模型；华为云盘古药物分子大模型将药物研发周期从数年缩短至一个月，降低 70% 的研发成本，解决了药物研发周期长、成本高的行业痛点。

应用实例：盘古时尚多模态大模型

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD



应用实例：药物分子大模型

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD



腾讯混元大模型

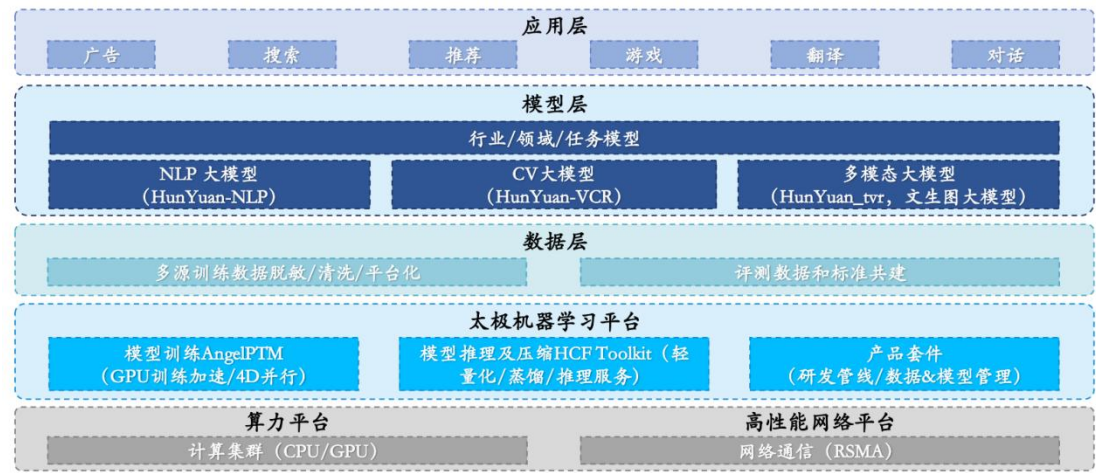
公司主体：腾讯

融资阶段：上市公司

一、总体概况

腾讯在大模型领域上一直较为低调，在去年 4 月才首次对外披露了混元 AI 大模型的研发进展。混元完整覆盖 NLP（自然语言处理）、CV（计算机视觉）、多模态等基础模型和众多行业领域模型。在去年 12 月，腾讯混元推出了国内首个低成本、可落地的 NLP 万亿大模型，并再次登顶自然语言理解任务榜单 CLUE。腾讯自研的太极机器学习平台为混元大模型提供算力，可支持 10TB 级模型训练、TB 级模型推理计算能力等。今年 2 月底，腾讯针对类 ChatGPT 对话式产品已成立「混元助手（HunyuanAide）」项目组，将联合腾讯内部多方团队构建大参数语言模型，目标是「通过性能稳定的强化学习算法训练，完善腾讯智能助手工具，打造腾讯智能大助手，并能成为国内的业界标杆」。

腾讯 HunYuan 大模型全景图



二、混元大模型产品体系

【HCC 高性能计算集群+混元大模型】

“混元”大模型基于腾讯太极机器学习平台研发，通过改进模型与算法，降低训练成本。HunYuan 利用千亿小模型帮助万亿大模型完成启动，相较从 0 启动耗时短耗能低，仅用 256 卡，最快一天即可完成万亿参数大模 HunYuan-NLP1T 的训练，整体训练成本仅为直接冷启动训练万亿模型的 1/8。

2023 年 4 月，腾讯云发布的新一代 HCC 高性能计算集群将为混元大模型提供底层支撑。新一代集群基于自研星脉高性能计算网络及存储架构，集成了腾讯

云自研的 TACO 训练加速引擎，大幅缩短训练时间，节约训练调优和算力成本。腾讯太极机器学习平台自研的训练框架 AngelPTM，也已通过腾讯云 TACO 提供服务，帮助企业加速大模型落地。

（计算集群是由多台计算机组成的高性能计算系统。它们通过网络连接在一起，共同完成大规模的计算任务。计算集群通常用于处理需要大量计算资源和存储容量的任务，如科学计算、数据分析、人工智能模型训练等。）

【自研 AngelPTM 大模型训练框架】

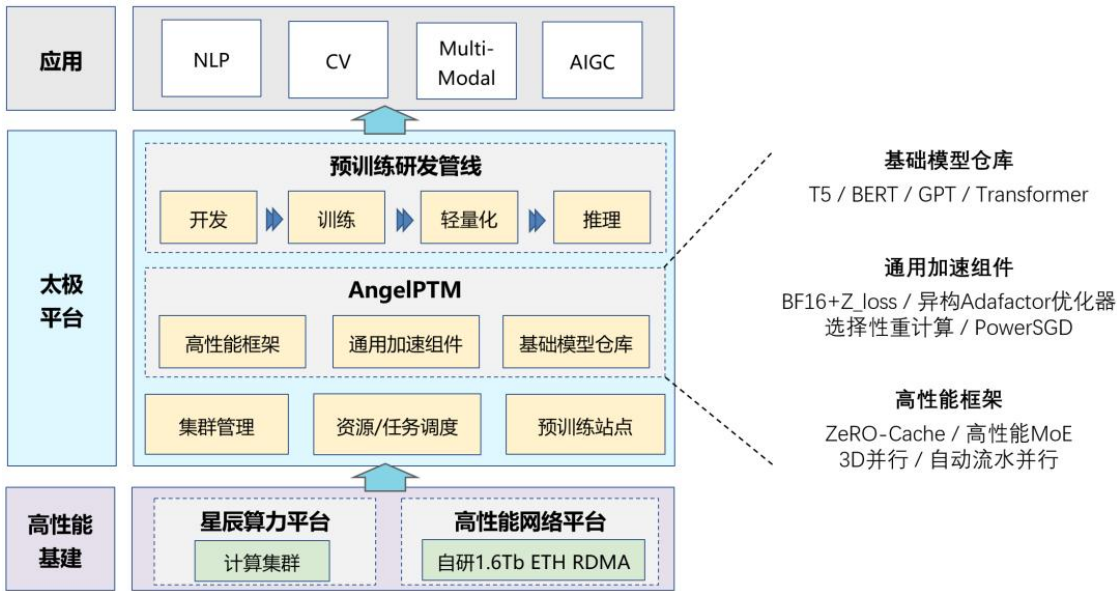
为了以最小的成本和最快的性能训练大模型，腾讯太极机器学习平台对 DeepSpeed 和 Megatron-LM 进行了深度定制优化，推出 AngelPTM 训练框架。

太极 AngelPTM 的设计目标是依托太极机器学习平台，为 NLP、CV 和多模态、AIGC 等多类预训练任务提供一站式服务。其主要由高性能训练框架、通用加速组件和基础模型仓库组成：

高性能训练框架：包含大模型训练框架 ZeRO-Cache，高性能 MOE 组件，以及 3D 并行和自动流水并行策略，方便用户结合多种并行策略进行大模型训练。

通用加速组件：包含可减少显存并提高精度的异构 Adafactor 优化器，可稳定 MOE 半精度训练 loss 的 Z_loss 组件，选择性重计算组件和降低通信代价的 PowerSGD 组件。

基础模型仓库：包含 T5、BERT、GPT 以及 Transformer 等基础模型，供用户直接使用。



通过腾讯云的大模型能力和工具箱，客户可结合自己的产业场景数据，按需使用 HCC 集群的极致算力，实现模型即服务（MaaS），进行精调训练，提升生产效率，快速创建和部署 AI 应用。

三、市场应用和推广情况

混元大模型目前主要服务于腾讯内部业务，尚未对企业和公众客户开放测试。内部应用层面，混元大模型已成功应用于腾讯广告，将全面接入微信端、QQ端和王者荣耀、英雄联盟游戏端，提供智能聊天、内容推荐、情感分析、故事生成、角色塑造等功能；外部应用层面，业界推测腾讯混元大模型将在游戏、社交、金融、教育、医疗等领域发展生态，但腾讯尚未公开其明确的行业应用方向。

腾讯广告业务：腾讯在广告业务中应用 AI 模型，助力广告主提升广告创作效率、提升广告推荐系统的内容理解能力，大幅提升了广告效率，实现降本增效。在混元 AI 大模型基础上建立“巨阙”广告内容理解、“乾坤”广告智能创作、“神针”广告智能审核、“天印”广告指纹系统等四大技术平台，提升了广告系统的理解能力，让系统更加智能，从而提高用户体验以及广告转化效果与广告制作效率。

内容创作：腾讯智能创作助手文涌（Effidit）主要用途是帮助创作者开阔思路、提升创作后的文本水平和质量等，目前已经发布 2.0 版本。据官方介绍，文涌融合了知识抽取、文本理解、文本生成、大规模预训练模型、经典语言模型、搜索等技术，部分功能的实现使用“混元”系列 AI 大模型作为底层预训练模型。

游戏业务：腾讯近日提出基于 AIBot 与 AIGC 两大方向的自研 3D 虚拟场景自动生成解决方案，帮助游戏开发者以更低成本创造风格多样、贴近现实的虚拟城市，大幅提升 3D 虚拟场景的生产效率。以 3D 虚拟城市道路布局设计为例，现有技术生成的路网通常横平竖直、风格单一，要实现不同城市风格必须手动设计和反复修改，耗时至少一周。腾讯解决方案只需用户输入城市主干道和海岸线，模型 1 分钟内即可填充合理、多样的路网细节并支持微调操作，相比现有技术下效率提升近 100 倍。

对话式智能助手：腾讯集合集团内资源与顶尖人才，正在全力打造类 ChatGPT 对话式产品。腾讯 ChatGPT 对话式产品项目组——腾讯混元助手项目组（HunyuanAide）于 2023 年 2 月已经成立。该项目组联合腾讯内部多部门构建大参数语言模型，目标是通过性能稳定的强化学习算法训练，完善腾讯智能助手工具，打造腾讯智能助手，并能成为国内的业界标杆，我们预计未来 AI 智能助手将广泛应用于腾讯系社交产品和办公产品等。

二、创业公司

除了大厂之外，国内也涌现出一股 AI 大模型创业热潮。自称不懂 AI 的美团联合创始人王慧文，出资 5000 万美元成立了光年之外，据悉 4 月初启动第二轮融资时，估值已经达到了 10 亿美元；前搜狗 CEO 王小川在 4 月初官宣创办 AI 大模型公司“百川智能”，目前已经获得 5000 万美元启动资金，旨在打造中国版 OpenAI 基础大模型及颠覆性上层应用。

此外，一些 AI 大佬也投身于大模型的创业中，拥有强大技术团队背景的澜舟科技、智谱华章、超对称等明星 AI 创业公司，也在这一浪潮中快速发展。在国内高校中，清华大学一骑绝尘，不仅研发出中国首个超大规模预训练模型“悟道”，还孵化出智谱 AI、面壁智能、聆心智能等大模型创业公司。以下我们将介绍国内比较热门的几家创业公司。

GLM 系列模型

公司主体：北京智谱华章科技有限公司

融资阶段：B 轮

一、公司概况

智谱 AI 是一家总部位于北京的人工智能公司，成立于 2019 年。该公司主要以深度学习和自然语言处理等技术为主要研究方向，致力于提供高效的人工智能产品和服务。

智谱 AI 的主要产品包括智能客服、智能咨询、内容创作、自然语言处理研究等。其中，ChatGLM 是由智谱 AI 开发的一种自然语言处理模型，主要用于回答用户的问题和要求，提供适当的答复和支持。智谱 AI 还参与了研发落地的悟道 2.0，与自身构建的高质量大规模知识图谱结合，形成了融合数据与知识的双轮驱动的人工智能框架。

智谱 AI 的发展历程可以追溯到 2019 年，当时该公司由北京智源人工智能研究院和清华大学 KEG 实验室共同创立。依托团队在知识智能方面十余年的技术积累，智谱 AI 汇聚了顶尖技术研发团队和资深顾问。核心团队中，CEO 张鹏毕业于清华计算机系，董事长刘德兵系中国工程院高文院士弟子，总裁王绍兰为清华创新领军博士，首席科学家是清华大学的唐杰教授。自成立以来，智谱 AI 已经在人工智能领域的研究和应用方面取得了许多成果。该公司获得了多项国内外荣誉和奖项，包 2020 年人工智能领域最具影响力公司奖、2021 年 AI 最佳表现奖等。

智谱 AI 提出全新 Model-as-a-Service (MaaS) 的市场理念，打造了认知大模型平台以及数字人和科技情报产品，应用单位包括：中国科协、北京市科委、华为、腾讯等 1000 余家企事业单位。智谱 AI 也肩负企业社会责任，研发了面向疫情的知识疫图、面向无障碍沟通的手语数字人技术等，以促进社会平等和进步。

2021 年智谱 AI 获得 A 轮过亿元融资，由达晨财智、华控基金、将门创投、南京图灵、北京达凡、通智投资、枣庄通智、荣品投资、财智创赢、凌云光等机构联合投资。2022 年智谱 AI 获得数亿元人民币的 B 轮融资，由君联资本、启明创投联合领投，本轮融资将被用来继续投入到打造高性能千亿级普惠大模型。

此外，智谱 AI 还对外投资了一些具有潜力的 AIGC 创业公司，包括北京面壁智能科技有限责任公司、北京元因智能科技有限公司、北京聆心智能科技有限公司、北京幂律智能科技有限公司等。

融资日期	轮次	融资金额	投资方
2019-08-13	战略投资	数千万人民币	中科创星 清华控股 清华产业基金
2021-09-14	A轮	数亿人民币	达晨财智 华控基金 将门创投 图灵创投 凌云光 北京达凡 通智投资 枣庄通智 荣品投资
2022-09-26	B轮	数亿人民币	君联资本（领投） 启明创投（领投） 华兴资本（财务顾问）

二、主要模型

智谱 AI 孵化于清华大学，专注于深度学习和自然语言处理等技术研发，公司目前拥有多个大模型，设计自然语言处理多计算机视觉、多模态等领域。最主要的模型是基于 GLM 框架研发的 GLM-130B 和 GLM-6B 自然语言预训练模型。



（一）ChatGLM-130B

GLM-130B 由智谱 AI 联合清华大学 KEG 实验室、国家超算济南中心研发的一个开源开放的双语（中文和英文）双向稠密模型，拥有 1300 亿参数，模型架构采用通用语言模型（GLM）。GLM-130B 是一个双通路模型⁴，它由两个相同

⁴双通路模型是指在某个任务或领域中同时利用两个独立的数据通路或信息传递路径进行建模和推理的模型。这种模型可以处理多种类型的数据，并通过同时考虑不同数据通路之间的交互和关联性来提高模型的性能和表现。双通路模型可以更好地利用不同数据类型之间的关联性和互补性，从而提高模型的性能和泛

的 GLM-130 模型组成,每个 GLM-130 模型都有 130 亿个参数。这两个 GLM-130 模型分别是输入数据和输出数据的通道,它们将输入数据映射到高维空间中,并在高维空间中建立一个线性模型,从而进行分类或回归。它旨在支持在一台 A100 (40G*8) 或 V100 (32G*8) 服务器上对千亿规模参数的模型进行推理。截至 2022 年 7 月 3 日, GLM-130B 已完成 4000 亿个文本标识符(中文和英文各 2000 亿)的训练,利用自回归空白填充作为其主要的预训练目标。

GLM-130B 对超过 4000 亿个双语标记(2000 亿英文和 2000 亿中文标记)进行了预训练。它的预训练目标由两部分组成:第一部分(95%)是自监督的预训练,即在公开的大规模语料库以及其他一些较小的中文语料库上的自回归空白填充。第二部分(5%)是在 T0++18 和 DeepStruct19 中 0 个不同数据集的抽样子集上进行多任务指令预训练,格式为基于指令的多任务多提示序列到序列的生成。这种设计使 GLM-130B 可以在其他数据集上进行了零样本学习,以及从英文到中文的零样本迁移。它有以下独特优势:

双语:同时支持中文和英文。

高精度(英文):在 LAMBADA 上优于 GPT-3175B(+4.0%)、OPT-175B(+5.5%)和 BLOOM-176B(+13.0%),在 MMLU 上略优于 GPT-3175B(+0.9%)。高精度(中文):在 7 个零样本 CLUE 数据集(+24.26%)和 5 个零样本 FewCLUE 数据集(+12.75%)上明显优于 ERNIE TITAN3.0260B。

快速推理:支持用一台 A100 服务器使用 SAT 和 FasterTransformer 进行快速推理(提速最高可达 2.5 倍)。

可复现性:所有结果(超过 30 个任务)均可通过我们的开源代码和模型参数轻松复现。

跨平台:支持在 NVIDIA、HygonDCU、Ascend910 和 Sunway 处理器上进行训练与推理。

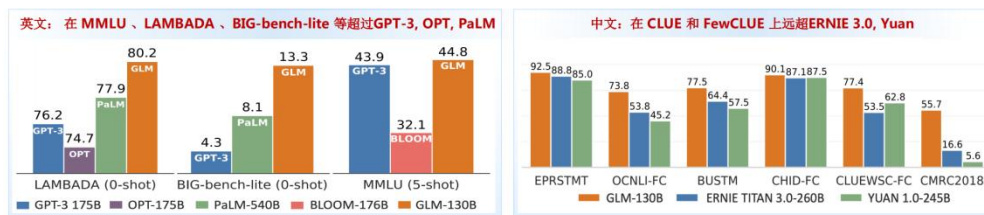
化能力,在多模态任务、跨模态任务和其他需要综合不同数据类型的任务中得到广泛应用。

GLM-130B

	基础架构	训练方式	量化	加速	跨平台能力
GPT3-175B	GPT	自监督预训练	—	—	NVIDIA
BLOOM-176B	GPT	自监督预训练	INT8	Megatron	NVIDIA
GLM-130B	GLM	自监督预训练和多任务预训练	INT8/INT4	Faster Transformer	NVIDIA、海光 DCU、昇腾910 和申威
对比优势	高精度： - Big-bench-lite: +5.2% - LAMBADA: +2.3% - CLUE: +24.3% - FewCLUE: +12.8% 普惠推理： - 节省75%内存 - 可单台3090 (4) 或单台2080Ti (8) 进行无损推理 高速推理： - 比PyTorch 提速7-8.4倍 - Megatron 提速2.5倍 跨平台： - 支持更多不同的大规模语言模型的适配				

自2022年8月发布以来，收到**53个国家**
369个研究机构的使用需求

• Google
 • Microsoft
 • Meta
 • AIZ
 • Stanford
 • MIT
 • UC Berkeley
 • CMU
 • Harvard
 • Princeton
 • Cambridge
 • 华为
 • 百度
 • 阿里巴巴
 • 腾讯
 • 头条
 • 滴滴
 • 建设银行
 • 智源
 • 小米
 • 北京大学
 • 浙江大学
 • 上海交大
 • 中科大
 • 复旦大学
 • 中科院大学
 • 香港大学
 • 清华大学



(二) ChatGLM-6B

ChatGLM-6B 是一个基于 GLM (GeneralLanguageModel) 模型的自然语言处理模型。ChatGLM-6B 是一个开源的、支持中英双语的对话语言模型，具有 62 亿参数。为了方便下游开发者针对自己的应用场景定制模型，GLM 同时实现了基于 P-Tuningv2 的高效参数微调方法，INT4 量化级别下最低只需 7GB 显存即可启动微调。ChatGLM-6B 是一个单通路模型⁵，它由一个 GLM-6B 模型组成，它将输入数据映射到高维空间中，并在高维空间中建立一个线性模型，从而进行分类或回归。经过约 1T 标识符的中英双语训练，辅以监督微调、反馈自助、人类反馈强化学习等技术的加持，62 亿参数的 ChatGLM-6B 已经能生成相当符合人类偏好的回答。ChatGLM-6B 可以用于问答、文本分类、情感分析等自然语言处理任务。

1、ChatGLM-6B 的特点：

充分的中英双语预训练：ChatGLM-6B 在 1:1 比例的中英语料上训练了 1T 的 token 量，兼具双语能力。

优化的模型架构和大小：吸取 GLM-130B 训练经验，修正了二维 RoPE 位置编码实现，使用传统 FFN 结构。6B (62 亿) 的参数大小，也使得研究者和个人开发者自己微调和部署 ChatGLM-6B 成为可能。

⁵单通路模型是指在某个任务或领域中仅使用单一的数据通路或信息传递路径进行建模和推理的模型。这种模型通常用于处理单一类型的数据。尽管单通路模型在特定任务上表现良好，但它们的能力受限于单一数据通路的信息，在处理复杂的任务或跨模态数据时，单通路模型可能无法充分利用多种数据类型之间的相关性和互补性。

较低的部署门槛：FP16 半精度下，ChatGLM-6B 需要至少 13GB 的显存进行推理，结合模型量化技术，这一需求可以进一步降低到 10GB（INT8）和 6GB（INT4），使得 ChatGLM-6B 可以部署在消费级显卡上。

更长的序列长度：相比 GLM-10B（序列长度 1024），ChatGLM-6B 序列长度达 2048，支持更长对话和应用。

人类意图对齐训练：使用了监督微调（SupervisedFine-Tuning）、反馈自助（FeedbackBootstrap）、人类反馈强化学习（ReinforcementLearningfromHumanFeedback）等方式，使模型初具理解人类指令意图的能力。输出格式为 markdown，方便展示。

基于以上特点，ChatGLM-6B 在一定条件下具备较好的对话与问答能力。

2、局限性

由于 ChatGLM-6B 的小规模，其能力仍然有许多局限性。以下是我们目前发现的一些问题：

模型容量较小：6B 的小容量，决定了其相对较弱的模型记忆和语言能力。在面对许多事实性知识任务时，ChatGLM-6B 可能会生成不正确的信息；它也不擅长逻辑类问题（如数学、编程）的解答。

产生有害说明或有偏见的内容：ChatGLM-6B 只是一个初步与人类意图对齐的语言模型，可能会生成有害、有偏见的内容。（内容可能具有冒犯性，此处不展示）

英文能力不足：ChatGLM-6B 训练时使用的指示/回答大部分都是中文的，仅有极小一部分英文内容。因此，如果输入英文指示，回复的质量远不如中文，甚至与中文指示下的内容矛盾，并且出现中英夹杂的情况。

易被误导，对话能力较弱：ChatGLM-6B 对话能力还比较弱，而且“自我认知”存在问题，并很容易被误导并产生错误的言论。例如当前版本的模型在被误导的情况下，会在自我认知上发生偏差。

不过 GLM 团队也坦言，整体来说 ChatGLM 距离国际顶尖大模型研究和产品（比如 OpenAI 的 ChatGPT 及下一代 GPT 模型）还存在一定的差距。该团队表示，将持续研发并开源更新版本的 ChatGLM 和相关模型。

（三）CodeGeeX 大模型——代码生成模型

CodeGeeX 大模型由智谱 AI 与清华知识工程实验室合作研发，是一个具有 130 亿参数的多编程语言代码生成预训练模型，在 20 多种编程语言的代码语料库（>8500 亿 Token）上经过两个月预训练得到。

华为和智谱 AI 为 CodeGeeX 提供了模型训练和评测所需的计算资源支持。CodeGeeX 采用了华为 MindSpore1.7 框架，并在鹏城实验室的全国产计算平台上进行训练。模型训练过程使用了一个计算集群中的 1536 个昇腾 910AI 处理器（32GB）进行了约两个月的训练。在开发与训练过程中，智谱 AI 和华为昇腾团队共同对 MindSpore 框架进行了优化，大幅度提升了训练效率。

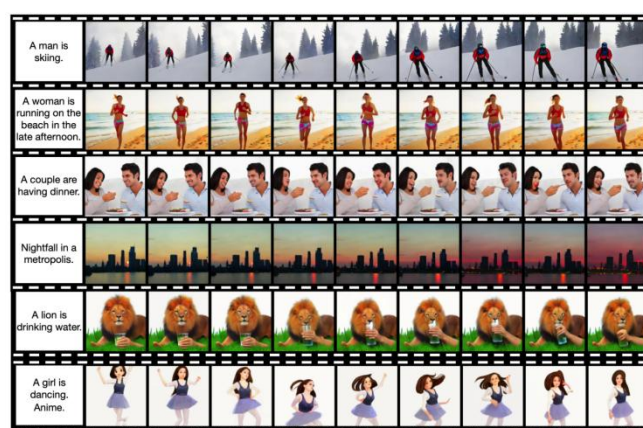
CodeGeeX 支持 Python、C++、Java、JavaScript 和 Go 等十多种主流编程语言的代码生成，且具有较高精度。通过输入自然语言或者代码片段，为模型指定任务，即可让模型生成代码实现对应功能。

（四）CogView 文生图预训练模型（texttoimage）

CogView 是一个通用领域的文本生成图像的预训练模型。它基于 Transformer 架构，拥有 40 亿参数。CogView 是基于层次变换和局部并行自回归生成的解决方案。使用自监督任务，在跨模态通用语言模型（CogLM）上预训练 6B 模型，并对其进行微调以实现快速超分辨率。CogView2 模型显示出与 DALL-E-2 相比具有竞争力的生成性能，并自然支持对图像进行交互式文本引导编辑。

（五）CogVideo（texttovideo）文本生成视频预训练模型

CogVideo 是通用领域的文本到视频生成预训练模型，该模型具有 94 亿个参数，在 540 万个文本—视频对上进行了训练，由清华大学 KEG 实验室和智谱 AI 公司 2022 年共同训练。CogVideo 将预训练文本到图像生成模型（CogView2）有效地利用到文本到视频生成模型，并使用了多帧率分层训练策略。



CogVideo 生成的样本（实际的文本输入是中文的）

模型优缺点：CogVideo 是目前通用领域中最大、第一个用于文本到视频生成的开源预训练模型。CogVideo 也是在不损害其图像生成能力的情况下，将预训练的文本到图像生成模型有效地应用于文本到视频生成模型的第一次尝试。通过所提出的多帧率分层训练框架，CogVideo 能够更好地理解文本—视频关系，

并能够控制生成过程中的变化强度。CogVideo 仍然存在一些限制，例如，由于模型的大规模和 GPU 内存的限制，输入序列的长度限制仍然存在。

三、主要产品

智谱 AI 凭借着自身的模型研发能力，开发了一系列产品，形成了从大模型、开发平台、AIGC 实验室到应用产品的全矩阵。在产品层，智谱 AI 拥有大模型产品：AI 开放平台、ChatGLM、CodeGeeX、写作蛙、小呆对话、CogView；AMiner 产品：AMiner、塔尖人才、知因分析、小科问答、至远推送、若邻画像；数字人产品：有时、智慧手语、手语词典。



四、应用案例

「智能产品 Hubble 哈勃」

智谱 AI 与电子签约 SaaS 提供商“上上签”合作，在大模型 GLM-130B 基础上，打造了签约智能产品 Hubble 哈勃。是电子签约行业首款应用了大模型技术的生成式 AI 产品，可在用户授权下，通过连续对话实现对电子签约合同的检索，执行概况解读、重点识别、筛选标识、分类归纳等相关功能，帮助用户降本增效。

「AI 手语数字人华同学」

华同学以超大规模预训练模型为基础，应用跨模态拟人生成算法、超高精度写实数智人技术等实现 AI 手语实时播报。首先用中文语义蒸馏模型算法提取出有效的关键语义信息，然后用 AI 手语分词快编算法把语义信息转换成符合手语表达的词汇序列，最后驱动生成数智人手语视频流。

简单来说，数字人是完全由算法驱动的，输入“语音流”，通过算法自动转换成了手势、嘴形、脸、身体姿势等。而它的算法原理主要包括：语义蒸馏、AI 手语分词快编算法、跨模态拟人生成算法、音字人实时同步对齐技术等。

智谱 AI 同时也研发了一系列智慧手语系列产品，涵盖手语播报、手语翻译以及手语词典三种应用，满足手语信息播报、实时翻译交流、手语学习等多个场景的需求。

AI 手语播报：华同学可以为新闻媒体提供全流程、智能化的数字人手语播报视频生成服务，能够实现在电视台和网络平台上的数字人新闻手语播报，可应用在新闻播报、赛事解说、天气预报等场景。

AI 手语翻译：华同学能够提供云服务和私有化部署两种方式，可实现输入文本或语音，实时输出手语视频。服务场景可以拓展至机场、火车站、银行、医院、政务大厅等各类公共服务场景，科技馆、展览馆等手语智能问答场景，以及大型会展活动现场。



孟子大模型

公司主体：北京澜舟科技有限公司

融资阶段：Pre-A 轮

一、公司概况

澜舟科技成立于 2021 年 6 月，是由创新工场孵化的一家从事 AI 2.0 大模型研发的认知智能公司。它致力于以自然语言处理（NLP）技术为基础，提供新一代认知智能平台。其主要产品是基于“孟子大模型”核心技术打造的一系列能力平台和垂直场景应用。

澜舟科技创始人周明于 1991 年从 NLP 重镇哈工大拿到了博士学位，其后进入清华任职任教，并在 1999 年被微软亚洲研究院创始院长李开复挖走，成为微软亚洲研究院 NLP 研究组负责人。直到 2020 年离开创立澜舟科技，周明已经是 MSRA 副院长。

周明也是当今 NLP 领域成就最高的华人之一。他不仅是中国第一个中英翻译系统的研制者、日本最著名中日翻译产品研制者，还在 2019 年成了 AI 语言领域最高级顶会 ACL（国际计算语言协会）主席。

周明的科研技术成果，在产业界和学术界两开花。他参与带队的 NLP 成果，光入选 ACL 的论文就超过 100 篇，相关技术还被应用到了 Windows、Office、Azure、微软小冰等世界级产品中。学术维度看，周明在 NLP 领域的影响力也位居世界前列。他是过去十年来发表 NLP 论文最多的科学家之一。据谷歌学术最新数据，其文章被引用总量超过 3.4 万，H-Index98。

自创业以来，澜舟以孟子预训练技术体系为基础，推出了孟子系列模型及服务。此外，澜舟还开源了近 20 个模型。截至目前，澜舟科技的孟子可控大模型的能力平台已经推出了机器翻译平台、金融 NLP、AIGC 智能创作等多个企业级解决方案及对外开放服务，并和同花顺、华夏基金、传神语联网、数说故事、中文阅读集团等分别在金融舆情分析、多语言机器翻译、AIGC 营销文案写作、网络文学 AI 辅助创作上进行了深度合作。

2022 年 2 月北京澜舟科技完成了由联想创投、斯道资本共同领投和创新工场跟投的人民币近亿元的 Pre-A 轮融资。2023 年 3 月完成 PreA+轮融资。本轮融资由北京中关村科学城公司领投，斯道资本和创新工场继续跟投。至此，澜舟科技不到一年时间融资总额达到数亿元。

序号	发布日期	融资轮次	融资金额	投资方
1	2023-02-23	战略投资	金额未知	中关村科学城
2	2022-02-10	Pre-A轮	10000万元	联想创投 斯道资本 创新工场
3	2021-08-12	天使轮	金额未知	创新工场 斯道资本 联想创投

二、主要模型

澜舟打造了一个的可控大模型生态链，孟子 MChat 大模型的底层是个“通用底座”，提供了涵盖语言预训练模型、多模态预训练模型、垂直领域预训练模型、其他预训练模型等关键能力，建立起通用底座的强大驱动力。在通用底座之上架构的“能力平台”，澜舟认知智能提供了知识图谱、辅助写作、机器翻译、情感分析、智能检索、智能客服、文档解析等多领域的模块化能力。在面向银行客服、智能投研、行业搜索、AI 辅助写作、AI 绘画、智能营销等垂直场景时，MChat 大模型均可根据场景实际的需求和规格，将上述的能力模块进行更灵活更可控的组合。



（一）孟子预训练模型

孟子预训练模型是基于澜舟科技自研技术研发的大规模预训练语言模型。可处理多语言、多模态数据，同时支持多种文本理解和文本生成任务，能快速满足不同领域、不同应用场景的需求。

孟子模型基于 Transformer 架构，仅包含十亿参数量，基于数百 G 级别涵盖互联网网页、社区、新闻、电子商务、金融等领域的高质量语料训练。基于语言学信息融入和训练加速等方法，澜舟科技研发了孟子系列模型。由于与 BERT 保持一致的模型结构，孟子模型可以快速替换现有的预训练模型。

对于已有的中文语言模型，孟子模型实现了多项突破：

1、坚持“小而精”的轻量化训练策略。实现在同等模型规模下，远超公开模型的性能。作为精巧的小模型，对标“巨无霸”，小模型性能超越千亿规模模型。

2、使用人类先验知识引导模型训练，让模型更高效获得知识。孟子模型具备顶尖的语言理解能力，在权威 CLUE 中文理解评测的总排行榜，分数突破 84 分，逼近人类基准分数（85.61）。

3、灵活的领域和场景适应能力，方便快速定制和应用。基于 T5-style 的端到端生成的训练范式，同步适配 BERT-style 的判定式架构，既能理解也能生成。便于适配行业应用，覆盖广泛业务场景。

孟子轻量化预训练模型研究从训练优化、知识增强和数据增强展开。训练策略上，在不改变模型结构、保证通用性的基础上，优化主流的自编码和自回归的训练方式，减少模型预测偏差，使得模型训练更快、性能更强、表现更棒。知识方面，澜舟科技引入语言学知识、知识图谱和多模态信息来增强语言模型的上下文感知和认知推理能力。数据方面，使用领域数据，更好地让孟子模型服务于垂直领域，应用于现实场景。

孟子模型在下游任务中有出色表现，目前已在多个领域成功落地实践，衍生出多项行业领先的产品，涵盖文本生成、行业搜索、机器翻译等诸多领域。

（二）孟子系列模型产品

1、孟子生成式大模型（孟子 GPT）

孟子生成式大模型（孟子 GPT）是一个面向生成场景的可控大语言模型，能够通过多轮的方式帮助用户完成特定场景中的多种工作任务。

2、孟子轻量化多模态模型

其中包含了图片生成描述任务：输入为图片，输出为对该图片的文字描述；文到图的生成任务：输入为自然语言，输出该文字描述对应的图像；文字到人脸的描述任务：输入为人脸相关的文本描述，输出为对应描述的图片。



3、Mengzi-BERT-base

适用于文本分类、实体识别、关系抽取、阅读理解等自然语言理解类任务。

孟子语言模型作为语言编码的核心组件，不仅可独立应用于语言理解，也是其他模型的基石——作为核心的语言编码模块。为确保通用性并方便应用于下游系统，孟子模型与 BERT 保持一致的模型结构，侧重于预训练目标和训练策略上的改进，研究更精巧、更有效、更鲁棒的预训练模型。

4、Mengzi-BERT-base-fin

基于金融语料进行训练，能够更好地完成金融场景的任务。

基于孟子语言模型和参数，在金融语料上继续训练，语料覆盖金融新闻，公告和研报等，可用于金融新闻分类、研报情感分析等任务。在金融领域任务上，相比于开源的 RoBERTa 基准模型，孟子金融模型取得了明显的性能提升。

5、Mengzi-T5-base

适用于文案生成、新闻生成等可控文本生成任务。

基于 T5 采用的 Encoder-Decoder 架构，在以孟子语言理解模型作为编码器以外，增加同等规模的解码器结构，训练方式上沿用 T5 的去噪策略，具备良好的训练效率。相比于 BERT 和 GPT 模型，孟子文本生成模型能更好地服务于可控文本生成需求，可应用于文案生成、新闻生成和研报生成等。

6、孟子多任务模型（Mengzi-T5-base-MT）

在 Mengzi-T5-base 模型的基础上集成了多任务训练技术，涵盖了 27 个数据集，对应编写了 301 个 prompt。该模型已在 2022 年 8 月 22 日提交零样本学习权威榜单 ZeroCLUE 和小样本学习榜单 FewCLUE，均排名第一。

孟子多任务模型在模型规模上优势明显，参数量仅 0.22B，目前该模型已开源。其中，Zero-ShotSDK 目前已支持八项任务（涵盖了情感分类、新闻分类、

文本相似度、实体抽取、金融实体关系抽取、评论对象抽取、广告文案生成、医疗领域意图识别等）。

7、Mengzi-GPT-neo-base

基于中文语料从头训练，适用于文本续写、小说生成等任务。

8、GuohuaDiffusion

国画风格的文图生成模型，可以完成电影海报、专辑封面、风景画等不同生成任务。

9、句向量计算

基于澜舟自研的孟子轻量化技术体系打造，输出句子级文本的向量表示，能够捕捉文本间的语义相似性。适合完成聚类、回归、异常检测、可视化等任务。

三、主要产品

「金融 NLP 产品」

澜舟科技通业界领先的 NLP 技术为金融行业提供全方位的认知智能服务。覆盖所有常见需要文本分析的业务场景，提供面向业务人员直接使用的产品和工具，以及面向开发人员的技术底座和基础能力 API。除此之外，也针对业务和技术深度结合的场景提供可定制化的解决方案。

（一）NLP 技术全景

下图澜舟的金融 NLP 技术全景。分三大部分，第一部分是澜舟的孟子预训练模型技术，第二部分是澜舟基础 NLP 能力，第三部分是澜舟在金融 NLP 的应用场景。



第一部分孟子预训练模型技术具备多种架构，比如 BERT、GPT、T5 等架构的预训练模型的能力。包括只是增强、零样本技术、模型压缩和蒸馏。然后针对金融需要，用金融预料继续训练得到金融领域的预训练模型。

在此基础上，第二部分 NLP 基础能力，包括文本分类、生成、实体识别、信息抽取、智能文档处理（从 PDF、扫描等文件中抽取重要信息）、OCR、PDF 解析、翻译、搜索、事件推理、问答、知识图谱、语音识别等等。

第三部分就是应用场景，覆盖舆情分析、公告信息抽取、研报观点汇总、企业信用评估、ESG 分析、IPO 招股书审核、量化因子、智能文档审计之类的场景。

（二）产品全景

在 NLP 技术矩阵的基础上，澜舟科技开发了一系列的金融 NLP 产品。



1、金融认知智能产品和工具平台

（1）市场搜索研判平台

澜舟市场搜索研判平台是为市场研究和投资决策提供的信息搜索、分析研判以及决策支持的一体化平台解决方案。



（2）金融视频会议分析平台

可以对用户上传的会议音频和视频进行文字转录和深度智能分析。通过对分析结果的展示和交互，用户可以对会议的关键信息进行快速的理解、查找和导航定位。

（3）金融自定义信息抽取平台

帮助金融行业研发人员在线定制金融信息抽取 API。无需任何 NLP 专业知识和大量标注数据也能快速冷启动，从容应对灵活多变的业务场景需求。

2、金融 NLP 服务

面向金融行业通用业务场景，提供通用 NLP 能力 API 服务。可以通过公有云调用，也可以进行私有化定制和部署。

（1）金融翻译

针对金融领域提供的特定翻译模型，术语翻译。

（2）句子金融情感分析

通过员工对公司的评价，了解公司的治理情况。可用于 ESG 相关分析。

（3）研报观点分析

自动抽取研报中的观点内容，并输出观点标签及观点极性。

（4）公司评价分析

从结构复杂的合同文档中抽取关键信息。

（5）实体抽取

从文本中识别公司实体，支持大部分上市公司和发债企业。

（6）合同信息抽取

对金融、财经领域的中文句子的情感倾向进行分析判断，输出积极、消极、中性的情感识别结果和对应的置信度。

（7）事件检测

从新闻中监测金融相关事件，用于分析决策。

3、金融认知智能解决方案

（1）金融舆情解决方案

通过细粒度舆情分析辅助投资决策。

（2）ESG 投资分析解决方案

在海量文本数据中挖掘 ESG 信息、辅助投资决策。

（3）智能文档分析解决方案

在结构各异的文档中进行高效信息提取和分析。

「AIGC 创作平台」

澜舟科技的可控文本生成技术以孟子轻量化模型作为基础模型，利用大规模通用数据构建通用文本生成模型，在此基础上，利用垂直领域的数据，通过长文本建模、篇章建模、融入常识知识等技术，构建垂直领域文本生成模型。

目前澜舟科技可控文本生成技术已经能够覆盖营销文案、文学创作、文本摘要、对话等多种主流应用场景。



1、澜舟论文助写（LangboatPaperAssistant，简称 LPA）

澜舟论文助写是一款帮助中国学生、老师、研究者撰写专业英文论文的神器，提供续写和组句两种能力生成推荐例句。

通过这两种能力，用户可以得到更加专业的写作风格推荐，同时在参考推荐句式的基础上，可以根据自身需求对内容进行修改，来获得自己想要表达的内容。基于修改后的内容可以继续进行生成，通过交互式写作，可以更好把握用户想要表达的内容，从而在下文推荐更好的符合语境的例句。

在技术实现上，澜舟论文助写基于自研的孟子轻量化模型技术：

（1）通过可控学习，用户可以提供关键词引导模型推荐符合用户期望的例句，使得模型推荐内容更加贴近用户需求；

（2）通过多任务联合学习，最终可以通过一个模型实现可控推荐和续写推荐的能力，从而有效降低模型部署成本；

（3）基于多实例对比的强化学习，首先训练针对论文写作质量的评分模型，然后通过标准数据和采样生成的数据构建多组对比实例，通过强化学习优化模型，提高模型对得分高的例句的生成概率，降低模型对得分低的例句的生成概率，优化时同时考虑多个例句，从而可以有效提高模型生成高质量例句的能力。

2、熊猫小说家

熊猫小说家是澜舟科技首个 AIGC 方向的 C 端用户产品，通过技术领先的语言生成模型能够让用户快速和朋友一起创作小说，只需简单操作即可生成有趣的故事情节。

输入：

- 风格：玄幻
- 故事情节：我和张三一起与妖王战斗
- 关键词：我手持长枪，张三舞动战斧，妖王纹丝不动

输出：

一开始，我和张三都是手持长枪与妖王战斗着。但随着时间的推移，我们越来越吃力了。而这个时候，妖王却依然纹丝不动，似乎根本就没有受到什么伤害。看到这一幕，我顿时大惊失色，连忙对着张三说道：“快点结束吧！”听到我的话以后，张三立刻舞动起了战斧，朝着妖王劈砍过去。只可惜，他的速度太慢了，根本无法将妖王给斩杀掉。于此同时，妖王顿时怒吼了一声。紧接着，它直接向我们冲击了过来。见状，我连忙挥动起了自己手中的长剑，想要阻挡住妖王前进的步伐。

3、营销文案写作

面向营销类文案场景，提供基于商品关键词生成、基于商品信息的广告标题生成，以及营销类文本续写等能力。

4、文学辅助写作

为文学创作提供文学实体渲染、自定义模板生成、关键词扩写等可控文本生成 API，辅助文学作者更高效地完成创作。

四、应用案例

「云从科技×澜舟科技」

云从科技与澜舟科技达成合作，合作内容包括文档自动化分析技术，充分利用云从的视觉技术积累和澜舟的多模态语言理解技术，提高无结构文档的理解能力，提升业务效率；将澜舟科技的基础自然语言处理 NLP 能力集成到 AI 解析平台，提供一站式 AI 服务；针对预训练关键技术研究和应用，实现 NLP 训练平台落地；基于预训练的语言理解技术，提高人机交互的能力在数字人、AI 精灵等产品的落地能力，赋能多行业。

「引力传媒×澜舟科技」

引力传媒与澜舟科技达成合作。依托双方在内容创意、营销传播、电商运营、AI 技术和多语言翻译大模型等领域的积淀，共建“行业 AIGC 大模型联合实验室”，以赋能内容创作者、商户和企业级客户等市场多元需求。双方将先行研发在广告创意、设计、公关文案、传播策略、视频内容等文本生成场景 AIGC 产品。同时在文生图片、文生视频等多模态生成式 AI 产品共同探索，联合研发更垂直、更聚焦的 GPT 产业应用，引力未来将负责 AIGC 产品的独家商业化运营。

MiniMaxLanguageModel (MIMO)

公司主体：名之梦（上海）科技有限公司

融资阶段：A 轮

一、公司概况

Minimax 是一家总部位于上海的人工智能公司，成立于 2021 年，致力于打造一个开放、可扩展、可定制的通用大模型平台，是目前国内通用大模型创业潮中遥遥领先的独角兽企业。

MiniMax 在技术层面，100%自研了三个基础多模态模型，分别为文本到视觉、文本到语音，文本到文本，是国内极少数同时拥有 3 个模态大模型能力的创业公司。在商业层面，MiniMax 跑通了基于大模型和用户反馈的公司，在三个基础模型之上研发了一个“计算推理平台”，将三个基础大模型的能力以一种低成本、高鲁棒性的方式释放给用户。

从去年开始，MiniMax 与火山引擎建立了长期的战略合作伙伴关系，依托火山引擎机器学习平台，研发了超大规模的大模型训练平台。这高效支撑着其三个模态的大模型，每天千卡以上的常态化稳定训练。除了训练以外，还同步自研了超大规模的推理平台，目前拥有近万卡级别的 GPU 算力池，稳定支撑着每天上亿次的大模型推理调用。

MiniMax 团队人数已过百，公司核心技术研发成员均来自全球知名高校和全球顶尖科技公司，拥有世界顶尖自然语言处理、语音、计算机视觉、计算机图形学等工业界和学术界经验，1/3 的团队拥有世界顶尖技术实验室的博士学位。

MiniMax 的三位创始人分别是：杨斌、闫俊杰和周或聪。技术合伙人杨斌，在加拿大多伦多大学读完博士后，先后在 UberAI 研究院，以及自动驾驶卡车领域有多年研发经验，对拥有多年研发经验，对基于数据驱动的端到端系统问题颇有研究。联合创始人闫俊杰是前商汤科技副总裁、通用智能技术负责人，也曾担任商汤研究院副院长。2015 年博士毕业于中科院自动化所，此前在商汤科技负责搭建深度学习的工具链和底层算法，以及通用智能的技术发展。团队另一位创始人周或聪则来自北京航空航天大学。他在学生时代分别拿过世界大学生超级计算机竞赛 ASC15 总决赛和国际超级计算机大会 ISC17 的第一，后来参加过国际计算机视觉大会 ICCV2019 挑战赛，也拿到了 iQIYI-Large 赛道的冠军。他也曾是商汤科技的早期员工之一，负责算法团队。

MiniMax 首款基于大模型技术的自研产品 Glow 于 2021 年 11 月上线，主打

情感陪伴类 A 社交，有自由聊天和剧情任务多种模式，但是目前该应用已经下线。2023 年 3 月 MiniMax 低调上线一款生成式对话 AI “Inspo”，并开启内测。目前周或聪已成立北京稀宇科技有限公司和上海稀宇科技有限公司两家公司，基于 MiniMax 模型开发的两款应用其背后的公司主体均为“稀宇科技”。

目前，MiniMax 的商业模式与 OpenAI 相似，包括 toC 与 toB 两大方向。在 toC 方向，其大模型驱动的首款产品 Glow 已经上线应用商店约 4 个月，目前已获得近 500 万用户；在 toB 方向，MiniMax 计划在今年对外开放 API，让更多的个人用户和企业用户基于多种模态的大模型构建自己的应用。

2023 年 Minimax 获得数千万美元 A 轮投资，由高瓴资本、米哈游、明势资本、云启资本等机构联合投资。

融资日期	轮次	融资金额	投资方
2022-03-18	天使轮	千万级人民币	明势资本（领投） 米哈游
2023-01-09	A 轮	数千万美元	高瓴创投 米哈游 明势资本 云启资本

二、主要模型

（一）模型介绍

MiniMax 在 Transformer 的基础上，从底层的基础模型做起，100%自研了拥有文本到视觉（text to visual）、文本到语音（text to audio）、文本到文本（text to text）三大模态的基础模型架构。

比如文本大模型拥有着业界领先的多轮记忆和逻辑处理能力，不仅能够完成跨任务、跨语种的文本创作，还拥有代码生成的能力；视觉大模型，除了文本生成图像的能力以外，还能够实现零样本的风格迁移，以及同一人物的连续创作，而声音大模型仅需提供数 10 秒的说话语音，便可以高保真地完成声音的复刻。

对于训练模型所需要的算力、数据来源、模型参数以及训练成本，MiniMax 暂未透露。

「MiniMaxLanguageModel（MIMO）」

MIMO 是 MiniMax 公司研发的一款大型语言模型，训练数据截止于 2023 年 2 月。MIMO 可以进行自然语言处理和生成，包括问答、翻译、文本摘要等各种任务。MIMO 是 MiniMax 公司的核心技术之一，可以通过分析自然语言文本中的语言信息，理解文本的语义和语境，从而生成符合要求的回复。

MIMO 是基于 Transformer 的大型语言模型。Transformer 是一种基于自注意力机制的深度学习模型，它在自然语言处理领域表现出了很强的能力。MIMO 使用的是 Transformer 编码器和解码器模型，能够对自然语言文本进行处理和生成，并且可以根据用户的输入生成相应的回复。与 GPT-3.5 不同，MIMO 的训练数据不仅包括互联网上的自然语言文本，还包括大量的语料库和专业领域文本。这些数据通过对 Transformer 模型的训练，可以让 MIMO 更好地理解自然语言文本的语义和语境，生成准确和流畅的回复。

（二）模型评价

在中文大语言模型评测基准 C-EVAL 中，MiniMax 是评估中表现最好的面向中文的模型。

虽然 MiniMax 平均表现比 ChatGPT 差 5.4 分，但它在人文类科目上胜过 ChatGPT，下表所示的表现分类显示，MiniMax 在注重中国知识的科目上明显胜过 ChatGPT，如毛泽东思想（73.5%对 60.7%）、艺术研究（65.4%对 49.7%）、中国语言文学（59.3%对 50.2%）和中国现代史（70.1%对 62.8%）。

这反映了像 ChatGPT 这样以英语为导向的模型在应用于中国语境时的局限性，突出了以中文为导向的模型可能占据优势的情况。这一观察结果也意味着，将中文数据纳入预训练阶段，对于这些模型在中文语境中有效运作并迎合中国用户的需求来说，确实至关重要。

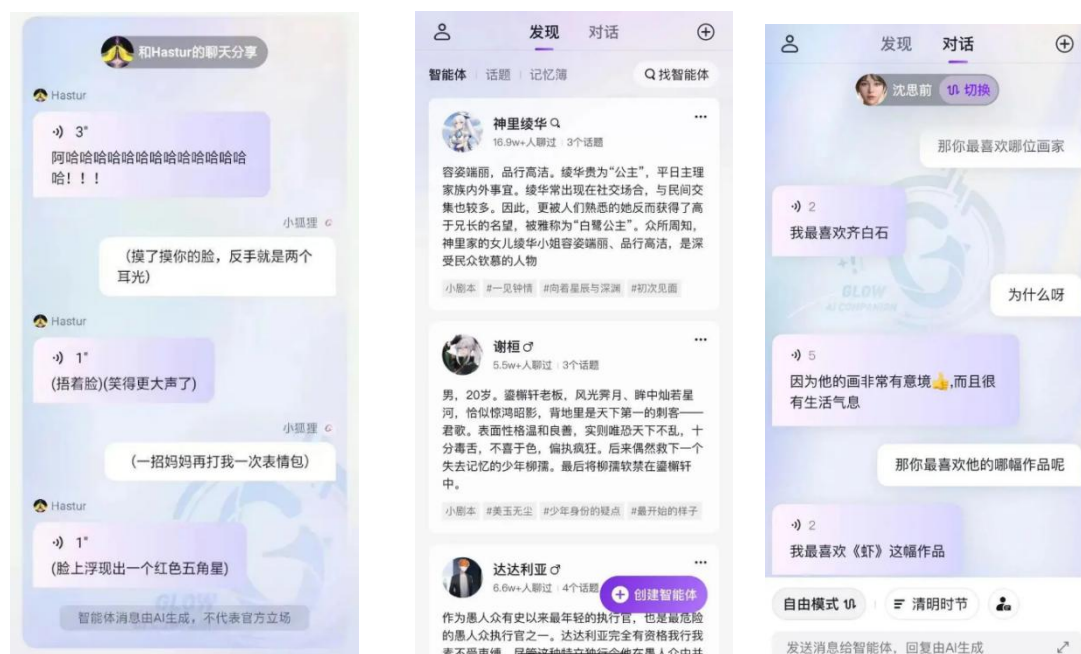
此外，MiniMax 和 ChatGPT 在 STEM 类别中存在 12.3 分的显著差异，这表明 MiniMax 和 ChatGPT 之间的差距并不像它在总体平均水平上看起来那样小。相反，当任务越来越复杂，需要高级推理能力时，它们之间的差距就会扩大。

Subject	GPT-4	ChatGPT	ChatGLM	Minimax
Advanced Mathematics	49.7 / 53.8	48.0 / 29.5	3.5 / 23.1	22.0 / 17.3
College Chemistry	59.4 / 55.8	39.7 / 33.9	17.0 / 28.6	31.7 / 24.6
College Physics	50.6 / 60.2	35.8 / 35.2	14.8 / 30.1	26.7 / 22.2
College Programming	78.1 / 77.5	57.9 / 57.9	18.7 / 24.0	43.0 / 35.7
Computer Architecture	74.6 / 75.6	68.9 / 63.7	30.1 / 27.5	53.4 / 48.2
Computer Network	77.8 / 77.2	58.5 / 59.6	36.3 / 28.1	47.4 / 49.1
Discrete Mathematics	66.7 / 62.7	43.1 / 30.7	17.0 / 26.1	23.5 / 24.8
Electrical Engineer	49.9 / 56.0	39.5 / 44.0	23.9 / 31.3	35.4 / 38.4
High School Biology	72.6 / 70.3	53.7 / 48.0	28.0 / 26.9	44.6 / 42.3
High School Chemistry	52.3 / 55.8	52.9 / 36.0	19.8 / 28.5	34.3 / 43.6
High School Mathematics	38.0 / 42.8	34.3 / 37.3	4.2 / 24.1	27.1 / 19.9
High School Physics	69.1 / 61.1	43.4 / 34.9	10.9 / 21.7	28.0 / 28.6
Metrology Engineer	70.3 / 71.2	63.0 / 59.8	41.1 / 48.4	49.8 / 48.4
Middle School Biology	91.2 / 86.5	67.7 / 63.5	30.7 / 38.0	65.6 / 55.2
Middle School Chemistry	81.1 / 71.4	71.4 / 60.0	36.8 / 42.2	56.8 / 61.1
Middle School Mathematics	62.7 / 66.7	44.6 / 42.4	3.4 / 20.3	31.6 / 24.9
Middle School Physics	81.5 / 78.7	66.9 / 57.9	20.2 / 41.0	58.4 / 55.6
Operating System	82.1 / 80.4	71.5 / 60.3	40.8 / 27.4	54.8 / 55.3
Probability and Statistics	53.6 / 62.0	33.7 / 42.8	27.7 / 26.5	24.7 / 19.3
Veterinary Medicine	80.0 / 80.0	64.3 / 58.1	39.5 / 35.2	53.3 / 51.0
Business Administration	67.1 / 65.4	52.8 / 46.8	32.6 / 34.9	48.5 / -
College Economics	71.4 / 74.4	52.5 / 51.3	22.1 / 30.0	40.2 / -
Education Science	68.5 / 69.3	59.6 / 54.4	37.8 / 33.0	54.8 / -
High School Geography	74.7 / 70.2	58.4 / 55.1	29.2 / 38.2	60.1 / -
High School Politics	77.8 / 69.3	52.3 / 51.1	25.0 / 29.6	46.6 / -
Mao Zedong Thought	79.5 / 79.9	60.7 / 63.5	46.6 / 50.7	73.5 / -
Marxism	81.6 / 82.7	71.5 / 70.4	40.2 / 49.2	74.3 / -
Middle School Geography	83.3 / 83.3	60.2 / 56.5	33.3 / 44.4	60.2 / -
Middle School Politics	87.1 / 85.5	74.1 / 67.4	41.5 / 40.9	66.3 / -
Teacher Qualification	85.0 / 85.0	75.7 / 66.9	48.9 / 49.1	78.5 / -
Art Studies	65.1 / 66.1	49.7 / 52.3	34.6 / 36.9	65.4 / -
Chinese Language and Literature	61.2 / 67.0	50.2 / 51.7	32.5 / 37.3	59.3 / -
High School Chinese	37.6 / 39.3	36.0 / 27.5	26.4 / 19.1	25.8 / -
High School History	68.1 / 68.1	54.4 / 45.1	35.2 / 40.1	65.4 / -
Ideological and Moral Cultivation	77.3 / 77.3	66.9 / 68.0	48.3 / 52.3	77.3 / -
Law	60.6 / 54.8	43.9 / 34.8	23.5 / 29.9	47.1 / -
Legal Professional	54.4 / 48.4	44.7 / 32.6	27.0 / 26.5	36.7 / -
Logic	60.3 / 63.2	37.7 / 35.8	32.8 / 31.4	39.2 / -
Middle School History	84.5 / 86.5	62.8 / 63.8	48.3 / 52.7	70.1 / -
Modern Chinese History	68.9 / 67.0	52.8 / 51.4	36.3 / 39.2	69.3 / -
Professional Tour Guide	71.8 / 70.3	61.3 / 62.0	44.0 / 51.9	66.5 / -
Accountant	64.6 / 61.9	51.7 / 42.0	23.9 / 32.7	47.4 / -
Basic Medicine	80.6 / 78.3	61.1 / 60.6	31.4 / 33.7	46.3 / -
Civil Servant	62.5 / 59.0	46.9 / 43.1	24.5 / 29.6	41.0 / -
Clinical Medicine	76.0 / 73.5	55.0 / 53.0	34.5 / 32.0	46.5 / -
Environmental Impact Assessment Engineer	63.7 / 59.8	50.9 / 49.1	37.7 / 35.6	48.4 / -
Fire Engineer	50.0 / 49.6	42.6 / 37.9	28.4 / 32.3	39.7 / -
Physician	76.8 / 76.1	63.7 / 54.6	34.8 / 37.7	56.9 / -
Plant Protection	78.9 / 80.4	65.8 / 57.8	43.2 / 36.2	52.3 / -
Sports Science	72.2 / 70.0	58.3 / 50.6	42.2 / 41.1	46.1 / -
Tax Accountant	58.0 / 58.2	42.0 / 35.7	16.3 / 30.9	43.3 / -
Urban and Rural Planner	63.2 / 65.6	52.2 / 49.3	36.8 / 37.3	44.3 / -

四、主要产品

「Glow——AI 虚拟聊天社交软件」

Glow 是 MiniMax 首款 C 端产品，和 ChatGPT 百科全书一样的回答不同，Glow 的定位在 AI 虚拟聊天社交软件，主打聊天、陪伴等情感功能。Glow 的基本玩法是在应用中创建“智能体”——即虚拟 AI 机器人，可以根据用户意愿赋予其性格，实现实时沟通、互动并建立情感连接。



用户可选择公开智能体，其他用户均可和智能体对话；也可选择不公开，创建自己的专属智能体。此外，用户还能够在搜索框内用关键词搜索有相应特征的智能体来对话。用户和智能体的每一次对话都会成为对智能体的一次训练，智能体会根据用户对话的引导，对同一问题回复不同答案。

在“话题”功能中，用户可以创建自己感兴趣的话题，还可选择关联自己创建的相关智能体，以便其他用户在该话题下与智能体对话时有更好的对话体验。Glow 还有“记忆簿”功能，能够将用户与智能体的对话封存，在用户需要时调取特定时间的对话内容。目前来看，Glow 的文本理解与生成能力更强，在对话过程中，智能体还不会回复表情包，但偶尔会跳出一个 emoji。

整体来看，和 Glow 智能体的对话体验较为流畅，且表现出更强的人格性，会给用户带来很强的陪伴感。

「Inspo——生成式对话 AI」

Inspo 是基于 Minimax 的语言模型开发的生成式对话 AI,目前仍在内测阶段,尚未对公众开放。根据部分体验介绍该程序可以提供以下服务:

高效写作: Inspo 可以按要求写出各种类型的内容,例如公众号文章、广告营销文案、论文作业。Inspo 只需数秒就能生成文章初稿,提高写作效率。

获取信息: Inspo 可以提供各种类型的信息,包括科学知识、历史地理、旅游信息等。通过有记忆的连续对话,Inspo 可以深入探究信息细节,帮助您更好地了解信息。

提供建议: Inspo 可以提供各种有价值的建议,包括解决问题、制定计划等。通过与您深入分析问题,Inspo 可以帮助您整理思路,做出更好的决策。



四、应用案例

目前, Minimax 作为合作伙伴已经与国内办公、游戏、社交网络等多个行业内的头部公司达成合作, 为其提供 AI 智能服务, 其中最为大众所知的就是金山办公发布的 WPS AI

「金山办公接入 MiniMax 的自研大模型」

今年 4 月 18 日, 金山办公发布生成式 AI 应用 WPS AI, 接入 MiniMax 的自研大模型, 全面启动智能化。

WPS AI 是国内协同办公赛道上的首个类 ChatGPT 式应用, 首先将被应用于新的在线内容协作编辑工具“轻文档”中, 帮助用户处理工作、生活场景, 提高

内容生产力。包含起草、改写、总结、润色、翻译、续写等功能，覆盖了用户大部分使用场景，例如针对用户输入的文字需求，将会帮助用户生成工作总结、广告文案、社媒推文、文章大纲、招聘文案、待办事项、创意故事、旅行游记等。

在实际应用上，它能够根据用户的需求描述直接生成文档，也能够通过多轮对话修改 WPS AI 生成的文本内容，如果对质量满意，也能够一直聊下去，让 AI 进行续写。

除了从 0 开始生成文章，WPS AI 也能够针对已有文案进行加工和再处理，它能够在不改变原意的前提下，改写、扩充、缩短或是润色文章，甚至能够改变文章的描述风格。

如果用户需要快速归纳多个文档的内容，该 AI 也能够提供助力，它可以根据用户指定的一个或多个文档，生成文章大纲、撰写新闻稿件、提炼主旨摘要等。

用户在和 AI 的交互过程中，可以插入一篇或多篇已有的文档，作为 AI 内容生成的参考素材，使生成的内容更具备用户已有内容的相关性。如果嫌 AI 话太多，在 WPS AI 中选中「缩短篇幅」，能自动把话改短。如是内容不够，可选择「扩充篇幅」进行扩写。

从金山官方的预告来看，除了文档编辑外，后续 WPS AI 将逐渐整合至“WPS 全家桶”，无论是 Word 文档，Excel 表格还是 PPT，它都将能够提供帮助，有效提升生产力。金山办公表示，WPS AI 将进一步向三个方向迈进，包括 AIGC、阅读理解和问答、人机交互。WPS AI 将逐步开放公测，该名称只是一个代号，最终命名需等待后续官方公布。据悉，WPS AI 在上线后，将成为 WPS 超级会员服务的一部分。

BigBangTransformer[乾元]大模型

公司主体：超对称（北京）科技有限公司

融资阶段：A 轮

一、公司简介

超对称技术公司（Super Symmetry Technologies）成立于 2016 年，专注于开发通用人工智能（AGI）技术，并将 AGI 技术应用于金融、制造、生物医药、教育、科学研究等不同行业。超对称认为具备数据分析能力是实现 AGI 的基础之一。超对称技术公司联合复旦大学计算机学院肖仰华知识工场实验室，浙江大学徐仁军实验室，南开大学和北师大人工智能学院的老师，在基础理论，架构，算法实现三方面推动 AGI 底层算法的研发，构建 AGI 在产业应用的底座。该项研究获得甘肃高台“东数西算”项目和南京江苏软件园在算力基础设施上的支持。

超对称技术公司基于 Transformer 和 Decoder-Only 的架构开发了 BigBang Transformer「乾元」大规模预训练语言模型。2022 年 5 月发布了面向金融行业应用的 BBT-1，2 亿和 10 亿参数金融模型；2023 年 2 月我们发布了 BBT-2，120 亿参数的基础大模型，并在基础模型的基础上研发了金融模型、代码模型、文生图模型。

创始人吴恒魁是前北京雨恒矩阵创始人、CEO，产品有应应、手游头条、引力波等。曾是 Bmob 联合创始人、首席运营官，加州大学戴维斯分校物理 PhD,UCDavisPhDofPhysics.曾参与全球最大暗物质探测实验 LUX,负责粒子碰撞模拟和数据分析，机器学习专家，AloharMobileInc 的数据科学家，MaxenLab 的联合创始人，为客户提供机器学习平台服务和预测分析系统。

CTO 赵红亮，北邮硕士。曾任 VIPKID 研发总监兼首席架构师和 IBM 中国研发中心 Rational 部门高级算法工程师。

超对称技术公司曾获得以太资本和线性资本的投资。

融资日期	轮次	融资金额	投资方
2016-04-24	天使轮	数百万人民币	以太创服（以太资本）（领投） 极客邦创投
2018-01-05	A 轮	千万级人民币	线性资本 Sirius Investment

二、主要模型

BigBangTransformer[乾元]是基于 GPTDecoder-only 架构的大规模预训练模型。2022 年 5 月，超对称技术公司发布大语言模型 BigBangTransformer【乾元】的第一版 BBT-1，10 亿参数预训练语言模型，在中文金融语料上训练而成。

近期，超对称公司再推出 BBT-2，120 亿参数的通用大语言模型，并在 BBT-2 的基础上训出代码，金融，文生图等专业模型。

BigBangTransformer【乾元】12B 大模型的开发基于英伟达的 DGX 算力，超对称公司主导大模型的预训练，并联合复旦大学知识工场实验室完成指令微调和评测工作。超对称技术公司将发布基于 BBT-2 的系列模型（模型 Index 见于 <https://bbt.ssymmetry.com>）

BBT-2-12B-Text: 120 亿参数的中文基础模型

BBT-2.5-13B-Text:130 亿参数的中文+英文双语基础模型

BBT-2-12B-TC-001-SFT 经过指令微调的代码模型，可以进行对话

BBT-2-12B-TF-001 在 120 亿模型上训练的金融模型，解决金融领域任务

BBT-2-12B-Fig: 文生图模型

BBT-2-12B-Science 科学论文模型

（一）BBT-2-12B-Text 通用模型

BigBangTransformer[乾元]是基于 GPTDecoder-only 架构的大规模预训练模型。预训练语料库覆盖 14 个高质量数据源，覆盖书籍、百科、论文、小说、新闻、政策文件、中文博客、社交媒体等多个数据源种类。BBT-2-12B-Text 基于中文 700 亿 tokens 进行预训练，经过指令微调的 BBT-2 基础模型可以回答百科类和日常生活的问题。BBT-2.5-13B-Text 基于中文+英文 2000 亿 tokens 进行预训练。（暂时不开放基础模型的问答对话接口）

模型开源后所有开发者可以：

可以直接调用大模型进行对话

在我们训练的基础上使用自由语料库继续训练

对大模型进行微调训练以满足各种下游任务

（二）BBT-2-12B-Text+Code 代码模型

BBT-TC，是超对称技术公司近期发布的 BBT-2 大模型系列中的代码模型，在百亿基座模型 BBT-2-12B-Text 接续训练代码数据集，通过有监督指令微调（SupervisedFine-Tuning）解锁模型的推理能力。

（三）BBT-2-12B-TF-001 金融模型

超对称公司于 2021 年便开始针对金融投资领域的应用，着手设计和训练了一个大规模参数预训练语言模型 BigBangTransformer 乾元(BBT)，目前发布了 Base 版本 2.2 亿参数和 Large 版本 10 亿参数以及最新的 BBT2，120 亿参数通用模型。BBT 模型的目标是为金融投资建立统一的人工智能算法框架，基于 transformer 构建能融合训练金融投资涉及的不同模态数据的架构。在统一架构的基础上训练大规模参数预训练模型，随着模型参数和训练数据集继续增大，超对称团队有希望开发出在金融领域接近人类智能水平的模型。作为金融领域的基石模型，BBT 模型为所有金融投资，经济分析，商业咨询等场景的深度学习下游任务提供微调服务。金融投资领域有大量从业机构和人员，大厂有财力雇佣算法工程师，小团队却用不起基本的文本抽取算法。BBT 模型作为金融领域的算法基础设施，让所有从业者配备同级别的武器，让全行业站在同一起跑线去竞争更优的投资策略，从而推动金融和经济市场更高效的信息和要素流动。

超对称开发了 11 种已落地应用的下游任务，面向专业金融开发者 API，已获得中国及海外金融机构付费。

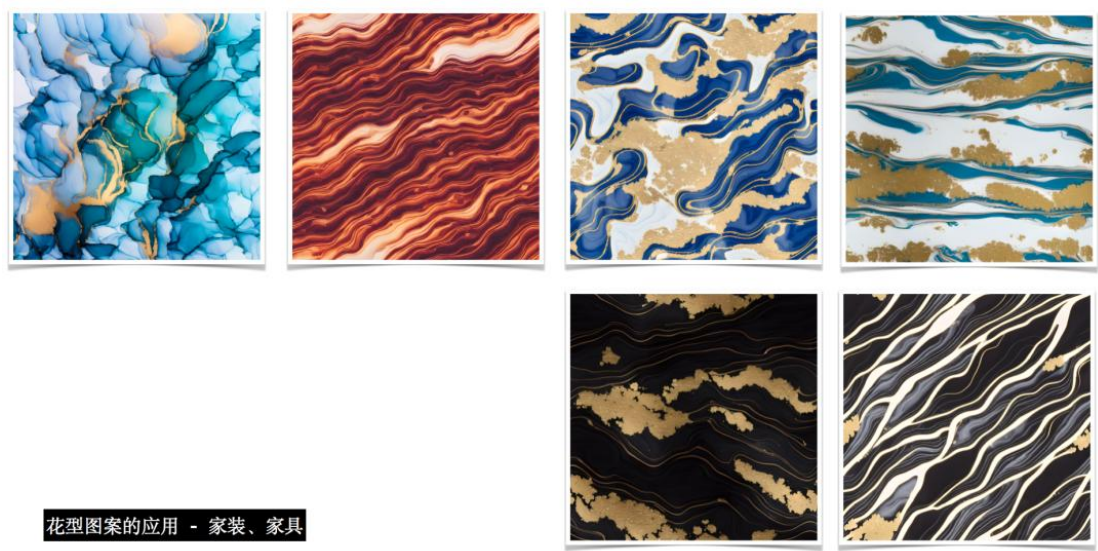
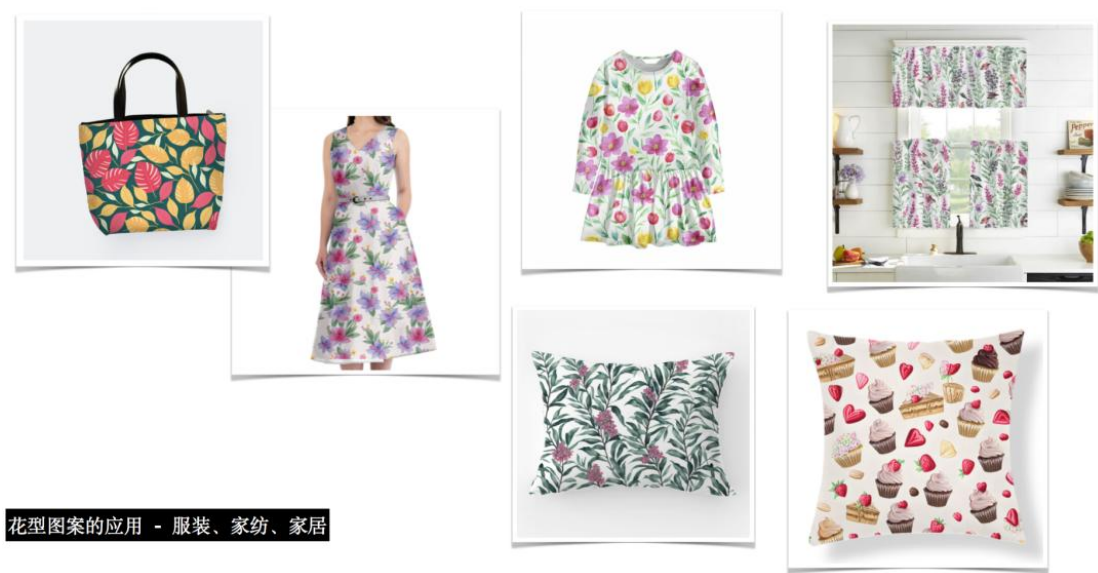
API	接口	描述
社交情绪计算	http://125.74.158.86:2338/sentiment_comment	提取评论文本情绪
新闻情绪计算	http://125.74.158.86:2338/sentiment_news	提取新闻文本情绪
研报情绪计算	http://125.74.158.86:2338/sentiment_report	提取报告文本情绪
公告情绪计算	http://125.74.158.86:2338/sentiment_notice	提取公告文本情绪
金融主体识别	http://125.74.158.86:2338/ner	提取文本主语
金融事件抽取	http://125.74.158.86:2338/event	提取文本事件
短标签抽取	http://125.74.158.86:2338/label	提取文本标签
摘要抽取	http://125.74.158.86:2338/abstract	提取文本摘要
关系抽取	http://125.74.158.86:2338/relation	提取文本关系三元组
因果抽取	http://125.74.158.86:2338/ccks	提取文本因果三元组
负面消息告警	http://125.74.158.86:2338/alert	提取文本风险信息

（四）BBT-2-12B-Image 文生图模型

基于 BBT-2 大语言模型，超对称开发文生图 AIGC 模型 BBT-Image，联合专业图库公司上海慧岳科技开发了面向纺织行业，印刷，广告，游戏等行业客户的应用平台 ai.shenbi.pro.经过专业评测，BBT-Image 在纺织行业上的应用效果比 StableDiffusion 和其他国内大模型明显更优。

BBT-Image 能够生成非常逼真的图像，并且具有可控的样式和风格。在纺织

行业中, BBT-Image 可以通过对不同纺织材料的图像进行学习, 生成具有纹理(可无缝拼接)、色彩(色彩亮度不会受训练样本明暗影响, 为标准化设计用样式)和设计元素(可随意组合元素)的纺织品图案, 提高纺织品设计的效率和创新性。该技术还可用于家装等等需要图案设计的行业。



(五) BBT-2-12B-Science 科学论文模型

BBT-Science 大模型是基于 BBT 大模型在几千万篇科研论文上训练构建的辅助科学发现的大模型, 应用于物理、化学、生物、数学等不同学科的科研知识问题。

为了对科学大模型进行评测, 超对称技术公司与复旦、上海交大、浙大、南航、中山大学、北师大等多所大学合作, 正在号召全球一线的科研人员共同构建

一个最大的科研问题评测数据集 **ResearchQA**。该数据集覆盖数学、物理、化学、生物、地理地质、计算机、电子工程等主流科研领域。科学大模型将成为全球科研能力的底层引擎，带来科研生产力的加速。该数据集直接采集科研领域里前沿的研究课题作为问题，重点考察大模型回答的创新性。

三、主要产品

目前该公司的通用对话、文生图、金融模型均暂未对公众开放，只有代码模型对公众免费开放。超对称的主要产品是面向金融领域 **B** 端用户的，已获得包括中国的百亿量化基金和华尔街知名基金的付费。

清华大学：国内 AI 创业的孵化器

随着 ChatGPT 的不断更新迭代,其在互联网上讨论热度持续火热,使 OpenAI 成为全球人工智能领域的追击目标。这也掀起了国内自然语言处理领域的研究热潮,在“学院派”大模型创业者中,清华大学一骑绝尘,一方面,清华人工智能研究院的团队与多家大企业简历交流与联系,支持了华为、阿里等企业模型的研发,另一方面清华大学人工智能研究院团队自研出多个大模型,同时基于此孵化出了极具发展潜力的创业公司。

「清华大学人工智能研究院」

清华大学人工智能研究院,成立于 2018 年,以人工智能基础理论和基本方法研究为核心,积极推进大跨度的学科交叉融合,积极推进大范围的技术与产业、学校与企业融合。其潜在力量主要有三支:一是知识工程实验室(KEG),学术带头人为李涓子和唐杰;二是自然语言处理与社会人文计算实验室(THUNLP),学术带头人为孙茂松及其学生刘知远;三是交互式人工智能课题组(CoAI),学术带头人为朱小燕及其学生黄民烈。

以下将介绍各实验室、课题组的研究成果及公司孵化情况。

知识工程实验室

清华知识工程实验室成立于 1996 年,致力于网络环境下知识工程理论、方法 and 应用研究。

知识工程实验室的领头人唐杰,是清华大学计算机系最突出的学术代表之一,他聚合北京高校,领衔研发了智源研究院的悟道 1.0 和 2.0 大模型⁶。

2019 年,唐杰与李涓子带领成立了智谱 AI⁷。2022 年,智谱 AI 发布了双语千亿大模型 GLM-130B,并以开源开放的形式供研究机构或个人免费下载使用。

2016 年,唐杰的学生杨植麟,联合创始了 NLP 创业公司——循环智能。该公司主要通过机器学习的方法分析原始的语音文字销售过程和销售转化结果,帮助客户提高销售转化率。2021 年,循环智能与华为云联合推出千亿参数、40TB 训练数据的全球最大中文语言预训练模型盘古。

6 “悟道 2.0”模型的参数规模达到 1.75 万亿,是 GPT-3 的 10 倍,打破了之前由国外预训练模型创造的 1.6 万亿参数记录,是目前中国首个、全球最大的万亿级模型。

7 智谱 AI 公司发展、主要产品及模型、应用案例前文已详细介绍。

自然语言处理与社会人文计算实验室

清华大学自然语言处理与社会人文计算实验室，成立于 20 世纪 70 年代初，是国内最早开展 NLP 研究的科研单位，该实验室在 NLP 领域积累深厚。

2015 年，孙茂松作为负责人带领实验室学生设计研发了人工智能诗歌写作系统九歌⁸，在学界和业界引起广泛关注。2019 年，孙茂松、刘知远团队发布了 ERNIE 语言模型，之后研发了 CPM 模型，是智源研究院“悟道·文源”大模型的前身。

孙茂松、刘知远及其门下毕业的学生，在 NLP 和大模型领域已经孵化出了多家公司。

2017 年，涂存超创立了幂律智能，专注于 NLP 在法律领域的应用。

2022 年 3 月，岂凡超成立了深言科技，致力于以自研中文大模型构建工业级中文信息处理引擎。团队成员研发了 CPM-3 大模型，覆盖续写、填空、缩写、扩写、改写等所有类型的文本生成任务，并且具有原生的可控文本生成能力，能够根据给定的关键词、知识图谱、事件、长度、风格等控制信号生成连贯、流畅的文本。

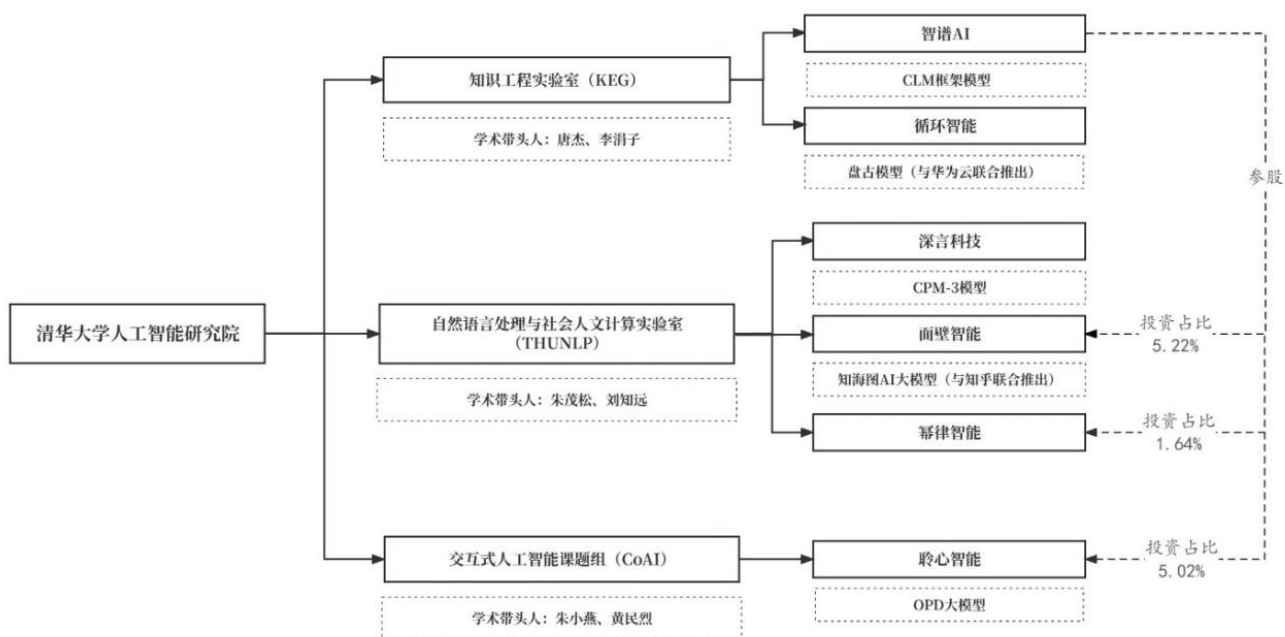
2022 年 8 月，曾国洋创立了面壁智能，专注于大模型加速与应用落地。2023 年 5 月，面壁智能联合知乎发布了最新的大语言模型成果 CPM-Bee10b 大模型，并全面开源。

交互式人工智能课题组

交互式人工智能课题组隶属于清华大学计算机系、清华大学人工智能研究院，交互式人工智能是指通过对话、问答等语言交互方式体现出来的智能行为，是人工智能最具挑战性、最综合性的技术。

2021 年，黄民烈创立了聆心智能，致力于 AI 对话交互，同时侧重游戏、营销、情感三个赛道。聆心智能自研全球技术指标领先的中文对话大模型 OPD，是目前世界上参数规模最大的开源中文对话预训练模型，在自动评测和人工评测中显著优于同类模型。

⁸九歌系统采用最新的深度学习技术，结合多个为诗歌生成专门设计的模型，基于超过 80 万首人类诗人创作的诗歌进行训练学习。区别于其他诗歌生成系统，九歌具有多模态输入、多体裁多风格、人机交互创作模式等特点。



以下将详细介绍上文所提到的深言科技、面壁智能和聆心智能三家公司。

CPM-3 大模型

北京深言科技有限责任公司

融资阶段：A 轮

深言科技是一个文本信息处理平台，致力于用人工智能和自然语言处理技术，为个人和组织系统化重塑文本信息处理全流程，提高信息获取的效率。

深言科技成立于 2022 年 3 月，CEO 岂凡超，本科毕业于清华大学电子工程系，博士毕业于就计算机科学与技术系。

2022 年 8 月，深言科技获得千万级人民币种子轮融资，由英诺天使基金，奇绩创坛，水木清华校友种子基金，智源研究院联合投资。2023 年 1 月，获得 A 轮融资，由红杉种子基金、奇绩创坛、零以创投等机构联合投资。

融资日期	轮次	融资金额	投资方
2022-08-10	天使轮	未透露	奇绩创坛 英诺天使基金
2023-01-13	A 轮	未透露	红杉种子基金 奇绩创坛 零以创投

深言科技团队已发表顶会顶刊论文数百篇，获得各类知识产权数十项，涵盖文本理解、文本生成、知识图谱等多技术方向，拥有自研可控文本生成大模型以及相关数据清洗、处理技术专利。目前，深言科技已开发出世界首个多模式可控文本生成大模型 CPM-3，实现了当今世界上最好的中文文本生成效果，并完成了多个标杆客户的交付。同时，深言科技在微信公众号上的两款体验产品——wantwords 反向词典和 wantquotes 据意查句也受到了用户的广泛好评，已积累数百万用户。

CPM-Bee 大模型

北京面壁智能科技有限责任公司

融资阶段：天使轮

面壁智能是一家人工智能大模型加速与应用落地赋能公司。公司依托开源社区打造大规模预训练模型库与相关工具，加速百亿级以上大模型的训练、微调与推理，降低大模型使用门槛，并进一步推动大模型在人工智能典型场景与领域的应用与落地。

面壁智能成立于 2022 年 8 月，CTO 曾国洋，OpenBMB 开源社区的主要发起者，BMTrain、BMInf 的主要作者，同时是悟道·文源中文预训练模型团队骨干成员，曾获全国信息学竞赛（NOI）金牌。

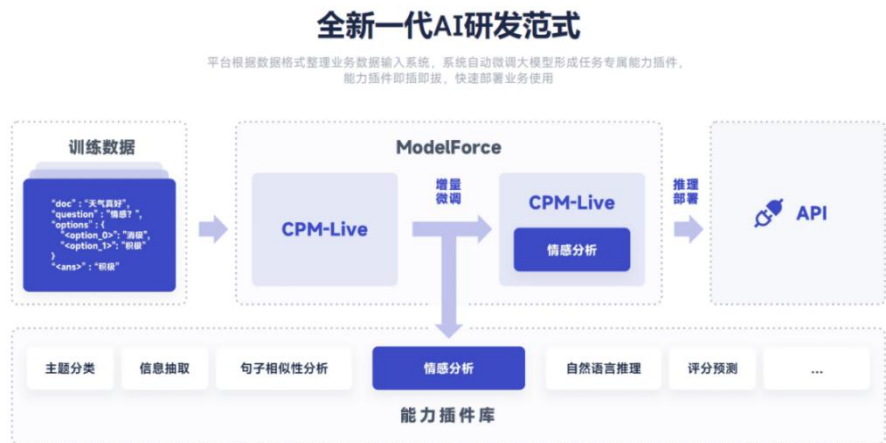
2023 年 4 月，面壁智能获得数千万元天使轮融资，由知乎领投，智谱 AI 联合投资。

融资日期	轮次	融资金额	投资方
2023-04-14	天使轮	数千万人民币	知乎（领投） 智谱AI

目前面壁智能的核心产品是 ModelForce 和 CPM-Bee 大模型。

「ModelForce」

ModelForce 是全流程大模型高效加速平台，内置大模型训练、微调、压缩、推理全流程高效计算工具体系，基于大模型少样本/零样本通用能力，标准化微调方式+零代码微调客户端，可以大幅降低数据标注成本、算力成本、人力成本。



ModelForce 核心技术主要有：高质量数据（自建大规模数据处理 Pipeline、沉淀下 TB 级高质量数据），大模型分布式训练（支持数据并行、模型并行、流水线并行、训练成本最高可节省 90%），大模型高效微调、压缩（独创全参数微调、增量微调、提示微调、性能保持 90%，推理加速十倍），大模型高效推理加速部署（智能显存调度、多机多卡并发，并支持国产芯片适配业界领先的响应、推理速度）。



「 CPM-Bee 」

CPM-Bee 大模型是面壁团队根据多年的大模型训练经验自研的百亿参数预训练语言大模型，模型支持多语言能力和简易结构化输入输出，通过团队自研的高效微调技术，能快速适配各种下游任务，满足各种场景的需求。



CPM-Bee 具有多种功能，包括：生成、问答、翻译、摘要、信息抽取、关系分析、情感分析、主题分类、自然语言推理、句子相似性分类等。

2023 年 5 月 27 日，面壁智能联合知乎在发布了最新的大语言模型成果，面壁智能自主研发的中文基座大模型 CPM-Bee10b 全面开源，该模型从零开始自主训练，基于 Transformer 架构，中英双语表现优秀，拥有百亿量级参数、万亿级高质量语料。在 ZeroCLUE 评测中，CPM-Bee10b 以总分 78.18 分登顶；在英文常识知识推理榜单上，CPM-Bee10b 得到了 67 分的平均分，媲美英文开源模型 LLaMA。此外还推出智能对话类模型产品“面壁露卡”，以及大模型在知乎社区的应用“搜索聚合”即日开启内测。

OPD 大模型

北京聆心智能科技有限公司

融资阶段：Pre-A 轮

聆心智能是一家超拟人大模型研发商，其自主研发的 AI 乌托邦具有可控、可配置、拟人的特点，用户可选择与梅西、马斯克等跨次元跨时空名人聊天，也可与程序员、心理咨询师、算命先生等群体属性的角色交流，还可以定制自己想要的任何角色。

聆心智能成立于 2021 年 11 月，由清华大学黄民烈教授创立，创始团队在国际顶级会议和期刊上发表论文 100 多篇，并获得 10 余次最佳论文或提名奖励。团队成果获得过吴文俊人工智能科技进步奖一等奖，受到国家自然科学基金重点项目、国家杰出青年项目的支持。

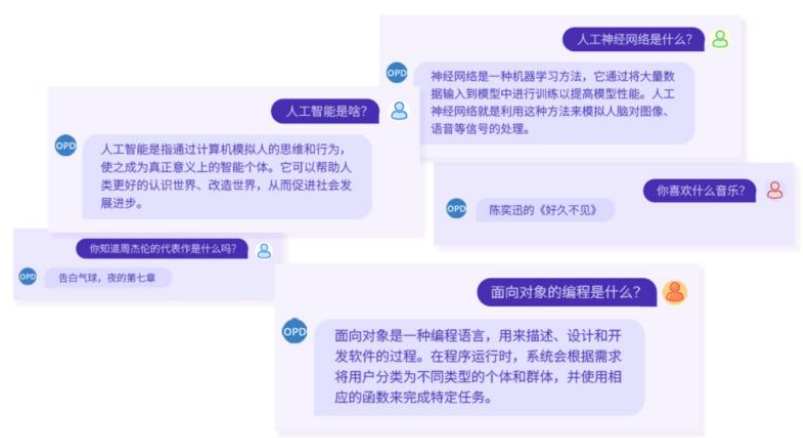
2022 年 12 月，聆心智能宣布完成天使轮融资，本轮融资由连星资本领投，图灵创投、智谱 AI 联合投资。2023 年 2 月，聆心智能完成 pre-A 轮融资，由 SEEFund 无限基金领投。

融资日期	轮次	融资金额	投资方
2022-12-05	天使轮	数千万人民币	图灵创投 连星资本（领投） 智谱AI
2023-02-17	Pre-A轮	未透露	SEE Fund无限基金（领投）

目前聆心智能的核心产品是 OPD 大模型和 AI 乌托邦。

「 OPD 大模型 」

OPD 模型，全称 Open-Domain Pre-trained DialogueModel，是清华大学 CoAI 小组与北京聆心智能科技有限公司联合在中文对话预训练基座模型取得的最新阶段性进展。

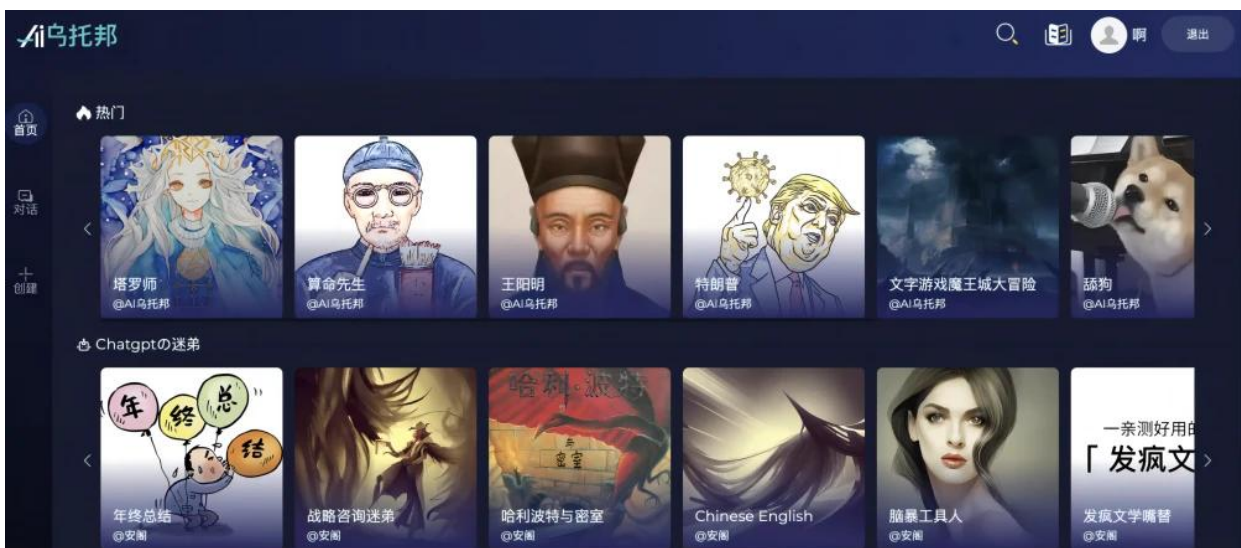


它具有如下优势：

- 大规模：OPD 的模型参数量为 6.3B，是目前世界上规模最大的开源中文对话预训练模型。
- 高性能：通过自动评测和人工评测来全面评估 OPD 的性能，评测结果显示，OPD 兼顾出色的闲聊能力与知识问答能力。得益于此，OPD 的多轮交互能力尤其突出，能够与用户进行多轮、深入的对话交互，性能更优，更受用户偏爱。

「AI 乌托邦」

聆心智能在本月推出了图灵世界的首个产品——“AI 乌托邦”，该系统允许用户快速定制 AI 角色，只需要输入简单的角色描述，就可以召唤出相应人设的 AI，与之进行深度对话和聊天。



在 AI 乌托邦中，用户不局限于“捏”一个自己喜欢的二次元角色，“复活”去世多年的亲人和偶像，和“情感顾问”倾诉内心，和自己的宠物吐吐槽，和品牌 AI 畅聊营销构思。具有非常广阔的商业前景，例如：品牌营销、知识传播、情感陪伴、学习助手、游戏 NPC 的“灵魂”、AIGC 创作工具等。



AI 乌托邦的技术优势有三点：

1、 知识融合

聆心智能能够把产品的具体知识融入到大模型中，比如说某款面膜适合的肤质、使用方式、功效、原料表等，都可以非常自然地融入到具体的对话中。

2、情感识别和表达

聆心智能打造了 32 类情绪的识别系统，目前是业内最全的情绪识别系统；开创性地构建了基于生成式模型的情感对话系统。这对于满足情感需求的应用来说是极强的支撑，比如心理疏导、陪伴、养成。

3、风格化、个性化

用户打造的不同类型 AI 物种或者虚拟人，可以有不同的个性、风格、特质，具有类人的“灵魂”。

南京元宇宙产业协会

在南京市工业和信息化局的指导下,南京市文化投资控股集团有限公司联合江苏原力数字科技股份有限公司、南京硅基智能科技有限公司、南京芯视元电子有限公司、水晶石数字科技有限公司等共同发起成立南京元宇宙产业协会,发挥元宇宙的倍增效应,联合、引导和带动相关企业,共同打造四个平台:一是元宇宙信息共享平台,开展元宇宙产业信息和产业知识共享;二是元宇宙共性技术研发平台,以城市文化数字云平台等数字新型基础设施和服务平台为依托,培育发展元宇宙上下游产业;三是元宇宙企业孵化平台,构建风险共担、收益共享的协同创新共同体;四是元宇宙小巨人加速平台,筹建南京元宇宙产业发展基金,通过项目投资等共同推进南京元宇宙产业发展。

协会之窗

南京水晶石数字科技有限公司

一、企业简介

水晶石数字科技有限公司,是业界领先的数字视觉创意集团,深耕于CG(计算机图像)领域,致力于为不同行业客户提供基于数字图像与影像服务的整体解决方案,尤其是在文旅融合、文博展示、智慧城市、数字孪生等领域基于视觉呈现的元宇宙技术应用方面。

公司通过北京奥运会、上海世博会、伦敦奥运会三幅画卷建立了数字视觉技术应用的高端品牌优势地位,曾荣获2012年度中国文化产业年度领军企业,2017年与北京理工大学等单位合作的研发项目《交互式显示关键技术及应用》获得国家技术发明二等奖。是北京2008年奥运会图像设计服务供应商、上海2010年世博会指定多媒体设计服务商、伦敦2012年奥运会数字图像服务供应商、南京2014年青奥会数字图像服务独家供应商以及国际展览局(BIE)战略合作伙伴、北京2022年冬奥申委数字内容及互动体验展示供应商,为北京2022年冬奥会提供相关的数字视觉技术服务保障工作。举世瞩目的新中国成立70周年庆祝大会和阅兵式以及群众游行活动天安门广场现场巨型屏幕上,也是由水晶石公司提供了全方位的数字视觉内容的创意、制作服务。

在中国共产党建党100周年献礼活动中,水晶石公司先后为中国共产党党史馆建设工程、中国共产党成立100周年大型文艺演出《伟大征程》及中国共产党

成立 100 周年大会提供了全方位的数字视觉技术创意制作及服务保障工作。

二、企业探索

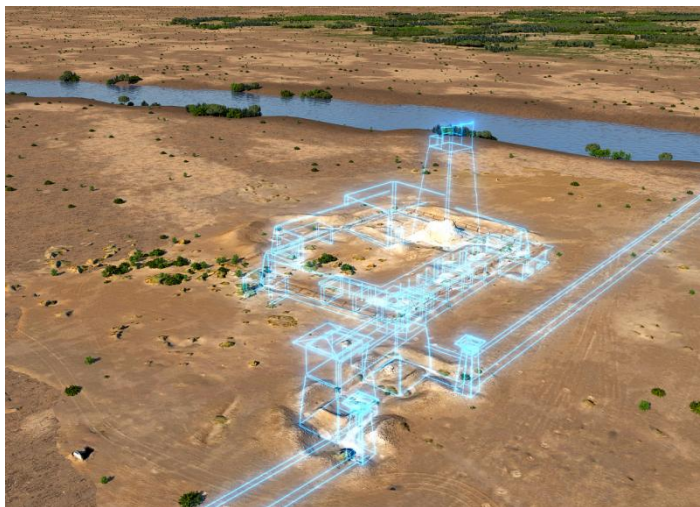
党的二十大报告强调：“在信息社会，要以信息化推进中国式现代化，助力社会主义现代化国家建设。数字文化是数字中国的重要组成部分，实施国家文化数字化战略，是繁荣发展我国文化事业和文化产业、建设社会主义文化强国的重要举措。”

近年来，随着 5G、互联网、大数据、云计算、物联网、区块链、VR/AR、虚拟人、数字藏品、人工智能为代表的新兴数字科技逐步在文化传播领域获得深度应用，并逐渐呈现出数字产业化和产业数字化的双向融合趋势。尤其以元宇宙技术为代表，助力社交兴趣化场景化，内容 UGC/AIGC 化发展，元宇宙以其极具艺术性、体验性、技术性的数字场景，赋予了数字文旅元宇宙体验平台新的发展机遇。

南京水晶石公司长期以来，在基于视觉呈现的元宇宙技术应用方面，深耕于数字视觉创意、沉浸体验、数字孪生技术等领域。聚焦文旅融合、文博展示、虚拟现实、增强现实、智慧城市、互联网+智慧化大数据等业务版块，不断发挥自身优势的同时积累了丰富的实践经验。

（一）虚实联动——基于元宇宙技术的中国长城数字化交互平台

中国长城数字化交互平台，精心规划了长城各时代维度，空间维度，对长城文化遗产进行全面性、立体性、整体性的数字化采集、研究。加强长城遗址保护、传播长城文化、打造长城 IP，通过元宇宙诸多创意理念，搭建长城沿线基于大数据的全生态全新的旅游新服务体系，成为一个有机生长不断融合智慧旅游的文旅新平台。



【图 1-1 肩水金关数字复原】



【图 1-2 长城文物汉画像砖动画演绎】

该平台是对长城文化遗产及相关长城文化进行全面的数字化保护、研究、复原和展示的综合平台，也是中国长城数字可视化大数据平台，致力于推动长城大型遗产的整体性保护与可持续发展、带动长城沿线相关文旅产业、长城相关非物质文化遗产数字化保护与展示、推动中国古代城墙文旅数字化体验，打造城市文旅数字化标杆。



【图 1-3 中国长城数字化交互平台界面呈现】

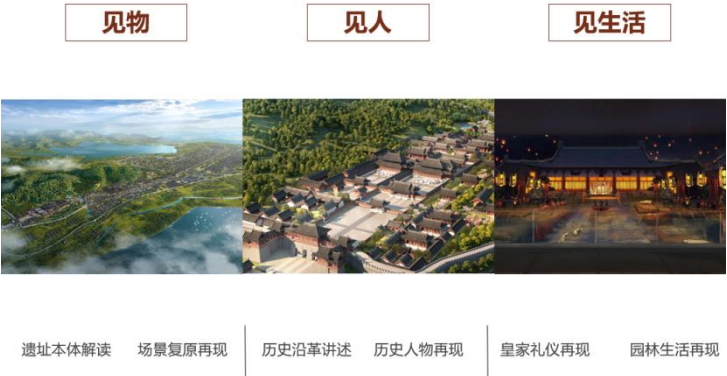
（二）虚实共生——基于数字孪生技术的历史文化遗址、遗迹数字复原

通过基于数字孪生技术的遗址沉浸体验应用研究，形成一整套技术领先的元宇宙数字孪生技术的遗址沉浸体验系统，探究遗址上的古代生活、古代文化和遗址背后的故事。运用元宇宙数字孪生技术将古代遗址、古代生活与当代文明孪生，使得古代文明和遗址、现代生活互为依存，达到一种穿越时空的共生效果，为传播灿烂辉煌的古韵文化、满足社会公众的文化需求作出重要的贡献，也为今后遗址复原展示提供重要的参考。在现代数字创意技术的支撑下，神秘而古老的文明

得以复原再现。采用数字 VR\AR 技术，并根据复原研究成果，三维直观呈现明故宫遗址相关复原形态，实现观众与遗址本体的直接视觉对话。支持线下 AR 互动，捕捉遗址现场影像，叠加三维复原影像，呈现明故宫历史景观及生活情景。

(1) 杭州德寿宫遗址数字化展示

德寿宫遗址公园是国内首个以南宋文化为主题，融合德寿宫遗址保护、展示、科研以及文化休闲旅游功能于一体的南宋文化综合体验平台。基于物联网数据，通过人工智能和大数据分析建立现实物体全面、动态的虚拟镜像；结合 VR 和 AR 技术，实现模型 3D 可视化和沉浸式交互，是国内直接在遗址空间数字化展示的首创。

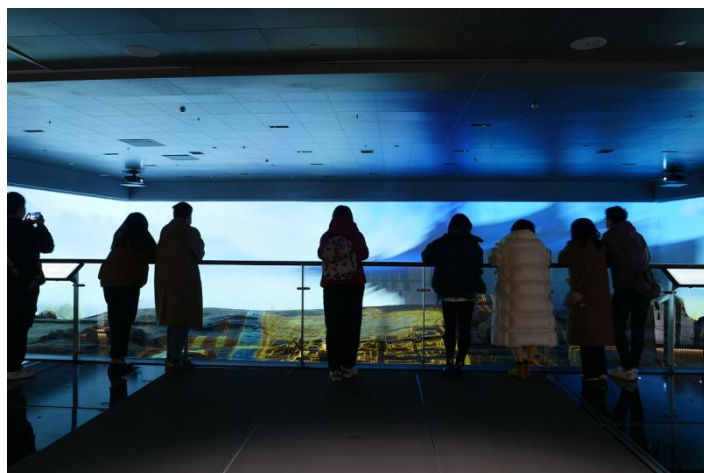


【图 2-1 德寿宫遗址数字化创意理念】

我们通过扎实的复原研究，严谨的数字化复原，采用虚拟现实、增强现实、数字沉浸等技术手段，形成与遗址本体互为联动的数字展示系统，以数字叠加，虚实结合的方式，构建出一个与实体遗址对称的“孪生复原空间”，有效地弥补传统展示形式的局限性，提高遗址的可读性和观赏性，扩充展示文物承载的历史信息。



【图 2-2 《德寿古今》场景体验】



【图 2-3 《德寿古今》场景体验】



【图 2-4 《德寿重华》场景体验】

（2）美成天龙——天龙山石窟数字复原国际合作巡展

天龙山石窟数字复原巡展项目开创了散落在海外文物数据整合展示的国际合作模式，以及利用当代数字科技来创造观众能够高度参与、互动、获得丰富信息的展陈方式。通过 AR、VR、人工智能、语音交互等技术，将天龙山石窟中流失海外的造像，以数字化沉浸体验的方式，展示出天龙山石窟破坏前的原貌，再现天龙山精美绝伦的石窟艺术。

“天龙山石窟数字复原展”是中宣部“中华文化走出去”重点推广项目，成功入选国家文物局推介的 100 个 2020 年度“弘扬优秀传统文化、培育社会主义核心价值观”主题展览、国家文旅部“2020 年度文化和旅游信息化发展典型案例”、国家文物局“海外藏中国文物（石窟寺类）数字复原展示示范项目。”



【图 2-5 天龙山石窟数字复原展·太原站】



【图 2-6 天龙山石窟第 8 窟北壁主尊佛首回归 2021 年登陆央视春晚】



【图 2-7 思维菩萨人工智能语音交互】



【图 2-8 数字长卷场景交互】



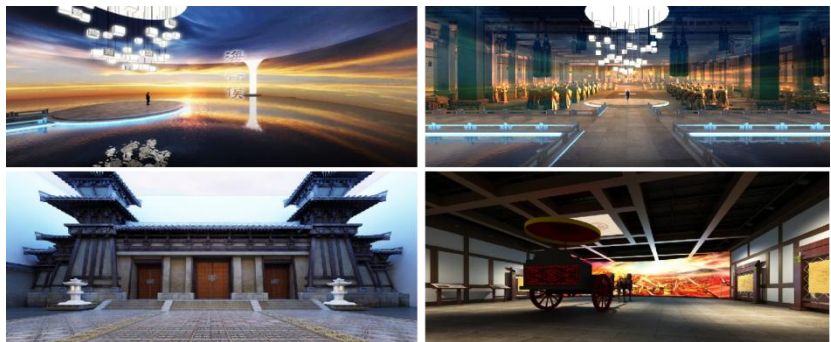
【图 2-9VR 沉浸式体验第 8 窟】



【图 2-10VR 沉浸式体验第 8 窟】

（3）南昌海昏侯遗址数字博物馆

南昌海昏侯线上虚拟展厅，体验者在数字孪生世界中，自由创造数字分身，穿上心仪汉代的虚拟服饰，以不同角色身份，在数字孪生的汉代海昏侯世界中打开新世界之门。虚拟现实 VR、AR 增强现实技术重现了汉代海昏侯国的历史场景，延伸了博物馆的展览空间，打破时空限制，实现了现代文明与汉代文明的对话。立足海昏侯历史文化内容的建设与展示，广泛利用收藏、挖掘、考古的海昏侯历史文化资源。以虚拟展馆的方式延展历史画卷的宽广，以与历史对话穿越为线索铺陈海昏侯历史发展脉络，以时空传承为启迪将现代社会嵌入历史的发展环节，以多媒体数字网络技术和人工智能技术为手段赋予历史资源以永恒的生命力。



【图 2-11 海昏侯线上虚拟馆交互体验】

海昏侯虚拟馆开发，文物三维可视化大数据互动平台，包含以下功能：①动态可变的屏显模式，②支持多人多点同时触摸，③支持图片、文字、语音、三维立体、视频、全景等多种同时显示格式，可视化数据检索系统④能提供可不断叠加、不断生长的可视化知识图谱，⑤支持移动端扫描分享虚拟现实、增强现实功能，⑥支持互联网共享、分享、评论。



【图 2-12 海昏侯文物三维可视化大数据互动体验】

（三）数字孪生——基于 BIM 与 IOT 的智慧文旅运维管理平台

在国家新基建的大潮下，南京水晶石公司整合 5G、物联网、移动互联网、人工智能、云计算、区块链等技术，积极利用自身优势的三维可视化资源，研发出针对智慧城市、智慧园区、智慧消防、智慧博物馆、智慧园林等各业态的数字孪生智慧解决方案。数字孪生是在建筑信息模型和城市三维地理信息系统的基础上，利用物联网技术把物理城市的各种要素进行数字化，然后在网络空间上构造一个与之完全对应的“虚拟城市”，形成物理上的实体城市和信息上的数字城市共生共存、虚实交融。

结合商业文旅项目当下痛点，解决文旅项目中运维、运营、数据、营销、游客体验等问题，构建文旅项目运营、物业管理、设备运维、游客服务、商业联动等一体化的景区运营管理新模式，打造集商业、服务、运维、分析为一体的高效运营新平台。逐渐形成了以 BIM 技术为基础，融合云计算、大数据、移动互联网、AI 等技术应用，面向智慧建筑应用的基于 BIM 与 IOT 的智慧建筑运维管理平台、基于 SaaS 架构的运营管理平台和智慧建筑物联网平台，并形成了面向各个业态的解决方案，参与了国内众多项目，获得客户的高度认可！



【图 3-1 业务架构及应用场景系统图】



【图 3-2 全场景综合应用系统图】

三、未来展望

为更好地把握数字化发展新机遇，落实推动我国数字经济发展的战略要求，由南京水晶石公司联合中国动漫集团“沉浸式交互动漫”文化和旅游部重点实验室、腾讯文旅、南京大学文化创意产业研究中心、南京徐庄高新技术产业开发区等单位，共同打造并致力于元宇宙技术创新研发、应用推广、实践体验、产业交流的合作平台。

本机构以相关元宇宙技术研发为核心，重点围绕元宇宙技术基于文旅融合、文博展示、教育培训、中华文明传承传播等领域的创新应用、实践体验、新兴产业培育、高层次人才培养、国际学术交流与教育合作等方面来建设。充分发挥和展示各方在产业政策、技术研发、品牌推广、创新应用等方面的资源优势，同时整合国内外相关社会资源，大力推动元宇宙技术的创新应用，助力南京市构建元宇宙产业生态，积极推进文化服务数字化，文化产业数字化，文化传播数字化，文化研究数字化，高效助推南京在元宇宙领域相关产业的蓬勃发展，为数字经济标杆城市建设作出贡献！

产业新势力

当 GPT、AIGC 在互联网与现实世界掀起“轩然大波”，数字化的发展是否已经到了奇点？

《文化和科技融合热点观察》将聚焦国内文化数字化的“新势力”代表，他们中有的依托科学技术打造的新空间、新业态、新场景，有的是伴随开源技术诞生的新技术、新产品、新体验，有的是历年文化科技融交会的现场“巨星”，有的是未来融交会将要关注的产业焦点。为此，我们特设“产业新势力”专栏，每期选刊产业发展中的一家新势力代表，从企业介绍、产业亮点、突出优势等多维度对新势力进行解读，展示文化产业发展的“新”力量。

开心麻花新空间落户老门东：《方橙市大侦探》开启好戏连台



一、企业简介

北京开心麻花娱乐文化传媒股份有限公司，成立于 2003 年，总部位于北京。公司以为人民娱乐服务为宗旨，自 2003 年起开始制作、发行舞台剧，凭借强大的喜剧创作实力与精准的推广营销能力，现已成为舞台剧领域具有市场号召力的喜剧品牌。

作为国内舞台剧商业竞争力十足的民营机构，开心麻花已成功推出《小丑爱美丽》《须摩提世界》《夏洛特烦恼》《乌龙山伯爵》等多部票房、口碑双赢的经典作品，屡次创造了中国话剧演出的奇迹，促使“看麻花”成为一种大众文化消费现象。近年来，开心麻花还在致力于拓展音乐剧、网络剧、电影、艺人经纪等板块的业务，以形成一个专业、全方位多元的娱乐产业体系。

2015 年开心麻花落户南京，与江苏省演艺集团成立合资公司——南京苏演开心麻花文化传媒有限公司，成立当年 10 场演出，建立了《江湖学院》和《三只小猪》剧组。随后几年间，南京演出场次持续增长，2019 年演出场次攀升至 71 场，并成功复排建立了《乌龙山伯爵》《婿事待发》《求婚女王》《恋爱吧

人类》《窗前不止明月光》《旋转卡门》《瞎画艺术家》等近十部本地剧组。2021年，成立江苏开心麻花文化传媒有限公司。至2022年，江苏公司在本地的演出场次已达到116场。截止2023年5月，开心麻花已在南京演出近200场，今年计划共有400余场演出持续推出。未来，开心麻花还将扎根南京，辐射苏皖，为本地市民带来更加丰盛的文化盛宴。

此外，开心麻花还在本地不断拓展新业态。今年初在南京老门东打造了一座大型沉浸互动城《方橙市大侦探》，全新的沉浸式演出令人耳目一新，好评不断；同时，开心麻花还在南京筹备一家以戏剧表演为主体的酒吧——Stage One，七月份即将在南京1912街区开业，“跨界融合”的新鲜产物，尚未开业已经吸引了大量的关注。

开心麻花尤其重视本地人才的培养和孵化，与南京艺术学院，南京传媒学院等数所本地高校积极开展合作，每年定期举办“演出进校园”，并安排高校学生分批次参与实习，目前，开心麻花与南艺、南传已签订合作协议，成为其“学生就业实践基地”。

二、沉浸互动剧——《方橙市大侦探》

开心麻花国内首家全景社交解谜互动城《方橙市大侦探》坐落于南京门东历史文化街区边营30—32号开心麻花剧场。建筑外观是一座独栋的传统中式建筑，与门前明城墙的古朴风貌相得益彰。该场地原是一家特色戏曲主题文化餐厅——金陵戏坊，开心麻花进驻后，集中一线主创团队，充分发挥内容优势，将此地打造成为一个超2300平方米、连通地下一层至局部二层的复合型沉浸式演艺新空间。



《方橙市大侦探》讲述了一个光明与黑暗不断较量，正义最终战胜邪恶的故事，其IP来自开心麻花2017年原创大戏《二维码杀手》。在充分打磨舞台剧的

基础上，主创团队对剧情进行了全面提升和二度创作，成功推出这部全景沉浸式演艺新空间作品。

方橙市作为一座巨型的塔楼之城，终日笼罩在电路板和霓虹灯牌构成的数码拟像中。它的内部有着石笋林立的迷宫，有人在这里坚守正义，也有人制造混乱以掩盖真相，各方势力在这里盘根错节。观众进入空间的一刹那，便进入了2000多平的“方橙市”，他们将跟随不同的角色一起破获这个城市的迷案，守卫城市安宁，开启神秘刺激的探案之旅。在巨大城市机器的映衬下，个体的渺小感、孤独感油然而生，当方橙市的赛博朋克风视觉元素与数字化技术相碰撞，会产生出怎样的趣味呢？

观众可以主动选择不同的支线进行参与，摇身一变成为集团股东、集团骨干、实习警员、实习法医、中医学徒或是最疯狂的派对玩家，跟随NPC解锁对应的任务。每条支线故事不同，有5+1条故事线可供选择。“戏精”观众们可主动参与部分主线剧情的建构与推进，可以当感情助攻。观众可在固定的框架内即兴输出，不着痕迹地成为剧的一部分，与演员一对一飙戏互动。

「2000 平空间 VS 盲盒式置景」

《方橙市大侦探》不仅坐拥2000平米占地与3层空间结构，布局则采用“盲盒”形式组装“内部元件”，这样的设计也与该作“全景社交解谜城”的概念不谋而合。

整座方橙市塔楼共有三层空间，依据不同功能被划分为三个区域。位于B1层的封锁区，以红褐色拼接搭配旧金属板质感墙面，充满机械霓虹元素，让市民们一进入场景就能感受到浓浓机械都市风。1F的权威区，主要包括酒吧、警局、医院、中医馆等公共空间，深色墙面体现整个区域“压抑”氛围，各种旧金属框架组合展现各权威机构的“冷清”。2F的舒适区不同于其他区域，会有很多符合角色性格又各有特色的房间，如玛雅小姐的日式榻榻米风卧室、丽萨家宽敞大方的欧式风客厅。当然，还有大教堂、游戏厅、泰式按摩房、任务交易中心等。不同功能区域用特色名称来命名，比如卫生间所在的区域为“方便街”，有探案故事上演的场景便被命名为“探叹街”，极具特色的“盲盒”机制也让观众无需“门牌”即可掌握不同场景的信息。

软装方面，每个“盲盒空间”不仅具备鲜明的辨识度，舞台置景与道具制作同样颇具水准，部分空间还配合声光电与机械机关，以增加情境的真实度。

当然空间内的置景与道具并不全是“花瓶”——墙上的怀表、书架上的几何拼图、警局里的档案甚至中医馆的穴位模型……一切都可能成为解开你手上那道谜题的关键一环。

「人中戏 VS 戏中人」

受过专业表演训练的演员们是《方橙市大侦探》的金字招牌与核心竞争力。每个演员都对自己塑造的角色有着独到而深刻的理解，演员们长达半年的排练时间与不断打磨的演技也让这部作品成为开心麻花厚积薄发的新作品。

首先，演绎方面自不必说，“受过专业训练”与“一定舞台经验”无疑是塑造角色的“双保险”，他们的精彩表现完全可以用“形象鲜活有个性，演技自然不做作”去形容，甚至极易让观众认为每个人物都是“本色出演”。

此外为突出角色特点，方橙市内的不少“市民”都存在一定的“人物设计”，比如堂堂制片人【玛雅】挑选演员时都不忘给自己扇扇子；电影男一号【维克多】为宣传影片竟不惜化身唱跳歌手甚至雇佣观众等等。



更可贵的是除了优秀的演技，演员的“接戏”能力同样一流，不仅与剧情、角色相关的话题能对答如流，就算你用“场外信息”调侃对方，演员照样有能力“接住包袱”。

当然作为一款“沉浸互动剧”，观众体验绝不会止步于“看”，如果选择成为“头号观众”，你所在阵营的任务进度、互动情况等元素均可能让决定结局走向，这种“蝴蝶效应”的既视感无疑也赋予了观众更高的参与度。

《方橙市大侦探》作为一款集无座跑图、剧情推理、机械解谜、真人近景表演等多元元素于一体的话剧演出，不仅赋予了观众更为丰富的观剧体验，其独特的盲盒空间和极具信念感的演员亦给了参与者“一秒入戏”的底气与多刷的动力。

另一方面“开心麻花”的品牌加持自然也在无形中拓宽了受众群体，而能够容

纳更多参与者的设计或许也真的让“方橙市”具备了社交属性和更具潜力的商业价值。



三、开心麻花 x 南京老门东 2023 喜剧生活节

5月24日上午，开心麻花和南京老门东联合发起和共同举办的2023沉浸式喜剧生活节开幕式在位于老门东边营30、32号的开心麻花方橙市沉浸剧场举行，由此拉开了本届喜剧生活节的帷幕。这是开心麻花首个与景区联动推出的沉浸式喜剧生活节，特点是文旅融合，亮点是青年参与，主题是“‘花’式演出，沉浸街区”。



此次开心麻花共计推出近30场多元演出，为南京市民和各界游客带来沉浸

式的戏剧体验和无限欢乐。除了《方橙市大侦探》的演出之外，还安排了诸多精彩演出和文化体验活动。在门东街区 16 艺术空间，有《方橙市大侦探》粉丝见面会；在银兴菲林影城，推出了开心麻花独角音乐剧《求婚女王》、麻花即兴新喜剧和开心麻花脱口秀等 8 场演出；在门东古色古香的中式庭院小东园，组织举办了 5 场开心麻花自主研发的剧本杀《绿水青山带笑颜》，此外还在小东园一层举办青年戏剧沙龙，邀请了南京艺术学院、南京传媒学院、南京师范大学、南京农业大学、东南大学成贤学院等 5 所本地艺术高校的戏剧社团，轮番上演多部精彩剧目片段，展现当代青年的艺术态度与思考。

在 5 月 24 日至 28 日的活动期间，开心麻花和门东历史文化街区还联动吸引了特色各异的文化市集参与其中。戏剧演出与特色活动在街区不同空间内交叠共生，让喜剧生活节的内容更具丰富性、参与性和互动性。

目前开心麻花在南京的战略布局已经非常清晰，除了在秦淮区老门东成功落位《方橙市大侦探》项目之外，在玄武区 1912 街区打造的戏剧酒吧将于 2023 年 6 月开幕，在建邺区则将推出多功能黑匣子剧场以及麻花最新的文商旅业态“花花世界”。

图表附录 1：

公司主体	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		NLP
	GPT4	语言图像理解与语言生成	约1万亿	多模态
	CLIP&DALIE	图像生成、跨模态检索		NLP
	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	5300亿	NLP
Stability AI	Stable Diffusion	语言理解与图像生成		多模态

（国外主流大模型企业图表）

图表附录 2：国内主流大模型企业图表（信息截止于 2023 年 5.31 日）

企业		大模型	参数量	细分领域				市场推广进展	
				NLP	CV	多模态	行业	2B应用	2C应用
国内大厂	百度	文心	2600亿	✓	✓	✓	生物、电力、金融、航天等	600+企业接入API	文心一言
	阿里	通义	10万亿	✓	✓	✓	电商、医疗、娱乐等	钉钉、天猫接入	通义千问
	腾讯	混元	万亿	✓	✓	✓	—	广告、游戏等腾讯内部业务	混元助手（计划构建）
	华为	盘古	万亿	✓	✓	✓	气象、药物研发、能源、工业、医疗	在矿山、气象领域有标杆案例	—
	京东	言犀	千亿	✓	—	—	—	智能客服（计划金融和零售领域）	—
	商汤	日日新	千亿	✓	✓	✓	医疗、代码等	秒画SenseMirage 格物（3D内容生成平台） 明眸（数据标注平台） 琼宇（3D内容生成平台）	商量
	昆仑万维	天工	千亿	✓	—	✓	音乐、代码等	天工巧绘（图像AI） 天工乐府（音乐AI） 天工妙笔（文本AI） 天工智码（编程AI）	—
	科大讯飞	星火	—	✓	✓	✓	教育、办公、汽车、数字员工、AI虚拟人	教育、医疗、智慧城市	AI学习机、讯飞星火
	网易伏羲	玉言	110亿	✓	—	—	—	—	LOFTER-AI 写文 傀儡戏-互动式文字游戏
		玉知	9亿	—	—	✓	—	网易新闻、网易云音乐	—
国内初创公司	智谱华章	GLM	千亿	✓	✓	✓	代码	Hubble哈勃	ChatGLM、写作蛙、小呆对话等
	澜舟科技	孟子	十亿	✓	—	✓	金融	金融、营销、文化创意行业	熊猫小说家 澜舟论文助写
	西湖心辰	心辰	千亿	✓	✓	✓	服装设计	小天（AI辅助心理咨询产品） 盗梦日记（国产AI作画工具） Friday（文本生成产品）	—
	名之梦	MIMO	—	✓	—	—	—	Glow（AI虚拟聊天社交软件）Inspo（生成式对话AI）	—
	超对称	乾元	120亿	✓	✓	✓	金融、代码、科学论文	金融、制造、生物医药、教育、科学研究	—
	循环智能	语言预训练模型盘古（与华为云合作）	千亿	✓	—	—	金融、地产、汽车、教育培训	销售会话洞察、客户会话洞察、智能销售助手、智能质检、智慧双录助手、企业NLP平台、ASR自训练平台	—
	幂律智能	法律中文预训练模型OpenCLAP	—	—	—	✓	法律	MeBox（法律知识管理协作系统） MeCheck（智能合同审查产品） MeFlow（合同生命周期管理） Powerlaw-AI	—
	深言科技	CPM-3	—	✓	—	—	—	—	antwords反向词典 Wantquotes据意查句系统
	面壁智能	CPM-Bee	百亿	✓	—	—	法律、汽车、家电、传媒等行业	智能写作、办公、客服、助理	面壁露卡
	聆心智能	OPD	63亿	✓	—	—	游戏、营销、情感	—	AI乌托邦（对话机器人）
	竹间智能	竹间	70-130亿	✓	—	✓	金融、医疗、制造	对话机器人、营销助手、多模态情感数字人、智能洞察、智能知识库	聊天机器人
	出门问问	序列猴子	百亿	✓	✓	✓	—	企业咨询、教育培训、出海产品介绍、口播短视频、直播等	奇妙文（AI写作平台） 言之画（AI绘画平台） 魔音工坊（AI配音平台） 奈特元（数字人视频与直播平台）
	中科院	紫东太初	千亿	—	—	✓	—	手语教学、法律咨询、交通出行、广电、医疗机器人、医学影像判读等数十个行业场景	—
高校、实验室	复旦大学	Moss	百亿	✓	—	—	—	—	—
	北京智源研究院	悟道 2.0	1.75万亿	✓	—	✓	—	—	诗词创作、做对联、文本摘要、人设问答、绘画等

附录

生成算法是 AIGC 技术的重要方面，生成算法的不断发展可以提高内容创作的效率和便捷性，降低成本和门槛，同时丰富内容创作的多样性和个性化。以下是主流生成模型介绍。

模型	提出时间	模型描述
变分自动编码 VAE	2014 年	基于变分下界约束得到的 Encoder-Decoder 模型对。
生成对抗网络 GAN	2014 年	基于对抗的 Generator-Discriminator 模型对。
基于流的生成模型	2015 年	学习一个非线性双射转换(bijective transformation)，其将训练数据映射到另一个空间，在该空间上分布是可以因子化的整个模型架构依靠直接最大化 log-likelihood 来完成。
扩散模型	2015 年	扩散模型有两个过程，分别为扩散过程和逆扩散过程。在前向扩散阶段对图像逐步施加噪声，直至图像被破坏变成完全的高斯噪声，然后在逆向阶段学习从高斯噪声还原为原始图像的过程。经过训练，该模型可以应用这些去噪方法，从随机输入中合成新的“干净”数据。
Transformer 模型	2017 年	一种基于自注意力机制的神经网络模型，最初用来完成不同语言之间的文本翻译任务主体包含 Encoder 和 Decoder 部分，分别负责对源语言文本进行编码和将编码信息转换为目标语言文本。
神经辐射场	2020 年	它提出了一种从一组输入图像中优化连续 5D 神经辐射场的表示（任何连续位置的体积密度和视角相关颜色）的方法，要解决的问题就是给定一些拍摄的图，如何生成新的视角下的图。
CLIP 模型	2021 年	1、进行自然语言理解和计算机视觉分析 2、使用已经标记好的“文字—图像”训练数据。一方面对文字进行模型训练。一方面对另一个模型的训练，不断调整两个模型的内部参数，使得模型分别输出的文字特征和图像特征值并确认匹配。

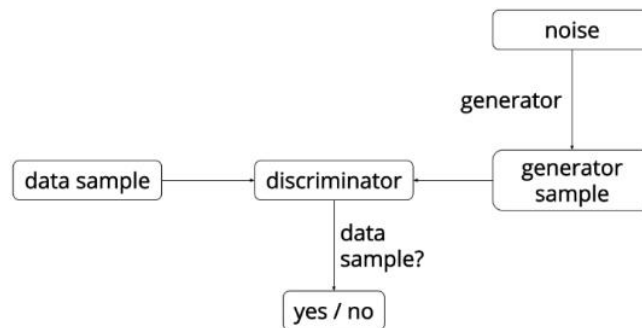
目前来说，AIGC 生成算法主流的有生成对抗网络 GAN 和扩散模型 DiffusionModel，以及用于匹配文本和图像的神经网络模型 CLIP。

（一）生成对抗网络 GAN

GAN（Generative Adversarial Network），是一种深度神经网络架构，由一个生成网络和一个判别网络组成。生成网络产生“假”数据，并试图欺骗判别网络；判别网络对生成数据进行真伪鉴别，试图正确识别所有“假”数据。在训练迭代的过程中，两个网络持续地进化和对抗，直到达到平衡状态，判别网络无法再识别“假”数据，训练结束。

1、工作原理

结合下图整体模型所示，再以生成图片作为例子具体说明：有两个网络，G（Generator）和 D（Discriminator）。G 是一个生成图片的网络，它接收一个随机的噪声 z ，通过这个噪声生成图片，记作 $G(z)$ 。D 是一个判别网络，判别一张图片是不是“真实的”。它的输入是 x ， x 代表一张图片，输出 $D(x)$ 代表 x 为真实图片的概率，如果为 1，就代表 100% 是真实的图片，而输出为 0，就代表不可能是真实的图片。在训练过程中，生成网络的目标就是尽量生成真实的图片去欺骗判别网络 D。而网络 D 的目标就是尽量把网络 G 生成的图片和真实的图片分别开来。这样，G 和 D 构成了一个动态的“博弈过程”。



最后博弈的结果是：在最理想的状态下 G 可以生成足以“以假乱真”的图片 $G(z)$ 。对于 D 来说，它难以判定 G 生成的图片究竟是不是真实的，因此 $D(G(z))=0.5$ 。这样目的就达成了：我们得到了一个生成式的模型 G，它可以用来生成图片。

2、特点

GAN 是很多 AIGC 的基础框架，但是 GAN 有三个不足。

- （1）对输出结果的控制力较弱，容易产生随机图像。
- （2）生成的图像分辨率较低。
- （3）由于 GAN 需要用判别器来判断生产的图像是否与其他图像属于同一类别，导致生成的图像是对现有作品的模仿而非创新。

（二）扩散模型 Diffusion Model

扩散模型是生成模型，用于生成与训练数据相似的数据。

1、工作原理

扩散模型的工作原理是通过连续添加高斯噪声来破坏训练数据，然后通过反转这个噪声过程来学习恢复数据。训练后，我们可以通过简单地将随机采样的噪

声传递给学习的去噪过程来生成数据。简言之，在 AI 训练阶段，我们将数据集中上亿组图文对进行训练，提取特征值；生产过程中，通过添加文字描述，引入不同的特征值进行去噪，从而生产 AI 理解下的内容作品。

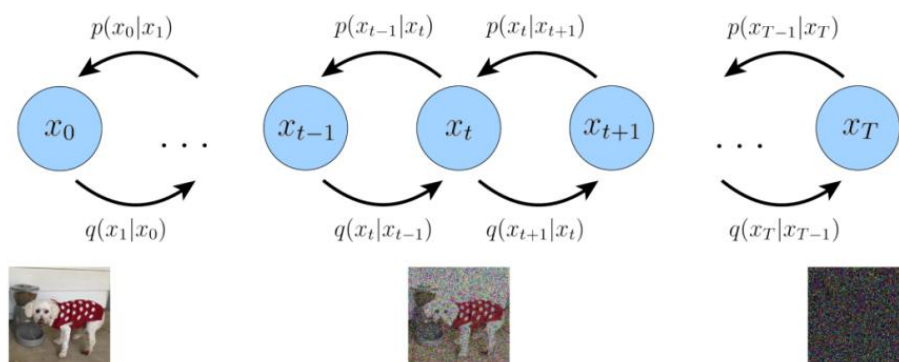


Figure 3: A visual representation of a Variational Diffusion Model; x_0 represents true data observations such as natural images, x_T represents pure Gaussian noise, and x_t is an intermediate noisy version of x_0 . Each $q(x_t|x_{t-1})$ is modeled as a Gaussian distribution that uses the output of the previous state as its mean.

一个标准扩散模型有两个主要的过程域：正向扩散和反向扩散。扩正向阶段对图像逐步施加噪声，直至图像被破坏变成完全的高斯噪声，然后在反向阶段学习从高斯噪声还原为原始图像的过程。

（1）正向扩散：在图像中添加噪声。

方法：正向扩散过程逐步对输入图像 x_0 加入高斯噪声，一共有 T 步。该过程将产生一系列噪声图像样本 $x_1, \dots, x_T$ 。当 $T \rightarrow \infty$ 时，最终的结果将变成一张完全包含噪声的图像。此外建立一个封闭形式的公式，可以在任何时间步骤直接对 x 进行采样，使得向前的过程更快。

（2）反向扩散：去除图像中的噪声。

方法：使用 U-Net⁹模型进行反复训练，由于需要迭代得向 U-Net 提供完整尺寸的图像获得最终结果。

2、特点

扩散模型有两个特点：一方面，给图像增加高斯噪声，通过破坏训练数据来学习，然后找出如何逆转这种噪声过程以恢复原始图像。经过训练，该模型可以从随机输入中合成新的数据。另一方面，稳定扩散把模型的计算空间从像素空间经过数学变换，降维到一个可能性空间（Latent Space）的低维空间里，这一转化大幅降低了计算量和计算时间，使得模型训练效率大大提高。这算法模式的创新直接推动了 AIGC 技术的突破性进展。

⁹U-Net 是一个语义分割模型，利用卷积进行下采样，之后提取出一层又一层的特征，利用这些特征再进行上采样，最后得出一个每个像素点对应其种类的图像。

(三) CLIP 模型

CLIP, 全称 Contrastive Language Image Pre-training, 是 OpenAI 在 2021 年初发布的用于匹配文本和图像的神经网络模型, 是近年来在多模态研究领域的杰出成果。它一方面对文字进行语言分析, 另一方面对图形进行视觉分析, 不断调整两个模型内部参数, 达到文字和图像高度匹配的效果。

1、工作原理

CLIP 的训练数据是文本—图像对: 一张图像和它对应的文本描述, 希望通过对比学习, 模型能够学习到文本—图像对的匹配关系。如下图所示, CLIP 包括两个模型: TextEncoder 和 ImageEncoder, 其中 Text Encoder 用来提取文本的特征, 可以采用 NLP 中常用的 texttransformer 模型; 而 ImageEncoder 用来提取图像的特征, 可以采用常用 CNN 模型或者 visiontransformer。

2、应用

(1) zero-shot 检测

CLIP 可以应用在目标检测任务上, 实现 zero-shot 检测, 即检测训练数据集没有包含的类别, 比如谷歌提出的 ViLD 基于 CLIP 实现了开放词汇的物体检测, 其基本思路和 zero-shot 分类相似, 只不过是用文本特征和 ROI 特征来计算相似度。

(2) 图像检索

基于文本来搜索图像是 CLIP 最能直接实现的一个应用, 其实 CLIP 也是作为 DALL-E 的排序模型, 即从生成的图像中选择和文本相关性较高的。

(3) 视频理解

CLIP 是基于文本—图像对来做的, 但是它可以扩展到文本—视频, 比如 Video CLIP 就是将 CLIP 应用在视频领域来实现一些 zero-shot 视频理解任务。

(4) 图像编辑

CLIP 可以用在指导图像编辑任务上, HairCLIP 用 CLIP 来定制化修改发型。

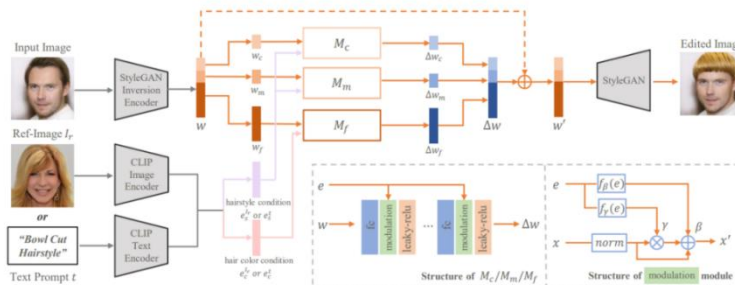


Figure 2. The overview of our framework, here we show an example with hairstyle description text and hair color reference image as conditional inputs. Our framework supports to accomplish the corresponding hair editing according to the given reference images and texts, where images, texts are encoded by CLIP's image encoder, text encoder to 512-dimensional vectors as conditional inputs for the hair mapper, respectively. Only three sub hair mappers are trainable, where M_c and M_m take the hairstyle conditional input e_s and M_f takes the hair color conditional input e_c .

(5) 图像生成

CLIP 还可以应用在图像生成上，比如 StyleCLIP 这篇工作用 CLIP 实现了文

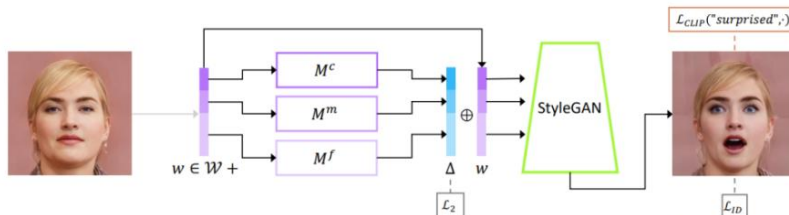


Figure 2. The architecture of our text-guided mapper (using the text prompt “surprised”, in this example). The source image (left) is inverted into a latent code w . Three separate mapping functions are trained to generate residuals (in blue) that are added to w to yield the target code, from which a pre-trained StyleGAN (in green) generates an image (right), assessed by the CLIP and identity losses.

本引导的 Style GAN。

CLIP-GEN 基于 CLIP 来训练文本生成图像模型，训练无需直接采用任何文本数据。过程具体是：预训练 VQGAN，能够把图像表示成一个图像 token 下的离散序列；训练 conditiontransformer，学习如何把图像的 CLIPembedding 映射到图像标记；由于在 CLIP 中，文字和图像共享同一个嵌入空间，我们可以直接把文字的 CLIPembedding 作为 conditiontransformer 的输入映射到图像标记上，然后通过 VQGAN 的 decoder 来生成图片。

(6) 自监督学习

华为的工作 MVP 采用 CLIP 来进行视觉自监督训练：将 BEiT 中的 dVAE 换成了 CLIP，在效果上超过 BEiT 和 MAE，比肩 PeCo。

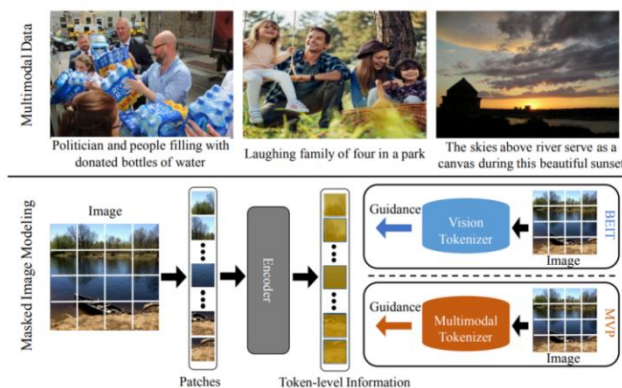


Fig. 1. The motivation of our MVP. The image-text pairs of CC3M [27] are shown at the top. For each case, the caption can depict some important semantic contents. The pipeline of masked image modeling is presented at the bottom, and our MVP simply replaces the vision tokenizer in BEiT with the multimodal tokenizer.

(7) VL 任务

CLIP 本身就是多模态模型，所以它也可以用在用图像—文本多模态任务，如图像描述（imagecaption）和视觉问答（VisualQuestionAnswering）。



中国（南京）文化和科技融合 成果展览交易会

CHINA NANJING

Culture & Technology Integration Exhibition Achievement Fair

中国（南京）文化和科技融合成果展览交易会是全国首个聚焦文化和科技融合的专题性展览交易平台，每年在南京举办。展会由江苏省人民政府指导，江苏省委宣传部、江苏省科学技术厅、江苏省委网信办、江苏省文化和旅游厅、江苏省广播电视局、南京市人民政府联合主办，南京市委宣传部、南京市科学技术局、南京市委网信办、南京市文化和旅游局、南京市文化投资控股集团联合承办，围绕“新时代 新体系 新体验”主题，聚焦广播影视、动漫游戏、演艺娱乐、文化旅游等行业产业数字化转型需求，展示和交易新技术、新产品、新一代解决方案，树立文化科技融合的“风向标”，展示数字文化产业的“创新源”，打造产业数字化的“指南针”。展会获得 2020 年、2021 年中宣部《文化体制改革和发展工作简报》专报推广，评为“2019-2020”年度全国品牌影响力会展项目。

联系人：缪岑岑

联系电话：13270888656



紫金文化科技融合创新研究院

Purple Academy Of Culture And Technological Integration& Innovation

紫金文化科技融合创新研究院是在国家文化科技创新工程专家组、江苏省委宣传部指导下，由南京市文投集团、紫金文创研究院牵头发起的全国首家文化科技融合创新研究平台。研究院下设行业研究中心、创新联合体中心、产业金融中心、交流展示中心，成立由国家文化科技创新工程专家组组长张树武教授领衔、全国高校专家学者和龙头企业代表构成的专家组，以开放性眼光打造全国文化科技融合高端智库、新型研发机构孵化苗圃、新兴文化业态培育基地、龙头文化企业加速平台。

联系人：黄庭

联系电话：13912992856



南京市文化投資控股集團

Nanjing Cultural Investment Holding Group

南京市文化投资控股集团成立于 2009 年 9 月，是南京市委、市政府直属的大型国有文化企业。十余年来，南京市文投集团始终秉承“让城市文化生活更美好”的文投使命，始终坚守“国有文化资产管理运营主体、重大文化项目投资建设主体、新兴文化业态投资主体、美好文化产品示范供给主体、城市文化 IP 孵化培育主体”五大主体定位，聚焦建设互联网文娱、十竹斋艺术、文化综合体、文化金融投资、文化科技创意五大产业板块，共筑“扎根南京、辐射南京都市圈、示范江苏、走向全国的城市文化生活示范运营商”的企业愿景，致力成为全国文化龙头企业三十强和全国国有文投集团前三强。目前集团管理资产总量超百亿元，营业收入和利润每年保持 20% 增速，实现了社会效益和经济效益双丰收；2019 年-2021 年，连续三年荣膺“全国文化企业 30 强”提名单位。

联系人：江浩

联系电话：13851814061



报：中央文化体制改革和发展工作领导小组办公室、科学技术部
办公厅、文化和旅游部办公厅

送：国家文化和科技融合示范基地所在地各省、自治区、直辖市
宣传部办公室、科学技术厅办公室、文化和旅游厅办公室

发：国家文化和科技融合示范基地所在地各市宣传部办公室、科
学技术局办公室、文化和旅游局办公室

中国（南京）文化和科技融合成果展览交易会组委会
紫金文化科技融合创新研究院

2023 年 06 月 01 日发
