计算机科学
COMPUTER SCIENCE

基于大小模型结合与迭代反思框架的电子病历摘要生成方法

钟博洋，阮彤，张维彦，刘井平

引用本文

钟博洋，阮彤，张维彦，刘井平．基于大小模型结合与迭代反思框架的电子病历摘要生成方法[J]．计算机科学，2025，52(9)：294-302.

ZHONG Boyang, RUAN Tong, ZHANG Weiyan, LIU Jingping. Collaboration of Large and Small Language Models with Iterative Reflection Framework for Clinical Note Summarization [J]. Computer Science, 2025, 52(9): 294-302.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241000114

基于大小模型结合与迭代反思框架的电子病历摘要生成方法

钟博洋　阮　彤　张维彦　刘井平
华东理工大学信息工程与科学学院　上海 200237
(a1561418501@163.com)

摘　要　在医疗人工智能领域,从医患对话中自动生成电子病历(EMR)是一项核心任务。传统主流方法多依赖于大规模语言模型(LLM)结合少量示例进行学习,然而,这些方法往往未能有效融入深度的医学专业知识,导致生成的EMR内容在专业性方面存在不足。针对这一挑战,提出了一种新颖的迭代反思框架,该框架融合了Error2Correct示例学习与领域模型监督,旨在提升EMR的总结质量。具体而言,首先设计了一种集成了Error2Correct示例学习机制的大规模语言模型,用于EMR的初步生成与持续优化,并在预生成阶段融入医学领域知识。然后利用一个经过微调的小规模医学预训练语言模型,对初步生成的EMR进行进一步的评估与优化,从而在后生成阶段再次深化领域知识的整合。最后,引入了一个迭代调度器,该调度器能够高效地引导模型在持续的反思与迭代过程中进行优化。实验结果显示,所提方法在两个公开的EMR数据集上均展现出了先进的性能。特别是在IMCS-V2-MRG和ACI-BENCH数据集上,与经过微调的大规模语言模型相比,所提方法分别实现了3.66个百分点和7.75个百分点的整体性能提升[1)]。

关键词:大规模语言模型;医疗预训练模型;摘要生成;大模型反思;大小模型结合
中图分类号　TP391

Collaboration of Large and Small Language Models with Iterative Reflection Framework for Clinical Note Summarization

ZHONG Boyang,RUAN Tong,ZHANG Weiyan and LIU Jingping
School of Information Science and Engineering,East China University of Science and Technology,Shanghai 200237,China

Abstract　Generating clinical notes from doctor-patient dialogues is a critical task in medical artificial intelligence. Existing methods typically rely on large language models(LLMs) with few-shot demonstrations but often struggle to integrate sufficient domain-specific knowledge,leading to suboptimal and less professional outputs. To address this problem,a novel iterative reflection framework is proposed,which integrates Error2Correct example learning and domain-model supervision,aiming to improve the summary quality of EMR. Specifically,a large-cale language model integrating the Error2Correct example learning mechanism is designed for the initial generation and continuous potimization of EMR,and the medical domain knowledge is integrated into the pre-generation stage. Then,this paper uses a lightweight medical pre-training language model,fine-tuned with domain data,to evaluate the refined content,integrating domain knowledge in post-generation. Finally,an iterative scheduler is introduced,which can effectively guide the model to optimize in the continuous process of reflection and improvement. Experimental results on two public datasets demonstrate that the proposed method achieves state-of-the-art performance. Compared with the fine-tuned large language models,the proposed method improves overall performance by 3.68% and 7.75% on IMCS-V2-MRG and ACI-BENCH datasets.

Keywords　Large language model,Medical pre-trained model,Summarization generation,Large model reflection,Collaboration of large and small models

1　引言

电子病历是医疗领域的重要资源[1],为医生提供了关键的参考信息,有助于他们做出准确的诊断和治疗决策[2]。目前,电子病历主要由医生根据医患对话手动撰写,这不仅增加了医生的工作负担,还占用了他们本可以用于护理患者的宝贵时间[3-4]。因此,从医患对话中自动生成电子病历是一项具有重要价值的任务。

目前针对这一任务主要有两种研究范式。第一种范式是微调预训练语言模型(PLMs),如BART[5]和T5[6],用于适应特定专业领域摘要生成任务[7-8]。在这一范式下,典型的研究方法通过设计分段-总结式摘要生成框架[9-10],或是将医学知

1) https://github.com/steins048596/REFLEXES

到稿日期:2024-10-21　返修日期:2025-01-024
通信作者:刘井平(jingpingliu@ecust.edu.cn)

识融入深度模型中以提高生成质量[11-12]。第二种范式是利用指令微调或少样本示例增强的大规模语言模型(LLMs)[13-14]，充分发挥其丰富的通用领域知识，来生成高质量的摘要文本。其中，指令微调指的是将训练样本转换为指令数据，对 Llama3 和 ChatGLM 等开源大规模语言模型[15]进行少量的参数调整。另一种方式是少样本示例，即在推理过程中为大规模语言模型(如 GPT-4)提供一个或多个任务示例，使其能够基于这些示例从医患对话中生成电子病历[16]。

本研究主要采用第二种范式，即使用大规模语言模型(LLMs)从医患对话中生成医学电子病历。通过分析此前基于 LLMs 方法生成的错误案例[16]，发现有两个关键方面需要进一步优化。1)该范式下生成的结果并不总是严格遵循官方医学文件中规定的明确要求。例如，官方文件要求在“主诉”部分必须包括症状、症状部位及发生时间等信息，而 LLMs 生成的结果往往缺少关键的时间信息，即使包含时间信息，也通常较为模糊，如“1－2 天”，而非官方文件要求的精确时间表达。2)该范式下的方法普遍没有关注到医学领域特有的语言风格方面的隐含要求，尤其是在中文电子病历中。例如，医患对话中常用的口语化表达“时不时咳嗽”，在电子病历中应正式记录为“阵发性咳嗽”，以符合医学术语规范。这些问题难以解决的主要原因在于：医疗领域有较高的专业性要求，能整合领域知识的正样本数量有限，难以为 LLMs 提供足够的学习资源，从而无法生成高质量的医学电子病历。

为了解决上述问题，本文提出了两种在大规模语言模型(LLMs)生成前和生成后阶段融入医学知识的策略。1)在 LLMs 的输入中引入错误-正确(Error2Correct)示例。这些示例包括一个错误示例、详细的错误原因及正确表达方式。其中的错误原因主要涉及两个方面：未能遵循明确的医学指南要求，以及未使用隐含的医学语言风格进行书写。通过这些示例，LLMs 能够模仿人类的认知过程，识别并纠正错误，从而提升自我优化能力。2)开发一个小规模语言模型的、领域特定的模型，该模型使用与任务相关的数据进行微调，用于评估 LLMs 生成的结果。此模型作为一个专门的知识库，涵盖了规范要求和语言风格等方面的专业知识，指导 LLMs 不断优化其输出，最终实现更高质量的结果。值得注意的是，与之前直接用于生成任务相比，将垂直领域的小规模语言模型用于判别电子病历的质量更为合理，因为判别任务比生成任务更简单，而小参数量的模型在处理简单任务时表现更为出色。

本文提出了一种创新的框架——结合 Error2Correct 示例和小规模语言模型领域模型监督的迭代反思框架(REFLEXES)，以实现 LLMs 在生成电子病历方面的持续优化。具体而言，该框架首先采用具备上下文学习(ICL)能力的 LLMs(如 GPT-4)生成完整的电子病历，作为初步的对话摘要；然后，对于初始对话摘要的每个部分，使用另一个带有设计好的 Error2Correct 示例的 LLMs(如 ChatGPT3)对其内容进行优化，确保其符合显性和隐性的要求；最后，采用一个小规模的医学预训练语言模型，并对其进行微调，使其能够对 LLMs 生成的内容进行有效评估，并提供有价值的反馈指导。

本文的主要贡献如下：

1)首次提出了一个用于电子病历总结的迭代反思框架。该框架使大规模语言模型(LLMs)能够持续优化其输出，以符合明确的医学指南和隐含的医学语言风格要求。

2)开发了两种在生成前和生成后阶段融入医学知识的策略。第一种策略是 Error2Correct 示例，通过帮助识别和纠正错误，增强自我优化能力；第二种策略是引入一个小规模的医学预训练语言模型，用于评估和提升专业性及领域任务性能。

3)在中文和英文数据集上的实验结果表明，REFLEXES 框架取得了先进的性能表现。尤其是在 IMCS-V2-MRG 和 ACI-BENCH 数据集上，相比之前的最佳方法，该方法的整体平均得分分别提高了 3.66 个百分点和 7.75 个百分点。

2 相关工作

与电子病历任务相关的研究可以分为两大类，即基于预训练语言模型(PLMs)的生成和基于大语言模型(LLMs)的生成，后者通常涉及少量示例演示或指令微调。本章将对这两类研究进行详细回顾。

1)基于 PLMs 的生成[17-18]。该技术旨在将 PLMs 适应于电子病历的摘要任务[19]。相关研究大致可分为两类。第一类采用端到端方法[20]。例如，HET[21]模型利用层次化标签，通过词元和话语级编码器识别医疗对话中的关键话语，从而增强端到端模型的效果。一些研究者在此基础上尝试将领域特定知识(如标准化医学实体[11]和语义类型[22])整合进模型。第二类基于流水线方法[23-25]。例如，Cluster2Sent[26]模型识别电子病历各部分的关键话语，将相似话语分组，并为每个组生成一个总结句。此外，有研究[9]提出了一种多阶段框架，该框架先从部分对话中创建摘要，随后将这些摘要重写为综合性总结。然而，领域特定标注数据的数量有限，并且质量普遍不高，这对高性能模型的训练构成了重大挑战，可能导致生成的电子病历质量较低。

2)基于 LLMs 的生成。鉴于 LLMs 在生成能力方面的先进性，一些研究者开始探索其在电子病历摘要任务中的应用。这一过程采用了少量示例示范[13,27]和指令微调[14,28]等技术，以提升 LLMs 的性能。例如，Giorgi[16]等利用 GPT-4 和少量示例示范，实现了从医生与患者的对话中自动生成电子病历的功能。该方法基于余弦相似度选择与输入对话相似的训练样本作为示例。Nair 等[29]也采用类似方法，通过患者的年龄、性别和查询点选择示例。此外，Van Veen[30]等探索了一种不同的模型定制方法，实施 QLoRA[31]参数高效微调方法，利用大量问答数据将 LLMs 调整至医疗领域。然而，第一种方法依赖于少量正样本引导 LLMs 生成医疗电子病历，面临着两个挑战：可能无法充分识别错误样本，且专门领域知识的整合通常不足。第二种方法则存在产生幻觉(即错误或虚构输出)的风险，因为它允许 LLMs 在单次迭代中生成输出，缺乏自我反思的机制。

3 总体概述

3.1 任务定义

本研究旨在生成一个总结医生与患者对话的电子病历。该病历由一个或多个部分组成，例如“主诉”和“现病史”。值得注意的是，这一任务超出了常见的开放域文本生成，主要是由

于内容中存在显性规范和隐性规范。显性规范要求每个部分的内容必须严格遵循官方医学文件中的专业指南，包括内容范围限制和书写规范。例如，“主诉”部分应详细描述症状、症状部位和持续时间，避免使用模糊的时间描述，如“一两天”。隐性规范则指对话中口语语言与电子病历中使用的正式语言之间存在显著的风格差异，尤其是在中文电子病历中。例如，对话中的“时不时咳嗽”需要转换为电子病历中的“阵发性咳嗽”。

3.2 算法框架

如图 1 所示，本研究的电子病历摘要解决方案主要包括以下 4 个步骤。

1）初始生成：针对医患对话，采用大语言模型（如 GPT-4），通过指令和示例指导模型生成初步摘要，生成完整的电子病历。

2）章节反思：设计了规则提示和 Error2Correct 示例，以优化生成的电子病历中每个章节的内容，确保其符合明确和隐含的规范。

3）反思结果评分：采用经过领域数据微调的小规模医学预训练语言模型，对 LLMs 生成的优化结果进行质量评估。

4）迭代调度：将当前迭代的输出（在正常循环中）或所有先前迭代中得分最高的输出（在死循环中）作为下一次迭代的输入。当满足预定义条件或达到迭代限制时，过程终止；否则，将继续按顺序迭代执行步骤 2）和步骤 3）。

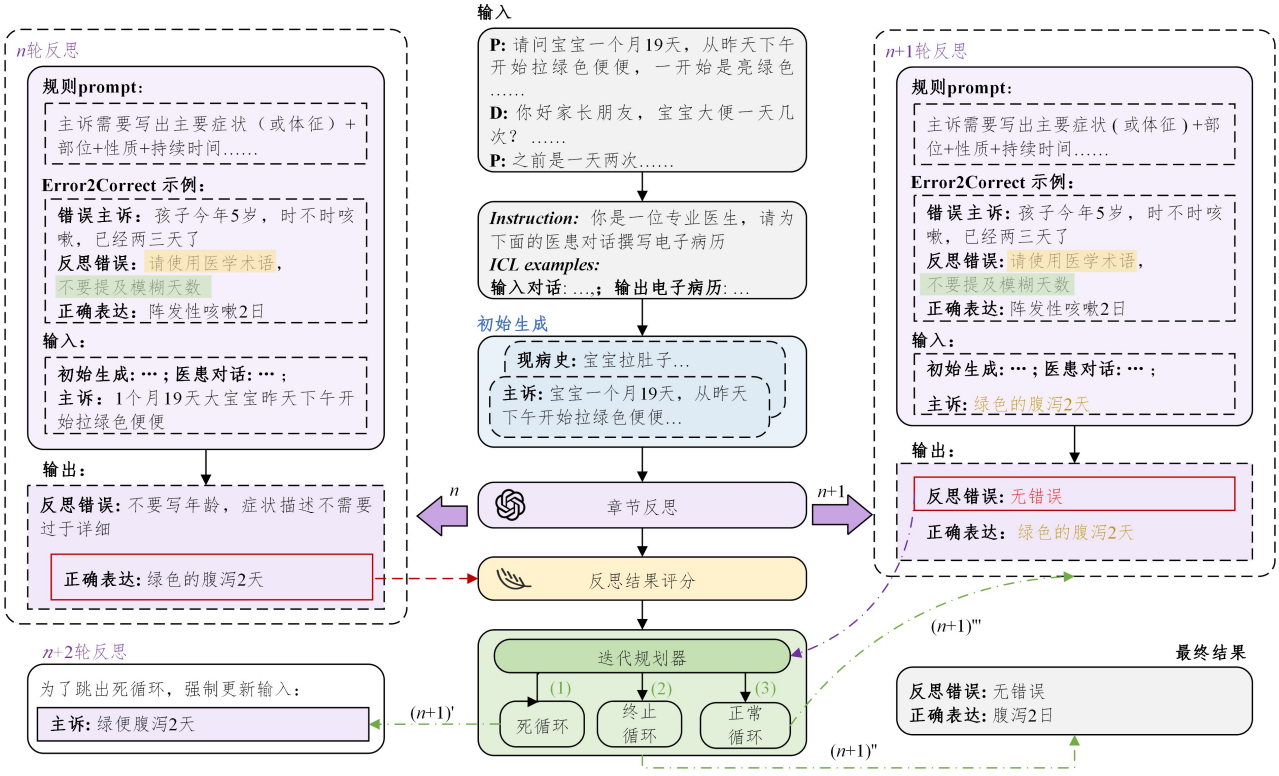

图 1 基于反思框架的电子病历生成流程（电子版为彩图）

Fig. 1 Framework of REFLEXES

4 REFLEXES

4.1 初始生成

在医患对话的基础上，初始生成旨在生成完整的初步电子病历。由于初步结果的质量对后续反思步骤有显著影响，采用能力较强的 GPT-4 模型进行初始生成。除了对话内容，这一环节还向 GPT-4 引入了两个额外输入：自然语言指令和示例学习（ICL）示例。指令表述为：“您是一位专业医生，请根据给定的医疗对话生成电子病历。”对于 ICL 示例，常见做法是从训练数据中挑选与输入对话相似的对话及相应的电子病历。然而，这种方法往往会导致模型在输出中复制这些示例的片段[32]。为了解决这个问题，采用了一种简单而有效的随机抽样策略，确保模型专注于基于输入对话生成摘要。

4.2 章节反思

在生成初始电子病历后，便是优化笔记中每个章节的内容。将每个章节视为独立单元，并采用相同的优化流程。为了便于说明，以“主诉”章节为例描述具体的优化过程。此步骤中，使用的大模型为 ChatGPT 而非 GPT-4，若使用 GPT-4 的迭代优化过程可能导致成本过高。生成的笔记应符合官方医学文件的明确指导方针和写作风格的隐含规范。因此，除了“主诉”的初始化内容及相关输入对话外，该步骤还引入了两个附加部分，如图 1 所示。第一部分是关于“主诉”的明确指导方针的规则提示，来自于卫健委发布的电子病历规范文件。第二部分是手动预定义的 Error2Correct 示例，其作为重要参考，确保模型遵循明确（图 1 中绿色背景的文本）和隐含（图 1 中黄色背景的文本）要求。每个示例均包括 3 个部分：1）一个违反“主诉”约束的例子（如“孩子已经咳嗽一两天了”）；2）一个 ErrorPrompt，用于提供对上述错误示例问题的描述性解释（如“避免模糊的疾病持续时间”）；3）错误示例的正确表达（如“患儿咳嗽 2 天”）。基于这些示例，LLMs 将生成与输入相应的 ErrorPrompt 和正确表达。

4.3 反思结果评分

这一步骤旨在评估在前一步中反思生成的每个章节内容的质量。输入为与章节相关的对话及其反思后的内容，输出为该内容的标量评分。本环节采用一个开源的医学预训练语言模型作为内容评分器的基础，并于此模型后额外添加一个线性层来输出评分。在模型微调过程中，本框架采用了类似于 InstructGPT[33] 中奖励模型的损失函数，如式(1)所示：

$$loss(\theta)=-E_{(y_g,y_n)\sim D}[\log(\sigma(r_\theta(x,y_g)-r_\theta(x,y_n)))] \tag{1}$$

在医疗领域中，评分函数 $r(\theta)(x,y)$ 输出一个标量值，用于评估由相关对话 x 和其摘要内容 y 组成的对。其中，x 是由 LLMs 根据章节指南和示例从原始对话中选择的，要求模型仅选择而不修改对话内容。集合 D 包含真实章节内容对(y_g)和负样本(y_n)。负样本是从初始生成的内容中选择的，选择依据两个标准：1)初始化内容与真实内容在字符串匹配上的低相似度；2)它们长度的最小差异。在推理阶段，经过训练的评分函数 $r(\theta)(\cdot)$ 用于评估生成的内容，得分越高，质量越好。

4.4 迭代规划器

该部分旨在增强 LLMs 的迭代反思能力，确保各章节内容的持续反思优化，因为单次反思往往不足以达到理想效果。该环节包含的迭代调度器重点作用于 3 个关键层面：为下一次迭代定义输入，制定策略避免死循环，以及确定迭代的终止条件。

在下一次迭代中，本框架将当前迭代中电子病历反思后的章节内容作为输入的一部分，并整合 4.1 节中的初始结果以及 4.3 节中与章节相关的对话。此外，在大模型的输入中添加了一条自然语言指令，用于引导 LLMs 在反思过程中参考这些额外的输入。此方法有助于防止 LLMs 在迭代过程中出现语义漂移，降低潜在幻觉风险。

在迭代过程中，模型存在陷入死循环的风险，从而无法生成最佳结果。在该环节中，存在两种类型的死循环：1)Error-Prompt 显示结果为“无错误”或“无需修改”，且出现次数最多，但是该结果在评分模型打分的所有生成结果中并不位于前 K 个；2)连续几次迭代中的反思生成结果保持一致，且其评分在所有轮次结果的得分最高的前 K 个之外。当出现死循环的情况时，模型会选择之前所有迭代中最佳的输出作为下一次迭代的输入，而不是使用当前迭代的输出。此外，为避免死循环的重复，本框架维护一个黑名单。当输出与黑名单中的条目匹配时，本框架会对提示做出微调整，例如增加换行符或调整示例的顺序，以确保模型生成多样化的结果。在结合大模型的医疗领域应用中，迭代反思过程有 3 个终止标准。

1)由大模型生成的错误提示(ErrorPrompt)中包含诸如“无错误”或“无需修改”等信号，同时，出现次数最多的结果在评分模型排名的 Top-K 中。如果在所有迭代中存在少于 K 个不的同结果，则对提示稍作修改以继续反思。

2)大模型在多次迭代中产生一致的结果，并且这些结果也位列 Top-K。

3)模型达到预设的最大迭代次数上限 N。

5 实验

本章首先在两个公共数据集上进行了广泛的实验，以评估所提方法的有效性；然后对该方法进行了详细的分析；最后提供了领域专家对该方法生成结果的评估意见。

5.1 实验设置

5.1.1 数据集

本实验使用了两个数据集：来自 CBLUE 基准测试的中文数据集 IMCS-V2-MRG[34] 和英文数据集 ACI-BENCH[35]。IMCS-V2-MRG 中的电子病历分为 6 个部分：“主诉”“现病史”“辅助检查”“既往史”“诊断”和“建议”。由于测试集中缺少真实标注数据，从验证集中选择前 200 个样本作为测试集，其余样本作为验证集。ACI-BENCH 中的电子病历则划分为 4 个部分：“主观”“客观检查”“结果”和“评估与计划”。统计数据如表 1 所列。

表 1 数据集详细信息

Table 1 Details of datasets

	IMCS-V2-MRG			ACI-BENCH		
	训练集	验证集	测试集	训练集	验证集	测试集
样本数	2039	524	200	60	20	120
对话平均长度	640	624	769	6443	6124	6582
病历平均长度	88	84	111	2649	2716	2703
对话平均轮数	13.83	15.21	13.48	26.69	26.17	24.00

5.1.2 对比基线

基于上述两个数据集，将所提方法与 3 类基线进行比较：1)在任务数据集上微调的预训练语言模型(PLMs)；2)在任务数据集上进行指令微调的开源大语言模型(LLMs)；3)使用 ICL(示例学习)或链式思维(Chain-of-thought，CoT)策略的闭源大语言模型。具体细节如表 2 所列。

表 2 在 IMCS-V2-MRG 数据集上的实验结果

Table 2 Experiment result on IMCS-V2-MRG dataset

Method	Rouge-1	Rouge-2	Rouge-AVG	Meteor	BertScore	All-AVG
BART+FT	51.13	32.58	41.86	22.49	75.48	45.42
T5-Pegasus	52.69	33.80	43.25	23.72	76.44	46.66
$ERNIE_h$+BART	51.26	32.87	42.07	22.26	75.64	45.54
$ERNIE_h$+T5	53.95	35.12	44.54	24.44	76.80	47.58
IDEA-CCNL	55.18	39.71	47.45	26.44	78.49	49.96
Qwen1.5-7B	53.60	34.31	43.96	35.40	78.03	50.34
ChatGLM3-6B	52.04	36.91	44.48	34.68	78.20	50.46
SumCoT	49.76	31.26	40.51	30.08	77.63	47.18
ChatGPT+ICL	51.18	32.51	41.86	35.07	75.64	48.61
GPT4+ICL	52.13	33.72	42.93	37.75	75.71	49.83
REFLEXES(Ours)	**58.42**	**39.86**	**49.14**	**39.44**	**78.75**	**54.12**

在 IMCS-V2-MRG 数据集上，用于比较的预训练语言模型(PLMs)主要包括 BART-Base-Chinese 和 T5-Pegasus，两者均在训练集上进行了微调。该对比实验首先使用 ERNIE-HEALTH[36]识别与不同部分相关的对话轮次，然后使用 BART 和 T5-Pegasus 为每个部分生成电子病历(分别记为 ERNIE-H＋BART 和 ERNIE-H＋T5)。对比模型还包括在多个中文摘要任务中表现优异的 IDEA-CCNL[37]。对于开源大语言模型(LLMs)，选择了在从训练集中转换的指令数据上微调的 Qwen1.5-7B 和 ChatGLM3-6B 作为基线。对于闭源大模型，评估了使用输入对话相似度最高的 ICL 示例的 ChatGPT 和 GPT-4，分别记为 ChatGPT＋ICL 和 GPT-4＋ICL。此外，还比较了 SumCoT[38]，该模型通过逐步提示生成摘要。

在 ACI-BENCH 数据集上，同样选择了 BART[5]作为基线模型，并包含两个变体：BioBART[39]和经过微调的 BART(BART＋FT)。其中，BioBART 在 PubMed 摘要[40]上进行预训练，BART＋F 则是在 SAMSum 语料库[41]上进行微调。此外，将本文方法与在任务相关数据上微调的开源大语言模型进行比较，包括 Mistral-7B，Vicuna-7B[42]，Llama3-8B 和 SCORE-INSTRUCT[43]。最后，评估闭源大语言模型 ChatGPT＋ICL 和 GPT-4＋ICL，以及较新的方法如 SumCoT。

5.1.3 评价指标

实验采用 6 个评价指标：Rouge-1，Rouge-2，Rouge-Avg(Rouge-1 和 Rouge-2 的平均值)，Meteor，Bertscore 以及 All-Avg(除 Rouge-Avg 外所有指标的总体平均值)。一个优秀的模型方法需要在这些指标上都取得较高的分数。

5.1.4 实验配置

在 IMCS-V2-MRG 数据集上，将本文方法应用于“主诉”“现病史”和“建议”部分，因为其他部分可以轻松总结或留空。为了确保实验的公平性，在评估过程中对所有部分计算指标，对于其他部分(辅助检查、既往史、诊断)，则使用初始摘要作为结果。本实验设置了 2 个示例，并将温度参数设为 0.2，保持 OpenAI API 的其他超参数为默认值。为了确保结果的可靠性，进行了 3 次重复实验，并报告了平均结果。对于内容评分模型，采用 ERNIE-HEALTH-CHINESE[36]，该模型使用 80 个训练样本和 10 个验证样本进行微调，以选择最佳参数。

在 ACI-BENCH 数据集上，本实验计算了“主观”部分中“现病史”以及“评估与计划”部分的反映结果，其余部分则使用初始结果作为评价依据。评分模型采用 LED-PubMed，这是一种医学预训练语言模型，本文使用 40 个训练样本和 10 个验证样本对其进行微调。将 K 值设置为 2，以应用于中英文数据集。为了公平比较，所有小规模语言模型的基线模型都在训练集上进行了微调。

5.2 主实验结果与分析

为了评估本文方法的有效性，在中文数据集 IMCS-V2-MRG 和英文数据集 ACI-BENCH 上，将其与之前提到的 3 类方法进行了比较。实验结果分别如表 2 和表 3 所列。

表 3 在 ACI-BENCH 数据集上的实验结果

Table 3 Experiment result on ACI-BENCH dataset

Method	Rouge-1	Rouge-2	Rouge-Avg	Meteor	Bertscore	All-Avg
BART	49.19	20.84	35.02	35.45	63.02	42.13
BioBART	45.81	18.40	32.11	31.09	62.73	39.51
BART＋FT	47.25	19.08	33.17	33.85	61.63	40.45
Mistral-7B	46.74	22.17	34.46	32.49	67.50	42.23
Vicuna-7B	41.82	18.29	30.06	28.70	68.53	39.34
Llama3-8B	48.32	22.73	35.53	35.98	69.17	44.05
SCORE-INSTRUCT	48.12	22.45	35.29	35.50	68.81	43.72
SumCoT	29.09	9.02	21.90	19.06	53.12	27.62
ChatGPT＋ICL	45.27	18.31	31.79	27.42	65.02	39.01
GPT-4＋ICL	51.10	21.91	36.54	33.83	67.69	43.65
REFLEXES(Ours)	**59.36**	**29.28**	**44.32**	**46.45**	**72.12**	**51.80**

根据表中的数据，得出以下结论：

1)本文方法在所有指标上都达到了最先进(State-of-the-Art，SoTA)的性能，其有效性得到验证。特别地，该方法在 IMCS-V2-MRG 数据集上相较于当前先进的模型 ChatGLM3-6B(经过微调)整体平均性能提升了 3.66 个百分点；在 ACI-BENCH 数据集上，相较于目前先进的 Llama3(经过微调)整体平均性能提升了 7.75 个百分点。

2)与开源的大语言模型(如 SumCoT，ChatGPT＋ICL 和 GPT-4＋ICL)相比，REFLEXES 显示出显著的性能提升。具体来说，在 IMCS-V2-MRG 和 ACI-BENCH 数据集上，相较于 GPT-4＋ICL，REFLEXES 在 All-Avg 上分别提升了 4.29 个百分点和 8.15 个百分点。这表明了本框架使用“Error2Correct”示例和领域模型监督的反思框架的有效性。

3)尽管闭源的大语言模型在所有基线比较中表现最好，但本文方法仍表现出相对于它们的优势。这表明，这些模型即使在任务特定数据上进行了微调，仍无法超过 REFLEXES 的性能。可能的原因是，闭源模型在生成输出时采用单次迭代，缺少自我反思的过程。主实验结果如表 2 所列。

5.3 分析消融实验

本节首先对 IMCS-V2-MRG 数据集上迭代反思框架的重要组成部分进行了详细分析，包括“Error2Correct”示例和领域模型监督；然后对所提框架进行了收敛性分析；最后通过案例研究，展示了 REFLEXES 相较于当前先进(SoTA)方法的优越性。

5.3.1 “错误-正确”示例分析实验

REFLEXES 框架提出了使用“错误-正确”示例来指导 ChatGPT 生成高质量的电子病历，为了验证这一策略的有效性，设计了两个实验。实验 1(变体 1)：使用了两个“Error2Correct”示例，但省略了“ErrorPrompt”，仅给出一个错误表达和与之对应的正确表达。实验 2(变体 2)：仅使用了

一个“Error2Correct”示例。实验结果如表 4 所列。从结果中可以观察到，当省略“ErrorPrompt”时，LLMs 生成的电子病历质量显著下降，这强调了为大语言模型提供错误原因的重要性。具体来说，变体实验 1 相较于本文方法，整体性能下降了 1.94 个百分点。此外，变体实验 2 减少示例数量也会对模型性能产生不利影响，这表明多个示例可以提供更广泛且有价值的参考。然而，考虑到模型输入长度的限制和成本因素，本实验将示例数量限制为 2 个。

表 4 Error2Correct 示例与领域模型监督分析实验

Table 4 Analysis of Error2Correct demonstrations and domain-model supervision

	Method	Rouge-1	Rouge-2	Rouge-Avg	Meteor	Bertscore	All-Avg
Error2Correct Demonstrations	Variant 1	56.24	36.81	46.53	38.17	77.50	52.18
	Variant 2	56.29	37.23	46.76	38.50	77.38	52.35
Domain-model Supervision	Variant 3	57.69	38.28	47.89	39.33	77.99	53.27
	Variant 4	57.57	38.51	48.04	39.43	78.10	53.40
	Variant 5	57.66	38.69	48.18	39.50	78.23	53.52
	REFLEXES(Ours)	**58.42**	**39.86**	**49.12**	**39.44**	**78.75**	**54.12**

5.3.2 小规模领域模型分析实验

为了验证领域模型监督是否必要，进行 3 个实验。1)实验 3(变体 3)：在生成初始结果后，对每个部分的内容进行一次反思，但不使用领域模型的监督。2)实验 4(变体 4)：在生成初始结果后，对每个部分的内容进行 5 次迭代反思，同样不使用领域模型监督，并选择最后一次迭代反思作为最终结果。3)实验 5(变体 5)：在初始结果之后，继续反思每个部分的内容，直到获得 3 个不同的结果。在迭代过程中，不涉及领域模型。完成迭代后，使用领域模型对每个结果进行评分，并选择得分最高的结果。实验结果如表 4 所列。结果显示，无论是反思迭代一次还是多次，缺乏领域模型监督的结果在大多数指标上都明显不如 REFLEXES，这凸显了领域模型监督的重要性。具体而言，变体 4 相较于 REFLEXES，性能下降了 0.72 个百分点。此外，使用领域模型在生成后选择最佳答案的效果不如将其评估过程整合到迭代过程中，这导致采用该方案的实验结果相比 REFLEXES 指标平均值下降了 0.6 个百分点。主要原因有两个：1)高质量的结果可能在第四次迭代或更晚出现，而在这个过程中可能会出现死循环；2)使用小模型来避免这些循环，可以使大语言模型生成更多样化且更高质量的结果。

为证明所设计的小规模领域模型的有效性，将其与多个基线模型进行比较分析，包括 ChatGLM，ChatGPT 和 GPT-4。如果某个部分的真实值得分超过其初始生成结果的得分，将其分类为正确；否则，标记为错误。实验结果如图 2 所示。

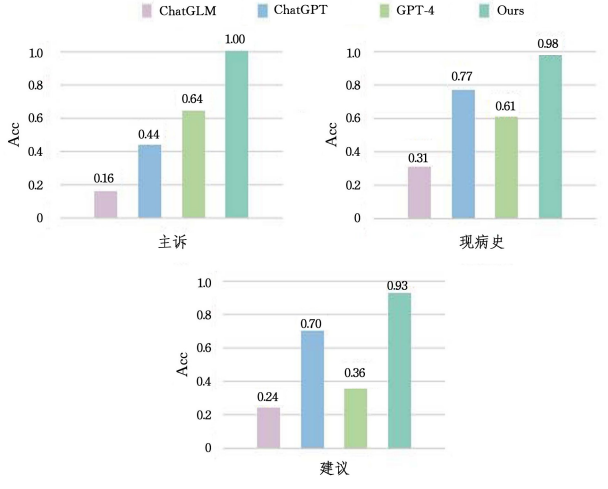

图 2 不同监督模型的结果对比

Fig. 2 Comparison results of different supervised models

从图 2 中可以看出，本框架中采用的领域模型在准确性方面优于所有竞争对手。这表明，用领域数据训练的小规模语言模型在电子病历摘要任务中比开源领域的 LLMs 更适合作为评分专家。尤其是，该模型对各个章节的质量判别的准确性几乎达到了 100%，甚至在“主诉”部分完全达到了 100%，而较大的模型在取得相似结果方面明显落后。值得注意的是，ChatGPT 在“现病史”和“建议”部分的表现优于 GPT-4，但在处理“主诉”时表现不佳。这种性能差异可能与每个模型使用的训练数据存在差异有关。

5.3.3 迭代轮数分析

为了分析本框架中的收敛迭代次数(记为 C-iterations)，记录了每个测试样本的迭代次数，结果如图 3 所示。

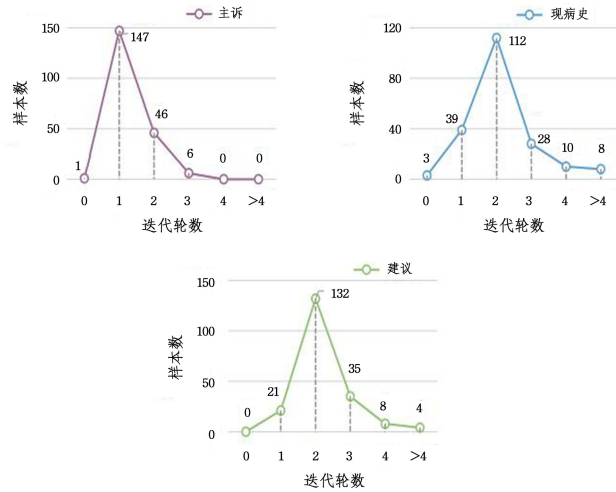

图 3 各章节最优结果出现所需的反思迭代轮数(C-iterations)

Fig. 3 Number of reflective iterations needed for optimal results to emerge in each section(C-iterations)

可以看出，“主诉”部分生成的结果通常在第一次或第二次迭代内收敛，而“现病史”和“建议”部分的结果则通常在第一、第二或第三次 C-iteration 时达到收敛，有时甚至需要第四次或更多轮次。这种现象的出现是因为“主诉”部分的输入和输出相对简短，而其他两部分的内容较长。所以，在反思过程中设定固定的迭代次数并不是一种实际的做法。相反，小规模领域模型进行监督，使本文的反思框架能够自适应地选择适当轮次中的最佳结果。

5.3.4 实例分析

本节中提供了几个案例，以直观了解当前较先进的ChatGLM3模型和REFLEXES在IMCS-V2-MRG数据集上的表现，以及实际的标准答案。这些案例如图4所示。

在“主诉”部分，ChatGLM3生成了模糊的时间信息，如“过去3或4天”，而REFLEXES的生成结果则明确为“4天”，提高了精准度。此外，本文框架有效去除了电子病历中的冗余信息，使结果更加简洁。在“现病史”部分，REFLEXES反思框架能够将非正式语言（如“鼻子不通气”）转换为更加正式且医学上准确的术语（“鼻塞”），这与当前先进方法不同。在“建议”部分，ChatGLM3的输出缺乏关键细节，而REFLEXES不仅包含了这些细节，甚至提供了比参考标准更详细的信息，展示了其在捕捉关键信息方面的优越性。

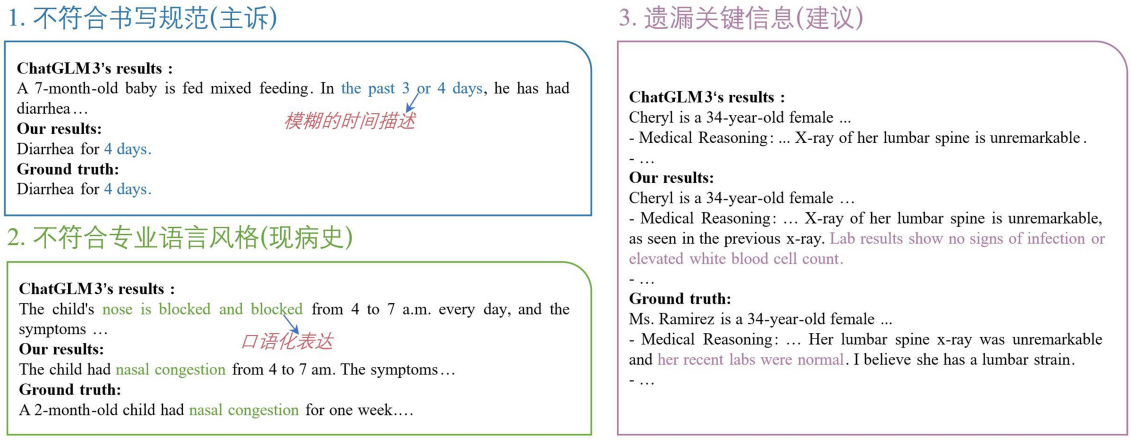

图4 案例分析

Fig. 4 Case studies

5.4 人工评估

与之前的研究[14]类似，本研究邀请了3位医学领域专家对IMCS-V2-MRG测试集中随机选取的50个对话进行质量评估。对于每个对话，本研究提供了由GPT-4，ChatGLM3（表2中显示其为该数据集上的最佳竞争模型），REFLEXES框架以及基准参考摘要生成的总结。为了确保公平比较，提供给评审人员的摘要并未显示其生成模型的来源。评估因素包括以下5方面：1）流畅性，评估生成文本的流畅度；2）相关性，衡量生成文本与对应部分的相关程度；3）完整性，检查是否缺失任何关键细节；4）幻觉，识别任何不准确或虚构的内容；5）风格，确保内容符合医学领域的专业写作标准。3位领域专家评估的平均结果如表5所列。

表5 人工评估结果

Table 5 Human evaluation result

Method	Fluency	Relev.	Comple.	Halluc.	Sysle
Reference	4.895	4.905	4.805	4.940	4.980
GPT4+ICL	4.800	4.545	4.845	4.905	4.270
ChatGLM3	4.275	4.750	4.765	4.785	4.775
REFLEXES(Ours)	4.835	4.845	4.835	4.940	4.850

从表5中得出以下结论：

1）在流畅性指标上，本文方法和GPT-4+ICL的表现与参考标准相当，表明基于GPT-4的方法能够生成连贯的语言。

2）在相关性方面，本文方法与参考标准非常接近，并且超过了GPT-4+ICL和ChatGLM3。这归功于“Error2Correct”示例和领域模型监督，其有助于生成的内容与显性和隐性标准对齐。

3）在完整性和幻觉指标上，本文方法和GPT-4+ICL均与参考标准一致，反映了基于GPT-4的方法具备生成详细且准确的摘要的能力。

4）关于风格，本文方法得分与参考标准相当，而GPT-4+ICL的得分明显较低，甚至落后于ChatGLM3。这表明，本文方法在隐性风格要求的遵循上有显著改进，这是GPT-4在处理较长文本输入时可能忽略的一个方面。相反，ChatGLM3这类经过训练的领域模型往往能更好地理解这些要求。

5）对于闭源模型ChatGLM3，尽管其经过大量数据（2039个样本）的微调，但其性能在所有评估指标上仍低于本文方法。

结束语 本文提出了一种新的迭代反思框架REFLEXES，该框架结合了“错误-正确示例”和领域模型监督，用于生成电子病历。其核心思想是将生成电子病历的任务交给LLMs，而一个经过领域特定数据微调的小规模医学预训练语言模型则负责评估生成的内容。为了提升LLMs在生成高质量结果方面的性能，设计了“Error2Correct”示例，包括错误示例、错误分析以及相应的正确表达，以赋予LLMs有效检测和纠正错误的能力。在中英文数据集上进行的实验表明，REFLEXES达到了最先进（SoTA）的效果。此外，详细的分析和人工评估也验证了REFLEXES方法的有效性。

参 考 文 献

[1] LIU Z J, WANG X L, CHEN Q C, et al. Temporal indexing of medical entity in Chinese clinical notes[J]. BMC Medical Informatics and Decision Making, 2019, 19: 1-11.

[2] YU H Y, ZUO X L, TANG J T, et al. Identifying causal effects of the clinical sentiment of patients' nursing notes on anticipated fall risk stratification[J]. Information Processing & Management, 2023, 60(6): 103481.

[3] LU X T, SUN L P, LING C, et al. Named Entity Recognition of

Chinese Electronic Health Records Incorporating Phonetic and Part-of-speech Features[J]. Journal of Chinese Computer Systems，2025，46(2)：330-338.

[4] LIU S S，NIE W J，GAO D F，et al. Clinical quantitative information recognition and entity-quantity association from Chinese electronic medical records[J]. International Journal of Machine Learning and Cybernetics，2021，12：117-130.

[5] LEWIS M. Bart：Denoising sequence-to-sequence pre-training for natural language generation，translation，and comprehension[J]. arXiv：1910.13461，2019.

[6] RAFFEL C，SHAZEER N，ROBERTS A，et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. Journal of Machine Learning Research，2020，21(140)：1-67.

[7] NI H Q，LIU D，SHI M Y. Semantic-aware Chinese Short Text Summarization Model[J]. Computer Science，2020，47(6)：74-78.

[8] XI T J，DUAN Z T，CAO J R，et al. Hybrid Summarization Method for Legal-Related Long Texts in Public Opinion Information[J]. Journal of Chinese Information Processing，2024，38(7)：63-72.

[9] ZHANG L，NEGRINHO R，GHOSH A，et al. Leveraging pretrained models for automatic summarization of doctor-patient conversations[J]. arXiv：2109.12174，2021.

[10] KRISHNA K，KHOSLA S，BIGHAM J P，et al. Generating SOAP notes from doctor-patient conversations using modular summarization techniques[J]. arXiv：2005.01795，2020.

[11] JOSHI A，KATARIYA N，AMATRIAIN X，et al. Dr. summarize：Global summarization of medical dialogue by exploiting local structures[J]. arXiv：2009.08666，2020.

[12] MICHALOPOULOS G，WILLIAMS K，SINGH G，et al. MedicalSum：A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations [C]// Findings of the Association for Computational Linguistics：EMNLP 2022. 2022：4741-4749.

[13] LU G L，JU X L，CHEN X，et al. GRACE：Empowering LLM-based software vulnerability detection with graph structure and in-context learning[J]. Journal of Systems and Software，2024，212：112031.

[14] WANG L F，ZHAO M，JI H R，et al. Dialogue summarization enhanced response generation for multi-domain task-oriented dialogue systems[J]. Information Processing & Management，2024，61(3)：103668.

[15] DU Z X，QIAN Y J，LIU X，et al. Glm：General language model pretraining with autoregressive blank infilling[J]. arXiv：2103.10360，2021.

[16] GIORGI J，TOMA A，XIE R，et al. Clinical note generation from doctor-patient conversations using large language models：Insights from mediqa-chat[J]. arXiv：2305.02220，2023.

[17] ZHOU W，WANG Z Y，WEI B. Generative Automatic Summarization Model for Legal Judgments[J]. Computer Science，2021，48(12)：331-336.

[18] KONG Y L，WANG Z Q，WANG H L. Research on Comment Summarization Combined with Evaluation Object Information [J/OL]. Computer Science，1-8 [2024-10-16]. http：//kns.cnki.net/kcms/detail/50.1075.TP.20241012.0929.010.html.

[19] GAO Y J，MILLER T，XU D F，et al. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models[C]// Proceedings of COLING. International Conference on Computational Linguistics. NIH Public Access，2022：2979.

[20] ENARVI S，AMOIA M，TEBA M D A，et al. Generating medical reports from patient-doctor conversations using sequence-to-sequence models[C]// Proceedings of the First Workshop on Natural Language Processing for Medical Conversations. 2020：22-30.

[21] SONG Y，TIAN Y H，WANG N，et al. Summarizing medical conversations via identifying important utterances[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020：717-729.

[22] MICHALOPOULOS G，WILLIAMS K，SINGH G，et al. MedicalSum：A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations [C]// Findings of the Association forComputational Linguistics：EMNLP 2022. 2022：4741-4749.

[23] CAI P S，LIU F，BAJRACHARYA A，et al. Generation of patient after-visit summaries to support physicians[C]// Proceedings of the 29th International Conference on Computational Linguistics(COLING). 2022：6234-6247.

[24] WU R S，WANG H L，WANG Z Q，et al. Short Text Summarization Method Based on Global Self-matching Mechanism[J]. Journal of Software，2019，30(9)：2705-2717.

[25] HUANG Y X，YU Z T，GUO J J，et al. Case Topic Summarization Based on Topic Interaction Graph[J]. Journal of Software，2023，34(4)：1796-1810.

[26] KRISHNA K，KHOSLA S，BIGHAM J P，et al. Generating SOAP notes from doctor-patient conversations using modular summarization techniques[J]. arXiv：2005.01795，2020.

[27] TANG X R，TRAN A，TAN J，et al. Gersteinlab at mediqa-chat 2023：Clinical note summarization from doctor-patient conversations through fine-tuning and in-context learning[J]. arXiv：2305.05001，2023.

[28] LONGPRE S，HOU L，VU T，et al. The flan collection：Designing data and methods for effective instruction tuning[C]// International Conference on Machine Learning. PMLR，2023：22631-22648.

[29] NAIR V，SCHUMACHER E，KANNAN A. Generating medically-accurate summaries of patient-provider dialogue：A multi-stage approach using large language models[J]. arXiv：2305.05982，2023.

[30] VAN VEEN D，VAN UDEN C，BLANKEMEIER L，et al. Clinical text summarization：adapting large language models can outperform human experts[J]. Research Square，2023，30(4)：1134-1142.

[31] DETTMERS T，PAGNONI A，HOLTZMAN A，et al. QLORA：Efficient finetuning of quantized LLMs [C]// Proceedings of the

37th International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc., 2023: 10088-10115.

[32] LYU X, MIN S, BELTAGY I, et al. Z-icl: Zero-shot in-context learning with pseudo-demonstrations[J]. arXiv: 2212.09865, 2022.

[33] 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.

[34] CHEN W, LI Z W, FANG H Y, et al. A benchmark for automatic medical consultation system: frameworks, tasks and datasets[J]. Bioinformatics, 2023, 39(1): 817.

[35] YIM W, FU Y, BEN ABACHA A, et al. Aci-bench: a novel ambient clinical intelligence datasetfor benchmarking automatic visit note generation[J]. Scientific Data, 2023, 10(1): 586.

[36] WANG Q, DAI S T, XU B F, et al. Building chinese biomedical language models via multi-level text discrimination[J]. arXiv: 2110.07244, 2021.

[37] ZHANG J X, GAN R, WANG J J, et al. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence[J]. arXiv: 2209.02970, 2022.

[38] WANG Y, ZHANG Z, WANG R. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method[J]. arXiv: 2305.13412, 2023.

[39] YUAN H M, YUAN Z S, GAN R, et al. BioBART: Pretraining and evaluation of a biomedical generative language model[J]. arXiv: 2204.03905, 2022.

[40] COHAN A, DERNONCOURT F, KIM D S, et al. A discourse-aware attention model for abstractive summarization of long documents[J]. arXiv: 1804.05685, 2018.

[41] GLIWA B, MOCHOL I, BIESEK M, et al. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization[J]. arXiv: 1911.12237, 2019.

[42] ZHENG L M, CHIANG W L, SHENG Y, et al. Judging llm-as-a-judge with mt-bench and chatbot arena[J]. Advances in Neural Information Processing Systems, 2023, 36: 46595-46623.

[43] RIBEIRO L F R, BANSAL M, DREYER M. Generating summaries with controllable readability levels[J]. arXiv: 2310.10623, 2023.

ZHONG Boyang, born in 2000, postgraduate. His main research interests include natural language processing and vertical domain large language model.

LIU Jingping, born in 1991, lecturer, master supervisor. His main research interests include natural language processing and vertical domain large language model.

（责任编辑：柯颖）