How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection

Biyang Guo[1†*], **Xin Zhang**[2*], **Ziyuan Wang**[1*], **Minqi Jiang**[1*], **Jinran Nie**[3*]
Yuxuan Ding[4], **Jianwei Yue**[5], **Yupeng Wu**[6]
[1]AI Lab, School of Information Management and Engineering
Shanghai University of Finance and Economics
[2]Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen)
[3]School of Information Science, Beijing Language and Culture University
[4]School of Electronic Engineering, Xidian University
[5]School of Computing, Queen's University, [6]Wind Information Co., Ltd

Abstract

*警告：该PDF由GPT-Academic开源项目调用大语言模型+Latex翻译插件一键生成，版权归原文作者所有。翻译内容可靠性无保障，请仔细鉴别并以原文为准。项目Github地址：https://github.com/binary-husky/gpt_academic/。项目在线体验地址：https://chatpaper.org。当前大语言模型：gpt-3.5-turbo，当前语言模型温度设定：1。为了防止大语言模型的意外谬误产生扩散影响，禁止移除或修改此警告。

ChatGPT的推出引起了学术界和工业界的广泛关注[2]。ChatGPT能够对各种人类问题进行有效的回应，提供流畅而全面的答案，其安全性和实用性显著超过以前的公开聊天机器人。一方面，人们对ChatGPT如何实现这种强大能力、与人类专家相距多远感到好奇。另一方面，人们开始担心像ChatGPT这样的大型语言模型(LLM)可能对社会产生的负面影响，例如虚假新闻、抄袭以及社会安全问题。在本研究中，我们收集了数以万计来自人类专家和ChatGPT的比较回答，涵盖了开放领域、金融、医疗、法律和心理等各个领域的问题。我们将收集到的数据集称为**H**uman **C**hatGPT **C**omparison **C**orpus（**HC3**）。基于HC3数据集，我们研究了ChatGPT回答的特点、与人类专家之间的差异和差距，以及LLM的未来发展方向。我们对ChatGPT生成的内容与人类生成的内容进行了全面的人类评估和语言分析，揭示了许多有趣的结果。然后，我们在如何有效检测某段文本是由ChatGPT还是人类生成方面进行了大量实验。我们构建了三个不同的检测系统，探索了影响其效果的几个关键因素，并在不同场景下对其进行了评估。数据集、代码和模型都可以在以下链接公开获取：https://github.com/Hello-SimpleAI/chatgpt-comparison-detection。

[*]Equal Contribution.
[†]Project Lead. Corresponding to guo_biyang@163.com
[+]Each author has made unique contributions to the project.
[2]于2022年11月由OpenAI发布。https://chat.openai.com/chat

1 Introduction

自2022年11月以来，OpenAI的ChatGPT在自然语言处理（NLP）社区和其他许多领域引起了巨大关注和广泛讨论。根据OpenAI的说法，ChatGPT是通过对GPT-3.5系列进行强化学习与人类反馈（RLHF；[7, 32]）进行微调得到的，使用了与InstructGPT [25] 几乎相同的方法，只是在数据收集设置上有些许差异。 GPT-3.5中的丰富知识和基于人类反馈的精细微调使得ChatGPT在许多具有挑战性的NLP任务上表现出色，例如将自然语言翻译为代码 [5]，完成极度掩盖的文本 [15]，或根据用户定义的要素和风格生成故事 [40]，更不用说典型的NLP任务如文本分类、实体提取、翻译等。此外，精心收集的人工编写演示还使得ChatGPT能够承认自己的错误，质疑不正确的前提，甚至拒绝不适当的请求，这是OpenAI所声称的[3]。

ChatGPT的惊人强大能力引起了许多兴趣和关注：

一方面，**人们对ChatGPT与人类专家有多接近很感兴趣**。与之前的语言模型（LLM）如GPT-3 [4]相比，通常无法恰当地回应人类查询的情况不同，InstructGPT [25] 和更强大的ChatGPT在与人类交互方面取得了巨大改进。因此，ChatGPT有很大潜力成为日常助理用于一般或专业咨询目的[20, 21]。从语言学或NLP角度来看，我们还对ChatGPT和人类之间的剩余差距以及它们之间的隐含语言差异感兴趣[14, 18]。

另一方面，**人们对ChatGPT等LLM可能带来的潜在风险感到担忧**。由于ChatGPT的免费预览演示广泛传播，大量由ChatGPT生成的内容涌入各种UGC（用户生成内容）平台，威胁到平台的质量和可靠性。例如，著名的编程问答网站Stack Overflow暂时禁止ChatGPT生成的内容[4]，因为它认为"*从ChatGPT获得正确答案的平均率太低，由ChatGPT创建的答案在很大程度上对网站及正在寻找正确答案的用户造成伤害*"。许多其他应用和活动也面临类似的问题，如在线考试[33]和医疗分析[20]。我们对ChatGPT在法律、医疗和金融问题上的实证评估还揭示了可能产生有害或虚假信息的问题。

考虑到ChatGPT的不透明性和模型误用可能带来的社会风险，我们对学术界和社会做出以下贡献：

1. 为促进与LLM相关的研究，特别是人类和LLM之间的比较研究，我们收集了近四万个问题及其对应的人类专家和ChatGPT的回答，涵盖了广泛的领域（开放领域、计算机科学、金融、医学、法律和心理学），并命名为**人类与ChatGPT比较语料库**（**HC3**）数据集。HC3数据集是分析人类和ChatGPT的语言和风格特征的宝贵资源，有助于探讨LLM未来改进的方向；

2. 我们对人类/ChatGPT生成的答案进行了全面的**人工评估**和**语言分析**，发现了人类和ChatGPT呈现出许多有趣的模式。这些发现可以帮助区分某些内容是否由LLM生成，并为语言模型未来发展提供洞察；

3. 基于HC3数据集和分析结果，我们开发了几个针对不同检测场景的**ChatGPT检测模型**。这些检测器在我们的留存测试集上表现出良好的性能。我们还总结了一些对检测器效果至关重要的关键因素；

4. 我们**开源**了所有收集的比较语料库、评估结果和检测模型，以促进未来学术研究和在线平台对AI生成内容的规范管理。

[3] https://openai.com/blog/chatgpt/

[4] https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned

HC3-English

	# Questions	# Human Answers	# ChatGPT Answers	Source
All	24322	58546	26903	
reddit_eli5	17112	51336	16660	ELI5 dataset [10]
open_qa	1187	1187	3561	WikiQA dataset [39]
wiki_csai	842	842	842	Crawled Wikipedia (A.1)
medicine	1248	1248	1337	Medical Dialog dataset [6]
finance	3933	3933	4503	FiQA dataset [23]

HC3-Chinese

	# Questions	# Human Answers	# ChatGPT Answers	Source
All	12853	22259	17522	
open_qa	3293	7377	3991	WebTextQA & BaikeQA [38]
baike	4617	4617	4617	Crawled BaiduBaike (A.1)
nlpcc_dbqa	1709	1709	4253	NLPCC-DBQA dataset [8]
medicine	1074	1074	1074	Medical Dialog dataset [6]
finance	689	1572	1983	ChineseNlpCorpus (A.1)
psychology	1099	5220	1099	from Baidu AI Studio (A.1)
law	372	690	505	LegalQA dataset (A.1)

表 1: HC3数据集的元信息。英文版本（中文版本）包含了5（分别是7）个子集。

2 Human ChatGPT Comparison Corpus (HC3)

ChatGPT基于GPT-3.5系列，它在超大规模语料库上进行了预训练，包括网络爬取的文本、图书和代码等，从而使其能够回答各种问题。因此，我们很想知道一个人（尤其是专家）和ChatGPT分别如何回应相同的问题。受[1]的启发，我们还想评估ChatGPT能否保持诚实（不虚构信息或误导用户）、无害（不会生成有害或冒犯性内容），以及相对于人类专家来说，它有多么“有帮助”（为用户的问题提供具体而正确的解决方案）。

考虑到这些因素，我们决定收集一个对比语料库，其中包含人类和ChatGPT对相同问题的回答。我们相信这样的对比语料库可以成为研究人类语言和语言模型本质的宝贵而有趣的资源。

2.1 Human Answers Collection

Inviting human experts to manually write questions and answers is tedious and unaffordable for us to collect a large amount of data, therefore we construct the comparison dataset mainly from two sources:

• Publicly available question-answering datasets, where answers are given by experts in specific domains or the high-voted answers by web users;

• Wiki text. We construct question-answer pairs using the concepts and explanations from wiki sources like Wikipedia[5] and BaiduBaike[6].

[5] https://www.wikipedia.org/

[6] https://baike.baidu.com/

The split-data source mapping is shown in Table 1, and please refer to Appendix A.1 for further detailed information.

2.2 ChatGPT Answers Collection

基于收集的人类问答数据集，我们使用ChatGPT来回答这些问题。由于ChatGPT目前仅通过其预览网站提供，我们手动将问题输入到输入框中，并借助一些自动化测试工具获取答案。ChatGPT的答案可能受到聊天历史的影响，因此我们针对每个问题刷新对话。

为了使答案更符合人类回答，我们为特定数据集给ChatGPT添加了额外的指令。例如，来自`reddit-eli5`数据集分割的人类答案是在“用我五岁的理解”(Explain like I'm five)的背景下，因此我们通过在原始问题的末尾添加“用我五岁的理解”来指导ChatGPT。更多细节可以在附录中找到。

ChatGPT可以在不同的线程中给出相同问题的不同答案，这可能是由于解码过程中的随机抽样。然而，我们发现差异可能非常小，因此对于大多数问题我们只收集一个答案。

2.3 Human ChatGPT Comparison Corpus (HC3)

对于每个问题，可以有一个以上的人类/ChatGPT回答，因此我们使用以下格式组织对比数据：

```
{
    "question": "Q1",
    "human_answers": ["A1", "A2"],
    "chatgpt_answers": ["B1"]
}
```

总体上，我们收集了英文版本的$24,322$个问题，$58,546$个人类答案和$26,903$个ChatGPT答案；中文版本的收集了$12,853$个问题，$22,259$个人类答案和$17,522$个ChatGPT答案。每个数据集划分的元信息如表1所示。

3 Human Evaluation & Summarization

在本节中，我们邀请了许多志愿者测试人员，并从不同角度进行了广泛的人工评估。在人工评估之后，我们将收集到的对比语料库提供给志愿者，并要求他们手动总结出一些特征。然后，我们将志愿者的反馈与我们的观察结果进行总结。

3.1 Human Evaluation

人类评估分为**图灵测试**和**有用性测试**。图灵测试[34]是一个评估机器是否能表现出与人类无法区分的智能行为的测试。我们邀请了17名志愿者，分为两组：8名专家（经常使用ChatGPT的用户）和9名业余者（从未听说过ChatGPT）。这是因为熟悉ChatGPT的人可能已经记住了ChatGPT展示的某些模式，帮助他们轻松区分角色。

我们设计了四种类型的评估，使用不同的查询格式或测试组。我们在接下来的部分介绍具体的评估设计和结果：

A. **专家图灵测试，成对文本（`pair-expert`）**
`pair-expert`测试是在**专家**组中进行的。每个测试员需要进行一系列的测试，每个测试包含

一个问题和一对答案（一个来自人类，另一个来自ChatGPT）。测试员需要确定哪个答案是由ChatGPT生成的。

B. **专家图灵测试，单一文本（single-expert）**
single-expert测试也在**专家**组中进行。每个测试员需要进行一系列的测试，每个测试包含一个问题和一个随机由人类或ChatGPT给出的**单一**答案。测试员需要确定答案是否由ChatGPT生成。

C. **业余图灵测试，单一文本（single-amateur）**
single-amateur测试在**业余**组中进行。每个测试员需要进行一系列的测试，每个测试包含一个问题和一个随机由人类或ChatGPT给出的**单一**答案。测试员需要确定答案是否由ChatGPT生成。

D. **有用性测试（helpfulness）**
我们还对ChatGPT生成的答案与人类答案在某个问题上的有用程度感到好奇。请注意，有用性是一个非常主观的度量标准，可以受到许多因素的影响，包括情感、测试员个性、个人偏好等等。因此，仅仅提供更准确的信息或更详细的分析并不总是能导致更有帮助的答案。

helpfulness测试是在**专家**组中进行的。每个测试员需要进行一系列的测试，每个测试包含一个问题和一对答案（一个来自人类，另一个来自ChatGPT）。每个测试员被要求假装该问题是由他/她自己提出的，并需要确定哪个答案对他/她更有帮助。

设置。 我们从每个数据集划分（例如reddit_eli5、wikipedia、medical等）中随机选取约30个<问题,人类答案,ChatGPT答案>三元组作为人类评估的样本。我们为每个划分分配2-5名测试员，并报告他们的平均结果。对于所有图灵测试，我们报告测试员能正确检测到*ChatGPT*生成答案的比例。对于有用性测试，我们报告测试员认为*ChatGPT*生成答案更有帮助的比例。

结果。 根据表2中显示的结果，我们可以得出几个结论。比较pair-expert和single-expert的结果，我们发现**当提供一个比较对时,更容易区分ChatGPT生成的内容,而不仅仅提供一个单独的答案。** 比较single-expert和single-amateur的结果，我们发现**专家的准确性要远高于业余人士**。关于helpfulness测试，它反映了志愿者认为ChatGPT的答案在多大程度上对他们有帮助。令人惊讶的是，结果表明，在超过一半的问题中，ChatGPT的答案**通常被认为比人类的答案更有帮助**，尤其是在金融和心理学领域。通过检查这些领域的具体答案，我们发现ChatGPT通常能够提供更具体和具体的建议。然而，在英文和中文的医学领域中，ChatGPT的帮助性表现较差。在我们收集的数据集中，ChatGPT常常对医学咨询给出冗长的答案，而人类专家可能直接给出简明扼要的回答或建议，这在一定程度上可以解释为什么志愿者认为人类的答案在医学领域更有帮助。

3.2 Human Summarization

在上述评估之后，我们向志愿者开放了我们收集到的HC3数据集，他们可以自由地浏览人类和ChatGPT之间的比较答案。所有的数据集分区都分配给不同的志愿者，每个志愿者被要求浏览至少100组比较数据。之后，我们请他们总结人类答案和ChatGPT答案的特点。最终，我们收到了200多份反馈，我们将这些发现总结如下：

ChatGPT独特的模式

Human Evaluation (En)

	Pair-expert	Single-expert	Single-amateur	Helpfulness
All	0.90	0.81	0.48	0.57
reddit_eli5	0.97	0.94	0.57	0.59
open_qa	0.98	0.78	0.34	0.72
wiki_csai	0.97	0.61	0.39	0.71
medical	0.97	0.97	0.50	0.23
finance	0.79	0.73	0.58	0.60

Human Evaluation (Zh)

	Pair-expert	Single-expert	Single-amateur	Helpfulness
All	0.93	0.86	0.54	0.54
open_qa	1.00	0.92	0.47	0.50
baike	0.76	0.64	0.60	0.60
nlpcc_dbqa	1.00	0.90	0.13	0.63
medicine	0.93	0.93	0.57	0.30
finance	0.86	0.84	0.84	0.75
psychology	1.00	1.00	0.60	0.67
law	1.00	0.77	0.56	0.56

表 2: 人类对ChatGPT生成的英文和中文答案进行的评估。

(a) **ChatGPT以有组织的方式编写,逻辑清晰.** 不失一般性，ChatGPT喜欢在问题中定义核心概念。然后，它会逐步提供详细答案，并在最后给出总结，按照推理和总结结构进行；

(b) **ChatGPT倾向于提供长而详尽的答案。** 这是人类反馈强化学习（RLHF）的直接产物，也在一定程度上与模式（a）相关，除非您提供提示，如“用一句话向我解释”；

(c) **ChatGPT显示出较少的偏见和有害信息**. ChatGPT在敏感话题上保持中立，几乎不表现出对政治领域或具有歧视性的有害对话的态度；

(d) **ChatGPT拒绝根据其知识回答问题**. 例如，ChatGPT无法回答需要2021年9月之后的信息的查询。有时，ChatGPT还会拒绝回答其认为自己不知道的问题。这也是RLHF隐式且自动地确定哪些信息在模型知识范围内，哪些不在的能力；

(e) **ChatGPT可能捏造事实**. 在回答需求特定领域的专业知识的问题时，为了给出回答，ChatGPT可能会捏造事实，尽管[25]提到InstructGPT模型在真实性方面已经显示出改进。例如，在法律问题中，ChatGPT可能会虚构一些不存在的法律条款来回答问题。这种现象提醒我们在使用ChatGPT进行专业咨询时要格外小心。此外，当用户提出了一个没有现成答案的问题时，ChatGPT也可能会捏造事实以提供回应。

上述许多结论如(b)、(c)和(d)也在Fu等人的研究[12]中进行了讨论。

人类与ChatGPT之间的主要差异

(a) **ChatGPT的回答通常严格聚焦于提出的问题,而人类的回答则多样化且容易偏离其他话题。**就内容的丰富程度而言，人类在不同方面更加多元化，而ChatGPT则更偏好关注问题本身。人类可以根据自己的常识和知识回答问题背后的含义，但ChatGPT却依赖于问题字面上的意思；

(b) **ChatGPT提供客观的答案,而人类更喜欢主观表达。**一般来说，与人类相比，ChatGPT生成的文本更加安全、平衡、中立和具有信息性。因此，ChatGPT在解释术语和概念方面表现出色。另一方面，人类的回答更加具体，并包含有关法律、书籍和论文等来源的详细引用，尤其是在提供医学、法律和技术问题的建议时；

(c) **ChatGPT的回答通常较为正式,而人类的表达更加口语化。**人类倾向于简洁，并使用口语缩写和俚语，如"LOL"、"TL;DR"、"GOAT"等。人类还喜欢运用幽默、讽刺、隐喻和例子，而ChatGPT从不使用讽刺。此外，人类的交流经常包含"互联网迷因"，作为一种特定而生动的表达方式；

(d) **ChatGPT在回答中表达的情感较少,而人类则选择使用多种标点符号和语法特征来传达自己的情感。**人类使用多个感叹号("!"")、问号("?"")、省略号("...")来表达强烈的情感，并使用各种括号("(", ")", "[", "]")来解释事情。相比之下，ChatGPT喜欢使用连词和副词来传递逻辑思路，例如"一般来说"、"另一方面"、"首先，...，其次，...，最后"等等。

总的来说，这些总结的特点表明ChatGPT在各个领域的问答任务方面有显著改进。与人类相比，我们可以将ChatGPT想象成保守的专家"团队"。作为一个"团队"，它可能缺乏个性，但可以对问题有更全面和中立的视角。

4 Linguistic Analysis

在本节中，我们分析了人类和ChatGPT的回答的语言特征，并试图找到一些统计证据，验证在第 3 节中总结的特点。

4.1 Vocabulary Features

在本部分中，我们分析了我们收集的语料库的词汇特点。我们对人类与ChatGPT在回答相同问题时选择词汇的差异感兴趣。

由于人类/ChatGPT回答的数量不平衡，我们在统计过程中随机抽取了一个人类回答和一个ChatGPT回答。我们计算了以下特征：**平均长度（L）**，即每个问题中平均单词数；**词汇大小（V）**，即所有回答中使用的唯一词汇数；我们还提出了另一个特征称为**密度（D）**，它由$D = 100 \times V/(L \times N)$计算，其中$N$是回答数量。密度衡量了文本中使用的不同词汇的「拥挤程度」。例如，如果我们写了一些总共1000个单词的文章，但只使用了100个不同的词，那么密度是$100 \times 100/1000 = 10$。密度越高，同样长度的文本中使用的不同词汇越多。

在表3中，我们报告了英文和中文语料库的词汇特征。从*平均长度*和*词汇大小*两个特征来看，我们可以看到：相比于ChatGPT，人类回答相对较短，但使用了更大的词汇量。这一现象在中文的`open_qa`分组和两种语言的`medical`分组中尤为明显，其中ChatGPT的平均长度几乎是人类的两倍长，但词汇大小显著较小。

这一现象也反映在*密度*因子上。人类的词汇密度在**每个分组**中均大于ChatGPT的，进一步揭示了**人类在表达时使用更多样化的词汇**。

	English	avg. len.	vocab size	density	Chinese	avg. len.	vocab size	density
human	**All**	142.50	**79157**	**2.33**	**All**	102.27	**75483**	**5.75**
ChatGPT		**198.14**	66622	1.41		**115.3**	45168	3.05
human	*reddit_eli5*	134.21	**55098**	**2.46**	*nlpcc_dbqa*	24.44	10621	**25.43**
ChatGPT		**194.84**	44926	1.38		**78.21**	**11971**	8.96
human	*open_qa*	35.09	9606	**23.06**	*open_qa*	93.68	**40328**	**13.13**
ChatGPT		**131.68**	**16251**	10.40		**150.66**	26451	5.35
human	*wiki_csai*	**229.34**	**15859**	**8.21**	*baike*	**112.25**	**28966**	**5.59**
ChatGPT		208.33	9741	5.55		77.19	14041	3.94
human	*medicine*	92.98	**11847**	**10.42**	*medicine*	92.34	**9855**	**9.94**
ChatGPT		**209.61**	7694	3.00		**165.41**	7211	4.06
human	*finance*	202.07	**25500**	**3.21**	*finance*	80.76	2759	**5.05**
ChatGPT		**226.01**	21411	2.41		**120.84**	**4043**	4.94
human	-	-	-	-	*psychology*	**254.82**	**16160**	**5.77**
ChatGPT	-	-	-	-		164.53	5897	3.26
human	-	-	-	-	*law*	28.77	2093	**19.55**
ChatGPT	-	-	-	-		**143.76**	**3857**	7.21

表 3: 我们的语料库上的平均回答长度、词汇量和密度比较。

4.2 Part-of-Speech & Dependency Analysis

在这一部分中，我们比较不同词性标签（POS）的出现次数以及依存关系的特征。

4.2.1 Part-of-Speech

图1展示了人类和ChatGPT在词性使用上的比较。在HC3-英文测试集中，ChatGPT使用了更多的"NOUN"、"VERB"、"DET"、"ADJ"、"AUX"、"CCONJ"和"PART"单词，而使用了较少的"ADV"和"PUNCT"单词。

大量使用名词（"NOUN"）通常表示文本更具有辩证性，表达了信息丰富和客观性[24]。因此，介词（"ADP"）和形容词（"ADJ"）单词也更常出现[11]。连词（"CCONJ"）与名词、动词和介词单词的频繁搭配表明了文章结构以及因果关系、进展或对比关系的清晰性。这些也是学术论文或官方文件的典型特征[29]。我们认为RLHF训练过程对ChatGPT的写作风格有很大影响，这在一定程度上解释了词性标签分布的差异。

4.2.2 Dependency Parsing

依存句法分析是一种通过识别单词之间的依赖关系来分析句子的语法结构的技术。我们对语料库中的答案进行分析，并比较不同依存关系及其相应的依存距离的比例。图2显示了人类和ChatGPT在HC3-English中的比较情况。由于篇幅限制，中文版本放在附录A.2中。

依存关系的比较呈现出与词性标签类似的特征，其中ChatGPT使用了更多的决定、并列和助词关系。在依存距离方面，ChatGPT在标点和依赖关系上的距离要长得多，这可能是因为CharGPT倾向于使用更长的句子。然而，ChatGPT的"conj"关系明显较短。根据词性标签的分析，ChatGPT通常使用更多的连词来使内容更具逻辑性，这可能解释了为什么ChatGPT的"conj"关系相对于人类而言较短。

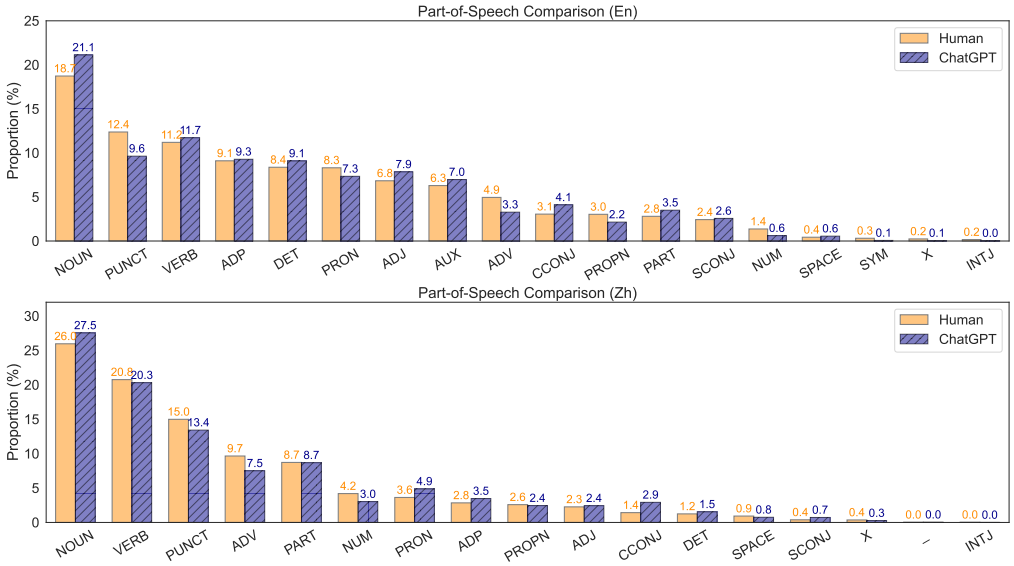


图 1: ChatGPT和人类回答之间的词性分布比较。结果按人类回答的词性比例排序。上图显示了HC3-English数据集的结果，下图显示了HC3-Chinese数据集的结果。

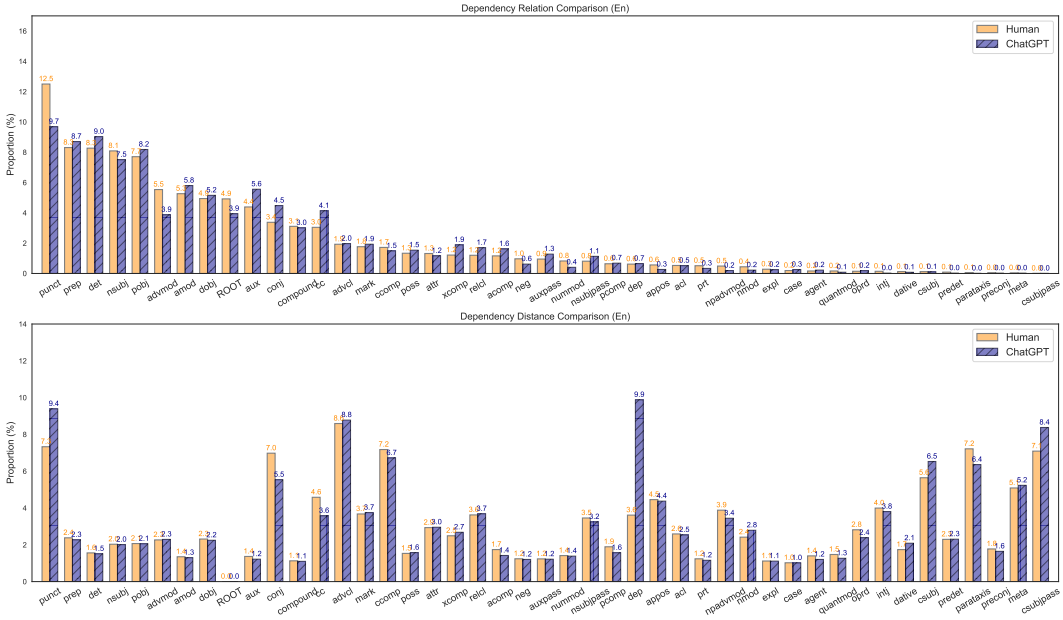
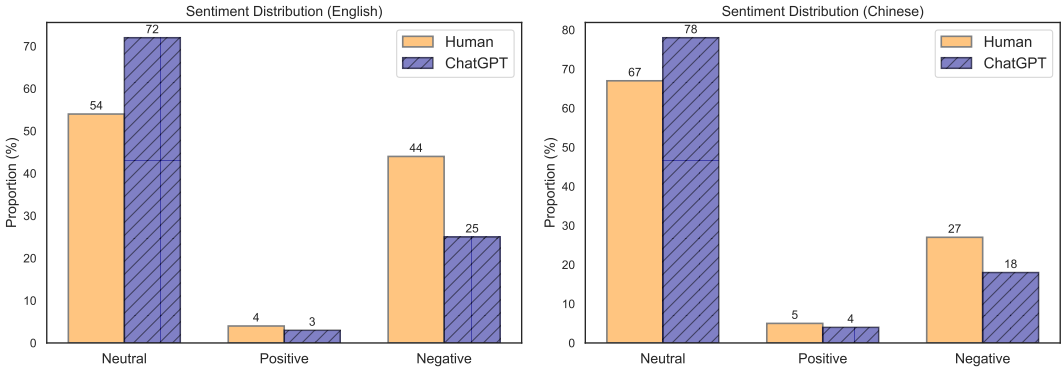


图 2: HC3-English中人类回答和ChatGPT回答之间的前30个依存关系的比较（上）以及对应的依存距离（下）。结果按照人类回答的关系比例排序。

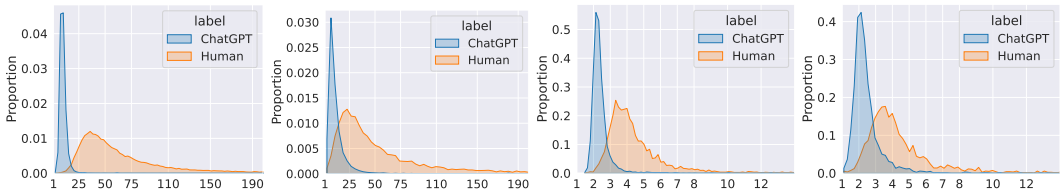
4.3 Sentiment Analysis

人类是情绪化的存在，我们的情绪在某种程度上自然地反映在我们的语言中。ChatGPT是通过大规模人类生成的文本进行学习的，但它经过了人类指导的进一步微调。因此，我们很好奇与人类相比，"情感化"的ChatGPT是如何的。



(a) Sentiment distribution of HC3-English (b) Sentiment distribution of the HC3-Chinese

图 3: 我们语料库中三种情感（中性、积极、消极）的比例。



(a) English text ppl (b) English sent ppl (c) Chinese text ppl (d) Chinese sent ppl

图 4: 在英语和中文数据以及文本和句子层面上对PPL分布进行分析。

我们使用了一个多语言情感分类模型[7]，对英文和中文对照数据进行情感分析。需要注意的是，基于深度学习的模型往往会受到一些暗示性词语的影响（比如"but"和"sorry"等词可以很容易地让分类器预测"negative"标签），从而导致预测结果具有偏见[16]。因此，分类器给出的情感仅仅是对文本背后真实情感的一个参考。

图3展示了人类和ChatGPT情感分布的比较结果。从结果中可以得出几点发现：首先，我们发现无论是人类还是ChatGPT，中性情绪的比例最大，这与我们的预期相符。然而，**ChatGPT一般表达的中性情绪比人类更多**。其次，负向情绪的比例显著高于正向情绪。值得注意的是，**人类表达的负向情绪明显比ChatGPT更多**。人类的正向情绪比例稍微高于ChatGPT的比例。总体而言，ChatGPT比人类更少情绪化，尽管它并非完全没有情感。

4.4 Language Model Perplexity

困惑度（Perplexity，PPL）通常被用作评估语言模型（LM）性能的指标。它被定义为LM下文本的负对数似然的指数值。较低的PPL表示语言模型对其预测更有信心，因此被认为是一个较好的模型。对LM的训练是在大规模的文本语料库上进行的，可以认为它已学习了一些常见的语言模式和文本结构。因此，我们可以使用PPL来衡量文本符合常见特征的程度。

我们使用开源的GPT-2 small模型[8]（用于中文的Wenzhong-GPT2-110M[9]）来计算所收集文本的PPL（包括文本级别和句子级别[10]的PPL）。图4显示了人类撰写的文本和ChatGPT生成的文本的PPL分布情况。

[7]https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

[8]`https://huggingface.co/gpt2`

[9]`https://huggingface.co/IDEA-CCNL/Wenzhong-GPT2-110M`

[10]对于英文文本,我们使用NLTK[3]进行句子分割（中文使用HarvestText）。

显然可见，无论是在文本级别还是句子级别，ChatGPT生成的内容相较于人类撰写的文本具有较低的PPL值。 ChatGPT捕捉到了它训练文本中的常见模式和结构，并且在复现它们方面非常出色。因此，ChatGPT生成的文本具有相对集中的低PPL值。

人类具有根据写作上下文、受众和目的而以各种方式表达自己的能力。这可能包括使用创意或想象力元素，如隐喻、比喻和独特的词语选择，这可能使GPT2更难进行预测。因此，人类撰写的文本具有更多的高PPL值，并呈现出长尾分布，如图4所示。

5 ChatGPT Content Detection

人工智能生成内容（AIGC）在互联网上变得越来越普遍，很难将其与人类生成的内容区分开来，正如我们在人工评估中所展示的（见第3.1节）。因此，需要AIGC检测器来帮助识别和标记由机器生成的内容，以减少因不当或恶意使用AI模型而对社会造成的潜在风险，并提高在线共享信息的透明度和问责性。

在本节中，我们进行了几个实证实验来研究ChatGPT内容检测系统。检测人工智能生成的内容是一个广泛研究的课题[19, 27]。我们基于[30, 13, 27]建立了三种不同类型的检测系统，包括基于机器学习和深度学习的方法，并在不同的细粒度和数据源上进行了评估。详细的结果和讨论将会提供。

5.1 Methods

Detection of machine-generated text has been gaining popularity as text generation models have advanced in recent years[19, 27]. Here, we implement three representative methods from classic machine learning and deep learning, i.e, a logistic regression model trained on the GLTR Test-2[13] features, a deep classifier for single-text detection and a deep classifier for QA detection. The deep classifiers for both single-text and QA are based on RoBERTa [22], a strong pre-trained Transformer [35] model. In fact, algorithms for OOD detection or anomaly detection [17] can also be applied to develop ChatGPT content detectors, which we leave for future work.

GLTR. [13] studied three tests to compute features of an input text. Their major assumption is that to generate fluent and natural-looking text, most decoding strategies sample high probabilities tokens from the head of the distribution. We select the most powerful Test-2 feature, which is the number of tokens in the Top-10, Top-100, Top-1000, and 1000+ ranks from the LM predicted probability distributions. And then a logistic regression model is trained to finish the classification.

RoBERTa-*sinlge*. A deep classifier based on the pre-trained LM is always a good choice for this kind of text classification problem. It is also investigated in many studies and demo systems [30, 9, 27]. Here we fine-tune the RoBERTa [22] model.

RoBERTa-*QA*. While most content detectors are developed to classify whether a single piece of text is AI-generated, we claim that a detector that supports inputting both a question and an answer can be quite useful, especially for question-answering scenarios. Therefore, we decide to also build a QA version detector. The RoBERTa model supports a text pair input format, where a separating token is used to join a question and its corresponding answer.

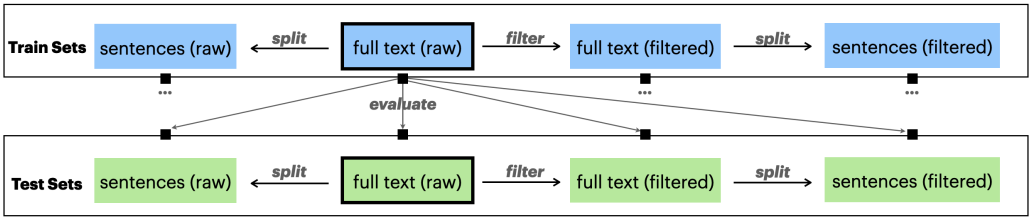


图 5: 实验设计用于检测器的训练和测试。通过过滤或分割，可以生成不同的数据集版本。

5.2 Implementation Details

对于GLTR使用的LM，我们在英文方面使用gpt2-small [28]，而在中文方面使用由[36]发布的Wenzhong-GPT2-110M，和第4.4节中的相同。对于基于RoBERTa的深度分类器，我们分别使用`roberta-base`[11]和`hfl/chinese-roberta-wwm-ext`[12]检查点进行英文和中文的训练。以上所有模型均来自huggingface的`transformers` [37]。

我们使用sklearn [26]在训练集的GLTR Test-2特征上训练逻辑回归模型，并按照[27]的代码搜索超参数。基于RoBERTa的检测器使用`transformers`的设施进行训练。具体来说，我们使用AdamW优化器，批大小设置为32，学习率为$5e-5$。我们对英文进行了1个时期的微调，对中文进行了2个时期的微调。

5.3 Experiment Design

HC3数据集包括问题及其对应的人类/ChatGPT答案。我们提取了所有的<问题，答案>对，并将标签0分配给人类答案的对，将标签1分配给ChatGPT答案的对。

简单地使用人类和ChatGPT的原始答案来训练一个二分类器是最直接的方法。然而，这样做可能会出现一些问题：

- 首先，根据第3节中的观察结果，无论是人类回答还是ChatGPT回答都可能含有一些明显的指示性词语，这些词语可能会影响模型的效果；
- 其次，用户可能想要识别出ChatGPT生成的单个句子，而不仅仅是整个文本。对于仅在完整文本上进行训练的分类器来说，这可能是相当困难的；
- 第三，考虑到答案相应问题可能有助于探测器进行更准确的判断，与仅考虑答案本身相比。这可以广泛应用于许多问答平台（如Quora，Stack Overflow和知乎），以查找在某个特定问题下哪个答案是由人工智能生成的。

因此，我们设计了不同组的实验来研究这些关键问题：
• 指示词会如何影响检测器？
• ChatGPT 检测器对于句级内容的检测更具挑战性吗？训练一个句级分类器更困难吗？
• 相应的问题是否能够帮助检测器更准确地检测答案的来源？

图 5 显示了我们如何生成不同类型的训练和测试集。具体而言，我们使用收集到的原始语料库构建了第一组训练-测试集（图中的“全文（原始）”），我们称之为**原始全文**版本。然后我们从文本中过滤掉指示词，得到**过滤全文**版本。通过将全文分割成句子，我们得到了**原始句子**版本和**过滤句子**版本。我们还将全文和句子合并成混合版本，即**原始混合**和**过滤混合**版本。总体而言，我们有六个不同版本的训练和测试集。

[11] `https://huggingface.co/roberta-base`
[12] `https://huggingface.co/hfl/chinese-roberta-wwm-ext`

在版本 B 的测试集上评估模型在版本 A 的训练集上训练的表现，可以被看作是一种超出分布（OOD）的泛化评估，这更具挑战性，因为它要求模型对样本风格的变化具有鲁棒性。

5.4 Results

按照上述实验设计，我们对各种衍生语料进行了全面的实证研究。表4显示了测试的F1分数。

		English							**Chinese**						
Test →		*raw*			*filtered*			Avg.	*raw*			*filtered*			Avg.
		full	sent	mix	full	sent	mix		full	sent	mix	full	sent	mix	
Train ↓		**RoBERTa**													
raw	full	99.82	81.89	84.67	99.72	81.00	84.07	88.53	98.79	83.64	86.32	98.57	82.77	85.85	89.32
	sent	99.40	98.43	98.56	99.24	98.47	98.59	**98.78**	97.76	95.75	96.11	97.68	95.31	95.77	**96.40**
	mix	99.44	98.31	98.47	99.32	98.37	98.51	98.74	97.70	95.68	96.04	97.65	95.27	95.73	96.35
filtered	full	99.82	87.17	89.05	99.79	86.60	88.67	91.85	98.25	91.04	92.30	98.14	91.15	92.48	93.89
	sent	96.97	97.22	97.19	99.09	98.43	98.53	97.91	96.60	92.81	93.47	97.94	95.86	96.26	95.49
	mix	96.28	96.43	96.41	99.45	98.37	98.53	97.58	97.43	94.09	94.68	97.66	95.61	96.01	95.91
Train ↓		**GLTR Test-2**													
raw	full	98.26	71.58	76.15	98.22	70.19	75.23	81.61	89.61	44.02	53.72	85.89	43.58	53.62	61.74
	sent	86.26	88.18	87.96	87.72	88.23	88.19	87.76	84.49	71.79	74.01	84.06	70.29	72.90	76.26
	mix	95.97	86.45	87.81	96.13	86.24	87.73	90.06	86.45	70.85	73.59	84.94	69.14	72.14	76.19
filtered	full	98.31	70.91	75.65	98.30	69.48	74.72	81.23	89.46	58.69	64.52	86.51	55.45	62.18	69.47
	sent	84.00	88.25	87.71	85.68	88.35	87.99	87.00	84.56	71.85	74.07	84.22	70.59	73.18	76.41
	mix	95.36	86.73	87.97	95.60	86.56	87.92	90.02	86.30	71.00	73.70	84.98	69.45	72.40	76.31

表 4: 在每个测试集上，不同模型的F1得分（%）以及每种语言的平均得分被报告。

5.4.1 Which detector(s) is more useful? ML-based or DL-based? and Why?

根据表4，我们可以得出以下结论：

首先，**基于RoBERTa的检测器的鲁棒性优于GLTR**。当通过比较*raw→raw*和*filtered→filtered*中的主对角线元素来划分句子时，RoBERTa的F1分数略微下降（英文数据集下降1.5-2%，中文数据集下降2-3%）。相反，GLTR在英文数据集中减少了超过10%，在中文数据集中减少了超过15%。总之，基于RoBERTa的检测器具有更强的抗干扰性能。

其次，**RoBERTa-based-detector受指示词的影响不大**。 RoBERTa的F1分数仅在英文*full*数据集中微降0.03%，在中文*full*数据集中降低0.65%，就是通过对比*raw→raw*与*filtered→filtered*中相关主对角线元素的差异。相反，基于GLTR的评估在中文数据集上下降了3.1%，尽管在英文数据集上稍微上升，这表明GLTR对指示词敏感，容易受到ChatGPT模式的影响。

最后，**RoBERTa-based-detector在处理超出分布情况时表现出效果**。与原始模型相比，它在GLTR的超出分布测试数据集上表现出显著下降，英文数据集下降了最多28.8%（*filtered-full→filtered-full − filtered-full→filtered-sent*），中文数据集下降了45.5%（*raw-full→raw-full − raw-full→raw-sent*）。然而，RoBERTa保持了一致的性能，其F1分数变化不超过19%。

5.4.2 How will the indicating words influence the detector?

我们首先收集了一系列人类和ChatGPT的指示性词语。例如，ChatGPT的指示性词语（或短语）包括“AI助手”，“很抱歉听到这个”，“有几个步骤…”等等，而人类的指示性词语可

能包括“嗯”，“不对”，“我的观点是”等等。在经过筛选的版本中，我们删除了回答中包含人类和ChatGPT指示性词语的所有句子。

根据表4，**删除指示性词语有助于基于全文训练的模型在不同内容粒度上表现更好**。例如，RoBERTa-*filter-full*在句子级和混合级评估方面表现明显优于RoBERTa-*raw-full*，平均F1得分提高了3%以上。然而，**筛选可能会稍微影响基于句子训练的模型的性能**。这可能是因为指示性词语在句子级文本中起到比全文更大的作用。删除指示性词语可能导致某些句子在字面上无法区分。

5.4.3 Which granularity is more difficult to detect? Full-text or sentence?

通过表 5中的大量实验结果，我们得出结论：**在单个句子中检测ChatGPT生成的文本比在整个文本中更困难**。这个结论可以通过以下两点来证明：首先，我们的结果显示，无论是英文还是中文的基于句子的检测器（即*raw-sent*和*filtered-sent*版本）在检测ChatGPT生成的段落或句子的测试任务上都取得了满意的结果，而相反的情况并不成立——当检测ChatGPT生成的句子时，*raw-full*和*filtered-full*相对较差。换句话说，在“困难样本”（即句子语料库）上训练的检测器更容易解决简单任务（即检测完整语料库），而“简单样本”（即完整语料库）可能对解决更困难的任务（即句子语料库）不太有用。

其次，我们观察到，虽然*raw-mix*和*filtered-mix*版本都提供了完整和句子两种语料库，但对它们来说，检测ChatGPT生成的单个句子仍然更困难。对于中文语料库而言，这一点甚至更为明显，其中*raw-mix*在测试原始句子答案时的F1分数为94.09%，而在测试原始完整答案时为97.43%。类似的结果也可以观察到对于过滤后的语料库，在检测过滤后的句子答案时，*filtered-mix*的F1分数为95.61%，而在测试过滤后的完整答案时为97.66%。一个可能的解释是，当提供文本段落时，ChatGPT的表达模式更明显（因此更容易被检测），而检测生成的单个句子则更困难。

	English							**Chinese**						
Test →	*raw*			*filtered*			Avg.	*raw*			*filtered*			Avg.
Train ↓	full	sent	mix	full	sent	mix		full	sent	mix	full	sent	mix	
full-*raw*	99.82	81.89	84.67	99.72	81.00	84.07	88.53	98.79	83.64	86.32	98.57	82.77	85.85	89.32
sent-*raw*	99.40	98.43	98.56	99.24	98.47	98.59	**98.78**	97.76	95.75	96.11	97.68	95.31	95.77	**96.40**
mix-*raw*	99.44	98.31	98.47	99.32	98.37	98.51	98.74	97.70	95.68	96.04	97.65	95.27	95.73	96.35
full-*filtered*	99.82	87.17	89.05	99.79	86.60	88.67	91.85	98.25	91.04	92.30	98.14	91.15	92.48	93.89
sent-*filtered*	96.97	97.22	97.19	99.09	98.43	98.53	**97.91**	96.60	92.81	93.47	97.94	95.86	96.26	95.49
mix-*filtered*	96.28	96.43	96.41	99.45	98.37	98.53	97.58	97.43	94.09	94.68	97.66	95.61	96.01	**95.91**

表 5: RoBERTa模型在全文模式、句子模式和混合模式下的F1得分（%）。

5.4.4 Which corpus is more helpful for model training? Full-text, sentence, or mix of the two?

我们发现，当模型训练中有细粒度语料数据时，无论是英文还是中文的基于RoBERTa的检测器都更加稳健。基于句子的检测器在F1分数方面优于基于完整文本的检测器，而当句子语料注入模型训练时，后者可以显著改善，我们观察到混合式检测器也能达到令人满意的结果。

对于英文语料，*raw-full* 仅在测试句子回答的F1分数中达到了81.89%，而 *raw-sent* 则显着优于后者，其F1分数为98.43%，如表 5 所示。此外，通过将句子回答注入检测器，相对较差的检测性能也能得到改善，我们发现 *raw-mix* 也能显著提高（F1分数为98.31%），相较于仅

使用完整回答进行训练的检测器。对于筛选版本，类似的结论也适用，无论是对英文还是中文语料，*filtered-sent* 和 *filtered-mix* 在F1分数方面都明显优于 *filtered-full* 版本。

我们指出，上述结论同样适用于其他类型的检测器，如GLTR Test-2基于特征的检测器，如表 4所示。对于GLTR Test-2，*raw-full* 和 *filtered-full* 的平均F1分数分别为61.74％和69.47％，而 *raw-sent* 和 *filtered-sent* 的F1分数分别为76.26％和76.41％，在混合语料训练的检测器性能接近基于句子的版本。

考虑到前一段关于完整回答和句子回答之间检测困难性的结论，我们指出，细粒度语料对于区分ChatGPT生成的文本很有帮助，因为它在模型训练中额外提供了指导和提示，以便检测ChatGPT隐藏在单个句子中的微妙模式。

5.4.5 Will a QA-style detector be more effective than a single-text detector?

表6展示了*raw-full*和*filtered-full*模型在所有测试数据集上的结果。

在英文数据集上，问答（QA）模型的F1分数表现优于单一模型，除了两个*full*测试数据集，在这两个数据集上，它的平均F1分数为97.48%，超过单一模型的5.63%。在中文数据集中存在一些差异，在*raw-full*训练数据集中，单一模型的表现优于QA模型。然而，QA模型仍然取得了94.22%的最佳评价。

总而言之，QA模型通常比单一模型更有效，并且适用于经过过滤的情景。而且，QA训练使模型对句子输入更加稳健。

	English							**Chinese**						
Test →	*raw*			*filtered*			Avg.	*raw*			*filtered*			Avg.
	full	sent	mix	full	sent	mix		full	sent	mix	full	sent	mix	
	Train → *raw* - full													
Single	99.82	81.89	84.67	99.72	81.00	84.07	88.53	98.79	83.64	86.32	98.57	82.77	85.85	89.32
QA	99.84	92.68	93.70	99.75	92.34	93.46	95.30	98.99	80.56	83.85	98.73	80.24	83.89	87.71
	Train → *filtered* - full													
Single	99.82	87.17	89.05	99.79	86.60	88.67	91.85	98.25	91.04	92.30	98.14	91.15	92.48	93.89
QA	99.70	96.14	96.64	99.70	96.07	96.61	**97.48**	97.29	92.10	93.01	97.18	92.40	93.31	**94.22**

表 6: 使用 QA Single 设置训练的 RoBERTa 模型的 F1 分数（%）。

5.4.6 Which data sources are more difficult for the ChatGPT detectors? and What are the conditions that make it easier to detect ChatGPT?

如表7所示，我们的HC3数据集中基于*filtered-full*模型的评估结果按照不同的数据源分为几类。

在英语数据集上，与没有例外的ChatGPT相比，人类回答的F1得分略高，无论是在全文测试数据集上使用RoBERTa还是GLTR。然而，ChatGPT在转移测试数据集上的F1得分在性能上存在高度不一致，尤其是在`open-qa`数据集上，呈现出不同的表现。**在数据资源方面,`reddit-eli5`和`finance-en`的数值较高，而`wiki-csai`对于检测器来说是一个挑战。**

在中文数据集上，人类和ChatGPT的F1得分相当，没有显著差异。这表明检测ChatGPT的困难程度取决于数据源。**可以观察到`open-qa`和`baike`的性能较好，而`nlpcc-dbqa`的性能较低。**

总之，与英语数据集相比，在中文数据集上的评估显示出更高的转移测试数据集稳定性。此外，不论数据集是英语还是中文，ChatGPT的F1得分都低于人类回答的F1得分。这表明**ChatGPT的检测器更依赖于In-Distribution模型**。

Model	Test	F1-hu	F1-ch	F1-hu	F1-ch	F1-hu	F1-ch	F1-hu	F1-ch	F1-hu	F1-ch
English											
		finance		medicine		open_qa		reddit_eli5		wiki_csai	
RoBERTa	full	99.34	99.28	99.69	99.62	99.53	98.60	100.00	100.00	96.59	96.37
	sent	78.84	85.84	84.06	80.45	70.74	26.78	77.27	93.31	68.91	84.12
GLTR	full	97.50	97.37	98.28	97.96	92.68	82.20	98.22	99.40	95.76	95.72
	sent	46.60	75.26	45.41	61.72	42.01	17.81	38.12	87.05	39.24	76.94
Chinese											
		finance		law		open_qa		nlpcc_dbqa		baike	
RoBERTa	full	98.87	97.99	97.78	98.50	98.75	99.33	97.42	95.42	94.61	93.99
	sent	95.00	80.46	93.77	86.23	91.17	93.77	90.10	63.29	86.08	88.88
GLTR	full	86.67	80.42	82.41	88.89	85.75	93.15	77.25	69.78	81.62	77.91
	sent	36.91	32.80	33.99	46.22	36.45	75.21	46.39	27.50	48.10	71.72

表 7: 人类 (F1-hu) 和 ChatGPT (F1-ch) 检测 F1 分数 (%) 关于不同的数据来源，模型被训练在过滤后的全文上，测试在过滤后的全文和句子上。在 HC3-Chinese 上，我们省略了 *medicine* 和 *psychology* 领域的结果，它们分别与 *finance* 和 *open_qa* 类似。

6 Conclusion

在这项工作中，我们提出了HC3（Human ChatGPT Comparison Corpus）数据集，它包含近40,000个问题及其对应的人类/ChatGPT回答。基于HC3数据集，我们进行了包括人类评估、语言分析和内容检测实验在内的广泛研究。人类评估和语言分析为我们提供了关于人类和ChatGPT之间的隐含差异的深入洞察，这激发了我们对LLMs未来方向的思考。ChatGPT内容检测实验展示了一些重要结论，可以为AIGC检测工具的研究和开发提供有益指导。我们公开提供所有数据、代码和模型，以促进相关研究和应用。网址为`https://github.com/Hello-SimpleAI/chatgpt-comparison-detection`。

7 Limitations

尽管我们对ChatGPT进行了全面的分析，但当前论文还存在几个限制，我们将在未来的工作中对其进行改进：

1. 尽管我们在数据收集方面做出了努力，但由于时间和资源有限，收集到的数据量和范围仍然不够，并且来自不同源头的数据存在不平衡的情况。为了进行更准确的语言分析和内容检测，我们需要更多具有不同风格、来源和语言的数据；
2. 目前，所有收集到的ChatGPT的回答都是在没有特殊提示的情况下生成的。因此，本文中的分析和结论是基于ChatGPT最通用的风格/状态。例如，使用特殊提示，如“假设你是莎士比亚...”可以生成绕过我们的检测器的内容，或使本文中的结论站不住脚；

3. ChatGPT（可能）主要是在英语语料库上进行训练，对中文的训练较少。因此，从HC3-Chinese数据集得出的结论可能并不总是准确的。

Acknowledgments

我们要感谢参与我们人类评价的志愿者们，其中很多是我们的好朋友和亲人。我们要感谢朱俊辉（北京语言大学）对语言分析进行有价值的讨论。郭必扬要感谢黄海良教授和韩松桥教授（上海财经大学人工智能实验室）对本项目的主题和方向进行了深刻的反馈。张鑫要感谢赵宇（新加坡国立大学和天津大学智慧城市创新中心）分享OpenAI账号。最后，我们要感谢本项目所有团队成员的独特贡献。我们共同使这成为可能。

参考文献

[1] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[2] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Language Resources and Evaluation Conference*, pages 258–266, Marseille, France, June 2022. European Language Resources Association.

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[6] Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. Meddialog: a large-scale medical dialogue dataset. *arXiv preprint arXiv:2004.03329*, 2020.

[7] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *neural information processing systems*, 2017.

[8] Nan Duan. Overview of the nlpcc-iccpol 2016 shared task: Open domain chinese question answering. In *Natural Language Understanding and Intelligent Applications*, pages 942–948, Cham, 2016. Springer International Publishing.

[9] Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415, 2021.

[10] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: long form question answering. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3558–3567. Association for Computational Linguistics, 2019.

[11] Zhihui Fang. The language demands of science reading in middle school. *International journal of science education*, 28(5):491–520, 2006.

[12] Yao Fu, Hao Peng, and Tushar Khot. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu' s Notion*, Dec 2022.

[13] Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy, July 2019. Association for Computational Linguistics.

[14] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for

language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

[15] Biyang Guo, Yeyun Gong, Yelong Shen, Songqiao Han, Hailiang Huang, Nan Duan, and Weizhu Chen. Genius: Sketch-based language model pre-training via extreme and selective masking for text generation and augmentation. *arXiv preprint arXiv:2211.10330*, 2022.

[16] Biyang Guo, Songqiao Han, and Hailiang Huang. Selective text augmentation with word roles for low-resource text classification. *arXiv preprint arXiv:2209.01560*, 2022.

[17] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[18] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*, 2022.

[19] Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. Automatic detection of machine generated text: A critical survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2296–2309, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[20] Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882*, 2022.

[21] Michael R King. The future of ai in medicine: a perspective from a chatbot, 2022.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[23] Macedo Maia, Siegfried Handschuh, Andr'e Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[24] William Nagy and Dianna Townsend. Words as tools: Learning academic vocabulary as language acquisition. *Reading research quarterly*, 47(1):91–108, 2012.

[25] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[27] Jiameng Pu, Zain Sarwar, Sifat Muhammad Abdullah, Abdullah Rehman, Yoonjin Kim, Parantapa Bhattacharya, Mobin Javed, and Bimal Viswanath. Deepfake text detection: Limitations and opportunities. In *Proc. of IEEE S&P*, 2023.

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[29] Mary J Schleppegrell. *The language of schooling: A functional linguistics perspective*. Routledge, 2004.

[30] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

[31] SophonPlus. Chinesenlpcorpus. `https://github.com/SophonPlus/ChineseNlpCorpus`, 2019.

[32] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize from human feedback. *neural information processing systems*, 2020.

[33] Teo Susnjak. Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*, 2022.

[34] Alan M Turing. Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer, 2009.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970, 2022.

[37] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[38] Bright Xu. Nlp chinese corpus: Large scale chinese corpus for nlp, September 2019.

[39] Yi Yang, Scott Wen-tau Yih, and Chris Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. ACL - Association for Computational Linguistics, September 2015.

[40] Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385, 2019.

A Appendix

A.1 HC3 Dataset Splits Creation

我们针对HC3英语和中文分别创建了5个和7个数据集划分。大部分数据来自公开的问答（QA）数据集，具体细节如下所列。对于这些QA数据，我们直接输入问题到ChatGPT并收集至少一个答案。

我们还从维基百科和百度百科上获取了一些概念和解释，并将解释视为人类专家的答案，概念用于构造问题，详细信息请参考下述段落。

对于HC3-英语，我们创建了五个数据集划分：

1. `reddit_eli5`. 从ELI5数据集中采样 [10]。
2. `open_qa`. 从WikiQA数据集中采样 [39]。
3. `wiki_csai`。我们收集了数百个与计算机科学相关的概念的描述，这些描述来自维基百科[13]，作为人类专家对问题的回答，如“请解释什么是<concept>?”
4. `medicine`。从医学对话数据集中采样 [6]。
5. `finance`。从FiQA数据集中采样 [23]，该数据集是通过在投资主题下爬取StackExchange[14]帖子构建的。

对于HC3-Chinese，我们创建了七个数据集拆分：

1. `open_qa`。从[38]的WebTextQA和BaikeQA语料库中抽样。
2. `baike`。我们从百度百科[15]收集了一千多个与信息科学相关的概念的描述，这些描述是人类专家对类似"的问题的回答。我有一个计算机相关的问题，请用中文回答，什么是<concept>
3. `nlpcc_dbqa`。从NLPCC-DBQA数据集中抽样[8]。
4. `medicine`。从医药对话数据集中抽样[6]。
5. `finance`。从FinanceZhidao数据集中抽样[31]。
6. `psychology`。从公开的中文心理问答数据集[16]中抽样。
7. `law`。从LegalQA数据集[17]中抽样。

A.2 Additional Results

下面我们展示了中文语料库中的依赖关系的附加结果，如图所示6。结论基本与主要论文一致。

Other detailed results, including vocabulary features, sentiment analyses, and dependency parsing results for each data source are all available at our project GitHub repository at `https://github.com/Hello-SimpleAI/chatgpt-comparison-detection`.

[13] https://www.wikipedia.org/
[14] https://stackexchange.com/
[15] https://baike.baidu.com/
[16] https://aistudio.baidu.com/aistudio/datasetdetail/38489
[17] https://github.com/siatnlp/LegalQA

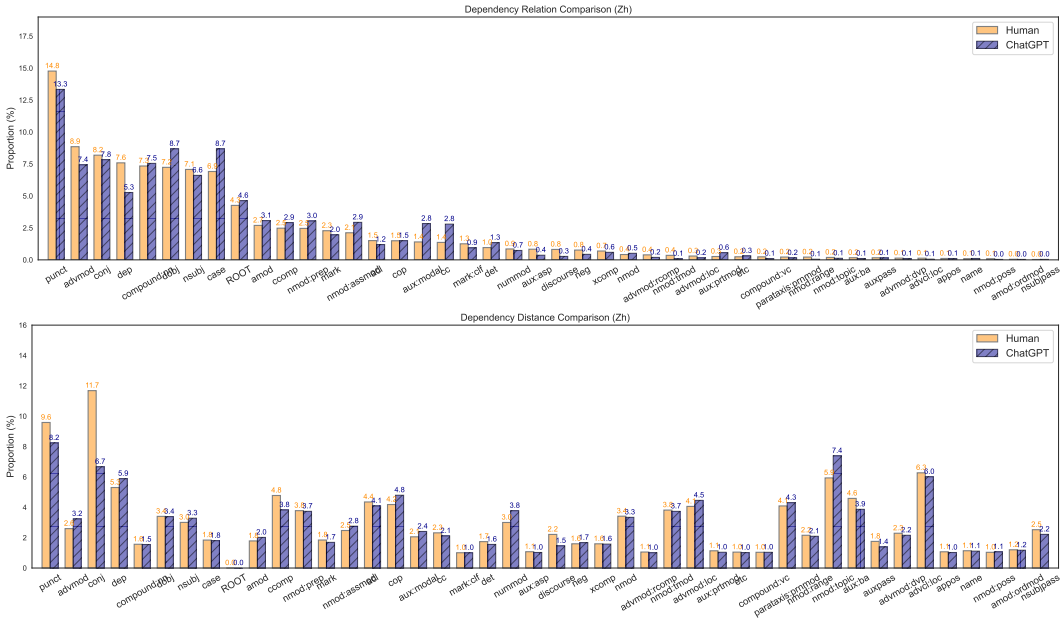


图 6: 人类和ChatGPT在HC3中文数据集上的前30个依赖关系（上）及相应的依存距离（下）比较结果。根据人类回答的依赖关系比例进行排序。

A.3 Human Evaluations Examples

请访问我们的项目GitHub存储库，查看我们人工评估的实例，网址为`https://github.com/Hello-SimpleAI/chatgpt-comparison-detection`。