评估　　ChatGPT 模型系列在生物医学推理与分类中的应用　　n

*Shan Chen,[1,2] *Yingya Li,[3] Sheng Lu,[4] Hoang Van,[3] Hugo JWL Aerts,[1,2,5]Guergana K. Savova,[3] Danielle S. Bitterman,[1,2]1.医学人工智能项目（AIM），Mass General Brigham，哈佛医学院，波士顿，马萨诸塞州2.放射肿瘤科，布里格姆妇女医院／丹娜-法伯癌症研究所，波士顿，马萨诸塞州3.计算健康信息学项目，波士顿儿童医院及哈佛医学院

，

波士顿，马萨诸塞州4.泛在知识处理实验室，达姆施塔特工业大学，达姆施塔特，德国 5. 放射学与核医学，GROW & CARIM，马斯特里赫特大学，荷兰 * 表示共同第一作者

摘要

研究目的：近年来，大语言模型（LLM）取得了显著进展，在生物医学问答方面展现出出色的能力，但其在更具体的生物医学应用中的表现尚未得到充分研究。本研究考察了 ChatGPT 模型系列（GPT-3.5、GPT-4）等大语言模型在问答之外的生物医学任务中的性能。

材料/方法： 由于患者数据无法传递至 OpenAI API 公共接口，我们使用 10,000+ 个样本作为代理，对模型在临床领域两项基本任务——分类和推理——上的性能进行评估。第一项任务是对科学文献中的临床和政策建议陈述进行分类，判断其是否构成健康建议。第二项任务是检测生物医学文献中的因果关系。我们使用 20% 的数据集，在有无思维链（CoT）的零样本和少样本设置下进行提示开发。每种设置中效果最佳的提示均在剩余的 80% 数据上进行评估。我们将大语言模型与采用简单特征（词袋模型（BoW））和逻辑回归的模型，以及经过微调的 BioBERT 模型进行了比较。

结果：微调 BioBERT 在分类（F1 0.800-0.902）和推理（F1 0.851）上取得了最佳结果。在大语言模型方法中，少样本思维链在分类（F1 0.671-0.770）和推理（F1 0.682）上取得了最佳结果，与词袋模型相当（分类和推理的 F1 分别为 0.602-0.753 和 0.675）。达到最佳大语言模型结果所需的总计时间为 78 小时，而开发表现最佳的 BioBERT 和词袋模型分别仅需 0.078 和 0.008 小时。

结论： 尽管围绕走红的 ChatGPT 令人兴奋，我们发现，针对两个基础自然语言处理任务进行微调仍然是最佳策略。简单的词袋模型表现得与最复杂的大语言模型提示相当。提示工程需要大量投入。

Funding: R01GM114355（YL）、R01LM013486（HV、GS）、伍兹基金会（SC、DB）

主论文

1.0 引言与背景

机器学习（ML）方法在自然语言处理（NLP）方面的进步，例如Transformer[1] 和强化学习[2], 结合丰富的数字文本和扩大规模的硬件能力，导致了许多预训练大语言模型（LLM）——也称为基础模型。将其中一些大语言模型与智能工程相结合，给世界带来了病毒式传播的ChatGPT，这反过来普及了这项技术并重新激发了人工智能（AI）/通用人工智能（AGI）的辩论。尽管大多数大语言模型被训练为聊天机器人，但主流媒体中的一些声称甚至说这些大语言模型是有感知的，甚至能够解决以前需要高水平人类专业知识和专门训练的任务。另一方面，描述大语言模型的科学论文则更为审慎[3] 概述了局限性："...除了故意滥用之外，还有许多领域，大语言模型应仅在极其谨慎的情况下部署，或者根本不部署。例子包括高风险领域，如医学诊断、基于受保护特征对人进行分类、确定信贷、就业或住房的资格、生成政治广告以及执法。"[3] 因此，科学界有责任理解大语言模型的优势和劣势，如何管理其局限性和风险，以及对我们未来的影响。[4,5]

医学是大语言模型风险最高的领域之一。围绕大语言模型的热潮已渗透到生物医学与临床界，推动了各种早期用例评估。两项研究[6,7] 评估了 ChatGPT 在美国医师执照考试（USMLE）测试中的表现，表明其达到及格分数。Zuccon 和 Koopman[8]研究了提示词对 ChatGPT 回答复杂健康信息问题的影响。Chen 等人[9]评估了 ChatGPT 在提供符合美国国家综合癌症网络（NCCN）指南的癌症治疗建议方面的性能和稳健性。Lyu 等人[10] 研究了使用 ChatGPT 将放射学报告翻译成患者和医疗保健提供者通俗语言的可行性。Google Research 和 DeepMind 的一篇论文[11]介绍了使用 Google 的 PaLM 系列大语言模型的实验，表明大语言模型在医学中的潜在效用，但也揭示了重要的局限性，强调了评估框架和方法开发的重要性。

与此同时，大语言模型的实际应用受到其庞大规模和高计算需求的限制，使大多数医疗机构和研究人员难以使用。因此，在大语言模型时代，研究人员一直在探讨更广泛的问题，例如专科临床模型的效用，尤其是那些规模更小、计算成本更低的模型。Lehman 等人[12] 表明，相对较小的专科临床模型，其性能显著优于更大的大语言模型，即使后者仅使用有限的标注数据进行微调。此外，他们还表明，在临床数据集上进行预训练，可以构建规模更小、参数效率更高的模型，其性能能够达到甚至超过那些规模大得多且计算需求极高的大语言模型。Wang 等人[13] 专注于

探索 ChatGPT 的稳健性，其中一个医疗诊断数据集代表了域外分布。结果与 Lehman 等人[12]的研究一致

我们的目标是促进对生物医学领域大语言模型的日益深入理解，并着重关注实际终端应用。对电子病历（EMR）中的临床叙述开展自然语言处理研究，可直接应用于转化科学[14], 临床决策支持、[15]以及医疗管理[16] ，此外还可用于直接的患者护理。支持这些应用的两项基础自然语言处理任务是分类（例如患者表型分析）和推理（例如药物不良事件分析）。因此，我们旨在评估最先进的（SOTA）大语言模型在需要理解上下文细微差别的分类和推理任务上的性能。尽管近年来已有众多大语言模型得到训练，但其中大多数属于专有模型，无法公开下载到本地（例如 GPT-3、ChatGPT），这使得人们无法在包含受保护健康信息（PHI）的临床数据集上对其进行评估，即使这些数据已经过匿名化处理。因此，我们使用生物医学代理数据开展研究。我们在典型用户所面临的限制条件下评估大语言模型，以了解其在现实世界中的实用性；具体使用 OpenAI API[17] 以及在大多数医院和临床实践的 IT 能力范围内具有计算可行性的大语言模型。

2.0 实验设置

2.1 任务和数据集

我们考察大语言模型在临床领域两项基本任务上的性能——分类和推理。[1] 具体而言，我们选择了两个针对健康建议和因果语言标注的开放数据集，以测试模型对医学文献发现及其对健康相关实践影响进行分类和推理的能力。

- **分类任务：HealthAdvice**[18] 是一个由10,000+个从医学研究文献的摘要和讨论/结论部分提取的句子的标注组成的数据集。该数据集采用多维分类法，将每个句子分类为“无建议”、“弱建议”和“强建议”，以捕捉临床和政策建议的出现及其强度。由于健康建议通常出现在摘要或讨论/结论部分，其语言风格可能因不同部分而异，标签进一步分为3个数据集：讨论中的建议、非结构化摘要中的建议和结构化摘要中的建议。

- **推理任务：因果关系**[19] 是一个多标签推理数据集，旨在识别医学研究文献发现中的相关性和因果主张。标注语料库包括从摘要中提取的超过3,000个PubMed研究结论句子。每个句子根据其确定性和推理类型被标记为“相关性”、“条件因果”、“直接因果”和“无关系”。

表1(A) 和1(B) 展示了数据集的分布情况。我们将两个数据集划分为开发集和测试集。开发集是原始数据集的 20% 的按比例抽样，而

[1] 所有代码，提示词 pts 和相关输出可以找到： https://github.com/shan23chen/HealthLLM_Eval

数据集剩余的 80% 用作测试集。最终评估在测试集上进行。

Task	Label	Advice in discussion sections	Advice in unstructured abstracts	Advice in structured abstracts
Classification: HealthAdvice	弱建议	162	28	1482
	强建议	135	16	925
	无建议	3635	890	3575
	总计	3932	934	5982

表 ***1(A)***。 *HealthAdvice* 数据集分布

Task	Label	Count
Reasoning: CausalRelation	Correlational	998
	Conditional causal	213
	因果	494
	无关系	1356
	总计	3061

表 ***1(B)***。*CausalRelation* 数据集分布

2.1 基线模型

我们将大语言模型的性能与经典机器学习方法和基于 Transformer 的预训练语言模型进行比较。对于经典机器学习方法，我们训练采用随机梯度下降（SGD）拟合的逻辑回归模型，并使用词袋模型（BoW）表示，以词频-逆文档频率作为向量化方法。鉴于基于 BERT 的预训练语言模型[20], 尤其是 BioBERT[21], 已在上述两项任务中展现出有效性，我们对 BioBERT 模型进行微调（超参数设置见附录表 A1）。基线模型在完整开发集上训练和微调，并在测试集上进行测试。为进一步考察数据量对模型性能的影响，我们使用开发集的 20%、50% 和 100% 训练并测试词袋模型和 BioBERT 模型。我们记录开发和评估词袋模型与 BioBERT 模型所需的时间。

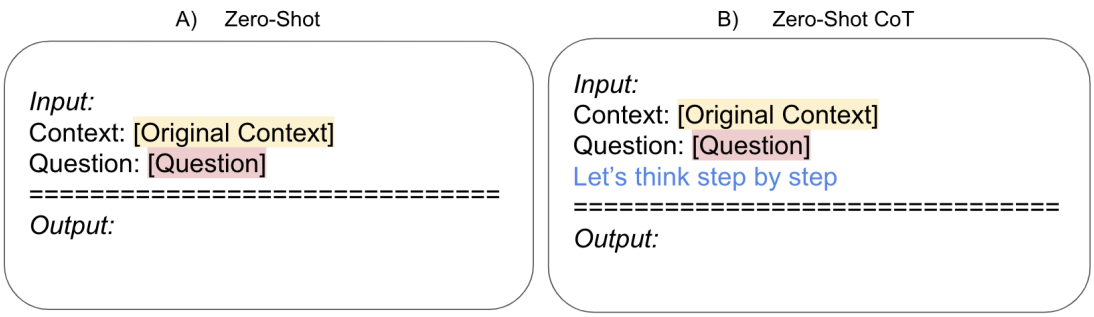
2.2 ChatGPT 模型系列

我们在这两项任务上评估 GPT-4 及其前代模型，包括 GPT-3.5-Turbo(20B)和 GPT-Davinci-003(175B))。对于GPT-3.5模型，我们考察了有思维链（CoT）和无思维链的零样本、单样本和少样本提示。鉴于计算成本限制，对于 GPT-4，我们考察了零样本（模仿普通用户的最简单提示策略）以及带 CoT的少样本（最复杂的提示策略）。CoT技术会将中间推理步骤明确列出，作为提供给大语言模型的提示词，以诱发多步推理行为。 [22]

为了设计提示，我们遵循既往研究中应用的提示结构。[22] 图1展示了分类任务提示模板的一个示例。对于零样本设置，五位作者各自独立开发了两个提示词。对于无思维链的单样本和少样本设置，示例直接从开发集中选择。对于有思维链的单样本和少样本设置，同样的五位作者独立为每个数据集的示例编写了思维链提示词。为了评估模型效率，我们跟踪设计提示词所花费的总计时间。鉴于健康建议（3个类别）和因果语言（4个类别）数据集中类别数量不同，我们分别对分类任务和推理任务的少样本设置应用3样本和4样本示例。我们使用正则表达式将模型输出与数据集中的标签进行匹配，并进行验证以核实正则表达式的准确性。为了评估模型的性能，我们将预测结果与数据集中的金标准标注进行比较。提示词的性能最初在开发集上评估，并选择性能最佳的提示用于测试数据集的最终评估。还评估了来自不同类别的示例的性能。评估的完整提示词集包含在附录表B1和B2中。

我们使用 OpenAI API 运行预测，并根据测试集上的四折交叉验证所得宏平均 F1 分数的平均值来衡量模型性能。F1 分数是自然语言处理中的经典指标，表示召回率/敏感度与精确率/阳性预测值的调和平均数。宏平均 F1 分数通过计算所有各类别 F1 分数的算术平均值获得。为比较模型效率，我们记录了使用 API 进行推理所需的时间和费用（附录表A2 a-c）

[Original Context] = Georgian public health specialists working in the HIV field should prioritize implementation of such interventions among HIV patients.
[Question] = Is this a 0) no advice, 1) weak advice or 2) strong advice?
[Answer] = 2) Strong advice.
[CoT Solution] =
1) The statement is a directive, so it is a strong advice.
2) The statement is specific and clear, so it is not a weak advice.
3) Therefore, the statement is a strong advice.
Answer: 2) Strong advice.



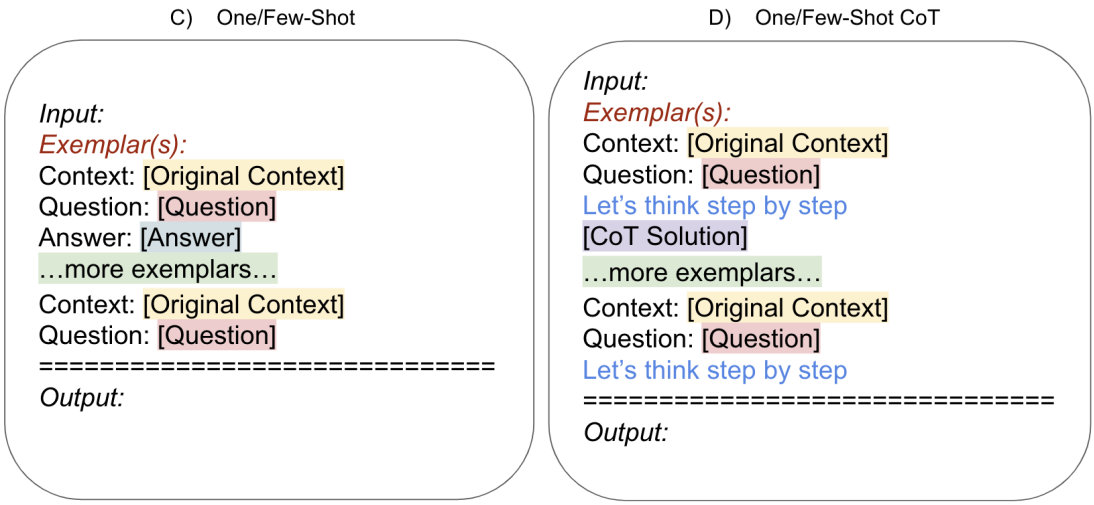


图***1.*** 用于分类和推理任务的提示模板示例。针对评估的每种提示设置都创建了提示模板：***(A)*** 零样本，***(B)***带思维链（*CoT*）的零样本，***(C)*** 单样本和少样本，以及***(D)*** 带思维链（*CoT*）的单样本和少样本。*"……更多示例……"*（以绿色突出显示）仅添加到带或不带思维链的少样本设置中。所有评估过的提示词均见附录。词袋 =词袋。

2.3 更小的大语言模型

我们采用相同的设置（零样本和少样本，分别搭配或不搭配思维链）来评估若干较小的大语言模型（参数少于 10B）在这些任务上的性能，包括 GPT-J、GPT-JT 和 Galactica[23]。GPT-J 基于 EleutherAI 的 60 亿参数模型 GPT-J-6B 构建，并经过 35 亿词元的微调。在多种零样本下游任务中，它的表现与 GPT-3 非常相似。GPT-JT 是 GPT-J-6B 的一个分支，经过 35.3 亿词元的微调；研究表明，它在某些分类任务上的表现甚至优于 GPT-3。Galactica 使用 4800 万个科学文章、网站、教科书、讲义和百科全书示例进行训练。对 GPT 系列模型采用的相同评估流程也适用于该模型。

2.4 统计分析

配对*t* 检验用于比较各任务的平均宏平均F1分数，并且双侧$p<0.05$被认为具有统计学意义。分析使用python版本3.9.7（Python软件基金会）进行。

3.0 结果

我们的研究结果总结见表2，图2、图3和图4展示了多项比较分析，以提供背景并促进对结果的解读。

模型	设置	分类			推理
		建议 discussion sections	Advice in unstructured abstracts	Advice in 结构式 摘要	因果 关系 检测
	随机	**0.209**	**0.174**	**0.313**	**0.231**
GPT-J ，GPT-JT ，Galactica	Best-6.7B 模型[2]	**0.276**	**0.284**	**0.358**	**0.244**[3]
GPT-4	零样本	0.509	0.478	0.517	0.301
	少样本思维链	**0.648**	**0.712**	**0.770**	**0.682**
GPT3.5s （Davinci-003/ ChatGPT-Turbo)	zero-shot	0.506	0.489	0.548	0.288
	one-shot	**0.513**	**0.554**	**0.593**	0.396
	few-shot	0.475	0.403	0.495	**0.542**
	零样本思维链	0.436	0.424	0.515	0.345
	单样本思维链	0.430	0.410	0.552	0.553
	少样本思维链	**0.671**	**0.670**	**0.718**	**0.649**
BoW	20% 词袋	0.551	0.566	0.701	0.465
	50% 词袋	**0.602**	0.609	0.735	0.581
	100% 词袋	0.59**3**	**0.640**	**0.753**	**0.675**
BioBERT	20% 微调 BioBERT	0.682	0.787	0.791	0.597
	50% 微调 BioBERT	0.703	0.810	0.864	0.760
	100% 微调 BioBERT	**0.800**	**0.821**	**0.902**	**0.851**

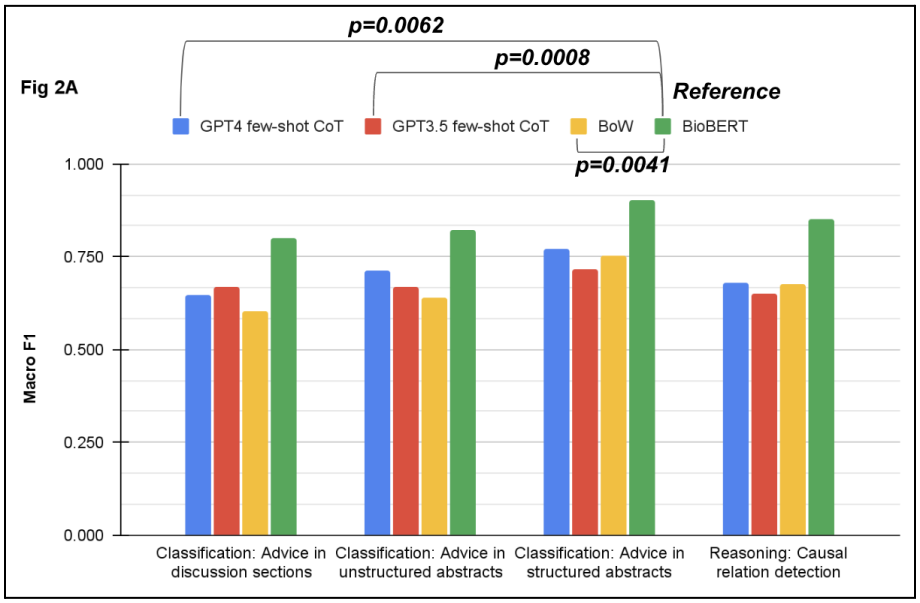
表***2***。 跨数据集和模型的最佳宏平均*F1* 结果汇总。*FT* = 微调。粗体文本表示最佳性能。对于词袋和 *BioBERT*，百分比表示在监督设置下用于训练的开发集数据量。“随机”反映了类别标签之间的均匀分布。完整结果见附录表 *A3-7*

3.1 微调和提示开发对模型性能、运行时间和时间投入的影响

[2] 由于参数规模低于100 亿的语言模型在开发阶段的不同设置下增益不显著，因此未呈现其综合结果。[3] 该结果是在零样本设置下从 Galactica 获得的；所有其他结果均来自少样本设置下的 GPT-JT。

对于推理和分类任务，微调 BioBERT 始终以相当大的优势优于最佳 GPT 设置（ΔF1 0.109 - 0.169）（图 2(A) 和表 2）。在所有数据集上平均性能，BioBERT 的宏平均 F1 分数显著优于其他模型（p<0.01 对于所有，图 2(A)。词袋、表现最佳的 GPT-3.5s 和 GPT-4 模型的平均性能之间没有统计学差异（附录表 A3，t 检验，见图 2a）。

此外，要设计出能取得最佳结果的大语言模型提示，需要投入大量时间。图 2(B) 展示了取得最佳结果所需的时间，其中考虑了在少样本设置下开发和确定提示策略所需的时间。即使在无需进行提示开发的零样本设置下，与词袋和 BioBERT 的性能相比，推理时间——这是运行时间的指标，也是计算预算的考量因素——更长，且性能更差（图 2(B)）。综上，通过经典微调方法训练任务专用神经网络不仅速度更快，性能也更好。



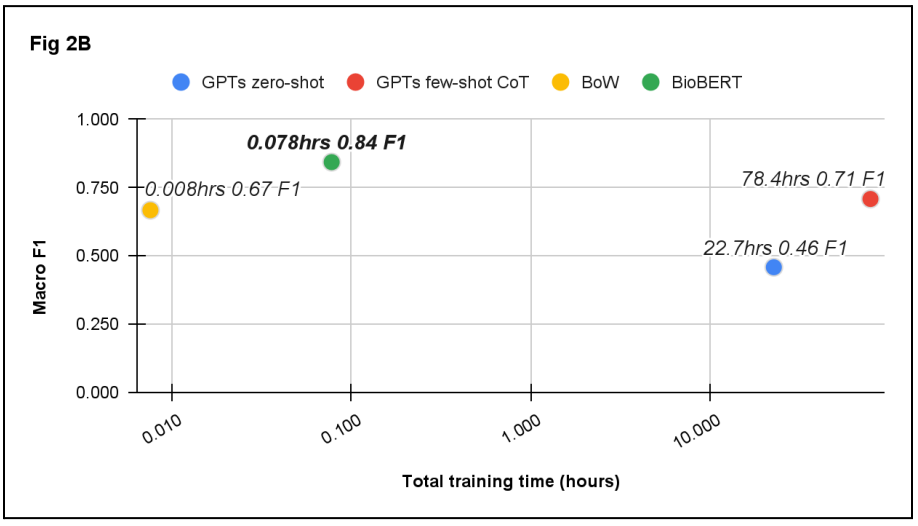


图 ***2. (A)*** 各数据集上的模型性能比较。以微调后的 *BioBERT* 为参照，其在所有数据集上的平均宏平均 *F1* 显著优于其他所有模型（所有比较的 *p<0.05*）；其他模型两两比较的平均性能之间没有统计学上的显著差异（附录表 *A3*）。***(B)*** 在所有数据集上，获得最佳结果所需的时间（小时）与平均性能的关系。词袋 = 词袋；思维链 = 思维链。

3.2 训练数据量对模型性能的影响

在本实验中，我们研究了从开发集中需要多少训练/微调数据，才能使词袋和 BioBERT 模型达到与最佳 GPT 设置（少样本思维链）相似或更好的性能。如图 3 所示，仅使用开发集的 20% 对 BioBERT 进行监督微调，在 4 个数据集中的 3 个上超过了最佳 GPT。在开发集的 50% 上进行微调，在所有数据集上都优于最佳 GPT。此外，使用开发集 100% 的简单且计算高效的词袋模型，在 4 个数据集中的 2 个上优于最佳 GPT。

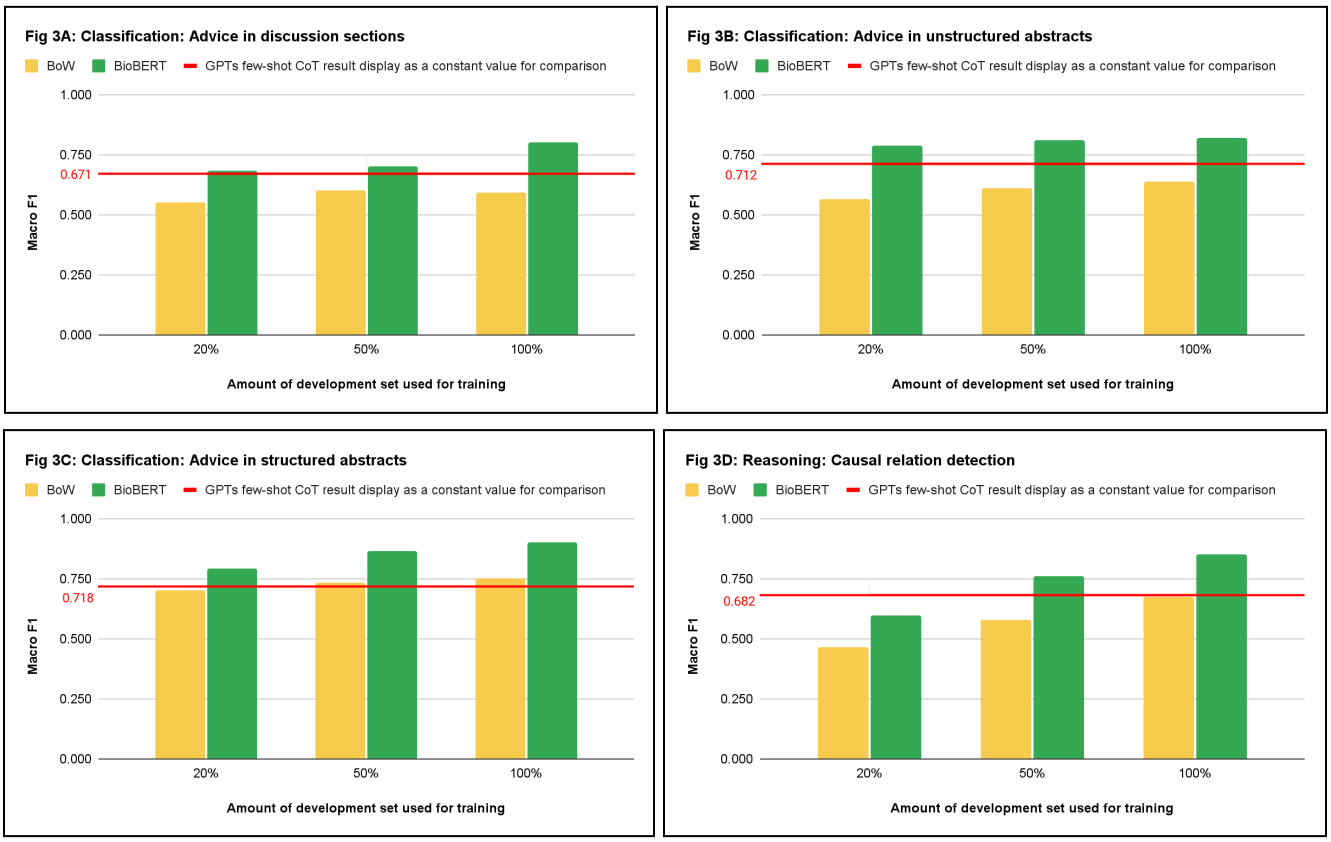


图 ***3***。 *(A-D)* 在 *4* 项任务中，对使用不同比例（*20%*、*50%*或*100%*）开发集数据微调的 *BioBERT*（绿色柱）和词袋（黄色柱）模型的性能，与 *GPTs* 的最佳设置（红线表示的少样本思维链）进行比较。***(A)*** 讨论部分中的建议比较，***(B)*** 非结构式摘要中的建议比较，***(C)*** 结构化摘要中的建议比较，***(D)*** 因果关系检测比较。词袋 =词袋；思维链 = 思维链

3.3 效果　示例提示词数量和思维链提示词数量对模型性能的　　e

我们考察了示例数量与思维链提示词之间的关系及其对 GPT 设置性能的影响。如图4(A)所示，在三个数据集中，从单样本到少样本设置时，我们观察到性能下降。因果关系检测的推理是唯一通过添加提示示例（无需思维链）而持续改进的任务。然而，如图4(B)所示，加入更多思维链示例会在各数据集上带来持续改进。这一观察突出了添加思维链示例的价值，尽管由于创建思维链提示词所需的时间精力而成本高昂。

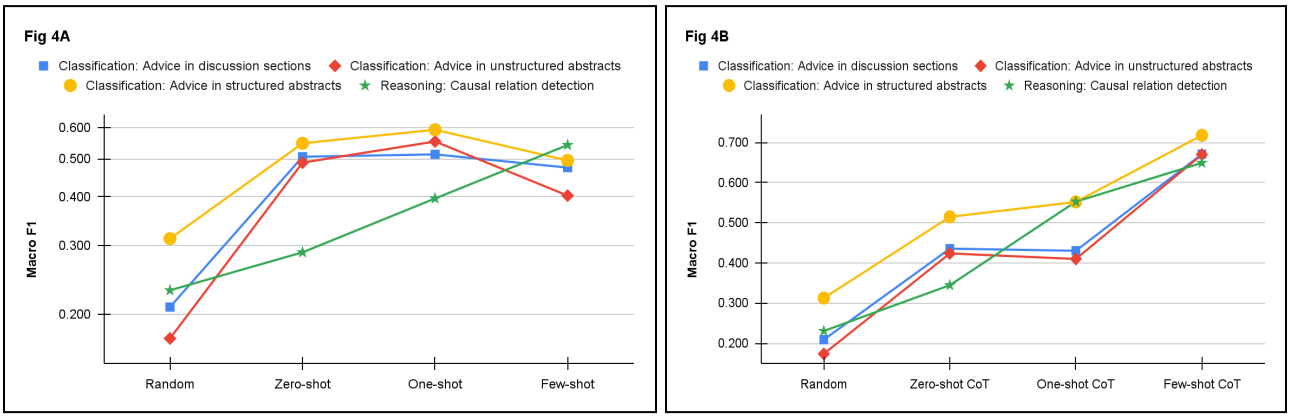


图***4***。 比较 *GPT-3.5s* 在每个数据集上随着示例数量增加，在无 （***A***） 和有 （***B***） 思维链（*CoT*）情况下的性能。对于两幅图，随机作为基线，表示类别标签上的均匀分布。这里，单样本对每个任务使用多数类示例，少样本对每个类别使用一个示例。分类数据集使用少样本 *= 3* 示例，推理数据集使用 *4* 个示例，分别对应每个任务的类别数量。

在所评估的大语言模型中，从单样本实验（包括使用和不使用思维链的实验）中还观察到，示例提示选择的不同会对性能产生影响。单样本提示中示例提示文本的变化，导致模型结果出现显著差异（附录表A6 a-c）。

4.0 错误分析

在所研究的 GPT 中，采用少样本思维链设置的 GPT-4 在 4 个数据集中的 3 个上表现最佳。因此，我们分析 GPT-4 生成的思维链，并找出常见错误模式，以更好地了解其优势和局限性。为保持一致，我们从每个数据集的测试集中随机选取 100 个采用少样本思维链设置的 GPT-4 预测错误。错误示例的选取遵循预测结果混淆矩阵中的相同错误类型比例。我们找出了两种常见错误模式。模式 A 是基于某个特定关键词产生的错误推理步骤。例如，如果文本中出现“重要性”一词，模型就会将其分类为强建议（附录表A8，第 4 行）。模式 B 是一种假阳性：在分类和推理任务中，模型分别错误地判定存在健康建议或关系，而实际上并不存在。例如，附录表 A11 中的第 3 行

展示了模式 B 错误：GPT-4 将提取实体之间的关系误分类为临床关系。附录表 A8-A11 展示了错误模式的示例。

5.0 讨论

在本研究中，我们发现，即使采用最佳的上下文学习方法，对于所有数据集而言，微调 BioBERT 的性能始终超过大语言模型的性能，宏平均F1 >0.100。事实上，仅使用开发数据集的 20% 进行微调，其性能就超过了所有分类数据集上 GPT-4 和 GPT-3.5 的最佳性能，并超过了推理数据集上 GPT-3.5 的最佳性能。令人惊讶的是，在不使用思维链的情况下，简单的词袋模型优于所有大语言模型上下文学习方法；在结构化摘要分类任务（宏平均 F1 -0.017）和推理任务（宏平均F1 -0.007）中，其表现与性能最佳的 GPT-4 方法相近。

我们的研究强调，在考虑临床自然语言处理任务的各种方法时，应将性能、时间和计算方面的权衡纳入考量。目前，我们的结果表明，总体平衡倾向于对特定任务的较小模型进行微调，这与 Lehman 等人[16]的研究一致。SOTA大语言模型（如 GPT-3.5/GPT-4）比传统语言模型大几个数量级，如果没有大量计算资源，就无法进行训练或微调。例如，GPT-3拥有1750 亿个参数，预训练需要数千个千兆亿次浮点运算秒日。[24] 另一方面，不进行微调即可直接使用的、更小且更易获取的大语言模型（<10 十亿个参数）在我们的任务中表现非常差，并且通过上下文学习也未能展现出改进，这与大语言模型涌现能力会随模型规模扩大而增强的发现一致。[25]

不过，如果大语言模型的零样本或少样本性能足够好，其计算需求在理论上可以得到抵消。我们的结果清楚地表明，零样本性能较差。尽管少样本思维链提示提升了性能，但对 BioBERT（一个拥有 1.1 亿参数的预训练语言模型）进行微调，始终能带来最佳性能。为了找出最佳提示策略而进行提示开发，本身就需要大量资源：既需要人力来设计提示词，也需要计算资源和时间进行评估。最终，每个数据集最多只需使用完整开发集的 50%，就足以微调 BioBERT 模型，使其性能超过基于完整开发集确定的最佳提示策略下的大语言模型性能。将提示开发和评估所需的资源考虑在内，获得性能最佳的大语言模型结果所需的时间，是微调我们性能最佳的 BioBERT 模型所需时间的 100 倍。

尽管性能不如经过微调的 BioBERT，上下文学习——尤其是思维链提示——显著提升了分类和推理任务的性能。此前已有研究表明，思维链能够引出推理，并提升大语言模型在通用领域的性能。[22,26] 有趣的是，尽管在思维链提示中提供更多示例始终能提升性能，但不带思维链的提示并非总是如此。对于 GPT-3.5，单样本提示在分类任务中效果最佳，而少样本提示在推理任务中效果最佳。这可能是

由于提供的噪声（包括信息量较低的提示词），并突显了基于所提供提示词的大语言模型性能的脆弱性。

这种缺乏鲁棒性还体现在提示所用示例的选择对性能有重大影响这一事实上。这些发现与Shi等人的研究一致，他们表明，对于算术任务，提示可能会提供无关上下文，从而降低性能。[27]显然，仅仅提供更多示例并不能解决大语言模型鲁棒性的挑战——这是临床领域关注和未来研究的一个主要领域，在该领域鲁棒性能至关重要。令人担忧的是，即使是对提示词的微小、看似非实质性的更改（如拼写错误）也已被证明会影响性能。[9,13] 目前已有越来越多的工作致力于开发提高鲁棒性和自洽性的策略。[13,28] 提高性能的方法在医学领域将尤为重要，因为该领域中行话、拼写错误、缩写和同义词很常见，这限制了我们开发可靠提示策略并稳健评估真实世界性能的能力。

应当指出，我们的任务间接解决了大语言模型在临床自然语言处理文本上的表现问题，并使用生物医学文本作为基本临床自然语言处理任务的代理。最先进的大语言模型如GPT-3.5/GPT-4不能与PHI一起使用，从而排除了对真实临床数据的评估。尽管如此，已知大语言模型的性能在域外任务上会下降，即包含不反映其训练内容的文本的任务，包括合成临床文本。[13] 与此同时，已证明在临床文本上训练的经典语言模型优于采用上下文学习的大语言模型。[12] 结合我们的发现，即在生物医学文本上预训练的BioBERT优于大语言模型，可以合理预期在临床数据集上的发现会类似，但一旦GPT-3.5/GPT-4能够安全且广泛地用于HIPAA保护的数据，将需要进一步评估。

大多数评估大语言模型在临床应用中表现的研究都聚焦于问答，但结果参差不齐。[6–9,11] 然而，问答并不能代表处理临床文本所需的自然语言处理任务范围，[29] 单独来看，其在临床和研究中的实际用途也十分有限。我们选择分类和推理作为任务，是因为它们对于开发能够支持临床护理和研究、且超越问答范畴的自然语言处理技术至关重要。用于分类的自然语言处理包括确定输入文本所属的类别。分类方法能够识别患者的电子病历是否包含某项目标特征或结局，这对于结局研究、临床试验匹配以及在护理点识别关键事件都具有重要意义。推理包括确定实体之间的关系，即自动识别患者病史中的不同事件如何相互关联。尤其是由于传达医学推理的细微信息往往只能通过自由文本表达，因此需要用于推理的自然语言处理方法来自动化更高层次的医学推理。在此，我们研究了因果关系；对于需要将任何临床结局与其致因因素联系起来的任务而言，这是必不可少的，例如将药物不良事件与其诱发因素联系起来。其他推理任务还包括时间推理，即确定医学事件随时间发生的先后顺序。我们选择这些任务的另一个好处是，与问答相比，它们更易于进行客观评估，从而能够更直接地评估大语言模型的表现

在高风险的临床领域。未来，将需要对临床领域中的大语言模型进行评估，以用于其他分类和推理任务，以及其他常见的自然语言处理任务，例如关系抽取、命名实体识别、共指消解、词义消歧和机器翻译。

6.0 结论

本研究表明，针对特定任务进行微调的经典自然语言处理模型仍将发挥持续作用，同时也为如何优化大语言模型以适用于生物医学领域提供了指导。即使经过广泛的提示开发，微调BioBERT（一种规模小得多的预训练语言模型）在生物医学分类和推理任务上的表现仍优于最先进的超大规模大语言模型。与零样本提示以及不使用思维链的提示相比，采用多个示例进行思维链提示能够提升大语言模型的性能。然而，开发思维链提示既耗时又耗费数据，而BioBERT 在这两方面都更加高效。此外，大语言模型对提示策略和提示词的选择非常敏感，这引发了人们对于开发可靠且安全的医学大语言模型方法之潜力的担忧。我们的工作深入揭示了这些快速兴起的方法在生物医学文本处理方面的潜力与缺陷。未来研究可以着重开发更高效的提示策略和大语言模型微调技术，以应用于生物医学领域，同时确保其可靠性与安全性；还可以探索结合经典自然语言处理模型与大语言模型优势的混合方法，进一步提升生物医学文本处理任务的性能。

References

1 Vaswani A, Shazeer N, Parmar N, 等。 注意力就是你所需要的一切。载于：NeurIPS 会议录。2017年。https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a 845aa-Paper.pdf（于2023年4月4日访问）。2 Sutton RS, Barto AG. 强化学习：导论，第 2 版。剑桥，马萨诸塞州：麻省理工学院出版社，2018年。 3 Ouyang L, Wu J, Jiang X, 等。 通过人类反馈训练语言模型以遵循指令。载于：NeurIPS Proceesings。2022年。https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805 a001731-Abstract-Conference.html。 4 Lee P, Bubeck S, Petro J. GPT-4作为医学AI聊天机器人的益处、局限性和风险。*NEngl J Med* 2023年； **388**：1233–9。 5 Reardon S. AI聊天机器人可以在家诊断医疗状况。它们有多好？*Scientific American* 萨拉·里尔登于2023年3月31日。https://www.scientificamerican.com/article/ai-chatbots-can-diagnose-medical-conditions-at-home-how-good-are-they/（于2023年4月1日访问）。 6 Gilson A, Safranek CW, Huang T, 等。ChatGPT在美国医师执照考试中的表现如何？大语言模型对医学的影响

arXiv [cs.CL]。2022；在线发表于7月17日。http://arxiv.org/abs/2207.08143。 8 Zuccon G、Koopman B。ChatGPT 医生，告诉我想听的答案：提示知识如何影响健康答案的正确性。arXiv [cs.CL]。2023；在线发表于2月23日。http://arxiv.org/abs/2302.13793。 9 Chen S、Kann BH、Foote MB、等。ChatGPT 在癌症治疗信息方面的实用性。medRxiv。2023；：2023.03.16.23287316。 10 Lyu Q、Tan J、Zapadka ME、等。将放射学报告翻译成通俗语言，使用

g
.

ChatGPT 和 GPT-4 结合提示学习：有希望的结果、局限性与潜力
arXiv [cs.CL]。2023年；在线发表于3月16日。http://arxiv.org/abs/2303.09038。11 Singhal K、AziziS、Tu T，等。 大语言模型编码临床知识。arXiv[cs.CL]。2022年；在线发表于12月26日。http://arxiv.org/abs/2212.13138。12Lehman E、Hernandez E、MahajanD，等。我们是否仍需要临床语言模型？arXiv [cs.CL]。2023年；在线发表于2月16日。http://arxiv.org/abs/2302.08091。13 Wang J、Hu X、Hou W，等。 论ChatGPT的稳健性：对抗性与分布外视角。arXiv [cs.AI]。2023年；在线发表于2月22日。http://arxiv.org/abs/2302.12095。14 Liao KP、Cai T、Savova GK，等。 使用电子病历并纳入自然语言处理开发表型算法。*BMJ* 2015年；**350**：h1885。15 ZhangY、Liu M、Zhang L，等。基于自然语言处理生成的胸部X线片描述与由放射科医师完成的描述之比较。*JAMA NetwOpen* 2023年；**6**：e2255113。16 Medori J、Fairon C。用于半自动ICD-9-C的机器学习与特征选择。

M

encoding. In: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents. Los Angeles, California, USA: 计算语言学协会, 2010: 84–9. 17 OpenAI API. http://platform.openai.com(accessed April 3, 2023). 18 Li Y, Wang J, Yu B. 检测医学研究文献中的健康建议. In: Proceedings of the 2021 Conference on Empirical Methods in 自然语言处理. Online and Punta Cana, Do minican Republic: 计算语言学协会, 2021: 6018–29. 19 Yu B, Li Y, Wang J. 检测科学发现中的因果语言使用. In: Proceeding

s
th

2019年自然语言处理实证方法会议及第9
自然语言处理国际联合会议（EMNLP-IJCNLP）。中国香港：计算语言学协会，2019年：4664–74。20 J. Devlin、M.-W. Chang、K. Lee、K. Toutanova。BERT：用于语言理解的深度双向变换器预训练。arXiv [cs.CL]。2018年；在线发表于10月11日。http://arxiv.org/abs/1810.04805。

21 J. Lee、W. Yoon、KimS，等。BioBERT：一种用于生物医学文本挖掘的预训练生物医学语言表示模型。生物信息学2020年；**36**：1234–40。22 J. Wei、X. Wang、D. Schuurmans，等。思维链提示激发大语言模型的推理能力。arXiv [cs.CL]。2022年；在线发表于1月28日。http://arxiv.org/abs/2201.11903。23 R. Taylor、Kardas M、Cucurull G，等。Galactica：面向科学的大语言模型。arXiv [cs.CL]。2022年；在线发表于11月16日。http://arxiv.org/abs/2211.09085。24 Brown TB、Mann B、Ryder N，等。语言模型是少样本学习者。arXiv[cs.CL]。2020年；在线发表于5月28日。http://arxiv.org/abs/2005.14165。25 J. Wei、Tay Y、Bommasani R，等。大语言模型的涌现能力。arXiv[cs.CL]。2022年；在线发表于6月15日。http://arxiv.org/abs/2206.07682。26 Kojima T、Gu SS、Reid M、Matsuo Y、Iwasawa Y。大语言模型是零样本推理器。arXiv [cs.CL]。2022年；在线发表于5月24日。http://arxiv.org/abs/2205.11916。27 Shi F、Chen X、Misra K，等。无关上下文很容易分散大语言模型的注意力。arXiv [cs.CL]。2023年；在线发表于1月31日。http://arxiv.org/abs/2302.00093。28 Wang X、Wei J、Schuurmans D，等。自洽性提升语言模型中的思维链推理。arXiv [cs.CL]。2022年；在线发表于3月21日。http://arxiv.org/abs/2203.11171。29 Savova GK、Danciu I、Alamudun F，等。使用自然语言处理从电子病历中提取临床癌症表型。癌症研究2019年；**79**：5463–70。

附录 A:

	Classification: Advice in discussion sections	Classification: Advice in unstructured abstracts	Classification: Advice in structured abstracts	Reasoning: Causal relation detection
random seed	42	42	42	42
batch size （每_设备_训练）	128	128	128	128
batch size (每_设备_评估)	512	512	512	512
epoch	50	50	50	200
learning rate	2e-5	2e-5	2e-5	2e-5
最大序列长度	512	512	512	512
预热_步	500	500	500	500
weight_decay	0.01	0.01	0.01	0.01

表 ***A1***。微调 *BioBERT* 的参数和超参数

表 A2 a-c 对比了使用不同模型和设置进行训练和推理所需的预计时间。

根据OpenAI 二月至三月的定价，OpenAI API 的推理总计支出为 1,299 美元，2023[4]

Model & Setting	Prompt creation time (in seconds)	Inference time for 在开发集上训练（在 seconds)	Inference time 在测试集上 （单位：秒）	Total （小时）	性能
GPT 零样本	2400.000	28296.000	11322.000	11.005	0.458
GPT 少样本思维链	18000.000	94320.000	37740.000	36.683	0.709
词袋分类器	0s	13.000	0.700	0.004	0.667
BioBERT 微调	0s	139.427	0.800	0.039	0.844
	创建零样本提示大约花费了240秒。 而编写少样本思维链提示大约花费了 1800 秒。 总计有 10 条提示词用于测试。				
	0 样本每次查询的推理时间为 1.2 秒，四样本思维链每次查询为 4 秒。				
	性能通过给定设置下四个任务的平均宏平均 F1 计算得出。				

表 ***A2a.*** 每种方法训练和推理所需的时间

[4] https://openai.com/pricing

模型 & 设置	Oracle 提示词编写者	假设有一位 Oracle 提示编写者，只创建一个提示， 以在测试集上取得最佳性能。因此，不存在多个提示 验证集所需。			
	Prompt creation time （以秒为单位）	用于 在开发集上训练（在 秒)	推理时间 在测试集上 （以秒为单位）	总计 （以小时为单位）	Performance
GPT 零样本	120.000	0.000	11322.000	3.145	0.458
GPT 少样本思维链	900.000	0.000	37740.000	10.483	0.709
词袋分类器	0.000	13.000	0.700	0.004	0.667
BioBERT 微调	0.000	139.427	0.800	0.039	0.844
	请注意：我们假设这样的 Oracle 提示编写者还会创建多出 100% 的提示词 比我们更高效。				

表 ***A2b.*** *Oracle* 提示词编写者训练和推理所需时间

Model & Setting	Oracle 提示器	批处理模式：512			
	提示 创建时间 （秒）	推理时间，用于 在开发集上训练（ seconds)	Inference time on test set (in seconds)	Total (in hours)	Performance
GPT 零样本	120.000	0.000	22.113	0.006	0.458
GPT 少样本思维链	900.000	0.000	73.711	0.020	0.709
词袋分类器	0.000	13.000	0.700	0.004	0.667
BioBERT 微调	0.000	139.427	0.800	0.039	0.844

表 ***A2c***。 能够以 *512* 为批次并行进行推理的机器上 *Oracleprompter* 的训练和推理所需时间

	GPT-3.5s	BoW	BioBERT
GPT4	0.4150	0.4228	0.0062**
GPT-3.5s	–	0.7969	0.0008**
BoW	–	–	0.0041**

**: *p<0.05*

表 ***A3.*** 模型显著性比较的*P*值（*t* 检验）

	Classification: Advice in discussion sections	Classification: Advice in 非结构式 摘要	分类：建议（适用于 结构式 摘要	推理：Causal relation detection
零样本 Davinci	0.503	0.461	**0.548**	0.242
零样本 Turbo	0.506	**0.489**	0.484	0.288
零样本 GPT4	**0.509**	0.478	0.517	**0.304**
少样本思维链 Davinci	**0.671**	0.617	0.718	0.649
少样本思维链 Turbo	0.603	0.670	0.699	0.566
少样本思维链 GPT4	0.648	**0.712**	**0.770**	**0.682**

表 ***A4***。*GPT3.5-Davinci* 与 *Turbo(ChatGPT)* 及 *GPT4*在测试集上的比较结果。最佳结果以粗体显示。

	Classification: Advice in discussion sections	Classification: Advice in unstructured 摘要	分类：结构化摘要中的 结构式 摘要	推理：因果 关系 检测
零样本 Davinci	0.503	0.461	0.548	0.242
零样本 Turbo	**0.506**	**0.489**	0.484	**0.288**
单样本 Davinci	0.513	0.501	0.593	0.210
单样本 Turbo	0.513	**0.554**	0.470	**0.396**
少样本 Davinci	0.468	0.403	0.495	0.542
少样本 Turbo	**0.475**	0.243	0.436	0.513
零样本思维链 Davinci	0.436	0.383	0.515	0.332
零样本思维链 Turbo	0.418	**0.424**	0.486	**0.345**
单样本思维链 Davinci	0.418	**0.423**	0.470	0.489
单样本思维链 Turbo	0.430	0.410	0.552	0.553
少样本思维链 Davinci	0.671	0.617	0.718	0.649
少样本思维链 Turbo	**0.603**	**0.670**	0.699	**0.566**

表 ***A5.*** *GPT3.5-Davinci* 与 *Turbo(ChatGPT)* 在测试集上的比较结果。加粗表示 *Turbo* 的表现优于 *Davinci*（*24* 种设置中有 *10* 种），尤其是在非结构化摘要的分类：建议任务中表现始终更佳。

表 A6（a-c）比较了使用 GPT3.5-Davinci 时，在单样本思维链提示中改变类别选择示例后，开发集/测试集上的性能。对不同类别示例进行的系统分析凸显了思维链方法中各种示例选择的有效性。表 A6b-c 展示了使用不同示例时，开发集/测试集上的性能变化。

类别	分类： 讨论中的 建议 部分	分类： 建议在 非结构化 摘要	分类： 建议在 structured abstracts	Reasoning: Causal relation detection
无 类别/**多数**	0.421	**0.410**	**0.552**	0.526
弱/相关性	**0.430**	0.283	0.381	0.455
强/条件性 Causal	0.343	0.319	0.447	**0.553**
直接因果	–	–	–	0.386

表***A6a***。 *GPT3.5-Davinci*，单样本思维链在测试集上的结果。粗体表示最佳性能。

类别	分类： 讨论中的建议 讨论 部分	分类： 建议在 非结构化 摘要	分类： 建议在 结构式 abstracts	Reasoning: Causal relation detection
无类别/**多数**	**0.462**	**0.510**	**0.552**	**0.521**
弱/相关性	0.460	0.333	0.391	0.495
强/条件性 Causal	0.393	0.339	0.448	0.453
直接因果	–	–	–	0.456

表***A6b***。 *GPT3.5-Davinci*，单样本在开发集上的结果。粗体表示最佳性能。

Class	分类： 讨论中的 讨论 部分	分类： 讨论中的 非结构式 摘要	分类： 建议于 结构式 摘要	推理： 因果 关系 检测
无类别/**多数**	**0.411**	0.390	**0.423**	0.526
弱/相关性	0.340	0.283	0.361	0.455
强/条件性 Causal	0.343	**0.430**	0.350	**0.553**
直接因果	–	–	–	0.456

表 ***A6c****. GPT3.5-Davinci*，单样本思维链在开发集上的结果。粗体表示最佳性能。

	分类：讨论中的讨论部分	分类：讨论中的非结构式摘要	分类：建议于结构式摘要	推理 Causal relation detection
零样本 GPT-J	0.248	0.230	0.302	0.211
零样本 GPT-JT	**0.272**	**0.262**	**0.321**	0.226
零样本 Galactica	0.234	0.218	0.294	**0.252**
少样本 GPT-J	0.157	0.142	0.292	0.191
少样本 GPT-JT	**0.276**	**0.284**	**0.358**	0.216
少样本 Galactica	0.246	0.205	0.293	**0.244**

表 ***A7a.*** 较小模型（少于*10B*参数）在测试集上的性能。在开发集实验中显示的最佳两种设置的结果。最佳结果为加粗。

	Classification: Advice in discussion sections	Classification: Advice in unstructured abstracts	Classification: Advice in structured abstracts	Reasoning: 因果关系检测
零样本 GPT-J	0.247	0.246	0.296	0.210
零样本 GPT-JT	**0.285**	**0.262**	**0.331**	0.226
零样本 Galactica	0.233	0.209	0.278	**0.250**
少样本 GPT-J	0.155	0.143	0.286	0.192
少样本 GPT-JT	**0.284**	**0.282**	**0.348**	0.217
少样本 Galactica	0.257	0.222	0.308	**0.257**
少样本思维链 GPT-J	0.211	0.200	0.323	–
少样本思维链 GPT-JT	0.172	0.129	0.269	–
少样本思维链 Galactica	0.241	0.195	0.368	–

表 ***A7b***。较小模型（参数少于 *10B*）在开发集上的性能。最佳结果以粗体显示。

Sentence	GPT-4 答案	Gold	Pred
对这一假设进行检验或许代表了一个方向。 未来研究。	1. “可能代表”一词表明上下文中存在一条建议。 2. “May represent”暗示一种可能性，这不是强烈意见。 3. 答案是弱建议。	no advice	weak advice
未来研究应尝试复现这些 在一项采用更……的大规模试验中的研究结果 先进、用户友好的项目及其他 在更多多样化样本中的改善指标。	1. 术语“should attempt”表明上下文中有一条建议。 2. “应该”表示指令性意见。 3. 答案是强建议。	no advice	strong 建议
我们的结果表明，孕妇的 糖化血红蛋白为5.4% (36 mmol/mol) 应进行 口服葡萄糖耐量试验。	1. 术语“suggest”表明上下文中存在建议。 2. “Suggest”表明一种指令性意见，但不如推荐那么强烈。 3. 答案是弱建议。	强 建议	weak 建议
这凸显了考虑 下呼吸道感染患者亚组以制定 RP 的重要性 循证实践。	1. 上下文提到了考虑亚组的重要性，这意味着一个建议或建议。 2. “重要性”一词表明，这种观点并非软弱无力。 3. 答案是强建议。	weak advice	强 建议
尽早进行手术有助于减轻 梗阻造成的作用力，并有助于 面部正常生长。	1. 上下文表明手术是有帮助的，这意味着一种建议。 2. “尽快”这一短语表示一种紧迫性，使建议更强。 3. 答案是强建议。	weak advice	strong advice
这项研究显示缺乏规范性共识 关于预立照护计划，并强调了一致性的必要性 关于……作用的教育项目 预立照护计划流程中的CBLTC服务提供者。	1. 上下文指出需要统一的教育项目，这是一条建议。 2.“highlights the need”一词表明该观点并不强烈。 3. 答案是弱建议。	strong advice	weak advice
Pattern A 基于特定关键词的错误推理步骤 模式 B 健康建议分类错误			

表 ***A8***： *HealthAdvice* 测试集结构化摘要中建议的三样本思维链生成示例（由 *GPT-4* 生成）。

句子	GPT-4 的回答	Gold	Pred
未来研究应寻求实施更多 循证 改进措施 生活质量在这个 population.	1. “shouldseek to implement”一词表明上下文中有一条建议。 2. “Should”表示强烈意见。 3. 答案是强建议。	no advice	strong advice
In conclusion, 对比增强全心脏心脏磁共振血管成像 在 3.0T 下可显示正常及变异的心脏 静脉解剖。	1. 上下文陈述了一个结论，这并非建议。 2. 答案是无建议。	weak 建议	no 建议
然而， 由于我们的分析基于 数量相对较少的基因，因此需要开展规模更大的研究， 以确定疾病结局的最佳预测因子 总体结果。	1. 上下文提到“需要一项更大规模的研究”，这表明这是一条建议。 2. “将需要”表明这是一种强烈意见。 3. 答案是强建议。	no 建议	强 建议
CDI 500 直列血气分析仪的制造商 监视器建议在使用前进行气体校准。	1. 上下文给出了一项推荐，这是一种建议。 2. “recommends”一词表示强烈意见。 3. 答案是强建议。	no advice	强 建议
健康 促进在交通运输行业工作场所 应该是未来研究的一个重要主题。	1. 术语“should be”表明上下文中有一个建议。 2. "应该"表示强烈意见。 3. 答案是强建议。	no advice	strong advice
用碘仿治疗将是有益的。	1. 术语“would be beneficial”在上下文中表示一条建议。 2. “将会是”表示强烈意见。 3. 答案是强建议。	weak advice	strong advice
全面的幼儿视力筛查 program in 利用夏威夷 自动手持式	1. 术语“may have”在上下文中表示建议。 2. “May have”表示一种弱观点。	no advice	weak advice

照片筛查仪可能具有公共卫生价值。	3. 答案是弱建议。		
尽管阿尔茨海默病相关遗传变异似乎 to 对早期认知变化有轻微影响， 将需要进行重复验证，以确定是否存在 is a 可辨别的影响 on cognitive status in 中年期.	1. 短语“重复验证将是必要的”表明了该语境中的一项建议。 2. “将被要求”表示强烈意见。 3. 答案是强建议。	no advice	strong advice
在具有普遍性的样本中进行重复验证试验和 标示了更长时间。	1. “有指征”一词暗示了上下文中存在一条建议。 2. “Are indicated”意味着强烈意见。 3. 答案是强建议。	no advice	strong advice
Pattern A 基于特定关键词的错误推理步骤 模式 B 健康建议的错误分类			

表***A9***： *GPT-4* 为*HealthAdvice*测试集中的非结构式摘要里的建议生成的三样本思维链示例。

句子	GPT-4 答案	Gold	Pred
In this 沙特阿拉伯样本中，平均牙弓宽度为 与土耳其样本相比更窄，[19]a 一个北美样本、一个南美样本、一个 韩国样本，[20] 以及一个埃及样本；[21]在其上 其他方面，它接近一个马来样本。[22]这些 族群在牙弓尺寸方面的差异解释了 每种情况都需要特定的正畸弓丝。 患者，基于初始牙弓形态。	1) 该声明正在讨论不同种族之间牙弓尺寸的差异 groups. 2) 该声明表明需要为每位患者使用特定的正畸弓丝 根据其初始牙弓形态。 3)该主张没有使用强烈或模棱两可的措辞，但确实暗示了一项可采取行动的 正畸治疗建议。 4) 因此，该声明具有弱建议。 答案：(2) 弱建议。	no 建议	weak 建议
In 比较， 患者3表现出显著的 PR3-ANCA抗体生成，但没有证据表明 外周B细胞再增殖，这强调了 需要考虑其中的肉芽肿 含有对驱动至关重要的三级淋巴结构 该复发性病程。 肉芽肿性多血管炎作为B细胞	1) 该论述正在讨论将GPA中的肉芽肿视为B细胞的重要性 含有三级淋巴结构。 2) 该主张使用了“强调需要考虑”这一短语，这表明一个 建议或推荐。 3) 然而，建议的强度并未明确为强或弱，而是更倾向于 由于使用了“emphasizes”一词，趋向于给出弱建议。 答案：（2）弱建议。	strong advice	weak advice
Despite considerable challenges that face the interpretation of 临床解剖学的 correlations in 神经退行性 疾病， the current results 表明至少左侧前颞叶的部分区域 新皮层应插入颞叶-外侧裂区 语言网络，在其中它们可能发挥关键作用 在将词语与其客体指称对象关联起来以及在 提升命名和词语的分类精确度 理解。	1）该论断包含一项可执行的建议，涉及将左侧 前颞新皮层的部分区域纳入颞叶-外侧裂语言网络。 2）在描述该建议时，作者使用了“suggest that”和 “likely to play critical roles”等措辞，这表明该论断的力度并不强，而是 较弱。 3）因此，该论断属于弱建议。 答案：（2）弱建议。	strong advice	weak advice
The results, coupled with 其他相关性和 准实验结果， 强调.....的重要性 providing high quality 循证 reading 贯穿低年级阶段的教学。	1) The claim has an actionable suggestion related to providing 高质量， 低年级的循证阅读教学。 2) 在描述该建议时，作者使用了短语“强调重要性” of”，这表明该主张的力度很强。 3) 因此，该主张具有强建议。 答案：（1）强建议。	weak advice	strong advice
护士了解抢救措施非常重要。 镇痛策略，例如患者自控镇痛，以保护 针对行走时的爆发痛、剂末 失效于控释阿片类药物，或自发性 无明显病理学的疼痛（Chiaretti & Langer， 2005年）。	1) 该主张有一个可操作的建议，让护士意识到补救镇痛 策略。 2）描述该建议时，作者使用了“it is important for”这一表述， 这表明该建议的力度很强。 3）因此，这一主张属于强建议。 答案：（1）强建议。	weak 建议	强 advice

体外受精诊所采用低氧环境是合理的 自胚胎培养开始时的浓度，并且它 将优化患者在一个 长期视角。	1) 该声明　　有一项与使用低氧相关的可操作建议 体外受精诊所中的浓度。 2) 在描述这项建议时，作者使用了“合理”一词，这 表明对这项建议有一定程度的信心，因此属于强建议。 3) 因此，这项主张属于强建议。 答案：(1) 强建议。	weak 建议	strong advice
尽管研究结果提供了有关生活质量的信息 and some influencing 其中的因素 艾滋病病毒感染者在 格鲁吉亚，目前研究的局限性可以是 在未来的研究中加以考虑并尽量减少。	1）该说法讨论的是研究结果及其局限性。 2) 该论断表明，未来可以考虑并尽量减少这些局限性 研究，这是对未来研究的一项可切实执行的建议。 3) 然而，该主张并未使用强烈的情态词或短语，使得建议 弱。 答案：（2）弱建议。	no 建议	weak 建议
模式 A　基于特定关键词的错误推理步骤 模式 B　健康建议的错误分类			

表 ***A10:*** 针对 *HealthAdvice* 测试集讨论中的建议，由 *GPT-4* 生成的三样本思维链样本。

句子	GPT-4 答案	Gold	Pred
现有文献支持这一观点，即 权威型教养可能对日后 超重和肥胖具有保护作用，但研究结果并不一致。	1）该句描述了“权威型教养”与“日后 超重和肥胖”之间的关系。 2）作者使用了“具有保护作用”这一语言线索，表明两者之间存在关系 在这两个实体之间。 3) 然而，作者还提到“研究结果不一”，这表明 这种关系并非明确或确定。 4) 因此，该句子表达了一种相关性，但带有一些不确定性。 答案：2 - 相关性	conditional 因果关系	相关性
AN诊断与CD之间的双向关联 值得在初步评估中予以关注 以及对这些疾病的随访，因为 对这些疾病的诊断不足和误诊 可能会导致病程迁延，并造成不必要的发病负担。	1) 该句描述了“AN诊断”与“CD”之间的关系。 2) 作者使用了“ 双向关联”这一表述，这暗示存在相关性。 在这两种条件之间。 3) 因此，该句表达的是相关性。 答案：2 - 相关性	conditional 因果关系	相关性
该数值将低于公认的国际 进行筛查性乳腺摄影的阈值，并且 引发了辐射防护问题。	1) 该句描述了“该数值低于公认的 国际阈值”和“引发辐射防护问题”。 2) 作者使用了动词“raises”，这表明存在直接效应 在它们之间。 3）在描述该影响时，作者未使用任何模糊限制语或情态动词 降低对该效应确定性的表述。 4) 因此，该句表达了直接因果关系。 答案：4 - 因果关系	无关系	因果关系
高危人群可能会接受将对比增强乳腺摄影作为 筛查检查，并且可能更倾向于选择它，而非筛查磁共振成像 如果正在进行的试验证明筛查用对比增强乳腺摄影 临床上不劣于 MRI。	1) 该句子描述了“将对比增强乳腺摄影作为筛查 检查”与“正在进行的试验证明筛查对比增强乳腺摄影临床上不劣于 MRI". 2) 作者使用了条件性语言“可能接受”和“如果正在进行的试验 证明”，这表明它们之间存在条件效应。 3) 因此，该句表达了条件因果关系。 答案：3 - 条件因果关系	no relation	conditional 因果关系
NAC后的3D MRI重建能够准确 检测新辅助后的残留肿瘤 化疗，并有助于选择那些 在NAC后接受保乳治疗并伴有 肿瘤降期。	1) 该句子描述了“NAC后3D MRI重建”之间的关系 以及“准确检测残留肿瘤”和“有助于选择患者进行 NAC 后进行保乳治疗并实现肿瘤降期" 2) 作者使用了“能够准确检测”和“有助于选择”这两个动词， 表明它们之间存在直接效应。 3) 在描述效果时，作者还使用了模糊限制语“could”来缓和语气 效应的确定性。 4) 因此，该句仅表达条件因果关系。 答案：3 - 条件因果关系	direct 因果关系	条件 因果关系

结果显示，芳香疗法没有影响。 生活质量、睡眠质量和生命体征。	1) 该句描述了“芳香疗法”与“生活质量、睡眠”的关系 质量，和生命体征 2) 作者使用了“无影响”这一短语，这表明不存在无关系 两个实体之间。 3) 因此，该句子表达无关系。 答案：1 - 无关系	direct 因果关系	no 关系
TCF7L2的rs7903146（C/T）多态性 基因可能与 喀麦隆人群中的肥胖无关。	1) 该句描述了“ TCF7L2基因的rs7903146（C/T）多态性”与“喀麦隆人群中的肥胖”之间的关系。 2) 作者使用了“可能与……无关”这一语言线索来描述这种 关系，这表明缺乏相关性。 3) “might not”的使用表示不确定性，但重点仍然在潜在的 不存在关系。 4) 因此，该句子表达无关系。 答案：1 - 无关系	correlation	no relation
该方案将节省资金、时间和工作量 而不损害患者安全和未来 管理。	1) 该句子描述了“该方案”与“节省资金、 时间和工作量”以及“不损害患者安全和未来 管理”。 2) 作者使用了动词“will save”和“without compromising”，这表明 方案与所提及的结果之间存在直接效应。 3) 效应的确定性很强且直接，因为作者使用了“will”一词来 描述该效应。 4) 因此，该句子表达了因果关系。 答案：4 - 因果关系	no relation	因果关系
我们应当进行重大政策变革以增加 未来筛查项目和健康教育 能够产生总体积极影响的项目 关于减轻疾病负担。	1) 该句描述了“重大政策变革”与 "减轻疾病负担" 2) 作者使用了语言提示“would havean overall positive impact on”，该提示 表明二者之间存在直接效应。 3) 在描述该效果时，作者使用了模糊限制语“would”来弱化这种 效应的确定性。 4) 因此，该句子仅表达条件因果关系。 答案: 3 - 条件因果关系	no relation	conditional 因果关系
Pattern A 基于特定关键词的错误推理步骤 模式 B 推理关系分类错误			

表 ***A11***：*GPT-4* 在 *CausalRelationtest* 数据集上进行推理时生成的四样本思维链示例。

附录 B：

提示模板	用于	ID	数据集
"上下文：{context} \n 问题：这是 2) 强建议、1) 弱建议，还是 0) 无建议？"	零样本	cs	HealthAdvice
"上下文：{context} \n 问题：上下文中是否包含任何医疗建议？：2：是，1：可能，0：no"		cs	HealthAdvice
上下文：{context} \n 问题：这是一条（0）无建议、（1）强建议，还是（2）弱建议陈述?		yy	HealthAdvice
上下文：{context} \n 问题：该主张是否包含（1）强建议、（2）弱建议，或者是(0) 无建议?	One-shot	yy	HealthAdvice
上下文：{context} \n 问题：这是（0）无建议、（1）强建议，还是（2）弱建议?	Few-shot	db	HealthAdvice
上下文：{context} \n 问题：这是什么类型的建议？仅从以下选项中选择一项：0 – 无建议，1 - 强建议，或 2 - 弱建议。		db	HealthAdvice
上下文：{context} \n 将句子标记为强烈医疗建议、弱医疗建议或无医疗建议	零样本思维链	gs	HealthAdvice
上下文：{context} \n 此句中医疗建议（如有）的力度是……。		gs	HealthAdvice
上下文：{context}-- 问题：这描述的是哪种类型的关系？只能从以下选项中选择一个：1) 无关系、2）相关性、3）条件因果关系或4）直接因果关系。		hv	HealthAdvice
上下文：{context} -- 问题：前一个陈述是否描述了一个（1）直接相关关系、（2）条件因果关系、（3）因果关系，或（0）无关系。relationship.		hv	HealthAdvice
上下文：{context} \n 问题：从以下因果关系中选择：0：无，1：相关性 , 2: 条件因果 , 3：直接因果?		cs	因果关系
上下文：{context}\n问题：这是：0) 无，1) 相关性，2) 条件因果，3) 直接因果因果？"		cs	因果关系
上下文：{context} \n 问题：句子表达的是 1 - 相关性、2 - 条件因果关系，还是 3 - 直接因果关系，还是属于 0 - 无关系句子?	少样本	yy	因果关系
上下文：{context} \n 问题：这是 1 - 无关系、2 - 相关性、3 - 条件因果关系，还是 4 – 直接因果关系?	One-shot	yy	因果关系
上下文：{context} \n 问题：这描述的是哪种关系？请仅选择一个：0 -无关系，2 - 相关性，3 - 条件因果关系，或 4 – 直接因果关系。	Zero-shot 零样本思维链	db	因果关系
Context: {context} \n Question: 这是否描述了 (1) 直接相关关系，(2) 条件因果关系，(3) 因果关系，还是 (0) 无关系。		db	因果关系
Context: {context} \n 将句子中表达的关系标记为相关性、条件因果关系、因果关系或其他		gs	因果关系
上下文：{context} \n 句子中的关系类型为……		gs	因果关系
上下文：{context} -- 问题：这描述的是哪种关系？只能从以下选项中选择一个：1) 无关系, 2) 相关性, 3) 条件因果关系, 或 4) 直接因果关系.		hv	因果关系
上下文：{context} -- 问题：前一个陈述是否描述了一个（1）直接相关关系、（2）条件因果关系、（3）因果关系，或（0）无关系。relationship.		hv	因果关系

表 ***B1***：提示模板，无思维链。第二列表示为每个数据集选择的提示。

Prompt Template	Dataset	Setting
“背景：第三，尽管我们的研究已考虑大量潜在混杂因素。 考虑到这一点，我们无法完全排除未测量混杂的可能性。 问题：这项主张属于（1）强建议、（2）弱建议，还是（0）无建议？ 让我们一步一步思考： 1) 该声明谈论的是当前研究的局限性。 2) 该声明没有任何关于健康相关临床或政策变化的可操作建议。 3) 因此，该声明没有强建议或弱建议。它是一个无建议声明。 答案：(0) 无建议。 上下文：在HIV领域工作的格鲁吉亚公共卫生专家应优先实施 HIV患者中的此类干预措施。 问题：该主张是否具有（1）强建议、（2）弱建议，还是（0）无建议？ 让我们逐步思考： 1）该主张包含与实施干预措施相关的可操作建议。 2）在描述该建议时，作者使用了情态动词“should”，这表明了 该建议的强度，且为强建议。 3) 因此，该主张有强建议。 答案：（1）强建议。 在此，我们证明，根治性手术后的癌症复发显著低于 ANP治疗患者中高于对照组患者，提示ANP可能可用于预防 手术后癌症复发。 问题：这项主张属于（1）强建议、（2）弱建议，还是（0）无建议？ 让我们一步一步思考： 1) 该主张提出了使用 ANP 预防手术后癌症复发的可操作建议。 2) 在描述该建议时，作者使用了模糊措辞“could potentially be used to”，这 表明该主张的强度不强而是弱的。 3) 因此，该主张具有弱建议。 答案：(2) 弱建议。 上下文：{context} 问题：此声明是否有 (1) 强建议、(2) 弱建议，还是 (0) 无建议？ 让我们一步一步思考：	分类： 建议在 讨论 sections	Few-shot CoT

背景：沙眼衣原体血清学抗体滴度与其他检测方法的相关性，依据的是 直接荧光抗体（DFA）检测和培养，而非用于 筛查的核酸检测。 问题：这是（0）无建议、（1）弱建议，还是（2）强建议？让我们逐步思考。 答案： 1. 背景中陈述了一个事实 , 这不是一项建议。 2. 答案是无建议。 背景：这些结果表明，LOS检测是一种有价值的工具，应纳入所有 用于筛查有平衡障碍主诉的创伤性脑损伤患者的客观平衡评估中。 问题：这是（0）无建议、（1）弱建议，还是（2）强建议？让我们一步一步地思考。 答案： 1. “应纳入”这一表述表示上下文中给出了建议。 2. “应当”表示强烈意见。 3. 答案是强建议。 背景：因此，在常规心脏手术中，当术中抗纤溶治疗可能不适用时 其他节血技术被采用。 问题：这是（0）无建议、（1）弱建议，还是（2）强建议？让我们逐步思考。 ANSWER: 1. 术语“may not be indicated”在上下文中表示一项建议。 2. “可能不”表示弱观点。 3. 答案是弱建议。 上下文: {context} QUESTION: 这是一个 (0) 无建议, (1) 弱建议, 还是 (2) 强建议? 让我们一步一步思考。	分类： 建议位于 非结构式 摘要	 少样本 CoT
背景：有必要在更大的队列中开展进一步的机制研究，以协调……的潜在作用 2型糖尿病在子宫肌瘤风险中。 QUESTION: 这是一个 (0) 无建议, (1) 弱建议, 还是 (2) 强建议? 让我们一步一步思考。 ANSWER: 1. 上下文指出了进一步研究的必要性，这不是一项建议。 2. 答案是无建议。 由于血压问题在黑人群体中的病程更为严重，我们建议予以鼓励 在偏好夜间服药的人群中，夜间服药可能有益；对于需要服用两种或更多药物的人群，建议应由一名…… 应在夜间服用。 QUESTION: 这是一个 (0) 无建议, (1) 弱建议, 还是 (2) 强建议? 让我们一步一步思考。 ANSWER: 1. “我们建议”和“建议”表示上下文中存在建议。 2. “推荐”和“建议”表示指令性意见。 3. 答案是强建议。 因此，该治疗方案可作为痤疮治疗的可行选择，无论是作为单药治疗还是 作为联合治疗。 问题：这是（0）无建议、（1）弱建议，还是（2）强建议？让我们逐步思考。 ANSWER: 1. “将是一个可行的选项”一词表明上下文中有一条建议。 2. “Would be”表示该观点并不强烈。 3. 答案是弱建议。 上下文：{context} QUESTION: 这是一个 (0) 无建议, (1) 弱建议, 还是 (2) 强建议? 让我们一步一步思考。	Classification: Advice in 结构式 摘要	 少样本 CoT

背景：短信的高使用率使得招募年轻成年男性进行前瞻性研究变得可行 其中使用个性化短信来促进健康行为。 问题：2 - 相关性、3 - 条件因果关系或 4 – 直接因果关系是否表达在 句子，还是它是一个 1 - 无关系句子? 让我们一步一步地思考： 1) 该句子没有描述两个实体之间的任何相关性或因果关系。 2）这是一个无关系句。 答案：1 - 无关系 上下文：跌倒发生率和生活质量差可能与……的存在部分相关 抑郁症。 问题：句子中表达的是 2 - 相关性、3 - 条件因果关系，还是 4 – 直接因果关系， 还是一个 1 - 无关系句? 我们一步一步来分析： 1) 这句话描述了“跌倒发生率和生活质量差”与 “抑郁症”之间的关系。 2) 它们之间是相关性，因为作者使用了“associated with”这一语言线索来 描述这种关系。“Associated with”是表示相关性的常用表达。 3) 因此，这句话表达的是相关性。 答案：2 - 相关性 背景：我们的研究提供了初步证据，表明水果摄入量较高的母亲和 妊娠期间油炸食品和腌制肉类摄入较低可能降低单侧视网膜母细胞瘤的风险在 它们的后代。 问题：在其中表达的是 2 - 相关性、3 - 条件因果关系，还是 4 – 直接因果关系 句子，还是一个 1 - 无关系句子? 让我们一步一步地思考： 1) 该句描述了“水果含量较高、油炸食品和腌制食品含量较低的饮食”之间的关系 肉类" 和 "的风险 "单侧视网膜母细胞瘤在后代中"。 2) 作者使用了动词“reduce”，这表明它们之间存在直接效应。 3) 在描述该效应时，作者还使用了模糊限制语“可能”来降低其表述的确定性。 effect. 3) 因此，该句仅表达条件因果关系。 答案：3 - 条件因果关系 背景：通过实施以下措施，可以改善接受结肠手术患者的营养疗程。 在麻醉后监护室早期使用口服营养补充剂。 问题：该句表达的是 2 - 相关性、3 - 条件因果关系还是 4 – 直接因果关系， 还是属于 1 - 无关系的句子? 让我们一步一步地思考： 1) 该句描述了“营养过程”与“口服营养风险之间的关系。 麻醉后监护室中的补充剂"。 2) 作者使用了“可以通过……得到改进”这一动词短语，这表明两者之间存在直接效应 them. 3）在描述效果时，作者还使用了情态动词“can”，这表明其确定性给主人留下些什么吧 影响强烈且具有直接因果关系。并且没有使用模糊限制语来弱化语气的确定性。 3）因此，该句表达了因果关系。 答案：4 - 因果关系	推理： 因果关系 检测	少样本 CoT

表 ***B2***： 思维链（思维链）提示模板