开 栏 语

为了更好地服务校友职业发展，服务学校科技成果转化，服务社会经济发展，清华校友总会于2016年开始举办清华校友三创大赛。

开赛8年以来，三创大赛已成为全方位、多层次、无间歇地支持校友、师生以及其他社会各界人士“创意创新创业”的重要平台；为有创新精神和创业能力的校友、师生及社会各界人士提供展示、交流、融资和成长的舞台；为愿意支持和投资创新创业项目的机构提供参与平台；为能够服务国家和地方产业升级及经济发展的优秀创新创业项目建立落地通道；为清华校友创新创业提供更大的势能和动能，打造清华创业生态圈服务国家和地方经济社会发展的支点。

经过8年积累，平台已吸引超过5000个项目参加，聘请了700多位创业导师参与，并与170多家投资机构、上百名天使投资人、70多个地方政府招商机构保持密切联系与合作。

2019年6月，三创大赛入选全国双创示范基地创新创业百佳案例。据不完全统计，2021年有40个三创项目获得融资，总融资额超过35亿。2022年有39个三创项目获得融资总融资额超过46亿。

这每一个优秀的项目背后，都有一位掌握核心技术且数十年如一日苦心钻研的掌舵人，也都有一段从筚路蓝缕到终于守得云开见月明的艰辛创业故事。本栏目将为读者重点呈现三创大赛当中涌现出的优秀企业和企业家们，以期让更多读者了解这些企业发展的故事及这些在路上的企业发展更深层次的需求。

张鹏：站在时代潮头，拥抱大模型新可能

特约记者 黄乐媛

8月底，首批通过《生成式人工智能服务管理暂行办法》备案的大模型产品已经公布并正式上线，智谱AI的首款生成式AI助手——智谱清言榜上有名。

“大模型”技术正是智谱清言的“灵魂”。简单来说，“大模型就是神经网络预训练模型，包括初始的语言模型以及衍生的预训练多模态模型，它具有语言学习能力、图像及视频的理解和生成能力等。”

立足于“大模型”，这家脱胎于清华实验室的人工智能企业正怀抱着更宏大的使命，奔走在时代的潮头之上。

张 鹏

清华大学计算机系1998级本科、2002级硕士、2018级创新领军工程博士

从清华园到中关村东路

回忆起在清华园读书的点点滴滴，张鹏最怀念的，就是实验室中的同窗情谊。年轻人们沉浸在国际顶尖的科技中，双眼因同

样的情怀而闪亮——希望学术成果可以走出实验室，为国家、为社会解决实际的问题。在这样的氛围中，智谱 AI 正悄然孕育着。

2006 年，AMiner 平台正式诞生于清华大学计算机系知识工程实验室中。2013 年，平台的商业化应用提上了日程。2018 年，国家部委发文鼓励科研人员将知识成果转化落地，并提出了指导意见，这为清华大学的科研人员们提供了新的思路，也鼓励他们做出尝试。张鹏和创始团队希望 AMiner 能够在他们手里发挥出更大的潜力。

2019 年 6 月，智谱 AI 正式诞生。公司成立后，许多曾经在清华学习过的同学选择加入，有人甚至为此辞掉了国外顶级公司的工作。

在 2020 年的清华校友三创大赛中，智谱 AI 获得了 TMT/AI 大数据全球总决赛成长组第一名。通过大赛，团队与许多政界、投资界、创业界的清华校友产生了联系，并坚定了实现“让机器像人一样思考”的目标。

从阵痛到飞跃

回忆起多年来创业的历程，张鹏对阵痛和挫折记忆犹新。“尽管回头去看，这件事情还蛮简单的，但是这个探索的过程是非常艰苦的。”

“大模型”面向认知域，应用场景十分广泛。然而，“大模型”技术门槛高，其训练需要专业团队提供大量的算力支持，成本也很高，个人和团体通常难以负担。因此，将训练好的“大模型”作为服务推出，可以降低其使用成本，让科研成果惠及更多的企业和团队。

起初，“大模型”的研发者普遍追求数量，参数由千亿甚至到万亿，然而智谱 AI 则将目光放在算法的优化上，通过训练让它的性能更高效，同参数规模达成一定的平衡。限制规模的好处在于模型投入使用的时候对算力的消耗更小，即使是算力有限的场景，仍然可以成功使用大模型，这样就做到了让“大模型”更具普适性。

2022 年，智谱 AI 联合清华大学打造了高精度双语千亿模型 GLM-130B，构建了高精度通用知识图谱，形成数据与知识双轮驱动的认知引擎。在训练 GLM-130B 时，智谱的理想是世界上任何一个人都可以免费下载千亿模型，并在一台低配的 GPU 服务器上就可以使用它。

在这期间，团队面临两个核心问题：一是缺乏高质量的预训练算法，针对双语的高质量预训练算法还有待验证和提升。二是缺乏快速推理方法，快速推理方法是保证模型能在低配 GPU 服务器上运行起来的基础，也是让每个人都能用得上千亿大模型的关键。

对于预训练模型架构算法，团队联合清华大学于 2021 年提出 了 GLM（General Language Model）算法框架，其在多个任务上表现出了不俗的性能。若 GPT 的原理可以被比作“根据上文做续写”，那么 GLM 的依据则从上文扩充到上下文，并可以同时完成续写和填空。理论上，GLM 的训练效率会比 GPT 更高，也能理解更复杂的场景。经过几轮激烈的争论，团队最终决定训练一个 1300 亿参数的 GLM 模型。一来千亿稠密模型能保证高精度，另一方面这个规模还可以在一台 A100 服务器上就进行单机推理。整个训练过程横跨两个月，在此期间，团队开始考虑训练完成后的推理解决方案，并在一台 V100（32G * 8）服务器上实现了合理速度的 130B 模型推理。

在训练过程中，团队遇到了很多挑战，预训练一个高精度的千亿模型与训练百亿模型完全不同——频繁的随机硬件故障、模型梯度爆炸、算法中意外的过多内存使用、新的 Megatron 和 DeepSpeed 框架中 3D 流水线的调试、无法从优化器状态中恢复、机器间 TCP 拥塞，以及许多许多意外的 “bug”，最终这些问题

AI 提效助手已经具备辅助工作的丰富能力

被一一攻克。

随后，智谱 AI 将 GLM-130B 模型开源，放到了 GitHub 上，让模型更快服务于产业发展，并在公共平台上通过 API 的方式让所有人方便地使用大模型；同时，还为客户提供了很多具体的服务，比如提供大模型的授权产品并将其部署至客户内网，以及为客户设计大模型培训课程，实现“授人予渔”。

除此之外，智谱 AI 在算法上做了新的开发，使智谱大模型系列能够支持更多国产化的算力平台，并积极探索大模型的商业化应用路径，打造更具商业应用前景、更具易用性的开放性生态平台（https://open.bigmodel.cn）。

从冲击中突破

2020 年 5 月，OpenAI 发布了 GPT-3，将预训练模型的参数规模推到了 1000 亿以上。模型表现出了超乎想象的智能水准，也激发了智谱 AI 对参数量的重视，坚定了他们投入更多资源和精力去做模型的决心。他们预感到，基于 GPT-3 这样一个优秀千亿基座的智能应用会迎来爆发。 2022 年末，ChatGPT 的发布传达给智谱 AI 一个明显的信号：预训练模型已经到了完全可使用且好用的阶段，是产品化很好的范例。

同行的启发、市场的期待、多年的积淀……智谱 AI 很快顺势推出了千亿基座的对话模型 ChatGLM，并开源单卡版模型 ChatGLM-6B，使得研究者和个人开发者进行微调和部署成为可能。在细分领域方面，团队打造了 AIGC 模型及产品矩阵，包括生成式 AI 提效助手“智谱清言”、高效率代码模型 CodeGeeX 等。

今年 6 月，智谱 AI 将千亿模型 ChatGLM 升级到二代，效果大幅提升，模型支持的上下文长度扩展到 32K，并大幅提高推理速度。基于基座模型能力的增强，AI 提效助手“智谱清言”已具备更强大的性能，在多轮对话当中，作为一个“有知识、有记忆”的 AI 助手，其对上下文理解长度已从 2K 拓展至 32K，储备了包括科学、技术、历史、文化、艺术、商业和其他垂直领域的丰富知识，以此保障用户人机对话体验，持续畅聊无压力。目前产品已具备

通用问答、多轮对话、创意写作、代码生成以及虚拟对话、多模态生成等丰富能力。

“中国没有自己的预训练模型框架。无论是 GPT、BERT，还是 T5，都是西方的科学家提出的底层技术，路径是被西方垄断的状态。”智谱 AI 希望在完整的模型生态和全流程技术支持下，打破垄断局面，走出有中国特色的人工智能之路，通过认知大模型链接物理世界的亿级用户，为千行百业带来持续创新与变革，加速迈向通用人工智能的时代。

张鹏在新一代模型发布会上演讲

拥抱社会责任

清华大学的张钹院士率先提出了“第三代人工智能”，即认知智能的概念。在这一过程中，传统基建将向数字基建转化，数字底座建设完毕后，数字化成果将通过智能化凸显。“大模型”在其中扮演数字世界引擎和桥梁的角色，是机器与人交流的纽带。

中国人工智能发展具有两大核心优势：移动互联网普及带来的数据优势，以及庞大的网民数量带来的用户优势。同时，国家政策对人工智能产业发展给予高度重视，国务院印发的《新一代人工智能发展规划》中提出了我国人工智能发展的三步走战略，其中提到在 2025 年人工智能将成为带动我国产业升级和经济转型的主要动力，智能社会建设取得积极进展。

在社会服务方面，“大模型”有不可估量的价值。2022 年北京冬奥会期间，智谱 AI 同清华大学、凌云光技术股份有限公司携手，在北京市残联和北京市聋人协会的帮助下，为北京电视台打造了专属手语数字主播，方便听障人士实时观看比赛盛况。

口语新闻主播的语速大概是两百字每分钟，但是手语一分钟最多只能打八十个字。数字手语主播的翻译速度不掉队，离不开“大模型”的帮助。利用“大模型”理解音频中的语义，再利用语义蒸馏模型和手语翻译模型，将语音转化为手语，最终通过 3D 驱动数字人的形象呈现出来，一个数字手语主播便活灵活现地出现在观众面前。GLM-130B 的应用算力需求更少，从而降低了整个流程的成本。

据人口普查数据，我国共有 2700 万的听障人士，他们背后有庞大的家属群体，也有同他人交流的情感需求。借冬奥会这一契机，智谱 AI 实现了其他更普惠的成果。微信小程序上，可以搜到智谱 AI 开发的手语词典，除了听障人群，每个人都可以在上面学习标准化手语，打破交流的障蔽；手语数字人的应用场景也不只在媒体平台上，很多线下景区、博物馆、展览馆等都已经配备了智谱 AI 出品的手语解说，比如北京门头沟区的潭柘寺等。

肩负着建设智能社会的企业责任，智谱展望未来，希望在人工智能发展的里程碑事件中，将会出现更多中国人的身影。“我觉得在原创性、甚至是基础理论的突破上面，我们有这样的责任，去培养更多的人才，发挥人才的创新创业能力。这也是清华大学和三创大赛一直在做的事情。”

北京智谱华章科技有限公司简介

北京智谱华章科技有限公司（简称“智谱 AI”）成立于 2019 年，是基于清华大学计算机系知识工程实验室（KEG 实验室）科研成果设立的成果转化企业。

智谱 AI 团队打造的产品“智谱清言”上线

智谱 AI 致力于打造新一代认知智能大模型，专注于做大模型的中国创新。公司于 2020 年底开始研发 GLM 预训练架构，2021 年训练完成百亿参数模型 GLM-10B，同年利用 MoE 架构成功训练出收敛的万亿稀疏模型，2022 年合作研发了中英双语千亿级超大规模预训练模型 GLM-130B。

2023 年，智谱 AI 推出了基于千亿基座的对话模型 ChatGLM，并开源单卡版模型 ChatGLM-6B，使得研究者和个人开发者进行本地微调和部署成为可能。6 月，ChatGLM2 系列模型推出，提供丰富尺寸，适用于多种场景。与此同时，团队打造了 AIGC 模型及产品矩阵，包括 AI 提效助手智谱清言（chatglm.cn）、高效率代码模型 CodeGeeX、多模态理解模型 CogVLM 和文生图模型 CogView 等。10 月，在全新升级的 ChatGLM3 赋能下，生成式 AI 助手智谱清言已成为国内首个具备代码交互能力的大模型产品。

公司践行 Model as a Service（MaaS）的市场理念，推出大模型 MaaS 开放平台（https://open.bigmodel.cn/），基于领先的千亿级多语言、多模态预训练模型，打造高效率、通用化的“模型即服务”AI 开发新范式，实现服务效率的提升。

通过认知大模型链接物理世界的亿级用户，智谱 AI 基于完整的模型生态和全流程技术支持，为千行百业带来持续创新与变革，加速迈向通用人工智能的时代。

企业诉求

生态合作：我们积极寻求建立长期互信的合作关系，以大模型核心技术推动产业创新，提供私有化部署、API 调用等多种大模型解决方案，以更好地支撑行业生态，赋能合作伙伴高速发展。

联系电话：(010) 8215 8853