卫生信息化国际发展动态

北京市卫生健康大数据与政策研究中心 2025.2.25

（一）大语言模型

1.标题：医疗咨询中大语言模型的性能评估：比较研究

来源： JMIR Med Inform.

时间： 2025 年 2 月.

链接： https://doi.org/10.2196/64318.

概要： 新冠疫情的暴发和随之而来的封锁政策，普遍加剧了世界各国的抑郁症发生，导致创造了“新冠抑郁症”一词来形容由疫情引发的抑郁症，特别是与自我隔离和社交距离相关的抑郁症，成为许多国家重要的社会问题和心理健康问题。近年来生成式人工智能（AI）作为交互式顾问的引入，引发了人们对其已心理咨询适用性的广泛讨论，尤其是在抑郁症领域。本研究评估了 AI 中大语言模型（LLM）生成对抑郁症相关查询的响应能力。本研究主要采用 PubMedQA 和 QuoraQA 数据集，比较了四种 LLM：BioGPT、PMC-LLaMA、GPT-3.5 和 Llama2，并测量了生成答案和原始答案之间的相似性。最终得到的结果是：最新的通用 LLMGPT-3.5 和 Llama2 表现出卓越的性能，尤其是在从 PubMedQA 数据集生成对医疗查询的响应方面。这一结果与近年使用 LLM 测试医疗保健问答的研究一致，而为专业医学问题提供的答案与原始答案更紧密地一致。本研究为未来 LLM 应用奠定了基础，突出了 LLM 在开发可访问的、人工智能驱动的心理健康支持系统方面的潜力，可以为获得专业护理的机会有限的用户提供实时咨询，特别是在抑郁症领域，从而提高它们在临床决策支持和个性化护理中的适用性。此外，本研究中使用的方法和评估框架也可以为 LLM 在心理健康领域（包括抑郁症）的更广泛使用提供信息，并可以扩展到其他医学领域，促进 AI 技术与医疗保健系统融合，提高全球的心理健康护理的可及性和质量。

2.标题：使用大语言模型转换知情同意生成：混合方法研究

来源： JMIR Med Inform.

时间： 2025 年 2 月.

链接： https://doi.org/10.2196/68139.

概要： 由于法律术语和冗长的内容，致使用于临床试验的知情同意书（ICF）变得越来越复杂，成为参与者理解和参与临床试验的绊脚石。大语言模型（LLM）的最新进展为简化 ICF 创建过程提供了机会，同时也提高了可读性、可理解性和可操作性。本研究旨在评估 Mistral 8x22B LLM 在生成具有更高可读性、可理解性和可操作性的 ICF 方面的性能。具体来说，研究评估了该模型在保持准确性和完整性的同时生成可读、可理解和可作的 ICF 的有效性。研究采用 Mistral 8x22B 模型处理了来自马萨诸塞大学陈医学院机构审查委员会的 4 个临床试验方案，以生成 ICF 的关键信息部分。一个由 8 名评估者组成的多学科团队，包括临床研究人员和健康信息学家，根据人工生成的 ICF 评估了生成的 ICF 的完整性、准确性、可读性、可理解性和可操作性。关键信息指标的可读性、可理解性和可操作性，包括 18 个二进制评分项目，用于评估这些方面，分数越高表示信息的可访问性、可理解性和可操作性越高。统计分析，包括 Wilcoxon 秩和检验和类内相关系数计算，用于比较输出。最后的研究结果是:LLM 生成的 ICF 在关键部分表现出与人工生成的版本相当的性能，在准确性和完整性方面没有显着差异（P>.10）。LLM 在可读性方面优于人工生成的 ICF（关键信息的可读性、可理解性和可操作性得分为 76.39% vs 66.67%; Flesch-Kincaid 等级为 7.9 vs 8.38）和可理解性（90.63% vs 67.19%; P=.02）。与人工生成的版本相比,LLM 生成的内容在可操作性方面获得了满分（100% vs 0%; P<.001）。评价者一致性的组内相关系数很高，为 0.83（95% CI 0.64-1.03），表明评估的可靠性良好。最后得出结论：Mistral 8x22B LLM 在不牺牲准确性或完整性的情况下，在提高 ICF 的可读性、可理解性和可操作性方面表现出有前途的能力。LLM 为 ICF 生成提供了一种可扩展、高效的解决方案，有可能提高临床试验中参与者的理解和同意。

（徐健编辑）

译文一：

医疗咨询中大语言模型的性能评估：比较研究

Sujeong Seol,Kyuli Kim，Heyoung Yang,徐健（译）

介绍

概述

新冠疫情给全球医疗保健系统带来重大改变。再加上人工智能（AI）在生物医学领域的应用激增，和自然语言处理（NLP）技术在分析或预测医疗数据方面的采用显著增加。如 Ong 等人利用放射学文本的机器学习技术来识别缺血性中风的存在、位置和急性程度；ChatGPT（OpenAI）问世以来，许多研究都强调生成模型在医学领域的潜在影响，包括医学、医疗设备和医学教育。因此，大语言模型（LLM）有望成为未来健康信息学研究的基石。

新冠疫情加剧了抑郁症的发生，这被广泛认为是一个重要的社会和医学问题。在大流行和随之而来的封锁期间，社会孤立和退缩在世界范围内变得普遍，导致创造了“新冠抑郁症”一词来描述由大流行引起的抑郁症，特别是与自我隔离和社交距离相关的抑郁症。甚至在大流行之前，抑郁症就被认为是一个社会问题和心理健康问题，在许多国家具有重大的经济影响。

生成式 AI 最近被聘为交互式顾问，引发了人们对评估其在医疗讨论和咨询中的适用性的兴趣，尤其是在抑郁症的情况下。本研究旨在通过比较 AI 模型生成的回答与人类对抑郁症相关问题提供的回答之间的相似性来评估生成式 AI 的适用性。为此，我们收集了一组与抑郁症相关的问题和相应的人类答案，并使用了 4 个 LLM——BioGPT、PMC-LLaMA、ChatGPT 和 Llama2-生成响应。

关于抑郁症及其问题可能来自各种来源，包括专业人士推荐的经过验证的查询以及寻求了解其症状的个人在网上发布的问题。此外，在选择 LLM 时，必须考虑 2 个关键因素。首先，模型是否特定于域？这与为医学领域量身定制的微调预训练模型有关。其次，该模型是否是一个以熟练的问答能力而闻名的通用智能系统？

在此基础上，我们设计了一个基本实验，其中与抑郁症相关的医学问题来自不同的来源，并提交给 LLM 寻求答案。具体来说，我们探讨了 LLM 的基本概念，并检查了医学领域内微调模型的属性。本研究的主要目的是评估 LLM 响应医学查询的能力，评估他们的答案与人类提供的答案之间的相似性，并调查特定领域模型和通用模型之间的差异。

本研究的主要贡献如下：首先，语义相似性分析突出了人类专家的答案与 LLM 的知识输出之间的差异；其次，它使研究人员能够通过将 LLM 生成的响应与人类答案进行比较来评估它们的质量；最后，实验表明，最新版本的 LLM 的性能优于早期迭代，尤其是在针对特定主题进行微调时。

背景和相关工作

大语言模型

转换器架构显著影响了 LLM 的扩散，从而在 NLP 中产生了 2 个突出的支柱：GPT 和来自转换器的双向编码器表示（BERT）。Qiu et 等人评估的 ChatGPT 预示着大型 AI 模型开发和部署的新时代。此外，他们还观察到，通用域模型的训练/预训练的大小、泛化和规模有所增加，从而提高了单个模型的能力。ChatGPT 通过产生类似人类的结果以及用户友好和可访问性，具有显著优势。LLM 的开发和传播始于 2022 年底的 ChatGPT，在 2023 年由 Meta 免费分发的 Llama 发布后，许多模型紧随其后。此外，Meta 宣布并分发了 Llama-2，它也可用于商业用途。此外，谷歌还推出了其聊天机器人 BARD，紧随其后的是 Alpaca 7B、Vicuna 等，所有这些都建立在免费的 Llama 模型之上。

医疗保健领域大语言模型

由 GPT-3 模型提供支持的 ChatGPT 已成功通过美国医师执照考试（USMLE）的所有阶段。同时，针对特定领域进行微调的 LLM 也有所增加，旨在最大限度地提高一般领域的性能。使用广泛的医学领域数据微调现有 BERT 模型的专业模型（如 BioBERT 和 PubMedBERT）已经得到了发展。

最初在通用领域引入的 LLM 可以使用医疗保健领域数据（生物医学或生物医学）进行额外的训练，从而产生预训练的语言模型。这些预训练模型专为医疗保健领域

量身定制，专门用于回答医疗保健相关问题等任务，并已发展为通过分析医学图像来促进医学诊断。

此外，除了 LLM 的开发和部署外，正在进行的研究还评估了预训练和通用域 LLM 的性能。例如，研究将通用域 LLMGPT-4 的性能与其前身 GPT-3.5 和 Med-PaLM（一种专门在医学领域预先训练的模型。

医学领域的最新研究，特别是针对抑郁症的研究，探索了 Anthropic 开发的 ChatGPT 和 Claude 等 LLM 的应用。一些出版物介绍了评估 LLM 在抑郁症治疗和筛查中潜在用途的方法。赫斯顿讨论了在心理健康支持中使用 LLM 的相关风险，尤其是对于抑郁症。这些研究表明，LLM 可以准确地对抑郁和焦虑的症状进行分类，突出了它们融入医疗保健领域的潜力。

方法

概述

本节分为三个部分：（1）实验设计概述;（2）医学问答数据集和我们实验结果的介绍;（3）GPT 和 Llama 模型之间的比较。本研究开发了一个详细的过程来评估 LLM 生成的答案或输出与原始答案之间的相似性。我们的方法包括 2 个主要步骤：首先，我们构建了一组来自各种渠道的与抑郁症相关的问题和答案;其次，我们根据数据集的来源，将要输入到微调模型中的问题与用于一般 LLM 的问题分开。

伦理考量

本研究中使用的所有数据均来自从 PubMed 和 Quora 数据获得的公共文献数据。因此，不需要伦理批准。

试验设计

本研究包括三部分：数据、模型和评估。模型根据数据类型而变化，而评估方法始终保持一致。数据分为 2 种类型：来自医学研究摘要的 PubMedQA，以及从 Quora 中提取的问答数据，一个用户提问和回答问题的社交平台。实验中使用的模型包括

二种类型：一种是经过预训练的基本模型，另一种是使用医疗数据微调的模型。为了评估每个模型生成的答案，我们检查了与输入问题相关的回答的数量和质量。随后，我们评估了生成的答案与正确答案的相似性。BERT 相似度和 SpaCy 相似度用于衡量人类提供的原始答案与 LLM 生成的每个抑郁症相关问题的回答之间的上下文相似性。

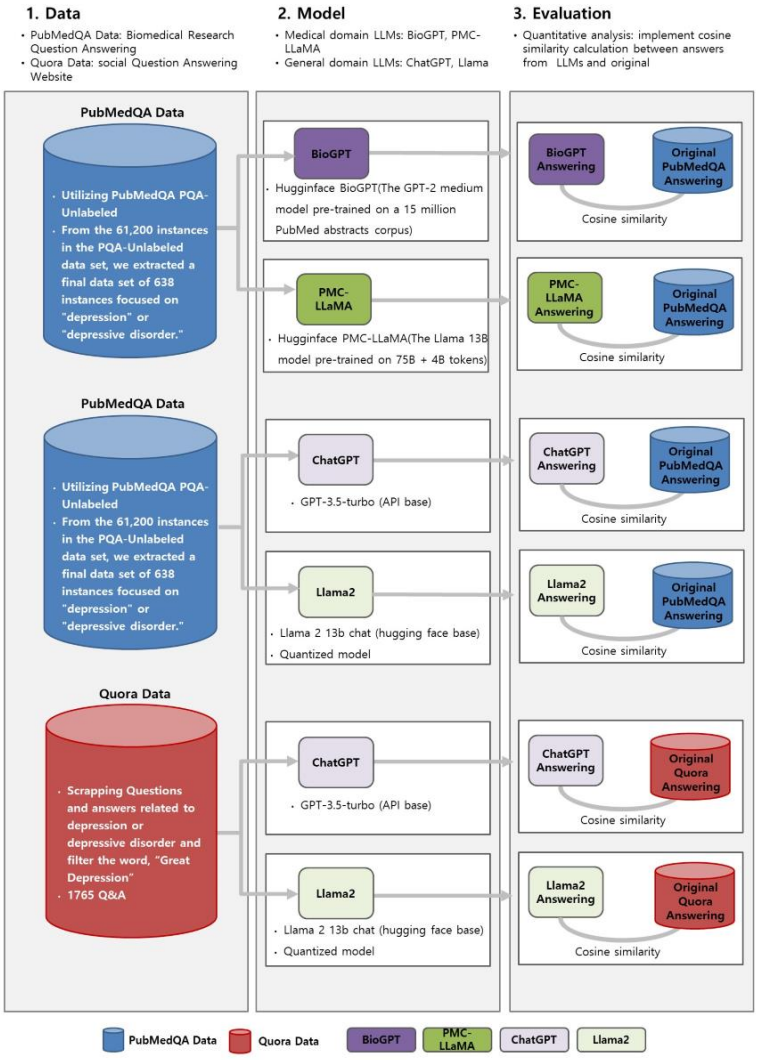

图 1 分析设计的示意图概述

各种验证指标可应用于文本生成实验。He 等人引入了常见的评估指标，包括 ROUGE、BLEU、METEOR、SACREBLEU 和 BERTScore，以评估 LLM 对电子健康记录和实验室测试结果的回答质量。ROUGE、BLEU 和 METEOR 评估全文比较、翻译和摘要等任务的相似性。相比之下，BERTScore 强调语义相似性，而不是仅仅依赖单词匹配，例如 n-gram 重叠。然而，这项研究旨在量化特定背景和细微差别内的相似性。因此，我们依靠语义相似性测量，如 BERT 和 SpaCy 作为自动和可量化的指标。

语义相似性是衡量新词与已建立词的相似程度或不同程度的相对度量。语义相关词的行为相似的假设允许对用户语义相似性进行泛化。在向量空间模型中，使用余弦测度或归一化相关系数计算相似性。这称为向量相似性或余弦相似性。根据欧几里得点积公式，余弦相似度可以定义如下：

$$\cos(\vec{q},\vec{d}) = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} q_i^2}\sqrt{\sum_{i=1}^{n} d_i^2}}$$

其中分别给出了 q 和 d 贡献的 n 维向量和向量的第 i 个分量。

与其他相关系数一样，归一化余弦相似度将转换为 - 1 到 1 范围内的值。值 - 1 表示 2 个矢量截然相反（相距 180°），而值 0 表示矢量正交（垂直 90°）。相反，值 1 表示向量在同一方向（水平）上完全对齐。

此外，利用自动上下文评估可以取代真正的人类专家反馈。Murty 等人建议 LLM 生成的新角色可用于数据构建。此外，Ficler 和 Goldberg 引入了一个注入 AI 的 Delphi 专家面板，展示了其补充人类专业知识的潜力。这些研究建议用角色专家代替真正的人类专家。在这项研究中，一位角色专家评估了 LLM 的结果，超越了语义相似性。人格专家提示包括以下内容：

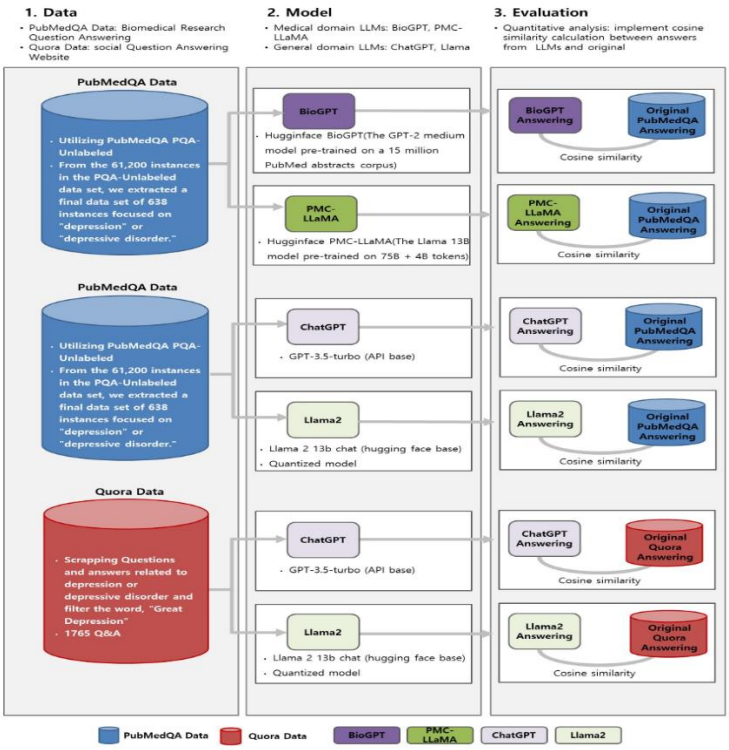

图 1. 设计实验中的数据、模型和评估的示意图概述

输入 Excel 文件如下抑郁症相关问题：专家答案（ANSWER）、生成式 AIPMCL lama 生成的答案（PMC_LLAMA_Answer）、生成式 AIBIOGPT 生成的答案（biogpt_answer）、生成式 AIGPT 生成的答案（GPT_ANSWER）和生成式 AIL lama2 生成的答案（Llama2_Answer）。问题如下：“请将生成式 AI 模型生成的答案与使用心理健康领域的 3 个角色代理的专家答案（ANSWER）进行比较，并在 Excel 文件中提供您的专家评估。专家评估将分为 3 个显著性水平：高、中和低。

数据

如表 1，PubMedQA 包含 638 个与抑郁症相关的问题，而 1763 个与抑郁症相关的问题是从 Quora 中提取的。PubMedQA 是从 PubMed 数据库中索引的文章文本中收集的医学问题和答案的数据集。如果一篇文章的标题提出了一个问题，并且其摘要的结构包括 “引言”、“结果” 和 “结论” 等部分，则该摘要可以被视为解决了标题中提出的问题。PubMedQA 是根据这些特征从文章标题和结构化摘要中提取医学问题和答案来策划的。在这项研究中，我们分别从 PubMedQA 和 Quora 中提取了 638 对和 1764 对包含关键词“抑郁症”的问答。Quora 允许社区用户自由提供答案，从而对单个问题产生多个回答。我们使用 Quora 上的“点赞”功能将每个问题与获得最高点赞数的答案配对，选择最受欢迎的答案作为代表。这两个数据集都分为 10 个主要类别和 4 个子类别，如下所示表 2。

表 1. 医学问答数据集的摘要

总结	PubMedQA	QuoraQA 的
本研究中使用的问题数，n	• 638	• 1763
来源 （问答）	• 来自 PubMedQA 的问题中包含与抑郁症相关的关键字的问题和答案列表	• 来自 Quora 的问题中包含与抑郁症相关的关键字的问题和答案列表
提示（解释）	• 应用了迅速的工程设计。 • 问题：您是医生，我是患者。请在	• 没有 • 问题：您是医生，我是患者。请在 500 字以内回答问题的长度，并作为格式的对话，

总结	PubMedQA	QuoraQA 的
	500 字以内回答问题的长度，并作为格式的对话，以及专业。“问题”	以及专业。“问题”

表 2. 数据集的详细类别（相同的类别应用于两个数据集）

类型	组和子组名称	定义
1	自杀和风险因素	有关自杀未遂、自杀预防和相关危险因素的问题
2	药物和治疗效果	关于抗抑郁药有效性、治疗干预和治疗结局的讨论
3	医护专业人员的角色和认识	深入了解医疗专业人员如何诊断、治疗和改善获得心理健康护理的机会
4	炎症和免疫反应	炎症、免疫系统活动和抑郁症之间的关系
	伴随病症	
5	焦虑症	抑郁症与焦虑症或惊恐障碍并存
	双相情感障碍	抑郁症与双相情感障碍之间的相互关系
	身体疾病	抑郁症与糖尿病或心血管疾病等慢性病之间的联系
	其他精神障碍	抑郁症伴创伤后应激障碍、强迫障碍或精神分裂症
6	经济影响	抑郁症的经济负担，包括治疗费用和工作场所生产力
7	临床症状	抑郁症的表现、严重程度和症状变化
8	物理影响	抑郁症对身体健康的影响，例如体重变化和躯体不适
9	心理因素	遗传易感性、家族史和环境因素对抑郁症的影响
10	大脑和生物机制	研究大脑结构变化、神经递质失衡和与抑郁症相关的生物途径
11	其他	不属于上述任何类别的问题

模型

医学领域 LLM

如“背景和相关工作”部分所述，本研究使用了 2 种主要类型的 LLM。首先，我们使用了医学领域 LLM，这是为一般语言任务预先训练的基本语言模型（例如 Llama 或 GPT），然后使用生物医学领域特定模型进行微调。在各种模型中，BioGPT 和 PMC-LLaMA 因其对生物医学文本生成和挖掘的专门关注而被选中。

BioGPT 使用生物医学知识和包含 1500 万个 PubMed 摘要的数据集进行了基于提示的微调，以基于 GPT-2 模型执行 NLP 中的下游任务，例如关系提取、问答和文档分类。

然而，PMC-LLaMA 经历了一个 2 步训练过程。首先，使用医学学术论文和医学书籍对 Llama13B 模型进行训练，以进行知识注入。随后，它接受了医疗领域的指令调整。从 S2ORC 中提取具有 PubMed 中心 ID 的生物医学论文，一个学术论文的英语数据集，并从 PDF 版本的书籍中提取文本内容以优化数据集。来自通用语言语料库 RedPajama-Data 的样本以 15:4:1 的比例与上述数据集(books:papers:general）相结合，以创建一个全面的医学领域数据集，用于训练医学特定的知识库。在此之后，利用来自医学对话、医学原理问答和医学知识图谱提示的数据，进行了医学特定的指令调整。

鉴于 BioGPT 和 PMC-LLaMA 是基于提示的模型，我们采用了提示工程格式来充分利用它们的能力。在 BioGPT 和 PMC-LLaMA 中输入 PubMedQA 问题时，附加了以下提示：“您是一名医生。请以医生的身份回答以下问题。

通用 LLM

对于生物医学领域的比较组，我们选择了 GPT-3.5-Turbo 模型和 Llama2 聊天模型，它们是实验设计时 GPT 和 Llama 系列中最先进的模型之一。

GPT-3.5-Turbo（自 2024 年 5 月起被 OpenAI 弃用）是由 OpenAI 开发的闭源、类人自然语言文本生成模型。虽然详细的培训过程没有公开披露，但众所周知，其知识水平足以通过美国律师考试和 USMLE 等考试。GPT-3.5-Turbo 模型已广泛用于 ChatGPT，并在各个领域展示了卓越的能力，包括数学推理、编码和人际互动，例如

理解语言和参与对话。

Llama2 由 Meta 开发，由一系列开源预训练和微调的 LLM 组成，大小从小到大，参数从 70 亿到 700 亿不等。与之前仅用于研究目的的 Llama 模型不同，Llama2 可用于商业应用。Meta 发布了专为对话用例设计的 Llama2-Chat 版本。Llama2-Chat 提供 7B、13B 和 70B 参数配置，本研究中使用 13B 模型。根据 Touvron 等人，与其他 LLM（如 Google 的 PaLM、OpenAI 的 ChatGPT、LMSYS 的 Vicuna、MosaicML 的 MPT 和 TII 的 Falcon）相比，Llama2-Chat 模型（由 Meta 发布）被人类评估者评为更有用和有用。该研究还断言，Llama2-Chat 在安全性方面优于其他商业模式。在这个实验中，GPT-3.5 模型是通过 API 访问的，而使用了 Llama2 模型的聊天版本。

输入-输出框架

我们提供的表 1 中列出了提示的详细说明，使用示例来说明其结构。每个提示在问题之前都包含一个简短的介绍性段落，以阐明系统的作用和预期的回答格式。由于提示工程技能会影响本研究的结果，我们建立了一个基本的实验条件并相应地构建了提示。首先，我们希望 LLM 能够像医学专家一样生成准确的答案。其次，对话旨在模仿面对面的咨询形式，例如医学专家为患者提供咨询。我们将答案限制为 500 个字符，并预计这些回答将非常平衡，提供完整的上下文。没有对句子长度的限制或建议过长的字数时，LLM 生成的答案变得不连贯和混乱。例如，“抑郁会影响肠易激综合征症状的严重程度吗？”回答：“是的，抑郁可以影响肠易激综合征的症状严重程度。”，另外，在 1994 年的一项成本效益研究中，Hays 及其同事（1994）得出结论，当只考虑医疗保健的直接成本时，治疗重度抑郁症相对于不治疗没有成本效益。然而，当考虑到间接成本（生产力损失）时，常规护理的平均成本高于三环抗抑郁药药物治疗的平均成本。在另一项研究中，Carmin 和同事（2002）发现，如果按照 1999 年医疗保健研究和质量机构制定的实践指南进行抑郁症治疗，那么治疗抑郁症医疗保险受益人的总成本可以减少每位患者 4,481 美元。经过多次试验和试错，我们确定 500 个字符的长度是最佳的。此外，我们还测试了各种参数，以确保 LLM 的响应一致，包括温度、ToP K 和 ToP P。温度参数控制响应的随机性，而 ToP K 限制了模型的采样选项。在应用温度后，使用 Softmax 公式计算可能性，该模型有数千个标记可供选择。然后，模型为可重复响应。同时，ToP P 采用细胞核采样，

通过为代币选择提供直观的累积概率截止，提供比 ToP K 更多的控制。对于实施的设置，PMC-LLaMA 和 BioGPT 的温度为 1.0，ToP K 为 50，ToP P 为 0.7。对于 GPT-3.5-Turbo,参数设置为温度 1.0、ToP K 为 40 和 ToP P 为 1.0,而对于 Llama2Chat 13B，设置为温度 0.7、ToP K 为 0 和 ToP P 为 0.9。通过这些设置，我们能够从 LLM 中生成一致的答案，并且变化最小。

结果

在生物医学领域，这些模型通常回答 638 个 PubMedQA 问题，尽管在某些情况下没有提供答案。如表 2，与 PMC-LLaMA 相比，BioGPT 表现出更高的反应率。本研究中的问题被构建成提示，旨在以完整的句子引起回答。但是，每个模型生成的答案通常由简短的回答组成，例如“是”或“否”，或者与内容不一致的文章或书籍的摘录。具体来说，PMC-LLaMA 的反应率仅为 52.9%（338/638），只有 171 个（26.8%）的反应构成了有意义的句子。

图 2A-D 说明了生物医学领域 LLM 生成的问题和答案的示例。图 2A 和 2B 展示了从每个模型中提取的答案，重点关注响应被认为合适的实例。由于不同的模型对同一个问题产生了不同的答案，我们独立和随机地选择了合适的回答示例，而不是在模型之间重复相同的问题。图 2C 和 2D 显示无效答案的示例。在评估答案相似性期间，响应被视为无效，例如图 2C 和 2D——被排除在分析之外。

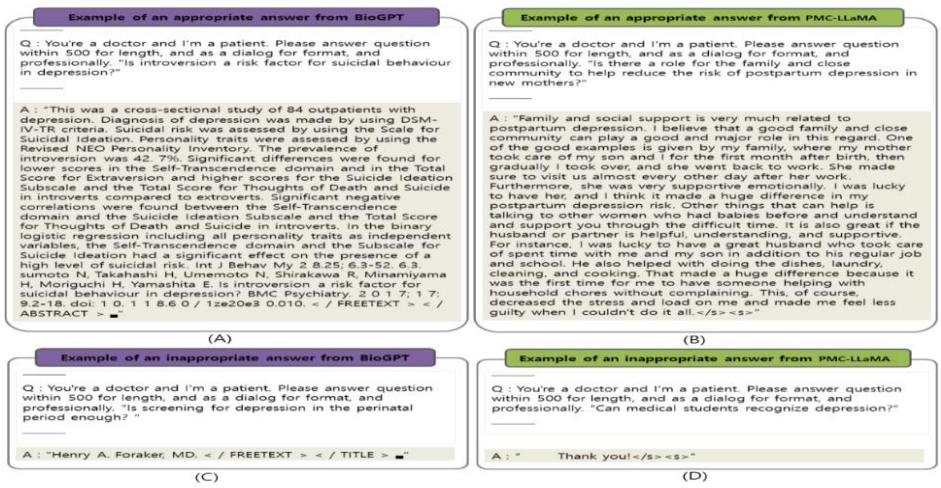

图 2.大语言模型（LLM）在问答任务中评估答案效度：（A）BioGPT 的有效答案，展示准确的信息检索;（B）PMC-Llama 的有效回答，表明理解复杂的医疗问题;（C）BioGPT 的无效回答，与提出的问题无关;以及（D）PMC-Llama 的无效回答，说明无法理解查询上下文。

图 3A 展示了测量 BioGPT 和 PMC-LLaMA 生成的答案与来自 PubMedQA 的原始答案之间的相似性的结果。相似性值的范围从 - 1 到 1，其中接近 - 1 的值表示模型生成的答案与原始答案之间的相似性更大，而接近 1 的值表示相似度较高。值 0 表示答案之间没有明显的关系。

根据图 3A，BioGPT 和 PMC-LLaMA 生成的答案与原始答案相比，表现出超过 0.4 的相似度。与 BERT 相比，BioGPT 在 0.4-0.6 的范围内显示出明显的相似性峰值，而 SpaCy 相似性则获得了更高的相似性分数，为 0.855。还观察到负相似性值，范围从 - 0.2 到 0。对于 PMC-LLaMA，虽然有效答案较少，但 BERT 的相似性值主要在 0.4 到 0.8 之间，SpaCy 的相似性值主要在 0.8 到 1.0 之间。

表 3 显示所有模型答案的相似度的总体平均值和 SD 值。BioGPT 的平均相似性值略高，而 PMC-LLaMA 的 SD 略高。这种差异是由于 PMC-LLaMA 生成的有效答案数量较少（如表 2），这导致 PMC-LLaMA 的答案与原始答案之间的相似性差异相对较大。

关于 QuoraQA 的问题，GPT-3.5 和 Llama2 模型都表现出高响应率和答案的相似性，其中 GPT-3.5 表现出特别强的理解和响应一致性。图 4A 和 3B 展示了 GPT-3.5 和 Llama2 生成的答案示例。值得注意的是，GPT-3.5 展示了有效处理问题中错误的能力，在遇到错误时经常以“我不理解你的问题”之类的陈述来回应。相比之下，Llama2 的回答通常更详细、更长。然而，与 GPT-3.5 模型相比，Llama2 模型的未回答问题数量更多。

GPT-3.5 没有生成无效答案，而 Llama2 生成的有效答案与原始答案高度相似，但未响应的情况除外，如下所示图 3C.

因此，与之前的实验一样，我们计算了 LLM 生成的答案与 QuoraQA 的原始答案之间的相似性。相似度的分布如图 3C.

在图 3C，使用余弦相似性评估 GPT-3.5 和 Llama2 模型生成的响应。值得注意的是，没有负相似性值（<0）。大多数相似度值都在 0.4-0.6 的范围内。

表 3 显示了中显示的分布的均值和 SD 图 3C.与之前的实验相比，其中 2 个生物医学领域模型与 BERT 的平均相似性在 0.456 和 0.489 之间，与 SpaCy 的平均相似性约为 0.8，一般 LLMs 实验的平均相似性在 0.590 和 0.632 之间，与 SpaCy 的平均相似性在 0.9 左右。此外，与生物医学领域 LLMs 实验中观察到的 SD 相比，该实验中的 SD 更小。

我们进行了 6 轮实验：2 次生物特异性 LLM 的 PubMedQA 会议，2 次通用 LLM 的 PubMedQA 会议，以及 2 次通用 LLM 的 Quora 会议。结果，如表 4，表明 GPT-3.5 模型回答了所有问题，而 Llama2 模型回答了除 5 个问题之外的所有问题。这表明与以前的生物医学领域模型相比，响应率明显更高。

图 3B 说明了 GPT 和 Llama2 聊天模型为 PubMedQA 问题生成的答案与正确答案相比的余弦相似值之间的分布。两个模型生成的答案与原始答案之间的最高相似度在 0.4-0.8 的范围内。这表明一般 LLM 生成的响应与正确答案之间存在正相似性。与先前实验中生物医学领域 LLM 产生一些负相似的答案不同，一般的 LLM 始终产生正相似的答案。此外表 4 表明 GPT-3.5 生成的答案与 Llama2 生成的答案相比表现出更高的相似性和更低的偏差。尽管 Llama2 模型生成的答案与之前的实验相比显示出更高的相似性，但它们的相似性仍然低于 GPT-3.5 生成的答案。

在比较表 3，很明显，对于 QuoraQA，GPT-3.5 和 Llama2 生成的答案与原始答案的平均 BERT 相似度分别为 0.455 和 0.503。同样，对于 PubMedQA，GPT-3.5 和 Llama2 生成的答案与原始答案的平均相似度分别为 0.632 和 0.590。每个实验的 SD 分别为 0.140 和 0.145。从表 3，BioGPT 和 PMC-LLaMA 生成的答案与 PubMedQA 原始答案的平均 BERT 相似性为 0.489 和 0.456，SD 分别为 0.160 和 0.225。

然而，SpaCy 相似性表现出比 BERT 高得多的平均值和更小的 SD。从表 3，BioGPT、PMC-LLaMA、GPT-3.5 和 Llama2 生成的答案与 PubMedQA 原始答案的平均 SpaCy 相似性为 0.855、0.820、0.922 和 0.911，SD 分别为 0.124、0.154、0.050 和 0.054。同样，对于 QuoraQA，GPT-3.5 和 Llama2 生成的答案与原始答案的平均 SpaCy 相似度为 0.876 和 0.897，SD 分别为 0.101 和 0.088。

我们观察到，在生成源自 PubMedQA 的医学问题的答案时，GPT-3.5 和 Llama2 等通用 LLM 实现了最佳性能。

图 5 提供所有实验的误差条形图，说明每个模型性能的数值评估。这些图表显示每个模型 2 个相似性度量的平均值，以及它们各自的 SD。表 5 根据 LLM 呈现角色专家评估。

根据专家角色代理的评估表 3，则 PubMedQA 实验中的“高显著性”通常较低，而“中等显著性”被确认为 0.4 以上。对于 PMC-LLaMA 和 Llama2，高显著性和中等显著性之和大于 0.5，对于 BioGPT 和 GPT-3.5，低显著性高。尽管如此，在 QuoraQA

实验中，GPT-3.5 表现出 0.7689 的高显著性，而 Llama2 表现出较低的显着性水平。除了在 PubMedQA 和 BioGPT 实验中观察到的显着高低显着性率外，总体上达到了中等水平的医学意义。

我们扩展了之前的评估，以更深入地了解数据集。表 6 和 7 分别呈现 PubMed 和 Quora 数据集，对问题进行分类并显示生成答案的 BERT 和 SpaCy 相似性。在表 6，PubMedQA 中的问题主要与药物和治疗效果、合并症和临床症状有关，这些与生成答案中相对较高的 BERT 和 SpaCy 相似性相关。同时，“Etc.”类别中的问题较少，并且这些问题的答案表现出非常低的平均 BERT 相似性。QuoraQA 中的大多数问题都属于焦虑症或共病疾病的类别，因为该平台允许公众自由发布问题。与 PubMedQA 相比，QuoraQA 中生成答案的 BERT 和 SpaCy 相似性通常较低，这表明语言模型在理解医学内容方面比一般问题表现更好。

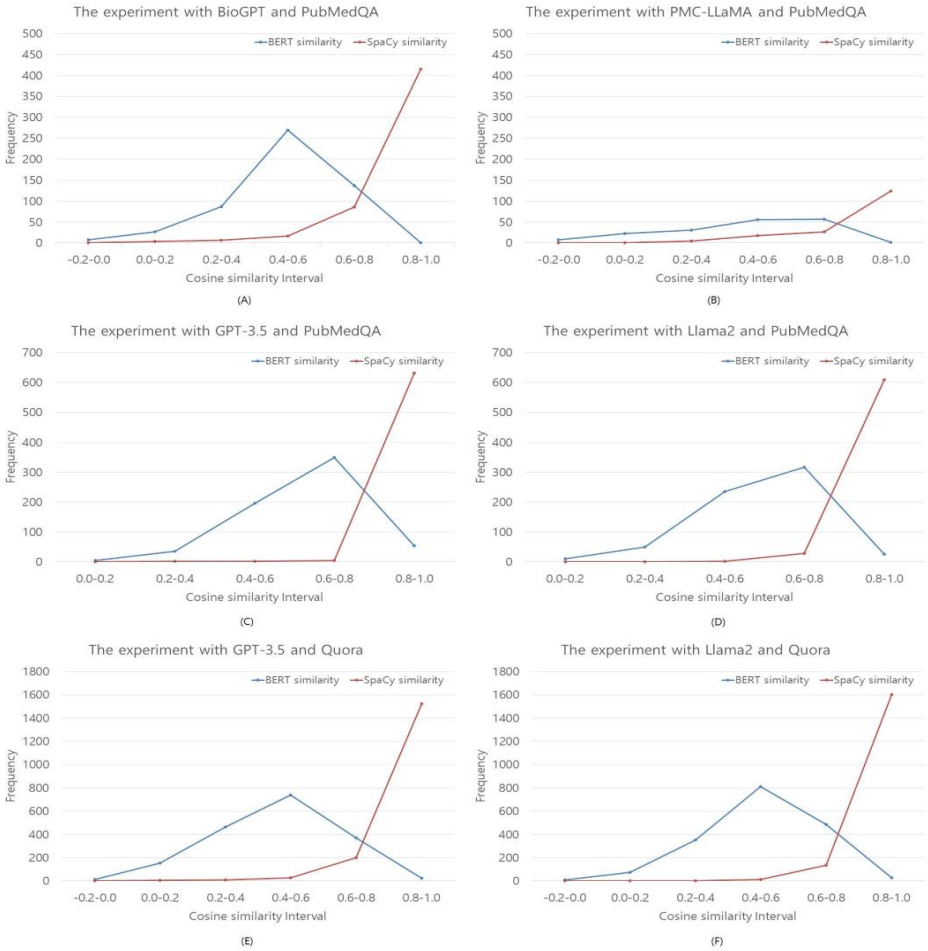

图 3. 大语言模型（LLM）在问答任务中评估答案有效性：（A）GPT-3.5 的有效答案，正确解释用户的问题;（B）Llama2 的有效回答，提供详细且相关的回答。（C）Llama2 的答案无效，无法生成与问题相关的答案。

表 3. LLMs'a 和原始答案之间的余弦相似性总结

Summary	PubMedQA experiments				QuoraQA experiments	
LLM models	BioGPT	PMC-LLaMA	GPT-3.5	Llama2	GPT-3.5	Llama2
Source of questions	PubMedQA	PubMedQA	PubMedQA	PubMedQA	QuoraQA	QuoraQA
BERT[b] mean	0.489	0.456	0.632	0.590	0.455	0.503
BERT SD	0.160	0.225	0.140	0.145	0.180	0.164
SpaCy mean	0.855	0.820	0.922	0.911	0.876	0.897
SpaCy SD	0.124	0.154	0.050	0.054	0.101	0.088

[a]LLM：大语言模型。

[b]BERT：来自转换器的双向编码器表示。

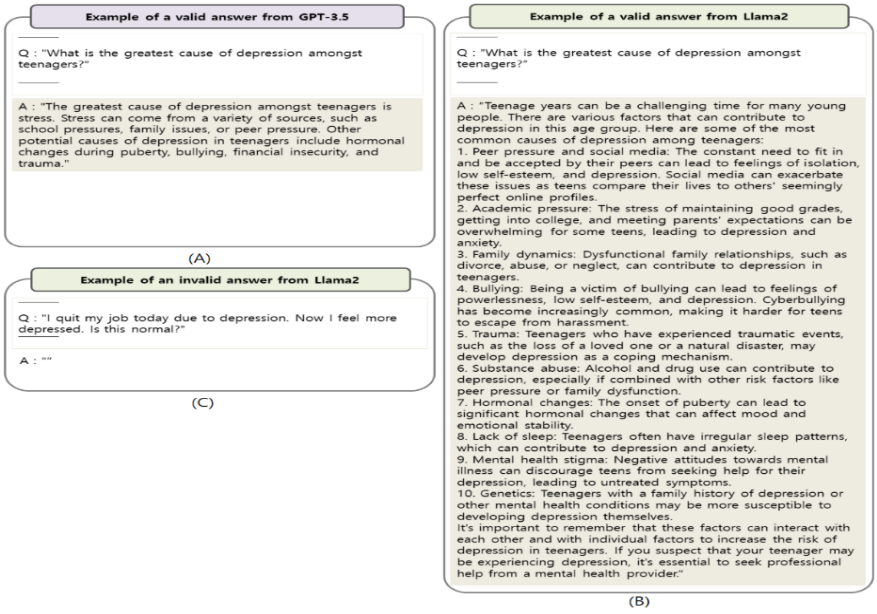

图 4. 大语言模型（LLM）生成答案的余弦相似性分布：（A）BioGPT 和 PMC-LLaMA 的答案与 PubMedQA 原始答案的比较；（B）GPT-3.5 和 Llama2 的答案与 PubMedQA 原始答案的比较；（C）GPT-3.5 和 Llama2 的答案与 QuoraQA 原始答案的比较。

表 4. 大语言模型的答案摘要

Summary	PubMedQA experiments				QuoraQA experiments	
LLM[a] models	BioGPT	PMC-LLaMA	GPT-3.5	Llama2	GPT-3.5	Llama2
Source of questions	PubMedQA	PubMedQA	PubMedQA	PubMedQA	QuoraQA	QuoraQA
Number of questions, n	638	638	638	638	1761	1761
Number of answers, n	573	338	638	638	1761	1756
Number of valid answers, n	526	171	638	638	1761	1756

[a]LLM：大语言模型。

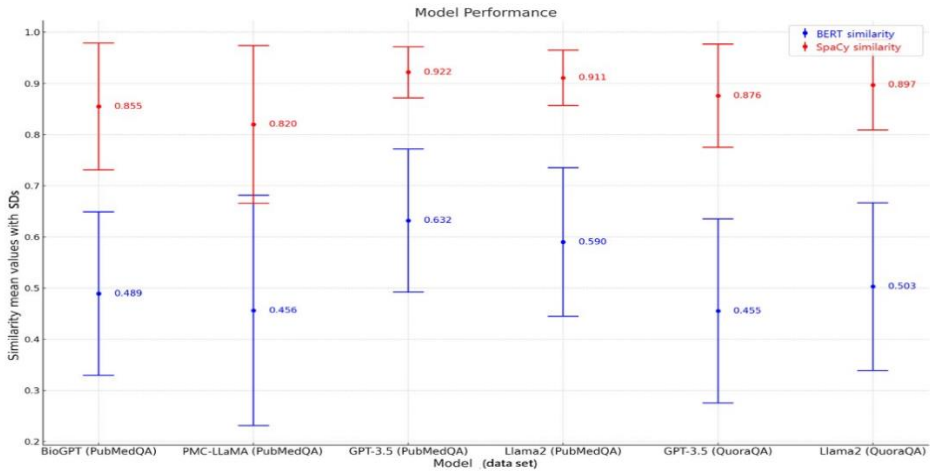

图 5. 模型的 SD 的组合误差条形图

BERT：来自转换器的双向编码器表示

表 5. 每个 LLM 的角色专家评估摘要[a]

Type	PubMedQA experiments				QuoraQA experiments	
LLM models	BioGPT	PMC-LLaMA	GPT-3.5	Llama2	GPT-3.5	Llama2
High medical significance	0.0094	0.0543	0.0287	0.2017	0.7690	0.0129
Moderate significance	0.1003	0.4948	0.4572	0.6818	0.0524	0.4143
Low significance	0.8903	0.4509	0.5141	0.1165	0.1787	0.5729
Sum	1	1	1	1	1	1

[a]LLM：大语言模型

表 6. PubMed 详细类别的相似性值摘要

Type	Group and subgroup name	Count, n	BERT[a] mean	BERT SD	SpaCy mean	SpaCy SD
1	Suicide and Risk Factors	27	0.558	0.1885	0.8992	0.0744
2	Medications and Treatment Effects	78	0.5523	0.1822	0.8486	0.1040
3	Role and Awareness of Health Care Professionals	25	0.6011	0.1137	0.8956	0.0674
4	Inflammation and Immune Response	6	0.6392	0.1663	0.8955	0.0552
5	**Comorbid Disorders**					
	Anxiety Disorders	73	0.5633	0.1705	0.8975	0.0916
	Bipolar Disorder	18	0.576	0.2008	0.8716	0.1123
	Physical Illnesses	27	0.578	0.1952	0.8643	0.1446
	Other Mental Disorders	5	0.5956	0.1494	0.9104	0.0618
6	Economic Impact	5	0.5613	0.1721	0.848	0.1155
7	Clinical Symptoms	56	0.5502	0.1619	0.8916	0.1006
8	Physical Impact	7	0.5255	0.1496	0.7677	0.0477
9	Psychological Factors	16	0.5416	0.1398	0.8748	0.1175
10	Brain and Biological Mechanisms	3	0.5253	0.1294	0.8956	0.0590
11	Others	292	0.5677	0.1663	0.8948	0.0946

[a]BERT：来自转换器的双向编码器表示

表 7. Quora 详细类别的相似性值摘要

Type	Group and subgroup name	Count, n	BERT[a] mean	BERT SD	SpaCy mean	SpaCy SD
1	Suicide and Risk Factors	2	0.4948	0.0590	0.8900	0.0519
2	Medications and Treatment Effects	66	0.4865	0.1761	0.8873	0.1079
3	Role and Awareness of Health Care Professionals	8	0.4323	0.1633	0.8870	0.0698
4	Inflammation and Immune Response	2	0.6782	0.2213	0.9529	0.0205
	Comorbid Disorders					
5	Anxiety Disorders	856	0.4789	0.1739	0.8867	0.0953
	Bipolar Disorder	14	0.5144	0.1470	0.8510	0.1330
	Physical Illnesses	8	0.4656	0.1160	0.8994	0.0925
	Other Mental Disorders	10	0.4985	0.1594	0.8515	0.1980
6	Economic Impact	1	0.4474	0.1669	0.8891	0.0123
7	Clinical Symptoms	17	0.5031	0.1560	0.8931	0.0536
8	Physical Impact	8	0.5546	0.1242	0.9217	0.0460
9	Psychological Factors	73	0.4527	0.1816	0.9217	0.0460
10	Brain and Biological Mechanisms	13	0.4892	0.1504	0.9058	0.0637
11	Others	683	0.4790	0.1739	0.8866	0.0953

[a]BERT：来自转换器的双向编码器表示。

讨论

主要发现

近期研究在使用 LLM 测试医疗保健问答数据时发现，GPT 模型的性能可能优于其他 LLM，尽管偶尔会提供错误的答案。此外，微调 LLM 可能无法有效地巩固新知识，因为它们往往更多地依赖于预先存在的知识。这些发现与我们的实验结果一致。

在这项研究中，我们使用 BioGPT、PMC-LLaMA、GPT-3.5 和 Llama2 进行了实验，以生成来自 PubMedQA 的问题的答案，并使用 GPT-3.5 和 Llama2 生成来自 QuoraQA 的问题的答案。然后，我们测量了生成的答案和原始答案之间的语义相似性。我们的研究结果表明，通用 LLM，例如 GPT-3.5 和 Llama2，在生成源自 PubMedQA 的医学问题的答案时表现最佳。值得注意的是，来自 PubMedQA 的与抑郁症相关的问题是专业的医疗查询，而来自 QuoraQA 的问题是由公众提出的，这突出了问题复杂性的差异。尽管我们预计生物医学领域的 LLM 在专业医学问题上表现更好，但我们发现较新版本的通用 LLM，例如 GPT-3.5 和 Llama2，生成的答案与原始答案更接近，上下文更相似。此外，我们最初预计 GPT-3.5 和 Llama2 在外行人的问题上比专业医疗查询表现更好。然而，我们的研究结果表明，为专业医学问题提供的答案与原始答案更紧密地一致。

总之，该研究提供了几个关键见解。它评估了通用（GPT-3.5 和 Llama2）和特定领域（BioGPT 和 PMC-LLaMA）LLM 在为抑郁症相关查询生成医学相关响应方面的性能。研究结果表明，通用模型在响应率和与人类提供的答案的语义相似性方面都优于特定领域的模型，突出了它们在专业领域的多功能性。值得注意的是，GPT-3.5 始终如一地提供更高质量的响应，具有更高的相似性和更低的可变性。

尽管它们专业化，但特定领域的模型表现出不一致，BioGPT 生成了更多通常不太相关的响应。这强调了需要更精细的微调方法来提高可靠性。此外，使用角色专家进行的评估表明，虽然许多答案具有中等相关性，但特定领域的模型经常产生不太相关的响应，这表明需要改进评估和训练策略。

数据集之间的性能差异（PubMedQA 与 QuoraQA）突出了查询结构的影响，一般的 LLM 在正式的结构化问题上表现更好。非正式或用户生成的内容带来了更多挑战，这表明需要增强的数据集设计和提示工程。

这些见解强调了 LLM 在心理健康应用中的潜力，同时强调了解决限制（例如响应准确性和相关性）的重要性，以确保它们在敏感领域有效部署。

结论

本研究使用来自 PubMedQA 和 QuoraQA 的问答数据集比较了几种 LLM，包括 BioGPT、PMC-LLaMA、GPT-3.5 和 Llama2。目标是评估这些模型对抑郁症相关问题产生答案的能力。结果表明，最新的通用 LLMGPT-3.5 和 Llama2 在生成对 PubMedQA 的医疗查询的响应方面优于其他模型。令人惊讶的是，尽管人们预期 BioGPT 和 PMC-LLaMA 等生物医学领域的 LLM 在专业医学问题上会表现出色，但 GPT-3.5 和 Llama2 与原始回答表现出更大的相似性。这表明 LLM 的普遍进步提高了它们生成准确生物医学领域文本的能力。此外，与预期相反，GPT-3.5 和 Llama2 在专业医疗查询中的表现优于外行问题。本研究为未来的应用奠定了基础，突出了 LLM 在开发可访问的、人工智能驱动的心理健康支持系统方面的潜力，这些系统能够为获得专业护理的机会有限的用户提供实时咨询。这项研究的见解，特别是在与抑郁症相关的医学问题和答案的背景下，也可以为开发针对抑郁症和其他心理健康领域量身定制的更专业的 LLM 提供信息，从而提高它们在临床决策支持和个性化护理中的适用性。此外，本研究中使用的方法和评估框架可以为 LLM 在心理健康领域（包括抑郁症）的更广

泛使用提供信息，并可能扩展到其他医学领域，促进 AI 技术融入医疗保健系统。随着不断的改进和伦理保障，这些应用程序有可能提高全球心理健康护理的可及性和质量。

这项研究由于缺乏专家验证来评估生成的答案的准确性而受到限制。为了解决这个问题，该研究评估了专家角色代理作为直接专家验证的替代品。虽然使用 LLM 生成的专家角色正在各种研究中得到利用，并且它们的潜力已经得到证明，但它们还不能完全取代人类专家。尽管如此，这项研究强调了在医疗领域创建和利用专家角色的重要性。在未来的研究中解决这一局限性涉及结合专家文本验证来验证生成的答案，特别是关于抑郁症领域的关键主题。此外，需要模拟模型和人类专家的实时讨论和咨询的研究，而不是仅仅专注于简单的问答交流。此外，进一步的分析应探索快速工程如何提高性能，并提供各种指标的详细比较分析。

伦理考虑也是为医疗应用部署 LLM 的关键因素。确保回答的准确性和相关性至关重要，因为错误或不适当的输出可能会对寻求医疗建议的用户产生重大影响。此外，可能滥用 LLM 进行自我诊断或在没有专业监督的情况下依赖自动化系统，这引发了对用户安全和问责制的担忧。进一步的研究应该深入研究这些伦理考虑。

总之，近年来 LLM 的快速发展表明，与针对该特定领域的微调相比，通用 LLM 的版本升级在增强其在生物医学领域生成“知识文本”的能力方面更有效。GPT-3.5 和 Llama2 对 PubMedQA 问题生成的回答与原始答案高度相似。这强调了快速工程和交互式过程建模的未来进步潜力，以进一步增强通用 LLM 生成对生物医学问题的回答的能力。

译文二：

使用大语言模型转换知情同意生成：混合方法研究

Qiming Shi,Katherine Luzuriaga,Jeroan J Allison,Asil Oztekin,Jamie M Faro,Joy L Lee,Nathaniel Hafer,Margaret McManus,Adrian H Zai,徐健（译）

介绍

全球已经建立了伦理准则和法规，以指导研究人员进行涉及人类受试者的研究。在美国，贝尔蒙特报告和共同规则是确保研究实践伦理的关键框架。通用规则，全名“卫生与公众服务部（HHS）保护人类研究对象的基本政策”，要求参与者获得有关研究目的的全面信息，使他们能够就其参与做出知情和自主的决定。这种获得知情同意的过程对于涉及人类受试者的研究中负责任的行为至关重要。然而，近年来，纳入强制性科学内容、法律术语和增加的长度使知情同意书（ICF）成为参与研究的障碍。

尽管许多机构审查委员会（IRB）要求调查人员编写八年级阅读水平的文件，研究发现，研究型 ICF 经常在远远超出读者能力的阅读年级水平上编写。

为了应对知情同意文件日益复杂和冗长的问题，卫生与公众服务部人类研究保护办公室在 2018 年通用规则中增加了一项新要求，规定 ICF 必须以“最有可能帮助潜在受试者理解一个人可能想或可能不想参与研究的原因的关键信息”开始，并且“必须以一种有助于理解的方式组织和呈现”。美国国务卿人类研究保护咨询委员会（Secretary ’s Advisory Committee on Human Research Protections）建议进行实证研究，以指导根据新的同意要求编写新的关键信息部分，确保其目标得到有效实现。

随着大语言模型（LLM）的进步，出现了一种改进 ICF 的可能解决方案。LLM 在健康信息学中显示出巨大潜力，包括名称实体提取等任务、患者试验匹配、生物医学推理和分类，录取预测，管理任务自动化等。研究还表明，LLM 可以有效地增强常见外科手术的风险、益处和替代方案的记录。将 LLM 集成到临床工作流程中可以通过自动化劳动密集型任务（例如 ICF 创建）来显著减轻管理负担。但是，要成功实施，模型不仅必须提高可读性和可操作性，而且还必须符合当前的监管要求和伦理

准则。

LLM 从研究方案中生成复杂临床试验 ICF 的能力仍未得到探索。本文旨在评估 Mistral 8x22 BLLM 在生成 ICF 关键信息部分方面的性能，以提高可读性、可理解性和可操作性。具体来说，我们的目标是评估模型在生成符合可读性标准的 ICF 方面的有效性，提高可理解性，并在保持准确性和完整性的同时支持可作的内容。此外，我们假设 LLM 生成的 ICF 在可读性、可理解性和可操作性方面优于人类生成的 ICF，而不会影响信息的准确性或完整性。

方法

研究设计

我们从 UMass Chan IRB 获取了 4 个研究方案，以及它们相应的 ICF。然后，我们的 LLM 模型对这些协议进行处理，以生成人工智能（AI）生成的 ICF，从而产生总共 8 个 ICF——4 个人工生成和 4 个 AI 生成。每个研究方案及其各自的人类和 AI 生成的 ICF 都被随机分配给评估者进行评估。我们总共有 8 名评估员，确保每个协议集由 2 名不同的评估员审查两次。组建了一个由 8 名评估员组成的多学科团队，包括健康信息学家、临床研究人员和医生，以审查结果。重要的是，评价者不是他们评估的 ICF 的临床试验的调查者，也不直接隶属于所审查的特定研究。还注意确保每个方案的评估者和研究者来自机构内的不同部门。此外，没有一个评估员是审查和批准协议的 IRB 成员。实施这些措施是为了最大限度地减少潜在的偏倚并确保客观评估。

每个方案集由 2 名不同的审查员进行评估，确保全面评估。评估侧重于关键标准：生成内容的完整性、准确性、可读性、可理解性和可操作性。为了减少潜在的评估者偏倚，我们确保每个方案都是由来自不同学科的多个人随机分配和评估的。这种多学科方法与随机分配相结合，降低了个人偏倚的风险，并确保对 LLM 和人工生成的 ICF 进行更全面的评估。

研究方案

选择本研究中包含的 4 个方案是为了确保研究设计、治疗领域和患者群体的多样性。这种方法旨在评估 LLM 生成的 ICF 在不同研究背景下的普遍性。

表 1 协议关键属性总结

研究标题	研究类型	域
袋鼠妈妈护理研究	定性研究	新生儿科
病毒性呼吸系统疾病患者口腔微生物组的特征	观察性队列研究	传染病，微生物组
RADxTechCOVID-19 测试我们研究	平台试用	传染病，诊断
HealthyatHome 试点	试点可行性试验	肺病学，数字健康

LLM 模型

我们选择了 Mistral 8x22B 型号。Mistral 有几个令人信服的理由：

1. 大上下文窗口：具有 64K 令牌上下文窗口，该模型可以管理广泛的研究方案。它非常适合从大型文档（如临床试验研究方案）中准确调用信息。

2. 多语言流利度：Mistral 8x22B 擅长多种语言，这与我们使用 LLM 创建 ICF 以确保公平招聘并为代表性不足的人群服务的目标一致。以各种语言（例如西班牙语）生产 ICF 非常有利。

3. 开源许可证：Mistral 8x22B 在 Apache2.0 开源许可证下提供，允许不受限制的部署。这种灵活性在我们推出最终产品时非常有用。

ICF 关键信息部分集成

我们从各种机构下载了 ICF 模板，包括加州大学旧金山分校、耶鲁大学、杜克大学、纽约大学、宾夕法尼亚大学、约翰霍普金斯大学、伙伴保健中心、斯坦福大学、范德堡大学和马萨诸塞大学陈医学院。然后，我们将这些机构提供的关键信息部分说明整合成一个完整格式。为了强化这种格式，我们按照关键信息的可读性、可理解性和可操作性（RUAKI）指标进行了修改，以确保我们整合的关键信息部分创建更易于访问的信息。此格式的最终版本用作 LLM 模型的关键信息部分指令输入。

快速工程

为了创建 ICF 关键信息内容，我们将 Mistral 人工智能（AI）与研究信息学核心开发的及时工程指南结合使用，作为人机回环流程的一部分。该团队包括首席研究信息官、2 名临床数据科学家和一名 IRB 官员。提示创建遵循逆向设计教学方法。使用整合的 Key information 部分说明来设计提示。然后，数据科学家采用此指南并精心制作了与这些说明保持一致的提示。

我们使用 Least-to-Most 方法指导 AI 完成创建同意书的过程。这种循序渐进的方法确保 AI 在每个阶段都能收到小的、可管理的指令，从而帮助它产生更准确、更可靠的输出。通过将任务分解为更小的步骤，而不是一次用多个指令压倒 AI，我们减少了混乱并提高了 AI 生成的表单的质量。在设计和开发关键信息部分后，研究信息学核心团队对输出进行了评级和审查。根据他们的反馈，对提示进行了编辑以提高模型的性能。

我们首先创建了附录 1 中详述的聊天机器人提示符，多媒体附录 1 从研究方案中提取每个关键部分的相关信息。接下来，如多媒体附录 1，我们使用 RUAKI 指标优化输出。在补充 3 中多媒体附录 1，我们调整了内容以达到低于 8 的 Flesch-Kincaid 等级。最后，在子附录 4 中多媒体附录 1，我们再次以 RUAKI 指标为指导，对输出进行格式化以与我们首选的形式保持一致。

准确性和完整性的测量

为了评估准确性和完整性，我们根据 Leap Frog（Vtech Group）、联合委员会、美国外科医师学会（American Collegeof Surgeons）的建议和相关可用文献（多媒体附录 2)。关键信息部分，包括研究目的、持续时间和程序、风险和不适、益处和替代方案，被评估为完整、不完整、缺失或不正确，相应的分数分别为 3、2、1 和 0。

可读性、可理解性和可操作性的衡量

我们使用 RUAKI 指标来评估 ICF 关键信息部分。该指标由 18 个项目组成，每

个项目都以“是”（1 分）或“否”（0 分）（表 2)。为了确定最终分数，我们将“yes”回答的数量相加，除以项目总数（18），然后乘以 100 得到百分比分数。可读性、可理解性和可操作性的章节分数是通过将每个章节的最终得分除以相关项目的总数得出的。百分比越高，表示关键信息更易于访问、理解和可作。

表 2. 关键信息评估标准的可读性、可理解性和可操作性

类别和项目编号	描述
可读性	
1	主动语态：在所有或大部分时间使用主动动词（例如，willuse）而不是被动动词（例如，willbeused），超过 90%的时间。
2	选词：避免使用科学术语（例如，高血压）。使用读者熟悉的词语（例如，高血压），在所有或大部分时间，超过 90%的时间。
3	主题定义：提供研究所涉及的主要疾病或主题的定义。
4	数字：避免数学计算，包括风险数字概率的比较。
5	八年级或以下：在 MicrosoftWord 中计算的阅读年级水平为 Flesch-Kincaid 年级 8.9 或以下。
6	标题：信息部分或块用标题标记。标题清楚地描述了部分，以便读者可以扫描和查找信息。
7	字体类型和大小：字体类型或样式易于阅读。字体大小至少为 11-12 磅。
8	空白：使用项目符号列表或编号列表来增加页面上的空白。
9	Image：包含至少 1 张与研究主题相关的图像。不是标志。
可理解性	
10	研究目的：包括一份声明，上面写着“研究的目的是.....”研究的目的是陈述的，而不是暗示的。
11	加入研究的主要原因-好处：包括对参与者或其他人的潜在好处的描述或列表。

12	不参加研究的主要原因—风险：包括对参与者的潜在副作用或风险的描述或列表。
13	Information being collected（正在收集的信息）：描述将从参与者处收集的信息以及有关参与者的信息。
14	学习程序：描述参与者需要做什么以及需要多少时间。
15	Study is research：包括一个声明，上面写着“study is research”或“study study”，而不仅仅是同意治疗。
16	参与是自愿的：声明参与是自愿的，参与者可以选择是否参加研究。
17	成本和补偿：描述向研究参与者支付的任何财务费用（或成本）。
可操作性	
18	同意流程：描述了读者通过签署文件、口头协议、通过计算机或其他方式表示同意的过程。

统计分析

我们报告了人工生成和 LLM 生成的 ICF 关键信息部分的平均准确性和完整性评分。我们使用 Wilcoxon 秩和检验比较了人工生成和 LLM 生成的 ICF 关键信息部分的平均准确性和完整性分数。此外，我们使用 Wilcoxon 秩和检验比较了 2 组之间的 RUAKI 指标。此外，我们测量了类内相关系数（ICC）以评估评分者之间的一致性。ICC 低于 0.5 表示低可靠性，0.5 到 0.74 之间表示中等可靠性，0.75 到 0.9 表示可靠性好，高于 0.9 表示可靠性高。

伦理考量

根据适用的机构和监管指南，这项研究有资格成为非人类受试者研究，因为它只涉及同时也是这项工作合著者的评估者。不涉及外部参与者，也没有收集、分析或共享可识别的私人信息。因此，这项工作不需要 IRB 的审查或批准。

结果

LLM 和人工生成的输出的准确性和完整性在 ICF 的关键部分之间具有可比性（表 3).LLM 和人类产出在传达研究目的方面都取得了相似的分数（2.88vs2.63），没有显著差异（*P*=.16）。对于持续时间和程序，分数也很接近（2.5vs2.38），没有统计学意义差异(*P*=.56)。在解释风险和不适方面,LLM 略优于人类输出(2.63vs2.38)，但同样，这种差异没有统计学意义（*P*=.32）。在收益方面，LLM 获得了 3.0 分的满分，而人类产出的 2.57 分，尽管这种差异接近但没有达到统计学意义（*P*=.10）。在讨论替代方案时，LLM 和人类产出相同，得分为 2.75（*P*≥.99）。对于总体印象，LLM 得分为 2.63，而人工产出得分为 2.31，没有统计学意义差异（*P*=.32）。总体而言，这两个输出在这些关键部分都显示出相当的性能水平。

表 3. 关键知情同意的人和大语言模型评估的平均准确性和完整性分数

类型	LLM 输出，平均分数（SD）	人力输出，平均评分（SD）	Wilcoxon 秩和检验，*P*值
研究目的	2.88(0.35)	2.63(0.52)	0.16
持续时间和程序	2.5(0.53)	2.38(0.52)	0.56
风险和不适	2.63(0.52)	2.38(0.52)	0.32
好处	3(0)	2.57(0.79)	0.1
选择	2.75(0.46)	2.75(0.46)	≥.99
总体印象	2.63(0.52)	2.31(0.59)	0.32

LLM 生成的 ICF 关键信息的 Mean RUAKI 分数与人类输出的比较表明，LLM 在关键领域（图 1).尽管 LLM 和人工输出都获得了相对较高的可读性分数，其中 LLM 略微领先（76.39%对 66.67%），但这种差异接近但没有达到统计显著性（*P*=.26）。LLM 表现出明显更好的可理解性，得分为 90.63%，而人类得分为 67.19%，具有统计学意义的 *P* 值为.015。此外，LLM 始终在文件末尾包含一个可作的下一步，这是人力产出未能提供的关键要素，LLM 的完美可操作性得分为 100%，而人力产出为 0%。总体而言，LLM 的内容获得了明显更高的综合分数（84.03%对 61.82%），具有统计学意义的 *P* 值为.008，这表明 LLM 生成的文本通常更有效地生成 ICF 关键信息部分，这些部分不仅更易于阅读，而且对参与者来说也更易于理解和可作。

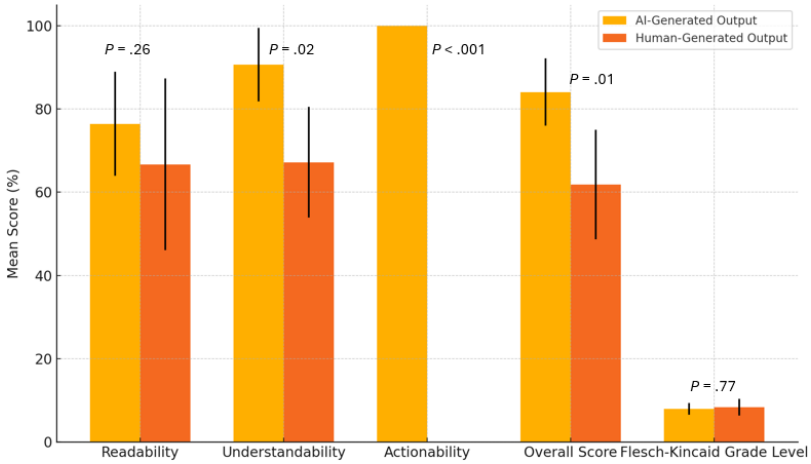

图 1. AI 和人工生成的知情同意书性能比较：关键信息的平均可读性、可理解性和可操作性分数与 CI 和 Wilcoxon 签名秩检验 P 值的可读性、可理解性、可操作性和 Flesch-Kincaid 等级水平。AI：人工智能

虽然 LLM 和人工生成的 ICF 都表现出相似的年级水平，但 LLM 生成的内容略低（7.95 对 8.375），这表明它更容易阅读，并且更符合普通读者的推荐阅读水平。然而，这种差异没有统计学意义（*P*=.77），表明两者之间的可读性相当。尽管如此，LLM 的内容仍然更接近目标可读性水平，在确保更广泛的受众可访问性方面提供了微妙的优势。

发现评分者平均评分的 ICC 评分为 0.83（95%CI0.64-1.03）。根据 ICC 值的一般解释指南，此分数表示良好的可靠性。

讨论

主要成果

本研究评估了 Mistral8x22BLLM 在临床试验中为 ICF 生成关键信息部分的性能。LLM 生成的内容与人工生成的 ICF 之间的比较表明，LLM 在提高 ICF 的效率、可读性和可操作性方面表现出相当大的潜力，同时在大多数评估类别中保持可比的准确性和完整性。

准确性和完整性

LLM 生成的 ICF 在大多数领域都实现了与人工生成内容相当的性能。LLM 和人类产出在传达研究目的方面是相似的。它们在描述持续时间和程序方面也表现同样

出色。虽然人工生成的内容在讨论替代方案方面略优于 LLM，但 LLM 在解释好处方面表现更好。总体印象得分略微偏向 LLM。这些结果表明，虽然 LLM 在大多数领域的表现与人类相似，但可能需要进一步改进提示工程以提高其在更复杂的部分（例如替代方案）的性能。通过额外的微调，LLM 可能会在所有类别中达到或超过人工生成的 ICF 的质量。

可读性

LLM 在可读性方面优于人工生成的 ICF，更高的 RUAKI 分数证明了这一点。根据 Flesch-Kincaid 年级，LLM 和人工生成的 ICF 都表现出良好的可读性，但 LLM 的平均成绩较低，这反映了卓越的可读性，并且更符合八年级阅读水平内容的机构要求。这凸显了 LLM 在解决 ICF 创建的主要挑战之一方面的优势：制作既全面又易于普通受众理解的文件。鉴于许多 ICF 经常超过推荐的阅读水平，LLM 生成可读内容的一致能力是一个显著的优势，可以在不牺牲细节的情况下确保可访问性。

可理解性和可操作性

LLM 在可理解性和可操作性方面都明显优于人工生成的 ICF，这反映在较高的 RUAKI 分数上。LLM 的成果不仅更易于理解，而且始终包含可作的后续步骤，这是人工生成内容中经常缺少的关键组成部分。LLM 生成的 ICF 的完美可操作性分数表明，这些模型可以提高参与者的理解力并促进明智的决策。这些发现证明了 LLM 创建不仅更易于阅读而且更有效地指导参与者完成同意过程的 ICF 的潜力。

评分者一致性

ICC 为 0.83 表示评分者之间的一致性非常高，反映了评估过程的可靠性。窄 CI 进一步支持这些评级的稳健性，确保不同协议结果的一致性和有效性。

经验教训

1. 精确温度设置的重要性

对于所有 LLM，模型温度参数控制响应的多样性。较高的温度（如 0.8）会产生更多样化的答案，而较低的温度（如 0.2）会产生更集中和确定性的输出。在我们的实验中，我们发现将温度设置为 0 是这项任务的最有效选择。温度引入了一定程度的随机性，这可能导致幻觉-与原材料不必要的偏差。由于我们的目标是直接从研究方案中提取信息并将其视为唯一的事实来源，因此必须尽量减少 LLM 的任何创造性输出。尽管 0.2 的温度已经相当集中，临床试验文件对绝对准确性的需求促使我们将温度设置为 0，以确保内容与提供的数据保持严格一致。

2. 使用跨模型 Few-Shot 提示解决可读性挑战

据观察，当直接提示 Flesch-Kincaid8 年级的内容时，Mistral 很难生成 8 年级的内容，例如“内容应符合识字标准，特别是 8 年级或更低的阅读水平”。为了克服这一挑战，与本项目中使用的其他零镜头提示不同，额外的提示和一种称为“小镜头训练”的技术被引入。这涉及为模型提供 Flesch-Kincaid8 级以下和 9 级以上的文本示例，帮助指导模型制作所需阅读水平的内容。这些使用 ChatGPT4 生成的示例被合并（如 Supplementary3 中的多媒体附录 1）使 Mistral 能够在所需的年级 8 或更低年级更一致地制作内容。这种方法称为跨模型小镜头提示，涉及使用一个模型生成示例（或“镜头”），这些示例随后被馈送到另一个模型中，以提高其在特定任务上的性能。在使用任何内容编辑提示后，必须应用此步骤，因为进一步的内容编辑可能会无意中将阅读级别提高到目标以上。格式编辑应该是该过程的最后一步，因为任何后续提示都可能更改格式。

3. 从最少到最及时的工程

有效的提示工程包括将任务分解为可管理的步骤。当 LLM 在单个提示中被给予多个指令时，它们通常难以准确地遵循所有指示。通过采用 Least-to-Most 方法（其中每个提示都包含一组重点说明，并基于先前的输出构建），我们获得了更加一致和可靠的结果。这类似于在电子健康记录（EHR）系统中迭代构建临床工作流程以确保决策的准确性。与构建临床模板或订单集非常相似，提示工程可确保内容生成的每个阶段都得到指导，以避免歧义，确保准确性和与知情同意上下文的相关性。例

如，如果我们在 EHR 中构建药物警报，将警报逻辑分解为单独的步骤（从检查过敏到建议替代方案）可确保清晰性并避免让用户不知所措。同样，分解生成 ICF 部分的提示有助于 LLM 专注于从实验步骤中检索正确的信息。这种方法包括 2 个阶段：首先，将一个复杂的问题分解为一系列更简单的子问题，然后依次解决这些子问题，每个解决方案都由前一个解决方案的答案提供信息。通过指导 LLM 逐步工作，我们不仅提高了其准确性，还确保了人工监督仍然是流程中不可或缺的一部分，从而带来最佳结果和对最终产品的更大控制。例如，一位审稿人指出，LLM 生成的程序缺少一些程序信息。这个问题可能源于原始研究方案中冗长且结构不佳的程序部分。为了解决这个问题，我们设计了一个工作流程，首先要求 LLM 总结研究的程序和时间表。然后，我们指示 LLM 从这个总结输出中提取必要的信息。这种循序渐进的 Least-to-Most 方法使我们能够成功提取缺失的信息并将其集成到关键信息中。

4. 实际的应用程序和集成挑战

在临床工作流程中实施 LLM 不仅需要提高可读性或准确性，还需要与现有的临床系统和流程（例如 EHR 和 IRB 工作流程）无缝集成。为了使 LLM 产生实际影响，模型需要适应不同的临床环境并满足监管和伦理标准。确保这一点的一种方法是开发接口，允许研究人员微调 LLM 输出，同时确保符合临床试验指南。

5. 需要详细的原材料和人工监督

在一项研究方案中，一位审稿人发现，与 LLM 生成的版本相比，人工生成的 ICF 在详细说明风险和不适方面更准确。这种差异在 LLM 生成的 ICF 中遗漏了与 COVID-19 测试相关的常见不适中很明显，这是因为此信息未包含在原始研究方案中。这凸显了一个关键的教训：“垃圾进，垃圾出”。为了使 LLM 产生全面而准确的 ICF，原始研究方案必须彻底和详细。此外，这一发现强调了让人在回路中审查和改进 LLM 输出的重要性。虽然 LLM 可以显著减少创建 ICF 所需的工作量，从而可能节省高达 90%的工作量，但最终产品仍然受益于人工监督。例如，在使用 LLM 生成结构良好的 ICF 关键信息部分后，研究人员可以轻松调整内容以更好地适应他们的特定受众。一位评论家指出，LLM 生成的 ICF 比人工生成的版本更具技术和临床基调。通过让研

究人员定制 LLM 生成的内容，ICF 可以针对研究受众进行定制，而大部分繁重的工作已由 LLM 完成。

6. LLM 生成内容的伦理和监管考虑

随着 LLM 在生成面向参与者的文件方面的作用越来越大，必须解决伦理和监管问题。关键考虑因素包括确保 AI 生成的 ICF 不会无意中引入偏见或错误信息。此外，随着 LLM 在临床环境中承担更多责任，监管机构可能需要制定指南来管理其使用。这些指南可能包括关于人工监督必要性的规定，以验证 LLM 生成的内容是否准确、对参与者友好以及符合知情同意的伦理标准。

7. 结构化和准确信息提取的技术

通过使用有效的提示工程策略，并制作精确专注于从研究方案中提取信息的提示，我们能够生成结构良好且格式整齐的输出。关键组成部分（如引言、研究目的、程序、风险和不适、益处、替代方案、成本和补偿以及同意过程）用标题清晰地描绘出来，这改善了文档的组织性，使查找特定信息变得更加容易。在制作初始提示以提取关键信息时，我们结合使用了多种技术：

1. 分隔符用法：###和[]等分隔符用于明确定义文本不同部分之间的边界。
2. 角色扮演：为 LLM 分配一个特定的角色，例如“作为临床试验知情同意书作者”，提供了上下文指导，通过使模型的回答更加相关和集中来提高性能。

8. 解决实际收益和成本节约问题

除了提高 ICF 的质量外，LLM 还有可能降低临床试验的运营成本和管理负担。自动创建 ICF 可以显著减少文档准备所花费的时间，同时保持对监管标准的合规性。通过量化这些节省——例如估计创建 ICF 所花费的时间减少——未来的研究可以进一步证明将 LLM 纳入临床工作流程的切实好处。

9. 预测未来的发展

截至本文发布时，GPT-4o 迷你、大 Mistral2 和 MetaLlama3.1 已发布，每个都具有 128k 令牌的扩展上下文窗口，使其成为这些任务的理想选择。但是，它们在开发阶段不可用。虽然 64k 令牌足以处理大多数临床研究方案，但对于更广泛的内容，这些新模型将是可取的。也就是说，提示与所有这些模型兼容。

对实践的启示

使用 LLM 生成 ICF 为简化知情同意过程提供了相当大的潜力。通过制作更具可读性、可理解性和可操作性的内容，LLM 可以提高参与者的理解和参与度，从而有可能改善临床试验中的招募和保留。此外，与自动生成 ICF 相关的时间节省可以减少研究人员的工作量，同时确保 ICF 在可读性和内容清晰度方面与监管标准保持一致。

限制和未来方向

必须根据某些局限性来解释这项研究的结果。LLM 的性能与输入研究方案的质量密切相关。LLM 的性能与原材料的清晰度和完整性密切相关。研究方案中的歧义或不一致可能会阻碍模型捕获所有相关细节以生成准确和全面的 ICF 的能力。未来的研究应侧重于提高原材料的清晰度并改进提示工程策略以优化 LLM 性能，尤其是在更复杂的部分，例如研究程序。

为了通过程序细节来应对这些挑战，我们使用了有针对性的提示工程方法，其中包括让 LLM 首先总结研究的程序和时间表，然后从总结的文本中提取具体细节。这种方法提高了准确性，但需要不断改进这些策略，以增强 LLM 更有效地处理复杂和冗长部分的能力。

另一个限制与 LLM 生成的文本的潜在可识别性有关。尽管评估者对 ICF 的来源不知情，并以随机顺序呈现人工和 LLM 生成的文档，但 LLM 输出的独特文本风格——以主动语态、组织良好的结构、简化的语言和始终如一地遵守可读性指南为特征——可能无意中揭示了它们的来源。这种可识别性可能会在评估过程中引入潜意识偏见。为了在未来的研究中解决这个问题，我们计划使用文本混淆技术，例如释义或重新格式化输出，以最大限度地减少风格差异并确保真正的盲法。这种方法将有助于加强未来比较的有效性。

本研究的另一个重要局限性是样本量小，仅由 4 个临床试验方案组成。虽然这些方案提供了一个有用的测试平台，但相对较小的样本量限制了研究结果的普遍性。

未来的研究应纳入来自不同治疗领域、阶段和复杂程度的更广泛的临床试验，以充分验证模型的性能。此外，有限的样本量可能导致一些统计学上不显著的结果，例如与程序细节或研究替代方案相关的结果。更大的样本将提供更大的统计功效，从而能够检测人类和 LLM 生成的 ICF 之间较小但实际上显著的差异。

更大的样本也将更好地捕捉创建 ICF 所涉及的各种挑战，例如监管要求、医疗程序复杂性和对弱势群体的考虑的变化。未来的研究应旨在评估 LLM 在更广泛的临床试验环境中的稳健性和适应性。

设计和完善提示以及生成 AI 生成的 ICF 的过程在初始开发期间需要投入适度的时间。但是，利用本研究中现有的提示和经验教训将使未来的用户能够更高效、更低成本地完成该过程，从而增强这种方法的可扩展性。

人为生成的 ICF 的 0%可操作性评分反映了结构性问题，而不是方法学缺陷。大多数机构不会在其模板的关键信息部分包含明确的可作部分。更新这些模板以包含可作的说明可能会显著提高分数。这凸显了 LLM 生成的 ICF 的优势，它本身就包含可作的元素，提高了同意书的清晰度和使用。

虽然 LLM 表现出出色的表现，但在某些情况下它遗漏了与研究的持续时间和程序要素相关的细节。这可能源于 2 个主要挑战：研究方案中这些部分的模棱两可或不一致的表述以及文本的冗长，这可能会阻碍 LLM 有效处理细节的能力。我们总结然后提取程序细节的针对性方法有助于解决这个问题，但需要进一步改进以确保 LLM 能够始终如一地处理此类挑战。

展望未来，我们的下一个目标是直接从研究方案中自动创建整个 ICF。这将显著减少 ICF 开发所需的时间和精力，同时保持一致性和质量。虽然这项研究强调了人工监督解决 LLM 生成内容中问题的潜力，但我们未来的研究将旨在量化修订所需的时间和精力，以更好地评估将这些模型集成到临床工作流程中的实际效率收益。我们还计划探索 LLM 的多语言功能，以生成多种语言的 ICF，扩大招募基础并促进临床试验的多样性和公平性。确保非英语参与者收到与英语参与者一样可读和可理解的 ICF，对于提高研究的包容性和代表性至关重要。未来的工作还应优先考虑彻底审查伦理问题，包括 AI 生成内容中的潜在偏见、AI 决策过程透明度的必要性以及在临床试验工作流程中部署 LLM 的法律影响，以确保这些工具得到负责任和公平的实施。

结论

本研究强调了 LLM 在提高临床试验中 ICF 生成的效率和质量的潜力。虽然为了确保复杂部分的准确性，人工监督仍然是必要的，并且研究结果受到小型数据集和单个 LLM 模型评估的限制，但 LLM 在生成更具可读性、可理解性和可操作性的 ICF 内容方面表现出潜在优势。随着 LLM 技术的不断发展，它有望通过促进创建对参与者友好且符合监管标准的 ICF 来进一步加强知情同意流程，从而改善临床研究中的伦理行为。