

医学
拾萃

IMPROVE REVIEW

智慧医疗篇
Wisdom medical

二〇二三年六月刊
June, 2023



Article Reading Guidance
文献导读

Article Abstract Collection
文献摘要

Featured Article
文献精读

Original Article Reading
文献原文

医学
拾萃

CONTENTS

 Article Reading Guidance
文献导读 | p01

 Article Abstract Collection
文献摘要 | p11

 Featured Article
文献精读 | p21

 Original Article Reading
文献原文 | p28



主编 邓冠华 博士
Chief Editor Dr. Guanhua Deng

联合主编 朱勇 博士
Co-Editor-in-Chief Dr. Yong Zhu

执行主编 高树勇
Executive Editor Shuyong Gao

专刊编辑 田艳丽 博士 李小林 博士 杨艳 博士
Special Editor Dr. Yanli Tian Dr. Jerry Li Dr. Yan Yang

李铁 罗江兰
Tie Li Jianglan Luo

审核编辑 卢嘉丽
Audit Editor Jiali Lu

协作单位 阳普学院
Partnership Improve College

广东省医用材料血液相容性研究企业重点实验室
Guangdong Enterprise Key Laboratory of
Blood Compatibility of Medical Materials

美国吉迪思诊断有限公司
GIMDX, Inc.

联系方式 yxsc@improve-medical.com
Contact Us

 **陽普醫療**
IMPROVE MEDICAL



电子文献



Article Reading Guidance

文献导读

GPT 技术对医疗信息行业的影响

李 铁 屈晓涵 谢桂连

阳普医疗科技股份有限公司

广东阳普智慧医疗信息科技有限公司

摘要

GPT-4 (Generative Pre-trained Transformer 4) 是 Open AI 旗下 GPT 系列模型的最新成果,这是一种大型多模态模型(可以接受图像和文本输入,并输出文本),在各专业和学术领域展现出了与人类水平相当甚至超越人类水平的能力。这类大型人工智能模型的出现不断地挑战人类对学习和认知的理解,迫使人们开始思考其可能产生的深远影响,以及其改变各领域的巨大潜力。随着互联网技术与深度学习(Deep Learning)技术的发展,目前在生物医学和医疗健康领域已经积累了海量的多模态数据,为后续开发、验证和推进大型人工智能模型(Large AI Model, LAM)在相关领域的突破提供了坚实基础。与此同时,对医疗信息行业而言,GPT 技术的出现开启了多模态通用 AI 大模型时代,并为方法论的设计带来了新的范式与灵感。另一方面,GPT 技术同样也存在深度学习模型本身固有的黑盒问题^①,以及涉及在数据安全、患者隐私等方面法律和伦理仍不完善的情况。本文旨在研究 GPT 技术对医疗信息行业可能产生的重要影响,并对其在医疗信息化领域的潜在未来方向和局限进行了批判性的分析与讨论。

一、引言

1. 背景介绍

1.1 医疗信息行业的发展

医疗信息化是指将包含计算机、数据库、网络等在内的信息技术赋能医疗行业,从而提高医疗工作效率和医疗质量,降低医疗风险和管理成本。医疗信息行业的发展历程主要分为五个阶段,早期阶段的重点为实现医疗数据的电子化。医院和诊所开始使用计算机管理患者的基本信息(如姓名,性别,年龄等)。此外,一些基本的医疗记录和处方也开始电子化。这一阶段的目标是简化数据记录和存储,提高数据管理的效率。在

中期阶段,医疗机构开始采用电子病历系统(Electronic Medical Record, EMR)来实现患者病历的完全电子化。与此同时,医疗信息系统,如医院信息系统(Hospital Information System, HIS),实验室信息系统(Laboratory Information System, LIS),影像归档和通讯系统(Picture Archiving and Communication System, PACS)等得到了广泛应用,实现了不同部门和系统之间的数据整合和共享。这一阶段的目标是提高数据的可用性和一致性,实现医疗资源的合理利用。在进阶阶段,医疗信息化开始拓展到远程医疗和移动医疗领域。利用互联网技术,患者可以通过在线平台进行预约、咨询和获取电子处方等。同时移动医疗应用(mHealth)开始涌现,患者和医生可以通过手机或平板电脑等移动电子设备随时随地访问医疗信息。这一阶段的目标是提高医疗服务的可及性和便捷性,缓

^① 深度学习模型有一个共同特点为内部结构非常复杂,其运作机制就像黑盒子,难以用人类可以理解的语言去描述,模型输出结果也难以被解释,使得其在一些有关生命安全或重要决策领域的应用受到巨大挑战。

解医疗资源的不平衡。在第四阶段，医疗信息化的重点转向了智能化和大数据分析。借助人工智能（Artificial Intelligence，AI），机器学习（Machine Learning，ML）和自然语言处理（Natural Language Producing，NLP）等技术，医疗信息化实现了诊断辅助、病理分析和知识获取等智能化应用。此外，大数据技术在医疗领域的应用，如疾病预测、流行病学研究和个性化治疗等，也取得了显著成果。当前，我们正在步入第五个阶段，医疗信息化将更进一步发展，实现更加个体化和智能化的医疗服务。基于基因组学，蛋白质组学和代谢组学等数据，个体化医疗将为患者提供更精确的诊断和治疗方案。此外，智能医疗机器人有望在手术，康复训练和护理等领域发挥重要作用，GPT（Generative Pre-training Transformer）技术和其他人工智能技术也愈发重要。

1.2 GPT 技术的兴起与发展

通过开发具有写作和对话能力的大规模自然语言模型，AI 已经向更好的自然语言理解（Natural Language Understanding，NLU）和人机交互（Human-Computer Interaction，HCL）迈出了一大步^[1]。截至 GPT-4 出现前，OpenAI 开发的 ChatGPT 是参数最多、规模最大、能力最强的语言的模型^[2]。GPT 技术作为一种基于 NLP 的人工智能技术，

本文将其兴起与发展经历的几个关键阶段整理为表 1。由表 1 可知，GPT 是基于当时表现性能最好的语言模型 Transformer 框架^[3]提出的，在利用了非监督学习进行预训练后，通过监督学习根据特定任务进行微调，使其在多个自然语言处理任务上取得显著的性能提升^[4]。2019 年推出的 GPT-2 开始在文本生成、摘要、翻译与问答等多个任务上展示出惊人的能力，引发了广泛关注^[5]。但由于担心其潜在的滥用风险，OpenAI 最初并未公开发布 GPT-2 的完整模型。2020 年公布的 GPT-3 拥有 1750 亿个参数，是当时世界上最大的语言模型。GPT-3 在各种自然语言处理任务上取得了令人瞩目的表现，仅通过少量的样本进行零样本学习，即在没有进行监督训练的情况下，生成极为真实和连贯的文本结果。GPT-3 的出现标志着语言模型的发展进入了一个新的阶段，且在某些情况下成为若干语言理解任务的优先选择^[6]，特别是其最后一个系列的 ChatGPT 引发了广泛开发研究和商业应用。近年推出的 GPT-4 除了掌握语言之外，其无需任何特殊提示就可以解决跨越数学、编程、视觉、医学、法律、心理学等领域的新颖而困难的任务。在所有这些任务中，GPT-4 的性能惊人地接近人类水平，并且远远超过以前的模型，包括 ChatGPT。其能力的升级和应用的拓展又一次引爆国内外的关注^[7]。

表 1. GPT 技术的发展历程

时间	提出的模型	表现
2017年	Transformer	采用自注意力机制，显著提高了神经网络在处理长距离依赖问题的能力，大幅提升训练速度；
2018年6月	GPT-1	基于Transformer架构提出，在多个自然语言处理任务上取得显著的性能提升；
2019年	GPT-2	相比GPT-1在模型规模，训练数据量和性能上都有了很大提升；
2020年5月	GPT-3	诞生之初为当时世界上最大的语言模型，由于强大的性能引发了广泛的研究和商业应用；
2022年11月	ChatGPT	GPT-3的最后一个系列，为用户而开发的模型；
2023年3月	GPT-4	OpenAI发布的最新GPT系列模型，输出任务依旧是一个自回归的单词预测任务，但据估拥有超过10万亿个参数，比ChatGPT更具有普遍性。

可以预见的是，未来的 GPT 系列模型将在规模、性能和多样性上持续提升。GPT 技术的兴起与发展为人工智能领域带来了巨大的机遇和挑战，并对医疗、教育、科研等领域在技术、商业应用和社会影响等方面产生非常重要的贡献，为各行

业带来了多种应用场景，具体有：

- 1) 多模态学习和跨领域应用；
- 2) 低资源语言的支持；

- 3) 模型压缩与高效计算;
- 4) 可解释性与道德伦理等等。



2. 研究的目的与意义

2.1 分析 GPT 技术在医疗信息行业中的应用场景

GPT 技术由于其强大的性能,同样对医疗信息行业产生了强大的冲击,由于语言模型会推动人机交互方式的变化,代码与文档的自动生成,以及数据的自动分析处理,这些改变在本质上会影响所有的信息系统。其中最重要的改变主要在三大类系统上,分别为自然语言记录与检索系统,知识库与临床决策支持系统和临床科研系统这三大类知识(或数据)密集型系统。

首先,自然语言信息记录与检索系统这一类原本与自然语言相关,带有模糊性和个性化的系统会有更大的变化与受益。比如电子病历系统、区域卫生共享平台、专病库系统等等,由于数据格式很难达到共识,给系统的构建和变迁带来困难。而大规模语言模型(Large Language Model, LLM)由于其强大的自然语言理解与生成能力,其衍生出的产物能够实现对于不同格式文档的理解与生成、数据格式的语义转换、数据的自动化清洗等等。更重要的是,这些能力的激发,并不需要大量语料,Zero-shot, Few-shot^①以及一些简单的自然语言提醒就可以达到比较好的效果。

其次,传统的知识库与临床决策支持类系统通常采用人工手段,从医学数据、临床指南和医学文献中抽取知识。这种抽取过程不仅费时费力,而且更新困难,另外加工后的知识库由于应用场景不一,和场景结合时,可能需要二次加工。LLM 对文本数据的解析能力极大方便了知识库系统的构建,而自然语言检索和问答提升了知识库的应用便利性,同时 LLM 的推理能力提供了智能医学诊断和治疗方案推荐的可能性。具体而

言,可以通过 ChatGPT 的开源框架 LangChain^②等系列工具,以及 ChatGPT 的各种 Zero-shot 抽取能力,将原始数据抽取成更为合适的形式。但是,LLM 本身固有的不稳定性以及缺乏严密逻辑推理等问题,也会给临床决策支持带来障碍,可以通过分解决策过程、外接知识库、对推理的每一步进行严格溯源与控制来得到缓解。

最后,在临床科研系统中,LLM 从选题到撰写论文,都将带来流程性的和时效性的重大变革。在传统的选题和构造病历报告表(Case Report Form, CRF)的过程中,需要调研文献,或是在科研数据中心(Research Data Repository, RDR)上分析临床数据找到潜在的方向,这个过程需要人类的思考和假设,但如果将思考过程记录在案,形成思考链的模板,就可以通过对文献的自动数据分析,获得可能的研究方向和字段。同时,对本身就受 AI 影响较大的影像和病理检查系统而言,GPT 为直接写出影像检查文本提供了可能性,但对于 HIS, LIS 以及涉及财务、药品、检验等交易与结算类医疗业务系统中的影响较小。



2.2 探讨 GPT 技术对医疗信息行业发展的影响

首先, GPT 技术对我国的医疗保健系统将产生巨大冲击。自 2017 年国务院提出《新一代人工智能发展规划》以来,我国就已经开始将人工智能作为重点政策进行推动,在近几年的两会中,关于 AI 的相关议题也反复被写入政府的工作报告中。在 2022 年的《“十四五”国民健康规划》中,推出一批融合人工智能等新技术的高质量医疗设备、推动符合条件的人工智能产品进入临床试验成为重点任务,而 GPT-AI 极有可能在未来进入我国,对我国当前固有的医疗保健相关政策法规产生重大冲击。

^① 在深度学习中,不进行样本标注学习,就是零样本学习(Zero-shot learning),如果只有一个标注样本,则称为 One-shot learning。

^② LangChain 是一种 LLMs 接口框架,允许用户围绕大型语言模型快速构建应用程序和管道,它直接与 OpenAI 的 GPT-3 和 GPT-3.5 模型以及 Hugging Face 的开源替代品集成。

其次，GPT 技术带动吸引大量资本进入医疗信息化行业。在过去五年时间里，已有近千亿资金注入医疗人工智能的赛道，AI 已经在医疗中的各个方面落地生根。科技巨头企业，包括微软、谷歌和亚马逊等也分别合作著名医疗机构和健康保险公司，共同研发和推动基于 GPT-AI 的医疗服务和保健方案，旨在提高诊断精准性，优化治疗方案，降低医疗成本，在保护患者隐私的同时提升可评估、可量化的综合医疗保健服务质量。

最后，也是最显而易见的是 GPT 技术将为医疗信息行业带来技术性变革。目前已经应用于临床的 GPT-AI 正在帮助医护人员减少文书工作和繁重的行政事务，把更多时间留给患者，该应用前景令人满意和期待。同时，GPT 技术也有望运用于专病领域，参与患者全病程健康管理，从患者记录中提取基本信息并进行分组处理，帮助医生更高效地评估患者目前的状况与需求。另一方面，GPT 可通过学习大量医疗知识、患者案例，成为医学检索学习的“百科全书”，为基层医学防治提供专业的医学指导。GPT 技术可能会彻底改变医疗信息行业原有的医学研究和 workflows。然而，使用这项技术也存在一些挑战，本文将在第四节进行详细阐述。

二、GPT 技术概述

1. GPT 技术原理与特点

GPT 是由 OpenAI 团队基于 Transformer 的解码器架构，在没有标注的海量数据下训练而成的语言模型，从而实现了与自然语言的高度理解和生成能力。从数学的角度来看，语言模型是对词语序列的概率相关性分布的建模，即利用已有的语句（语句可以视为数学中的向量）作为输入条件，预测下一个时刻不同语句甚至语言集合出现的概率分布。GPT 的核心特点包括：Transformer 架构、预训练和微调、大规模数据训练、生成式模型、高度可扩展性和多任务学习能力等。

在 GPT 的早期版本（GPT-1，GPT-2，GPT-3）中，OpenAI 通过不断扩大训练数据规模和网络结构等方式来提升模型能力，通过 Zero-shot、Few-shot，特别是有 Prompt 情况下，模型的效果很好，并逐渐演变为通用模型，从而实现了与自然语言高度理解和生成的能力。更进一步地，Open-AI 为了提升模型效果，从 GPT-3.5 开始引入人类强化学习技术（RLHF，Reinforcement Learning from Human Feedback）进行训练，这种方法通过人类干预来增强机器学习以获得更好的效果。在训练过程中，人类训练者扮演着用户和人工智能助手角色，并通过近端策略优化算法（PPO，Proximal Policy Optimization）进行微调，从而使模型输出更符合人类的喜好。



2. GPT 在各领域的应用

GPT 技术中的 GPT-3 作为一个语言生成模型，专注于对输入产生适当的文本反应，虽然可以被改编为解决更传统的机器学习任务，进行情感分析，但由于其文本生成能力可以作为各种问题的潜在解决方案所引起了广泛的关注。此外，ChatGPT 和 GPT-4 的出现对于多模态的人工智能生成内容（AIGC，AI Generated Content）应用具有重要意义，与此前 Web1.0、Web2.0 时代的 UGC（用户生产内容）和 PGC（专业生产内容）相比，代表人工智能构思内容的 AIGC，是新一轮内容生产方式的变革，会对 AI 产业上下游产生重大影响，如创建增强型聊天机器人^[8]、回答复杂问题、根据涉及规范生成计算机代码^[9]以及编写新闻文章^[10]。因此，在使用 GPT-3 进行各种文本生成应用方面有很多研究和商业兴趣。然而，在医疗保健中应用 GPT 技术仍然是一个新兴的话题。本文将在后续的第三节和第四节中，对医疗保健环境中整合 GPT 技术和部署此类解决方案所面临的诸多开发和实施挑战进行了讨论分析。

三、GPT 在医疗信息化中的应用

1. 诊断辅助与病历分析

这是 GPT 技术作为卫生系统内一项服务的典型实施。聊天机器人连接到医院网络后，与分流文本摘要服务相结合，GPT 通过对大量医学文献病历和检查报告进行深度学习后，进行总结并存储在电子健康记录中。为了避免或解决潜在的偏见^[11]、纠正错误与增加对病人数据使用和模型的控制，可以通过在前端使用报告反馈机制来实现人在环路模式（human-in-the-loop model）^[12]，或者后端让临床团队监督医院 EHR 系统中的 GPT-3 整合过程。此外，文本中的错误更正和调整可用来微调 GPT-3 模型，以提高其准确性和有效性。与此同时，依靠海量数据的训练，可提示危重病风险、用药及治疗方案等相关建议，并由医生根据经验做出决策，实现“AI + 专家”双重把关，提高诊断效率及准确率。再者，GPT 技术应用在基层、乡镇诊所或可提高全科医生诊断能力，从而提升医疗水平。具

体而言，GPT 技术在这一领域的应用主要包括以下几个方面：初步诊断建议、辅助影像诊断、病历相似度检索、病历自动生成与归档、个性化治疗建议、预后评估与风险预测、临床决策支持、医学知识问答、临床研究辅助、跨学科协作等等。



2. 医学文献检索与研究辅助

在医疗信息化领域中，利用大规模的医学信息和信号来创建能够识别、总结、生成医学和临床内容的人工智能模型是一个长期关注的话题。在过去几年，随着大型语言模型（Large Language Model, LLM）的发展，以及 EHR 和 PubMed 等公共医学文本渠道的丰富，已经提出和开展了 BioLLMs 的研究。自从 Bio BERT^[13]推出以来，一个开创性的 BioLLM 在各种生物医学文本挖掘任务（如生物医学命名实体识别）上的表现超过了既往最先进的水平，许多不同的 BioLLM 源于其一般的 LLM 对应物，包括 Clinical BERT^[14]、BioMegatron^[15]、BioMedRoBERTa^[16]、Med-BERT^[17]、BioELECTRA^[18]、PubMedBERT^[19]、BioLinkBERT^[20]、BioGPT^[21]和 Med-PaLM^[22]。最近拥有 89 亿个参数的 GatorTron^[23]模型用约 820 亿字的临床文本进行了预训练，发现扩大临床 LLM 规模会导致不同医学语言任务的改进，特别是对于一些复杂任务，如医学问题回答和推理，其改进更为显著。Agrawal 等人^[24]发现，最近的 LLMs，如 InstructGPT^[25]和 GPT-3，可以很好地提取临床信息。由于提示工程已经成为研究和改进 LLM 的关键技术，Lievin 等人^[26]也将各种提示工程应用于最近的 GPT-3.5 系列，如 InstructGPT^[25]，以了解其在医学问题回答上的能力。GPT 技术作为 BioLLMs 在医学语言任务上令人印象深刻的表现，显示拥有其用于协助临床研究和医学文献检索的巨大潜力。此外，GPT 能够根据反馈和输入的信息进行学习、分析、归类，根据实际情况向医疗人员生成数据洞察、重要指标统计结果的仪表盘，为接下来的医疗和科研工作指明方向。

3. 自动生成医学文档与报告

编写电子病历作为医生的日常主要任务之一，消耗了医生大量的时间和精力，GPT 有可能通过承担重复性的工作来缓解这一职业倦怠。Peng 等人^[27]通过对海量真实的 EMR 数据训练，研发了 CMedGPT2 模型用于自动生成理想的中文电子病历以辅助医生工作，减轻医生编写负担。Liu 等人开发一个新的基于 GPT-4 的 Deid-GPT 来自动识别和删除识别信息，在掩盖非结构化医疗文本中的私人信息的同时，保留了文本的原始结构和意义，为进一步研究和开发诸如 Chat GPT / GPT-4 等 LLM 在医疗领域的应用提供了启示^[28]。GPT 甚至可以制作进展记录、病人护理计划、随访、投诉处理等服务性工作和其他必须为课堂或病房准备的文件初稿^[29]。因此，医生通过 GPT 解放文书工作，提升了工作效率，并减轻医生的倦怠感，最终让更多患者享受到专业的医疗资源，从而改善医患关系。

4. 患者教育与健康咨询

GPT 技术在改善患者体验与满意度方面有着显著潜力。Lyu 等人研究使用 ChatGPT 将放射学报告翻译成通俗易懂的语言，以供病人和医疗机构使用，以便其接受教育，改善医疗服务^[30]。患者虽然可以通过使用 GPT 的医疗资源受益，但 GPT 给出的建议必须是经过医生过滤的，否则虚假或误导性的信息可能会危及生命安全。从责任角度看，临床医生是否了解人工智能是如何运作的并不重要，重要的是其提供的护理是否达标。AI 产生的不准确建议和专业人员传播错误信息并无不同，当两者都经过专业人员的判断过滤时，意味着临床医生要对结果负责。因此，在彻底验证之前，我们不应该比其他医疗工具更信任 GPT，应该仔细检查和验证 GPT 的输出，以保护病人和自身。



5. 自动化任务处理与团队协作

GPT 技术在自动化任务处理和团队协作方面有着显著的潜力。以下是一些具体的应用场景，说明了 GPT 如何协助医疗团队进行自动化任务处理与协作：自动分配任务、任务进度跟

踪与管理、自动化沟通与协作、文档自动生成与管理、智能提醒与通知、数据分析与决策支持、跨部门与跨机构协作、在线培训与跨机构协作、在线培训与知识共享、虚拟助手与智能问答、基于人工智能的预警与风险管理等等。通过以上应用场景，我们可以看到 GPT 技术在自动化任务处理和团队协作方面具有广泛的应用潜力。然而，GPT 也因为将以前由人类工作者执行的任务自动化，从而减少工作量和人力，同时也创造一些可能替代人类教师和辅导员的工作^[31]。

6. 知识获取与创新思维

随着大型 AI 模型（Large AI Model, LAM）科学基础的研究不断加强，未来的医学教育很可能也会受到 LAM 的影响。ChatGPT 的生成能力可能会增强医学生的学习力，并帮助他们从人工智能生成的内容中获得更多洞察力^[32]。最近的 GPT-4 模型^[33]可作为导师，引导学生一步步自己找到答案，这是 LAM 在教育领域实际应用的重要一步。ChatGPT 等 LLM 的句子转述能力也可以帮助有阅读障碍的学生学习^[34]。可以预见 GPT 技术在知识获取和创新思维方面具有广泛的应用潜力，但也需要警惕对其非法使用。对这一问题，Mitchell 等人^[35]进行了一项试验性研究，提出了一个名为 DetectGPT 的零点检测器，用来区分人类和 LLM 生成的文本。基于对其进行判别和管制的前提下，在未来 GPT 技术有望为医疗行业提供更加高效、精准和个性化的知识获取和创新思维支持，推动医疗行业的发展和创新。



7. 高效获取信息并准确分诊

GPT 可以极大提升在线问诊效率、改善患者体验，甚至可能取代在线医生的工作。GPT4 可以更机动灵活地跟患者对话，并从患者的描述中提取详细的信息，可以完全胜任智能助

手的工作，包括获取和整理患者的症状、基本信息、过往用药史，甚至可以在线提取患者过往拍片中的信息，取代专业医生的部分工作。这样极大缓解了线上沟通的压力，降低了线上沟通成本。以往很多症状线上沟通不畅还需要线上转线下，GPT4 的应用更有利于在线医疗的发展，带来效率提升的颠覆性转变。

四、GPT 技术的挑战与限制

1. 数据安全与隐私保护

医疗保健的数字化促进了医疗数据的共享和再利用，也引起了对保密性和隐私的关注。HIPAA 法案^③规定，在传播医疗记录前，必须删除可重新识别的信息。因此非常需要有效的解决方案来去除医疗数据的身份信息，特别是在以自由文本形式存在的数据中。虽然，各种包括基于规则和基于学习方法的计算机辅助识别方法已经被开发出来并用于先前的实践中，但这些方案仍然缺乏普遍性，或者需要根据不同的情况进行微调，以克服在更广泛的使用中可能会造成的巨大限制。

2. 法规与合规性问题

面对 GPT 和其他 LLM 的这一类潜在技术，极有可能带来新的法律风险或新的监管模式，理解围绕新技术的伦理和法律问题的关键很少关于技术本身，而是关乎这些技术将如何影响用户之间的社会关系^[36]。人们必须首先确定这些关系，定义这些关系的价值观以及相关的法律框架。如果 GPT 被用来增强而不是取代专业判断，那么将其引入临床实践就不太可能影响病人与医生的关系。使用 LLM 就像医生早期采用智能收集、电子医疗记录等等。因此，GPT 可作为常规的专业建议被监管。更令人担忧和在法律上不确定的是面向患者的 GPT 使用，因为这时 AI 脱离了专业关系，例如为病人提供基本的心理健康护理，或取代通常进行分流的诊所工作人员，这时软件可能会取代专业判断。由于缺乏专业监督而产生的责任，以及对准确性和可靠性的担忧，GPT 不应该用于这些角色，但这些风险并没有阻止一些公司在医疗保健的前线和边缘地区使用 GPT。

3. 技术准确性与可靠性

在模型层面，GPT 是在一个大型数据集上训练而来，当这些数据集本身存在许多与种族和性别歧视的刻板印象有关的问题特征时，GPT 模型也会学习某些针对边缘化身份的偏

^③ HIPAA 是“Health Insurance Portability and Accountability Act 1996”的首字母缩写，即健康保障流通与责任法案，提供数据保密和安全条款以保护患者信息。

倚^[37-38]，而其中有些偏倚在临床上很可能是有害的。Korngiebel 和 Mooney^[11]强调了在卫生保健中使用 GPT 的风险，并指出了其很多时候会输出带有偏倚的观点，甚至会肯定自杀的想法^[39]。除在初始训练中可能存在固有偏倚以外，在微调过程中，引入的历史医疗数据也可能存在无意识的偏倚，如对边缘化亚人群的测试不足，会影响基础临床数据，并在预测模型的训练中引入偏倚^[40]。医护人员的隐形偏倚会影响诊断和治疗，并反映在临床记录中^[41]，如果将其用于训练 GPT，将有可能影响所开发的模型。

此外，部署 GPT 还有方法上的难题，Watson 等人^[42]指出模型的可解释性是判断 GPT 是否部署成功的当务之急，理想情况下，应通过使用评价方法和框架来跟踪和评估模型性能。但 GPT 所解决的任务的评估却极为困难，标准指标，如预测灵敏度和阳性预测值并不能明确地反映文本生成的正确性。更加专业的文本生成指标，如 BLEU^[43]和 METEOR^[44]试图说明文本的变化，但仍然只在词的层面上检查文本，没有捕捉到基本含义。试图在文本评价中纳入文本意义的方法，都是依靠其他黑盒深度学习模型来评价黑盒模型，这样并不能提高可解释性^[45]，反而只会导致整体信任度降低，减少 GPT 被部署的可能性。

五、结论与展望

1. GPT 技术在医疗信息化中的积极影响

就 GPT 能力的表层而言，它可以编写处理和可视化数据的计算机程序、翻译外语、为不熟悉每种语言的用户解读社会福利政策说明和实验室检验结果。就其更深层面而言，GPT 技术本身并不是一个目的，而是打开了一扇通往新的可能性以及新的风险的大门，在未来会出现更强大，更有能力的人工智能系统，作为工具积极辅助人类的医疗健康。



2. 面临的主要挑战

首先，GPT 与更为传统的人工智能模型不同，其在内存方面的需求要大得多，且计算量更大。用于模型训练和执行的专

用硬件（包括图形处理单元和张量处理单元）都是可扩展的实施所必需的。对于任何医院系统来说，可能需要对基础设施进行额外投资，以弥补处理需求。

其次，鉴于医院网络和 EHR 系统现状，GPT 技术需要复杂的系统和高技术知识才能有效部署，而且操作成本高昂。为减轻负担，一个潜在的解决方案是与医院系统内的云计算平台整合。许多云计算供应商提供运行此类模型所需的专门硬件，并能轻松处理现有的网络和动态负载平衡，从而减轻 GPT 部署主要组成部分的负担。然而，外包的云计算平台同样有可能会增加运营成本。

最后，GPT 作为一个设计模型，其训练方案的目标是达到通用的认知能力，并未专门为医疗保健应用进行训练。当前 GPT 系统完全是根据从互联网上的开源数据（如公开发表的医学文献、研究论文、卫生系统网站、卫生信息视频等等）进行训练的，训练数据中并不包括诸如医疗机构电子健康记录系统中等任何私人限制的数据，或仅存在于医学院或其他类似组织的私人网络中^[46]。

3. 未来发展趋势与应用前景

在未来，可以从政府、市场、医疗机构三个方面共同努力推动 GPT 技术在医疗信息化领域的发展。首先，在政府层面，需要完善诸如美国 HIPAA 法案等对于使用数据保密和患者隐私的相关规范与政策引导。考虑到政策制定相对滞后于市场，需要一段时间逐步推进，应尽早为 GPT 做出相应的准备，如果没有相关的合规性引导，盲目地将 GPT 应用于医疗行业可能成为一个错误的开始。其次，就市场而言，技术供应商应不断改善对于 GPT 技术作为工具的使用性能，并提供一套完善的评估指标体系。技术公司需要获得政府和医疗机构的充分信任，以获取客户数据来进一步改善服务或实现额外的商业价值。最后，医疗机构需要在临床数据、电子病历的收集整理中最小化可能存在的潜在偏倚，提升数据集本身的质量，以训练出更优质的模型。

虽然目前一些公司已经将 GPT-3 模型预测整合到自身的产品中，但同时也限制了开放性的研究和开发，并最终会限制改进。在未来的迭代中，一旦在医疗环境中使用 GPT-技术的价值得到保证，可及性的责任就可以委托给医疗和政府机构，通过安全的云平台建立一个联合学习机制来运行分散的培训服务^[47]，将减少各个卫生系统在建设，培训和部署自己 GPT 平台的负担，降低成本，这些优势对于低资源环境下的医院尤为有利。总体而言，如果将 GPT 技术应用于医疗信息化领域，其最终将与医疗保健政策和临床医学方法一同演变，互相影响。

诚然，GPT等一类人工智能模型目前在各行各业仍被视为一把“双刃剑”，那么是否应用以及如何将这把“双刃剑”应用于生命健康领域更是需要我们“慎之又慎”！

参考文献

- Zhang M, Li J. A commentary of GPT-3 in MIT Technology Review 2021[J]. Fundamental Research, 2021, 1(6): 831-833.
- Mijwil M, Aljanabi M, Ali A H. ChatGPT: Exploring the Role of Cybersecurity in the Protection of Medical Information[J]. Mesopotamian journal of cybersecurity, 2023, 2023: 18-21.
- Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.
- Ethayarajh K. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings[J]. arXiv preprint arXiv:1909.00512, 2019.
- Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
- Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: Early experiments with gpt-4[J]. arXiv preprint arXiv:2303.12712, 2023.
- Sezgin E, Sirrianni J, Linwood S L. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model[J]. JMIR medical informatics, 2022, 10(2): e32875.
- Langston J. From conversation to code: Microsoft introduces its first product features powered by GPT-3[J]. The AI Blog, 2021.
- GPT-3. A robot wrote this entire article. Are you scared yet, human? [J]. The Guardian, 2020.
- Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery[J]. NPJ Gigit Med, 2021, 4(1):93.
- Miller K. When algorithmic fairness fixes fail: The case for keeping humans in the loop[R]. Stanford University. 2020, URL: <https://hai.stanford.edu/news/when-algorithmic-fairness-fixes-fail-case-keeping-humans-loop>.
- Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.
- Alsentzer E, Murphy J R, Boag W, et al. Publicly available clinical BERT embeddings[J]. arXiv preprint arXiv:1904.03323, 2019.
- Shin H C, Zhang Y, Bakhturina E, et al. BioMegatron: Larger biomedical domain language model[J]. arXiv preprint arXiv:2010.06060, 2020.
- Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: Adapt language models to domains and tasks[J]. arXiv preprint arXiv:2004.10964, 2020.
- Rasmy L, Xiang Y, Xie Z, et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction[J]. NPJ digital medicine, 2021, 4(1): 86.
- Raj Kanakarajan K, Kundumani B, Sankarasubbu M. BioELECTRA: pretrained biomedical text encoder using discriminators[C]//Proceedings of the 20th Workshop on Biomedical Language Processing. 2021: 143-154.
- Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing[J]. ACM Transactions on Computing for Healthcare (HEALTH), 2021, 3(1): 1-23.
- Yasunaga M, Leskovec J, Liang P. Linkbert: Pretraining language models with document links[J]. arXiv preprint arXiv:2203.15827, 2022.
- Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining[J]. Briefings in Bioinformatics, 2022, 23(6).
- Singhal K, Azizi S, Tu T, et al. Large Language Models Encode Clinical Knowledge[J]. arXiv preprint arXiv:2212.13138, 2022.
- Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records[J]. npj Digital Medicine, 2022, 5(1): 194.
- Agrawal M, Hegselmann S, Lang H, et al. Large language models are few-shot clinical information extractors[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 1998-2022.
- Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.
- Liévin V, Hother C E, Winther O. Can large language models reason about medical questions? [J]. arXiv preprint arXiv:2207.08143, 2022.
- J. Peng, Ni P, Zhu J et al. Automatic Generation of Electronic Medical Record Based on GPT2 Model[C]//Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), 2019:6180-6182.
- Liu Z, Yu X, Zhang L, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4[J]. arXiv preprint arXiv:2303.11032, 2023.
- Haupt CE, Marks M. AI-Generated Medical Advice—GPT and Beyond[J]. JAMA. Published online March 27, 2023. doi:10.1001/jama.2023.5321
- Lyu Q, Tan J, Zapadka M E, et al. Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: Promising results, limitations, and potential[J]. arXiv preprint arXiv:2303.09038, 2023.
- Qadir J. Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education[J]. 2022.
- Kung T H, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models[J]. PLoS digital health, 2023, 2(2): e0000198.
- O. Gpt-4 technical report[R]. arXiv:2303.08774, 2023.

34. Dai H, Liu Z, Liao W, et al. Chataug: Leveraging chatgpt for text data augmentation[J]. arXiv preprint arXiv:2302.13007, 2023.
35. Mitchell E, Lee Y, Khazatsky A, et al. Detectgpt: Zero-shot machine-generated text detection using probability curvature[J]. arXiv preprint arXiv:2301.11305, 2023.
36. Haupt C E. Artificial professional advice[J]. Yale JL & Tech., 2019, 21: 55.
37. Lucy L, Bamman D. Gender and representation bias in GPT-3 generated stories[C]//Proceedings of the Third Workshop on Narrative Understanding. 2021: 48-55.
38. Bender E M, Gebru T, McMillan-Major A, et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? [C]//Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 2021: 610-623.
39. Daws R. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves[J]. AI News, 2020.
40. Wiens J, Price W N, Sjoding M W. Diagnosing bias in data-driven algorithms for healthcare[J]. Nature Medicine, 2020, 26(1): 25-26.
41. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review[J]. BMC medical ethics, 2017, 18(1): 1-18.
42. Watson J, Huttyra C A, Clancy S M, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? [J]. JAMIA open, 2020, 3(2): 167-172.
43. Papineni K, Roukos S, Ward T, et al. BLEU: a Method for Automatic Evaluation of Machine Translation [C]//Proceeding of the 40th Annual Meeting of the ACL.2002:311-318.
44. Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. 2005: 65-72.
45. Celikyilmaz A, Clark E, Gao J. Evaluation of text generation: A survey[J]. arXiv preprint arXiv:2006.14799, 2020.
46. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine[J]. New England Journal of Medicine, 2023, 388(13): 1233-1239.
47. Sadilek A, Liu L, Nguyen D, et al. Privacy-first health research with federated learning[J]. NPJ digital medicine, 2021, 4(1): 132.

降低成本

提升医护体验

智慧医院

改善患者体验

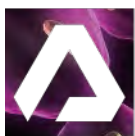
微服务分布式架构

提高医疗质量

医院高质量发展

国际标准

搭乘医院信息系统赛车
智慧医疗一路驰骋



Article Abstract Collection

文献摘要

参考文献摘要翻译

1. A commentary of GPT-3 in MIT Technology Review 2021

麻省理工学院 2021 年技术审查中对 GPT-3 的述评

通过开发具有协作和对话能力的大规模自然语言模型，人工智能（AI）已经向更好的自然语言理解（NLU）和人机交互（HCI）迈出一大步。截至目前，由 OpenAI 开发的 GPT-3 模型是参数最多、规模最大、能力最强的语言模型。通过利用大量的互联网文本数据和数以千计的文献进行模型训练，GPT-3 可以几乎完美地模仿人类的自然语言模式。由于极为逼真，GPT-3 被认为是截止目前最令人深刻的模型。

虽然 GPT-3 具有强大的建模和描述能力，但也存在很大的问题和局限性。首先，GPT-3 模型不能很好地理解写作（自然语言生成），有时会产生不可控的内容。其次，训练 GPT-3 模型需要大量的算力、数据和资本投入，并会产生大量的二氧化碳排放。开发此类模型只有在拥有足够资源的实验室才可能实现。此外，由于 GPT-3 模型是用充斥着错误信息和偏倚的互联网文本数据来训练的，经常会产生与训练数据一样带有偏倚的章节和段落。



2. A Large Language Model for Electronic Health Records

为电子病历而生的大型语言模型

人们对开发人工智能（AI）系统来处理 and 解释电子健康记录（EHR）的越来越感兴趣。由预训练的语言模型驱动的自然语言处理（NLP）是利用临床叙述的医疗人工智能系统的关键技术。然而临床语言模型则很少，其中在临床领域训练的最大语言模型只有 1.1 亿个参数（与一般领域的数十亿个参数相比相对较小）。目前还不清楚具有数十亿个参数的大型临床语言模型如何帮助医疗 AI 系统利用非结构化的 EHR。在这项研究中，我们利用超过 900 亿字的文本（包括超过 820 亿字的去识别化的临床文本）从头开始开发了一个大型临床语言模型-GatorTron，并在五个临床 NLP 任务中进行了系统的评估，包括临床概念提取、医学关系提取、语义文本相似性、自然语言推理（NLI）和医学问答（MQA）。我们研究了（1）扩大参数数量和（2）扩大训练数据规模如何有利于这些 NLP 任务。GatorTron 模型将临床语言模型从 1.1 亿个参数扩展到 89 亿个参数，并改善了五个临床 NLP 任务（例如，NLI 和 MQA 的准确性提高了 9.6% 和 9.5%），这可以应用于医疗 AI 系统以改善医疗服务。GatorTron 模型的公开网站 https://catalog.ngc.nvidia.com/orgs/nvidia/teams/clara/models/gatortron_og。

3. AI-Generated Medical Advice-GPT and Beyond

人工智能生成的医疗建议 —— GPT 及其他

多年来，专家们一直在猜测人工智能（AI）在医疗保健领域的未来作用。一些人工智能工具在放射学、皮肤学和其他领域的特定任务上可以胜过医生，这引起了人们

对人工智能可能使某些专家过时的担忧。一些人担心人工智能可能会使病人和临床医生面临新的风险。还有人想知道，如果医生不了解人工智能的工作原理，他们是否能凭良心使用人工智能，或者未能采用人工智能的临床医生是都会被指控提供不合理的医疗服务。



4. Attention is All you Need

注意力是你所需的一切

当前占据主导地位的序列转导模型，是基于复杂的递归或卷积神经网络的编码器-解码器配置而成的。性能最好的模型通过注意力（Attention）机制将编码器和解码器连接起来。本文提出了一个新的简单网络结构—转化器（Transformer），只基于注意力机制，完全不需要递归和卷积。通过在两个机器翻译任务上的实验表明，这些模型在质量上更胜一筹，同时具有更高的并行性，所需的训练时间也大大缩减。本文中的模型在 WMT 2014^① 的英德翻译任务中达到了 28.4 BLEU^②，比现有最佳结果（包括合集）提高了 2 BLEU 以上。

5. Artificial Professional Advice

人工的专业建议

提供专业建议意味着什么？当各种形式的技术，如决策支持软件或预测性建议生成算法，被插入到专业建议的提供过程中时，情况会有什么变化？专业建议对客户来说是有价值的，因为非专业人士和专家的知识是不对称的，专业人员拥有客户所缺乏的知识。但技术正在日益改变传统的专业建议提供过程。

这篇文章考虑了将人工智能（AI）引入到医患关系中。医学建议的技术创新发生在一个密集的监管空间。管理

专业咨询的法律框架是为了保护医患关系背后的价值观。本文首先概述了专业咨询的监管情况，重点关注现有法律框架所保护的价值观。然后考虑了对提供咨询关系的各种技术干预，确定了由此产生的变化。最后概述了旨在将基于人工智能的创新融入医疗建议的法律对策，同时维护专业建议提供关系的基本价值观。如果现有的监管框架能够应对这些变化，那么就应该保留下来。但当人工智能引入医疗建议的过程中，关系的动态变化威胁到根本的价值观时，就需要新的监管对策。

6. AugGPT: Leveraging ChatGPT for Text Data Augmentation

AugGPT: 利用 ChatGPT 进行文本数据增援

在许多自然语言处理（NLP）任务中，文本数据增强是克服有限样本量挑战的有效策略。这一挑战在小样本学习的场景中尤为突出，因为目标领域的数据通常要少得多且质量较低。缓解这种挑战的一个自然和广泛使用的策略是进行数据扩增，以更好地捕捉数据的不变性并增加样本量。然而，目前的文本数据增强方法要么不能确保生成的数据正确标记（缺乏忠实性），要么不能确保生成的数据有足够的多样性（缺乏紧凑性），或者两者兼有。受最近大型语言模型的成功启发，特别是 ChatGPT 的开发，其展示了改进的语言理解能力，本文提出了一种基于 ChatGPT 的文本数据增强方法（名为 AugGPT）。AugGPT 将训练样本中的每个句子重写为多个概念上相似但语义上不同的样本。然后这些增强的样本可以用于下游的模型训练。在少量学习文本分类任务上的实验结果表明，与最先进的文本数据增强方法相比，所提出的 AugGPT 方法在测试精度和增强样本的分布方面有更好的性能。



^① WMT2014 数据集是从统计学机器翻译研讨会（Workshop on Statistical Machine Translation）评测中产生，于 2014 年发布，数据以新闻为主，也包含医疗相关的语料。

^② Bleu 是 IBM 在 2002 年提出，用于评价机器翻译任务的指标，后续出现许多变种，但总体思想就是统计学中分类混淆矩阵的查准率。本系列中第 11 篇为 Bleu 文献的翻译摘要。

7. BioBERT: a pre-trained biomedical language representation model for biomedical text mining

BioBERT: 用于生物医学文本挖掘的预训练生物医学语言表示模型

随着生物医学文献数量的迅速增长，生物医学文本挖掘变得越来越重要。随着自然语言处理（NLP）的进步，从生物医学文献中提取有价值的信息已经得到了研究人员的青睐，而深度学习也推动了有效的生物医学文本挖掘模型的发展。然而，直接将 NLP 的进展应用于生物医学文本挖掘，往往会产生不尽如人意的结果，这是因为词汇分布从通用领域的语料库转变到生物医学语料库。本文研究了最近引入的预训练语言模型 BERT 如何适用于生物医学语料。我们介绍了 BioBERT，这是一个针对特定领域的语言表示模型，在大规模生物医学语料库上进行了预训练。在不同任务中，BioBERT 的架构几乎是相同的，当在生物医学语料库上进行预训练时，BioBERT 在各种生物医学文本挖掘任务中的表现基本上超过 BERT 和以前最先进的模型。虽然 BERT 获得的性能与以往最先进的模型相当，但在以下三个代表性的生物医学文本挖掘任务中明显优于它们：生物医学命名实体识别（0.62% 的 F1 分数提高），生物医学关系提取（2.80% 的 F1 分数提高）和生物医学问题回答（12.24% 的 MRR 提高）。我们的分析结果表明，在生物医学语料库中预先训练 BERT 有助于理解复杂的生物医学文本。

8. BioELECTRA: Pretrained Biomedical text Encoder using Discriminators

使用判别器重新训练的生物医学文本编码器

最近在 NLP 预训练策略方面的进展表明，模型在各种文本挖掘任务上的表现有了明显的改善。我们应用 ELECTRA 提出的“替换标记检测”预训练技术，利用生物医学文本和词汇从头预训练生物医学语言模型。本文介绍了 BioELECTRA，一个针对生物医学领域的语言编码器模型，使 ELECTRA 适用于生物医学领域。在 BLURB^③和 BLUE 生物医学 NLP 基线上评估我们的模型。在 BLURB 基线的所有 13 个数据集和 BLUE 基线的所有 4 个临床数据集上，BioELECTRA 在 7 个不同的 NLP 任务上的表现优于以前的模型，达到了最先进的水平

（SOTA）。BioELECTRA 在 PubMed 和 PMC 全文章上进行了预训练，在临床数据集上也表现得非常好。BioELECTRA 在 MedNLI 上取得 86.34%（准确率提高 1.39%）的新 SOTA 分值，在 PubMedQA 数据集上取得分值为 64%（准确率提高 2.98%）。

9. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining

BioGPT: 用于生物医学文本生成和挖掘的生成式预训练转化器

预训练的语言模型因其在通用自然语言领域的巨大成功，在生物医学领域引起了越来越多的关注。通用语言领域的预训练模型主要存在两个分支，即 BERT（及其变体）和 GPT（及其变体），第一个分支已经在生物医学领域得到了广泛的研究，如 BioBERT 和 PubMedBERT。虽然在各种鉴别性的下游生物医学任务上取得了巨大的成功，但由于缺乏生成能力，限制了其应用范围。本文中提出了 BioGPT，一个针对特定领域的生成式转换器语言模型，在大规模的生物医学文献上进行了预训练。我们在六个生物医学 NLP 任务上评估了 BioGPT，并证明本文的模型在大多数任务上优于以往的模型。特别在 BC5CDR，KD-DTI 和 DDI 的端到端关系提取任务上分别取得 44.98%，38.42%和 40.76%的 F1 得分，在 PubMedQA 上得到 78.2%的准确率，创造了一个新的记录。我们关于文本生成的案例研究进一步证明了 BioGPT 在生物医学文献上的优势，即为生物医学术语生成流畅的描述。代码可在 <https://github.com/microsoft/BioGPT> 网址上找到。

10. BioMegatron: Larger Biomedical Domain Language Model

BioMegatron: 大型生物医学领域语言模型

针对生物医学领域语言模型大量涌现的这一现象，表明在生物医学文本上预训练的语言模型在生物医学领域的基线上比在通用领域文本语料（如维基百科和书籍）上训练的语言模型表现更好。然而大多数工作并没有深入研究影响每个领域语言应用的因素。此外，对特定领域模型规模的研究也几乎没有。本文实证研究并评估了几个可能影响领域语言应用性能的因素，如子词汇集、模型大小、预训练语料库和领域转移。展示了在较大的领

^③ BLURB 是生物医学语言理解和推理基准（Biomedical Language Understanding and Reasoning Benchmark）的缩写，即为 BLURB 的命名规则，包括 13 个公开可用的数据集，涉及 6 个不同的任务。

域语料库上训练的大型 BioMegatron 模型在基线上的一致改进，有助于我们对领域语言模型应用的理解。在命名实体识别、关系提取和问题回答的标准生物医学 NLP 基线上，展示了比以往最先进模型（SOTA）的明显改进。模型检查点和代码可在 <https://catalog.ngc.nvidia.com/?filters=&orderBy=weightPopularASC&query=> 和 <https://github.com/NVIDIA/NeMo> 获得。

11. **Bleu: a Method for Automatic Evaluation of Machine Translation**

Bleu: 一种自动评估机器翻译的方法

对机器翻译的人工评估往往是广泛而昂贵的。人工评估可能需要几个月的时间才得以完成，且涉及到不能重复使用的人力。本文提出了一种自动机器翻译评估方法，该方法快速、廉价、与语言无关、与人工评估高度相关，而且每次运行的边际成本很小。我们将这种方法作为熟练的人类评判员的自动替补，在需要快速或频繁的评估时替代他们。

12. **Considering the Possibilities and Pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery**

探讨生成式预训练转化器 3 (GPT-3) 在医疗服务中的可能性和隐患

自然语言的计算机应用正变得越来越复杂，随着最近生成式预训练转化器 3 的发布，可以被部署在由人与人之间交互组成的医疗保健相关的背景中。然而，考虑将 GPT-3 和类似的应用程序用于与健康相关的环境，需要对可能性和陷阱进行深思熟虑的探索。本文中我们简要介绍了部署在电子医疗中关于高级自然语言处理应用的一些机会和注意事项。

13. **Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing**

用于生物医学自然语言处理的特定领域语言模型预训练

预训练大型神经语言模型如 BERT，已经在许多自然语言处理（NLP）任务中取得了令人印象深刻的成果。然而，大多数预训练工作都集中在通用领域的语料库上，如新闻和网络。一个普遍的假设是，即使是特定领域的预训练也可以从一般领域的语言模型中受益。在本文中我们对这一假设提出了挑战，表明对于具有丰富未标记

文本的领域，如生物医学，从头开始预训练语言模型会比持续预训练通用领域的语言模型带来更大受益。为了促进这项调查，我们从公开的数据集中编制了一个全面的生物医学 NLP 基准。我们的实验表明，特定领域的预训练为广泛的生物医学 NLP 任务打下了坚实的基础，引导了全面而先进的结果。此外，在对预训练和特定任务微调的建模选择进行彻底评估时，我们发现一些常见的做法对于 BERT 模型来说是不必要的，例如在命名实体识别（NER）中使用复杂的标签方案。为了帮助加速生物医学 NLP 的研究，我们已经为社区发布了最先进的预训练和特定任务的模型，并在 <https://microsoft.github.io/BLURB> 网址上创建了一个以本文中的 BLURB 基线（生物医学语言理解与推理基线）为特色的排行榜。



14. **Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education**

ChatGPT 时代的工程教育：用于教育的生成式人工智能的希望和陷阱

工程教育正在不断发展，以跟上最新的技术发展，满足工程行业不断变化的需求。这个领域中一个有前途的发展是使用生成式人工智能技术，如 ChatGPT 对话代理。ChatGPT 有可能提供个性化和有效的学习经验，为学生提供定制化的反馈和解释，并为动手学习创造现实的虚拟模拟。然而，重要的是也要考虑这项技术的局限性。ChatGPT 和其他生成性人工智能系统只和它们的训练数据一样好，可能会延续偏倚，甚至产生和传播错误信息。此外，在教育中使用生成式人工智能会引起道德方面的担忧，比如学生可能会不道德或不诚实地使用，以及人类可能会因为技术而被裁员。虽然以 ChatGPT 为代表的生成式人工智能技术的现状令人印象深刻，但也有缺陷，这只是未来的一个预演。工程教育者必须了解这项技术的影响，并研究如何调整工程教育生态系统，以确保下一代工程师能够利用生成式人工智能提供的好处，同时尽量减少任何负面影响。

15. From Conversation to code: Microsoft Introduces its First Products Features Powered by GPT-3

从对话到代码：微软推出首批由 GPT-3 支持的产品功能

在 Build 开发者大会上^④，微软公布了由 OpenAI 开发的强大自然语言模型 GPT-3 驱动的用户产品首批功能，这将帮助用户在不需要知道如何编写计算机代码或公式的情况下构建应用程序。

GPT-3 将被整合到微软的 Power Apps 中，这是一个低代码的应用开发平台，可以帮助没有或少有编码经验的人——即“素人开发者”到具有深厚编程专业知识的职业开发者建立应用，以提高企业生产力或流程。这包括审查非营利性礼品捐赠的应用程序，在 COVID-19 期间管理旅行或减少维护风力涡轮机所需的加班。

例如，新的人工智能驱动的功能将允许构建电子商务应用程序的员工使用对话式语言描述一个编程目标，如“找到名称以‘孩子’开头的产品”。然后一个经过微调的 GPT-3 模型提供选择，将命令转化为微软的 Power Fx 公式，即 Power 平台的开源编程语言，如“Filter (‘BC Orders’ Left (‘Product Name’, 4) =” Kids”)”。

微软表示，这是展示 GPT-3 如何在微软 Azure^⑤上运行并由 Azure 机器学习提供支持的首批实施方案之一，也是其新的管理端点能力的首批内部使用方案之一，可以解决企业规模上的实际业务需求。

16. Gender and Representation Bias in GPT-3 Generated Stories

GPT-3 生成文本中的性别和代表性偏见

使用主题建模和基于词库的词汇相似性，我们发现由 GPT-3 生成的故事表现出许多已知的性别定型观念。根据 GPT-3 对提示中人物性别的认知，生成的故事描绘出不同的主题和描述，女性角色更可能与家庭和外表相关，并即使在提示中与高力量的动词相关，仍被描述为比男性角色更缺少力量。我们的研究提出了这样的疑问：在使用大型语言模型生成文本故事时，如何才能避免无意识的社会偏见。

17. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings

语境化的单词表征是怎样的？比较 BERT, ELMo 和 GPT-2 嵌入的几何结构

通过使用语境化的单词表征取代静态的单词嵌入，让许多 NLP 任务得到了显著的改善。然而 ELMo 和 BERT 等模型所产生的上下文表征到底有多大的关联？每个词是否有无限多的特定语境表征，或者说词基本上是被分配到有限数量的单词表征中的一个？首先，我们发现，所有单词的语境化表征在语境化模型的任何一层都不是各向同性的。虽然同一个词在不同语境中的表征仍然比两个不同词的表征有更大的余弦相似性，但这种自身相似性在上层要低得多。这表明上下文模型的上层产生更多的上下文特定表征，就像 LSTM 的上层产生更多的任务特定表征。在 ELMo, BERT 和 GPT-2 的所有层中，平均来说，一个词的上下文表征中只有不到 5% 的方差可以用该词的静态嵌入来解释，这为上下文表征的成功提供了一些理由。



18. Implicit Bias in Healthcare Professionals: a Systematic Review

有关医护人员内隐偏倚的系统性回顾

背景：内隐偏倚是指在意识之外的联想，导致对一个人基于无关特征（如种族或性别）的负面评价。本文研究了医护人员对病人表现出内隐偏倚的证据。

方法：检索了 PubMed、PsychINFO、PsychARTICLE 和 CINAHL 中 2003 年 3 月 1 日至 2013 年 3 月 31 日期间经同行评审发表的文章。两名审稿人根据精确的内容和质

④ 由微软公司于 2021 年线上举办，其目的并非指定新的规则或创建成果，而是希望帮助开发者实现自己的创新。

⑤ 微软 Azure 是微软的云平台，在中国通过授权给世纪互联由其代为运营，微软承诺中国版的公有云服务于全球版的公有云服务提供一样的 SLA，可高达 99.95%。

量标准评估确定的论文资格。对符合条件论文的参考文献进行了审查，以确定更多符合条件的研究。

结果：其中有 42 篇文章被确认为符合条件。17 篇文章使用了内隐测量（15 篇为内隐联想测试，2 篇为潜意识引诱），以测试医护人员的偏倚。25 篇文章采用了受试者之间的设计，用小故事来研究病人的特征对医护人员的态度，诊断和治疗决定的影响。第二种方法被包括在内，尽管因被专门研究内隐认知心理学界认为是检测可能存在内隐偏倚的一种方法，而不能脱离内隐态度。27 项研究考察了种族（或民族）偏倚；还调查了其他 10 中偏倚，包括性别、年龄和体重。35 篇文章发现了医护人员存在内隐偏倚的证据；所有调查相关性的研究都发现内隐偏倚水平和较低的护理质量之间存在明显的正相关。

讨论：证据表明，医护人员表现出的内隐偏倚水平与更广泛的人群相同。多种患者特征之间以及医护人员与患者特征之间的相互作用揭示了内隐偏倚现象的复杂性及其对临床医生与患者互动的影响。在本文的回顾中，最有说服力的研究是结合了 IAT[®]和测量实际治疗质量的方法。相关证据表明，偏倚在某些情况下可能会影响诊断和治疗的决定以及护理水平，但需要进一步调查。本文还发现在一些被测试的特征中，公正性的规范与医护人员接收公正性的程度之间有时可能存在着差距。

结论：我们的研究结果强调了医护行业需要解决的内隐偏倚在医疗保健中的作用。需要在实际的护理环境中进行更多的研究，并在测试医疗保健中的内隐偏倚时采用更多的同质性方法。

19. Improving Language Understanding by Generative Pre-Training

通过生成式预训练提高语言理解能力

自然语言理解包括如文本相关性分析、问答、语义相似性评估和文档分类等一系列不同的任务。虽然大量的无标签文本语料库很丰富，但用于学习这些特定任务的标签数据却很稀少，这使得经过鉴别性训练的模型很难有充分的表现。我们证明，通过在不同的无标签文本语料库中对语言模型进行生成式预训练，然后对每个特定任务进行鉴别性的微调，可在这些任务上取得巨大的受益。与以往的方法相比，我们在微调的过程中利用了任

务意识的输入转换来实现有效的转移，同时要求对模型结构进行最小的改变。在广泛的自然语言理解基础上证明了我们方法的有效性。我们的一般任务诊断模型优于使用专门为每项任务设计的架构的鉴别性训练模型，在所研究的 12 项任务中，有 9 项明显地改善了技术水平。例如，我们在常识推理方面取得了 8.9% 的绝对改进，在问题回答方面取得了 5.7% 的绝对改进，在文本关联方面取得了 1.5% 的绝对改进。

20. Language Models are Few-Shot Learners

语言模型是小样本学习者

最近的研究工作表明，在许多的 NLP 任务和基线上，通过对大型文本语料库进行预训练和特定任务的微调，可获得巨大的收益。虽然在结构上通常与任务无关的，但这种方法仍然需要几千或几万个样本用于特定任务的微调数据集。与其相反的是，人类通常可以从几个例子或简单的指令中完成一项新的语言任务，而目前的 NLP 系统在很大程度上仍然难以做到这一点。本文展示了通过扩大语言模型的规模，大大改善与任务无关的、少量的性能，有时甚至达到可与之前最先进的微调方法相匹配的竞争力。具体而言，我们训练了 GPT-3，一个具有 1750 亿个参数的自回归语言模型，比以往的任何非稀疏语言模型多 10 倍，并测试了其在少数情况下的性能。对于所有任务，GPT-3 的应用没有任何梯度更新或微调，纯粹通过与模型的文本交互来指定任务和少量演示。GPT-3 在许多 NLP 数据集上取得了强大的性能，包括翻译、问答和 cloze 任务，以及一些需要即时推理或领域适应的任务，如解读单词在句子中使用一个新词或进行 3 位数的运算。同时，我们也发现 GPT-3 在一些数据中的小样本学习仍然困难，面临着与大型网络语料库训练有关的方法学问题。最后我们发现 GPT-3 可以生成人类评价者难以区分的新闻文章样本。讨论了这一发现和 GPT-3 总体上的更加广泛的社会影响。

21. Large Language Models Encode Clinical Knowledge

为电子病历而生的大型语言模型

大型语言模型（LLM）在自然语言理解和生成方面表现出令人印象深刻的能力，但对医疗和临床应用的质量要求很高。今天评估模型的临床知识尝试通常依赖于对有限基准的自动评估。目前还没有一个标准来评估模型的

[®] IAT 为内隐联想测验（Implicit Association Test），以反应时为指标，通过一种计算机化的分类任务来测量两类词（概念词和属性词）之间自动化联系的紧密程度，继而对个体的内隐态度等内隐社会认知进行测量。

预测和推理的广泛任务。为了解决这个问题，我们提出了 MultiMedQA，一个结合了六个现有公开问答数据集的基线，涵盖了专业体检、研究和消费者查询以及 HealthSearchQA，新的在线搜索医学问题的自由回答数据集。我们提出一个框架用于沿着多个轴线对模型答案进行人工评估，包括事实性、精确性、可能的伤害和偏见。此外，我们还在 MultiMedQA 上评估了 PaLM（540 亿个参数的 LLM）及其指令调整的变体 Flan-PaLM。通过结合是同提示策略，Flan-PaLM 在 MultiMedQA 的每一个多选数据集（MedQA，MedMCQA，PubMedQA，MMLU 临床题目）上都达到了最先进的准确率，包括在 MedQA（美国医学执照考试题目）上达到 67.6% 的准确率，超过之前最先进水平 17% 以上。然而，人类评估揭示了 Flan-PaLM 反应中的关键差距。为解决这个问题，我们引入了指令提示调整，这是一种参数高效的方法，可以使用一些范例将 LLM 与新的领域对齐。由此产生的模型 Med-PaLM 表现令人鼓舞，但仍比临床医生差。表明理解力、知识回忆和医学推理随着模型规模和指令提示的调整而改善，这表明 LLM 在医学中的潜在效用。本文的人类评估揭示了当今模型的重要局限性，加强了评估框架和方法开发在为临床应用创造安全、有益的 LLM 模型方面的重要性。



22. LinkBERT: Pretraining Language Models with Document Links

LinkBERT：用文档链接预训练语言模型

语言模型（LM）预训练可以从文本语料库中学习各种知识，帮助下游任务。然而现有方法如 BERT，只对单一文件进行建模，并没有捕捉到跨文件的依赖关系或知识。在这项工作中，我们提出了 LinkBERT，一种利用文档之间的链接（如超链接）的 LM 预训练方法。给定一个文本语料库，我们将其视为一个文件图，并通过将链接的文件放在同一语境中来创建 LM 输入。然后用两个共同的自我监督目标对 LM 进行预训练：掩码语言建

模和我们的新建议文档关系预测。我们表明，LinkBERT 在两个领域的各种下游任务中的表现优于 BERT：通用领域（在带有超链接的维基百科上进行预训练）和生物医学领域（在带有引文链接的 PubMed 上进行预训练）。LinkBERT 对多跳推理和少见的 QA 特别有效（在 HotpotQA 和 TriviaQA 上的绝对改进为 5%），而我们的生物医学 LinkBERT 在各种 BioNLP 任务上设定了新的技术状态（在 BioASQ 和 USMILE 上为 7%）。我们在这个 https://github.com/michiya_sunaga/LinkBERT 网址上发布了预训练模型，LinkBERT 和 BioLinkBERT 的代码和数据。

23. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction

Med-BERT：在大规模的结构化电子健康记录上对疾病预测进行情景化嵌入训练

基于深度学习（DL）的电子健康记录（EHR）预测模型在许多临床任务中提供了令人深刻的性能。然而这些模型往往需要大量的训练队列来实现高精度，这阻碍了在训练数据有限的情况下采用基于 DL 的模型。最近由转化器的双向编码器表示法（BERT）和相关模型在自然语言处理领域取得了巨大的成功。BERT 在一个非常大的训练语料库上的预训练产生了语境化的嵌入，可以提高在较小数据集上训练的模型的性能。受 BERT 启发，我们提出了 Med-BERT，将最初为文本领域开发的 BERT 框架调整到结构化的 EHR 领域。Med-BERT 是一个在 28490650 名病人的结构化 EHR 数据集上预训练的上下文嵌入模型。微调实验表明，Med-BERT 大幅提高了预测的准确性，在两个临床数据库的两个疾病预测任务中，接收者操作特征曲线下面积（AUC）提高了 1.21-6.14%。特别是与没有 Med-BERT 的深度学习模型相比，预训练的 Med-BERT 在具有小的微调训练集的任务中训练的模型一样高的 AUC。我们相信，Med-BERT 将有利于具有小型本地训练数据集的疾病预测研究，减少数据收集费用，并加快人工智能辅助医疗的步伐。

24. On the Danger of Stochastic Parrots: Can Language Model Be Too Big?

论鹦鹉学舌的危险：语言模型会不会大而不实？

在过去三年中，NLP 工作的特点是开发和部署特别对英语越来越大的语言模型。BERT，及其变体 GPT-2/3 以及

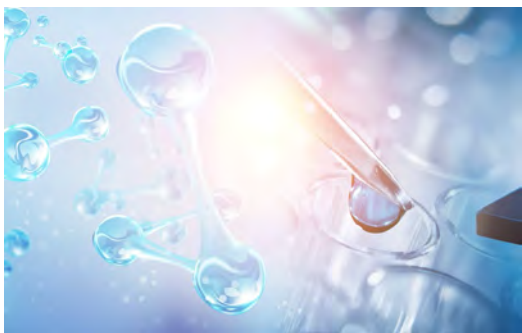
其他等等，和最近的 Switch-C 通过架构创新和庞大规模，推动了可能的边界。使用这些预训练模型和为特定任务微调模型的方法，研究人员已经在一系列的任务上扩展了技术水平，并通过英语的特定基准的排行榜来衡量。本文中我们退一步提问：多大才算大？与这项技术相关的可能风险是什么，有什么途径可以减轻这些风险？

本文提供的建议包括：首先权衡环境和财务成本，将资源投入到策划和仔细记录数据集上，而不是摄取网络上的一切，开展预开发工作，评估计划的方法如何符合研究和开发目标，并支持利益相关者的价值观，以及鼓励研究方向超越越来越大的语言模型。

25. Performance of ChatGPT on USMLE: Potential for AI-assisted Medical Education Using Large Language Models

USMLE 考试中 ChatGPT 的表现：使用大型语言模型进行人工智能辅助医学教育的潜力

我们评估了一个名为 ChatGPT 的大型语言模型在美国医学执照考试（USMLE）中的表现，该考试由三个考试组成：步骤 1，步骤 2CK^⑦和步骤 3。ChatGPT 在没有任何专门训练或强化的情况下，在所有三次考试中的表现都达到或接近通过门槛的水平。此外，ChatGPT 在其解释中表现出高水平的一致性和洞察力。这些结果表明，大型语言模型可能有潜力协助医学教育，并可能协助临床决策。



26. Privacy-first Health Research with Federated Learning

以联邦学习为基础的隐私优先健康研究

在进行健康研究时，隐私保护是最重要的。然而研究往往依赖于存储在集中式存储库中的数据，在进行分析时

可以完全接触到敏感的基本内容。联邦学习的最新进展使得建立复杂的机器学习模型成为可能，这些模型以分布式方式进行训练。这些技术有助于计算研究的终点，使私人数据永远不会离开特定的设备或医疗系统。通过在一组不同的单站点和多站点健康研究中显示，联邦模型可以实现类似的准确性、精确性和可推广性，并导致与标准的集中式统计模型相同的解释，同时实现更强的隐私保护，且没有显著提高计算成本。这项工作第一次将现代和通用的联邦学习方法应用于临床和流行病学研究—跨越联邦单位、模型架构、学习任务的复杂性和疾病的范围。因此，使健康研究参与者能够克服以往的矛盾点，保持对其数据的控制同时仍有助于推动科学的发展。

27. Publicly Available Clinical BERT Embeddings

开源的临床 BERT 嵌入模型

近几个月来，ELMo（Peters 等人，2018）和 BERT（Devlin 等人，2018）等语境词嵌入模型极大地提高了许多自然语言处理（NLP）任务的性能。然而这些模型在专业语料库（如临床文本）上的探索很少；此外，在临床领域，还没有公开可用的预训练 BERT 模型存在。在这项工作中，通过探索和发布临床文本的 BERT 模型来满足这一需求：一个用于一般的临床文本，另一个专门用于出院的总结。本文证明与非特定的嵌入相比，使用特定领域的模型在三个常见的临床 NLP 任务上产生了性能上的改进。这些特定领域的模型在两个临床识别任务上的表现不尽如人意，并认为这是去识别的源文本和合成的非去识别任务文本之间差异的自然结果。

28. Sparks of Artificial General Intelligence: Early experiments with GPT-4

通用人工智能的火花：GPT-4 的早期实验

由于大型语言模型（LLM）在各种领域和任务中表现出非凡的能力，挑战着我们对学习和认知的理解，人工智能（AI）研究人员一直以来在从事的开发和完善这类模型的相关工作。由 OpenAI 开发的最新模型，GPT-4 是使用前所未有的计算和数据规模来训练的。本文中报告了我们对当时还在 OpenAI 的积极开发中 GPT-4 早期版本的调查。我们认为，（这个早期版本的）GPT-4 是新一批 LLM 的一部分（例如，与 ChatGPT 和谷歌的 PaLM

^⑦ USMLE 考试中步骤 2 分为 CK（Clinical Knowledge，临床知识）和 CS（Clinical Skill，临床技能）两个部分。

一起），比以前的 AI 模型表现出更多的通用智能。讨论了模型的提升能力和意义。证明除了对语言的掌握，GPT-4 还能解决跨越数学、编码、视觉、医学、法律、心理学等领域的新颖而困难的任务，而不需要任何特别的提示。此外，在这些全部任务中，GPT-4 的表现都惊人地接近人类水平，且大大超过了 ChatGPT 等以往模型。鉴于 GPT-4 能力的广度和深度，我们可以合理地将其视为人工通用智能（AGI）系统的早期（但仍不完整）版本。在我们对 GPT-4 的探索中，特别强调发现其局限性，并讨论了在向更深入和更全面的 AGI 版本迈进时面临的挑战，包括可能需要追求一种超越单词预测的新范式。最后，我们对最近技术飞跃的社会影响和未来的研究方向进行了思考。



29. Training Language Models to Follow Instructions with Human Feedback

训练语言模型以遵循人类反馈的指令

让语言模型变得更大并不意味着能更好地遵循用户的意图。例如，大型语言模型可以产生不真实的、有害的或根本对用户没有帮助的输出。换句话说，这些模型没有与用户保持一致。本文展示了一个途径，通过人类反馈的微调，在广泛的任务中使语言模型与用户的意图保持一致。从一组标签员写的提示语和通过 OpenAI API 提交的提示语开始，我们收集了一组标签员对所需模型行为的演示数据集，使用监督学习对 GPT-3 进行微调。然后，我们收集模型输出的排名数据集，利用人类反馈的

强化学习来进一步微调这个监督模型。我们将产生的模型称为 InstructGPT。在人类对我们的提示分布评估中，尽管参数少了 100 倍，但 1.3B 参数的 InstructGPT 模型的输出比 175B 的 GPT-3 输出更受欢迎。此外，InstructGPT 模型显示了真实性的改善和有害输出生成的减少，同时在公共 NLP 数据集上的性能回归最小。尽管 InstructGPT 仍然会犯一些简单的错误，但我们的结果表明，利用人类反馈进行微调是使语言模型与人类意图相一致的一个有希望的方向。

30. Translating Radiology Reports into Plain Language using ChatGPT and GPT-4 with Prompt Learning: Promising Results, Limitations, and Potential

使用 ChatGPT 和 GPT-4 将放射学报告翻译成普通话的提示性学习：有希望的结果，局限性和潜在性

被称为 ChatGPT 的大型语言模型因其类似人类的表达和推理能力引起了广泛的关注。在这项研究中，我们调查了使用 ChatGPT 将放射学报告翻译成通俗易懂的语言，供病人和医疗机构使用，以便其接收教育和改善医疗服务。本研究在 2 月上旬收集了 62 个低剂量胸部 CT 肺癌筛查扫描和 76 个脑部 MRI 转移筛查扫描的放射学报告。根据放射科医生的评价，ChatGPT 可以成功将放射科报告翻译成通俗易懂的语言，在 5 分制中平均得分为 4.27 分，有 0.08 处信息缺失，0.07 处信息错误。就 ChatGPT 提供的建议而言，都是一般性的相关建议，如保持对医生的随访，密切关注任何症状，对于总共 138 个案例中约 37% ChatGPT 根据报告中的发现提供了具体的建议。ChatGPT 在回复中也呈现出一定的随机性，偶尔会出现过度简化或忽略的信息，这可以通过更详细的提示来缓解。此外，ChatGPT 的结果与新发布的大型模型 GPT-4 进行了比较，显示 GPT-4 可以显著提高翻译报告的质量。我们的结果表明，在临床教育中利用大型语言模型是可行的，还需要进一步开展研究以解决局限性，并最大限度地发挥其潜力。

医疗软件，医护助理 生命所系，质量保障



构建权威高效质量测评体系，
医疗软件规范发展必经之路



Featured Article

文献精读

在美国医疗系统中运行和实施预训练的大型人工智能语言学模型：生成式预训练转化器 3（GPT-3）作为一种服务模式的前景

Emre Sezgin^{1*}, 博士; Joseph Sirrianni^{1*}, 博士; Simon L Linwood², 工商管理硕士、医学博士

1 阿比盖尔-韦克斯纳研究中心, 全美儿童医院, 哥伦布市, 俄亥俄州, 美国

2 医学院, 加州大学河滨分校, 河滨市, 加利福尼亚州, 美国

* 以上作者贡献相等

摘要

生成式预训练转化器模型由于其增强的能力和性能, 最近大受欢迎。与许多现有的人工智能模型相比, 生成式预训练转化器模型可以在非常有限的训练数据下执行。生成式预训练转化器 3 (GPT-3) 是这个管道的最新版本之一, 展示了对提示性语言的类似人类的逻辑和智力反应。一些例子包括文章写作、回答复杂的问题、将代词与名词相匹配, 以及进行情感分析。然而关于其在医疗保健中的实施, 特别是在操作化和在临床实践和研究中的使用方面, 仍然存在问题。本文简要介绍了 GPT-3 及其功能, 并通过一个用例概述了其在临床实践中实施和操作的注意事项。实施方面的考虑包括: (1) 处理需求和信息系统基础设施; (2) 运行成本; (3) 模型偏见; (4) 评价指标。此外, 还概述了以下三个推动美国医疗系统采用 GPT-3 的主要操作因素: (1) 确保符合健康保险流通和责任法案; (2) 与医疗服务提供者建立信任; (3) 建立更广泛的 GPT-3 工具的使用权。这一观点可以为医疗保健从业人员、开发人员、临床医生和决策者提供信息, 以了解集成在医院系统和医疗保健中的强大人工智能工具的使用情况。

(JMIR Med Inform 2022;10(2): e32875) doi:10.2196/32875

关键词: 自然语言处理; 人工智能; 生成式预训练转化器; 临床信息学; 聊天机器人

引言

2020 年, OpenAI 公布了第三代语言生成模型, 也就是所谓的生成式预训练转化器 3 (GPT-3) 模型^[1]。GPT-3 是一系列通过使用转换器 (Transformer) 架构为理解和产生自然语言而设计的大型预训练模型中最新的一个, 转换器架构于 3 年前发

布^①, 与之前架构上建立的模型相比, 大大改善了自然语言理解任务的性能^[2]。无论如何, GPT-3 的发展是引人注目的, 因为它导致了模型规模的大幅增加, 其在 1 年内增加了 10 倍以上, 达到了 1750 亿个权重^[1-3]。GPT-3 模型规模的增加使其大大超过了先前的模型; 将其语言能力推进到接近人类的水平; 并且在某些情况下, 使其成为若干语言理解任务的最优选择^[1]。

^①此文于 2021 年初投稿, 故文中时间为 3 年前。

通常情况下，深度学习任务需要大量标记的训练数据。这一要求通常限制了深度学习可以有效应用的任务。然而，随着模型规模的扩大，GPT-3 与之前的模型相比，具有更强的能力，可以进行所谓的少量学习（Few-shot Learning）、单次学习（One-shot Learning）和零次学习（Zero-shot Learning）^[1,4]，以上这些学习方法涉及如何在明显数量较少的训练数据上训练一个模型。在这些方法中，给模型一个任务的描述，如果适用的话，还有少量的样本可用于学习，在数百到数千的样本上进行少量训练，在只有一个样本上进行单次训练，在只存在任务描述时进行零次训练。

GPT-3 被设计为一个语言生成模型，专注于对输入的文本产生相应的答复。尽管 GPT-3 可以用来处理更传统的机器学习任务，如回答“是”或“否”的问题、将代词与名词相匹配，以及进行情感分析^[1]，但 GPT-3 的文本生成能力作为各种问题的潜在解决方案引起了广泛关注，如创建增强型聊天机器

人^[5]、回答复杂问题、根据设计规范生成计算机代码^[6]，以及撰写新闻文章^[7]。因此，在使用 GPT-3 进行各种文本生成应用方面有很多研究和商业利益。文本框 1 在 3 个问答的案例中说明了 GPT-3 的实力^[8]。在这 3 个案例中，GPT-3 提供了类似人类的逻辑和智力反应。但在医疗保健领域中应用 GPT-3 仍是一个新兴的课题。

大量的研究文章概述了在各种医院和医疗保健环境中整合人工智能（AI）和机器学习模型的潜在机会^[9-11]。然而，作者也很快指出了部署此类解决方案所面临的许多开发和实施方面的挑战，如缺乏透明度^[11]、潜在的模型偏见^[9]和融资问题^[10]。GPT-3 的成功部署需要克服所有这些人工智能实施的标准障碍，以及 GPT-3 本身带来的额外挑战^[12]。在这个观点中，我们通过一个案例来介绍 GPT-3 在医疗环境中的应用，并讨论其在临床实践中的实施和操作注意事项。

文本框 1. 生成式预训练转化器 3（GPT-3）的问答案例

问与答

- 人类提出的问题：“美国的人类预期寿命为多少？”
- GPT-3 产生的答复：“美国的人类预期寿命是 78 岁。”

文本摘要

- 人为生成的提示：“木星是来自太阳的第五颗行星，也是太阳系中最大的行星。它是一个气体巨行星，质量为太阳的千分之一，但却是太阳系所有其他行星质量总和的 2.5 倍。木星是夜空中肉眼可见的最明亮的天体之一，早在有历史记载前，就已被古文明所熟知。以罗马神朱庇特命名。从地球上，木星的亮度足以让其反射光投下可见的阴影，而且是夜空中继月亮和金星之后平均第三亮的自然天体。”
- GPT-3 产生的答复（以一个二年级学生的语气）：“木星是一个由气体组成的巨大行星。它非常大，非常明亮。它是夜空中继月球和金星之后第三个最亮的物体。木星有许多卫星。其中一颗卫星比水星还大。”

一个聊天机器人（讽刺的）

- 人类：“为什么天空是蓝色的？”
- GPT-3 聊天机器人：“我不确定，我会问我的朋友谷歌”

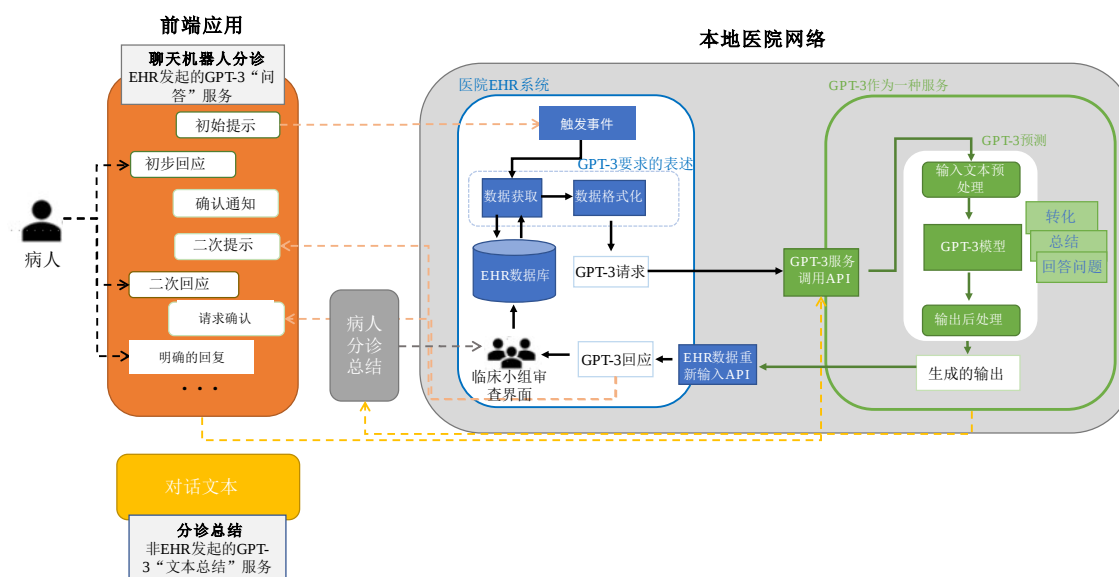


图 1. GPT-3 用例（聊天机器人分诊和病人记录总结）。API：应用程序接口；EHR：电子健康记录；GPT-3：生成式预训练模型转化器 3。

一个使用案例：医疗保健中的 GPT-3

GPT-3 生成有意义文本的能力使其成为人机对话的理想工具，特别是通过聊天机器人进行的对话。在我们的案例中，说明了 GPT-3 在医院网络中的应用。在假设的案例中，医院正在为入院的病人提供一个聊天机器人分诊机制，以减少诊所的开销，并在 COVID-19 大流行期间提高护理的安全性和质量。聊天机器人必须连接到医院网络，与要审查的分诊文本摘要服务相结合，并存储在电子健康记录（图 1）。抛开这个工作流程中的前端细节（图 1），这个案例概述了 GPT-3 作为卫生系统的一项服务的典型实施情况。

在这个例子中，分诊可以由病人或医院发起，以进行健康筛查。前端应用通过一个聊天机器人机制进行操作，该机制可以是患者门户应用、语音助手、电话或 SMS 短信。一旦建立连接，医院系统就会通过收集病人的健康信息，并将其格式化为可解释的 GPT-3 模型，从而制定 GPT-3 请求。在安全的医院网络内，GPT-3 位于 EHR 之外，作为“GPT-3 即服务”平台提供。应用编程接口实现了互操作性，并作为请求和响应的数据传输的守门员。一旦收到请求，“GPT-3 即服务”平台将对数据和请求进行预处理，分配需要完成的任务，以可解释的格式生成输出，并将输出发送给用户。分配的任务类型取决于请求，在我们的案例中，请求是回答问题、文本生成或在文化上对应的语言翻译，以及文本总结。响应被送回 EHR 系统，然后送到前端应用程序。在分诊结束时，类似于诊后总结，对话文

本被总结出来。为了减少阅读整个对话的额外临床负担，GPT-3 对文本进行总结（类似于数字划线器）并将其存储在病人的健康记录中。为了避免或解决潜在的偏见^[12]、纠正错误与增加对病人数据使用和模型的控制，可以通过在前端使用报告反馈机制来实现人在环路模式^[13]，或者在后端让临床团队监督医院 EHR 系统中的 GPT-3 整合过程。此外，文本中的错误更正和调整可以用来微调 GPT-3 模型，以提高其准确性和有效性。

为了能够在现实环境中执行这一用例，医疗保健从业人员和决策者应考虑并解决以下操作和实施方面的挑战。

实施方面的考虑

处理需求和信息系统基础设施

与更传统的人工智能模型不同，GPT-3 在内存需求方面要大得多，而且计算量更大。用于模型训练和执行的专用硬件——无论是图形处理单元还是张量处理单元——都是可扩展实现所必需的。对于任何医院系统来说，可能都需要对基础设施进行额外的投资，以补偿处理需求。

鉴于其规模、依赖性和硬件要求，GPT-3 解决方案可能需要作为一项服务来运行。对于这种服务，医院系统需要向 GPT-3 解决方案服务提交一个服务请求，该服务将处理请求并将结果返回给医院系统。图 1 中的医院本地网络显示了一个实施的样本工作流程图。该设置需要在配置、联网和监控方面有

细致和大量的准备工作，以确保服务的可访问性并提供有意义的价值。

运行成本

鉴于医院网络和 EHR 系统的现状，GPT-3 解决方案的整合需要复杂的系统和高技术知识才能有效部署，且运行成本高昂。为减轻 GPT-3 部署的负担，一个潜在的解决方案是与医院系统内的云计算平台整合。许多云计算供应商提供运行此类模型所需的专门硬件，并能轻松处理现成的网络和动态负载平衡。这将减轻 GPT-3 部署的主要组件的负担；然而，外包的云计算平台有可能增加运行成本。

模型偏见

偏倚的几个来源可以在 GPT-3 驱动的解决方案中的不同层面上表现出来。在模型层面，GPT-3 是在一个大型数据集上训练的，该数据集有许多与种族和性别歧视刻板印象有关的问题特征，因此，该模型学习了某些针对边缘化身份的偏倚^[14,15]。这些偏倚存在于 GPT-3 中，在临床上可能是有害的。Korngiebel 和 Mooney^[12]强调了在卫生保健服务中使用 GPT-3 的风险，指出了 GPT-3 鹦鹉学舌^[16]和肯定自杀意念的具体例子^[17]。

除了 GPT-3 初始训练的固有偏见外，对医疗数据进行微调也可能引入历史医疗数据中存在的无意识偏倚。实际的偏倚，如对边缘化亚人群的测试不足，会影响基础临床数据，并在预测模型的训练中引入偏倚^[9]。此外，医护人员的内隐偏倚会影响诊断和治疗，并反映在临床记录中^[18]，如果用于微调 GPT-3，将有可能影响所开发的模型。

鉴于这些偏倚，在没有主动偏倚测试的情况下，部署 GPT-3 或任何其他可观的语言模型都是不明智的^[15]。应制定明确的流程来监测、报告和应对 GPT-3 预测可能产生的潜在偏倚。这些机制将确保 GPT-3 能被有效地使用，而不对病人造成伤害。在我们的用例中（图 1），我们还增加了一个人在环路中的机制，它可以授权控制、评估和训练协议，并产生可解释和可管理的结果。

评估指标

除了物理实施外，部署 GPT-3 还有方法上的考虑。正如 Watson 等人^[10]在学术医疗中心的模型部署调查中指出，临床效用是机构的主要关注点。理解接受和解释模型结果的最佳方式是成功部署的必要条件，理想情况下，应通过使用评估方法和框架来跟踪和评估模型性能。

对文本生成任务的评估，也就是 GPT-3 所要解决的任务，

这是众所周知的困难。标准指标，如预测灵敏度和阳性预测值，并不能明确地反映文本生成的正确性，因为想法可以在文本中以多种方式表达。更加专业的文本生成指标，如 BLEU（双语评估研究）^[19]和 METEOR（明确排序的翻译评估指标）^[20]，试图解释文本的变化，但仍然只在词的层面上检查文本，而没有捕捉到根本含义。那些试图在文本评价中纳入文本意义的方法，则依靠其他黑盒深度学习模型来产生数值^[21]。依靠黑盒评价方法来评价黑盒模型并不能提高可解释性。这样的方法只会导致整体信任度降低，从而降低模型被部署的可能性。

因此，需要一个针对卫生保健的文本生成任务的评估方法和框架。在广泛采用 GPT-3 技术之前，需要开发更强大的方法来评估医疗保健领域的文本生成任务。数据科学家、信息学家、开发人员、临床医生和医疗保健从业人员必须合作制定评估措施，以确保 GPT-3 的成功实施。

操作方面的考虑：合规、信任和访问

除了实施之外，还有 3 个主要操作因素推动 GPT-3 在医疗领域的应用，具体如下：

- 1) GPT-3 需要在符合《健康保险流通和责任法案》（HIPAA）的情况下工作；
- 2) 技术供应商需要赢得医疗机构的信任；
- 3) 技术供应商应改善对该工具的使用（图 2）。

与 GPT-3 类似，亚马逊 Alexa（Amazon.com Inc）语音助手在 2014 年发布时，人们对其在医疗服务中的应用有着巨大的热情。然而，在当时，Alexa 还不能合法的存储或传输私人健康信息。亚马逊花了 5 年时间才符合 HIPAA 的要求，并能与医疗服务提供者签署商业伙伴协议^[22]。发布的 Alexa 技术数量有限，仍有一长串其他 Alexa 技术在等待符合 HIPAA 标准。这个例子表明，在将新技术纳入医疗保健方面，立法变更和法规更新的进展缓慢，这表明应该尽早为 GPT-3 做出努力。如果不符合 HIPAA 标准，GPT-3 在医疗领域的采用可能是一个错误的开始^[23]。然而，尽管 HIPAA 合规性可能不会立即实现，但可能会逐步推进。GPT-3 是一个黑盒模型，这使 HIPAA 合规过程变得复杂，因为与其他类型的程序化解决方案不同，其更难破译数据是如何被模型内部处理的。然而，假设 GPT-3 在未来可以部署，操作将从实施 GPT-3 的有限能力开始 [即存储和传输数据，在特定硬件的防火墙后面运行（安全规则），以及分析特定数据集或患者群（隐私规则）]。同时，需要进一步的实践来优化适应 GPT-3 的支付模式，并寻求机会来满足

美国食品和药物管理局对软件作为医疗设备的要求^[24]，即在临床应用中使用人工智能。

除法律要求外，必须在患者、医疗机构和技术公司之间建立信任，以采用 GPT-3^[25]。技术公司通常声称他们有权使用客户的数据来进一步改善他们的服务或实现额外的商业价值。此外，临床医生对人工智能持怀疑态度的文化会给模型的可解释性带来沉重负担，导致临床护理的信任度低于其他行业^[10]。与商业实施不同，GPT-3 需要明确讨论它将使用患者的数据做什么和不做什么。医疗机构的数据管理委员会在与 GPT-3 签署服

务协议时，需要了解并适应这一点。鉴于 GPT-3 的黑盒性质，在解读通过人在环路模型产生的评估报告和结果时，有必要采取一种操作性的战略方法。

接入也需要得到保证。训练像 GPT-3 这样的大型语言模型可能需要花费数千万美元。因此，GPT-3 正在对接入的商业模式进行创新。目前，GPT-3 由 OpenAI 私人控制，医疗机构可以远程运行该程序，并按代币支付使用费（1000 个代币大约相当于 750 个单词）^[26]。2020 年 9 月，微软购买了 GPT-3 的独家授权，并计划将其整合到其现有产品中。

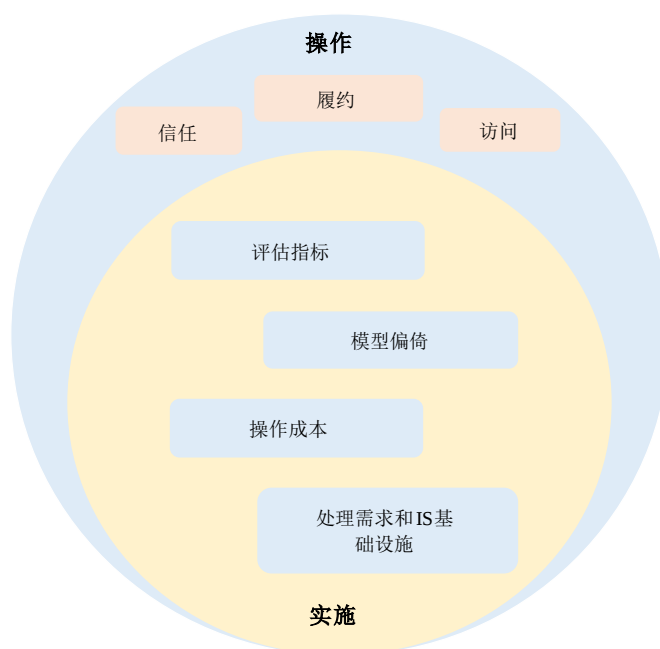


图 2. 生成式预训练转化器 3 的操作和实施考虑模型。IS：信息系统

同样，一些公司已经在将 GPT-3 模型预测整合到他们的产品中。然而，这种商业模式也限制了开放性的研究和开发，并最终会限制改进，如翻译机制的提升和对话代理开发中的全包裹、公平驱动的方法。在这些早期阶段，开源的替代方案，如 GPT-J^[27]，可能会帮助医疗保健开发商和机构评估运行的可行性。在未来的迭代中，一旦在医疗环境中使用 GPT-3 的价值得到保证，可及性的责任就可以委托给医疗和政府机构。这些机构可以通过安全的云平台分发“GPT-3 即服务”平台，并建立一个联合学习机制来运行分散的培训服务，同时为 GPT-3 模型提供帮助^[28]。这也将减少各个卫生系统在建设、培训和部署自己的 GPT-3 平台时的负担，并降低成本。这些优势对低资源环境下的医院尤其有利。

结论

在本文观点中，我们简要介绍了 GPT-3 及其功能，并通过一个用例概述了其在临床实践中实施和操作的注意事项。在 Korngiebel 和 Mooney^[12]对不现实的、现实的、可行的和现实但具有挑战性的用例进行评述的基础上，我们提供了在临床实践中实施和操作 GPT-3 的注意事项。我们相信，我们的工作可以为医疗保健从业人员、开发人员、临床医生和决策者提供信息，使他们了解整合到医院系统和医疗保健中的强大人工智能工具的使用。

致谢

这项研究得到了以患者为中心的结果研究所（PCORI）的部分支持（奖项编号：ME-2017C1-6413），名称为“通过使用知识库和单词嵌入的查询优化来破解 EMR 中的临床文本”。本报告中所有陈述，包括其发现和结论，仅代表作者本人，而非 PCORI、其理事会或其他方法委员会的观点。

作者贡献

ES, JS 和 SLL 构思了提出的想法。ES 和 JS 起草了手稿。SL 监督并严格审查了手稿。所有作者都批准了手稿的最终版本。

利益冲突

无。

参考文献

- Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on July 22, 2020 [FREE Full text]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017; Long Beach, California, USA URL: <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
- Hutson M. Robo-writers: the rise and risks of language-generating AI. Nature 2021 Mar;591(7848):22-25. [doi: 10.1038/d41586-021-00530-0] [Medline: 33658699]
- Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What makes good in-context examples for GPT-3? arXiv. Preprint posted online on January 17, 2021 [FREE Full text]
- Emerson. GPT-3 Demo. URL: <https://gpt3demo.com/apps/quickchat-emerson> [accessed 2021-12-14]
- Langston J. From conversation to code: Microsoft introduces its first product features powered by GPT-3. The AI Blog. 2021 May 25. URL: <https://blogs.microsoft.com/ai/from-conversation-to-code-microsoft-introduces-its-first-product-features-powered-by-gpt-3/> [accessed 2021-12-14]
- A robot wrote this entire article. Are you scared yet, human? GPT-3. The Guardian. 2020 Sep 08. URL: <https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3> [accessed 2022-02-01]
- Examples - OpenAI API. OpenAI. URL: <https://beta.openai.com/examples/> [accessed 2021-12-14]
- Wiens J, Price WN2, Sjoding MW. Diagnosing bias in data-driven algorithms for healthcare. Nat Med 2020 Jan;26(1):25-26. [doi: 10.1038/s41591-019-0726-6] [Medline: 31932798]
- Watson J, Huttyra CA, Clancy SM, Chandiramani A, Bedoya A, Ilangoan K, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? JAMIA Open 2020 Jul 10;3(2):167-172 [FREE Full text] [doi: 10.1093/jamiaopen/ooz046] [Medline: 32734155]
- Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif Intell Med 2020 Apr;104:101822 [FREE Full text] [doi: 10.1016/j.artmed.2020.101822] [Medline: 32499001]
- Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digit Med 2021 Jun 03;4(1):93 [FREE Full text] [doi: 10.1038/s41746-021-00464-x] [Medline: 34083689]
- Miller K. When algorithmic fairness fixes fail: The case for keeping humans in the loop. Stanford University. 2020 Nov 02. URL: <https://hai.stanford.edu/news/when-algorithmic-fairness-fixes-fail-case-keeping-humans-loop> [accessed 2021-12-14]
- Lucy L, Bamman D. Gender and representation bias in GPT-3 generated stories. In: Proceedings of the Third Workshop on Narrative Understanding. 2021 Jun Presented at: Third Workshop on Narrative Understanding; June 11, 2021; Virtual p. 48-55 URL: <https://aclanthology.org/2021.nuse-1.5.pdf> [doi: 10.18653/v1/2021.nuse-1.5]
- Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? 2021 Mar Presented at: FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency; March 3-10, 2021; Canada. [doi: 10.1145/3442188.3445922]
- McGuffie K, Newhouse A. The radicalization risks of GPT-3 and advanced neural language models. arXiv. Preprint posted online on September 15, 2020 [FREE Full text]
- Daws R. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. AI News. 2020 Oct 28. URL: <https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/> [accessed 2022-02-01]
- FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. BMC Med Ethics 2017 Mar 01;18(1):19 [FREE Full text] [doi: 10.1186/s12910-017-0179-8] [Medline: 28249596]
- Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. 2002 Jul Presented at: 40th Annual Meeting of the Association for Computational Linguistics; July 7-12, 2002; Philadelphia, Pennsylvania, USA URL: <https://www.aclweb.org/anthology/P02-1040.pdf> [doi: 10.3115/1073083.1073135]

20. Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. 2005 Jun Presented at: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; June 29, 2005; Ann Arbor, Michigan URL:<https://www.aclweb.org/anthology/W05-0909.pdf> [doi: 10.3115/1626355.1626389]
21. Celikyilmaz A, Clark E, Gao J. Evaluation of text generation: A survey. arXiv. Preprint posted online on May 18, 2021 [FREE Full text]
22. Jiang R. Introducing new Alexa healthcare skills. Amazon Alexa. 2019 Apr 04. URL: <https://developer.amazon.com/blogs/alexa/post/ff33dbc7-6cf5-4db8-b203-99144a251a21/introducing-new-alexa-healthcare-skills> [accessed 2022-02-01]
23. McGraw D, Mandl KD. Privacy protections to encourage use of health-relevant digital data in a learning health system. NPJ Digit Med 2021 Jan 04;4(1):2 [FREE Full text] [doi: 10.1038/s41746-020-00362-8] [Medline: 33398052]
24. Artificial intelligence and machine learning in Software as a Medical Device. U.S. Food & Drug Administration. URL: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> [accessed 2021-12-14]
25. Patient trust must come at the top of researchers' priority list. Nat Med 2020 Mar;26(3):301. [doi: 10.1038/s41591-020-0813-8] [Medline: 32161404]
26. Brockman G, Murati M, Welinder P, OpenAI. OpenAI API. OpenAI. 2020 Jun 11. URL: <https://openai.com/blog/openai-api/> [accessed 2021-12-14]
27. Romero A. Can't access GPT-3? Here's GPT-J — its open-source cousin. Towards Data Science. 2021 Jun 24. URL: <https://towardsdatascience.com/cant-access-gpt-3-here-s-gpt-j-its-open-source-cousin-8af86a638b11> [accessed 2021-12-14]
28. Sadilek A, Liu L, Nguyen D, Kamruzzaman M, Serghiou S, Rader B, et al. Privacy-first health research with federated learning. NPJ Digit Med 2021 Sep 07;4(1):132 [FREE Full text] [doi: 10.1038/s41746-021-00489-2] [Medline: 34493770].

缩写词

AI: 人工智能

BLEU: 双语评估研究 (Bilingual Evaluation Understudy)

EHR: 电子健康记录

GPT-3: 生成式预训练转化器 3

HIPAA: 健康保险可携性和责任法案

METEOR: 明确排序的翻译评估指标

PCORI: 以病人为中心的结果研究机构

编辑: C Lovis; 提交时间: 12.08.21; 同行评议: A Trojan, T Gaze, G Carot-Sans, J Ayre, W Zhang; 2021 年 10 月 24 日给作者的意见; 2021 年 12 月 14 日收到修订版; 2022 年 1 月 9 日接收; 2022 年 2 月 10 日发表。

请按照以下格式引用:

Sezgin E, Sirrianni J, Linwood SL Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model JMIR Med Inform 2022;10(2): e32875

URL: <https://medinform.jmir.org/2022/2/e32875>

doi: 10.2196/32875

© Emre Sezgin, Joseph Sirrianni, Simon L Linwood, 原文发表于 JMIR 医学信息学 (<https://medinform.jmir.org>) , 10.02.2022. 这是一篇开放性文章, 根据知识共享署名许可协议(<https://creativecommons.org/licenses/by/4.0/>)的条款发布, 允

许在任何媒体上不受限制地使用、传播和复制, 但必须正确使用首次发表于 JMIR 医学信息学的原作。必须包括完整的书目信息、原始出版物的链接 (<https://medinform.jmir.org/>) , 以及本版权和许可证信息。

Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model

Emre Sezgin^{1*}, PhD; Joseph Sirrianni^{1*}, PhD; Simon L Linwood², MBA, MD

¹The Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, OH, United States

²School of Medicine, University of California Riverside, Riverside, CA, United States

*these authors contributed equally

Corresponding Author:

Emre Sezgin, PhD
The Abigail Wexner Research Institute
Nationwide Children's Hospital
700 Children's Drive
Columbus, OH, 43205
United States
Phone: 1 6143556814
Email: esezgin1@gmail.com

Abstract

Generative pretrained transformer models have been popular recently due to their enhanced capabilities and performance. In contrast to many existing artificial intelligence models, generative pretrained transformer models can perform with very limited training data. Generative pretrained transformer 3 (GPT-3) is one of the latest releases in this pipeline, demonstrating human-like logical and intellectual responses to prompts. Some examples include writing essays, answering complex questions, matching pronouns to their nouns, and conducting sentiment analyses. However, questions remain with regard to its implementation in health care, specifically in terms of operationalization and its use in clinical practice and research. In this viewpoint paper, we briefly introduce GPT-3 and its capabilities and outline considerations for its implementation and operationalization in clinical practice through a use case. The implementation considerations include (1) processing needs and information systems infrastructure, (2) operating costs, (3) model biases, and (4) evaluation metrics. In addition, we outline the following three major operational factors that drive the adoption of GPT-3 in the US health care system: (1) ensuring Health Insurance Portability and Accountability Act compliance, (2) building trust with health care providers, and (3) establishing broader access to the GPT-3 tools. This viewpoint can inform health care practitioners, developers, clinicians, and decision makers toward understanding the use of the powerful artificial intelligence tools integrated into hospital systems and health care.

(*JMIR Med Inform* 2022;10(2):e32875) doi: [10.2196/32875](https://doi.org/10.2196/32875)

KEYWORDS

natural language processing; artificial intelligence; generative pretrained transformer; clinical informatics; chatbot

Introduction

In 2020, OpenAI unveiled their third-generation language generation model, which is known as the generative pretrained transformer 3 (GPT-3) model [1]. This model was the latest in a line of large pretrained models designed for understanding and producing natural language by using the transformer architecture, which was published only 3 years prior and significantly improved natural language understanding task

performance over that of models built on prior architectures [2]. However, GPT-3's development was remarkable because it resulted in a substantial increase in the model's size; it increased by more than 10-fold in 1 year, reaching 175 billion weights [1-3]. GPT-3's increased model size makes it substantially more powerful than prior models; propels its language capabilities to near-human-like levels; and, in some cases, makes it the superior option for several language understanding tasks [1].

Ordinarily, deep learning tasks require large amounts of labeled training data. This requirement usually limits the tasks to which deep learning can be effectively applied. However, with its increased model size, GPT-3 has an enhanced capability for so-called *few-shot*, *one-shot*, and *zero-shot learning* when compared to prior models [1,4]. These learning methods involve training a model on significantly smaller amounts of training data. In these methods, the models are given a description of the task and, if applicable, a handful of examples to learn from, with few-shot training on only hundreds to thousands of instances, one-shot training on only 1 example, and zero-shot training on only the task description.

GPT-3 was designed as a language generation model, focusing on producing appropriate text responses to an input. Although it can be adapted to address more traditional machine learning tasks, such as answering yes-no questions, matching pronouns to their nouns, and conducting sentiment analyses [1], GPT-3's text generation capabilities have attracted much attention as a potential solution for a variety of problems, such as creating enhanced chatbots [5], answering complex questions, generating

computer code from a design specification [6], and writing news articles [7]. As such, there is much research on and commercial interest in using GPT-3 for a variety of text generation applications. [Textbox 1](#) illustrates the power of GPT-3 in 3 question and answer examples [8]. In 3 cases, GPT-3 provided human-like logical and intellectual responses. However, applying GPT-3 in health care is still an emerging topic.

A plethora of research articles have outlined the potential opportunities for integrating artificial intelligence (AI) and machine learning models in a variety of hospital and health care settings [9-11]. However, the authors also quickly pointed out the many development and implementation challenges to deploying such solutions, such as the lack of transparency [11], potential model biases [9], and financing [10]. A successful deployment of GPT-3 would need to overcome all of these standard barriers to AI implementation alongside the additional challenges that GPT-3 itself introduces [12]. In this viewpoint, we introduce the use of GPT-3 in the health care setting with a use case and discuss implementation and operational considerations for its use in clinical practice.

Textbox 1. Generative pretrained transformer 3 (GPT-3) question and answer examples.

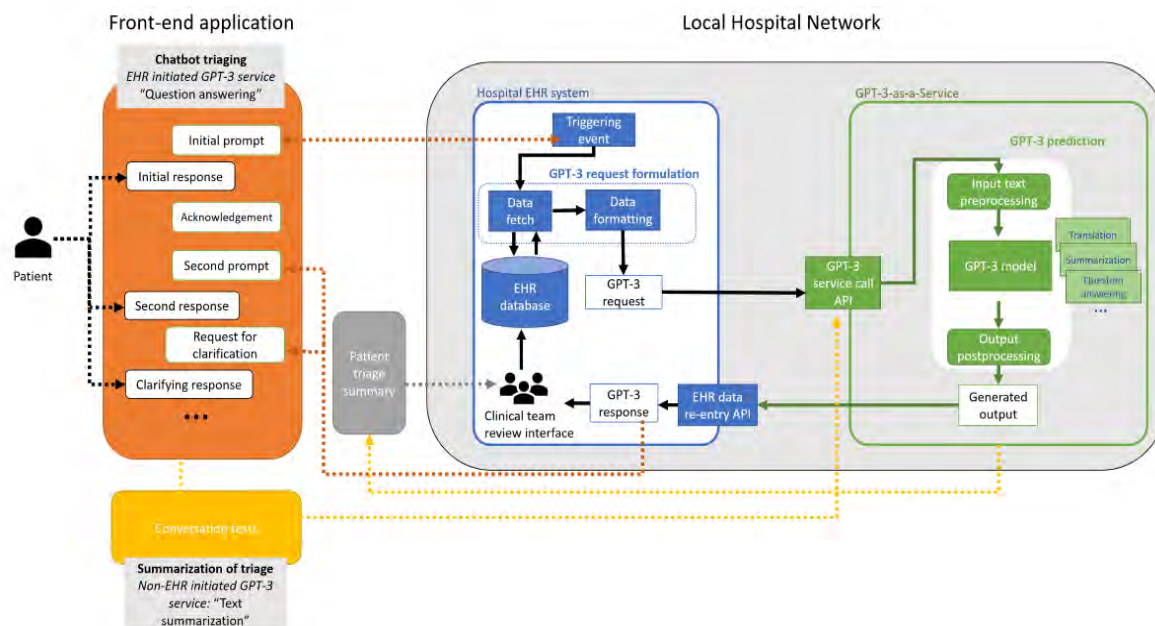
Question and answer - Human-generated question: "What is human life expectancy in the United States?" - GPT-3-generated response: "Human life expectancy in the United States is 78 years." Text summary - Human-generated prompt: "Jupiter is the fifth planet from the Sun and the largest in the Solar System. It is a gas giant with a mass one-thousandth that of the Sun, but two-and-a-half times that of all the other planets in the Solar System combined. Jupiter is one of the brightest objects visible to the naked eye in the night sky, and has been known to ancient civilizations since before recorded history. It is named after the Roman god Jupiter. When viewed from Earth, Jupiter can be bright enough for its reflected light to cast visible shadows, and is on average the third-brightest natural object in the night sky after the Moon and Venus." - GPT-3-generated response (a second grader's language): "Jupiter is a huge planet made of gas. It is the fifth planet from the sun. It is very big and bright. It is the third brightest object in the night sky after the moon and Venus. Jupiter has many moons. One of the moons is bigger than the planet Mercury." A chatbot (sarcastic) - Human: "Why is the sky blue?" - GPT-3 chatbot: "I'm not sure. I'll ask my friend Google."
--

A Use Case: GPT-3 in Health Care

The power of meaningful text generation by GPT-3 makes it an ideal tool for human-machine conversations, especially those conducted via chatbots. In our case, we illustrate the use of GPT-3 within a hospital network. In this hypothetical example, the hospital is providing a chatbot triaging mechanism for

incoming patients to reduce the overhead at clinics and increase the safety and quality of care during the COVID-19 pandemic. The chatbot has to be connected to the hospital network, combined with a triage text summary service that is to be reviewed, and stored in the electronic health record (EHR; [Figure 1](#)). Putting aside the front-end details in this workflow ([Figure 1](#)), this use case outlines a typical implementation of GPT-3 as a service within a health system.

Figure 1. GPT-3 use case (chatbot triaging and patient note summarization). API: application programming interface; EHR: electronic health record; GPT-3: generative pretrained transformer 3.



In this example, triage could be initiated by a patient or a hospital to conduct a health screening. The front-end application is operationalized through a chatbot mechanism over a front-end application, which could be a patient portal app, voice assistant, phone call, or SMS text messaging. Once a connection is established, the hospital system formulates GPT-3 requests by gathering patient health information and formatting this information to be interpretable with the GPT-3 model. Within the secure hospital network, GPT-3 is located outside of the EHR and provided as the "GPT-3-as-a-Service" platform. The application programming interface enables interoperability and acts as a gatekeeper for the data transfer of requests and responses. Once a request is received, the "GPT-3-as-a-Service" platform preprocesses the data and requests, allocates the tasks to be completed, produces outputs in an interpretable format, and sends the outputs to users. The type of tasks allocated depends on the requests, which, in our case, are question answering, text generation or culturally appropriate language translation, and text summarization. The response is sent back to the EHR system and then to the front-end application. At the end of triage, similar to the after-visit summary, the conversation text is summarized. To reduce the additional clinical burden of reading the whole conversation, GPT-3 summarizes the text (similar to a digital scribe) and stores it in the patient's health records. To avoid or address potential biases [12], correct errors, and increase the control over patient data use and the model, the human-in-the-loop model [13] can be implemented by using a report back mechanism at the front end, or the clinical team can be given oversight of GPT-3 integrated process in the hospital EHR system at the back end. Furthermore, the error corrections and adjustments in the text can be used to fine-tune the GPT-3 model to increase its accuracy and effectiveness.

To be able to execute this use case in a real-world setting, health care practitioners and decision makers should consider and address the following operational and implementation challenges.

Implementation Considerations

Processing Needs and Information Systems Infrastructure

Unlike more traditional AI models, GPT-3 is considerably larger in terms of memory requirements and is more computationally intensive. Specialized hardware for model training and execution—either graphics processing units or tensor processing units—is required for a scalable implementation. For any hospital system, additional investments for infrastructure to compensate for processing needs could be required.

Given its size, dependencies, and hardware requirements, a GPT-3 solution would likely need to be run as a service. For this service, hospital systems would need to submit a service request to the GPT-3 solution service, which would process the request and return its results back to the hospital system. The hospital local network in Figure 1 shows a sample workflow diagram for such an implementation. Such a setup would require diligent and significant provisioning, networking, and monitoring to ensure that the services are accessible and provide meaningful value.

Operating Cost

Given the current state of hospital networks and EHR systems, the integration of GPT-3 solutions would require complex systems and high technical knowledge for effective deployment and be costly to operationalize. One possible solution to ease the burden of GPT-3 deployments is integration with cloud computing platforms within hospital systems. Many cloud computing providers offer the specialized hardware needed to run such models and can easily handle off-the-shelf networking and dynamic load balancing. This would ease the burden of the major components of GPT-3 deployment; however, outsourcing cloud computing platforms can potentially increase the operating cost.

Model Bias

Several sources of bias can manifest themselves in a GPT-3-powered solution at different levels. At a model level, GPT-3 is trained on a large data set that has many problematic characteristics related to racial and sexist stereotypes, and as a result, the model learns certain biases against marginalized identities [14,15]. These biases, which are present in GPT-3, can be harmful in clinical settings. Korngiebel and Mooney [12] highlight the risks of using GPT-3 in health care delivery, noting specific examples where GPT-3 parrots extremist language from the internet [16] and affirms suicidal ideation [17].

Aside from the inherent bias of GPT-3's initial training, fine-tuning on medical data could also introduce the unintentional biases present in historic medical data. Practical biases, such as the undertesting of marginalized subpopulations, can influence underlying clinical data and introduce bias during the training of predictive models [9]. Additionally, the implicit biases of health care professionals can influence diagnoses and treatments and are reflected in clinical notes [18], which, if used to fine-tune GPT-3, would potentially affect the developed model.

Given these biases, it would be unwise to deploy GPT-3 or any other sizable language model without active bias testing [15]. Explicit procedures should be put in place to monitor, report, and react to potential biases produced by GPT-3 predictions. These mechanisms would ensure that GPT-3 can be used effectively without introducing harm to the patient. In our use case (Figure 1), we also added a human-in-the-loop mechanism, which can mandate the control, assessment, and training protocols and yield interpretable and manageable results.

Evaluation Metrics

Aside from physical implementation, there are methodological considerations for deploying GPT-3. As Watson et al [10] notes in their investigation of model deployment in academic medical centers, clinical utility is a major concern for institutions. Understanding the best way to receive and interpret model results is imperative for a successful deployment, and ideally, model performance should be tracked and assessed by using evaluation methodologies and frameworks.

The evaluation of text generation tasks, that is, those that GPT-3 is designed to address, is notoriously difficult. Standard metrics, such as prediction sensitivity and positive predictive value, do not cleanly reflect correctness in text generation, as ideas can be expressed in many ways in text. More specialized text generation metrics, such as BLEU (Bilingual Evaluation Understudy) [19] and METEOR (Metric for Evaluation of Translation with Explicit Ordering) [20], try to account for text variation but still only examine text at a word level without capturing the fundamental meaning. Methods that do try to incorporate the meaning of text in text evaluation rely on other black-box deep learning models to produce a value [21]. Relying on a black-box evaluation method to evaluate a black-box model does not increase interpretability. Such a method would only result in lower trust overall and thus decrease the likelihood of the model being deployed.

Health care-specific evaluation methods and frameworks for text generation tasks are therefore needed. The development of more robust methodologies for evaluating text generation tasks in the health care domain is required before the significant adoption of GPT-3 technology can be achieved. It is imperative that data scientists, informaticists, developers, clinicians, and health care practitioners collaborate in the development of evaluation measures to ensure a successful implementation of GPT-3.

Operational Considerations: Compliance, Trust, and Access

In addition to implementation, there are 3 major operational factors driving the adoption of GPT-3 in health care, as follows: (1) GPT-3 needs to work in compliance with the Health Insurance Portability and Accountability Act (HIPAA), (2) technology providers need to earn trust from health care providers, and (3) technology providers should improve access to the tool (Figure 2).

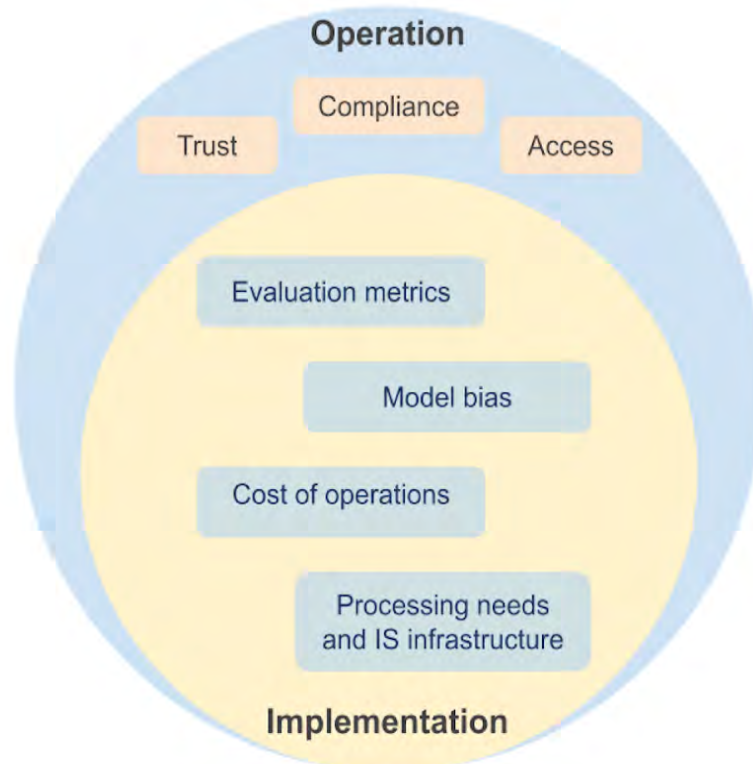
Similar to GPT-3, there was huge enthusiasm to use the Amazon Alexa (Amazon.com Inc) voice assistant in health care delivery when it was released in 2014. However, at the time, Alexa was not yet legally able to store or transmit private health information. It took Amazon 5 years to become HIPAA compliant and to be able to sign business associate agreements with health care providers [22]. A limited number of Alexa skills was released, and there is still a long list of other Alexa skills waiting to become HIPAA compliant. This example shows the slow progress of legislation changes and regulation updates for including new technologies in health care, suggesting that efforts should be put forward as early as possible for GPT-3. Without HIPAA compliance, the adoption of GPT-3 in health care can be a false start [23]. However, although HIPAA compliance may not be immediate, it may be gradually progressing. GPT-3 is a black-box model, which complicates the HIPAA compliance process because unlike with other types of programmatic solutions, it is harder to decipher how data are processed internally by the model itself. However, assuming that GPT-3 will be deployable in the future, operations will start with implementing the limited capabilities of GPT-3 (ie, storing and transmitting data, running behind the firewalls of specific hardware [security rules], and analyzing a specific data set or patient cohort [privacy rules]). In parallel, further practices are needed to optimize the payment models for accommodating GPT-3 and seek opportunities for satisfying the US Food and Drug Administration's requirements for software as a medical device [24] with regard to using AI in clinical applications.

In addition to legal requirements, trust must be established among patients, health care providers, and technology companies to adopt GPT-3 [25]. It is common for technology companies to claim the right that they can use their customers' data to further improve their services or achieve additional commercial value. Additionally, the culture of skepticism toward AI among clinicians can place a heavy burden on model interpretability and result in lower trust in clinical care than in other industries [10]. Unlike commercial implementations, GPT-3 needs to be explicitly discussed in terms of what it will and will not do with

a patient's data. Health care providers' data governance committees need to be aware and comfortable when they sign the service agreement with GPT-3. Given the black-box nature

of GPT-3, an operational strategic approach will be necessary for interpreting the evaluation reports and outcomes that are generated through the human-in-the-loop model.

Figure 2. A model of operational and implementation considerations for generative pretrained transformer 3. IS: information systems.



Access also needs to be ensured. Training large language models like GPT-3 can cost tens of millions of dollars. As such, GPT-3 is innovating the business model of access. Currently, GPT-3 is privately controlled by OpenAI, and health care providers can remotely run the program and pay for usage per token (1000 tokens are approximately equivalent to 750 words) [26]. In September 2020, Microsoft bought an exclusive license to GPT-3, with plans to integrate it into its existing products. Similarly, a number of companies are already integrating GPT-3 model predictions into their products. However, this business model also limits open-access research and development and will eventually limit improvements, such as advancements in translation mechanisms and all-inclusive, equity-driven approaches in conversational agent development. In these early stages, open-source alternatives, such as GPT-J [27], may help health care developers and institutions assess operational viability. In future iterations, once the value of using GPT-3 in the health care setting is assured, the responsibility of accessibility could be delegated to health care and government agencies. Such agencies may distribute the “GPT-3-as-a-Service” platform through secure cloud platforms

and establish a federated learning mechanism to run decentralized training services while collaboratively contributing to the GPT-3 model [28]. This would also reduce the burden on individual health systems when it comes to building, training, and deploying their own GPT-3 platforms and reduce costs. These advantages are especially beneficial for hospitals in low-resource settings.

Conclusion

In this viewpoint, we briefly introduce GPT-3 and its capabilities and outline considerations for its implementation and operationalization in clinical practice through a use case. Building on top of Korngiebel and Mooney’s [12] remarks toward unrealistic, realistic, feasible, and realistic but challenging use cases, we provide consideration points for implementing and operationalizing GPT-3 in clinical practice. We believe that our work can inform health care practitioners, developers, clinicians, and decision makers toward understanding the use of the powerful AI tools integrated into hospital systems and health care.

Acknowledgments

This study is partially supported through a Patient-Centered Outcomes Research Institute (PCORI) award (award number: ME-2017C1-6413) under the name “Unlocking Clinical Text in EMR by Query Refinement Using Both Knowledge Bases and Word Embedding.” All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the PCORI, its Board of Governors, or its Methodology Committee.

Authors' Contributions

ES, JS, and SLL conceived the presented ideas. ES and JS drafted the manuscript. SL supervised and critically reviewed the manuscript. All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on July 22, 2020 [FREE Full text]
2. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017; Long Beach, California, USA URL: <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
3. Hutson M. Robo-writers: the rise and risks of language-generating AI. Nature 2021 Mar;591(7848):22-25. [doi: [10.1038/d41586-021-00530-0](https://doi.org/10.1038/d41586-021-00530-0)] [Medline: [33658699](https://pubmed.ncbi.nlm.nih.gov/33658699/)]
4. Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What makes good in-context examples for GPT-3? arXiv. Preprint posted online on January 17, 2021 [FREE Full text]
5. Emerson. GPT-3 Demo. URL: <https://gpt3demo.com/apps/quickchat-emerson> [accessed 2021-12-14]
6. Langston J. From conversation to code: Microsoft introduces its first product features powered by GPT-3. The AI Blog. 2021 May 25. URL: <https://blogs.microsoft.com/ai/from-conversation-to-code-microsoft-introduces-its-first-product-features-powered-by-gpt-3/> [accessed 2021-12-14]
7. A robot wrote this entire article. Are you scared yet, human? GPT-3. The Guardian. 2020 Sep 08. URL: <https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3> [accessed 2022-02-01]
8. Examples - OpenAI API. OpenAI. URL: <https://beta.openai.com/examples/> [accessed 2021-12-14]
9. Wiens J, Price WN2, Sjoding MW. Diagnosing bias in data-driven algorithms for healthcare. Nat Med 2020 Jan;26(1):25-26. [doi: [10.1038/s41591-019-0726-6](https://doi.org/10.1038/s41591-019-0726-6)] [Medline: [31932798](https://pubmed.ncbi.nlm.nih.gov/31932798/)]
10. Watson J, Hutrya CA, Clancy SM, Chandiramani A, Bedoya A, Ilangovan K, et al. Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? JAMIA Open 2020 Jul 10;3(2):167-172 [FREE Full text] [doi: [10.1093/jamiaopen/ooz046](https://doi.org/10.1093/jamiaopen/ooz046)] [Medline: [32734155](https://pubmed.ncbi.nlm.nih.gov/32734155/)]
11. Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif Intell Med 2020 Apr;104:101822 [FREE Full text] [doi: [10.1016/j.artmed.2020.101822](https://doi.org/10.1016/j.artmed.2020.101822)] [Medline: [32499001](https://pubmed.ncbi.nlm.nih.gov/32499001/)]
12. Kornigebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digit Med 2021 Jun 03;4(1):93 [FREE Full text] [doi: [10.1038/s41746-021-00464-x](https://doi.org/10.1038/s41746-021-00464-x)] [Medline: [34083689](https://pubmed.ncbi.nlm.nih.gov/34083689/)]
13. Miller K. When algorithmic fairness fixes fail: The case for keeping humans in the loop. Stanford University. 2020 Nov 02. URL: <https://hai.stanford.edu/news/when-algorithmic-fairness-fixes-fail-case-keeping-humans-loop> [accessed 2021-12-14]
14. Lucy L, Bamman D. Gender and representation bias in GPT-3 generated stories. In: Proceedings of the Third Workshop on Narrative Understanding. 2021 Jun Presented at: Third Workshop on Narrative Understanding; June 11, 2021; Virtual p. 48-55 URL: <https://aclanthology.org/2021.nuse-1.5.pdf> [doi: [10.18653/v1/2021.nuse-1.5](https://doi.org/10.18653/v1/2021.nuse-1.5)]
15. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? 2021 Mar Presented at: FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency; March 3-10, 2021; Canada. [doi: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922)]
16. McGuffie K, Newhouse A. The radicalization risks of GPT-3 and advanced neural language models. arXiv. Preprint posted online on September 15, 2020 [FREE Full text]
17. Daws R. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves. AI News. 2020 Oct 28. URL: <https://artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/> [accessed 2022-02-01]
18. FitzGerald C, Hurst S. Implicit bias in healthcare professionals: a systematic review. BMC Med Ethics 2017 Mar 01;18(1):19 [FREE Full text] [doi: [10.1186/s12910-017-0179-8](https://doi.org/10.1186/s12910-017-0179-8)] [Medline: [28249596](https://pubmed.ncbi.nlm.nih.gov/28249596/)]
19. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. 2002 Jul Presented at: 40th Annual Meeting of the Association for Computational Linguistics; July 7-12, 2002; Philadelphia, Pennsylvania, USA URL: <https://www.aclweb.org/anthology/P02-1040.pdf> [doi: [10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135)]
20. Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. 2005 Jun Presented at: ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization; June 29, 2005; Ann Arbor, Michigan URL: <https://www.aclweb.org/anthology/W05-0909.pdf> [doi: [10.3115/1626355.1626389](https://doi.org/10.3115/1626355.1626389)]
21. Celikyilmaz A, Clark E, Gao J. Evaluation of text generation: A survey. arXiv. Preprint posted online on May 18, 2021 [FREE Full text]

22. Jiang R. Introducing new Alexa healthcare skills. Amazon Alexa. 2019 Apr 04. URL: <https://developer.amazon.com/blogs/alexa/post/ff33dbc7-6cf5-4db8-b203-99144a251a21/introducing-new-alexa-healthcare-skills> [accessed 2022-02-01]
23. McGraw D, Mandl KD. Privacy protections to encourage use of health-relevant digital data in a learning health system. NPJ Digit Med 2021 Jan 04;4(1):2 [FREE Full text] [doi: [10.1038/s41746-020-00362-8](https://doi.org/10.1038/s41746-020-00362-8)] [Medline: [33398052](https://pubmed.ncbi.nlm.nih.gov/33398052/)]
24. Artificial intelligence and machine learning in Software as a Medical Device. U.S. Food & Drug Administration. URL: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> [accessed 2021-12-14]
25. Patient trust must come at the top of researchers' priority list. Nat Med 2020 Mar;26(3):301. [doi: [10.1038/s41591-020-0813-8](https://doi.org/10.1038/s41591-020-0813-8)] [Medline: [32161404](https://pubmed.ncbi.nlm.nih.gov/32161404/)]
26. Brockman G, Murati M, Welinder P, OpenAI. OpenAI API. OpenAI. 2020 Jun 11. URL: <https://openai.com/blog/openai-api/> [accessed 2021-12-14]
27. Romero A. Can't access GPT-3? Here's GPT-J — its open-source cousin. Towards Data Science. 2021 Jun 24. URL: <https://towardsdatascience.com/cant-access-gpt-3-here-s-gpt-j-its-open-source-cousin-8af86a638b11> [accessed 2021-12-14]
28. Sadilek A, Liu L, Nguyen D, Kamruzzaman M, Serghiou S, Rader B, et al. Privacy-first health research with federated learning. NPJ Digit Med 2021 Sep 07;4(1):132 [FREE Full text] [doi: [10.1038/s41746-021-00489-2](https://doi.org/10.1038/s41746-021-00489-2)] [Medline: [34493770](https://pubmed.ncbi.nlm.nih.gov/34493770/)]

Abbreviations

AI: artificial intelligence
BLEU: Bilingual Evaluation Understudy
EHR: electronic health record
GPT-3: generative pretrained transformer 3
HIPAA: Health Insurance Portability and Accountability Act
METEOR: Metric for Evaluation of Translation With Explicit Ordering
PCORI: Patient-Centered Outcomes Research Institute

Edited by C Lovis; submitted 12.08.21; peer-reviewed by A Trojan, T Caze, G Carot-Sans, J Ayre, W Zhang; comments to author 24.10.21; revised version received 14.12.21; accepted 09.01.22; published 10.02.22

Please cite as:

Sezgin E, Sirrianni J, Linwood SL
 Operationalizing and Implementing Pretrained, Large Artificial Intelligence Linguistic Models in the US Health Care System: Outlook of Generative Pretrained Transformer 3 (GPT-3) as a Service Model
 JMIR Med Inform 2022;10(2):e32875
 URL: <https://medinform.jmir.org/2022/2/e32875>
 doi: [10.2196/32875](https://doi.org/10.2196/32875)
 PMID:

©Emre Sezgin, Joseph Sirrianni, Simon L Linwood. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org/>), 10.02.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.

数据
实时交互，质量管控
——集成平台腾飞之双翼



阳普医疗科技股份有限公司出版
仅供阳普医疗内部使用。
本刊所连载及翻译文章, 版权属原作者及
阳普医疗所有, 请勿用于其他用途!

MKDAPC-B161:202305

医学
拾萃

IMPROVE REVIEW

 阳普醫療科技股份有限公司
IMPROVE MEDICAL INSTRUMENTS CO., LTD

地址：珠海市横琴新区环岛东路3000号ICC大厦南塔23层

服务热线：4001-300030

电话：+86 20-3231-2666

传真：+86 20-3231-2667

网站：www.improve-medical.com

邮编：510530

