Beyond Text-Intrinsic Features: An Agentic Framework for Evidence-Based AI-Generated Text Detection

Chi Cui, Yixin Wu, Michael Backes, Yang Zhang*

CISPA Helmholtz Center for Information Security
{chi.cui, yixin.wu, director, zhang}@cispa.de

Abstract

With the rapid advancement of Large Language Models (LLMs), AI-generated text has become increasingly difficult to distinguish from human-authored content, raising serious concerns about online misinformation and academic integrity. Although plenty of methods have been proposed, they often suffer from high false positives on formal or highly structured human-written texts. Moreover, while external contextual information can provide valuable signals for verification, it remains largely underutilized in existing detection frameworks. In this paper, we propose an agentic framework driven by LLM for AI-generated text detection that integrates multiple standalone detectors with an online search module to retrieve and leverage external evidence. Evaluated on a diverse real-world dataset spanning four distinct sources, our agent significantly outperforms standalone detectors, achieving an overall accuracy of 93.90%. We show that online search acts as an effective error-correction mechanism under detector conflict and incorrect agreement, improving accuracy by 10.81% and 90.91%, respectively. We further conduct a failure analysis, identifying key limitations and providing directions for future research. [1]

1 Introduction

With the rapid advancement of large language models (LLMs), such as ChatGPT (OpenAI, 2026a), Gemini (Google DeepMind, 2026), and DeepSeek models (DeepSeek, 2026), these models have demonstrated a remarkable ability to generate fluent, coherent, and human-like text. Though these capabilities have significantly improved productivity in a wide range of applications, the widespread adoption of AI-generated text also introduces potential risks and raises serious concerns regarding trust, authenticity, and misuse. For example, AI-generated text can be exploited to produce fake news articles, manipulate online reviews, or create misleading comments on social media platforms, thereby influencing public opinion and distorting information ecosystems. Recent reports (Spero, 2025) indicate that 74% of AI-generated reviews gave products a five-star rating on Amazon, potentially misleading users into overestimating product quality. In addition, AI-generated text can serve as a powerful tool for academic dishonesty and other forms of cheating (Olander, 2025). These emerging risks have made it increasingly urgent to develop effective and reliable methods for distinguishing AI-generated text from human-written text.

Currently, most approaches for identifying AI-generated text rely on supervised detectors (e.g., MPU (Tian et al., 2024), RADAR (Hu et al., 2023)) or statistical scoring methods (e.g., DetectGPT (Mitchell et al., 2023) and Binoculars (Hans et al., 2024)). In general, researchers collect large-scale datasets of human-written and AI-generated text, and then either train classifiers or design scoring methods to capture distributional differences between the two, with strong performance primarily demonstrated on benchmark datasets.

However, while these approaches have demonstrated promising results under controlled experimental settings, they also suffer from several inherent limitations. First, many existing detectors are trained or calibrated on fixed and closed datasets, making them vulnerable to distribution shifts. One consequence is that they can exhibit systematic false positives on formal or highly structured human-written texts. For instance, prior work shows that AI detectors can misclassify real scientific abstracts written before the emergence of modern LLMs as AI-generated (Rashidi et al., 2023). Similarly, OpenAI reports that an AI-generated content detector incorrectly labeled clearly human-written texts, such as Shakespeare's literary works

*Corresponding author.

[1] Code and data are available at https://github.com/TrustAIRLab/AIGT-Agent

and the 1776 Declaration of Independence, as AI-generated (OpenAI, 2026c). These examples suggest that scientific writing, legal documents and literary works are particularly challenging because they often contain standardized phrasing, dense terminology, and low stylistic variation. As a result, detectors that rely primarily on surface-level or distributional text features may incorrectly flag authentic human-written documents as AI-generated. Second, current detectors predominantly focus on features intrinsic to the text itself, while largely ignoring auxiliary information beyond the text. In reality, contextual information related to the text can provide strong evidence about its origin. For example, text appearing in Shakespeare's works can be attributed to human authorship, whereas text found in an AI-generated novel is likely AI-generated. However, to the best of our knowledge, such contextual information is not incorporated into existing detectors.

Motivated by these observations, we propose a novel approach that reframes AI-generated text detection as an agent-based reasoning problem. Specifically, we construct an LLM-based agent designed to identify AI-generated text (AIGT) by integrating multiple complementary information sources. Our proposed AIGT detection agent consists of two primary modules: (1) an **Ensemble Detection Module** and (2) an **Online Search Module**. The ensemble detection module, consisting of multiple standalone AI-generated text detectors, provides predictions based on intrinsic textual features, while the online search module retrieves and summarizes external evidence from the web that may indicate whether the text is AI-generated or human-written. The main agent then aggregates and reasons over the outputs of these two modules to produce a final decision.

To evaluate the effectiveness of our approach in realistic settings, we construct a real-world AIGT dataset spanning four writing domains: Legal, Literature, Scientific, and News. The human-written portion is collected from sources such as the Federal Register (2026), Project Gutenberg (2026), SciXGen (Chen et al., 2021), and XSum (Narayan et al., 2018), while the AI-generated portion is produced by prompting multiple LLMs to continue human-written passages while preserving the original topic and style. This construction creates a challenging evaluation setting: human texts often come from formal or historically authored sources, while generated texts closely follow the topic, style, and register of their corresponding human passages. Experimental results demonstrate that our agent consistently outperforms standalone detectors across all evaluated datasets, achieving an overall accuracy of 93.90%. Compared with the two detectors integrated into the agent, our agent improves over Binoculars by 25.85% and over MPU by 11.45%. Moreover, the online search module contributes substantially, especially when the integrated detectors produce conflicting predictions, where contextual evidence from online search improves accuracy by 10.81%. Even when the detectors reach consensus, online search remains beneficial by correcting cases where both detectors are incorrect, allowing the agent to function as an effective error-correction mechanism. We further conduct a failure analysis showing that the remaining errors mainly arise from unavailable external evidence, ambiguous webpage signals, and domain-specific detector bias.

Our main contributions are as follows:

- We propose the first agentic framework for AI-generated text detection, which integrates ensemble detection and online search modules to reason about the origin of a given text rather than relying on a standalone detector.
- We are the first to incorporate contextual information beyond the text itself into AI-generated text detection, demonstrating that external evidence can effectively complement text-intrinsic features and improve detection robustness.
- We construct the evaluation dataset by collecting human-written and AI-generated text from diverse real-world sources. Extensive experiments show that our proposed agent significantly outperforms existing detectors. Component-level analysis highlights the contribution of each module.

2 Preliminary

2.1 AI-Generated Text Detection

Existing approaches for AI-generated text detection generally fall into two categories: metric-based and model-based methods. Metric-based methods leverage statistical properties of language model outputs, such as log-likelihood, entropy, and perplexity, to distinguish AI-generated text from human-written text. For example, DetectGPT (Mitchell et al., 2023) exploits geometric properties of the

log-probability function. While computationally efficient, these methods rely on assumptions about model distributions and often degrade when applied to text generated by unseen LLMs.

Model-based methods treat detection as a supervised classification task, typically fine-tuning pretrained language models on labeled datasets (e.g., OpenAI detector (OpenAI Community, 2019)). Although effective in-distribution, they often suffer from overfitting and limited generalization to out-of-distribution (OOD) data. Prior work has explored contrastive or adversarial learning to improve robustness (Guo et al., 2024; Liu et al., 2024; Hu et al., 2023), as well as using LLMs themselves as detectors (Bhattacharjee and Liu, 2023), but these approaches remain constrained by supervised training or exhibit limited reliability.

In contrast, our work introduces an LLM-agent-based framework that integrates existing detectors with contextual evidence from external sources, enabling more robust and generalizable AI-generated text detection in real-world settings.

2.2 LLM Agent

An LLM agent typically consists of an LLM as the core reasoning engine, augmented with modules such as memory, tool use, and planning, as well as prompt engineering techniques like chain-of-thought reasoning (Wei et al., 2022). Recently, LLM agents have been widely applied across domains, including scientific research (He et al., 2025; Schmidgall et al., 2025; Jin et al., 2024), software development (Qian et al., 2024; Yang et al., 2024), and medical diagnosis (McDuff et al., 2025; Tu et al., 2025; Tang et al., 2024).

In this work, we extend LLM agents to AI-generated text detection, leveraging their reasoning abilities to retrieve and analyze web-based information for more reliable decisions.

3 Framework

As illustrated in Figure 1, the proposed agentic framework is composed of a main agent equipped with two customized modules: the **Ensemble Detection Module** and the **Online Search Module**. Upon receiving an input text, the two modules process it independently. The agent then aggregates the returned insights to formulate a comprehensive judgment on whether the text is AI-generated or human-written.

3.1 Ensemble Detection Module

Nowadays, LLMs can generate text that closely resembles human writing, making it difficult for humans to distinguish between the two. However, existing text-based AIGT detectors (hereafter referred to as *detectors*) can still extract useful discriminative signals from text and provide predictions about its origin. For this reason, we incorporate these detectors into our agent. We select a set of detectors as tools available to the main agent. For each input query, all detectors independently determine whether the text is AIGT or human-written, and return their predictions to the main agent. These outputs serve as important reference signals that the main agent considers when making its final decision.

3.2 Online Search Module

As we discussed in Section 1, although detectors demonstrate strong capabilities, they still have limitations due to their dependence on specific training datasets. As a result, their performance may degrade when facing human-written formal texts. To compensate for these limitations, we incorporate external information to support the detection process. Thus, we introduce an online search module (the detailed workflow is shown in Figure 3 in Appendix A) that retrieves information related to the query text from the web and extracts evidence that can help determine its origin. In this section, we describe the two main stages of the online search module: web retrieval and evidence extraction.

Web retrieval. In this stage, our goal is to collect webpages that contain the query text, since such pages may provide useful external evidence for determining its origin. To find webpages containing the exact text, we first perform an *exact search*. We extract the first fifteen words from the input text, enclose them in quotation marks, and query them using Google Search. This forces the search engine to return only webpages that contain the exact same sequence of words, which can then be directly passed to the next stage. However, not all texts appear online with identical wording. Slight variations in formatting, punctuation, or wording may cause the exact search to fail. Thus, when no exact match is found, we proceed to a *blur search* to find the webpages containing similar text. In the blur search, we extract the first two hundred characters of the text and submit them to Google Search without quotation marks. This retrieves

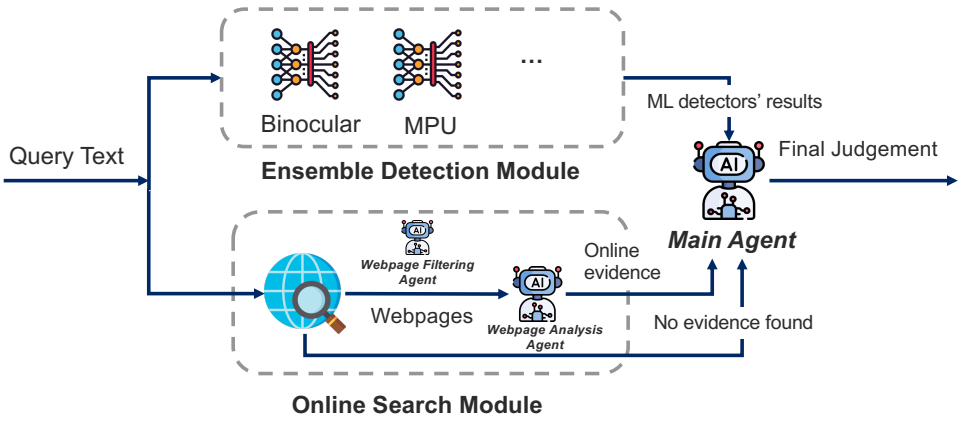

Figure 1: Overview of the proposed agentic framework.

webpages that contain relevant keywords or semantically similar content rather than an exact match. Next, we filter the webpages returned by the search process to locate webpages with a high likelihood of containing similar content. We employ a *Webpage Filtering Agent* to identify webpages that are likely to be relevant to the query text. For example, for research-related text, webpages from arXiv are preferred over webpages from BBC News. After selecting the relevant webpages, we extract their textual content and measure its similarity to the query text. Only webpages with a similarity score above 0.8 are considered to contain sufficiently similar content, and these pages are passed to the next stage. However, some texts still do not correspond to any webpages online. In such cases, we conclude that no online evidence is available and terminate the online search module.

Evidence extraction. After identifying the webpages that contain the query text, we extract useful evidence from these pages to help determine the origin of the text. To accomplish this, we design a *Webpage Analysis Agent* whose task is to analyze each webpage and extract relevant signals. This agent is equipped with two tools: `visit_webpage` is to retrieve the full textual content of the webpage, `get_last_modified` is to obtain the last modified time of the webpage. When a webpage URL is provided to the agent, it uses these tools to obtain both the webpage content and its modification time. Based on this information, the agent outputs all evidence found on the webpage, together with a reasoning explanation.

4 Experiment

In this section, we conduct comprehensive experiments to evaluate the proposed agent. Section 4.1 introduces our evaluation dataset, which comprises texts from four domains. Section 4.3-Section 4.5 then report the main experimental results and analyze the agent's two key components. In Appendix G, we provide further ablation studies investigating the effects of the LLM backbone, prompt composition, and removal of the online search module. Finally, in Appendix H, we evaluate the agent's robustness in the more challenging setting of adversarial paraphrasing.

4.1 Dataset

To evaluate our agent in real-world settings, we construct a diverse benchmark covering four major writing domains: Legal, Literature, Scientific, and News. For the Legal domain, we collect passages from the Federal Register (Federal Register, 2026), restricting the source documents to entries published before 2023. This domain contains formal regulatory and administrative language, which is often highly templated. For the Literature domain, we collect English literary classics from Project Gutenberg (Project Gutenberg, 2026). These books are typically published no later than 1930, and thus substantially predate modern LLMs. They represent canonical human-written Literature with diverse styles. For the Scientific and News domains, we use the corresponding subsets from MAGE (Li et al., 2024), an existing AIGT detection benchmark, whose human-written texts are sourced from SciXGen (Chen et al., 2021) and XSum (Narayan et al., 2018), respectively. The Scientific domain captures technical academic writing, while the

News domain contains concise news summaries.

To construct the AI-generated portion, we follow the MAGE construction protocol, prompting LLMs to continue existing human passages while preserving the original style, register, and topic. (Details shown in Appendix B.) For each dataset, we construct the corresponding AIGT samples using four generator models: `Qwen3-8B` (Qwen, 2025), `GPT-3.5-turbo` (OpenAI, 2023), `GPT-5-mini` (OpenAI, 2025), and `Gemini-2.5-flash` (Google, 2025).

For each domain, we sample 500 human-written texts and produce 500 AI-generated texts, evenly distributed across the four generator models. In total, we build a dataset containing 4,000 texts across four domains and four generator models.

4.2 Experiment Setup

Agent configuration. The framework is implemented using the `smolagent` (Hugging Face, 2025) architecture. All agents in our framework are powered by `gpt-5-mini`. Detailed system prompts are provided in Appendix E.

Ensemble detection module settings. For detection, we employ Binoculars (Hans et al., 2024) and MPU (Tian et al., 2024) as the underlying engines (details in Appendix D). Notably, we configure Binoculars with a low-FPR (False Positive Rate) threshold to prioritize the precision of AI-generated content identification. For MPU, we utilize the `AIGC_text_detector_env3` model to extract detection signals. Additionally, in Section 4.4, we further analyze how different choices of detectors influence the agent's performance.

Baselines. We compare our agent with four baselines. First, we include two standalone detectors, Binoculars and MPU, which are also used in our ensemble detection module. Second, we evaluate direct classification with `GPT-5-mini`, since the model may leverage its internal knowledge and reasoning ability to identify generated text. Finally, we construct a majority vote baseline that aggregates the predictions from the two detectors and directs `GPT-5-mini` classification.

Evaluation metrics. We evaluate all methods using three metrics: accuracy, F1 score, and balanced error rate (BER). Throughout our experiments, we label AIGTs as 1 and human-written texts as 0. Accuracy measures the overall proportion of correctly classified samples. F1 score reflects the balance between precision and recall for identifying AIGT. BER measures the average of the false positive rate (FPR) and false negative rate (FNR) (Zhao et al., 2013). A lower BER indicates more balanced performance across human-written and AI-generated texts.

4.3 Main Result

AIGT agent outperforms all baselines across all datasets. As shown in Table 1, our agent outperforms all baselines, achieving an overall accuracy of 93.90%. This is 7.93% higher than the majority vote, and improves over the two detectors by 25.85% compared with Binoculars and 11.45% compared with MPU. Moreover, we observe that our agent achieves the best performance in every domain. For example, in the Legal domain, where formal regulatory language can be difficult to distinguish from AIGT, Binoculars and MPU achieve accuracies of 61.50% and 65.20%, respectively, while our agent improves the accuracy to 88.70%. In the Scientific domain, direct `GPT-5-mini` classification nearly fails, achieving only 51.20% accuracy, while our agent reaches 96.70%. These results suggest that the agent is not merely tuned to a single domain; its effectiveness is consistent across different writing styles and data sources.

This success can be attributed to two main factors. First, the ensemble detection module enables the agent to integrate multiple complementary signals, effectively mitigating the domain-specific biases of individual detectors. Second, the online search module enhances the agent's accuracy by providing crucial external evidence; by cross-referencing text fragments with existing web sources, it helps compensate for the inherent uncertainty of standalone detectors.

The agent achieves stronger AIGT detection while maintaining the lowest balanced error across domains. Beyond accuracy, the F1 score results show that the agent is consistently better at identifying AIGT. As shown in Table 1, our agent achieves the highest F1 score across all four datasets, with a total F1 score of 94.14%, outperforming majority vote at 86.70%. This indicates that our agent is more effective at balancing precision and recall when detecting AIGT, rather than simply favoring one type of prediction.

Additionally, our agent shows the lowest BER across all datasets. As shown in the table, the agent reduces the total BER to 6.10%, compared with 17.55% for MPU and 31.95% for Binoculars. We further observe that MPU tends to suffer from a relatively high FPR, while Binoculars exhibits a high

Methods	Legal	Literature	Scientific	News	Total
Binoculars	61.50 / 39.37 / 38.50	77.40 / 71.17 / 22.60	63.20 / 42.32 / 36.80	70.10 / 57.71 / 29.90	68.05 / 53.76 / 31.95
MPU	65.20 / 74.03 / 34.80	82.00 / 83.96 / 18.00	86.60 / 88.14 / 13.40	96.00 / 96.13 / 4.00	82.45 / 84.82 / 17.55
GPT-5-mini	66.20 / 73.22 / 33.80	79.60 / 79.01 / 20.40	51.20 / 67.16 / 48.80	65.10 / 72.41 / 34.90	65.53 / 72.34 / 34.48
Majority vote	75.00 / 78.92 / 25.00	88.20 / 87.23 / 11.80	86.40 / 87.99 / 13.60	94.30 / 94.17 / 5.70	85.97 / 86.70 / 14.03
Agent	**88.70 / 89.76 / 11.30**	**93.80 / 93.84 / 6.20**	**96.70 / 96.79 / 3.30**	**96.40 / 96.50 / 3.60**	**93.90 / 94.14 / 6.10**

Table 1: Accuracy / F1 score / BER (%) of the proposed agent compared with four baselines across four domains.

FNR, making their errors biased toward different failure modes (details shown in Appendix C). In contrast, our agent better balances these two types of errors and achieves a substantially lower BER. This indicates that it not only detects AIGT more effectively, but also better controls false alarms on human-written text and missed detections of AIGT.

4.4 Ensemble Detection Module Analysis

The ensemble detection module plays a central role in our agent by providing key signals for the final decision. To ensure robust system design, we analyze this module in detail, examining the performance of different detectors and the effect of varying the number of detectors.

4.4.1 Performance of Different Detectors

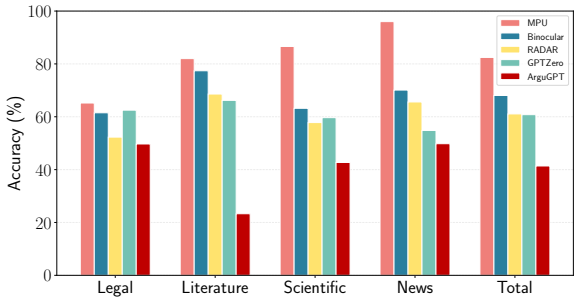

Figure 2: Accuracy of five different detectors across four different domains.

In order to investigate the capability of different detectors to distinguish AI-generated text from human-written text in real-world scenarios, we evaluate five representative detectors on the datasets introduced in Section 4.1. The details of each detector are shown in Appendix D.

As we show in Figure 2, MPU and Binoculars rank first and second in overall accuracy among the evaluated detectors. Specifically, MPU and Binoculars achieve overall accuracies of 82.45% and 68.05%, respectively. These results indicate that both MPU and Binoculars provide strong detection signals, making them a promising choice for supporting robust decision-making in the agent framework.

Detector configurations	Accuracy (%)
M	93.80
M+B	**93.90**
M+B+R	93.12
M+B+R+G	89.20
M+B+R+G+A	88.60

Table 2: Agent accuracy under different ensemble detector configurations. **M** denotes MPU, **B** denotes Binoculars, **R** denotes RADAR, **G** denotes GPTZero and **A** denotes ArguGPT. The performance follows the order **M > B > R > G > A**.

4.4.2 Performance of Different Agent Variants

We then evaluate agent variants on the datasets described in Section 4.1, where each variant corresponds to a different detector ensemble constructed by adding detectors in order of their individual performance. This design allows us to systematically examine the effect of progressively incorporating multiple detectors into the ensemble, without exhaustively enumerating all possible detector combinations. Based on the experimental results shown in Table 2, we make the following observations:

Complementary detectors improve agent performance over a single detector. As shown in Table 2, the agent achieves the best overall accuracy when using the combination of MPU and Binoculars, reaching 93.90%. This is slightly higher than using MPU alone, which achieves 93.80%, suggesting that Binoculars provides complementary signals to MPU. This complementarity is consistent with our analysis in Appendix C: MPU tends to have a higher FPR, while Binoculars tends to have a higher FNR. Combining these two detectors therefore provides the agent with more balanced signals, helping it compensate for the biased error patterns of each individual detector.

Adding more detectors does not necessarily improve the agent. Further increasing the number of detectors beyond two leads to a degradation in performance, with accuracies dropping to 93.12%, 89.20%, and 88.60%. This decline can

be explained by two factors. First, some of the additional detectors are less reliable, introducing noisy or misleading signals. Second, an excessive number of detectors' signals can weaken the relative influence of the online search module in the agent's decision-making process, reducing the effectiveness of external evidence even when it is informative.

4.5 Online Search Module Analysis

In this section, we examine how the online search component influences the performance of our agent. First of all, we investigate what types of online evidence help the agent make correct decisions. Then, we analyze the proportion of texts that can be retrieved through online search. Furthermore, we study how the contribution of the online search module varies across different ensemble detection module outcomes. Additionally, we show that removing the online search module substantially degrades the agent's performance, as detailed in Section G.3.

4.5.1 Evidence Analysis

To understand which types of online evidence guide the agent's decisions, we examine texts that can be retrieved through online search and manually analyze the reasoning behind the agent's judgments. We categorize the key evidence into three types, each one with an example of agent reasoning shown in Table 10 in Appendix F.

- **Human authorship.** If an article's author appears to be a well-known writer or has a real name with a detailed profile, the agent tends to assume the author is human and thus classifies it as a human-written text.
- **Platform.** The platform on which a text appears also influences the agent's judgement. Texts from platforms where human writing is more common, such as the government, BBC News, or research websites, are often considered human-written.
- **Publication time.** When the retrieved webpage provides a publication date before the widespread adoption of modern LLMs, the agent treats it as strong evidence of human-written text.

These three types of evidence show that the effectiveness of online search depends on the availability of explicit, reliable, and contextually meaningful signals. When such evidence is present, online search can effectively supplement or correct ensemble detection results; otherwise, its guidance is limited.

Dataset	Legal	Literature	Scientific	News	Total
Hit Rate (%)	36.60	34.70	36.50	26.20	33.50

Table 3: Online search hit rate of each dataset.

Dataset	Conflict	Consensus	
		Correct	Wrong
Legal	97.85 / 78.27	100.00 / 100.00	100.00 / 0.00
Literature	98.33 / 86.09	99.11 / 100.00	66.67 / 0.00
Scientific	96.33 / 93.32	99.61 / 100.00	– / 0.00
News	95.24 / 91.72	97.10 / 100.00	– / 0.00
Total	97.52 / 86.71	98.82 / 100.00	90.91 / 0.00

Table 4: Agent accuracy (%) under **Hit/Miss** conditions, grouped by ML detector outcomes.

4.5.2 Hit Rate

To analyze whether a text can be found online, we define a successful online search as a **Hit** and an unsuccessful one as a **Miss**. Table 3 reports the hit rates across different datasets. We observe that, overall, 33.50% of the texts can be found online. The Legal domain shows the highest hit rates at 36.60%, followed closely by Scientific at 36.50%. Although News has the lowest hit rate, 26.20% of its texts can still be found online. These results indicate that a substantial portion of the data can be retrieved through online search, demonstrating the feasibility of using online evidence as part of the detection process.

4.5.3 Online Search Contribution

In this section, we analyze how the online search module influences the final judgment by examining its performance under two conditions: when the detectors in the ensemble conflict and when they agree. We evaluate the impact of successful online searches (**Hit**) versus unsuccessful ones (**Miss**) on the agent's accuracy.

Scenario-I: detector conflict. In this scenario, the agent faces contradictory internal signals and must rely on external evidence to resolve the ambiguity. As shown in Table 4, a successful online search (**Hit**) leads to a total accuracy of 97.52%, an improvement of 10.81% compared to the 86.71% accuracy when the search is a **Miss**. When detectors conflict, the agent gives more weight to the online search results. A successful search provides the necessary evidence to make an informed deci-

sion, while a miss leaves the agent with insufficient information to make a reliable judgment.

Scenario-II: detector consensus. As shown in Table 4, when both detectors provide the same prediction, we analyze the online search module's impact by examining scenarios where these detectors are (as revealed by ground truth) both correct and both wrong. When both detectors are correct, their consensus prediction already offers a strong signal. In such cases, a successful online search (**Hit**) results in a total accuracy of 98.82%, lower than the 100.00% accuracy achieved when the search yields no results (**Miss**). This suggests that while the agent still considers online evidence, its additional value is limited and may introduce noise when detectors are already reliably correct. Conversely, in situations where both detectors are wrong, the online search module becomes crucial for error correction. A successful search (**Hit**) salvages the prediction in 90.91% of cases, whereas an unsuccessful search (**Miss**) results in 0.00% accuracy, as the agent has no other information to override the incorrect consensus. This underscores the continued importance of the online search module as a critical error correction mechanism, even when primary detectors show agreement.

5 Failure Analysis

Although the proposed agent substantially improves over standalone detectors, it still makes errors under several recurring conditions. We analyze these failure modes to better understand the boundary of evidence-based AIGT detection.

Failure due to unavailable external evidence. The first failure mode occurs when the online search module cannot retrieve reliable webpages related to the input text. In these cases, the agent loses access to its main source of contextual evidence and must rely primarily on the detector outputs. This limitation is most evident for human-written news texts, where 49.6% are classified as *Not Found*, likely because older news webpages may be moved, archived, or no longer stably indexed. This shows that online search is constrained by webpage availability and search engine coverage. For texts without retrievable online evidence, the agent's performance is limited because it must rely primarily on detector outputs.

Failure due to ambiguous or misleading online evidence. Even when a relevant webpage is retrieved, the evidence may not directly indicate whether the text is human-written or AI-generated. For example, a webpage may contain the same passage but provide no reliable author identity, publication time, or explicit disclosure of AI generation. In such cases, the agent may rely on less reliable cues, such as the apparent credibility of the webpage or the type of platform on which the text appears, and can therefore be misled. For example, for some News texts sourced from the XSum dataset, the online search module retrieves matching text from Hugging Face or Kaggle dataset pages; because these two platforms are strongly associated with AI and machine learning datasets, the agent may incorrectly treat these matches as evidence of AI generation. As a result, a successful search hit is not always equivalent to reliable evidence. This shows that the agent remains limited when retrieved webpages contain the target text but do not provide clear authorship or generation evidence.

Failure due to domain-specific detector bias. The remaining errors often reflect systematic biases in the detector signals. As shown in Appendix C, MPU tends to produce more false positives, while Binoculars tends to produce more false negatives. These biases become especially harmful in domains with highly regular or formal language. For instance, legal and scientific texts often contain formulaic phrasing, dense terminology, and standardized structure, which can resemble model-generated text even when written by humans. Conversely, fluent AI-generated continuations in literary or news domains may appear natural enough to evade detectors. Although the agent can correct many such cases when strong online evidence is available, it remains vulnerable when detector bias and weak external evidence point in the same direction.

Overall, these failure modes show that the proposed framework is most reliable when detector outputs and external evidence provide complementary signals. When external evidence is missing or misleading, biased detector outputs can still lead the agent to incorrect judgments. However, the impact of these cases is limited relative to the full evaluation set, and the agent maintains strong performance across all four domains.

6 Conclusion

We propose an evidence-based agentic framework for AI-generated text detection that integrates detector outputs with contextual evidence retrieved

through online search. Evaluations across four real-world domains show that our agent consistently outperforms standalone detectors, direct LLM classification, and majority voting baselines. Our analysis shows that complementary detectors provide useful initial signals, while online search helps resolve detector disagreement and correct cases where detector consensus is wrong. At the same time, failure analysis reveals that the agent remains limited when external evidence is unavailable, ambiguous, or misleading. Overall, these findings demonstrate that agentic frameworks can improve AI-generated text detection by moving beyond text-only features and incorporating external contextual evidence, thereby helping overcome the limitations of conventional AI detection methods.

Limitations

Dependency on external evidence. Our agent relies on external online information to support its decisions. For texts that cannot be retrieved through online search, such as unpublished, private, or newly generated content, the online search module becomes ineffective, which can lead to performance degradation. In addition, online evidence may contain noise or misleading information, which can negatively influence the agent's judgment. Addressing these issues will require more robust evidence filtering and reasoning mechanisms in future work.

Limited coverage of online AIGT. Although our agent is designed to consider the possibility that retrieved online text may itself be AI-generated, our evaluation does not fully cover this scenario. In practice, online AIGT is still relatively sparse compared with human-written content, and many AI-generated webpages do not provide explicit tags or disclosure. As a result, our dataset doesn't contain the cases where the same text is found online and clearly identified as AI-generated. Future work should evaluate evidence-based detection on more online AIGT examples, especially unlabeled or weakly labeled AI-generated content.

Computational and operational cost. Despite its improved accuracy, the proposed agentic framework incurs higher computational latency and operational costs compared to standalone detectors. Coordinating multiple detectors and performing real-time online searches requires additional computation time and API resources, which may limit scalability in cost-sensitive or real-time applications.

Ethics Considerations

This work aims to improve AI-generated text detection to support content verification and mitigate misinformation. The proposed system leverages publicly available online information and does not intentionally collect or infer sensitive personal data. Detection results should be treated as probabilistic signals rather than definitive judgments, and the system should not be used as the sole basis for high-stakes decisions due to potential noise or misleading external evidence.

References

Amrita Bhattacharjee and Huan Liu. 2023. Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text? *ACM SIGKDD Explorations Newsletter*.

Hong Chen, Hiroya Takamura, and Hideki Nakayama. 2021. SciXGen: A Scientific Paper Dataset for Context-Aware Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*, pages 1483–1492. ACL.

Yize Cheng, Vinu Sankar Sadasivan, Mehrdad Saberi, Shoumik Saha, and Soheil Feizi. 2025. Adversarial Paraphrasing: A Universal Attack for Humanizing AI-Generated Text. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS.

DeepSeek. 2026. DeepSeek. https://deepseek.com/en/.

Federal Register. 2026. https://www.federalregister.gov/.

Google. 2025. Gemini 2.5 flash. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash.

Google. 2026. Gemini 3.1 Pro preview. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview.

Google DeepMind. 2026. https://deepmind.google/technologies/gemini/#introduction.

Xun Guo, Yongxin He, Shan Zhang, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. 2024. DeTeCtive: Detecting AI-generated Text via Multi-Level Contrastive Learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs With Binoculars: Zero-Shot Detection of Machine-Generated Text. In *International Conference on Machine Learning (ICML)*, pages 17519–17537. PMLR.

Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. 2025. PaSa: An LLM Agent for Comprehensive Academic Paper Search. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11663–11679. ACL.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. RADAR: Robust AI-Text Detection via Adversarial Learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS.

Hugging Face. 2025. https://huggingface.co/docs/smolagents/en/index.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. AgentReview: Exploring Peer Review Dynamics with LLM Agents. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1208–1226. ACL.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated Text Detection in the Wild. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 36–53. ACL.

Shengchao Liu, Xiaoming Liu, Yichen Wang, Zehua Cheng, Chengzhengxu Li, Zhaohan Zhang, Yu Lan, and Chao Shen. 2024. Does DetectGPT Fully Utilize Perturbation? Bridging Selective Perturbation to Fine-tuned Contrastive Learning Detector would be Better. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1874–1889. ACL.

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023. ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models. *CoRR abs/2304.07666*.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, and 9 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finnthor. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *International Conference on Machine Learning (ICML)*, pages 24950–24962. PMLR.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1797–1807. ACL.

Samantha Olander. 2025. https://nypost.com/2025/07/05/lifestyle/gen-z-turns-to-ai-for-homework-essays-and-college-apps/.

OpenAI. 2023. GPT-3.5 Turbo. https://platform.openai.com/docs/models/gpt-3-5-turbo.

OpenAI. 2025. GPT-5-mini. https://developers.openai.com/api/docs/models/gpt-5-mini.

OpenAI. 2026a. ChatGPT. https://chat.openai.com/chat.

OpenAI. 2026b. GPT-5.6 Luna. https://developers.openai.com/api/docs/models/gpt-5.6-luna.

OpenAI. 2026c. How can educators respond to students presenting ai-generated content as their own? https://help.openai.com/en/articles/8313351-how-can-educators-respond-to-students-presenting-ai-generated-content-as-their-own.

OpenAI Community. 2019. roberta-base-openai-detector. https://huggingface.co/openai-community/roberta-base-openai-detector.

Project Gutenberg. 2026. Project Gutenberg. https://www.gutenberg.org.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative Agents for Software Development. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15174–15186. ACL.

Qwen. 2025. https://huggingface.co/Qwen/Qwen3-8B.

Hooman H. Rashidi, Brandon D. Fennell, Samer Albahra, Bo Hu, and Tom Gorbett. 2023. The ChatGPT conundrum: Human-generated scientific manuscripts misidentified as AI creations by AI text detection tool. *Journal of Pathology Informatics*.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. 2025. Agent Laboratory: Using LLM Agents as Research Assistants. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*, pages 5977–6043. ACL.

Max Spero. 2025. https://customerthink.com/a-plague-of-ai-generated-reviews-is-coming-to-online-retail-what-does-this-mean-for-the-customer-experience/.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *Findings of the Association for Computational Linguistics: ACL (ACL Findings)*, pages 599–621. ACL.

Edward Tian and Alexander Cui. 2023. GPTZero: Towards Detection of AI-Generated Text Using Zero-Shot and Supervised Methods. https://gptzero.me.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, Qinghua Zhang, Ruifeng Li, Chao Xu, and Yunhe Wang. 2024. Multiscale Positive-Unlabeled Detection of AI-Generated Texts. In *International Conference on Learning Representations (ICLR)*.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, and 7 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS.

Ming-Jie Zhao, Narayanan Unny Edakunni, Adam Craig Pocock, and Gavin Brown. 2013. Beyond Fano's inequality: bounds on the optimal F-score, BER, and cost-sensitive risk and their implications. *Journal of Machine Learning Research*.

A Online Search Module Workflow

We demonstrate the workflow of the online search module in Figure 3.

B AIGT Generation Process

We describe the process used to construct the AI-generated portion of our dataset. For each human-written passage, we prompt an LLM to continue the text while preserving its original topic, genre, register, and writing style. For each domain, we generate 500 AIGT samples using four generator models: `Qwen3-8B`, `GPT-3.5-turbo`, `GPT-5-mini`, and `Gemini-2.5-flash`, with samples evenly distributed across the four models. The prompt used for AIGT generation is shown in Figure 4.

C FPR and FNR Analysis

We further analyze the FPR and FNR of each method across the four datasets. As shown in Table 5, standalone detectors exhibit highly asymmetric error patterns. Binoculars achieves very low FPR across all domains, with an overall FPR of only 1.05%, but suffers from a high overall FNR of 62.85%. This indicates that Binoculars is conservative in predicting AI-generated text: it rarely misclassifies human-written text as AI-generated, but misses many AIGT samples. In contrast, MPU shows the opposite tendency. It achieves a very low overall FNR of 1.90%, but has a much higher overall FPR of 33.20%, suggesting that it is more aggressive in labeling texts as AI-generated and therefore produces more false alarms on human-written samples. By comparison, our agent achieves a more balanced error profile, reducing the overall FPR to 11.15% while keeping the FNR at 4.20%. Although the agent does not achieve the lowest FPR or FNR in isolation, it provides the best trade-off between the two types of errors.

D Detectors

In this study, we evaluate five representative detectors, each employing distinct methodologies to identify AI-generated content. We introduce each detector as follows:

- **Binoculars** (Hans et al., 2024): This detector utilizes a threshold-based approach based on the ratio of two perplexity measurements. Specifically, it calculates the ratio of a text's perplexity to its cross-perplexity, which is derived from a pair of pre-trained LLMs to distinguish between human and machine-generated patterns.
- **MPU** (Tian et al., 2024): This detector is specifically designed for short machine-generated texts. In particular, MPU treats detection as a partial Positive-Unlabeled (PU) learning problem. It incorporates a length-sensitive Multiscale PU Loss and is fine-tuned on a RoBERTa-based backbone to enhance its sensitivity to shorter text fragments.
- **ArguGPT** (Liu et al., 2023): This detector focuses on English essays. It is trained on a dedicated corpus of human and AI-generated argumentative essays and leverages a RoBERTa-based architecture to capture stylistic features prevalent in academic and formal writing.
- **GPTZero** (Tian and Cui, 2023): This detector is a widely adopted commercial AI detector that employs multiple statistical measures to assess text authenticity. It is marketed as a leading solution for high-accuracy detection across diverse domains, particularly in educational and professional settings.
- **RADAR** (Hu et al., 2023): This detector is built upon adversarial training between a paraphraser and a detector. The paraphraser attempts to generate high-quality AI text that evades detection, while the detector learns to identify these attempts. This co-training process makes RADAR particularly robust against sophisticated paraphrasing techniques.

E Agent Prompts

We present the prompts used by the agents in this work. The prompts of the main agent, webpage filtering agent, and webpage analysis agent are shown in Figure 5, Figure 6, Figure 7, respectively.

F Examples of Online Evidence

In this paper, we summarize three types of online evidence that can be retrieved from webpages and present an example for each type in Table 10

G Additional Ablation Studies

We conduct additional ablation studies to examine three aspects of our framework: (1) the choice of LLM backbone, (2) the design of the main agent prompt, and (3) the contribution of external online

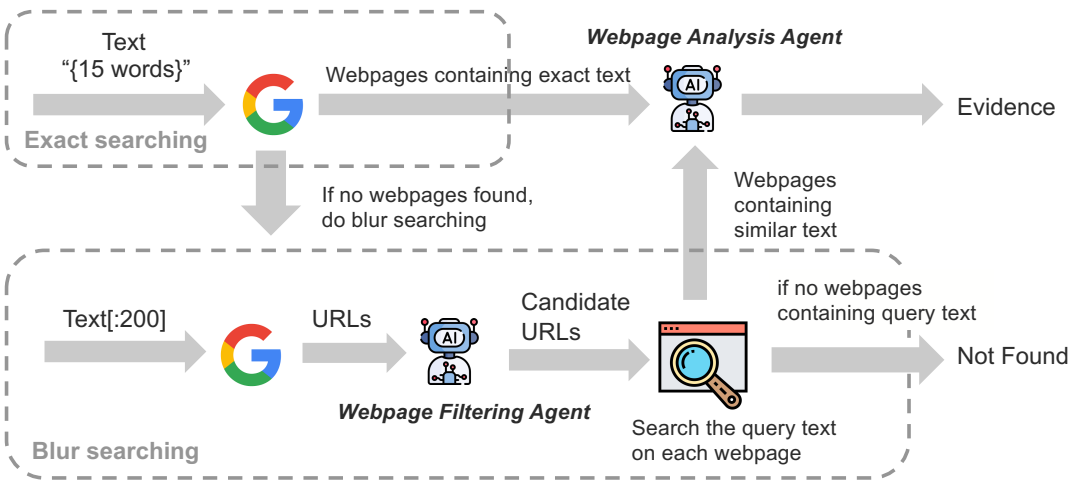

Figure 3: Online search module workflow

AIGT Generation Prompt

You are a continuation-writing engine for dataset generation.
Continue the provided passage in the same language, genre, register, and style.
Do not summarize, explain, quote the instruction, or mention that you are an AI.
Only output the continuation text.
Write complete sentences and end with a complete sentence.
Continue the following passage naturally.
Do not repeat the original text.
End with a complete sentence.
PASSAGE:
`{source_text}`
CONTINUATION:

Figure 4: Prompt used to generate AIGT samples.

Methods	Legal	Literature	Scientific	News	Total
Binoculars	2.00 / 75.00	1.00 / 44.20	0.60 / 73.00	0.60 / 59.20	1.05 / 62.85
MPU	68.80 / 0.80	30.20 / 5.80	26.40 / 0.40	7.40 / 0.60	33.20 / 1.90
GPT-5-mini	60.00 / 7.60	17.60 / 23.20	97.40 / 0.20	61.40 / 8.40	59.10 / 9.85
Majority Vote	43.60 / 6.40	4.20 / 19.40	26.80 / 0.40	3.40 / 8.00	19.50 / 8.55
Agent	**21.60 / 1.00**	**6.80 / 5.60**	**6.00 / 0.60**	**6.40 / 0.80**	**10.20 / 2.00**

Table 5: FPR / FNR (%) of the proposed agent compared with standalone detectors, direct GPT-5-mini classification, and majority vote across four datasets.

evidence. All experiments are conducted on the dataset in Section 4.1.

G.1 Different LLM Backbones

To evaluate whether the proposed prompt design generalizes across different LLM backbones, we replace the original `GPT-5-mini` backbone of the main reasoning agent with `GPT-5.6 Luna` (OpenAI, 2026b) and `Gemini 3.1 Pro Preview` (Google, 2026). The detector outputs, online evidence, prompt, and all other experimental settings are kept fixed.

As shown in Table 6, the agent achieves consistently strong performance across all three backbones. `Gemini 3.1 Pro Preview` obtains the best overall results, reaching 94.33% accuracy and 94.54% F1 with a BER of 5.68%. `GPT-5-mini` achieves 93.90% accuracy, while `GPT-5.6 Luna` achieves 92.40%. The relatively small variation

LLM Backbone	Accuracy	F1	BER
GPT-5-mini	93.90	94.14	6.10
GPT-5.6 Luna	92.40	92.82	7.60
Gemini 3.1 Pro Preview	**94.33**	**94.54**	**5.68**

Table 6: Performance (%) of the proposed agent with different LLM backbones. F1 is computed with AI-generated text as the positive class, and lower BER is better.

across backbones indicates that the effectiveness of the proposed framework is not limited to a particular reasoning LLM.

G.2 Different Prompting Strategies

We compare four prompt settings:

- **S1** is the full agent prompt used in our main experiments. As shown in Figure 5, it includes STEP 1 online-evidence taxonomy, STEP 2 detector-agreement-based decision logic, and STEP 3 structured JSON output constraints.
- **S2** removes the STEP 2 decision logic.
- **S3** removes the STEP 1 online-evidence taxonomy.
- **S4** removes STEP 1 and STEP 2, uses a minimal prompt that only asks the agent to make a final human-versus-AI judgment from the detector results and online-search evidence, without the structured evidence taxonomy or decision rules.

Prompt Setting	Accuracy	F1	BER
S1	**93.90**	**94.14**	**6.10**
S2	89.55	89.03	10.45
S3	93.58	93.85	6.43
S4	80.80	76.70	19.20

Table 7: Agent performance (%) under different prompting strategies. Lower BER is better.

The full prompt S1 achieves the best aggregate performance, with 93.90% accuracy and a BER of 6.10%. Removing the online search verification in S3 causes only a 0.33% reduction in accuracy. In contrast, removing the explicit decision logic in S2 reduces accuracy by 4.35%. The minimal prompt in S4 results in the largest degradation, reducing accuracy by 13.10% and increasing BER to 19.20%.

Further inspection shows that the FNR increases from 2.00% under S1 to 15.20% under S2 and 36.80% under S4. Thus, without explicit decision rules, the agent is substantially more likely to misclassify AI-generated text as human-written. These results indicate that the structured decision logic is the most important tested prompt component, while the explicit evidence taxonomy provides a smaller but still positive contribution.

G.3 Removal of External Evidence

To quantify the contribution of the online search module, we construct an ablation agent with the same detector inputs, full prompt, LLM backbone, and experimental settings as the proposed agent. The only difference is that the online-search result is always set to `Not Found`, preventing the agent from using external evidence.

As shown in Table 8, removing online search decreases overall accuracy from 93.90% to 82.12%, corresponding to a reduction of 11.78%. Overall BER increases by the same amount, from 6.10% to 17.88%. Performance decreases consistently across all four domains, with accuracy reductions of 24.50%, 11.60%, 10.40%, and 0.60% in the Legal, Literature, Scientific, and News domains, respectively.

The degradation primarily results from an increase in false positives on human-written text: the number of false positives increases from 204 with online search to 672 without online search. This result demonstrates that external evidence plays an important role in correcting detector errors, particularly when formal human-written texts are incorrectly flagged as AI-generated.

H Robustness Under Adversarial Paraphrasing

Beyond normal AI-generated text, recent adversarial paraphrasing methods have been specifically designed to rewrite AI-generated content so that it evades existing detectors. We further evaluate the robustness of our agent under adversarial paraphrasing. We apply the adversarial paraphrasing method of Cheng et al. (2025) to all 2,000 AI-generated texts while leaving the 2,000 human-written texts unchanged.

We use `openai_roberta_base` as the guidance classifier during adversarial paraphrasing, and `Meta-Llama-3-8B-Instruct` as the paraphraser. We set the top_p=0.99 and adversarial=1.0. We

Method	Legal	Literature	Scientific	News	Total
w/o Online	64.20 / 73.08 / 35.80	82.20 / 84.22 / 17.80	86.30 / 87.91 / 13.70	95.80 / 95.95 / 4.20	82.12 / 84.55 / 17.88
w/ Online	**88.70 / 89.76 / 11.30**	**93.80 / 93.84 / 6.20**	**96.70 / 96.79 / 3.30**	**96.40 / 96.50 / 3.60**	**93.90 / 94.14 / 6.10**

Table 8: Accuracy / F1 / BER (%) of the agent with and without online search across the four domains. Lower BER is better.

Method	Setting	Accuracy	F1	BER	AI Recall
Agent	Original	**93.90**	**94.14**	**6.10**	98.00
	Adversarial	**94.88**	**95.12**	**5.13**	**99.95**
Binoculars	Original	68.05	53.76	31.95	37.15
	Adversarial	52.40	10.94	47.60	5.85
MPU	Original	82.45	84.82	17.55	**98.10**
	Adversarial	83.38	85.74	16.63	**99.95**

Table 9: Performance (%) of the proposed agent and standalone detectors on the original and adversarial paraphrasing settings. AI Recall is the recall for the AI-generated class. Lower BER is better.

compare two settings:

- **Original**: the original 2,000 human-written and 2,000 AI-generated texts used in the main experiments.
- **Adversarial**: the same human-written texts combined with adversarially paraphrased versions of the same 2,000 AI-generated texts.

As shown in Table 9, adversarial paraphrasing substantially degrades Binocular. Its AI recall decreases from 37.15% to 5.85%, while its overall accuracy decreases from 68.05% to 52.40%. In contrast, MPU and the proposed agent do not exhibit performance degradation under this attack. The agent correctly identifies 99.95% of the adversarial AI-generated texts and achieves an overall accuracy of 94.88%, compared with 93.90% on the original evaluation set.

These results show that the proposed agent remains robust to the evaluated adversarial paraphrasing attack and substantially outperforms the standalone detectors in overall accuracy.

Main Agent

You are an expert AI-generated text detector. Detector results and online search evidence have already been collected and are provided below. Your task is to reason over this evidence and produce a final judgment.
Pre-computed Detector Results: `{detector_results}`
Online Search Result: `{online_search_result}`
—
STEP 1: ONLINE SEARCH VERIFICATION
Classify the provided online search result:
A. STRONG HUMAN EVIDENCE (weight: 90%):
- Text found on verified platforms, which could include government sites, news sites, academic journals, personal blogs, even if published after 2023, this should be a strong signal of human authorship.
- Text publication date BEFORE November 2022 (pre-ChatGPT era)
- Exact match with verifiable human author attribution
B. WEAK HUMAN EVIDENCE (weight: 60%):
- Partial match, or found on social media platforms with unclear or post-2023 date
C. AI PLATFORM EVIDENCE (weight: 90% towards AI):
- Text found on AI content platforms, ChatGPT forums, or AI writing showcases
- Text explicitly labeled as AI-generated in search results
D. NO MATCH FOUND:
- Inconclusive
- does NOT indicate AI-generated; fall back to detector consensus
STEP 2: DECISION LOGIC
First evaluate detector agreement from the pre-computed results above:
- Strong Agreement (all same) → high confidence baseline
- Partial Agreement (majority) → medium confidence baseline
- Conflict (split) → low confidence; online evidence becomes critical
Case A — All detectors agree:
Strong Human Evidence → Human (0), High confidence
AI Platform Evidence → AI (1), High confidence
Weak / No Evidence → follow detector consensus, Medium confidence
Case B — Majority agree:
- Strong Human Evidence → Human (0), High confidence
- AI Platform Evidence → AI (1), High confidence
- Weak Evidence → follow majority (weight 70%), Medium confidence
- No Evidence → follow majority (weight 65%), Low-Medium confidence
Case C — Detectors conflict:
- Strong Evidence → follow online evidence (weight 95%), Medium-High confidence
- Weak Evidence → weigh evidence type and detector confidence, Low-Medium confidence
- No Evidence → do NOT default to Human. Use detector-specific reliability:
If YHTDetector=1 and Bino=0, lean AI-generated (1) with low confidence unless there is strong human evidence.
If Bino=1 and YHTDetector=0, lean AI-generated (1) with low confidence unless there is strong human evidence.
If no detector predicts AI, lean Human (0).
STEP 3: OUTPUT FORMAT
Respond with valid JSON only: `{{ "judgment": <0 or 1>, "confidence": <"high" | "medium" | "low">, "online_search_result": <"Strong Human Evidence" | "Weak Human Evidence" | "AI Platform Evidence" | "Not Found">, {det_results}, "reasoning": "1. Detector Analysis: [...] 2. Online Evidence: [...] 3. Decision Logic: [...] 4. Confidence Justification: [...]" }}`
Critical Reminders:
- "Not Found" online ≠ AI-generated. Many humans write original content with no web presence.
- However, "Not Found" also must not override an AI-positive detector.
- Publication dates before November 2022 are strong human indicators.
- Be explicit about which case (A/B/C) applies.
—
Text to Analyze: `{text}`

Figure 5: Prompt of main agent

Webpage Filtering Agent

Your task is to pick 3 most relevant webpages from the following search results based on their relevance to the text.
You need to follow these steps:
1. Analyze the text category, it may be an abstract of a scientific paper, a news article, a blog post, an advertisement, a novel, a review, etc. You could choose one category from but not limited to these examples.
2. Then, pick three webpages' URLs which you think are most related with the type of the text you identified in step 1, for example, if you think the text is an abstract of a scientific paper, you should pick URLs that contain scientific papers.
3. After that, you could return the final result.
You need to return the result in the following json format:

```
 {"Text category":<>,
"Webpages": [{"link":<>, "reason":<>},
{"link":<>, "reason":<>},
{"link":<>, "reason":<>}]
}
```

Here is the text:

```
[text]
```

Here are the search results:

```
{results_text}
```

Figure 6: Prompt of webpage filtering agent

Webpage Analysis Agent

Your task: Determine if the given text is AI-generated or human-written based on webpage evidence.
CRITICAL RULE: Time is the strongest evidence
  If last modified date is BEFORE November 2022: DEFINITELY human-written
  If date is 2023 or later: Requires further analysis
—
Step 1: Gather Evidence
Use available tools to collect:
1. Webpage content (use `visit_webpage`)
2. Last modified date (use `get_last_modified`)
3. Platform context and metadata
—
Step 2: Evaluate Evidence (in priority order)
Priority 1: Temporal Evidence (Highest Weight)
  Last modified date < Nov 2022 → Human-written (STRONG)
  Explicitly mentions "written in 2020" → Human-written (STRONG)
Priority 2: Explicit Statements (High Weight)
  "AI-generated" / "Created with ChatGPT" → AI-generated (STRONG)
  "Written by [human author]" with bio → Human-written (MEDIUM)
  Platform tags: #AI, #ChatGPT → AI-generated (MEDIUM)
Priority 3: Platform Context (Medium Weight)
  Known AI-content platforms (weight: MEDIUM):
    Character.AI, ChatGPT playground, AI writing communities
    If hosted on these → Likely AI-generated
  Traditional human-content platforms (weight: LOW):
    Personal blogs, academic sites, news outlets (pre-2023)
    If on these → Slightly lean human-written
Priority 4: Author Behavior (Low Weight)
  Personal anecdotes, photos, creative process → Human-written (WEAK)
  High posting frequency + generic content → AI-generated (WEAK)
—
Step 3: Make Final Judgment
Use this decision tree:
1. If temporal evidence points to pre-2023 → HUMAN-WRITTEN
2. Else if explicit AI statement → AI-GENERATED
3. Else if explicit human authorship + evidence → HUMAN-WRITTEN
4. Else if on AI-content platform + no counter-evidence → LIKELY AI-GENERATED
5. Else → UNCERTAIN
—
Output Format
Provide:
1. Platform Type: `[AI-content / Traditional / Other]`
2. Last Modified Date: `[Date or "Not found"]`
3. Explicit Authorship Statement: `[Yes/No and what it says]`
4. Supporting Evidence: `[List 2-3 key pieces]`
5. Final Judgment: `[Human-written / AI-generated / Uncertain]`
6. Confidence Level: `[High / Medium / Low]`
7. Reasoning: `[2-3 sentence justification based on priority order]`
If page is inaccessible: Return "No relevant information found online."
—
URLs to analyze: `{url_text}`

Figure 7: Prompt of webpage analysis agent

Type	Example of agent reasoning
Human Authorship	1. Detector Analysis: Both pre-computed detectors return 0 (Bino=0, YHTDetector=0), so there is strong agreement among detectors (Case A). 2. Online Evidence: The Stanford publications page is a traditional academic personal webpage with explicit authorship (Ramesh Johari) and bibliographic entries spanning many years, which is classified as Strong Human Evidence. The LinkedIn page was inaccessible and provides no usable content; this does not override the Stanford evidence. 3. Decision Logic: Under Case A (all detectors agree) with Strong Human Evidence, the rules call for a Human (0) judgment with high confidence. 4. Confidence Justification: Detector consensus plus a verifiable academic source on a university domain with explicit author attribution provides robust support for human authorship. No detector indicates AI, and no AI-platform evidence was found. Therefore the final judgment is Human (0) with high confidence.
Platform	1. Detector Analysis: The detectors conflict (Bino=0, YHTDetector=1), which places this in Case C (Detectors conflict). 2. Online Evidence: Google search returned multiple government and regulations websites (federalregister.gov, regulations.gov) containing PRA/notice language. These are verified platforms and constitute Strong Human Evidence despite a downstream webpage analysis error. 3. Decision Logic: Under Case C with Strong Human Evidence, follow the online evidence (human-authored), so the judgment is Human (0). 4. Confidence Justification: The presence of authoritative government sources strongly supports human authorship; because detectors conflict, confidence is set to medium (per guidance: Medium-High for Case C with strong evidence, mapped to 'medium').
Publication Time	1. Detector Analysis: The two detectors conflict (Bino=0, YHTDetector=1), so this is Case C (Detectors conflict). 2. Online Evidence: The provided search results include direct matches to an academic paper titled 'BEAR: Sketching BFGS Algorithm for Ultra-High Dimensional ...' on arXiv (snippet shows date 26 Oct 2020) and a proceedings.mlr.press page. These are verified, human-authored publications dated before November 2022, constituting Strong Human Evidence. 3. Decision Logic: For conflicting detectors with strong human evidence, the guideline says to follow the online evidence (weight 95%). Therefore we judge the text as Human-written (0). 4. Confidence Justification: Because the text matches a peer-reviewed/arXiv paper from 2020 (pre-ChatGPT era) the match is a strong human indicator; this outweighs the single AI-positive detector. Hence confidence is set to 'high'.

Table 10: Examples of agent reasoning for three types of online evidence.