上海科学技术情报研究所
上海市前沿技术发展研究中心
TISC技术与创新支持中心

数字化城市
DIGITAL CITY BRIEFING

2023 年

第 14 期

大模型商用提速

编者按

当前，人工智能大模型的技术突破使其在广泛行业领域展现出巨大的应用潜力，大模型进入应用落地加速期。大型研究机构和企业纷纷抢抓生成式人工智能（AIGC）产业机遇，国内外头部企业相继推出行业大模型。其中，国外企业以 Open AI 为代表，其最新推出的大模型 GPT-4 Turbo 功能再次更新，预计将获得更广泛的行业应用。国内企业以百度为代表，其推出的百度文心大模型在国际数据公司发布的大模型能力评估报告中表现突出。大模型在生物医药、能源、半导体等垂直领域的应用带来了生产力的提升。制药公司英矽智能利用 AIGC 研发新药；能源公司能量奇点利用强化学习实现可控核聚变新突破；半导体公司英伟达利用内部数据训练定向行业模型，提高研发芯片的工作效率，等等。与此同时，主要国家和组织越来越重视对 AI 模型落地的应用基准工作，加强对其安全风险的前瞻研究和规范治理。

目录

技术

美国顶尖人工智能企业推出新模型 GPT-4Turbo

2023 年 11 月初，美国顶尖人工智能企业 OpenAI 前首席执行官 Sam Altman 宣布推出 ChatGPT 的新模型，GPT-4Turbo。**以下为最新模型 GPT-4 Turbo 相较 GPT-4 的 5 个关键更新**。

1. 新的信息截止日期

ChatGPT 之前版本的信息截止日期为 2021 年 9 月。“我们和你们所有人一样感到恼火，甚至可能更恼火，因为 GPT-4 对世界的了解在 2021 年就结束了”——Sam Altman 曾经在 OpenAI 大会这样说过。而新模型的信息截止日期为 2023 年 4 月的信息，因此它可以针对用户的提示提供更多上下文。

2. 输入更长的提示

GPT-4Turbo 支持多达 128 000 个上下文标记，可以帮助用户分析大量文档并提供摘要。

3. 更好地遵循指令

据 OpenAI 称，新模型将成为更好的倾听者。在需要仔细遵循指令的任务中，例如生成特定格式（例如，“始终以 XML 进行响应”），GPT-4Turbo 的表现比以前的模型更好。这对于在聊天机器人的帮助下编写代码的人来说可能特别有用。

4. 开发商价格更便宜

对于大多数 ChatGPT 用户来说，这可能不是首要考虑的问题，但对于开发人员来说，使用 OpenAI 的应用程序编程接口可能会非常昂贵。因此，新的定价是 1 美分/1 000 个提示。这意味着 GPT-4Turbo 开发人员输入信息和接收答案的成本可能会更低。

5. 自动选择工具

GPT-4 的下拉菜单功能让用户可以在其中选择要使用的聊天机器人工具。例如，如果想要一些人工智能生成的图像，可以选择 Dall-E 3 beta Browse with Bing 版本。而使用 GPT-4Turbo 更新的聊天机器人将自动为用户选择相应的工

具，例如，如果要获得图像，它会自动使用 Dall-E 3 来回答用户的提示。

这并不是 OpenAI 第一次为 ChatGPT 提供新模型。今年早些时候，OpenAI 将 ChatGPT 的算法从 GPT-3.5 更新为 GPT-4。GPT-4 相较 GPT-3.5 的改变主要有 6 点，分别是：可以处理文字和图片；更好地应对语言游戏；处理更多文字以及更复杂的问题类型；更好地应对标准化测试；更好的推理能力；更强的角色扮演等。新模型 GPT-4Turbo 相比以往的模型，在速度、经济、时效性等方面皆有很大的跃升。

参考文献：

[1] ROGERS R. 5 Key Updates inGPT-4Turbo, OpenAI's Newest Model[EB/OL].（2023-11-07）[2023-11-20]. https://www.wired.com/story/5-updates-gpt-4-turbo-openai-chatgpt-sam-altman/.

基于扩散模型的图像生成算法取得最新进展

自从 Stable Diffusion 从去年下半年发布之后，基于扩散模型的生成式图像模型获得了业界极大的关注。扩散模型一般模型都较大，而且需要运行多步的采样过程，之前虽然也有运行在手机上的例子，但是因为运行时间过程（10 秒左右），尚未得到真正大规模应用。然而，**随着 2023 年 10 月份中国清华团队发表了 latent consistency model（LCM）的研究论文，在手机上运行高性能图像生成式模型已经不再遥不可及**。

LCM 模型和 Stable Diffusion 的模型结构类似，但是 LCM 通过数学上的优化，可以把一次生成需要的模型执行次数从 Stable Diffusion 的 50 次降低到 2~4 次，相当于把端到端的运行速度提升了 10 倍，而且生成图像的质量和 Stable Diffusion 接近。目前，LCM 已经在人工智能社区得到了广泛的关注和应用，业内认为很快 LCM 就会成为手机上运行图像生成式模型的首选，而且 LCM 的低延迟可以真正实现全新的用户体验；例如，高质量的实时超分辨可以让数字变焦得到的拍摄质量和光学变焦相似，但是同时又避免了厚重的镜头；又如，inpainting/outpainting 可以让用户快速编辑拍摄的照片并分享，能实现在手机上拥有和 photoshop 相似的效果，大幅提升用户体验。

对于智能助理应用来说，目前主要还处于探索阶段，如何将多模态的信息（包

括用户的短信、备忘录、日历等）整合在一起并不容易，但是行业人士认为最终模型的形态最有可能还是类似 GPT 这样的大语言模型，通过海量数据与训练来实现对于用户数据的深入理解并且给出相应帮助。这类智能助手的第一步落地应用可能是用户消息编辑和改写，例如用户可以让智能助手去改写一条短信以改变语气，这样的应用预计在明年就会落地。

参考文献：

[1] 李飞. 手机端生成模型爆发在即，芯片迎来巨变？[EB/OL].（2023-11-20）[2023-11-20].https://mp.weixin.qq.com/s/HOCVxT1WdjnEL3M3ju7PIQ.

CVP 架构帮助解决大模型落地关键问题

目前，业界对生成式人工智能（AIGC）的焦点有所转移：从对 AIGC 大模型技术的狂热，转移到对大模型商业化落地效果的审视。

生产力工具是目前大模型落地的重要方向。但面向垂域应用，仅依靠大模型自身的训练数据集做支撑，难以达到生产可用的效果。围绕这一问题，也形成了两大流派，传统流派将垂域/私域内容补充至训练集，即单模型架构；**新兴流派引入向量数据库为大模型提供长短期记忆，集成领域知识库，即 ChatGPT+VectorDB + Prompt（CVP）架构**。

CVP 架构的兴起让向量数据库这个新的数据库品类站上了风口浪尖。全球最火的开源向量数据库项目 Milvus 在 Github 的标星已经突破 2 万，官方显示，目前 Milvus 已经拥有超过 1 000 家中大型企业用户。其母公司 Zilliz 已经完成 1.13 亿美元融资，并全面启动商业化步伐，为大模型落地提供向量数据库全栈产品与服务。CVP 架构相比单模型架构在灵活性、可扩展性、实时性、成本四个维度都有明显优势。最关键的原因是在 CVP 架构中，领域知识以数据入库的形式进行更新，而非重新训练或微调模型。

目前，国内大模型的综合能力与 GPT-4 还有代差，但现在已经走到了产业落地的早期。CVP 架构的端到端效果明显优于 GPT-4。在这个框架下，除了模型能力，还需要知识库的构建能力、模型与知识库的集成水平，这也为国产化大模型提供了一次在落地阶段弯道超车的机会。

参考文献：

[1] 腾讯云. 大模型商用新解法：CVP 架构崛起，向量数据库破圈[EB/OL].（2023-09-07）[2023-11-20]. https://cloud.tencent.com/developer/article/2325245.

产业

国内外头部人工智能企业相继推出行业大模型

随着大模型从技术到应用落地的加速，国内外头部 AI 企业相继推出行业大模型，大模型商用生态初现。

1. OpenAI 正式发布 GPT-4，具有多模态能力，应用范围更广

2023 年 3 月 15 日，OpenAI 正式发布 GPT-4 预训练大模型，相比于 GPT-3.5 性能表现显著提升，在一些专业和学术领域上已经达到了人类水平。GPT-4 具有一定的多模态能力，能够接收图文结合的输入，并输出文本回复，应用范围得到进一步拓展。基于 GPT-4 对话交互的特性，我们认为，GPT-4 将率先在教育、医疗、企业经营管理办公等领域实现落地，场景与人工智能的结合方式值得期待。

2. 百度的对话式大模型“文心一言”正式发布，多模态生成能力亮眼

2023 年 3 月 16 日下午，百度正式揭开了“文心一言”的面纱。“文心一言”具备五大能力，中文理解能力强，并且支持从文本生成图像、音频和视频，多模态能力十分亮眼。目前已有 650 家企业成为“文心一言”的首批生态合作伙伴，落地场景涵盖各行各业。我们认为，国产大模型的发布将带来诸多产业机会，MaaS 未来将有望成为大模型落地的新形态，中国生成式 AI 市场有望迎来需求的大幅增长。

3. 微软推出 Microsoft 365 Copilot，办公场景根本性变革

2023 年 3 月 16 日晚，微软宣布将 GPT 大模型引入 Office 应用程序，推出了 Microsoft 365 Copilot，帮助用户提高办公生产力。 Copilot 打通了微软的办公产品线，数据在各个产品中能够自由流通。作为核心的流程编排引擎，Copilot 大幅提升了用户在办公场景、业务协作场景、流程自动化场景的效率。

4. 英伟达 GTC 2023 召开，展现 AI 多领域应用

2023 年 3 月 21 日晚，英伟达 CEO 黄仁勋为 GTC2023 进行了主题演讲，对英伟达在 AI 应用、加速计算等领域的最新动态进行了介绍。在本次演讲中，英伟达推出了 DGX 云服务，方便企业客户更快地访问英伟达 AI 算力与应用库。对想要建立独有的垂直领域行业模型的客户，英伟达推出了 AIFoundations 一站式云服务，协助客户快速构建、优化和运营大模型，把制造大模型的能力传递到每一个用户。

大语言模型的诞生带来了新的知识表示和调用方式的变迁，用户仅需自然语言交互即可实现知识调用。当下人工智能在生成和通用两条主线上不断发展，两条赛道相互交织并行，AI 领域进入了“双 G 时代”。随着百度“文心一言”的发布，我国的大模型达到了“能用”的标准，在 B 端、垂直行业端都有广阔的应用空间。我们认为，我国在 AI 的“双 G 时代”中将持续扮演追赶者的角色，争取和国际领先水平缩小差距，百花齐放的 AI 新时代正在到来。

参考文献：

[1] 东方证券. 大模型应用百花齐放，AI 发展进入新时代[EB/OL].（2023-03-27）[2023-11-20]. https://pdf.dfcfw.com/pdf/H3_AP202303281584600914_1.pdf.

大模型带来的主要产业机会

大模型目前核心的竞争点在于人工智能最重要的核心能力，也就是理解、生成、逻辑和记忆能力。这四大能力也为通用人工智能的实现带来了曙光。

大模型技术对于加快形成新质生产力的贡献将主要体现在两个层面，**第一是对企业经营效益的提升**，大模型技术可以极大地解放生产力，将人从重复性的工作内容中脱离出来，从而将更多的创造力和精力投入到更高价值的生产工作中，最终整体提升行业生产力。**第二是对整体社会经济的增长拉动**，各行各业通过与大模型技术的结合应用，相信可以焕新生产模式，推动产业智能化过程，加速 AI 普惠，让通用人工智能成为可能。

大模型的产业机会主要有三个方向：

一、新型云计算公司

文心一言将根本性地改变云计算行业的游戏规则。之前企业选择云厂商更多

看算力、存储等基础云服务。未来，更多会看框架好不好、模型好不好，以及模型、框架、芯片、应用这四层之间的协同。

二、行业模型精调

这是通用大模型和企业之间的中间层，他们可以基于对行业的洞察，调用通用大模型能力，为行业客户提供解决方案。这方面，百度文心大模型已经在电力、金融、媒体等领域应用。

三、基于大模型底座的应用开发

对于大部分创业者和企业来说，真正的机会并不是从头开始做基础大模型。基于通用大语言模型抢先开发重要的应用服务，才是真正的机会。目前，基于文本生成、图像生成、音频生成、视频生成、数字人、3D 等场景，已经涌现出很多创业明星公司，可能就是未来的新巨头。

参考文献：

[1] 何珊珊. 百度吴昊：大模型技术发展迅速，将进入产业应用高速发展期[EB/OL].（2023-11-19）[2023-11-20]. https://time-weekly.com/post/306839.

首份《AI 大模型技术能力评估报告》发布

2023 年 8 月，国际数据公司 IDC 发布了首份《AI 大模型技术能力评估报告》，围绕产品技术、服务生态以及行业应用三个维度，考察大模型的 10 余项指标。这是 IDC 首次提出 AI 大模型技术能力评估框架，国内主流大模型，包括百度、阿里、腾讯、华为、科大讯飞、360、商汤等 14 家厂商参与了本次评估。

从三大维度的评分分布来看，**百度文心大模型、阿里通义千问分数遥遥领先。**其中，百度文心大模型靠着独占算法模型、行业覆盖两点满分，在这次评比中胜人一筹，成为该报告的大优势方。在产品维度，百度的算法模型、通用能力、创新能力满分；具体到行业，百度在能源和行业覆盖上满分；在服务方面，百度的生态合作指标满分。总体来看，百度文心大模型 3.5 拿到 12 项指标的 7 个满分。

一、产品能力是行业落地的基础

作为大模型的基础关口，产品能力是服务和行业落地的基础，对企业而言显得至关重要。

IDC 将产品维度进一步细分为算法模型、通用能力、创新能力、平台能力和安全可解释五方面，百度是所评估企业中综合评分最高的企业，除安全可解释层面获得 4 分（满分 5 分）外，其余各项均为满分；阿里稍逊一筹，在算法模型和安全可解释层面均拿到 4 分。

IDC 报告本次评估的 14 家企业，既包括阿里、百度大厂玩家，还包括智谱 AI、科大讯飞、第四范式等 AI 公司。可以看到，在产品维度上，百度的文心大模型在算法模型上依托先发优势和技术领先性，显现出明显的领先优势。

二、产品是基础，服务是重点

IDC 将服务维度细化为服务能力和生态合作两个主要方面，前者主要包括为开发者提供的配套服务、对客户的理解力等；后者则不仅包括合作伙伴的数量，也包括结构分布、生态支持等。

在生态合作上，百度和阿里均得到满分。这一方面体现了，百度不仅布局大模型产品本身，也提前大力发展生态伙伴。另一方面也说明了，大模型时代下的生态也面临着巨变，会在很大程度上区别于传统云计算的服务生态。

在服务能力方面，阿里基于过去深厚积累的客户服务经验拿到 5 分。百度在文心一言发布半个月后，便推出了大模型服务平台文心千帆——全球首个一站式的企业级大模型生产平台，不但为客户提供包括文心一言在内的大模型服务及第三方大模型服务，还提供大模型开发和应用的整套工具链。

三、研发大模型的最终目的：产业实践

无论是产品维度，还是服务能力，企业研发大模型的最终目的还是要落地于产业实践，这也就是 IDC 评估的第三个重要维度：行业覆盖。

综合来看，各家企业都依托于自身优势在部分垂直行业取得一定分数，但百度和阿里再次成为业内的领先企业，是少有的能够在部分行业拿到满分的企业。百度文心大模型领先优势明显，在综合指标行业覆盖上拿下唯一的满分。

IDC 在报告中指出，百度文心大模型形成了支撑大模型产业落地的关键路径：在模型层，文心大模型包含数 30 多个大模型，涵盖基础大模型、任务大模型、行业大模型的三级体系，全面满足产业应用需求。

参考文献：

[1] 代聪飞. 深度解读首份大模型报告：技术只是门槛，落地才是赛点[EB/OL].（2023-07-24）[2023-11-20]. https://www.leiphone.com/category/ai/bYcsTkSK93AEk3rl.html.

政策

各主要国家/组织加强人工智能规范治理

如今，在 AI 技术推动社会全要素生产率提高的同时，数据、算法、算力等各方面的局限性也会带来诸如虚假信息、AI 幻觉等问题，尤其是 AI 的侵权风险正日益威胁正常的社会秩序。目前世界各国都投入了巨大的精力和资源，通过各种监管方式来制约 AI 技术风险并逐步形成自身的监管模式。

一、目的性监管

目的性监管用于打击 AI 网络犯罪、恐怖主义等与 AI 相关的非法活动。例如欧盟主要运用自上而下的监管，即有目的地针对目标人群进行监管。 对 AI 监管态度最为严格的欧盟，在 2021 年就提出制定《人工智能法案》（*Artificial Intelligence Act*），提出了大量自上而下的规则，包括禁止使用欧盟认为可能会带来严重风险的 AI 技术和应用，强调要通过建立法规监管体系管控风险，保证 AI 发展安全、符合道德规范且值得信赖。这个法案对不同领域运用 AI 存在的风险进行评估，并禁止在“不可接受的风险（Unacceptable risk）”领域使用 AI 技术。不过，并非所有欧盟国家都认同这一法案的风险划分，例如法国的 AI 视频监控计划，就与欧盟法案相悖。目前，欧盟的《人工智能法案》已进入通过的最后阶段。

二、分散式监管

分散式监管主要由不同地方政府、组织和科技公司进行自我约束和控制。美国主要运用分散型监管规范 AI 技术。作为 AI 领域领导者，虽然联邦政府层面尚没有限制 AI 的法规，但近期有不少官员表达了更强的监管意愿；同时，美国不同的州、组织和领先企业中，也存在着不同程度的监管措施。例如，微软公司在其“管理 AI：未来蓝图”中提出了一种全面的方法，将建立一个新的政府 AI 机构，一个新的 AI 法律框架，要求控制关键基础设施的 AI 系统必须安全制动，

并为运行关键 AI 的数据中心颁发许可证。同样使用分散性监管的国家还有加拿大，其正在推动《AI 与数据法案》（*The Artificial Intelligence and Data Act*），明确 AI 对个人存在的威胁并监管其发展。

三、分散性立法

分散性立法旨在出台统一法律加强对 AI 技术的道德约束和规范治理。中国在制定复杂人工智能法规方面被认为是领导者。近两年，中国提出多部法律法规，从深度伪造、演算法、生成式 AI 等方面对相关技术的发布者提出监管要求。中国采用分散性立法对 AI 技术作出监管，譬如《生成式人工智能服务管理暂行办法》，指出“对生成式 AI 服务实行包容审慎和分类分级监管”“坚持社会主义核心价值观”。 2022 年 9 月，深圳、上海先后通过了《深圳经济特区人工智能产业促进条例》《上海市促进人工智能产业发展条例》，两部法规主要着眼于产业发展和应用，涉及部分监管和治理内容。但目前中国还没有制定统一的人工智能法。

总的来说，各个主要国家和组织都在积极出台政策，重视对 AI 技术进行监管。唯有在大力发展 AI 技术的同时加强对大模型的规范治理，才能在创新和安全之间取得平衡，确保 AI 技术造福全人类。

参考文献：

[1] 根据 2023 竞争情报上海论坛王丹教授发言整理。

中国信通院启动大模型技术及应用基准工作

当前大模型已呈现出产业生态不断扩大、场景应用快速涌现、商业路径逐步清晰等特点，成为现阶段人工智能产业的技术创新主线。为进一步推动我国大模型技术创新发展及工程化应用落地，中国信息通信研究院（以下简称“中国信通院”）现启动大模型技术及应用基准构建工作，针对当前主流数据集和评估基准多以英文为主，缺少中文特点、文化以及难以满足关键行业应用选型需求等问题，联合业界主流创新主体共同构建一套涵盖多任务领域、多测评维度的基准及测评工具 AISHPerf-LargeModel，推动我国大模型技术及应用的引领创新。

AISHPerf-LargeModel 属于人工智能软硬件基准测评体系范畴（Performance

Benchmarks of Artificial Intelligence Software and Hardware，以下简称 AISHPerf），AISHPerf 由中国信通院牵头构建，旨在面向自主生态建设和产业实际需求，围绕 AI 芯片性能、算法任务表现，构建国产人工智能软硬件领域的测试基准任务，推动相关技术发展。

中国信通院一直以来围绕大模型展开深入的研究及产业培育，在此之前已开展多项工作：

一、大模型能力测评，围绕自然语言处理、多模态技术能力以及系统平台功能构建共计 30 余项细分任务的评估体系，推动完成百度、中科院自动化所及武汉人工智能研究院、联汇科技等创新主体的大规模预训练模型系统能力测评。

二、发布大模型产业生态透视图 V1.0，对当前大模型领域算力设施、基础大模型、平台及服务、大模型典型应用四大关键环节的主要创新主体进行了梳理，同时分析了大模型落地的关键应用模式及演进趋势，目前，已启动产业生态透视图 V2.0 编制，欢迎业界共同参与。

三、依托工信部、科技部等主办的“兴智杯”全国人工智能创新应用大赛，设置基于大规模预训练模型的创新应用方案赛，构建大模型资源池，构建一批能够充分发挥大模型优势、解决领域痛点的创新解决方案。

参考文献：

[1] 罗懿. 中国信通院启动大模型技术及应用基准工作[EB/OL].（2023-03-28）[2023-11-20]. https://m.yicai.com/news/101714405.html.

案例

英矽智能利用生成式人工智能研发新药

英矽智能（Insilico Medicine）利用 AIGC 技术，缩短药物研发周期，提高研发效率，降低研发成本，为生物医药研究提供颠覆性的解决方案。使用传统手段和人类经验发明小分子创新药物需要投入巨大的时间和资金，而 **AIGC 可以帮助研发人员利用患者的组学数据寻找更新颖、更精确的靶点，设计效果更好的小分子化合物，并通过数据预测临床试验成功率**。

用传统手段去做小分子创新药有着很大的瓶颈。首先，平均一款药的出现需要 10 年以上的研发时间，需要耗费 20 亿～60 亿美元以上的研发投入。之所以有这个瓶颈，那是因为在传统创新药物研发主要是依靠人的经验去寻找靶点做分子，依靠人的经验不能很好解决以下三个方面的问题：

（1）纯粹依靠人的经验很难很高效去找到合适新颖、准确的靶点，运用生物医学靶点或者生物医学机制来做出创新药。

（2）依靠人的经验很难高效找到成药性很好的小分子化合物推进到临床。

（3）依靠个人经验很难设计出最好的临床试验方案，让分子在临床上最大可能地获得成功。

近年来出现的以数据和算法为基础的 AI 模型，有望帮助制药公司对这三方面的问题提供颠覆性的解决方案，以英矽智能为例，该公司发明了三个以数据和 AIGC 为基础的 AI 平台。

（1）Biology42，利用病人的组学数据帮助研究者寻找更准确、更新颖的靶点。

（2）Chemistry42，以生成式对抗神经网络等 AIGC 帮助研究者设计更好的小分子化合物。

（3）Medicine42，可以帮助研究者利用数据预测临床试验二期到三期的成功率，优化临床试验方案，增加药物研发成功率。

利用人工智能，英矽智能完美地将生物学和化学结合起来，针对未被满足的临床需求做出新药。

（1）治疗特发性非纤维化的药物

特发性非纤维化被定义成罕见病，在全球影响 70 多万的病人，是致死性很强的疾病，目前市场上有两款药，一款是尼达尼布，一款是吡非尼酮，这两款药 2021 年的销售额加起来达到 30 多亿美金，是一个庞大的市场，但这两款药有比较严重的毒副作用，可致 10%～40%的病人由于没有办法耐受毒副作用被迫停药，或者是在所耐受的剂量范围之内起不到药效，让这批病人没有药可以医治。而通过运用 AI 模型分析病人的组学数据，分析对比病人和健康人之间的组学差异，找到病人的哪些基因产生了变化或缺失、激活之后能让症状有所改善、预后比较好。通过这些方式找到了 20 个蛋白，利用内部的方法进行验证，最后固定在一个靶点上，即 T-neit，一种激酶抑制剂，目前全球只有英矽智能一家公司针对这个靶点有临床化合物。

（2）治疗溃疡性结肠炎

炎性肠炎在全球影响将近 2 000 万的病人，分为 UC 和 CD。目前市场上治疗炎性肠炎的药主要是抗炎或免疫一致，英矽智能希望脱离这两种机制，而是保护肠黏膜，因此该公司通过 AI 模型寻找到一种叫 PHD 蛋白的靶点，可以起到保护肠壁黏膜的作用；但研究者不希望药物进入血液，因为进入血液之后会让病人的血液变黏稠，药物只在肠道里，不进入血液，这是传统的制药方案达不到的。通过大模型的辅助分析，英矽智能公司发明了一种化合物，只在肠道里有暴露，而不进入血液，通过做药代动力学可以发现，肠道里面的暴露量是血液里面暴露量的几十倍，血液里面的可以忽略不计，活性达到预期效果。在噁唑酮小鼠模型的肠炎模型中，研究者发现在比较低的剂量下这种药物就可以改善肠道作用，同时实验人员将肠壁拿出来做了切片，发现新药可以改善肠壁，提高肠壁的紧密度。

总体而言，英矽智能公司在 AI 药物研发中的成果取决于两点：**一是专注于以 AIGC 搭建平台和持续优化；二是利用 AI 平台的赋能，持续加速药物研发速度**。

参考文献：

[1] 吴斯旻. 对话英矽智能联合 CEO：“AI 制药”有望打破创新药内卷[EB/OL].(2023-11-20)[2023-11-20]. http://www.stcn.com/article/detail/1040253.html.

能量奇点利用强化学习实现核聚变新突破

可控核聚变是解决人类能源问题以及环境问题的重要促进力量之一，有可能为全球碳中和作出重大贡献。根据国际能源展望，如果化石能源、天然气和煤炭逐步退出全球能源市场，到 2050 年人类将面临超过 50%的能源空缺，而风能、光伏目前只占约 10%的能源市场，无法满足人类的能源需求。

目前，各主要国家都在加快聚变能商用进程。美国在 2020 年就开始大幅增加聚变投资，计划在建设首个紧凑型聚变电站，宣称要用 10 年的时间重回世界第一。英国也制定了聚变能发展战略，计划在 2040 年前建成商业上可行的聚变发电站。日本在今年 4 月份发布了聚变和发展商业战略，把聚变能源列为解决人类问题的终极能源。韩国通过《聚变能发展与促进法案》。中国也计划在 2040 年演示并网发电。

国际聚变界正在持续探索新的技术路线。目前主要的两种路径，分别是先进磁约束方式以及简化磁约束方式。第一种方式主要依靠环形磁场系统来提高等离子体的稳定性，包括托卡马克、仿星器和反场箍缩等。第二种方式是没有环向场线圈的较简单磁约束系统，包括场反位形、球马克，以及磁惯性约束系统（MIF）等。不过，该途径还不够成熟，需要进行进一步的基础物理研究，以显著提升高温等离子体的约束性能。

但是，托卡马克面临的问题是，它的装置规模非常庞大复杂，建造和运营成本很高。所以如何降低成本，并使磁约束系统高效且紧凑是该途径面临的首要挑战。近年来，人工智能等技术突破使得发展经济型托卡马克成为可能。

能量奇点聚变能源开发公司利用 AI 技术，提升聚变性能，提前避免风险，解决关键材料问题。**强化学习的最新架构和算法大幅缩短了机器学习新任务的时间，提高了托卡马克装置中等离子体的形状精度，同时降低了电流稳态误差，实现能量持续稳定的输出。**

第一，人工智能可以节约核聚变的时间。

第二，可以帮助解决核聚变的关键科学和技术问题。

首先，与先进托卡马克相结合可以预测一些先进运行模式，提升聚变的性能。例如 Open AI 注重 AI 应用于聚变的投资。托卡马克装置中的电流很大，容易破裂，一旦破裂就会引起整个装置的毁坏。而 AI 可以帮助预测出现破裂风险的时间，提前避免风险。

其次，与先进控制相结合，保证装置运行。

最后，与先进制造相结合，解决聚变面临的材料问题。材料对聚变来说是一个非常关键的问题，到现在还没找到解决方案，每三到五年内就需要更新一次，严重影响运行成本，通过 AI 模型，可以从原子的层面设计材料，针对性地解决问题。

能量奇点公司将先进高温超导技术、先进托卡马克、人工智能相结合，制定三步走的发展规划：第一，计划在今年年底建造全世界首台高温超导托卡马克，利用最先进的第二代高温超导材料建造高温超导托卡马克。第二，计划 2027 年建成洪荒 170，这个装置对标美国的高温超导装置，计划实现 10 倍的能量增益，在 2030 年和 2035 年实现聚变发电。

参考文献：

[1] 郑闻文. 核聚变能源研发趋势、技术路线是什么？能量奇点创始人层层剖析[EB/OL].（2023-05-20）[2023-11-20]. https://j.021east.com/p/1684562099045269.

[2] 根据 2023 竞争情报上海论坛郭后扬博士发言整理。

英伟达推出内部定制模型测试芯片

半导体设计是一项极具挑战性的工作。在显微镜下，诸如英伟达 H100 这样的顶级芯片，看起来就像是一个精心规划的大都市，其中的数百亿个晶体管则连接在比头发丝还要细一万倍的街道上。为了建造这样一座数字巨城，需要多个工程团队长达两年时间的合作。其中，一些小组负责确定芯片的整体架构，一些小组负责制作和放置各种超小型电路，还有一些小组负责进行测试。每项工作都需要专门的方法、软件程序和计算机语言。因此，芯片从设计到生产，需要耗费大量的时间、人力资源和资金投入。

面对算力供应短缺的现状，领先的科技企业致力于提高芯片产量。2023 年 11 月，半导体公司英伟达推出 AI 模型测试芯片 ChipNeMo，将 AI 模型技术应用于半导体设计这种极具挑战性的工作中，开辟了设计芯片的新思路。ChipNeMo 以公司内部数据为基础进行训练，用于生成和优化软件，为人类设计师提供帮助，在一系列设计任务中实现了类似或更好的性能；同时这种“内部定制”模型的思路，也将帮助组织更好地维护数据安全。英伟达的尝试表明，对于高度专业化的领域，完全可以利用其内部数据来训练定向的 AIGC 模型，从而提高生产力和竞争力。

一、数据

为了构建领域自适应预训练（DAPT）所需的数据，研究人员同时结合了英伟达自己的芯片设计数据，以及其他公开可用的数据。经过采集、清洗、过滤，内部数据训练语料库共拥有 231 亿个 token，涵盖设计、验证、基础设施，以及相关的内部文档。就公共数据而言，研究人员重用了 Llama2 中使用的预训练数据，目的是在 DAPT 期间保留一般知识和自然语言能力。在代码部分，则重点关注了 GitHub 中与芯片设计相关的编程语言，如 C++、Python 和 Verilog。在监督微调 (SFT) 过程中，研究人员选取了可商用的通用聊天 SFT 指令数据集，并制作了的特定领域指令数据集。为了快速、定量地评估各种模型的准确性，研究人员还构建了专门的评估标准——AutoEval，形式类似于 MMLU 所采用的多选题。

二、训练

ChipNeMo 采用了多种领域适应技术，包括用于芯片设计数据的自定义分词器、使用大量领域数据进行领域自适应预训练、使用特定领域任务进行监督微调，以及使用微调检索模型进行检索增强。首先，预训练分词器可以提高特定领域数据的分词效率，保持通用数据集的效率和语言模型性能，并最大限度地减少重新训练/微调的工作量。其次，研究人员采用了标准的自回归语言建模目标，并对特定领域的数据进行了更深入的预训练。针对大模型的幻觉问题，研究人员选择了检索增强生成（RAG）的方法。研究人员发现，在 RAG 中使用与领域相适应的语言模型可以显著提高特定领域问题的答案质量。此外，使用适量的特定领域

训练数据对现成的无监督预训练稠密检索模型进行微调，可显著提高检索准确率。

在 DAPT 之后，则进一步利用监督微调（SFT）来实现模型的对齐。

三、性能

研究人员并没有直接部署现成的商业或开源 LLM，而是采用了以下领域适应技术：自定义分词器、领域自适应持续预训练（DAPT）、具有特定领域指令的监督微调（SFT），以及适应领域的检索模型。结果表明，与通用基础模型相比（如拥有 700 亿个参数的 Llama 2），这些领域适应技术能够显著提高 LLM 的性能——不仅在一系列设计任务中实现了类似或更好的性能，而且还使模型的规模缩小到五分之一（定制的 ChipNeMo 模型只有 130 亿个参数）。

具体来说，研究人员在三种芯片设计应用中进行了评估：工程助理聊天机器人、EDA 脚本生成，以及错误总结和分析。其中，聊天机器人可以回答各类关于 GPU 架构和设计的问题，并且帮助不少工程师快速找到了技术文档。代码生成器已经可以用芯片设计常用的两种专业语言，创建大约 10~20 行的代码片段。而最受欢迎的分析工具，可以自动完成维护更新错误描述这一非常耗时的任务。

参考文献：

[1] 新智元. 专攻芯片设计，英伟达推出定制版大语言模型 ChipNeMo[EB/OL].（2023-11-01）[2023-11-20]. https://36kr.com/p/2499490883639430.

地址：上海市永福路 265 号
邮编：200031
编辑：金旸
责编：曹磊
编审：林鹤
电话：021-64455555
邮件：istis@libnet.sh.cn
网址：www.istis.sh.cn