arXiv:2304.03325v1 [cs.CL] 6 Apr 2023

CHATGPT-CRAWLER：探究 CHATGPT 是否真的懂自己在说什么。*

Aman Rangapur, Haoran Wang学
生研究人员，伊利诺伊理工学院，美国
芝加哥。{arangapur,
hwang219}@hawk.iit.edu

摘要

大语言模型在各种任务上的出色表现引起了广泛关注。在这些模型中，由 OpenAI 开发的 ChatGPT 在早期采用者中极受欢迎，他们甚至将其视为客户服务、教育、医疗保健和金融等诸多领域的一项颠覆性技术。了解这些初期用户的看法至关重要，因为这能为该技术在不同领域的潜在优势、劣势以及成功或失败提供宝贵见解。本研究考察了 ChatGPT 针对不同对话式问答语料库生成的回答。研究采用 BERT 相似度分数将这些回答与正确答案进行比较，并获取自然语言推理（NLI）标签。此外，研究还计算并比较了评估分数，以确定 GPT-3 和 GPT-4 的整体表现。研究还识别了 ChatGPT 未能正确回答问题的实例，从而为模型可能容易出错的领域提供了见解。

关键词 ChatGPT · 自然语言推理 · BERT · 语料库 · 对话式问答

1 引言

近年来，大型语言模型彻底改变了自然语言处理领域。它们拥有复杂的神经网络模型，能够生成具有特定语调和内容的文本。它们接受海量数据的训练，以预测最适合接续给定提示的文本，从而产生听起来自然的输出。

在这些模型中，由 OpenAI 开发的 ChatGPT 凭借其在各种语言任务中的出色表现而广受欢迎。ChatGPT 是一种大型预训练语言模型，利用深度学习技术针对自然语言查询生成回答[1]。它理解并生成连贯回答的能力，使其成为广泛应用中的宝贵工具，包括聊天机器人、语言翻译和问答系统。ChatGPT 在多个方面不同于传统聊天机器人：它能够回忆与用户先前的对话、拒绝不合适的请求，并纠正不准确的回答。ChatGPT 可以针对复杂查询提供详细的答案、建议和解释，例如编程、优化和布局问题[2]。凭借卓越的能力，ChatGPT 在发布后的第一周内便获得了超过一百万名用户，超过了其他知名在线平台。ChatGPT 已在海量文本数据语料库上完成预训练，并展现出针对自然语言输入生成类人文本的非凡能力[3]。

ChatGPT 的预训练过程包含三个阶段：无监督预训练、监督微调，以及通过“人在回路”进一步微调模型，使其更好地理解人类指令。在无监督预训练阶段，ChatGPT 会在海量文本数据集上接受训练，以学习自然语言的模式和结构[4]。这一过程包括让模型接受多种语言任务的训练，例如语言建模、掩码语言建模和下一句预测。

*https://github.com/IIT-NLP-RESEARCH/ChatGPT-Crawler

ChatGPT 是一项突破性技术，有望改变我们与机器互动的方式。它可用于多种应用，包括聊天机器人、语言翻译和文本摘要生成[5]。许多行业已经采用这项技术，包括电子商务、客户服务和医疗保健，以提供个性化、高效的客户支持[6]。

本文旨在探究 ChatGPT 的性能及其在不同领域的潜在用途。研究分析了 ChatGPT 对不同数据集作答的准确性和一致性，并考察了模型容易出错的方面。我们旨在分析 ChatGPT 在对话式问答任务中的输出可靠性[7, 8, 9, 10]。为此，我们开发了一套能够大规模生成回答的流程，并对 ChatGPT 的回答与现有问答语料库进行了全面比较。我们计算了 ChatGPT 回答的 Jaccard[11], BLEU[12],ROUGE[13], 美国国家标准与技术研究院[14], METEOR[15], BART[16] 和 TER[17] 分数，以评估其输出的黄金比例和流畅性。

2 背景研究

对于涉及自然语言处理的挑战，基础模型如今已成为常见的研究与应用范式。基础模型经过海量数据训练，在情感分析、问答、自动诊断、逻辑推理和序列标注等多种下游任务上的表现显著优于早期模型。早期研究从多个方面评估了 ChatGPT[18]。[19]建议对 ChatGPT 在多任务、多模态和多语言的各类任务上的表现进行评估。他们发现，尽管 ChatGPT 在大多数任务上的表现尚可，但在低资源任务上仍有困难。 [20] 在2023年进行了类似的实证评估。 [21] 具体开展了多项评估。关于研究结果， [21] 发现 ChatGPT 在序列标注等细粒度下游任务上的表现不佳。 [22] 和 [23]指出，ChatGPT 及其他大型模型如同双刃剑，应受到监测。 [24]开展了伦理方面的研究。人机交互（HCI）[25], 、教育 [26],[27], [28], [29], 、医疗 [30], 和写作 [31] 均在 [32]的研究中得到讨论和反思。据我们所知，目前针对对话式问答语料库的研究并不多。对话式问答语料库旨在模拟人类对话，通常包含闲聊、幽默和情感等多种对话元素。这使聊天机器人更难作答，因为它们不仅需要理解所说词语的字面含义，还要理解其背后的语境、语气和意图。此外，对话式问答往往存在一定程度的歧义和不确定性，而这在常规问答中并不常见。例如，在对话中，一方可能会要求澄清或补充信息，也可能表达不确定或困惑。聊天机器人需要能够应对这类情况并作出恰当回应。本文的研究结果揭示了 ChatGPT 在对话式问答语料库上的擅长之处与不足。

优势： 理解上下文的能力：ChatGPT 能够理解问题提出时所处的上下文，并生成与该上下文相适宜的回答。

自然语言处理：ChatGPT 能够理解自然语言并生成回复，让对话更加自然、有吸引力。

灵活性：ChatGPT 能够处理广泛的主题和问题，并生成既信息丰富又引人入胜的回答。

弱点： 缺乏具体知识：尽管 ChatGPT拥有海量知识，但它可能并不具备某些主题的具体知识。这会导致回答不准确。

缺乏常识：ChatGPT 不具备与人类相同水平的常识，这会导致其给出.....
那些在技术上　y 正确，但与对话语境不符。
难以应对歧义　　：ChatGPT 难以理解含糊不清的问题并作出回应
　　或陈述，因此　导致回答不准确或毫无意义。

3 方法论

为了开展这项研究，我们设计并实现了一条利用 ChatGPT 生成大规模回答的处理流程。该流程由两个主要模块组成，即问题生成模块和回答生成模块。问题生成模块会生成一组多样化的问题，以代表典型的对话式问答任务。为确保涵盖广泛的主题，我们采用了多种技术，例如改写、扩充，以及从现有问答语料库中进行采样。随后，我们使用生成的问题向 ChatGPT 发起查询，以生成回答。
响应生成模块利用 ChatGPT 的语言生成能力，为问题生成模块提出的问题生成响应。为确保响应的质量，我们采用了束搜索和top-k 采样等多种技术。随后，我们对生成的响应与问题的相关性和具体性进行了评估。为评估我们流程的有效性，我们在四个广泛使用的数据集上对其进行了评估，

即 CoQA [7], DialFact [8], FaVIQ [9], 和 CoDAH [10]。这些数据集是对话式问答任务的热门基准，涵盖广泛的主题和领域。

为了评估ChatGPT 响应的质量，我们采用了 BLEU、METEOR、BART、美国国家标准与技术研究院、Jaccard、ROUGE-L和TER分数等多种指标。这些指标用于衡量生成响应的准确性、流畅性和连贯性。我们还在这些数据集上将我们的流程与现有的最先进模型进行了性能比较。

我们的流程还展现出很强的可扩展性和灵活性，能够处理各种各样的对话式问答任务。总体而言，我们的研究证明了利用 ChatGPT 为对话式问答任务生成大规模回答的有效性。我们的流程提供了一种稳健且可扩展的解决方案，可用于生成高质量回答，应用范围广泛，包括虚拟助手、客户服务聊天机器人和对话代理。

4 个数据集

CoQA（对话式问答）是一个用于开发和评估对话式问答系统的数据集[7]。CoQA 语料库包含超过127,000 个问题和答案，其中包括人类与机器围绕给定文本段落展开的对话。这些对话旨在类似于自然对话，提问者可以提出后续问题，以澄清自己对该段落的理解。

DialFact 是一个于 2020[8] 推出的自然语言处理 (NLP) 数据集。它专为对话场景中的事实核查而设计，旨在判断对话中提出的主张是否属实。该数据集包含 10,221 段对话，每段对话都包括一位参与者提出的主张，以及另一位参与者给出的回应，表明该主张为真、为假或未知。这些对话收集自互联网，涵盖广泛的主题，包括政治、健康和科学。DialFact 数据集的独特之处在于，它聚焦于对话事实核查，而非传统的新闻文章或其他书面文本事实核查。因此，它非常适合用于开发和评估对话代理，帮助用户判断对话中提出的主张是否属实。

FaVIQ（隐式查询中的事实验证）是一个自然语言处理（NLP）语料库，旨在用于对话场景中的事实验证[9]。该数据集包含 3,772 段对话，每段对话都包含一个或多个需要验证的隐式主张。这些主张涉及多个主题，包括科学、政治和娱乐。该数据集包含真假主张，旨在对 NLP 模型构成挑战。

CoDAH（用于自动幽默识别的常识数据）是一个用于开发和评估幽默识别模型的自然语言处理（NLP）语料库[10]。该语料库于 2018 年推出，其独特之处在于关注常识幽默识别，而非较为直白的幽默形式。CODAH 数据集包含 38,269 条短文本，每条都包含一个幽默句子或短语。该数据集分为两部分：包含 28,269 条文本的训练集，以及包含 10,000 条文本的测试集。这些文本收集自社交媒体平台，涵盖体育、政治和娱乐等广泛主题。CODAH 语料库对幽默识别模型而言颇具挑战，因为模型需要具备对常识知识的深入理解，以及识别微妙幽默形式的能力。表 1 展示了从 ChatGPT 获得的回答数量。

CoQA	DialFact	FaVIQ	CoDAH
991	3147	58	1887

表1：在每个语料库上获得的响应数量。

5 个结果

我们的研究评估了 ChatGPT 在对话式问答任务中的潜力，并确定了其局限性。结果显示，ChatGPT-3 能够生成高质量的回答，平均 BLEU 得分为0.79，平均 ROUGE- L 得分为0.53。然而，我们还观察到，ChatGPT-3 的回答可能较为笼统且不相关，从而降低了其在实际应用中的实用性。

为应对这些局限性，我们研究了新发布的 GPT-4模型在生成更具相关性和针对性的回答方面的有效性。我们的评估结果表明，GPT-4在准确性、相关性和一致性方面均优于 ChatGPT-3。GPT-4 模型在生成更加连贯且与上下文相关的回答方面展现出显著改进，使其成为对话式问答任务的有力候选模型。

表 2和表3展示了ChatGPT-3 和GPT-4 在各种对话式问答语料库上的评估分数，包括 CoQA、DialFact、FaVIQ和CoDAH。我们的结果表明，GPT-4 在所有指标上都取得了更高的分数，包括 BLEU、ROUGE-L和METEOR，证明了其生成高质量回答的优越性。由于这些指标并不完美，并且并不总是与人工判断的文本相似度或准确性保持一致。因此，计算了人工评估分数（1-含义相似，0-含义不相似），以便对生成的回答进行更全面的评估。

尽管 ChatGPT-3 和 GPT-4 取得了令人鼓舞的成果，我们也观察到这些模型存在一些局限性。具体而言，我们注意到，ChatGPT-3 的回答可能前后不一致，有时还会产生误导，尤其是在依据相同上下文回答同一个问题时。这种不一致性可能会降低模型在实际应用中的可靠性，而准确且一致的回答至关重要。不过，我们发现 GPT-4 解决了这一问题，能够生成更加一致、可靠的回答。图 1 展示了 GPT-4 的改进。
总之，我们的研究展现了 ChatGPT 及其后续版本 GPT-4 在为对话式问答任务生成大规模回答方面的潜力。尽管 ChatGPT-3 展现了令人鼓舞的成果，我们也发现了有待解决的局限性。我们对 GPT-4 的评估表明，它在生成相关性更高、更具体且更加一致的回答方面取得了显著进步，因此有望应用于对话式问答任务。这些发现对开发依赖自然语言处理与理解的对话代理和虚拟助手具有重要意义。

	CoQA	DialFact	FaVIQ	CoDAH
BLEU	0.04	0.02	0.017	0.004
ROUGE-1	0.11	0.04	0.032	0.066
ROUGE-L	0.13	0.04	0.034	0.062
METEOR	0.214	0.191	0.018	0.13
TER	12.47	10.19	3.36	15.73
Jaccard	0.0005	0.0001	0.0002	0.065
BART	-4.061	-4.05	-6.556	-5.257
NIST	0.006	0.029	0.106	0.0001
人工评估	0.84	0.91	0.93	0.78

(a) ChatGPT-3 在不同对话问答上的得分。

	CoQA	DialFact	FaVIQ	CoDAH
BLEU	0.09	0.06	0.033	0.006
ROUGE-1	0.30	0.06	0.04	0.068
ROUGE-L	0.31	0.06	0.042	0.079
METEOR	0.403	0.338	0.039	0.24
TER	10.23	10.01	3.11	12.89
Jaccard	0.0008	0.0001	0.0002	0.065
BART	-3.90	-4.11	-5.91	-5.01
NIST	0.008	0.032	0.223	0.0002
人工评估	0.93	0.95	0.98	0.82

(b) ChatGPT-4 在不同对话问答上的得分。

表 2：每个对话问答语料库的评估分数。

	CoQA	DialFact	FaVIQ	CoDAH
BERT	0.87	0.90	0.96	0.75

(a) ChatGPT-3 在不同语料库上的平均 BERT 分数。

	CoQA	DialFact	FaVIQ	CoDAH
BERT	0.79	0.88	0.91	0.66

(b) ChatGPT-4 在不同语料库上的平均 BERT 分数。

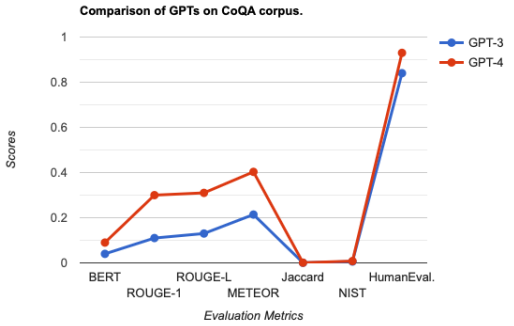
表3：平均 BERT 相似度分数。

5.1 案例研究

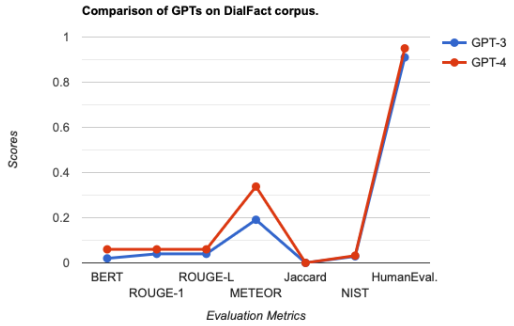
自然语言处理（NLP）系统的发展，因 GPT-3 和 GPT-4 等大规模预训练语言模型的出现而加速。这些模型能够在各种领域和任务中生成流畅、连贯的文本。然而，评估这些模型的质量和性能并非易事。人们提出了不同的评估指标，用以衡量文本生成的多个方面，例如流畅性、连贯性、相关性、信息量、多样性和事实一致性。然而，对于哪些指标最适合用于比较不同模型或设置，目前尚未达成共识。在本案例研究中，向 GPT-3 和 GPT-4 提供了一个背景，以比较 GPT-4 的能力得到了多大程度的提升。

从前，在一座农舍附近的谷仓里，住着一只名叫 *Cotton* 的白色小猫。*Cotton* 住在谷仓高处一个舒适温暖的地方，农夫的马儿都睡在她下面。不过，*Cotton* 并非独自住在谷仓上方的小窝里，哦，不是的。她和妈妈以及另外五个姐妹共睡一张干草床。她的姐妹们都像 *Cotton* 一样可爱又毛茸茸的。但她是其中唯一一只白猫。其余的姐妹全都是橘色的，身上有漂亮的白色虎斑纹，就像 *Cotton* 的妈妈一样。

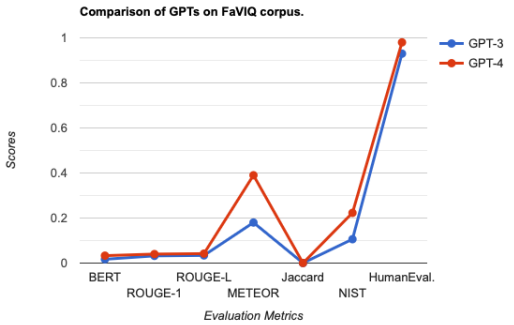
与众不同让 *Cotton* 很伤心。她常常希望自己看起来和家人一样。所以有一天，*Cotton* 发现了一罐老农夫的橙色油漆，就用它把自己涂成了和家人一样的颜色。当妈妈和姐妹们发现了她.....



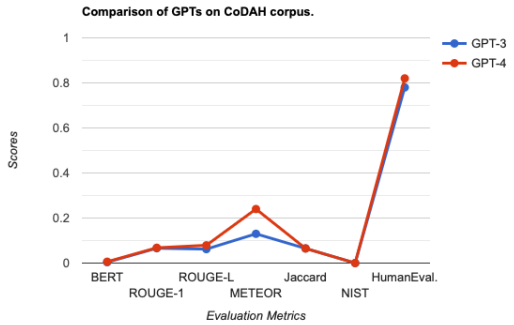
（a）GPT-3 和 GPT-4 在 CoQA 语料库上的性能折线图。



（b）GPT-3 和 GPT-4 在 Dial-Fact 语料库上的表现折线图。



（c）GPT-3 和 GPT-4 在 FaVIQ 语料库上的折线图。



(d)GPT-3 和 GPT-4 在 CoDAH 语料库上的表现折线图。

图1：GPT-3 和 GPT-4 在不同语料库上的性能比较。

她突然笑了起来。

“你在做什么，*Cotton*？！”“我只是想变得更像你。”*Cotton* 的妈妈用脸蹭了蹭 *Cotton*的脸，说道：“哦，*Cotton*，可你的毛皮是如此漂亮而特别，就像你一样。我们绝不会希望你变成其他样子。”说完，*Cotton* 的妈妈把她抱起来，扔进了一大桶水里。

当 *Cotton* 出来时，她又恢复了本来的样子 . 她的姐妹们舔她的脸，直到 *Cotton* 的毛全都干了

.“别再那样做了，*Cotton*！”他们都喊道。“下次你可能会弄乱你那漂亮洁白的毛，而我们可不想那样！”然后*Cotton*想道：“我改变主意了。我喜欢与众不同。”

表 4 展示了操纵 ChatGPT 时对查询的响应。

6 讨论

我们的研究结果表明，ChatGPT 在对话式问答领域展现出潜力，但也揭示了需要改进之处，以提高其回答的准确性和具体性。为实现这一目标，未来的研究可以考虑纳入外部知识来源，例如知识库，并提出对 ChatGPT 生成文本进行事实核查的方法。此外，研究针对对话式问答任务微调 ChatGPT 的替代方法也可能有所裨益，因为这或许能够带来更理想的结果。

7 结论

在本研究中，我们对 ChatGPT 的回答与现有问答语料库进行了全面比较，以分析其输出在对话式问答任务中的可靠性和适用性。我们开发了一套流程，用于生成大规模回答，并计算 ChatGPT 回答的 BLEU、ROUGE 和 TER 分数。我们的结果表明，ChatGPT 在对话式问答任务中潜力巨大，同时也凸显了最新 GPT-4 模型的改进。我们

查询	回答	ChatGPT-3 回复
Cotton觉得自己看起来不一样而感到开心吗? 不，你错了，Cotton 很伤心。	Yes	是的，Cotton 很开心。 是的，Cotton 很难过。
她独自居住吗? 你能纠正自己吗?	No	No, Cotton并非独自生活。 Cotton 独自住在她的小房子里。

(a) 与 ChatGPT-3 的对话

查询	回答	ChatGPT-4 的回复
Cotton 对自己看起来不一样感到高兴吗? 不，你错了，Cotton 当时很伤心。	Yes	是的，Cotton 很开心。 抱歉，但 Cotton 当时很开心。
她是独自生活的吗? 你能纠正一下吗?	No	No, Cotton 并非独自生活。 Cotton 和她的家人住在一起。

(b) 与 ChatGPT-4 的对话

表 4：基于 GPT-3 和 GPT-4 段落提出的查询的回答。

希望我们的研究将有助于开发基于 ChatGPT 等大规模语言模型的更有效、更可靠的对话式问答系统。

参考文献

[1] Mubin Ul Haque、IsuruDharmadasa、Zarrin Tasnim Sworna、RoshanNamalRajapakse和Hussain Ahmad。“我认为这是最具颠覆性的技术”：使用推特数据探索ChatGPT早期采用者的情感，2022年。[2] Andrianos Michail、Stefanos Konstantinou和Simon Clematide。Uzh_clyp在SemEval-2023 任务9中：用于推文亲密度预测跨语言学习的头优先微调和ChatGPT数据生成，2023年。[3] Haocong Rao、Cyril Leung和Chunyan Miao。ChatGPT能评估人类人格吗？一个通用评估框架，2023年。[4] Aakash Ahmad、MuhammadWaseem、Peng Liang、Mahdi Fehmideh、Mst Shamima Aktar和Tommi Mikkonen。迈向利用ChatGPT进行人机协作软件架构设计，2023年。[5] 什么是人工智能聊天机器人现象ChatGPT？它能取代人类吗？，2022年12月。[6] EdFelten、Manav Raj和Robert Seamans。像ChatGPT这样的语言模型开发者将如何影响职业和行业？，2023年。[7] Siva Reddy、Danqi Chen和ChristopherD.Manning。CoQA：一项对话式问答挑战，2018年。[8] PrakharGupta、Chien-Sheng Wu、WenhaoLiu和Caiming Xiong。DialFact：一个用于对话事实核查的基准，2021年。[9] Jungsoo Park、Sewon Min、Jaewoo Kang、LukeZettlemoyer和HannanehHajishirzi。FaVIQ：从信息寻求型问题进行事实验证，2021年。[10] Michael Chen、Mike D’Arcy、AlisaLiu、JaredFernandez和DougDowney。CoDAH：一个由对抗性方式编写的常识问答数据集，2019年。[11] Zifu Wang和MatthewB. Blaschko。Jaccard度量损失：使用软标签优化Jaccard指数，2023年。[12] KishorePapineni、Salim Roukos、ToddWard和Wei-Jing Zhu。BLEU：一种机器翻译自动评估方法。在计算语言学协会第*40*届年会论文集中，ACL’02，第311–318页，美国，2002年。计算语言学协会。[13] Chin-Yew Lin。ROUGE：一个用于自动评估摘要的工具包。在文本摘要分支拓展中，第74–81页，巴塞罗那，西班牙，2004年7月。计算语言学协会。[14] Y.Zhang、StephanVogel和AlexanderH. Waibel。解读BLEU/NIST分数：我们需要多大的改进才能获得更好的系统？在国际语言资源与评估会议中，2004年。[15] Satanjeev Banerjee和Alon Lavie。METEOR：一种用于机器翻译评估的自动度量指标，与人工判断具有更高的相关性。在*ACL*内在与外在评估指标研讨会论文集中

机器翻译和/或摘要生成，第65–72页，安娜堡，密歇根州，2005年6月。计算语言学协会。[16] Weizhe Yuan、GrahamNeubig 和 Pengfei Liu。Bartscore：将生成文本作为文本生成进行评估，2021年。[17] Abhaya Agarwal 和 AlonLavie。Meteor、M-BLEU和 M-TER：与机器翻译输出的人工排名高度相关的评估指标。载于第三届统计机器翻译研讨会论文集，第115–118页，哥伦布，俄亥俄州，2008年6月。计算语言学协会。[18] Eva A M vanDis、Johan Bollen、Willem Zuidema、Robert van Rooij和Claudi L Bockting。ChatGPT：研究的五项优先事项。《自然》，614(7947):224–226，2023年2月。[19] Yejin Bang、Samuel Cahyawijaya、Nayeon Lee、Wenliang Dai、Dan Su、Bryan Wilie、Holy Lovenia、Ziwei Ji、Tiezheng Yu、Willy Chung、Quyet V. Do、YanXu和Pascale Fung。对ChatGPT在推理、幻觉和交互性方面进行多任务、多语言、多模态评估，2023年。[20] Roberto Gozalo-Brizuela 和 Eduardo C. Garrido-Merchan。ChatGPT并非你所需要的一切。大型生成式人工智能模型的前沿综述，2023年。[21] ChengweiQin、Aston Zhang、Zhuosheng Zhang、Jiaao Chen、MichihiroYasunaga 和Diyi Yang。ChatGPT 是通用型自然语言处理任务求解器吗？，2023年。[22] Philipp Hacker、Andreas Engel 和 MarcoMauer。监管 ChatGPT 和其他大型生成式人工智能模型，2023年。[23] Yiqiu Shen、Laura Heacock、Jonathan Elias、Keith D Hentel、Beatriu Reig、George Shih和Linda Moy。ChatGPT 和其他大语言模型是一把双刃剑。《放射学》，第230163页，2023年1月。[24] Terry Yue Zhuo、Yujin Huang、Chunyang Chen 和 ZhenchangXing。探索ChatGPT 的人工智能伦理：一项诊断性分析，2023年。[25] Wilbert Tabone和JoostdeWinter。使用ChatGPT 进行人机交互研究：入门指南，2023年03月。[26] Mohammad Khalil和 ErkanEr。ChatGPT 会让你被发现吗？重新思考抄袭检测，2023年。[27] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4312358/。探索人工智能在提升学业表现方面的作用：ChatGPT 案例研究。2023年1月。[28] TeoSusnjak。ChatGPT：在线考试诚信的终结？，2022年。[29] Biyang Guo、Xin Zhang、Ziyuan Wang、Minqi Jiang、Jinran Nie、Yuxuan Ding、JianweiYue 和Yupeng Wu。ChatGPT 与人类专家有多接近？比较语料库、评估与检测，2023年。[30] Katharina Jeblick、Balthasar Schachtner、Jakob Dexl、AndreasMittermeier、AnnaTheresa Stüber、Johanna Topalis、Tobias Weber、Philipp Wesp、Bastian Sabel、Jens Ricke和MichaelIngrisch。ChatGPT让医学更易于理解：关于简化放射学报告的探索性案例研究，2022年。[31] Biyang Guo、Xin Zhang、Ziyuan Wang、Minqi Jiang、Jinran Nie、Yuxuan Ding、JianweiYue 和 Yupeng Wu。ChatGPT 与人类专家有多接近？比较语料库、评估与检测，2023年。[32] Som Biswas。ChatGPT与医学写作的未来。《放射学》，第223312页，2023年2月。

8 附录

8.1 评估指标

8.1.1 BLEU

双语评估替补（BLEU）分数是一种用于评估机器生成文本质量的指标[12]。BLEU 用于衡量机器生成文本与参考文本之间的相似度。它常用于机器翻译，通过将译文与参考译文进行比较来评估译文质量。BLEU 根据机器生成文本与参考文本之间共有的 n 元语法（连续的词序列）计算分数。要计算 BLEU 分数，系统生成的输出会使用一种改进的精确率计算方式，与一个或多个参考文本进行比较。这种改进的精确率会考虑某个 n 元语法在参考文本中出现的次数，以及它在系统生成的输出中出现的次数。分数范围为 0 到 1，分数越高，表示系统生成的输出与参考文本越匹配。

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

其中：

- BP 是简洁度惩罚项,
- N 是最大 n-gram 长度,
- w_n 是分配给 n 元语法精确率得分的权重，并且
- p_n 是 n-gram 精确率得分。

简洁度惩罚项 BP 的计算方式如下:

$$\text{BP} = \begin{cases} \exp(1 - \frac{ref}{out}) & \text{if } out < ref \\ 1 & \text{if } out \geq ref \end{cases} \tag{2}$$

n-gram 精确率 p_n 计算如下:

$$p_n = \frac{\sum_{\text{n-grams } g \in \text{output}} \min(count_{ref}(g), count_{output}(g))}{\sum_{\text{n-grams } g \in \text{output}} count_{output}(g)} \tag{3}$$

其中， $count_{ref}(g)$是 n-gram g 在参考译文中出现的次数，而 $count_{output}(g)$ 是 n-gram g 在机器生成的译文中出现的次数。
权重 w_n 通常设为 $1/N$，其中 N 是最大的 n 元语法长度。

8.1.2 ROUGE

面向召回率的摘要评估助手（ROUGE）是一组用于自然语言处理、评估机器生成的摘要质量的评估指标[13]。ROUGE 计算机器生成的摘要与一个或多个参考摘要之间的重叠程度。它通过比较机器生成的摘要与参考摘要之间重叠的n 元语法（连续的词序列）数量来衡量摘要质量。与 BLEU一样，ROUGE分数范围为0 到 1，分数越高表示系统生成的摘要与参考摘要的匹配程度越高。ROUGE 分数通常以 F 值报告，同时考虑精确率和召回率。

$$\text{ROUGE} = \frac{\sum_{r \in R} \sum_{n \in N} count_{match}(n, r)}{\sum_{r \in R} \sum_{n \in N} count_{ref}(n, r)} \tag{4}$$

其中:

- R 是参考摘要的集合,
- N 是机器生成的摘要中的 n 元语法集合,
- $count_{match}(n,\ r)$ 是 n 元语法 n 在参考摘要 r中出现的次数，并且
- $count_{ref}(n,\ r)$ 是 n 元语法 n 在所有参考摘要中出现的总次数。

8.1.3 TER

翻译编辑率（TER）衡量机器生成的译文与参考译文之间的编辑距离[17]。编辑距离是将机器生成的译文转换为参考译文所需的更改次数。这些更改可能包括插入、删除、替换和移位。TER 使用一组预定义的编辑操作来计算编辑距离，并将机器生成的译文与参考译文之间的词序和短语结构差异纳入考量。随后，将编辑距离除以参考译文的长度，得出 TER 分数。

$$\text{TER} = \frac{S + D + I}{N} \tag{5}$$

其中:

- S 是将机器生成的摘要转换为某个参考摘要所需的词语替换次数，
- D 是为了使机器生成的摘要与参考摘要匹配，需要从机器生成的摘要中删除的词数，而
- I 是为了使机器生成的摘要与参考摘要匹配，需要插入机器生成的摘要中的词数。
- N 是参考摘要中的词语总数。

8.1.4 美国国家标准与技术研究院

美国国家标准与技术研究院（NIST）评分公式常用于评估机器翻译输出的质量[14]。公式如下：

$$NIST = \frac{1}{N} \sum_{n=1}^{N} \frac{c_n}{n} \tag{6}$$

其中：

• N 是评估中使用的 n-gram 的最大阶数；c_n 是机器翻译输出和参考译文中同时出现的 n-gram 的数量； n 是正在评估的 n-gram 的阶数（单元语法为 1，二元语法为 2，依此类推）

8.1.5 METEOR

METEOR（显式排序翻译评估指标）分数是另一种常用的机器翻译输出质量评估指标[15]。其公式如下：

$$\text{METEOR} = (1 - \alpha) \cdot \frac{match_{seg}}{num_{seg}} + \alpha \cdot penalty \tag{7}$$

其中：

- $match_{seg}$ 是机器翻译输出与参考译文之间匹配的单元语法、二元语法和三元语法的总数，并按其重要性加权
- num_{seg} 是机器翻译输出中的词数
- α 是一个用于控制精确率和召回率重要性的参数（通常设为 0.5）
- $penalty$ 是一个惩罚项，用于计算机器翻译输出与参考译文之间的块数、块移位和未对齐词数

8.1.6 BARTScore

BARTScore 是一种用于 评估生成文本质量的指标[16]。BARTScore 的公式如下 ：

$$\text{BARTScore}(x, y) = \exp \frac{1}{n} \sum_{i=1}^{n} \log p_{\text{BART}}(x_i | x_{<i}, y) - \frac{1}{m} \sum_{j=1}^{m} \log p_{\text{BART}}(y_j | y_{<j}, x) \tag{8}$$

在哪里：

- x 是输入文本
- y 是生成的输出文本
- $p_{\text{BART}}(x_i | x_{<i}, y)$ 是在给定输入序列中的前文词语和生成的输出序列的条件下，预测输入序列中 i^{th} 词语的条件概率
- $p_{\text{BART}}(y_j | y_{<j}, x)$是在给定生成的输出序列中的前序词语和输入序列的情况下，预测生成的输出序列中的 j^{th} 词语的条件概率。
- n是输入序列的长度。
- m 是生成的输出序列的长度。