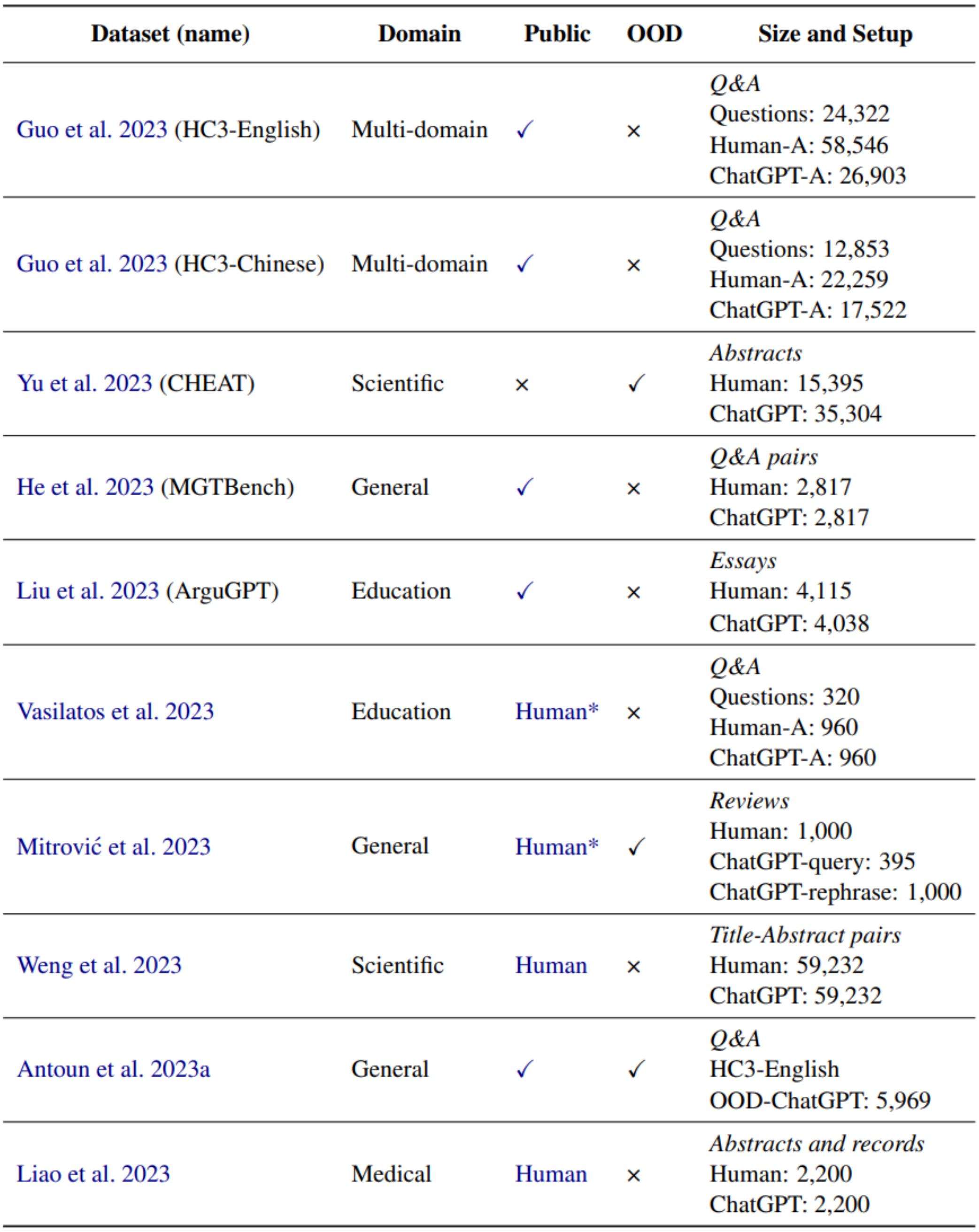
Detecting ChatGPT: A Survey of the State of Detecting ChatGPT-Generated Text

> https://arxiv.org/pdf/2309.07689.pdf

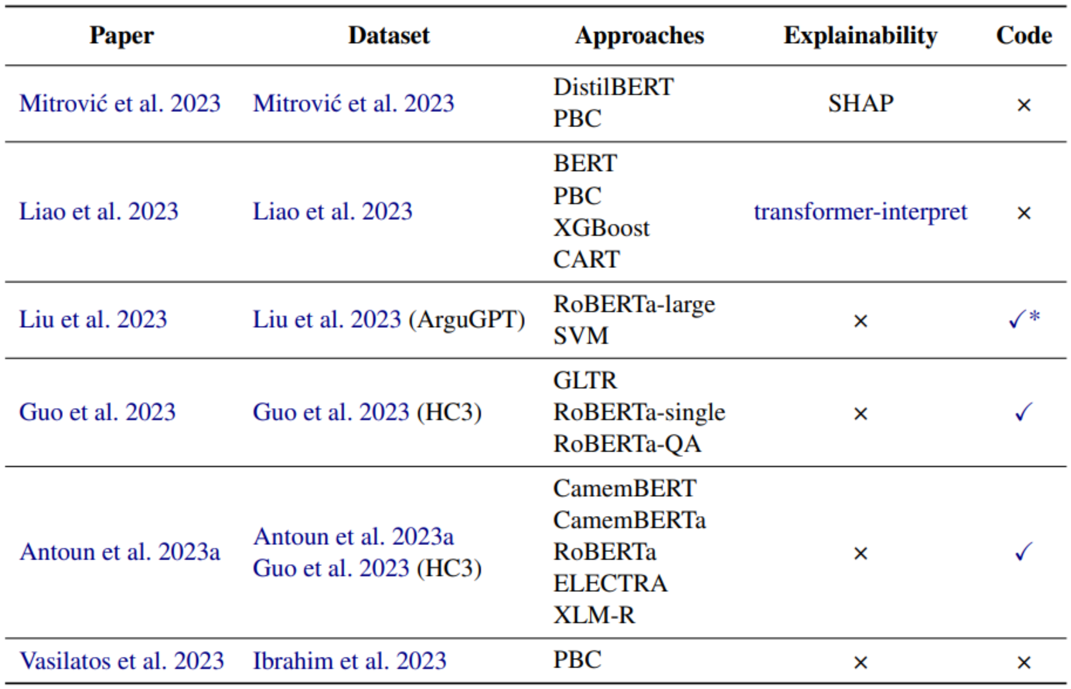
这是一篇聚焦辨别chatgpt生成文本的文章，选择看这篇survey是因为ChatGPT没开源，对它只有API level的access，所以很多方法用不了，想去看一看对于这种强但神秘的模型应该怎么去甄别

- Dataset



Dataset (name)	Domain	Public	OOD	Size and Setup
Guo et al. 2023 (HC3-English)	Multi-domain	✓	×	*Q&A* Questions: 24,322 Human-A: 58,546 ChatGPT-A: 26,903
Guo et al. 2023 (HC3-Chinese)	Multi-domain	✓	×	*Q&A* Questions: 12,853 Human-A: 22,259 ChatGPT-A: 17,522
Yu et al. 2023 (CHEAT)	Scientific	×	✓	*Abstracts* Human: 15,395 ChatGPT: 35,304
He et al. 2023 (MGTBench)	General	✓	×	*Q&A pairs* Human: 2,817 ChatGPT: 2,817
Liu et al. 2023 (ArguGPT)	Education	✓	×	*Essays* Human: 4,115 ChatGPT: 4,038
Vasilatos et al. 2023	Education	Human*	×	*Q&A* Questions: 320 Human-A: 960 ChatGPT-A: 960
Mitrović et al. 2023	General	Human*	✓	*Reviews* Human: 1,000 ChatGPT-query: 395 ChatGPT-rephrase: 1,000
Weng et al. 2023	Scientific	Human	×	*Title-Abstract pairs* Human: 59,232 ChatGPT: 59,232
Antoun et al. 2023a	General	✓	✓	*Q&A* HC3-English OOD-ChatGPT: 5,969
Liao et al. 2023	Medical	Human	×	*Abstracts and records* Human: 2,200 ChatGPT: 2,200

- Method：基本都需要fine-tune模型



Paper	Dataset	Approaches	Explainability	Code
Mitrović et al. 2023	Mitrović et al. 2023	DistilBERT PBC	SHAP	×
Liao et al. 2023	Liao et al. 2023	BERT PBC XGBoost CART	transformer-interpret	×
Liu et al. 2023	Liu et al. 2023 (ArguGPT)	RoBERTa-large SVM	×	✓*
Guo et al. 2023	Guo et al. 2023 (HC3)	GLTR RoBERTa-single RoBERTa-QA	×	✓
Antoun et al. 2023a	Antoun et al. 2023a Guo et al. 2023 (HC3)	CamemBERT CamemBERTa RoBERTa ELECTRA XLM-R	×	✓
Vasilatos et al. 2023	Ibrahim et al. 2023	PBC	×	×

感觉最大的问题还是可解释性，没有可解释性的预测很可能不鲁棒，稍微rewrite一下，变一变prompt可能就会出问题