基于思维链的跨语言多文档摘要生成技术研究

祁天[1]，杨建安[2]，赵铁军[1]，杨沐昀[1]
[1]哈尔滨工业大学 计算学部，中国 黑龙江 哈尔滨，150001
[2]东北大学 信息科学与工程学院，中国 辽宁 沈阳，110819
qitian@stu.hit.edu.cn，20206422@stu.neu.edu.cn，{tjzhao,yangmuyun}@hit.edu.cn

摘要

随着全球化的加速发展，跨语言信息的高效传递与理解变得尤为重要。传统的多文档摘要生成技术可以提升信息获取效率，然而往往忽视了跨语言场景下的特殊挑战。为了缓解这一问题，本文提出了跨语言多文档摘要生成任务。我们首先构建了一个全面的跨语言多文档摘要测试集作为评估基准，其次提出了一种基于思维链技术的跨语言多文档摘要生成方法，并对其进行了实验验证。在实验中，我们使用了几种典型的大语言模型，并通过人工评估和自动评估来验证我们的方法。结果表明，我们提出的基于思维链的方法在跨语言多文档摘要生成任务上取得了显著的性能提升，为解决语言障碍下的信息获取问题提供了有效的解决方案。

关键词： 多文档摘要；跨语言摘要；大语言模型；思维链

Cross-lingual Multi-document Summarization Based on Chain-of-Thought

Tian Qi[1], Jianan Yang[2], Tiejun Zhao[1], Muyun Yang[1]
[1]Faculty of Computing, Harbin Institute of Technology, Harbin 150001,China
[2]College of Information Science and Engineering, Northeastern University, Shenyang,110819, China
qitian@stu.hit.edu.cn, 20206422@stu.neu.edu.cn, {tjzhao,yangmuyun}@hit.edu.cn

Abstract

With the accelerating development of globalization, efficient transmission and understanding of cross-lingual information have become particularly important. Traditional multi-document summarization techniques can improve information retrieval efficiency but often overlook the challenges posed by cross-lingual scenarios. To address this issue, this paper proposes a task for cross-lingual multi-document summarization generation. We first construct a comprehensive cross-lingual multi-document summarization test set as an evaluation benchmark. Secondly, we propose a cross-lingual multi-document summarization generation method based on Chain-of-Thought technology and validate it through experiments. In the experiments, we use several typical large language models and verify our method through manual and automatic evaluation. The results show that our proposed method based on Chain-of-Thought significantly improves performance in cross-lingual multi-document summarization generation tasks, providing an effective solution for information retrieval issues under language barriers.

Keywords: Multi-Document Summarization , Cross-Lingual Summarization , Large Language Model , Chain-of-Thought

1 引言

随着全球化的加速和信息技术的飞速发展，人们面临着信息过载的挑战。在这个背景下，多文档摘要技术成为了解决信息过载问题的关键工具之一。多文档摘要旨在从一组相关文档中提取和整合关键信息，生成简洁、连贯的摘要，帮助用户快速理解和获取大量文本资料的精髓。随着互联网上多语言内容的不断增长，跨语言多文档摘要（Cross-Lingual Multi-Document Summarization, CLMDS）的需求日益增加，它不仅要求模型能够处理多文档信息，还要求能够跨越语言障碍，从不同语言的文档中提取和整合信息。

传统的多文档摘要方法在处理单一语言的文档集合时已取得一定的进展，但在跨语言环境下，它们面临着新的挑战。不同语言之间的语法结构、词汇表达和文化背景差异使得跨语言信息提取和整合变得更加复杂。此外，不同语言之间的资源分布不均，如英语等资源丰富的语言拥有大量的训练数据，而其他语言则可能缺乏足够的数据支持，这进一步加剧了跨语言多文档摘要的难度。

近年来，大规模语言模型（Large language models, LLMs）如GPT-3、Llama等在自然语言处理（NLP）领域取得了显著的成就，它们在单文档摘要任务中表现出色，显示出强大的语言理解和生成能力。然而，当应用于多文档摘要，尤其是跨语言场景时，这些模型仍然面临着理解文档间复杂关系、生成高质量摘要的挑战。为了克服这些挑战，本文提出了基于思维链（Chain-of-Thought, CoT）的跨语言多文档摘要生成方法，该方法通过模拟人类解决问题时的逻辑推理过程，引导模型逐步生成摘要，从而提高了处理复杂任务的能力。本文通过人工构建数据集分析了该方法在处理多文档关系、提高摘要质量方面的潜力和挑战，并探讨未来的研究方向。本文主要有以下三个贡献:

- 提出了跨语言多文档摘要任务。明确了该任务的定义，为解决语言障碍下的信息获取问题提供了新的研究方向。
- 构建了跨语言多文档摘要测试集。我们通过人工的方式，构建了一个全面的跨语言多文档摘要测试集，为研究人员在该任务上提供了一个可靠的评估基准。
- 提出了基于思维链的跨语言多文档摘要生成方法，并通过实验验证了该方法在跨语言多文档摘要生成上的性能优势。

2 相关工作

2.1 多文档摘要

多文档摘要（Multi-Document Summarization, MDS）是指从同一主题下的多篇文档中提取出一份能够表达这些文档核心内容的综合性摘要(黄文彬and 倪少康, 2017)。与单文档摘要（Single Document Summarization, SDS）相比，多文档摘要提供了基于某个主题的文档集的重要内容概述，考虑了单个文档的特有信息及文档之间的冗余信息，极大地提高了主题相似的多篇文档的处理效率。因此，多文档摘要成为自动文本摘要领域的一个热点。Wang et al. (2020)提出了异构图神经网络，专门用于提取多文档摘要。这种异构图包括代表单词、句子和文档的节点。基于共现单词节点建立句子节点和文档节点之间的连接，促进了对多文档之间关系的更全面表示。Liu et al. (2018)利用Transformer进行多文档摘要任务，从一系列的参考文献中生成英文维基百科文章。在此工作的基础上，Liu and Lapata (2019)提出了一个基于Transformer结构以层级方式编码文档的模型，通过注意力机制表示跨文档之间的关系，允许多个文档信息共享而不是简单拼接文本为一个平行化的长序列。Li et al. (2020)提出一个利用图结构学习跨文档关系的端到端模型结构，该模型结构可以学习字级别的嵌入也可以学习到句子级别的嵌入。为了从不同粒度级别得到更丰富的语义表示，Jin et al. (2020)为抽取式和生成式多文档摘要提出一个多粒度交互网络模型，该模型能够联合学习词级别、句子级别和文档级别三个粒度的表示。然而目前多文档摘要的研究集中于单语言上，即输入文档集合与输出摘要是同一种语言，在跨语言上的研究较少 (Pontes et al., 2020)。

基金项目：国家自然科学基金项目“知识驱动的文本生成模型研究”(62376075)

2.2 跨语言摘要

跨语言摘要（Cross-lingual summarization, CLS）是一项对源语言文档核心信息进行内容归纳，以目标语言的形式组织成摘要的任务。早期的工作通常集中在流水线方法上 (Leuski et al., 2003; Orăsan and Chiorean, 2008; Wan et al., 2010; Wan, 2011; Yao et al., 2015)，即先翻译再摘要或先摘要再翻译。随着深度学习技术在自然语言处理任务中的不断应用，许多研究者将研究重点转移到了端到端的跨语言摘要模型上。Zhu et al. (2019)使用round-trip翻译策略，构建了大规模的跨语言摘要平行语料库，并首次训练了基于神经网络的跨语言摘要模型。Zhu et al. (2020)提出了将Transformer模型与外部概率双语词典结合的融合翻译模式，降低了模型对于外部数据的过分依赖。Cao et al. (2020)使用一个多任务学习框架，共同学习摘要和上下文级别的对齐表示。Xu et al. (2020)采用多语言预训练的方法进行跨语言摘要任务，该方案利用大规模的预训练来提高跨语言摘要的质量。随着大模型研究的兴起，Wang et al. (2023b)探索了利用大模型引导生成跨语言摘要，通过使用各种提示引导大模型在端到端和流水线不同的范式上执行zero-shot生成跨语言摘要。然而，目前跨语言摘要的研究集中于单文档，针对多文档的跨语言摘要并没有系统性研究。

2.3 思维链

思维链（Chain-of-Thought，CoT）是将问题分解为一系列中间推理步骤的完整思考过程，形式上体现为输入→推理链→输出的映射 (Zhang et al., 2023)。这种方法通常比传统的直接推理更为有效。许多研究已经证明了思维链在广泛的领域中的有效性，包括算术推理、常识推理和符号推理 (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022)。Wang et al. (2023c)将思维链引入到摘要领域，提出了SumCoT，旨在引导大模型以逐步的方式生成摘要。首先指导大模型使用手动设计的问题提示从源文档中提取核心新闻要素。随后将提取的要素以及源文档中的额外细节整合起来，以产生全面且信息丰富的新闻摘要。

3 方法

3.1 问题描述

跨语言多文档摘要，是指从一组主题相近的多篇源语言文档中，生成以目标语言形式表达的、能够综合地概括这些文档核心内容的摘要。跨语言多文档摘要可以形式化描述为：给定一个多文档集合$D_{\text{src}} = \{d_{\text{src},1}, d_{\text{src},2}, ..., d_{\text{src},N}\}$，其中$N$是文档数量，$d_{\text{src},i} = \{s_{i,j}|j \in [1, M_{d_i}]\}$ 是用源语言编写的包含M_{d_i}个句子的文档，每个集合中的文档可能包含不同数量的句子。任务的目标是生成一个关于目标语言的多文档摘要S_{tgt}，使得该摘要可以用来概括整个源语言文档集合中的内容。

3.2 整体架构

图 1展示了本文提出的基于思维链的跨语言多文档摘要生成的整体架构。该方法包括三个主要阶段。首先，使用单文档摘要思维链（Chain1）对文档集合中的每一篇文档生成摘要；然后，通过多文档摘要思维链（Chain2）来组织多个文档之间的联系，生成关于源语言的多文档摘要；最后，通过跨语言思维链（Chain3）利用上述信息生成关于目标语言的跨语言多文档摘要。该方法可形式化为下面的公式，其中各参数的含义见后文：

$$CLMDS = \text{Chain3}(\text{Chain2}(\text{Chain1}(D; Q; A; P_1), P_2); L; P_3)$$

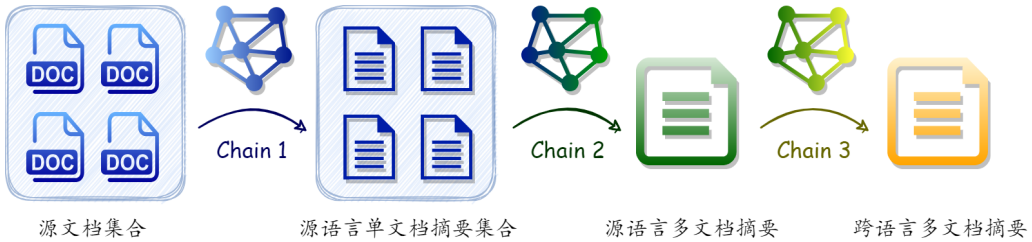

图 1. 基于思维链的跨语言多文档摘要生成整体架构

3.2.1 第一阶段：单文档摘要生成

基于思维链生成单文档摘要如图 2所示，将摘要任务拆分为以下两个主要步骤：

第一步是关键信息提取。针对学术论文的写作特点，我们创建了一系列引导性问题$Q = [q_1, q_2, q_3, q_4]$，以引导大模型提取论文中的核心内容，如研究对象、研究方法、研究结果等。假设源文档为D_i，输出的提取答案为A_i，则该步骤可形式化为：

$$A_i = LLM(D_i; Q)$$

第二步是多信息整合与摘要。在该步我们将提取的内容与源文档中更详细的信息进行整合，具体而言，将源文档、提取答案和整合生成摘要提示词P_1连接起来，以引导大模型生成该文档的摘要SDS_i，则该步骤可形式化为：

$$SDS_i = LLM(D_i; A_i; P_1)$$

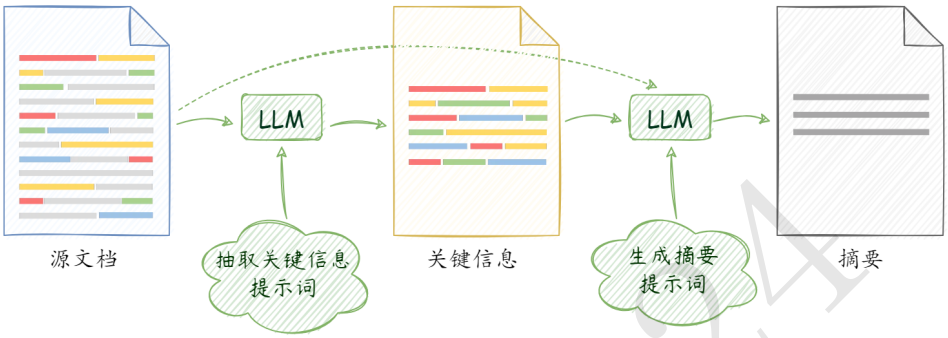

图 2. 单文档摘要生成思维链

3.2.2 第二阶段：多文档摘要生成

在本阶段，我们利用在第一阶段对每篇文档单独生成的摘要，通过思维链将这些摘要合并成一个整体的多文档摘要。合并过程可以采用不同的策略，例如简单直接地连接单篇文档的摘要后输入大模型；或者通过分层合并(Hierarchical merging)和增量更新(Incremental updating)的思维链策略进行生成。这两种策略已在书籍摘要领域进行了研究(Wu et al., 2021; Chang et al., 2023)，但目前尚未在多文档摘要领域进行深入探讨。

在层次化生成多文档摘要时，我们首先合并相邻两篇文档的摘要，然后逐层地合并摘要，直到只剩下一个摘要（整个文档集合的摘要），如图 3所示。这种方法类似于一种递归的过程，每一步都在合并低一层摘要以覆盖更多信息。层次化生成多文档摘要需要两个提示词引导大模型生成，一个用来合并单篇文档的摘要，另一个用来合并两个多文档集合子集的摘要。这种方法类似于将一个较大的任务分解成更小的部分，然后逐步合并这些部分来构建最终结果。这种策略的优点在于降低了模型处理的复杂性，与直接生成的方法相比，大幅降低了输入的上下文长度。我们将每篇文档的摘要作为叶子节点，构建初始的摘要节点集合$L = \{SDS_1, SDS_2, \ldots, SDS_n\}$，$P_2$为该阶段提示词，设BuildTree是一个递归函数，用于构建二叉树，则层次化生成多文档摘要的形式化公式可以表示为：

$$MDS = \text{BuildTree}(L, P_2)$$

在增量化生成多文档摘要时，我们首先生成一个全局摘要，然后逐步将每个单篇文档的摘要集成到这个全局摘要中，如图 4所示。这种方法类似于一种迭代的过程，每一步都在原有的摘要基础上进行更新，以适应新的信息。增量化生成多文档摘要需要两个提示词引导大模型生成，一个用来对初始的两篇文档生成一个全局摘要，另一个用来将新输入的单篇摘要合并到全局摘要中去。这种策略的优点在于可以更好地处理文档块之间的依赖关系，避免了一次性处理整个文档可能引入的不连贯性和错误。假设某文档集合中包含n篇文档，其中第i篇文档摘要为SDS_i，P_2为该阶段提示词，则增量化生成策略可形式化为：

$$MDS_0 = \emptyset, \forall i \in \{1, \ldots, n\}, MDS_i = \text{LLM}(MDS_{i-1}, SDS_i, P_2)$$

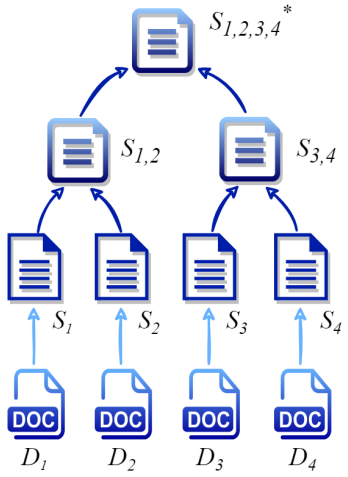

图 3. 层次化生成多文档摘要示意图

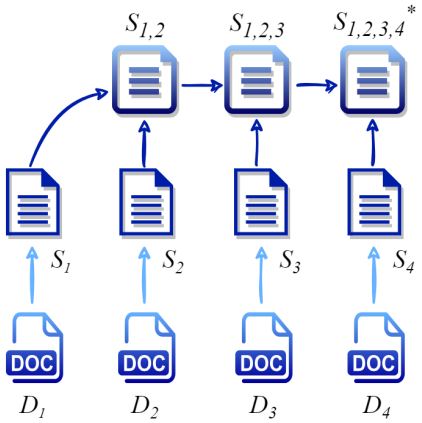

图 4. 增量化生成多文档摘要示意图

3.2.3 第三阶段：跨语言多文档摘要生成

在这一阶段，我们利用第二阶段生成的关于源语言的多文档摘要，通过跨语言多文档摘要思维链生成关于目标语言的跨语言多文档摘要。具体来说，我们将源语言的多文档摘要MDS、文档集合中每篇文章的摘要SDS_i与跨语言摘要生成的提示词P_3连接起来，以引导大模型生成该文档集合的跨语言摘要$CLMDS$，即：

$$CLMDS = LLM(MDS; L; P_3)$$

其中$L = \{SDS_1, SDS_2, \ldots, SDS_n\}$，$n$为文档集合长度。

4 实验

4.1 数据集

针对本文提出的跨语言多文档摘要任务，目前并没有现成的相关数据集可供使用。因此，本文精心构建了一个新的测试集，旨在评估我们提出的方法在这一任务上的表现。我们选择了医学和人工智能两个特定领域的学术论文作为文档来源。为了保证数据集的质量和代表性，我们通过关键词筛选、时间范围限定和质量评估从Pubmed医学论文数据库[-1]和AAAI人工智能论文集[0]中分别选择了一定数量的医学和人工智能相关论文。同时，在创建多文档摘要测试集时，我们通过人工筛选的方式确保簇内文章主题保持一致。考虑到输入上下文长度的限制，我们对于每篇学术论文只截取了部分关键内容作为源文档输入。我们将此数据集公开至github供相关研究使用[1]。

我们构建的测试集包括单文档和多文档、单语言和跨语言的摘要任务。在构建过程中，我们邀请了三位领域专家根据完整的写作协议编写了源文档的专业摘要。编写摘要时，特别强调了摘要的全面性、客观性和风格的一致性，以确保其质量和可靠性。测试集的相关统计指标如表 1所示，#Sentence、#Words、#Tokens分别表示句子、单词/中文字符和tokens的平均长度，DOC表示源文档或源文档集合，Sum-en表示英文参考摘要，Sum-zh表示中文参考摘要。

4.2 实验设置

为了验证本文提出的方法在跨语言多文档摘要任务上的有效性，我们选择了几个具有代表性的大语言模型进行实验：

- ChatGLM (Du et al., 2022)是智谱AI研发的一个开源的、支持中英双语的对话语言模型，基于General Language Model (GLM) 架构，具有6.2B参数。在本文中我们使用的是chatglm3-6b-32k模型 [2]。

[-1]https://pubmed.ncbi.nlm.nih.gov/
[0]https://ojs.aaai.org/index.php/AAAI/issue/archive
[1]https://github.com/yja2576/CLMDS-Dataset
[2]https://huggingface.co/THUDM/chatglm3-6b-32k

Task	Domain	#Docs	#Cluster	#Sentences			#Words			#Tokens		
				Doc	Sum-en	Sum-zh	Doc	Sum-en	Sum-zh	Doc	Sum-en	Sum-zh
SDS	AI	1.00	60	51.23	7.48	4.90	1,103.93	196.83	246.28	1,335.82	223.05	253.77
	Medical	1.00	61	18.33	4.15	4.23	474.69	161.29	320.57	630.52	197.39	375.31
	Overall	1.00	121	34.64	5.80	4.56	786.71	178.91	283.27	980.26	210.11	315.04
MDS	AI	2.52	50	125.36	7.34	7.34	2,712.72	261.96	496.78	3,274.94	296.22	514.42
	Medical	2.72	50	49.48	6.82	6.74	1,302.00	273.28	514.16	1,749.58	328.82	628.28
	Overall	2.62	100	87.42	7.08	7.04	2,007.36	257.62	505.47	2,512.26	312.52	584.80

Table 1: 本文构建的跨语言多文档摘要任务测试集的数据统计情况

- Qwen（通义千问）(Bai et al., 2023)是阿里云发布的大语言模型，参数规模从1.8B到72B，在本文中我们使用的是Qwen1.5-7B-Chat [3]模型。
- Llama-2 (Touvron et al., 2023)是Meta推出的一种Decoder-Only架构的大模型，模型参数范围从7B到70B。在本文中我们使用Llama-2的7B模型 [4]，当涉及跨语言任务时，我们使用了基于Llama-2的中文微调模型 [5]。
- GPT-3.5-turbo是OpenAI发布的GPT-3.5系列模型 (Ouyang et al., 2022)之一。该模型可以理解和生成自然语言或代码，并经过优化以在聊天中使用，但也适用于非聊天任务。本文使用的版本是*gpt-3.5-turbo-0125*。

为了全面评估性能情况，我们选择了多种评价指标。首先，我们考虑了基于参考摘要的指标。其中，ROUGE (Lin, 2004)用于衡量生成摘要与参考摘要之间的词重叠程度，我们主要关注ROUGE-1、ROUGE-2和ROUGE-L指标。另外，我们还采用了BERTScore (Zhang et al., 2019)），它利用BERT (Devlin et al., 2018)嵌入来衡量语义相似性。此外，在无参考摘要的情况下，我们利用GPT-3.5（gpt-3.5-turbo-0125）作为摘要评估器。GPT-3.5被证明是一个强大的自然语言生成评估器 (Jain et al., 2023; Wang et al., 2023a; Shen et al., 2023)，我们向大模型提供源文本和生成的摘要（适用于GPT-3.5的16k上下文窗口），并要求输出一个从1到10的分数，然后乘以10得到最终的分数。完整的提示模板见附录A。

同时，本文进行了人工评价，综合考虑以往的工作所评价的维度，我们从以下四个维度对生成的摘要进行了评估：

1. 流畅性（Fluency）：评估摘要的自然度和可读性。一个流畅的摘要应该具有正确的句子结构、无语法错误、用词恰当，读起来通顺。
2. 连贯性（Coherence）：评估摘要中信息的逻辑顺序。一个连贯的摘要应该具有清晰的结构，句子之间有合理的过渡，没有逻辑上的跳跃或混乱。
3. 一致性（Consistency）：评估摘要中的事实、观点和信息是否与原文一致。摘要应该准确地反映原文的内容，不包含原文中未提及的信息，也不包含错误的信息。
4. 相关性（Relevance）：评估摘要是否聚焦原文最重要的信息，排除不相关或次要的细节。

4.3 实验结果与分析

4.3.1 第一阶段：单文档摘要生成

我们通过Zero-shot和Chain-of-Thought两种生成方式生成单文档摘要，实验结果如表 2所示。在该表中，带有灰色背景的数据 表示Zero-shot和Chain-of-Thought两种生成方式中得分较高的一种。观察实验结果发现，在使用思维链后，摘要整体质量有了明显提升。在整个实验中，GPT-3.5模型在多个领域和指标下都表现出色，这可能归因于其在自然语言生成领域的强大性能。

[3]https://huggingface.co/Qwen/Qwen1.5-7B-Chat
[4]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[5]https://huggingface.co/FlagAlpha/Llama2-Chinese-7b-Chat

具体而言，对于医学领域，所有模型在使用思维链生成时，各项指标均比Zero-shot方式有所提升。这表明，思维链在医学领域的单文档摘要生成中是有效的。对于人工智能领域，尽管在一些指标上（如Qwen、Llama和GPT-3.5上的ROUGE-1/2/L），使用思维链生成方式并未带来得分提升。然而，根据 Goyal et al. (2022)的研究发现，自动评估结果与人类偏好并不一致。实际上，尽管GPT-3.5生成的摘要在自动评估指标上表现不佳，但其更受人类喜爱。

Dataset	Model	Zero-shot					Chain-of-Thought				
		R-1	R-2	R-L	BS	GPT-3.5	R-1	R-2	R-L	BS	GPT-3.5
Medical	ChatGLM3-6b	40.54	18.65	25.68	88.60	80.49	46.86	23.18	32.78	90.62	85.08
	QWen1.5-7b	46.70	22.19	32.64	91.00	85.57	47.40	23.13	34.92	91.42	87.21
	Llama2-7b	47.06	23.70	33.00	90.84	86.39	47.53	24.18	33.09	90.93	86.72
	GPT-3.5-turbo	**52.39**	**28.10**	**37.92**	**92.21**	**87.38**	**53.21**	**29.70**	**40.06**	**92.31**	**88.20**
AI	ChatGLM3-6b	31.55	12.12	18.84	86.22	83.00	**36.89**	**14.32**	**21.71**	87.74	88.00
	QWen1.5-7b	32.30	8.81	18.02	87.52	88.50	30.29	7.82	16.69	87.47	89.33
	Llama2-7b	**35.97**	**13.53**	**21.71**	87.58	87.83	35.58	13.45	20.97	87.78	88.17
	GPT-3.5-turbo	33.84	10.58	19.50	**87.91**	**89.17**	33.62	11.45	19.58	**87.79**	**89.67**

Table 2: 不同方法在单文档摘要测试集上的实验结果

因此，为验证生成摘要的质量，我们进行了人工评估。我们从流畅性、连贯性、一致性和相关性四个维度（具体标准参见第 4.2节）对摘要进行人工打分，评分结果见图 5和图 6。观察图表结果，我们发现不同大模型在使用了思维链技术后，生成的摘要在四个维度上都有不同程度的分数提升。其中，流畅性和一致性的提升较小，而连贯性和相关性的提升较为显著。这主要归因于我们设计了一个逐步引导大模型的提示词(Prompt)，使其遵循人类生成摘要的逻辑来构建摘要，并指示其需要保留哪些重要内容（例如，研究的对象、背景、方法、结果等）。这一做法在一定程度上弥补了大模型在逻辑组织和内容重要度判断能力上的不足。人工评估再次证明了思维链技术（Chain-of-Thought）的有效性。

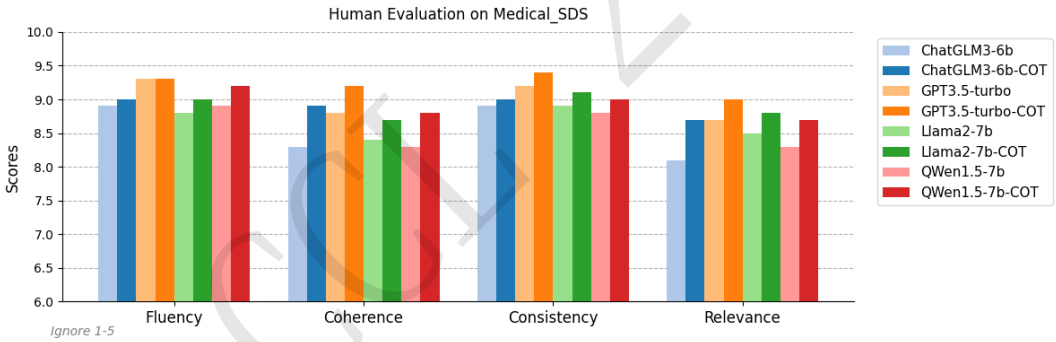

图 5. 医学领域单文档摘要生成结果的人工评分

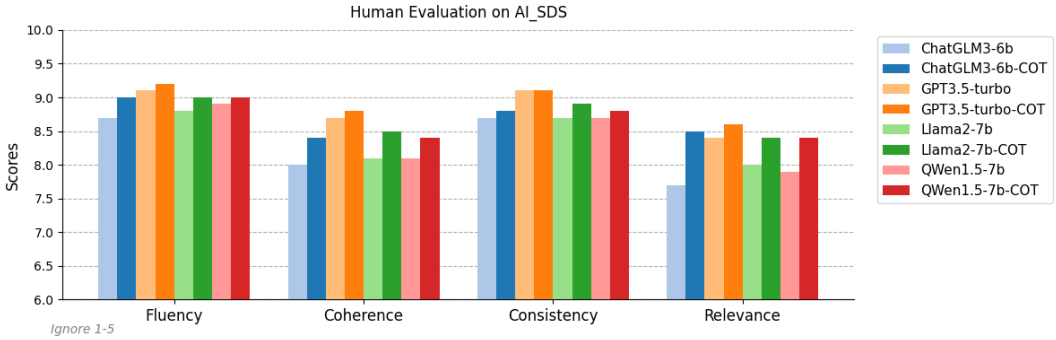

图 6. 人工智能领域单文档摘要生成结果的人工评分

4.3.2 第二阶段：多文档摘要生成

在第二阶段的多文档摘要生成任务中，我们采用了三种不同的生成方法：直接生成（Direct）、层次化生成（Hierarchical）和增量化生成（Incremental）。这些方法利用第一阶段不同方式生成的单文档摘要，其中w/o表示第一阶段Zero-shot生成的单文档摘要，w/表示第一阶段Chain-of-Thought生成的单文档摘要，并据此生成多文档摘要。表 3展示了在医学和人工智能多文档摘要测试集（Medical MDS和AI MDS）上使用以上三种方法进行实验的结果。带有灰色背景的数据表示在每一种生成方法下，w/o与w/生成分数较高的结果。

Metric	Model	Medical MDS						AI MDS					
		Direct		Hierarchical		Incremental		Direct		Hierarchical		Incremental	
		w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/	w/o	w/
ROUGE-1	ChatGLM3-6b	31.90	34.03	33.64	34.44	33.19	**35.41**	29.56	30.51	29.58	31.06	29.82	**31.96**
	QWen1.5-7b	38.45	38.21	39.07	38.08	38.93	**39.48**	33.77	35.14	33.75	**35.46**	33.97	34.93
	Llama2-7b	38.87	38.93	38.51	38.33	**39.70**	38.95	33.41	**34.92**	32.99	34.34	34.22	34.80
	GPT-3.5-turbo	40.10	39.77	41.08	40.00	**42.80**	42.33	34.73	34.47	33.89	34.28	**35.32**	35.26
ROUGE-2	ChatGLM3-6b	11.38	13.21	12.27	13.45	12.16	**13.89**	8.22	9.63	8.38	**10.11**	8.20	10.10
	QWen1.5-7b	14.80	14.64	15.44	14.50	15.06	**15.46**	10.16	10.53	10.21	**10.87**	10.02	10.53
	Llama2-7b	**16.65**	16.58	16.36	16.22	16.62	16.57	11.35	**11.79**	11.20	11.35	11.64	11.52
	GPT-3.5-turbo	15.66	15.13	16.69	15.99	**18.25**	17.95	11.46	11.37	11.32	11.01	11.55	**11.73**
ROUGE-L	ChatGLM3-6b	18.97	19.98	19.53	**20.57**	18.98	20.16	17.45	18.44	17.57	**18.70**	17.30	18.41
	QWen1.5-7b	23.41	23.01	23.64	23.15	23.19	**23.67**	19.48	20.05	19.91	**20.52**	18.59	19.37
	Llama2-7b	24.12	23.60	**24.17**	23.55	24.01	23.49	20.83	**21.52**	20.66	21.29	20.16	20.62
	GPT-3.5-turbo	24.41	24.03	25.27	24.14	**25.99**	25.83	**21.25**	20.74	19.90	20.39	20.22	20.39
GPT-3.5	ChatGLM3-6b	76.80	**81.60**	77.40	80.00	78.20	81.20	82.80	86.00	82.40	**86.60**	83.20	84.80
	QWen1.5-7b	86.20	85.60	**86.80**	86.00	84.60	84.20	86.80	87.60	87.80	**88.20**	85.40	86.40
	Llama2-7b	83.80	**86.00**	84.40	84.60	85.00	85.40	87.20	87.40	87.00	**87.60**	86.40	87.20
	GPT-3.5-turbo	86.00	86.60	87.00	**87.20**	86.20	86.20	88.60	88.80	88.20	**89.20**	88.40	89.00

Table 3: 不同方法在多文档摘要测试集上的实验结果

我们对表 3中使用第一阶段生成结果得到的三种策略下，不同模型生成的多文档摘要指标进行了平均值计算和分析，结果如表 4所示。针对医学领域的多文档摘要生成任务，层次化生成策略相较于直接生成策略在ROUGE-2和ROUGE-L指标上均有所提升，仅在ROUGE-1略微下降，而增量化生成策略的提升效果更为显著。对于人工智能领域，层次化生成策略在所有指标上均表现出提升，而增量化生成策略在ROUGE-L指标上的降低可能与生成文本长度减少有关。总体而言，层次化和增量化生成策略均表现出更好的性能，显著提高了多文档摘要生成的质量，这进一步证明了思维链方法在多文档摘要生成中的有效性。

Metric	Medical MDS			AI MDS		
	Direct	Hierarchical	Incremental	Direct	Hierarchical	Incremental
ROUGE-1	37.74	37.71(↓0.03)	**39.04(↑1.30)**	33.76	**33.79(↑0.02)**	**34.24(↑0.48)**
ROUGE-2	14.89	**15.04(↑0.15)**	**15.97(↑1.08)**	10.83	**10.84(↑0.01)**	**10.97(↑0.14)**
ROUGE-L	22.66	**22.85(↑0.19)**	**23.29(↑0.63)**	20.19	**20.23(↑0.04)**	19.70(↓0.49)

Table 4: 不同生成策略在第二阶段的多文档摘要生成结果比较

Metric	Medical MDS			AI MDS		
	Direct	Hierarchical	Incremental	Direct	Hierarchical	Incremental
ROUGE-1	**+0.4050**	-0.3625	**+0.3875**	**+0.8925**	**+1.2325**	**+0.9050**
ROUGE-2	**+0.2675**	-0.1500	**+0.4450**	**+0.5325**	**+0.5575**	**+0.6175**
ROUGE-L	-0.0725	-0.3000	**+0.2450**	**+0.4350**	**+0.7150**	**+0.6300**

Table 5: 第一阶段思维链在第二阶段生成中使用情况的效果分析

从表 3的结果还可以清晰看出，使用第一阶段生成的思维链摘要（w/）后，大多数模型在所有生成策略下的性能都有所提升，尤其在人工智能领域的测试集中表现更加显著。进一步的计算发现，如表 5所示，在第二阶段使用第一阶段思维链后生成结果的分数变化。相对于直接生成单文档摘要的方法（w/o），使用第一阶段思维链生成的摘要（w/）在人工智能领域的平均提升分数更高，这一趋势在医学领域的直接生成策略和增量化方法中也得到了体现。

为了深入研究在医学多文档摘要测试集上，第一阶段思维链在层次化生成策略下性能降低的原因，我们进行了人工评估，评估结果如表 6所示，w/o表示没有使用第一阶段的思维链。结果显示，第一阶段生成的思维链摘要在第二阶段的多文档摘要生成任务中性能显著提升，表明思维链对于帮助模型理解文本内容并生成更准确、连贯的摘要起到了关键作用。人工评价再一次验证了 Goyal et al. (2022)的研究结论，即自动评估结果与人类偏好的不一致性。

Model	Medical_MDS	AI_MDS
	Flu/Coh/Con/Rel	*Flu/Coh/Con/Rel*
ChatGLM3-6b	**9.10/8.70/8.60/8.40**	**8.80/8.50/8.70/8.40**
w/o chain1	8.90/8.70/8.10/8.00	8.70/8.40/8.50/7.80
QWen1.5-7b	**8.80/8.70/8.80/8.50**	**9.00/8.80/8.90/8.50**
w/o chain1	8.80/8.50/8.30/8.20	8.70/8.60/8.70/8.10
Llama2-7b	**9.00/8.90/8.70/8.60**	**8.90/9.00/8.80/8.70**
w/o chain1	8.80/8.60/8.40/8.10	8.80/8.70/8.70/8.30
GPT-3.5-turbo	**9.00/8.90/8.90/8.70**	**9.20/9.00/9.10/8.80**
w/o chain1	9.00/8.90/8.80/8.60	9.10/9.00/8.80/8.80

Table 6: 在医学和人工智能领域多文档摘要测试集上的层次化生成策略的人工评价结果

4.3.3 第三阶段：跨语言多文档摘要生成

在本节中，我们利用 4.1节构建的跨语言多文档摘要的测试集（CLMDS）进行了跨语言多文档摘要生成的实验。我们分别评估了两个模型（ChatGLM3-6b和GPT-3.5-turbo）在不同条件下的性能，包括是否使用了三个阶段的思维链生成。实验结果如表 7 所示。

我们对未使用思维链直接生成的方法，以及使用了第一阶段和第三阶段思维链的方法，以及使用了三个阶段所有思维链的方法进行了比较。结果显示，在医学领域的CLMDS任务中，无论是在ROUGE-1、ROUGE-2、ROUGE-L还是GPT-3.5得分方面，使用了思维链的模型都表现出了显著的提升。特别是当使用了所有三个阶段的思维链时，性能提升更为明显，这进一步验证了思维链在帮助模型理解文本内容并生成更准确、连贯的摘要方面的有效性。对于人工智能领域的CLMDS任务，我们观察到类似的趋势。使用思维链的模型在所有指标下均实现了提升，尤其是在使用了所有三个阶段思维链的情况下，性能得到了最大程度的提升。这些结果强调了思维链方法在跨语言多文档摘要生成任务中的有效性和重要性。

综上所述，本阶段的实验结果表明，在跨语言多文档摘要生成任务中，思维链技术对于提升模型性能具有积极作用，可以有效改善生成摘要的质量。

Model	Medical CLMDS				AI CLMDS			
	R-1	R-2	R-L	GPT-3.5	R-1	R-2	R-L	GPT-3.5
ChatGLM3-6b	35.82	11.94	22.13	68.60	36.48	11.68	22.29	79.80
w/ chain1,3	37.79	13.05	22.79	69.40	37.36	12.30	22.75	80.00
w/ chain1,2,3	**46.47**	**18.40**	**27.67**	**73.60**	**44.68**	**15.82**	**25.31**	**80.60**
GPT-3.5-turbo	22.43	9.08	14.89	76.80	38.26	14.06	26.14	80.40
w/ chain1,3	40.68	17.22	26.43	78.80	42.41	16.46	27.69	80.60
w/ chain1,2,3	**49.11**	**21.97**	**29.94**	**81.60**	**45.48**	**18.24**	**27.69**	**82.20**

Table 7: 使用不同思维链在跨语言多文档摘要生成的实验结果

5 结论

在本文中，我们提出了一种基于思维链的跨语言多文档摘要方法，并在医学和人工智能领域的学术论文上进行了实验验证。我们构建了一个新的测试集，用于评估我们方法在这一任务上的表现，并选择了一些典型的大语言模型进行实验。实验结果表明，无论是在单文档摘要还是多文档摘要生成阶段，无论是单语言还是跨语言，思维链技术都能显著提高摘要的质量。

这些发现对于跨语言多文档摘要领域具有重要意义。首先，我们的方法为解决跨语言多文档摘要生成任务提供了一种有效的思路。其次，我们构建的跨语言多文档摘要测试集也可为未来的研究提供参考。此外，我们的实验结果表明，不同的大语言模型在跨语言多文档摘要任务上表现不同，因此未来可以进一步研究模型选择和优化。

在未来的工作中，我们将继续优化思维链技术，探索更好的生成方法和模型选择策略。我们还计划扩展测试集，涵盖更多领域和语言，以更全面地评估方法的通用性和鲁棒性。我们相信，这项工作将为跨语言多文档摘要领域的研究和应用提供有价值的参考和启示。

参考文献

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6220–6231.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. *arXiv preprint arXiv:2306.01200*.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6244–6254.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. 2003. Cross-lingual c* st* rd: English access to hindi information. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3):245–269.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging graph to improve abstractive multi-document summarization. *arXiv preprint arXiv:2005.10043*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. *arXiv preprint arXiv:1905.13164*.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Constantin Orăsan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual romanian-english multi-document summariser.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Elvys Linhares Pontes, Stéphane Huet, Juan-Manuel Torres-Moreno, and Andréa Carneiro Linhares. 2020. Compressive approaches for cross-language multi-document summarization. *Data & Knowledge Engineering*, 125:101763.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? *arXiv preprint arXiv:2305.13091*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926.

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1546–1555.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. *arXiv preprint arXiv:2004.12393*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023b. Zero-shot cross-lingual summarization via large language models. *arXiv preprint arXiv:2302.14229*.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023c. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. 2020. Mixed-lingual pretraining for cross-lingual summarization. *arXiv preprint arXiv:2010.08892*.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 118–127.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, et al. 2023. Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. *arXiv preprint arXiv:2311.11797*.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. *arXiv preprint arXiv:1909.00156*.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 1309–1321.

黄文彬and 倪少康. 2017. 多文档自动摘要方法的进展研究. 情报科学, 35:160–165.

附录

A 本文中用到的提示词

A.1 第一阶段：单文档摘要生成

1.抽取关键信息提示词:

Read and analyze the preceding content carefully, anwser the question below:

what is the subject of the study?
what are the research methods of the study?
what are the primary findings of the study?
what are the innovations introduced in the study?

2.生成摘要提示词:

Let's integrate the above information and summarize the study in just one paragraph:

A.2 第二阶段：多文档摘要生成

1.直接生成提示词:

Summary 1:
{*summary_1*}

Summary 2:
{*summary_2*}

Summary n:
{*summary_n*}

Please carefully review the provided summaries. Analyze the similarities and differences between these studies. Subsequently, creating a cohesive, unified summary in just one paragraph. This new summary should be crafted in the style of a survey. Remember, when referencing the studies, avoid using labels like "Summary n." Instead, directly address the content of each research study.

New summary:

2.层次化生成提示词:

Consider two sets of scientific articles. Your task is to merge the summaries of these articles into a single comprehensive summary, capturing the connections between different documents and presenting the merged summary in one paragraph.

Summary of the first set of articles:

{*summary_1*}

Summary of the second set of articles:

{*summary_2*}

Merging Process:
1.Capturing Connections: Identify and analyze the connections between different documents in both summaries. Look for common themes, findings, methodologies, or implications.
2.Creating a One-Paragraph Summary: Merge the summaries of both sets of articles into a single paragraph that flows logically and smoothly. Ensure that the merged summary effectively communicates the overarching insights and contributions of the entire document collection.

Objective: Your goal is to create a merged multi-document summary in one paragraph that effectively captures the connections between the articles in both sets and presents them in a coherent and cohesive manner.

Updated summary:

3.增量化生成提示词:
Below is a summary of a science article:

{*science_article_summary*}

The following is a summary of the many scientific articles that have been generated:

{*overall_summary*}

Your task is to merge the summary of the science article into the overall multi-document summary. Ensure that the merge is done in a way that maintains coherence and consistency with the overall summary. Introduce any new concepts, terms, or references if they are being mentioned for the first time in the summary.

Step-by-step Process:
1.Merge: Incorporate the summary of the science article into the overall multi-document summary as a single paragraph. 2.Similarities and Differences: Analyze the similarities and differences between the articles, if applicable, and integrate these observations into the multi-document summary to create a smooth and logical narrative.

Objective: Your goal is to create a comprehensive, cohesive, and informative multi-document summary presented as a single paragraph that accurately represents the key aspects of the science article while maintaining coherence with the overall summary.

Updated summary:

A.3 第三阶段：跨语言多文档摘要生成

Summary 1:
{*summary_1*}

Summary 2:
{*summary_2*}

Summary n:
{*summary_n*}

Multi-document Summary:
{Multi-document Summary}

Integrate the summaries of each document and their multi-document summaries together, and generate summaries of these documents in Chinese. During the integration process, identify common themes, findings, methodologies, or impacts to ensure coherence and consistency. Introduce new concepts, terms, or references in the summary if they are mentioned for the first time.

Objective: Your goal is to create a comprehensive, coherent, and informative multi-document summary in Chinese that accurately reflects the main content of the articles mentioned above.

Updated summary:

A.4 单文档摘要评估提示词

Score the following summary generated by another system given the source on a scale from 1 to 10 with regards to overall general summary quality. 1-point indicates a low quality summary, and 10-points indicates a very high quality summary. A high quality summary is grammatical, fluent, informative, relevant, coherent and factually consistent with the source.

Let's think step-by-step and just output one number representing the score.

Source:
{source_text}

Summary:
{summary_llm}

Your score: (What you must observe is that your output must contain only a number and no other content)

A.5 多文档摘要评估提示词

Score the following multi-document summary generated by another system given the source on a scale from 1 to 10 with regards to overall general summary quality. 1-point indicates a low-quality multi-document summary, and 10-points indicates a very high-quality multi-document summary. A high-quality multi-document summary is concise in language, grammatical, fluent, informative, relevant, logically coherent and factually consistent with the source, successfully integrates all documents, and contains important content from all documents.

Furthermore, when you are scoring the summary, be strict and demand high standards, don't be too conservative with your scoring.

Let's think step-by-step and just output one number representing the score.

Source:
{source_text}

Multi-Document Summary:
{summary_llm}

Your score: (What you must observe is that your output must contain only a number and no other content)

A.6 跨语言多文档摘要评估提示词

Score the following Chinese multi-document summary generated by another system given the English source on a scale from 1 to 10 with regards to overall general summary quality.

1-point indicates a low-quality multi-document summary, and 10-points indicates a very high-quality multi-document summary. A high-quality multi-document summary is concise in language, grammatical, fluent, informative, relevant, logically coherent and factually consistent with the source, successfully integrates all documents, and contains important content from all documents.

Furthermore, when you are scoring the summary, be strict and demand high standards, don't be too conservative with your scoring.

Let's think step-by-step and just output one number representing the score.

English Source:
{source_text}

Chinese Multi-Document Summary:
{summary_llm}

Your score: (What you must observe is that your output must contain only a number and no other content)

B 生成实例分析

基于思维链的单文档摘要生成示例1：

Source Document:

Title: Ranking age-specific modifiable risk factors for cardiovascular disease and mortality: evidence from a population-based longitudinal study

Cardiovascular disease (CVD) remains a paramount contemporary health challenge.....

ChatGLM3 Without COT:

Cardiovascular disease (CVD) remains a major health challenge, and this study examined age-specific effects of 14 risk factors on CVD and mortality in different age groups. The study analyzed data from 226,759 CVD-free participants aged 40 years and older in the UK Biobank during the period from baseline time (2006-2010) to September 30, 2021. The primary CVD outcome was a composite of incident coronary artery disease, heart failure, and stroke. The study found that age-specific disparities were observed in the risk factors contributing to CVD, and the overall PAF declined with age. Metabolic factors had the highest PAF in each age group, with hypertension (14.04% of the PAF) and abdominal obesity (9.58% of the PAF) being prominent. Behavioral factors had the highest PAF in the middle-aged group (10.68% of the PAF), and smoking was the leading behavioral factor in all age groups. In socioeconomic and psychosocial risk clusters, low income contributed most among middle-aged (3.74% of the PAF) and elderly groups (3.66% of the PAF), while less education accounted more PAF for quinquagenarian group (4.46% of the PAF). The study concluded that a large fraction of CVD cases and deaths were associated with modifiable risk factors in all age groups, and targeted efforts should focus on the most impactful risk factors, as well as age-specific modifiable risk factors. These findings may inform the development of more precise medical strategies to prevent and manage CVD and related mortality.

ChatGLM3 With COT:

The study examined the age-specific effects of 14 modifiable risk factors on cardiovascular disease (CVD) and mortality in a large sample of 226,759 CVD-free participants aged 40 years and older in the UK Biobank. The study found that a large fraction of CVD cases and deaths were associated with modifiable risk factors in all age groups. Metabolic factors had the highest PAF in each age group, with hypertension and abdominal obesity being prominent. Behavioral factors had the highest PAF in the middle-aged group, and smoking was the leading behavioral factor in

all age groups. In socioeconomic and psychosocial risk clusters, low income and less education contributed most among middle-aged and elderly groups, respectively. The study highlights the need for targeted efforts to focus on the most impactful risk factors and age-specific modifiable risk factors to inform the development of more precise medical strategies to prevent and manage CVD and related mortality.

Reference:

This study investigates the age-specific impacts of 14 modifiable risk factors on cardiovascular disease (CVD) and mortality by analyzing data from 226,759 CVD-free participants aged 40 and above from the UK Biobank, spanning from 2006-2010 to September 2021. It identifies significant age-related disparities in the contribution of these risk factors to CVD, revealing that metabolic factors, especially hypertension and abdominal obesity, are the most influential across all age groups, while the importance of behavioral and socioeconomic factors varies with age. The research highlights that a substantial proportion of CVD cases and deaths could be attributed to these modifiable risk factors, suggesting that targeted preventive measures focusing on the most impactful factors for each age group could enhance strategies for CVD prevention and management. This approach marks a significant innovation in public health efforts, offering a more nuanced understanding of how age-specific risk factor modifications can inform more precise medical interventions.

基于思维链的单文档摘要生成示例2:

Source Document:

Title: Safety and efficacy of inhaled interferon-β 1a (SNG001) in adults with mild-to-moderate COVID-19: a randomized, controlled, phase II trial

With the emergence of SARS-CoV-2 variants resistant to monoclonal antibody therapies and limited global access to therapeutics, the evaluation of novel therapeutics to prevent progression to severe COVID-19 remains a critical need.....

GPT-3.5 Without COT:

In a randomized, controlled, phase II trial evaluating the safety and efficacy of inhaled interferon-β 1a (SNG001) in adults with mild-to-moderate COVID-19, it was found that SNG001 was safe with a non-statistically significant decrease in hospitalization for COVID-19 pneumonia compared to placebo. The trial, conducted between February 10 and August 18, 2021, showed no significant differences in adverse events, time to symptom improvement, or viral clearance between the SNG001 and placebo groups. However, there was a notable reduction in hospitalizations with SNG001 compared to placebo. The findings suggest that SNG001 may be a promising therapeutic option for preventing severe COVID-19 outcomes, although further research is needed to confirm its efficacy.

GPT-3.5 With COT:

The study evaluated the safety and efficacy of inhaled interferon-β 1a (SNG001) in adults with mild-to-moderate COVID-19 through a phase II randomized controlled trial. Adult outpatients with confirmed SARS-CoV-2 infection were randomized to receive either SNG001 or placebo for 14 days. The primary findings showed no significant differences in adverse events, time to symptom improvement, or viral RNA levels between the two groups. However, there was a trend towards fewer hospitalizations in the SNG001 group, with an 86% relative risk reduction compared to placebo. The study suggests that SNG001 was safe and may have potential benefits in reducing hospitalizations for COVID-19 pneumonia, although this reduction was not statistically significant in this trial.

Reference:

This study evaluates the safety and efficacy of inhaled interferon-β 1a (SNG001) for treating adults with mild-to-moderate COVID-19 through a phase II randomized controlled trial conducted between February 10 and August 18, 2021. Utilizing the ACTIV-2/A5401 platform, 220 adult

outpatients with confirmed SARS-CoV-2 infection were randomized to receive either SNG001 or a placebo once daily for 14 days. The findings revealed no significant differences in adverse events, time to symptom improvement, or viral RNA quantification between the two groups. However, SNG001 was associated with a non-statistically significant reduction in hospitalizations, suggesting an 86% relative risk reduction compared to the placebo. This study introduces the potential of inhaled interferon-β 1a as a novel therapeutic approach for COVID-19, highlighting its safety and a possible benefit in reducing hospitalizations, despite the need for further research to confirm its efficacy.

基于思维链的多文档摘要生成示例1：

Source Document:

Article1: *A Distillation Approach to Data Efficient Individual Treatment Effect Estimation*

.....

Article2: *Estimation of Local Average Treatment Effect by Data Combination*

.....

Article3: *Learning Instrumental Variable from Data Fusion for Treatment Effect Estimation*

.....

Article4: *Treatment Effect Estimation with Data-Driven Variable Decomposition*

.....

Qwen1.5 Direct:

This survey presents four innovative studies in the field of causal inference and treatment effect estimation, addressing various challenges in data efficiency and precision. "Data Efficient Individual Treatment Effect Estimator" (DEITEE) presents DEITEE, a method that combines a base model for confounding adjustment with a decision tree to select critical variables, reducing the number of required inputs while maintaining accuracy. It particularly shines in scenarios with limited data. The study on local average treatment effect (LATE) adaptation adapts a weighted least squares estimator to estimate LATE in merged datasets, relaxing positivity assumptions and outperforming IPW methods. The Group Instrumental Variable (GIV) approach, introduced in the Meta-EM algorithm, leverages representation learning to identify latent group labels without predefined IVs, overcoming limitations of traditional methods. Lastly, "Data-Driven Variable Decomposition" (D2VD) utilizes a regularized integrated regression model to separate confounders from adjustment variables, enhancing ATE estimation precision in high-dimensional data. These studies collectively demonstrate the advancement in causal inference techniques, addressing challenges in data handling and improving the accuracy of treatment effect estimation in various real-life settings.

Qwen1.5 Hierarchical:

Both sets of scientific articles delve into the realm of efficient and accurate treatment effect estimation in mixed observational datasets. The first set presents "Data Efficient Individual Treatment Effect Estimator" (DEITEE), a cutting-edge method that combines a base model for confounding adjustment with a decision tree to optimize predictions with limited data, particularly in scenarios with sparse individual information. This study complements the second one, which addresses non-compliance and missing treatment in a combined dataset by employing a weighted least squares estimator with estimated propensity-score-difference (PSD) to mitigate instability and relax positivity assumptions. Both works showcase their effectiveness through synthetic and real-world examples, highlighting their practicality in applications such as online-to-offline marketing and panel data analysis. The second set contributes to this area with its Group Instrumental Variable (GIV) method, a Representation Learning and Meta-EM algorithm that bypasses the need for predefined IVs by reconstructing latent source labels, relaxing linear

regression assumptions and outperforming traditional summary IV methods. Meanwhile, Data-Driven Variable Decomposition (D2VD) proposes a regularization-based approach that separates confounders from adjustment variables in high-dimensional data, enhancing ATE estimation while incorporating orthogonality and sparsity constraints. These innovative methods demonstrate the potential of data-driven causal inference in various sectors, including social marketing, healthcare, and public policy, and highlight the importance of overcoming challenges in mixed observational datasets. Together, these studies emphasize the significance of developing efficient and robust techniques for estimating treatment effects in diverse settings.

Qwen1.5 Incremental:

This study introduces the Data-Driven Variable Decomposition (D2VD) algorithm, a novel approach to precise ATE estimation in observational studies by leveraging a regularized integrated regression model. D2VD effectively separates confounders from adjustment variables in high-dimensional data through orthogonality and sparsity constraints, outperforming traditional methods that treat all variables as confounders. It enhances accuracy and reduces variance in causal inference applications, particularly in areas like social marketing, healthcare, and public policy. In contrast, the research on Group Instrumental Variables (GIV) using representation learning and the Meta-EM algorithm in mixed observational datasets offers a breakthrough in causal effect estimation, eliminating the need for predefined IV candidates and learning latent source labels without linear regression assumptions. Although both studies demonstrate improved performance, GIV focuses on group differences and relaxes linear regression requirements, while DEITEE addresses limited data by combining a decision tree for confounding adjustment and individual treatment effect estimation. LATE estimation, another work, handles non-compliance and missing data with a weighted least squares estimator for Local Average Treatment Effect. These studies, along with their empirical validation, highlight their practical applications in various real-world scenarios and suggest potential for combined methods in complex causal inference tasks.

Reference:

In the realm of estimating treatment effects from observational data, recent advancements have introduced a variety of innovative methodologies to tackle the inherent challenges associated with data fusion, practical limitations of data acquisition, and the complexity of high-dimensional variables. Among these, the Data-Driven Variable Decomposition (D2VD) algorithm stands out by offering a solution to automatically separate confounders and adjustment variables in observational studies, thereby enabling more accurate treatment effect estimation with tighter confidence intervals, especially in scenarios characterized by high-dimensional data. This approach complements the Data Efficient Individual Treatment Effect Estimation (DEITEE) method, which reduces data requirements by identifying a minimal set of variables for accurate estimation, and a novel strategy that addresses the estimation of the local average treatment effect (LATE) in the presence of incomplete compliance and disjointed datasets through a weighted least squares estimator. Furthermore, the introduction of a method to overcome challenges posed by omitted source labels and unmeasured confounders in mixed datasets, through reconstructing the source label and modeling it as a Group Instrumental Variable (GIV) within a Meta-EM framework, signifies a significant leap forward. This framework estimates individual treatment assignment probability and infers treatment effect by mapping raw data into a representation space, estimating distribution differences, and modeling the GIV for different treatment assignment mechanisms through an alternating training strategy. Collectively, these studies not only expand the toolkit available for researchers and practitioners aiming to derive meaningful insights from observational data to inform treatment decisions and policy-making but also showcase the diversity of approaches to overcoming the practical challenges of data-related limitations and complex data structures in the estimation of treatment effects. The integration of the D2VD algorithm into this body of work highlights the importance of distinguishing between confounders and adjustment variables in enhancing the precision of treatment effect estimates,

thereby contributing to the broader field of causal inference with a nuanced understanding of variable impacts in observational studies.

基于思维链的多文档摘要生成示例2：

Source Document:

Article1: *Efficacy and safety of zapnometinib in hospitalised adult patients with COVID-19 (RESPIRE): a randomised, double-blind, placebo-controlled, multicentre, proof-of-concept, phase 2 trial*

.....

Article2: *Safety and effectiveness of COVID-19 vaccines in patients with IgA nephropathy: a retrospective cohort study from the TriNetX global collaborative networks*

.....

Article3: *Efficacy of a monovalent (D614) SARS-CoV-2 recombinant protein vaccine with AS03 adjuvant in adults: a phase 3, multi-country study*

.....

Article4: *Safety and efficacy of inhaled interferon-β 1a (SNG001) in adults with mild-to-moderate COVID-19: a randomized, controlled, phase II trial*

.....

GPT-3.5 Direct:

The studies examined various interventions for COVID-19 in different patient populations. One study investigated the safety and efficacy of zapnometinib in hospitalised COVID-19 patients, showing potential benefits in improving clinical severity scores, especially in non-Omicron variants. Another study focused on COVID-19 vaccination in IgA nephropathy patients, revealing reduced infection risks and normalized renal function post-vaccination. A third study assessed a SARS-CoV-2 recombinant protein vaccine, demonstrating efficacy in non-naïve individuals. Lastly, a study on inhaled interferon-β 1a in mild-to-moderate COVID-19 patients suggested a trend towards reduced hospitalizations. These findings collectively underscore the importance of diverse interventions in managing COVID-19, highlighting the need for further research to optimize treatment strategies and outcomes.

GPT-3.5 Hierarchical:

The collective findings from a diverse set of scientific articles shed light on various strategies in combating COVID-19. One study explored the safety and efficacy of zapnometinib in hospitalized COVID-19 patients, indicating potential benefits in improving clinical severity scores, especially in severe cases and non-Omicron variants, despite recruitment challenges. Concurrently, research on COVID-19 vaccination in IgAN patients revealed reduced infection risks post-vaccination, with a noted temporary decline in eGFR, emphasizing the importance of renal function monitoring. Another study delved into a monovalent SARS-CoV-2 recombinant protein vaccine with AS03 adjuvant, showcasing varied efficacy based on prior virus exposure, while a separate investigation on inhaled interferon-β 1a (SNG001) for mild-to-moderate COVID-19 suggested a potential decrease in hospitalizations. These studies collectively underscore the necessity for continued exploration of diverse approaches in combating COVID-19, with implications ranging from targeting specific pathways like Raf/MEK/ERK to monitoring renal health and addressing challenges in vaccine efficacy among naïve individuals, ultimately offering promising avenues for mitigating severe outcomes and advancing our understanding of effective interventions against the virus.

GPT-3.5 Incremental:

In a comprehensive exploration of COVID-19 interventions, a global Phase 3 trial assessed a monovalent SARS-CoV-2 recombinant protein vaccine with AS03 adjuvant, revealing varied

efficacy rates based on participants' serostatus. The vaccine demonstrated 15.3% efficacy in SARS-CoV-2 naïve individuals, 32.9% irrespective of serostatus, and 52.7% in non-naïve participants, emphasizing challenges in protecting naïve individuals against the D614 strain but showing promise for those previously exposed to the virus. Concurrently, research on zapnometinib in hospitalised COVID-19 patients highlighted potential benefits in improving clinical severity scores, particularly in severe cases and non-Omicron variants, with minimal adverse events. Additionally, a study focusing on COVID-19 vaccination in IgAN patients indicated reduced infection risks and complications post-vaccination, alongside a temporary eGFR decrease that normalized over time, stressing the importance of monitoring renal function in this population. The study evaluated the safety and efficacy of inhaled interferon-β 1a (SNG001) in adults with mild-to-moderate COVID-19 through a phase II randomized controlled trial. Adult outpatients with confirmed SARS-CoV-2 infection were randomized to receive either SNG001 or placebo for 14 days. The primary findings showed no significant differences in adverse events, time to symptom improvement, or viral RNA levels between the two groups. However, there was a trend towards fewer hospitalizations in the SNG001 group, with an 86% relative risk reduction compared to placebo. The study suggests that SNG001 was safe and may have potential benefits in reducing hospitalizations for COVID-19 pneumonia, although this reduction was not statistically significant in this trial. These findings underscore the need for continued investigation into zapnometinib's therapeutic potential in COVID-19 treatment and the importance of optimizing outcomes through tailored interventions and vigilant monitoring.

Reference:

Recent research efforts have been directed towards identifying novel strategies to combat COVID-19 and mitigate its impacts, particularly in individuals with pre-existing health conditions. Among these, a pivotal study evaluated the safety, efficacy, and reactogenicity of a monovalent (D614) SARS-CoV-2 recombinant protein vaccine with AS03 adjuvant across 53 research centers in eight countries. This Phase 3 trial, randomized and placebo-controlled, aimed to assess the vaccine's effectiveness in preventing symptomatic COVID-19, revealing a nuanced efficacy profile: 15.3% in SARS-CoV-2 naïve participants, 32.9% regardless of initial serostatus, and 52.7% in previously exposed individuals, suggesting its potential utility as a booster despite challenges in protecting against new variants. The vaccine demonstrated an acceptable safety profile, aligning with the broader goal of identifying safe and effective COVID-19 interventions. Another innovative approach explored was the use of inhaled interferon-β 1a (SNG001) in a phase II randomized controlled trial for treating adults with mild-to-moderate COVID-19. Although the study, conducted between February 10 and August 18, 2021, did not show significant differences in adverse events, time to symptom improvement, or viral RNA quantification between the treatment and placebo groups, SNG001 was associated with a non-statistically significant reduction in hospitalizations, suggesting an 86% relative risk reduction compared to placebo. This introduces the potential of inhaled interferon-β 1a as a novel therapeutic approach. Additionally, zapnometinib, an oral MEK1/MEK2 inhibitor, was investigated for its potential immunomodulatory and antiviral effects in hospitalized adults with moderate to severe COVID-19. Despite early trial termination due to recruitment challenges, preliminary results indicated possible benefits in reducing clinical severity, especially in severe cases and those caused by non-Omicron variants, without significant adverse effects. Research utilizing the TriNetX collaborative networks also assessed the renal safety and effectiveness of COVID-19 vaccines in patients with Immunoglobulin A nephropathy (IgAN), finding a significant reduction in infection risk and complications, alongside a transient, reversible decline in renal function. Collectively, these studies underscore the critical importance of developing targeted therapeutic and preventive measures against COVID-19, emphasizing the need for ongoing research to refine treatment protocols and vaccination strategies in response to evolving viral variants and diverse patient health profiles.

基于思维链的跨语言多文档摘要生成示例1：

Source Document:

Article1: *FiLM: Visual Reasoning with a General Conditioning Layer*

.....

Article2: *Hierarchical ConViT with Attention-Based Relational Reasoner for Visual Analogical Reasoning*

.....

Article3: *Learning the Dynamics of Visual Relational Reasoning via Reinforced Path Routing*

.....

Article4: *Visual Concept Reasoning Networks*

.....

ChatGLM3 Direct:

本文综述了近年来在视觉推理领域的研究，这些研究主要集中在利用深度学习模型实现视觉推理。这些模型通常包含两个模块：感知模块和推理模块。感知模块通过卷积神经网络提取图像特征，而推理模块则利用这些特征进行逻辑推理。一些研究还探索了如何将条件化线性调制(FiLM)层引入视觉推理模型，以实现更复杂的推理任务。此外，一些研究还提出了基于注意力机制的模型，以更好地模拟人类视觉推理过程。这些模型通常包含多个模块，如卷积模块、转换器模块和注意力模块，以捕捉不同尺度的视觉信息。这些模型在解决视觉类比推理问题方面取得了良好的效果。此外，一些研究还探索了强化学习在视觉推理中的应用，以实现更准确和透明的视觉推理过程。总的来说，这些研究为视觉推理领域的发展提供了新的思路和方法，有望在未来的视觉理解和智能交互中发挥重要作用。

ChatGLM3 chain1,3:

本文综述了近年来在视觉推理领域的研究进展，包括使用特征线性调制层（FiLM）进行视觉推理，提出基于层次卷积视觉变换器（HCV-ARR）解决Raven's Progressive Matrix（RPM）问题，使用强化路径路由方法进行视觉关系推理，以及提出视觉概念推理网络（VCRNet）实现高级视觉概念之间的推理。这些研究都致力于开发更有效、更通用、更易于应用的视觉推理方法，以应对日益增长的数据和任务需求。这些方法在不同的视觉推理任务和数据集上表现出良好的性能，证明了这些方法的有效性和通用性。

ChatGLM3 chain1,2,3:

视觉概念推理网络（VCRNet）和两项最近的研究，其中一项引入了强化路径路由（RPR）和另一项引入了层次化卷积视觉变换器（HCV-ARR），这两项研究都为视觉推理和图像处理领域的发展做出了贡献。VCRNet提出了一种模块化的多分枝架构，通过推理高级视觉概念来捕获全局上下文，在多个识别任务上表现出优越性能，同时保持参数数量最小。另一方面，RPR利用强化学习在视觉图中学习多步路径，通过结合语言和历史注意力实现透明推理，在视觉关系推理任务上实现了最先进的准确度。HCV-ARR，一种结合卷积神经网络（CNNs）和视觉变换器的层次模型，强调了捕捉细粒度和全局上下文的重要性，通过有效地建模属性关系在RPM问题上超过了其他模型。这些研究展示了深度学习技术在处理复杂图像相关任务方面的灵活性和效率，每项研究都为视觉推理提供了独特的视角。尽管VCRNet专注于对高级概念进行推理，但RPR依赖于基于路径的推理和适应性，而HCV-ARR采用了一种结合卷积神经网络（CNNs）和变换器的层次方法。它们共同突显了视觉推理领域研究的不断发展和互补性。

Reference:

近期在神经网络调节和架构方面的进步显著提升了视觉推理能力，这是一个对深度学习模型传统上构成挑战的领域。特征线性调制（Feature-wise Linear Modulation, FiLM）层的引入标志着一个重大进步，提供了一种多功能的调节方法，使神经网络能够以提高的效率处理复杂的多步骤推理任务，如在CLEVR基准测试中所示。同时，分层ConViT与基于注意力的关系推理器（Hierarchical ConViT with Attention-based Relational Reasoner, HCV-ARR）的开发，解决了Raven's Progressive Matrices (RPMs)中视觉感知和逻辑推理的复杂性，将卷积块用于

低级视觉模式与变换器块用于高级图像语义相结合，并通过一个基于注意力的关系推理器增强推理能力，从而在RPM数据集上表现出色。此外，另一项研究提出通过将视觉关系推理概念化为路径路由任务来学习其动态性。这种方法采用加强的路径路由方法，从输入图像构建一个结构化的视觉图，并利用基于强化学习的模型来导航图中的路径（节点序列），由输入句子指导以推导出推理结果。在这些进步的基础上，视觉概念推理网络（Visual Concept Reasoning Networks, VCRNet）的引入采用了在卷积神经网络中用于视觉识别任务的分裂-转换-合并策略，以实现高级视觉概念之间的推理。通过将每个分支与一个视觉概念关联，并通过注意力模块派生紧凑的概念状态，VCRNet在通过基于图的交互更新概念状态后，适应性地调节局部描述符。这个模型通过分裂-转换-关注-交互-调节-合并阶段描述，展示了高度模块化架构在增强视觉识别任务方面的有效性，同时参数增加最小。总体而言，这些方法强调了整合多样化的神经网络机制——如特征调制、层次处理、基于注意力的推理和基于图的交互——以及像加强路径路由这样的创新方法，以应对视觉推理的多方面挑战的重要性。这一趋势突显了向更细腻和有效模型的转变，强调了推理的动态性和在模拟人类推理能力中建模认知过程的关键作用。

基于思维链的跨语言多文档摘要生成示例2：

Source Document:

Article1: *Efficacy and safety of sintilimab plus albumin-bound-paclitaxel in recurrent or metastatic cervical cancer: a multicenter, open-label, single-arm, phase II trial*

.....

Article2: *Comparative effectiveness of ustekinumab vs. vedolizumab for anti-TNF-naïve or anti-TNF-exposed Crohn's disease: a multicenter cohort study*

.....

Article3: *Efficacy and safety of second-line therapy of docetaxel plus ramucirumab after first-line platinum-based chemotherapy plus immune checkpoint inhibitors in non-small cell lung cancer (SCORPION): a multicenter, open-label, single-arm, phase 2 trial*

.....

Article4: *Fecal microbiota transplantation plus tislelizumab and fruquintinib in refractory microsatellite stable metastatic colorectal cancer: an open-label, single-arm, phase II trial (RENMIN-215)*

.....

GPT-3.5 Direct:

这些研究分别探讨了不同治疗方案在不同类型的癌症患者中的疗效和安全性。第一篇研究评估了Sintilimab联合白蛋白结合紫杉醇治疗宫颈癌的效果，显示出良好的抗肿瘤活性和可管理的毒性。第二篇研究比较了Ustekinumab和Vedolizumab在克罗恩病患者中的疗效，发现Ustekinumab在临床和客观结果上优于Vedolizumab。第三篇研究探讨了在非小细胞肺癌患者中，第一线免疫治疗后使用多西紫杉醇加拉姆西鲁单抗的二线治疗的疗效和安全性。最后一篇研究研究了在微卫星稳定性结直肠癌患者中，粪便菌群移植联合替雷利珠单抗和富奎替尼的疗效，显示出改善生存和可管理的安全性。这些研究为不同癌症类型的治疗提供了新的治疗选择，并强调了进一步进行大规模随机对照试验的重要性。

GPT-3.5 chain1,3:

这些研究均探讨了不同治疗方案在晚期癌症患者中的疗效和安全性。一项研究评估了在复发或转移性宫颈癌患者中将sintilimab与白蛋白紧密结合紫杉醇联合应用的效果，显示出潜在的治疗前景。另一项研究比较了ustekinumab和vedolizumab在克罗恩病患者中的效果，结果表明ustekinumab可能是更有效的选择。还有一项研究探讨了docetaxel加ramucirumab作为非小细胞肺癌患者的二线治疗选择，显示出良好的抗肿瘤活性和可管理的安全性。最后一项研究则研究了将粪便菌群移植与tislelizumab和fruquintinib结合应用于难治性微卫星稳定性结直肠癌患者的疗效，强调了肠道菌群在免疫治疗疗效中的重要作用。这些研究为癌症治疗领域提供了新的治疗选择和见解，需要进一步的临床验证和研究。

GPT-3.5 chain1,2,3:

在一系列探索不同癌症治疗方案的疗效和安全性的多样化研究中，涌现出了创新的方法。一项研究调查了将粪便微生物移植（FMT）与替雷利珠单抗和富奎替尼结合在难治性微卫星稳定（MSS）转移性结直肠癌（mCRC）患者中的应用，显示出有希望的结果，中位无进展生存期（PFS）为9.6个月，并且在反应者中发现了肠道微生物组成的改变，突显了肠道菌群在免疫治疗疗效中的作用。另一项研究关注了多西紫杉醇加拉莫西单抗作为非小细胞肺癌（NSCLC）患者的二线治疗，显示出34.4%的客观缓解率和可控的不良事件，而另一项研究探讨了辛地利单抗和白蛋白结合紫杉醇用于复发性或转移性宫颈癌，显示出有希望的缓解率和可耐受的副作用。此外，一项比较克唑替尼单抗和维多利珠单抗在克罗恩病患者中的研究发现，对于TNFi未曾接受和TNFi曾接受的个体，克唑替尼单抗在第26周和第52周具有更优越的临床结果。这些研究共同呈现了各种癌症的宝贵新治疗选择，强调了个性化治疗策略的重要性，并为在临床实践中进一步探索提供了有希望的途径。

Reference:

近期的临床研究为慢性疾病的创新疗法开辟了道路，展示了针对性治疗在各种癌症疾病中的潜力。其中，日本进行的第二阶段SCORPION试验在非小细胞肺癌（NSCLC）的治疗中显示出了希望，该试验通过将多西他赛与雷莫芦单抗结合作为二线疗法，达到了34.4%的客观响应率（ORR），且安全性可控。同样，将针对PD-1的抗体西妥昔单抗与白蛋白结合型紫杉醇用于复发或转移性宫颈癌的组合，显示出了显著的抗肿瘤活性，ORR为44.4%，疾病控制率为88.9%，尽管近半数参与者出现了可管理的高级别不良事件。在胃肠疾病领域，一项多中心队列研究强调了在治疗克罗恩病方面，乌司他单抗相比于维多利单抗的优越效果，分别在26周和52周时达到了临床和无需激素的缓解，且安全性相当。在这一系列针对性疗法中，研究“粪便微生物群移植加上替雷利珠单抗和复仑替尼治疗难治性微卫星稳定型转移性结直肠癌”引入了一种针对难治性微卫星稳定型转移性结直肠癌（mCRC）的创新治疗方案，结合了粪便微生物群移植（FMT）与抗PD-1和抗血管生成疗法。该试验报告了中位无进展生存期（PFS）为9.6个月，总生存期（OS）为13.7个月，且安全性可控，凸显了肠道微生物组在提高免疫疗效中的重要作用。这些研究共同强调了结合针对性疗法以改善晚期癌症和慢性疾病患者结果的潜力，暗示了向更个性化和有效的治疗策略转变。

C 数据集示例

C.1 人工智能领域

下面簇中包含4篇文档:

Article1:

Title: FiLM: Visual Reasoning with a General Conditioning Layer

（该文档共904个单词）

Article2:

Title: Hierarchical ConViT with Attention-Based Relational Reasoner for Visual Analogical Reasoning

（该文档共1192个单词）

Article3:

Title: Learning the Dynamics of Visual Relational Reasoning via Reinforced Path Routing

（该文档共918个单词）

Article4:

Title: Visual Concept Reasoning Networks

（该文档共693个单词）

跨语言多文档摘要:

近期在神经网络调节和架构方面的进步显著提升了视觉推理能力，这是一个对深度学习模型传统上构成挑战的领域。特征线性调制（Feature-wise Linear Modulation, FiLM）层的引入

标志着一个重大进步，提供了一种多功能的调节方法，使神经网络能够以提高的效率处理复杂的多步骤推理任务，如在CLEVR基准测试中所示。同时，分层ConViT与基于注意力的关系推理器（Hierarchical ConViT with Attention-based Relational Reasoner, HCV-ARR）的开发，解决了Raven's Progressive Matrices (RPMs)中视觉感知和逻辑推理的复杂性，将卷积块用于低级视觉模式与变换器块用于高级图像语义相结合，并通过一个基于注意力的关系推理器增强推理能力，从而在RPM数据集上表现出色。此外，另一项研究提出通过将视觉关系推理概念化为路径路由任务来学习其动态性。这种方法采用加强的路径路由方法，从输入图像构建一个结构化的视觉图，并利用基于强化学习的模型来导航图中的路径（节点序列），由输入句子指导以推导出推理结果。在这些进步的基础上，视觉概念推理网络（Visual Concept Reasoning Networks, VCRNet）的引入采用了在卷积神经网络中用于视觉识别任务的分裂-转换-合并策略，以实现高级视觉概念之间的推理。通过将每个分支与一个视觉概念关联，并通过注意力模块派生紧凑的概念状态，VCRNet在通过基于图的交互更新概念状态后，适应性地调节局部描述符。这个模型通过分裂-转换-关注-交互-调节-合并阶段描述，展示了高度模块化架构在增强视觉识别任务方面的有效性，同时参数增加最小。总体而言，这些方法强调了整合多样化的神经网络机制——如特征调制、层次处理、基于注意力的推理和基于图的交互——以及像加强路径路由这样的创新方法，以应对视觉推理的多方面挑战的重要性。这一趋势突显了向更细腻和有效模型的转变，强调了推理的动态性和在模拟人类推理能力中建模认知过程的关键作用。

C.2 医学领域

下面簇中包含3篇文档:

Article1:

Title: Estimates of the global burden of non-Hodgkin lymphoma attributable to HIV: a population attributable modeling study

（该文档共383个单词）

Article2:

Title: The burden of non-communicable diseases among people living with HIV in Sub-Saharan Africa: a systematic review and meta-analysis

（该文档共267个单词）

Article3:

Title: Burden and risk factors of chronic obstructive pulmonary disease in Sub-Saharan African countries, 1990-2019: a systematic analysis for the Global Burden of disease study 2019

（该文档共355个单词）

跨语言多文档摘要:

近期研究揭示了撒哈拉以南非洲（SSA）面临的紧迫健康挑战，重点关注了艾滋病毒感染者（PLHIV）中非传染性疾病（NCDs）的负担以及慢性阻塞性肺疾病（COPD）构成的日益严峻的公共卫生挑战。一项综合性的元分析揭示，2019年全球新增非霍奇金淋巴瘤（NHL）病例中有6.92%可归因于HIV，东部和南部非洲以及东欧和中亚的流行率显著，凸显了PLHIV中NHL风险增加。此外，一项系统评价和元分析发现，撒哈拉以南非洲的PLHIV中高血压、抑郁症、糖尿病和其他NCDs的比率值得注意，强调了加强卫生系统和综合医疗服务的必要性。与此同时，一项分析1990年至2019年全球疾病负担研究数据的研究发现，SSA的COPD流行率急剧增加，2019年的流行病例达到1030万，自1990年以来增加了117%。该研究指出，固体燃料的家庭空气污染是COPD的主要风险因素，尤其是在中部撒哈拉以南非洲，通过残疾和过早死亡显著地加剧了疾病负担。这两组发现强调了制定全面健康策略的迫切需要，这些策略旨在应对HIV等传染病和NCDs（包括COPD）日益增长的影响的双重挑战，以改善SSA的整体健康结果。这种综合方法与联合国艾滋病规划署的"90-90-90"快速行动目标相一致，并为该地区针对性的公共卫生干预和政策制定提供了关键见解，凸显了传染病和NCDs在塑造公共卫生优先事项中的相互关联性。