HC3 Plus: A Semantic-Invariant Human ChatGPT Comparison Corpus

Zhenpeng Su[1,2]*, **Xing Wu**[1,2]*, **Wei Zhou**[1,2]†, **Guangyuan Ma**[1,2], **Songlin Hu**[1,2]†
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{suzhenpeng,wuxing,zhouwei,maguangyuan,husonglin}@iie.ac.cn

Abstract

ChatGPT has garnered significant interest due to its impressive performance; however, there is growing concern about its potential risks, particularly in the detection of AI-generated content (AIGC), which is often challenging for untrained individuals to identify. Current datasets used for detecting ChatGPT-generated text primarily focus on question-answering tasks, often overlooking tasks with semantic-invariant properties, such as summarization, translation, and paraphrasing. In this paper, we demonstrate that detecting model-generated text in semantic-invariant tasks is more challenging. To address this gap, we introduce a more extensive and comprehensive dataset that incorporates a wider range of tasks than previous work, including those with semantic-invariant properties. In addition, instruction fine-tuning has demonstrated superior performance across various tasks. In this paper, we explore the use of instruction fine-tuning models for detecting text generated by ChatGPT.

1 Introduction

In recent years, artificial intelligence has made significant advancements. Since the release of ChatGPT, large-scale language models (LLMs) fine-tuned on GPT-3.5 using Reinforcement Learning from Human Feedback (RLHF) have garnered extensive attention. ChatGPT has demonstrated strong performance across various tasks and excels at generating human-like text. As shown in (Guo et al., 2023), the Turing test reveals that it is challenging for individuals unfamiliar with ChatGPT to distinguish between texts generated by ChatGPT and those written by humans.

As more and more AI-generated content appears on social media, it could increase the risk of AI-generated disinformation and harmful text spreading due to the difficulty for humans to distinguish them (Guo et al., 2023). Therefore, how to detect the text generated by the large language model has become a recent research hotspot. In fact, a good detector can be used for information supervision and accountability, making the source of information clearer. Guo et al. (2023) introduced HC3, a dataset containing both ChatGPT-generated text and human-written text, designed for training AI text detection models. They first collected nearly $40,000$ questions and responses from human experts through social media and wiki text. ChatGPT was then used to generate answers to these questions. Using this collected data, they established two detection methods: training a logistic regression model based on GLTR (Gehrmann et al., 2019) for classification and using RoBERTa (Liu et al., 2019) to train a binary classification model.

question-answer
source:
when was the world of coca cola built
target:
It opened to the public on May 24, 2007, relocating from and replacing the original exhibit.

summarization
source:
(CNN) More than 100 schools in Pakistan have been renamed in honor of the children killed in a Taliban siege at a Peshawar school last year, ...
target:
More than 100 schools renamed in honor of students killed in a Taliban schoolhouse attack . A six-hour siege at a Peshawar school killed 145 people in December.

Figure 1: Examples of QA and summarization, the same color indicates fragments that appear in both source and target sentences.

While their detector excels in performance on their test set, it is primarily focused on the question-answering (QA) task. In the QA task, the responses produced by the language model merely need to answer the requested question, without strict adherence to the semantics of the raw question. However, for tasks such as summarization, translation, and paraphrasing, the LLM must carefully consider the semantic nuances of the input sentences. The generated sentences must stay true to the meaning of the raw sentences without deviating from them. In other words, the output sentences need to maintain semantic invariance with the raw sentences. At the same time, the raw sentence used as input can be used as a reference for the output of the LLM. In the example shown in Figure 1, when comparing summarization with QA, there is an amount of overlapping vocabulary between the source sentence and the target sentence, and they have similar semantics. That may make it more difficult to recognize that the text is generated by the LLM.

Dataset	Human		ChatGPT		
	Precision	Recall	Precision	Recall	Accuracy
CNN/DailyMail	0.63	0.99	0.99	0.42	0.71
LCSTS	0.52	0.99	0.88	0.09	0.54
WMT’19 de→en	0.55	0.66	0.58	0.46	0.56
WMT’19 fr→en	0.56	0.66	0.59	0.48	0.57
WMT’19 en→zh	0.60	0.76	0.67	0.49	0.63
HC3-Paraphrase-en	0.76	0.98	0.97	0.70	0.84
HC3-Paraphrase-zh	0.57	0.85	0.70	0.35	0.60
HC3-en	0.99	0.99	0.98	0.99	0.99
HC3-zh	0.99	0.99	0.99	0.99	0.99

Table 1: Experimental results of RoBERTa detectors in Chinese and English respectively. Human and ChatGPT are ground-truth labels for text.

We focus on translation, summarization, and paraphrasing datasets to validate the aforementioned observations. We begin by selecting several commonly used human-annotated datasets and inputting their raw sentences into ChatGPT to generate corresponding outputs. To verify whether the detector can effectively distinguish between these types of data, we mix the human-labeled target sentences from the datasets with the sentences generated by ChatGPT. These combined sentences are then provided as input to the detector, which determines whether each text is generated by ChatGPT or written by a human. We use chatgpt-detector-RoBERTa-chinese[1] to detect Chinese and chatgpt-detector-RoBERTa[2] to detect English respectively, as proposed by (Guo et al., 2023). As shown in Table 1, we find that current detectors have difficulty distinguishing the source of sentences. The prediction results are mainly presented in two cases. In the first case, as shown in the translation dataset WMT’19 en→zh results, the detector performs poorly in terms of precision and recall for both classes. The second case, as shown in the summarization dataset CNN/DailyMail (See et al., 2017) results, where the detector is always biased towards predicting text as written by a human, exhibits very high recall in the human class but very low recall in the ChatGPT class. For direct comparison with detection results on the QA samples of HC3, we propose HC3-Paraphrase, which contains paraphrases of HC3 questions generated using ChatGPT. Although ChatGPT takes roughly the same input (both question text) for QA and paraphrasing tasks, as shown in Table 1, the detection performance on the HC3-Paraphrase is greatly degraded compared to the HC3. The result validates our hypothesis that current detectors struggle to detect semantic-invariant tasks.

In order to fill the gap of HC3 under semantic-invariant tasks, we extend HC3 and propose a larger ChatGPT-generated text dataset covering translation, summarization, and paraphrasing tasks, called HC3 Plus. In addition, language models fine-tuned on a large number of tasks have demonstrated impressive abilities and generalization (Wang et al., 2022). In this paper, to explore the effectiveness of instruction fine-tuning models in detecting machine-generated text, we investigate the use of the T*k*-instruct (Wang et al., 2022) model for detecting text generated by ChatGPT. Our main contribution can be summarized as follows:

- We experimentally demonstrate that current detectors cannot detect semantic-invariant samples.
- We introduce a more extensive and comprehensive dataset, including semantic-invariant tasks, for ChatGPT-generated text detection.
- We use instruction fine-tuning to train a stronger ChatGPT-based text generation detection model.

2 Dataset Construction

We first select several widely used high-quality corpora that were annotated by humans, encompassing translation, summarization, and paraphrasing tasks.

- The CNN/DailyMail (See et al., 2017) dataset is an English-language dataset containing a large number of unique news articles written by reporters from CNN and the Daily Mail. Each example contains an article and highlights.
- The Xsum (Narayan et al., 2018) is an English-language dataset, covering a wide range of domains, and consists of a BBC article with an accompanying one-sentence summary.

[1] https://huggingface.co/Hello-SimpleAI/chatgpt-detector-roberta-chinese

[2] https://huggingface.co/Hello-SimpleAI/chatgpt-qa-detector-roberta-chinese

- The LCSTS (Hu et al., 2015) is a large-scale Chinese short-text summarization dataset utilizing naturally annotated web resources on Sina Weibo, which is a Chinese social media like Twitter.
- The news2016 corpus, proposed by CLUEbenckmark (Xu et al., 2020), is captured from the Chinese We Media (self-media) platform. Each sample contains an article and its corresponding title.
- The WMT translation dataset is a widely used collection of data for machine translation. In this paper, we follow HC3 and mainly focus on the construction of English and Chinese detectors. Therefore, we consider the translation set to include WMT'19 En→Zh, WMT'19 Zh→En, WMT'16 Ro→En, WMT'16 De→en and WMT'14 Fr→En.
- For the paraphrasing dataset, we choose to use questions proposed by HC3 as the original text of paraphrasing, which can be directly compared with the QA samples proposed by HC3.

We use `GPT-3.5-Turbo-0301` via the OpenAI API to generate target texts. Our proposed dataset, HC3-SI (HC3 Semantic-Invariance), is approximately twice the size of HC3. Details about the composition and size of HC3-SI can be found in Appendix B. We then combine HC3-SI with HC3 to create a dataset referred to as HC3 Plus for convenience.

3 Method

The language model fine-tuned on a large number of instructions shows strong generalization ability on unseen tasks. We propose instructDGGC, to utilize extensive instruction fine-tuning LLM for **D**etecting Chat**GPT**-**G**enerated **C**ontent. Our method is based on the Tk-instruct model and further instruction fine-tuning on the HC3 Plus dataset. InstructDGGC is a generative model for discriminating whether an input text is generated by ChatGPT given in-context instructions (task definition and few-shot).

First, we finetune a Language model on a multitask dataset, named Super-NaturalInstructions (Wang et al., 2022) to obtain a model LM_{inst}. The obtained LM_{inst} already has a good instruction following ability. Then we further conduct instruction fine-tuning on HC3 Plus. Our task instructions follow the same uniform schema proposed by Tk-instruct which is composed of the following parts:

- DEFINITION is the description of the task, denoted as DEF. In this paper, our current task is to detect whether the text is produced by a machine or not.
- POSITIVE EXAMPLES are input samples and their correct output, denoted as POS. In this paper, we give 2-shot each time, one is a sample generated by ChatGPT, and the other is a sample written by a human.

For a given instance, we first concatenate it with instructions as input to the model. Then, the model directly outputs the corresponding label (human or model) in an auto-regressive manner. We use the T denoted the input text, the above process can be expressed as:

$$Out = LM_{Inst}(DEF, POS, T) \qquad (1)$$

4 Experiment

First, we follow (Guo et al., 2023) to use the `Roberta-Base`[3] model for training the detector as a baseline. We train two RoBERTa models based on HC3 and HC3 Plus, named RoBERTa-HC3 and RoBERTa-HC3 Plus, respectively. Subsequently, we train the InstructDGGC model, which is divided into two steps. First, we follow the Tk-instruct to conduct instruction tuning on a large number of tasks, and then further fine-tune the model on HC3 Plus. In this section, we introduce the experimental details.

4.1 Experiment Setting

Before training the detector, we need to perform instruction fine-tuning on a large number of tasks, so that our generative model has good instruction-following ability. For the English model, we use `T`k`-instruct-base-def-pos`[4] released by (Wang et al., 2022), which has been fine-tuned on 757 English tasks. However, they did not publish a base-size model that can handle Chinese. Therefore, we conduct instruction tuning a language model LM using instruction-equipped data. We used the code released by (Wang et al., 2022), followed their settings, and trained an instruction fine-tuning model on Super-NaturalInstructions(including 24 Chinese tasks) based on mt5-base (Xue et al., 2021).

We then further conduct instruction fine-tuning on the HC3 Plus data. At this stage, for both Chinese and English models, we use the same parame-

[3] https://huggingface.co/FacebookAI/roberta-base

[4] https://huggingface.co/allenai/tk-instruct-base-def-pos