arXiv:2305.02220v2 [cs.CL] 3 Jun 2023

WangLab参加MEDIQA-Chat 2023：使用大语言模型根据医患对话生成临床病历

John Giorgi[1, 2, 3*] **Augustin Toma**[1, 3, 4*] **RonaldXie**[1, 2, 3, 4*]
Sondra S. Chen[1,5] **KevinR. An**[1,6] **GraceX.Zheng**[1,5] **BoWang**[1, 3, 4*]

[1]多伦多大学 [2]Terrence Donnelly细胞与生物分子研究中心[3]Vector人工智能研究所 [4]大学健康网络 [5]桑尼布鲁克健康科学中心[6]心脏外科系，多伦多大学

{john.giorgi, augustin.toma, ronald.xie, bowang.wang}@mail.utoronto.ca

摘要

本文介绍了我们参加MEDIQA-Chat 2023共享任务的提交方案，该任务旨在根据医患对话自动生成临床笔记。我们报告了两种方法的结果：第一种是在共享任务数据上微调预训练语言模型（PLM），第二种则使用大语言模型（LLM）进行少样本上下文学习（ICL）。两种方法在自动指标（如ROUGE、BERTScore）上的表现均很出色，在所有共享任务提交方案中分别排名第二和第一。专家人工评估表明，使用GPT-4的基于上下文学习的方法生成的病历，其受青睐的频率与人工撰写的记录大致相当，这使其成为根据医患对话自动生成病历的一条颇具前景的途径。[1]

1 引言

临床文档日益沉重的负担已成为医疗保健领域的一个关键问题，导致临床医生的工作满意度下降、职业倦怠率上升，并对患者的就医体验产生负面影响（Friedberg 等人，2013；Babbott 等人，2014年；Arndt 等人，2017年）。另一方面，及时、准确地记录患者就诊情况，对于确保安全、有效的医疗护理以及专科医生之间的沟通至关重要。因此，利用自动生成问诊记录的方式为临床医生提供辅助，正受到越来越多的关注（Finley 等人，2018年；Enarvi 等人，2020年；Molenaar 等人，2020年；Knoll 等人，2022年）。

为进一步推动从医患对话中自动生成临床病历的研究，研究者提出了 MEDIQA-Chat Dialogue2Note共享任务（Ben Abacha，2023年）。本文介绍我们在子任务 B 中的提交：从医患对话生成完整临床病历。我们探索了两种方法；第一种方法对预训练语言模型（PLM，§3.1）进行微调，

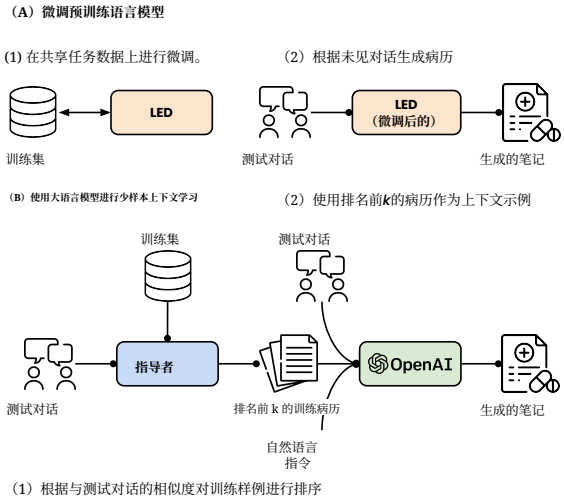

图 1：（**A**）对预训练语言模型（预训练语言模型，Longformer 编码器-解码器（LED，Beltagy 等人 2020年）进行微调。（**B**）使用大语言模型（大语言模型）进行上下文学习（上下文学习）。我们使用指导者（Su 等人，2022a）根据训练样例与测试对话之间的相似度对训练样例进行排序。随后，将排名前k 的最相似样例的病历作为上下文示例，与自然语言指令共同构成提示，并输入 GPT-4（OpenAI，2023年），以生成临床病历。

第二种则采用少样本上下文学习（上下文学习，§3.2）。根据自动自然语言生成指标（§4）衡量，两种方法均表现出色，在该共享任务的所有提交中分别排名第二和第一。在由三位专家医生参与的人工评估中，基于上下文学习的方法使用 GPT-4 生成的病历，其受青睐的频率与人工撰写的病历大致相同（§4.3）。

2 共享任务和数据集

MEDIQA-Chat 2023 提出了两项共享任务：

1.**Dialogue2Note 摘要**：给定一段医患对话，任务是生成一份临床病历，总结对话内容，并包含一个或多个病历部分——

*核心贡献者。参见作者贡献

[1]https://github.com/bowang-lab/MEDIQA-Chat-2023

示例医患对话

【医生】 您好，Thompson女士。我是Moore医生。您好吗?
【患者】 你好，Moore医生。
【医生】 你好。
【患者】 除了膝盖不太舒服之外，我感觉还好。
【医生】 好，嘿，Dragon，Thompson女士今年43岁。
女性，因右膝疼痛前来就诊。那么，告诉我你的膝盖发生了什么事?
你的膝盖怎么了?
【患者】 嗯，我当时，呃，正想换一个灯泡，然后我
正站在梯子上，不小心踉跄了一下
为了稳住身体，我试着接住自己，结果扭到了膝盖。
【医生】 好的。你还有其他地方受伤吗?
【患者】 不，不。看起来只是膝盖。
【医生】 好的。这是什么时候发生的?
【患者】 是昨天。
【<风格 id='2'>医生</风格>】 好的。那么，呃，主要是哪里疼?
【患者】 我的膝盖内侧疼，就在膝盖内侧。
【医生】 好的。
【患者】 就在这里。
【医生】 好的。有什么会让情况好转或恶化吗?
【患者】 我一直在冰敷，呃，而且我还一直在
服用布洛芬，但似乎没什么效果。
【医生】 好的。听起来你是几天前摔倒的，
而且右膝里面受了伤。
【患者】 嗯。
【医生】 而且你一直在用一点冰敷，呃，敷在
上面，但实际上并没有太大帮助，还服用了一些布洛芬。是
这样吗?

-------- 已截断 ---------

【医生】 那么总结一下，根据我的检查，呃，观察你的
膝盖，呃，从X光片和检查结果来看，你有一些
内侧半月板处有压痛，所以我认为您有
目前可能是急性内侧半月板扭伤或拉伤。
呃，目前我的建议是让您接受一种
护膝，呃，我们接下来会让您用一些搭配的用品
接下来几天暂时使用拐杖。我们会给你安排。
大约一周后回来复查，看看你的情况如何，如果
情况没有好转，我们到时就做磁共振成像检查。
【患者】 好的。
【医生】 我建议给你开一些布洛芬，
每次800毫克。呃，你大约每六小时可以服用一次，呃，
随餐服用。呃，我们会给你开大约两周的药量。
【患者】 好的。
【医生】 好的。嗯，你有什么问题吗?
【患者】 没有，我觉得我没问题了。
【医生】 好的。嘿，Dragon，请订购药物并
完成讨论过的操作，并定稿报告。好的，跟我来
我们一起去办理出院手续。

临床病历示例

主观

~~主诉~~
右膝疼痛。

~~现病史~~
Thompson女士，43岁，女性，前来就诊
今天接受右膝疼痛评估。她表示
她当时正试图站在梯子上更换灯泡 [...]

~~当前用药~~
布洛芬、地高辛。

~~既往病史~~
心房颤动。

~~既往手术史~~
鼻整形术。

客观检查

EXAM
右膝检查显示，屈曲时疼痛。
内侧关节线处有压痛。小腿无疼痛。
外翻应力测试时疼痛。感觉正常。

客观结果

~~结果~~
右膝X光片未见明显急性
骨折或脱位。可见轻度积液。

评估与计划

~~诊断意见~~
右膝急性内侧半月板扭伤。

PLAN
此时，我与患者讨论了诊断和治疗
方案。我建议佩戴护
膝。她将每 6 小时服用 800 毫克布洛芬，随
食物，持续两周。接下来她会使用拐杖
几天。她将在1周后复诊 [...]

图 2：子任务 B 验证集中的一组医患对话与临床病历示例。为便于阅读，对对话进行了轻微清理（例如删除末尾空格）。对话和病历的部分内容已截断。评估时，各部分归入以下四类之一：“主观”“客观检查”“客观结果”和“评估与计划”（详情见 §2.1 ）。

部分（例如评估、既往病史）。

2. **记录到对话生成**：给定一份临床病历，任务是生成一段与病历所述内容相关的合成医患对话。

我们专注于 Dialogue2Note，该任务分为两个子任务。在子任务 A（Ben Abacha 等人，2023年）中，目标是根据部分医患对话生成病历的特定部分。在子任务 B（Yim 等人，2023 年）中，目标是根据完整对话生成完整的病历。本文其余部分将聚焦于子任务 B；有关我们针对子任务 A 的方法，请参见附录 A。该方法在共享任务的所有提交中也排名第一。

2.1 任务定义

每个 k 示例均由一段医患对话和 $D = d_1$, ……, d_k 及相应的临床病历组成，$N = n_1$, ……, n_k。目标是根据一段对话 n_i 自动生成病历 d_i。每份病历包含一个或多个部分，例如“主诉”和“家族史”。评估时，各部分归入以下四类之一：“主观”、“客观检查”、“客观结果”和“评估与计划”。[2] 示例见图 2，其中展示了一组医患对话及其对应的临床病历。

[2]参见 此处 了解映射关系

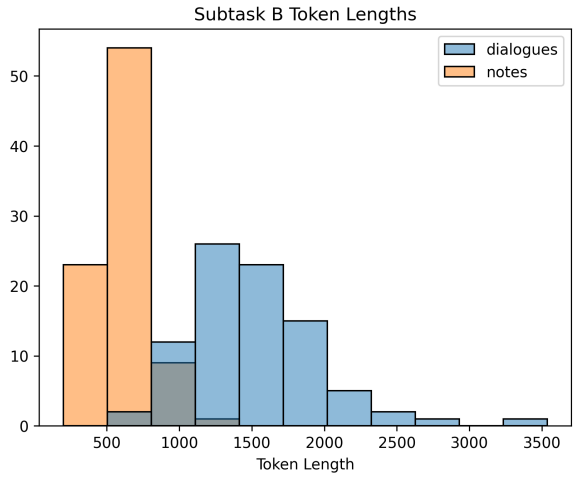

图 3：子任务 B 训练集和验证集的词元长度直方图。对话和病历使用`tiktoken` 的“gpt-4”编码进行词元化。

2.2 数据集

该数据集包含 67 个训练样例和 20 个验证样例，内容包括医患会诊的转录对话及其对应的临床医生撰写的笔记。每个样例均标注了“数据集来源”（dataset source），用于指明生成病历时所采用的对话转录系统。

3 方法

我们采用两种高性能方法完成共享任务。第一种方法是在提供的训练集（§3.1）上微调预训练语言模型（PLM）。第二种方法是使用大语言模型（LLM，§3.2）进行上下文学习（ICL）。

3.1 微调预训练语言模型

作为第一种方法，我们按照标准的序列到序列训练流程，在训练集上微调预训练语言模型（图 1 A；详情请参见附录 C）。鉴于输入对话的长度（图 3），我们选择使用 Longformer 编码器—解码器（LED，Beltagy 等，2020年），其最大输入长度为 16,384 个词元。我们从一个在 PubMed 摘要数据集上调优的 $\text{LED}_{\text{LARGE}}$ 检查点开始进行微调（Cohan 等，2018年），该检查点在初步实验中表现最佳。[3] 模型使用 HuggingFace Transformers（Wolf 等，2020年）在单块 NVIDIA A100-40GB GPU 上进行微调。我们在验证集上对超参数进行了少量调整。[4]

[3] https://huggingface.co/patrickvonplaten/led-large-16384-pubmed [4]详情请参见附录B.1 。

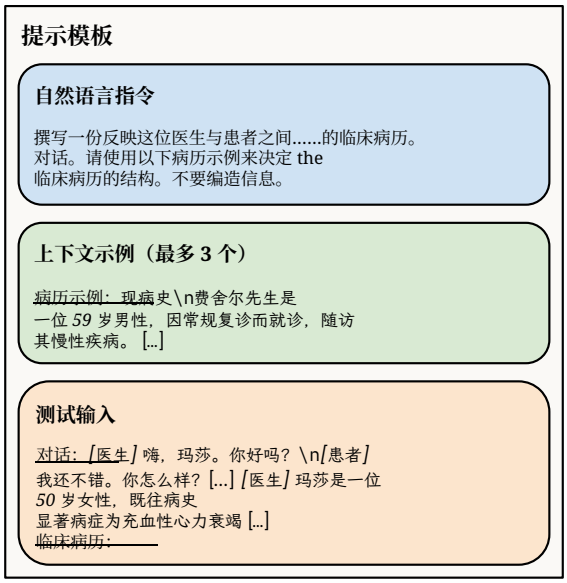

图 4：我们基于上下文学习（ICL）的方法所使用的提示模板。每个提示都包含自然语言指令、最多 3 个上下文示例，以及一段作为输入、未见过的医患对话。

3.2 使用大语言模型进行上下文学习

作为第二种方法，我们尝试使用上下文学习完成子任务 B。我们选择 GPT-4（OpenAI，2023年）[5] 作为大语言模型，并设计了一个简单的提示，其中包括自然语言指令和上下文示例（图 4）。我们将提示长度限制为 6,192 个词元——为 2,000 个输出词元留出空间，因为模型的最大词元长度为 8,192——并在此限制内尽可能多地使用上下文示例，最多不超过 3 个。我们将`temperature` 参数设为 0.2，其余OpenAI API 的超参数均保持默认值。

自然语言指令 在初步实验中，我们发现 GPT-4 对提示中自然语言指令的确切措辞并不十分敏感。因此，我们选择使用简短、简单的指令（图 4）。

上下文示例选择 每个上下文示例都是训练集中的一份病历。为了选择这些病历，我们首先对每个训练样本中的对话以及输入对话进行嵌入。随后，根据与输入对话的余弦相似度对训练对话进行排序；选择排名前 k 个训练样本对应的病历作为上下文示例（见 图 1，B）。对话使用 Instructor（Su 等人，2022年a）进行嵌入，这是一种文本编码器，

[5]具体而言，2023年3月14日的快照“`gpt-4-0314`”

支持自然语言指令。[6] 最后，我们将上下文示例限制为与输入对话来自同一“数据集来源”（见§2.2），并假设这可能会提升性能。[7]

3.3 评估

使用官方评估脚本对模型进行评估[8]，评估数据为验证集（因为未提供测试笔记）。使用 ROUGE（Lin，2004年）、BERTScore（Zhang 等人，2020年）和 BLEURT（Sellam 等人，2020年）将生成的笔记与提供的真实笔记进行评估。我们将 ROUGE-1 F1、BERTScore F1 和 BLEURT-20 的算术平均值作为性能报告指标（Pu 等人，2021年）。

4 结果

4.1 微调预训练语言模型

我们在表1中展示了微调 LED 的结果。由于 LED 实现具有非确定性，[9] 我们报告了三轮训练的平均结果。不出所料，我们发现，将 LED 模型规模从$_{\text{BASE}}$（12 层，~162M 个参数）扩大到 LED$_{\text{LARGE}}$（24 层，~460M 个参数）可显著提升性能。使用在 PubMed 摘要数据集上微调后的检查点初始化模型，可进一步提升性能（LED$_{\text{LARGE-PubMed}}$）。这很可能是因为：(1) Dialouge2Note 与摘要任务相似；(2) PubMed 文本比用于预训练 LED 的通用领域文本更接近临床文本。[10] 我们使用此方法提交的共享任务结果总排名第二，平均分比排名其后的提交高出 2.7 分；这一差距与我们将模型规模扩大一倍所带来的性能提升相当（参见 LED$_{\text{BASE}}$ 与 LED$_{\text{LARGE}}$，表1）。

4.2 使用大语言模型进行上下文学习

我们展示了使用 GPT-4 进行上下文学习的结果，见表 2。我们注意到一些有趣的趋势，并按影响程度排序。首先，选择上下文示例

表1：微调LED。展示了三轮训练的均值和标准差（SD）。扩大模型规模并在相关任务上进行预训练可提升性能。**粗体**：最佳分数。

模型	ROUGE-1 F1	BERTScore F1	BLEURT	Avg.
LED$_{\text{BASE}}$	57.0$_{0.4}$	67.3$_{0.1}$	36.9$_{0.0}$	53.8
LED$_{\text{LARGE}}$	59.8$_{0.2}$	70.0$_{0.6}$	41.1$_{0.8}$	57.0
LED$_{\text{LARGE-PubMed}}$	**61.7$_{0.4}$**	**70.7$_{0.2}$**	**41.5$_{0.6}$**	**57.9**

基于对话相似度筛选示例有显著的正面影响，通常可使平均分提高 4 分或以上。 与对话-笔记对相比，仅使用病历作为上下文示例也有正面影响，通常可使平均分提高 ~1。令人惊讶的是，增加上下文示例的数量对表现的影响微乎其微。这些结果共同表明，上下文示例的主要作用是提供有关预期笔记结构、风格和长度的指导。最后，将上下文示例筛选为与输入对话来自同一“数据集来源”，对表现的影响可以忽略不计。

最佳策略比LED高出近3分（平均分为60.8，而57.9，见表1&表2），在所有共享任务提交结果中名列第一，比排名第二的结果高出 > 9平均分。我们得出结论：（1）使用GPT-4进行少样本上下文学习，即使仅使用一个示例，也是从医患对话生成病历的高效方法；（2）使用语义相似的对话-笔记对中的病历，是选择上下文示例的有效策略。

4.3 人工评估

像 ROUGE、BERTScore 和 BLEURT 这样的自动评估指标并不完美，且可能无法与人类判断的各个方面相关联。[11]因此，我们进行了专家人工评估，以验证我们的结果。为使标注可行，我们使用表现最佳的微调模型 LED，在验证集（20 个示例）上进行了评估$_{\text{LARGE-PubMed}}$ (表1)，并采用表现最佳的基于上下文学习的方法：3-shot、相似且仅包含病历的示例，并按数据集类型进行筛选（表2）。

向三名高年资住院医师[12] 展示每个示例的真实笔记、微调模型生成的病历，以及基于上下文学习的方法生成的病历（以临床病历“A”、“B”和“C”的形式随机呈现），

[6]我们使用了以下指令：“为聚类表示医疗对话：{dialogue}”[7]人工审查发现，数据集来源可以预测病历结构和风格；这很可能是因为它表明了生成病历的临床医生或电子健康记录系统[8]

h

ttps://github.com/abachaa/MEDIQA-Chat-2023/blob/main/scripts/evaluate_summarization.py[9]

https://github.com/huggingface/transformers/issues/12482[10]LED 基于 BART 初始化（Lewis 等人，2020年），BART 在维基百科和 BooksCorpus 文本的组合上进行了预训练（Zhu 等人，2015）

[11]参见第6节，了解更详细的讨论[12]三位标注者是作者中的一部分，他们在标注前未与模型或模型输出进行交互

表2：使用 GPT-4 进行上下文学习。展示三次运行的 ROUGE-1 F1、BERTScore F1 和 BLEURT 均值。根据与输入对话的相似度选择上下文示例可提升性能。将对话-笔记对作为上下文示例（因令牌长度限制，省略 3-shot 结果）的表现不如仅笔记。将上下文示例筛选为与输入对话具有相同“数据集来源”，几乎没有影响。**粗体**：最佳分数。所有情况下的标准差为 < 0.1。

示例选择策略	未筛选				按数据集来源筛选			
	0-shot	1-shot	2次示例	3次示例	零次示例	1次示例	2次示例	3次示例
对话-笔记对作为上下文示例								
随机	52.2	54.5	53.9	–	–	54.8	54.5	–
相似对话	–	59.4	59.4	–	–	60.1	60.3	–
仅病历作为上下文示例								
随机	–	56.3	56.7	56.7	–	56.3	56.5	56.7
相似对话	–	60.7	60.6	60.4	–	**60.8**	60.4	**60.8**

表3：人工评估。三位医师从人工撰写的真实病历（GT）、微调模型生成的病历（FT）和上下文学习（ICL）生成的病历中选出偏好的病历。胜率是指剔除平局后，病历获选偏好的案例所占百分比。

	偏好			Ties		胜率（%）		
医师	GT	FT	ICL	FT/ICL	全部	GT	FT	ICL
1	9	1	4	2	4	64	7	29
2	5	0	14	0	1	26	0	74
3	9	0	6	0	5	60	0	40
总计	23	1	24	2	10	48	2	50

并要求他们根据一段对话和一些简单的说明，选择自己偏好的病历：

说明：请根据**提供的医患对话**，评估**临床笔记 A**、**B** 和**C** 。对于每组笔记，请选择您更喜欢的一份（“A”、“B”或“C”）。如果您对两份笔记的偏好大致相同，请选择（“A/B”、“B/C”或“C/A”）。如果您没有偏好，请选择“A/B/C”。一份“优质”笔记应包含对话中提及的所有**关键**、大多数**非关键**信息，并且尽可能少地包含**无关**信息：

- **关键**：根据医学法律要求，必须记录用于诊断和治疗决策的事项；若缺失或记录有误，之后可能导致错误的诊断和治疗。例如，在疑似胸部感染的问诊中，症状“咳嗽”就是此类信息。这是病历必须准确记录的关键信息，以免误导临床医生。

- **非关键**：应记录在完整病历中的项目，但缺少这些项目不会影响后续治疗或诊断，例如在胸部感染咨询中记录“患者与谁同住”。

- **无关信息**：咨询中涉及的医学无关信息，例如一名疑似胸部感染患者的宠物刚刚去世。

关键、非关键和无关信息的定义取自此前关于生成临床笔记人工评估的研究（Moramarco 等人，2022年；Savkov 等人，2022年）。

简而言之，评估者强烈偏好上下文学习生成的病历，而非微调模型生成的病历；平均而言，略微偏好于人工撰写的病历（表3），这验证了自动指标所报告的高性能。不过，我们注意到标注者间一致性较低，并在第6节推测了可能的原因。

5 相关工作

近年来，基于医患对话的自动化病历生成受到了越来越多的关注（Finley 等人，2018年；Enarvi 等人，2020年；Molenaar 等人，2020年；Knoll 等人，2022年）。研究人员提出了多种方法，例如抽取式与抽象式相结合的方法（Joshi 等人，2020年；Krishna 等人，2021年；Su 等人，2022b），以及对预训练语言模型进行微调（Zhang 等人，2021年，与我们在 §3.1 中采用的方法类似）。其他研究则侧重于整理用于训练和基准测试的数据（Papadopoulos Korfiatis 等人，2022年），包括利用大语言模型生成合成数据（Chintagunta 等人，2021年）。最后，还有一些研究致力于改进生成的临床笔记的评估方式，包括采用自动指标（Moramarco 等人，2022年）和人工评估（Savkov 等人，2022年）。尽管近期文献已探讨了上下文学习在病历生成方面的潜力（Lee 等人，2023年），但我们的工作是最早对这一方法进行严格评估的研究之一。

6 结论

我们介绍了我们在 MEDIQA-Chat 医患对话笔记生成共享任务中的提交方案。我们评估了基于 LED 的微调方法和基于 GPT-4 的上下文学习方法，在所有提交方案中分别排名第二和第一。由三位医师进行的人工评估显示，相比 LED 生成的病历，评估者明显更偏好 GPT-4 通过上下文学习生成的病历；平均而言，他们也略微更偏好这些病历，而非人工撰写的记录。我们得出结论：上下文学习是实现医患对话临床笔记生成的一条很有前景的路径。

局限性

生成文本的评估很困难对自动生成的文本（包括临床笔记）进行评估通常很困难，因为输出质量的许多方面本质上都具有主观性。ROUGE 和 BERTScore 等自动评估指标并不完美（Deutsch 等人，2022年），且未必能与专家判断的各个方面相关联。然而，这些指标常用于评估模型生成的临床笔记，并且确实与某些质量方面相关（Moramarco 等人，2022年）。为进一步验证我们的发现，我们还邀请了三位专家医生进行人工评估（§4.3）。如前所述（Savkov 等人，2022年），即便是对临床笔记进行人工评估，也远非完美；标注者间一致性通常较低，这可能是因为医师对每条患者陈述的重要性以及是否应将其纳入问诊记录有不同看法。我们在人工评估中也发现标注者间一致性较低，并推测这部分是由于医师的专科背景不同所致。来自家庭医学专业的医师1和医师3彼此之间的一致性较高，但与医师2（心脏外科）的意见一致性较低，见表3。如何改进临床笔记（以及更广泛意义上的生成文本）的自动指标并制定最佳评估实践，是一个活跃的研究领域。我们希望在未来的工作中采用新颖且性能优异的指标。

数据隐私 虽然我们基于 GPT-4 的解决方案性能最佳，但它不符合健康保险流通与责任法案等数据保护法规；尽管 Azure 确实宣传其符合健康保险流通与责任法案的规定。[13] 从隐私角度来看，可能更倾向于在本地部署 LED 等模型；然而，我们的结果表明，要使这种方法达到可接受的性能水平，仍需开展更多工作（参见 表3）。无论采用哪种方式，在实施自动化临床病历生成系统时，医疗保健提供者和开发人员都应确保整个系统——包括文本转语音、数据传输与存储以及模型推理——符合隐私和安全要求，以维护信任并防止临床环境中的隐私泄露。

伦理声明

开发用于根据医患对话生成临床病历的自动化系统，会引发多项伦理考量。首先，知情同意至关重要：必须让患者知晓其谈话会被录音，并优先保障数据所有权。公平获取同样重要；系统必须适用于不同背景的患者，包括残障人士、技术素养有限者或面临语言障碍者。必须解决数据偏见和公平性问题，以避免某些患者群体受到不公平待遇或遭到误诊。系统必须采取严格的安全措施，保护患者数据免遭未经授权的访问或泄露。对于使用自动化系统进行笔记生成所导致的错误或伤害，明确责任归属至关重要。披露使用系统时已知的局限性或潜在风险，对于维护医患关系中的信任必不可少。最后，必须持续开展评估，确保系统性能不会下降，也不会对护理质量造成负面影响。

致谢

本研究部分得益于加拿大数字研究联盟（alliancecan.ca）和安大略计算中心（www.computeontario.ca）提供的支持。我们感谢所有内部和外部审稿人提出的宝贵意见，这些意见帮助改进了本稿的早期版本。

作者贡献

John Giorgi (JG)、AugustinToma(AT) 和 Ronald Xie (RX) 总体负责该项目，包括

[13] https://azure.microsoft.com/en-us/products/cognitive-services/openai-service#security

数据清洗与处理、模型实现以及开展实验。JG 撰写了初稿，并在 AT 和 RX 的反馈下设计了人工评估。AT 和 RX 邀请了 Sondra S. Chen、Kevin R. An 和 Grace X. Zheng，他们作为专家医生参与了人工评估。Bo Wang 提供了高层次的反馈和建议。

参考文献

Brian G. Arndt、John W. Beasley、M. Watkinson、Jonathan L Temte、Wen-Jan Tuan、Christine A. Sin- sky 和 Valerie J. Gilchrist。2017年。与电子健康记录紧密相连：使用电子健康记录事件日志数据和时间动作观察评估初级保健医生工作负荷。《家庭医学年鉴》，15：419–426。

Stewart F. Babbott、Linda Baier Manwell、Roger L. Brown、Enid N. H. Montague、Eric S. Williams、Mark D. Schwartz、Erik P. Hess 和 Mark Linzer。2014年。初级保健中的电子病历与医生压力：MEMO研究的结果。美国医学信息学协会杂志：*JAMIA*，21 e1:e100–6。

Iz Beltagy、Matthew E. Peters 和 Arman Cohan。2020年。Longformer：长文档Transformer。*arXiv*，abs/2004.05150。

Asma Ben Abacha、Wen wai Yim、Griffin Adams、Neal Snider 和 Meliha Yetisgen。2023年。关于医患对话摘要与生成的 MEDIQA-Chat 2023 共享任务概述。载于*ACL-ClinicalNLP 2023*。

Asma Ben Abacha、Wen-wai Yim、Yadan Fan 和 Thomas Lin。2023年。基于医患会诊生成临床病历的实证研究。载于计算语言学协会欧洲分会第*17*届会议论文集，第2283–2294页，克罗地亚杜布罗夫尼克。计算语言学协会。

Bharath Chintagunta、Namit Katariya、Xavier Amatri- ain 和 Anitha Kannan。2021年。具备医学知识的GPT-3作为医疗对话摘要的数据生成器。载于第二届医疗对话自然语言处理研讨会论文集，第66–76页，线上。计算语言学协会。

Hyung Won Chung、Le Hou、S. Longpre、Barret Zoph、Yi Tay、William Fedus、Eric Li、Xuezhi Wang、Mostafa Dehghani、Siddhartha Brahma、Albert Web- son、Shixiang Shane Gu、Zhuyun Dai、Mirac Suz- gun、Xinyun Chen、Aakanksha Chowdhery、Dasha Valter、Sharan Narang、Gaurav Mishra、AdamsWei Yu、Vincent Zhao、Yanping Huang、Andrew M. Dai、Hongkun Yu、Slav Petrov、Ed Huai hsin Chi、Jeff Dean、Jacob Devlin、Adam Roberts、Denny Zhou、Quoc V. Le 和 Jason Wei。2022年。《扩大指令微调语言模型规模》。*arXiv*，abs/2210.11416。

Arman Cohan、Franck Dernoncourt、Doo Soon Kim、Trung Bui、Seokhwan Kim、Walter Chang 和 Nazli Goharian。2018年。一种用于长文档抽象式摘要的语篇感知注意力模型。载于《*2018*年北美计算语言学协会分会年会：人类语言技术论文集》第 *2* 卷（短篇论文），第 615–621 页，路易斯安那州新奥尔良。计算语言学协会。

Daniel Deutsch、Rotem Dror 和 Dan Roth。2022年。重新审视自动摘要评估指标的系统级相关性。载于《*2022*年北美计算语言学协会分会年会：人类语言技术论文集》，第 6038–6052 页，美国西雅图。计算语言学协会。

Seppo Enarvi、Marilisa Amoia、Miguel Del-Agua Teba、Brian Delaney、Frank Diehl、Stefan Hahn、Kristina Harris、Liam McGrath、Yue Pan、Joel Pinto、Luca Rubini、Miguel Ruiz、Gagandeep Singh、Fabian Stemmer、Weiyi Sun、Paul Vozila、Thomas Lin 和 Ranjani Ramamurthy。2020年。使用序列到序列模型从患者与医生的对话中生成医疗报告。载于《首届医疗对话自然语言处理研讨会论文集》，第 22–30 页，线上。计算语言学协会。

Gregory Finley、Erik Edwards、Amanda Robinson、Michael Brenndoerfer、Najmeh Sadoughi、James Fone、Nico Axtmann、Mark Miller 和 David Suendermann-Oeft。2018年。用于记录临床诊疗过程的自动化医疗记录员。载于*2018*年计算语言学协会北美分会年会论文集：演示，第 11–15 页，路易斯安那州新奥尔良。计算语言学协会。

Mark W. Friedberg、Peggy G. Chen、Kristin R Van Busum、Frances Aunon、Chau Pham、John P. Caloy-eras、Soeren Mattke、Emma Pitchforth、Denise D. Quigley、Robert Henry Brook、Francis J. Crosson 和 Michael A. Tutty。2013年。影响医生职业满意度的因素及其对患者护理、卫生系统和卫生政策的影响。兰德健康季刊，3 4:1。

Alex Graves. 2012年。使用循环神经网络进行序列转导。*arXiv*预印本，abs/1211.3711。

Anirudh Joshi、Namit Katariya、Xavier Amatriain 和 Anitha Kannan。2020年。Dr.summarize：通过利用局部结构对医疗对话进行全局摘要。载于计算语言学协会自然语言处理经验方法会议（*EMNLP 2020*）论文集，第 3755–3763 页，线上。计算语言学协会。

Tom Knoll、Francesco Moramarco、Alex Papadopoulos Korfiatis、Rachel Young、Claudia Ruffini、Mark Perera、Christian Perstl、Ehud Reiter、Anya Belz 和 Aleksandar Savkov。2022年。用户驱动的病历生成软件研究。载于计算语言学协会北美分会*2022*年会议：人类语言技术论文集，第 385–394 页，美国西雅图。计算语言学协会。

Kundan Krishna、Sopan Khosla、Jeffrey Bigham 和 Zachary C. Lipton。2021年。 使用模块化摘要技术从医患对话生成SOAP病历。载于计算语言学协会第*59*届年会暨第*11*届自然语言处理国际联合会议论文集（第*1*卷：长篇论文），第4958–4972页，线上。计算语言学协会。

Peter Lee、Sébastien Bubeck 和 Joseph Petro。2023年。GPT-4 作为用于医学的人工智能聊天机器人的优势、局限性和风险。《新英格兰医学杂志》，388 13：1233–1239。

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020年。BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。载于 第*58*届计算语言学协会年会论文集，第7871–7880页，线上。计算语言学协会。

Chin-Yew Lin。2004年。 ROUGE：一个用于自动评估摘要的软件包。载于 文本摘要分支拓展，第74–81页，巴塞罗那，西班牙。计算语言学协会。

Ilya Loshchilov 和 Frank Hutter。2019年。解耦权重衰减正则化。发表于 第七届国际学习表征会议，*ICLR 2019*，新奥尔良，路易斯安那州，美国，*2019*年*5*月*6*日至*9*日。OpenReview.net。

Sabine Molenaar、Lientje Maas、Verónica Burriel、Fabi- ano Dalpiaz 和 Sjaak Brinkkemper。2020年。医疗保健自动化报告的医疗对话摘要。高级信息系统工程研讨会，382:76–88。

Francesco Moramarco、Alex Papadopoulos Korfiatis、Mark Perera、Damir Juric、Jack Flann、Ehud Reiter、Anya Belz 和 Aleksandar Savkov。2022年。诊疗记录生成中的人工评估及其与自动指标的相关性。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第5739–5754页，爱尔兰都柏林。计算语言学协会。

OpenAI。2023年。Gpt-4技术报告。*arXiv*，abs/2303.08774。

Alex Papadopoulos Korfiatis、Francesco Moramarco、Radmila Sarac 和 Aleksandar Savkov。2022年。Pri- Mock57：一个初级保健模拟咨询数据集。载于计算语言学协会第*60*届年会论文集（第*2*卷：短篇论文），第588–598页，都柏林，爱尔兰。计算语言学协会。

Amy Pu、Hyung Won Chung、Ankur Parikh、Sebastian Gehrmann 和 Thibault Sellam。2021年。面向机器翻译的紧凑指标学习。载于《*2021*年自然语言处理实证方法会议论文集》，第751–762页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Aleksandar Savkov、Francesco Moramarco、Alex Papadopoulos Korfiatis、Mark Perera、Anya Belz 和 Ehud Reiter。2022年。会诊核对清单：规范化医疗病历生成的人工评估。载于《*2022*年自然语言处理实证方法会议论文集：产业专题》，第111–120页，阿联酋阿布扎比。计算语言学协会。

Thibault Sellam、Dipanjan Das 和 Ankur Parikh。2020年。BLEURT：学习用于文本生成的稳健指标。载于《计算语言学协会第*58*届年会论文集》，第7881–7892页，线上。计算语言学协会。

Hongjin Su、Weijia Shi、Jungo Kasai、Yizhong Wang、Yushi Hu、Mari Ostendorf、Wen tau Yih、Noah A. Smith、Luke Zettlemoyer 和 Tao Yu。2022a。《One embedder, any task: Instruction-finetuned text embeddings》。《style id='1'>arXiv，abs/2212.09741。

Jing Su、Longxiang Zhang、Hamidreza Hassanzadeh 和 Thomas Schaaf。2022b。《使用 BART 从医患对话中提取并摘要临床病历》。《style id='1'>Interspeech。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017年。《Attention is all you need》。载于《神经信息处理系统进展》第*30*卷：神经信息处理系统年会，*2017*年*12*月*4*日至*9*日，美国加利福尼亚州长滩，第5998–6008页。

Thomas Wolf、Lysandre Debut、Victor Sanh、J ulien Chaumond、Clement Delangue、Anthony Moi、Pier- ric Cistac、Tim Rault、Remi Louf、Morgan Funtow- icz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020年。Transformers：最先进的自然语言处理。载于 *2020*年自然语言处理实证方法会议论文集：系统演示，第 38–45 页，线上。计算语言学协会。

Wen-wai Yim、Yujuan Fu、Asma Ben Abacha、Neal Snider、Thomas Lin 和 Meliha Yetisgen。2023年。《The》

aci 演示语料库：一个开放数据集，用于评测基于医患对话的自动笔记生成领域的最新技术。已投稿至《自然·科学数据》。

Longxiang Zhang、Renato Negrinho、Arindam Ghosh、Vasudevan Jagannathan、Hamid Reza Hassanzadeh、Thomas Schaaf 和 Matthew R. Gormley。2021 年。利用预训练模型自动总结医患对话。载于计算语言学协会论文集：*EMNLP 2021*，第 3693–3712 页，蓬塔卡纳，多米尼加共和国。计算语言学协会。

Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q. Weinberger 和 Yoav Artzi。2020年。BERTScore：使用 BERT 评估文本生成。载于第八届国际学习表征会议（*ICLR 2020*），埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。OpenReview.net。

Yukun Zhu、Ryan Kiros、Richard S. Zemel、Ruslan Salakhutdinov、Raquel Urtasun、Antonio Torralba 和 Sanja Fidler。2015。对齐书籍与电影：通过观看电影和阅读书籍，生成类故事视觉解释。载于*2015* 年 *IEEE* 国际计算机视觉会议（*ICCV 2015*），智利圣地亚哥，*2015* 年 *12* 月 *7—13* 日，第 19–27 页。IEEE计算机学会。

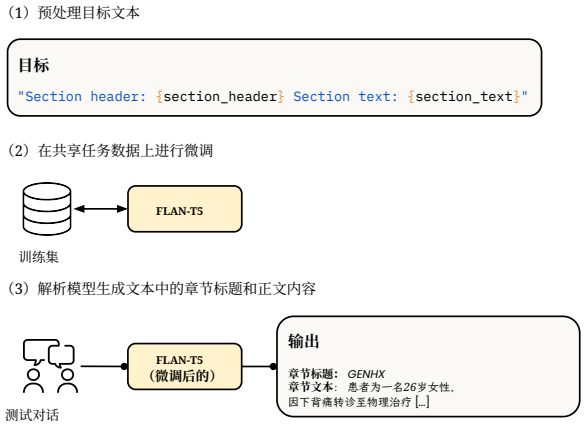

图 5：针对子任务 A 对 FLAN-T5（Chung 等人，2022年）进行微调。训练前，将目标文本预处理为“章节标题：{section_header} 章节文本：章节文本：{section_text}”。解码后，使用正则表达式解析章节标题和章节文本。

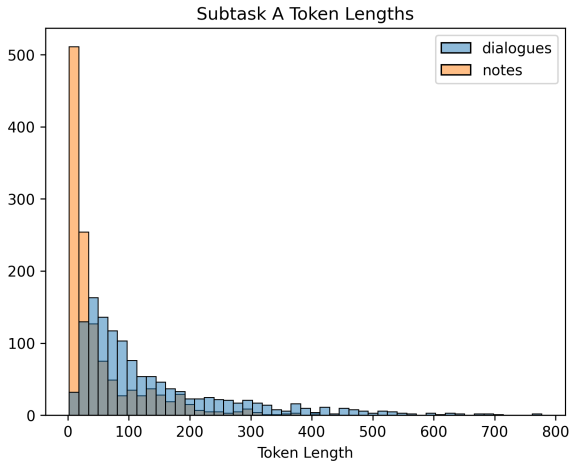

图 6：子任务 A 训练集和验证集的词元长度直方图。对话和病历使用HuggingFaceTokenizers进行词元化，所用模型为“google/flan-t5-large”。为便于查看，省略了大于第 99 百分位数的长度。

子任务 A

在 Dialogue2Note 摘要共享任务的子任务 A 中，给定一段不完整的医患对话，目标是：（1）预测合适的章节标题，例如“既往病史”；（2）生成病历中的相应章节。我们通过在提供的训练集上对预训练语言模型进行微调来完成此任务，并采用经典的序列到序列训练流程（详情见附录 C）。在初步实验中，我们发现指令微调的 FLAN-T5（Chung 等人，2022年）在这项任务上表现尤为出色。

我们假设，同时学习预测章节标题和生成章节文本能够提升整体表现。为此，我们对训练集进行了预处理，使目标文本格式如下：“章节标题：{section_header}章节文本：{section_text}”。解码后，我们使用正则表达式解析章节标题和文本，并分别进行评估（图 5）。章节标题预测的评估指标为预测标题与真实值相符的比例（准确率）；章节文本则采用与子任务 B 类似的方式进行评估（见 §3.3）。如果模型输出了无效的章节标题，[14] 我们会将其替换为“GENHX”（一般病史），因为它通常能够概括其他章节的内容。模型在单张 NVIDIA A100-40GB GPU 上进行了微调。在验证集上对超参数进行了少量调优（表 4）。

我们展示了我们的方法在验证集上的结果见表 5。与子任务 B（见 §4.1）类似，我们发现，或许并不令人意外的是，将模型规模从 FLAN-T5$_{\text{BASE}}$（24 层、~250M 个参数）扩展到 FLAN-T5$_{\text{LARGE}}$（48 层、~780M 个参数）后，性能得到大幅提升。通过联合学习章节标题预测和病历章节生成，性能进一步提升。我们基于该方法提交的共享任务系统在章节标题预测方面并列第一（准确率为 78%），在病历章节生成方面排名第一（ROUGE-1、BERTScore 和 BLEURT 的平均 F1 分数为 57.9）。

B 子任务 B

B.1 LED超参数调优

我们针对 ROUGE-1 F1、BERTScore F1 和 BLEURT-20 的平均分数，对 LED 的超参数进行了轻微调整$_{\text{LARGE-PubMed}}$，调整所依据的是子任务 B 验证集。获得的最佳超参数见表 6。微调 LED$_{\text{BASE}}$ 和 LED$_{\text{LARGE}}$ 时，我们在 §4.1 中使用了相同的超参数。

B.2 LED 输出后处理

在实践中，我们发现微调后的 LED 模型有时会生成无效的章节标头；值得注意的是，使用 GPT-4 的基于上下文学习的方法并未出现此问题。因此，我们使用一个简单的脚本对 LED 的输出进行了轻度后处理，该脚本会识别模型生成的、不属于真实值集合的章节标题，并进行模糊匹配。

[14] 在实际应用中，我们发现微调模型几乎不会生成无效的章节标题。

表 4：在 Dialogue2Note 子任务 A 中使用 FLAN-T5 的超参数

超参数	值	注释
最大值_源_长度	1024	将输入序列截断至此最大长度
最大_目标_长度	512	将输出序列截断至此最大长度
源_前缀	“总结以下患者- 医生对话。包含所有医学 相关信息，包括家族 病史、诊断、既往病史（及 手术史）、免疫接种情况、实验室检 查结果及已知过敏史。你应 首先预测最相关的临床 病历章节标题，然后概 述对话。对话：”	添加到所有输入前的指令文本
训练_批次_大小	8	训练期间的批次大小
评估_批次_大小	12	推理期间的批次大小
学习_率	1e-4	训练期间的学习率
优化器	AdamW（Loshchilov 和 Hutter，2019年）	训练期间使用的优化器
num_训练_轮次	20	训练轮次总数
预热_比例	0.1	训练步数中用于线性提高学习 率至学习_率
学习率_调度器	带预热的线性调度	学习率在最初的预热_比例内线性提高 训练步数的一部分，之后线性递减至 0
权重_衰减	0.01	未应用于偏置和层归一化权重。
标签_平滑处理	0.1	训练期间使用的标签平滑系数
bf16	true	是否在训练期间使用 BF16
num_束	2	集束搜索解码时使用的束宽

表 5：微调 FLAN-T5。表中展示了预测章节标题的准确率和生成的笔记各章节的得分。联合学习预测章节标题和生成病历可提升性能。**粗体**：最佳得分。

		笔记生成			
模型	标题预测（%）	ROUGE-1 F1	BERTScore F1	BLEURT	Avg.
随机标题	8.0	–	–	–	–
多数标题栏	22.0	–	–	–	–
FLAN-T5$_{\text{BASE}}$	71.0	40.1	70.5	52.7	54.5
FLAN-T5$_{\text{LARGE}}$	**79.0**	**49.8**	**74.5**	**58.0**	**60.8**
↪不含标题预测	–	48.0	74.3	57.6	59.9

字符串匹配[15]，将其替换为最接近的有效章节标题。例如，在一次运行中，该过程将预测的错误章节标题“HISTORY OF PRESENT”转换为最接近的有效章节标题“HISTORY OF PRESENT ILLNESS”。

C 微调序列到序列模型

在为子任务 A（附录 A）和 B（§3.1）训练序列到序列（seq2seq）模型时，我们遵循了标准的监督微调（监督微调）流程。我们从一个经过预训练的、基于编码器-解码器 Transformer 的语言模型开始（Vaswani 等人，2017年）。首先，编码器将输入中的每个词元映射为上下文嵌入。然后，自回归解码器逐个词元生成输出，并在每个时间步关注编码器的输出。解码过程持续进行，直到生成特殊的“序列结束”词元（例如 </s>），或达到生成词元数量的上限。形式化地说，X 是输入序列，在我们的案例中即医患对话；Y 是相应的输出序列，其长度为 T，在我们的案例中即临床病历。我们对条件概率进行建模：

$$p(Y|X) = \prod_{t=1}^{T} p(y_t|X, y_{<t}) \quad (1)$$

在训练期间，我们针对模型参数 θ 优化序列交叉熵损失：

$$\ell(\theta) = -\sum_{t=1}^{T} \log p(y_t|X, y_{<t}; \theta) \quad (2)$$

最大化训练数据的对数似然。与常见做法一样，我们在训练期间使用教师强制，将先前的真实值输入馈送给 the

[15]我们使用了 https://github.com/seatgeek/thefuzz

表 6：Dialogue2Note 子任务中 Longformer 编码器-解码器 (LED) 使用的超参数 B

超参数	值	说明
最大_源_长度	4096	将输入序列截断至此最大长度
最大_目标_长度	1024	将输出序列截断至此最大长度
源_前缀	“总结以下患者- 医生对话。包含所有医学 相关信息，包括家族 史、诊断、既往病史（以及 手术史）、免疫接种、实验室检 查结果和已知过敏史。对话：”	附加在所有输入前的指令文本
训练_批次_大小	8	训练期间的批次大小
评估_批次_大小	6	推理期间的批次大小
学习_率	3e-5	训练期间的学习率
优化器	AdamW（Loshchilov 和 Hutter，2019年）	训练期间使用的优化器
num_训练_轮次	50	训练轮次总数
预热_ratio	0.1	用于线性增加学习率的训练步数比例 学习率_学习率
学习率_调度器	带预热的线性调度	学习率在最初的预热_比例期间线性增加 训练步数的一部分，之后线性下降至 0
权重_衰减	0.01	不应用于偏置和层归一化权重
标签_平滑	0.1	训练期间使用的标签平滑因子
fp16	true	训练期间是否使用 FP16
数量_束	4	集束搜索解码时使用的束宽
最小_长度	100	生成序列的最小长度
最大_长度	1024	生成序列的最大长度
长度_惩罚	2.0	值 > 0 会促进生成较长的输出序列
不_重复_n 元语法	3	此大小的 n 元语法只能出现一次

解码器在预测序列中的下一个词元时。在推理过程中，我们使用集束搜索生成输出（Graves，2012年）。根据应用长度惩罚后的平均词元对数概率对集束进行排序。使用 HuggingFace Transformers 库对模型进行微调。[16]

[16] https://github.com/huggingface/transformers/blob/main/examples/pytorch/summarization/run_summarization.py