人工知能生成テキストにおける人間関与度の検出と解釈
Detection and Explanation of Human Involvement in AI-Generated Texts

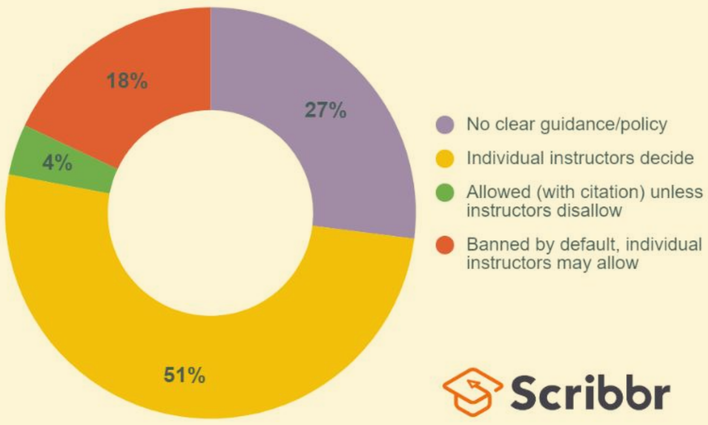
Guo Yuchen[1,2], Dou Zhicheng[1,2], Ching-Chun Chang[2], 越前 功[1,2]

[1]The University of Tokyo, [2]National Institute of Informatics

Motivation

- 73.6% students are using AI tools to help their research.
- Most students prefer to **cooperate with AI** which make it difficult to be detected.

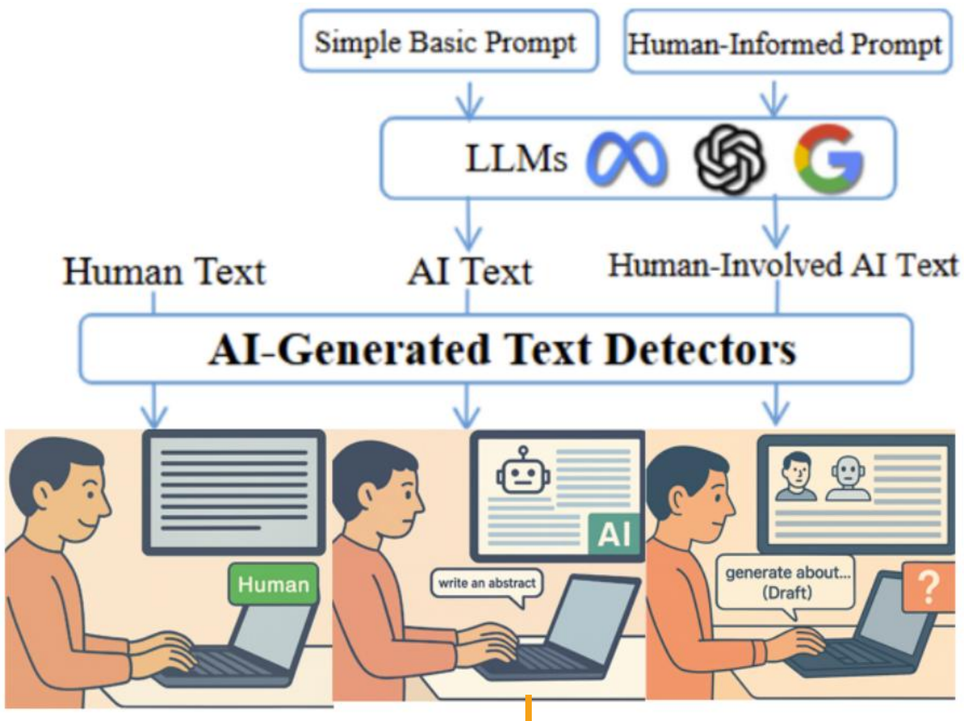
-> We need a tool to help instructors avoid **academic cheating**.



Contribution

- Created a dataset with different levels of human-machine collaboration.
- Quantifying the human contribution to generated text.
- Present a **regression detector**(MSE=0.004) with a **token classification module**(ACC=95.14).
- Have ability to generalize to other LLMs.

Hard to define what is AIGC in a complex generation



Simple basic prompt : Write an abstract of an academic paper whose title is "Attention is all you need".

Detector : **AI Generated**

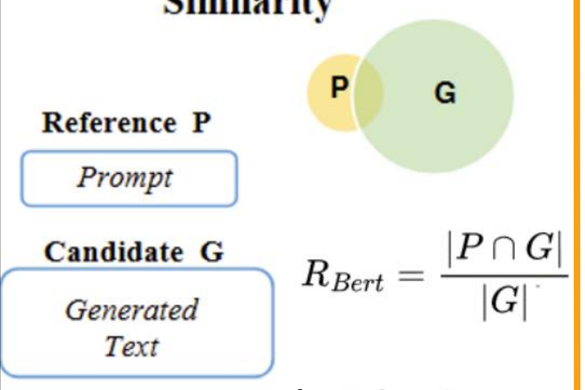
Human-informed prompt: I have written a draft/idea abstract " The dominant sequence transduction models are based on complex recurrent..." Please help me finish the whole abstract.

Detector : **AI or Human???**

Divide it to 2 class is unreasonable.

Method : BERTScore Label & RoBerTa-based Dual-head Detector

Embedding Similarity



$$R_{Bert} = \frac{|P \cap G|}{|G|}$$

How much P in G
Regression Label
0% ~ 100%(Human)

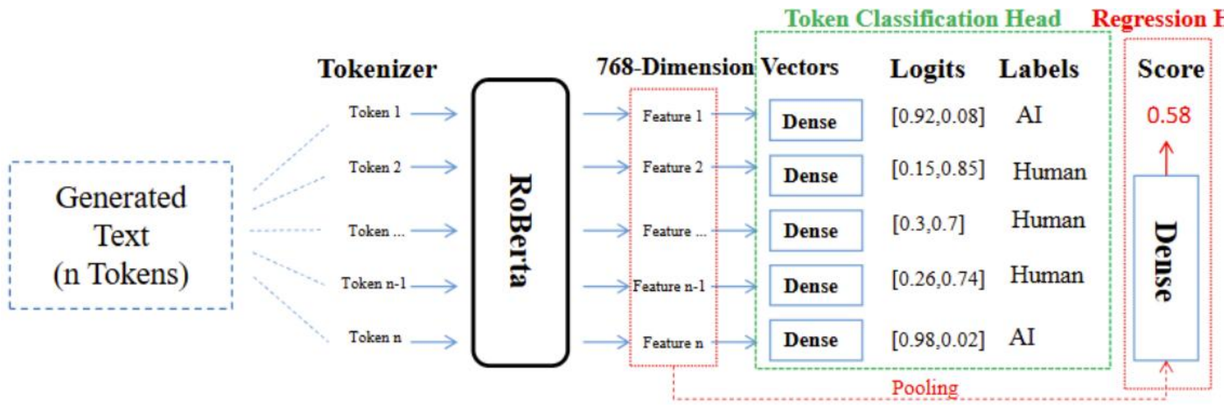
Lemmatization Mark

P: He likes **writting**.
G: I **wrote** a book.

[0 **1** 0 0]

Which token in G appears in P
Token Classification Label
0(AI) 1(Human)

Detector Structure



Without knowing P in advance, detector can output two labels.

Results & Demo

	Training set (ChatGPT)	Testing set (ChatGPT)	GPT-4	Claude-3
MSE	0.004	0.0065	0.009	0.034
ACC	99.7%	98.3%	96.9%	69.7%
	Gemini (Bard)	**Falcon-7B**	**GPA**	**Sa**
MSE	0.025	0.02	0.0073	0.03
ACC	78.3%	83.1%	97.3%	68%

MSE : Average squared difference between predicted and true values, suitable for continuous labels from **0 to 1**.

検出結果の原文

黄色の部分は人間が提供した言葉です

This paper introduces a novel approach to **learning** a **dialogue** system that **independently parameterizes** **different dialogue skills** and **learns** to **select** and **combine** them using **Attention** over Parameters (**AoP**). Our experimental **results** demonstrate that this approach **achieves competitive performance** on a **combined** **dataset** of **MultiWOZ**, In-**Car Assistant**, and **Persona**-**Chat**. Furthermore, we provide evidence that each **dialogue skill** is effectively **learned** and can be **combined** with other **skills** to produce **selective** responses. This work contributes to the advancement of AI **dialogue systems** by offering a method for independent **parameter**ization and selective combination of **dialogue skills**.

Future Work

1.Extend to other domains.
2. Processing long texts, such as full papers.

NII 越前研究室
国立情報学研究所

https://research.nii.ac.jp/~iechizen/official/index-e.html
guoyuchen@nii.ac.jp