

A Memory and Time Consumption

The details of time and resource consumption of our resources and baselines are listed in the following table.

Table 4: The memory and time consumption of DPIC and baselines.

Methods	Time Cost	Memory Consumption
Fast-DetectGPT	0.273s	25447 MB
RADAR	0.120s	1355 MB
RADAR	0.120s	1355 MB
Ghostbuster	2.493s	<500MB
DPIC	2.665s	28105 MB

B Details of the prompt templates

Both p_{rc} and p_{rg} are shown in Table 5.

Table 5: The prompt templates we use to obtain the re-constructed prompt and regenerated text from ChatGPT. <TEXT> represents the text to be detected, and <QUESTION> is the reconstructed prompt of the text to be detected.

Prompt template	
p_{rc}	I want you to play the role of the questioner. I will type an answer in English, and you will ask me a question based on the answer in the same language. Don't write any explanations or other text, just give me the question. <TEXT>
p_{rg}	<QUESTION>

C Examples of DPIC

We show an example of the detection process using DPIC. Specifically, given a candidate text, we utilize a fixed intermediary LLM, leveraging the LLM's powerful inductive capabilities, to regenerate prompt based on the candidate text. The prompt are then used to obtain the regenerated text by the intermediary LLM. The temperature we use in re-generation process is 0.2.

D Detection performance of other datasets

To fully test the detection performance of DPIC, we constructed three datasets. The three datasets were collected from scratch as a test set for detecting the algorithm's generalization ability. Human texts were taken from Wikipedia , community split from CovidQA dataset [26], and abstract data from SUMMAC Computation and Language (cmp-lg) corpus , respectively. To ensure that all comparative methods have not encountered the test data before testing, we use prompts on different datasets to collect LLM-generated texts from target language models. These datasets encompass diverse topics such as encyclopedic content, medical information, and academic discourse and contain both specialized writing and colloquial text, serving as a testbed to evaluate the generalization capabilities of various detectors for identifying out-of-domain text.

If not otherwise specified, when generating data corresponding to these four datasets using LLM, set the sampling temperature of LLM to 0.7, as has been done in previous research [19]. The prompts used for generating text for these four generalization datasets are as follows:

- **Wiki:** Please explain what is <concept>?
- **CovidQA Community:** <question>?

<https://www.wikipedia.org/>

https://www-nlpir.nist.gov/related_projects/tipster_summac/cmp-lg.html

Table 6: An example of the detection process using DPIC

Please explain what is “Spatial index”?		
	Human	ChatGPT
Candidate Text	A spatial database is a general-purpose database (usually a relational database) that has been enhanced to include spatial data that represents objects defined in a geometric space, along with tools for querying ...	A spatial index is a data structure that is used to efficiently store and query data that represents objects defined in a geometric space. It is designed to support spatial queries, which are queries ...
reconstructed Prompt	What is a spatial database and what types of objects can it represent?	What is a spatial index and how is it used to efficiently store and query data that represents objects defined in a geometric space?
regenerated Text	A spatial database is a database that is optimized to store and manage spatial data, which is data that has a geographic or spatial component. It is designed to handle data that is related to physical locations on the earth’s surface ...	A spatial index is a data structure that is used to efficiently store and query data that represents objects defined in a geometric space. It is a type of index that is designed to handle spatial data, ...

- **SUMMAC Abstract:** Please write an abstract with the title <title> for the research paper.

Both human texts and LLM-generated texts may contain some obvious indicating words that may influence the effectiveness of models [14]. Therefore, we conduct data cleaning on all data in the aforementioned datasets, removing indicating words corresponding to human-written and machine-generated text. In the experiment, we randomly selected 200 instances from each dataset for testing purposes.

The results are shown in Table 7. From the results, it can be seen that DPIC still achieves the best detection performance on three different types of datasets, which confirms the effectiveness of our method.

Table 7: The detection performance of baselines and DPIC on three datasets generated by ChatGPT.

Methods	Wiki	CovidQA Community	SUMMAC Abstract	Avg.
RoBERTa-base	0.9533	0.9164	0.9346	0.9347
RoBERTa-large	0.9229	0.8442	0.9483	0.9051
Likelihood	0.8206	0.9301	0.9998	0.9168
Entropy	0.6208	0.2980	0.0886	0.3358
LogRank	0.8486	0.9385	1.0000	0.9290
LRR	0.8717	0.9297	0.9747	0.9253
DNA-GPT(ChatGPT)	0.7893	0.7709	0.9178	0.8260
NPR	0.7365	0.8895	0.9211	0.8490
DetectGPT	0.1047	0.6891	0.6911	0.4949
Fast-DetectGPT	0.9611	0.9671	1.0000	0.9760
DPIC (ChatGPT)	0.9879	0.9923	1.0000	0.9934