Beyond Checkmate:
Exploring the Creative Choke Points for AI Generated Texts

Nafis Irtiza Tripto[1] **Saranya Venkatraman**[2] **Mahjabin Nahar**[1] **Dongwon Lee**[1]
[1] The Pennsylvania State University [2] Amazon
[1]{nit5154, mfn5333, dongwon}@psu.edu [2] saranvn@amazon.com

Abstract

The rapid advancement of Large Language Models (LLMs) has revolutionized text generation but also raised concerns about potential misuse, making detecting LLM-generated text (AI text) increasingly essential. While prior work has focused on identifying AI text and effectively *checkmating* it, our study investigates a less-explored territory: portraying the nuanced distinctions between human and AI texts across text segments (introduction, body, and conclusion). Whether LLMs excel or falter in incorporating linguistic ingenuity across text segments, the results will critically inform their viability and boundaries as effective creative assistants to humans. Through an analogy with the structure of chess games, comprising opening, middle, and end games, we analyze segment-specific patterns to reveal where the most striking differences lie. Although AI texts closely resemble human writing in the body segment due to its length, deeper analysis shows a higher divergence in features dependent on the continuous flow of language, making it the most informative segment for detection. Additionally, human texts exhibit greater stylistic variation across segments, offering a new lens for distinguishing them from AI. Overall, our findings provide fresh insights into human-AI text differences and pave the way for more effective and interpretable detection strategies. Codes available at https://github.com/tripto03/chess_inspired_human_ai_text_distinction.

1 Introduction

When Garry Kasparov, then world chess champion, lost to IBM's Deep Blue, a chess-playing supercomputer, in 1997 (Pandolfini, 1997), it marked a turning point in AI history, the moment machines overtook humans in a game long considered a symbol of strategic mastery. A similar shift occurred with the public debut of ChatGPT in late 2022, as Large Language Models (LLMs) captured global attention and began reshaping the landscape of communication, creativity, and cognition. With models like *GPT-4* passing professional exams (Katz et al., 2024) and even approaching Turing test benchmarks (Jones and Bergen, 2025), these advancements raise critical questions about distinctiveness of human intellect. Interestingly, AI chess engines and LLMs share a remarkable similarity. While chess engines determine the best move from a given board state, LLMs predict the next token based on preceding text. This shared mechanism of context-driven prediction has even led to the development of transformer-based chess engines capable of achieving Grandmaster-level performance (Ruoss et al., 2024).

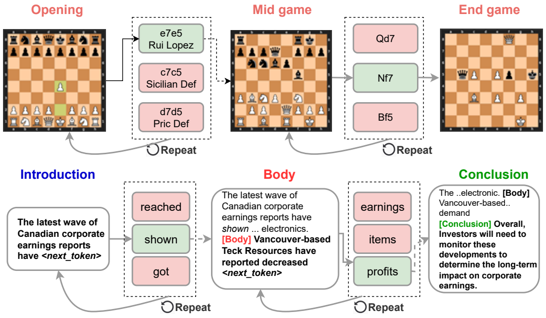
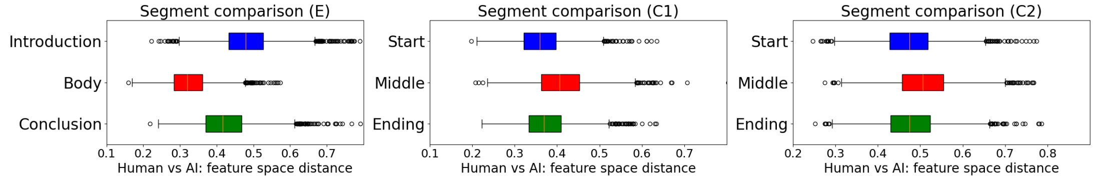

Opening
e7e5
Rui Lopez
c7c5
Sicilian Def
d7d5
Pric Def
Repeat
Mid game
Qd7
Nf7
Bf5
Repeat
End game
Introduction
The latest wave of Canadian corporate earnings reports have <next_token>
reached
shown
got
Repeat
Body
The latest wave of Canadian corporate earnings reports have shown ... electronics. [Body] Vancouver-based Teck Resources have reported decreased <next_token>
earnings
items
profits
Repeat
Conclusion
The ..electronic. [Body] Vancouver-based.. demand [Conclusion] Overall, Investors will need to monitor these developments to determine the long-term impact on corporate earnings.

Figure 1: An illustration of the resemblance between chess and AI text generation. In chess, players select the optimal move from valid options given a board state; in text generation, LLMs similarly choose the next word/-token from the vocabulary based on context. Both processes can be divided into three distinct segments, each serving a specific role in shaping the outcome of the game or the meaning of the text.

Inspired by this transformation, we revisit the metaphor of chess to investigate a new frontier: understanding how human and AI-generated texts differ across *segments*. In both chess and writing, structure matters. A chess match progresses through the opening, middlegame, and endgame, each demanding different levels of strategic reason-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11953–11970
November 4-9, 2025 ©2025 Association for Computational Linguistics



(a) (b) (c)

Figure 2: Segment comparison results using LIWC (Boyd et al., 2022) and WritePrint (Abbasi and Chen, 2008) features: **(a)** In the original setting *(E)*, the body segment shows less difference between human and AI texts, likely due to its greater length. Under length-controlled conditions: **(b)** *C1* (equal segmentation) and **(c)** *C2* (body matched to introduction/conclusion length), the body/middle segment exhibits the highest divergence.

ing. Likewise, written texts often follow a tripartite structure: an introduction to set the stage, a body to deliver core arguments, and a conclusion to synthesize insights. Chess opening and endgame moves are often heavily studied, analyzed, and codified into established theories for AI chess engines, like IBM DeepBlue (Campbell et al., 2002) or StockFish (Romstad et al., 2008). However, it is the dynamic middlegame where the true mastery of players is put to the test (Znosko-Borovski, 1922). As Brian Christian (Christian, 2011) explores in his book "*The Most Human Human*", the middlegame represents the crucible where creativity, strategy, and adaptability separate humans from AI.

Just as in the middlegame of chess, one critical question arises: can LLMs move beyond following the typical opening and ending from their training data to navigate the fluid "middlegame" of text generation with the same linguistic ingenuity as humans? While recent studies have made substantial progress in distinguishing LLM-generated (AI text) from human-written text using stylometric features (Muñoz-Ortiz et al., 2024; Rosenfeld and Lazebnik, 2024; Guo et al., 2024; Reinhart et al., 2025), thus *checkmating* them, they often overlook the *structural* context of the text. Do different text segments contribute differently to AI detection? And more importantly, do humans and LLMs exhibit similar patterns of stylistic variation across these segments? The answer has important implications, as limitations in this area could hinder their effectiveness in creative domains, while success would reinforce their role as versatile writing assistants.

Therefore, in this paper, we explore these questions through a comprehensive computational analysis of human and AI texts, focusing on three domains, news articles, essays, and emails, all of which naturally follow a structured format (Henry and Roseberry, 1997; Medvid and Podolkova, 2019; Matruglio, 2020). Our dataset includes both human texts and generations from four prominent LLMs: ChatGPT (*GPT-3.5*), PaLM (*text-bison-001*), LLaMA2 (*llama2-chat-7b*), and Mistral (*mistral_7b*). We introduce two core analyses:

1. **Segment Comparison:** Do differences between human and AI texts vary across segments?
2. **Source Comparison:** Do internal stylistic variation across segments differ between humans and AI texts ?

Our findings are both surprising and insightful. While body segments initially appear more similar between human and AI texts (Figure 2), this is largely due to their greater length (Révész, 2014). In length-controlled settings, the body (or middle) consistently reveals the most significant differences. Moreover, it plays a dominant role in AI text detection. We also find that humans exhibit more variation across text segments than LLMs, reinforcing that LLMs tend to maintain a consistent stylistic fingerprint throughout. To further ground our analogy, we also analyze over 166K chess games to examine how human and AI players differ across game phases, showing that divergence peaks in the middlegame, the creative core of a match. Overall, our research sheds new light on the nuanced distinctions between human and AI text, offering a compelling step toward understanding the subtle yet defining elements that make human writing authentically human.

2 Related Works

Stylometry difference between human and AI text Stylometry features have long been effective in text classification and authorship analysis tasks, and can be proxies for creative *chokepoints* in text (Neal et al., 2017). With the growing availability of LLM-generated text datasets (Dugan et al.,

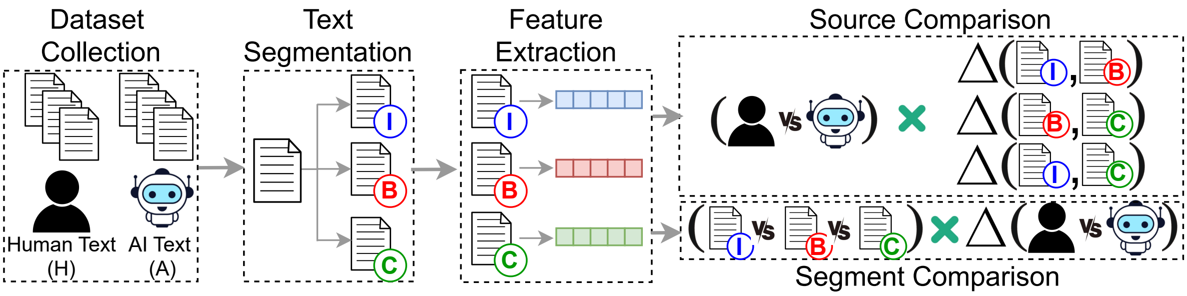


Figure 3: Overview of our methodology. Given a dataset of parallel human and AI texts, we divide each document into three segments and extract a comprehensive set of features from each segment. We perform statistical significance tests for **segment** and **source** comparisons for each feature, considering all possible combinations.

2024; Tripto et al., 2023; Verma et al., 2024), recent research has applied these features to distinguish between human and AI text. For example, AI texts often differ from human writing in vocabulary diversity (Muñoz-Ortiz et al., 2024), distinctive word choices (Berriche and Larabi-Marie-Sainte, 2024), formality (Al Hosni, 2024), and rhetorical styles (Reinhart et al., 2025). Therefore, several studies have leveraged linguistic features for AI text detection (Casal and Kessler, 2023; Guo et al., 2024; Rosenfeld and Lazebnik, 2024), citing their explainability (Muñoz-Ortiz et al., 2024) and strong statistical performance (Herbold et al., 2023). Despite these relevant studies, LLMs become increasingly adept at mimicking human writing styles, and their difference is narrowing (Toshevska and Gievska, 2025).

AI text detection With the rapid advancement of LLMs, interest in detecting AI-generated text has surged across domains. Beyond stylometry-based methods, current detection approaches include fine-tuned models like the RoBERTa-based OpenAI Detector (Solaiman et al., 2019), GROVER (Zellers et al., 2019), MAGE (Li et al., 2024), RADAR (Hu et al., 2023), and LLM-DetectAIve (Abassy et al., 2024), which use supervised learning on binary classification tasks (human vs. AI). In contrast, statistical and zero-shot detectors, such as DetectGPT (Mitchell et al., 2023), DetectLLM (Su et al., 2023), GPT-who (Venkatraman et al., 2023), and Binoculars (Hans et al., 2024a) leverage distributional differences, often via perplexity, to offer more robust cross-domain performance. Commercial tools like GPTZero (Tian, 2023), Originality.ai [1], and Turnitin's AI detector [2] also provide user-facing solutions. While many of these methods highlight important tokens for interpretability, they generally overlook which text segments contribute most to detection. By analyzing how different linguistic differences vary across text segments, our study offers a novel and necessary extension to the current literature, advancing the theoretical understanding and practical methodologies for AI text detection.

3 Methodology

Motivated by the chess middlegame analogy, we examine how human and AI texts differ across different segments. Figure 3 presents an overview of our methodological framework.

3.1 Dataset creation

We compile datasets from three domains (news articles, emails, and essays), each containing human-authored texts paired with corresponding LLM-generated versions. Our study includes four LLMs: ChatGPT (*gpt-3.5-turbo*) from OpenAI, PaLM (*text-bison-001*) from Google, LLaMA2 (*llama2-chat-7b*) from Meta, and *Mistral_7b* from Mistral AI, representing both open-source and proprietary models. For the essay domain, we use the *Persuade* corpus (Crossley et al., 2022), consisting of argumentative essays written by US students (grades 6-12) across different prompts. Our dataset includes approximately *1700* human samples and corresponding LLM generations from the Kaggle competition (King et al., 2023). For news, we employ the Ghostbuster dataset (Verma et al., 2024), which contains *Reuters* articles and existing LLM generations (we generate missing samples using identical prompts).

For the email domain, we draw on a curated subset of the Enron corpus (Klimt and Yang, 2004),

[1] https://originality.ai/ai-checker
[2] https://www.turnitin.com/campaigns/clarity//

which still remains one of the most widely used and publicly available resources for authorship and stylistic analysis (Tyo et al., 2022; Nini et al., 2024). Contemporary email datasets are scarce and often unsuitable, as many focus on spam detection, security leaks (e.g., Clinton emails (Shane and Schmidt, 2015)), or bulk announcements (e.g., DBWorld (Filannino, 2011)), rather than personal correspondence. To ensure suitability, we perform extensive preprocessing: filtering for one-to-one internal emails to retain a personal tone, excluding automated, forwarded, or bulk messages, and removing emails with attachments. We also discard emails that are too short or excessively long, selecting only users with at least ten messages to ensure sufficient representation. This process yields a clean subset balancing realism with consistency. Using this data, we prompt LLMs to generate responses based on the original email's header, sender/recipient information, and a concise content summary, with full prompt details provided in the Appendix A. Table 1 summarizes key statistics across domains.

Dataset (Domain)	**Source**	**# texts**	**Avg. # words**	**Avg. # sentences**	**I-B-C ratio(%)**
Reuter (News)	Human	989	310.90	10.98	13-67-20
	AI	4741	288.05	10.87	15-57-28
Enron (Email)	Human	1632	173.34	8.78	17-70-13
	AI	6289	144.61	8.63	17-63-20
Persuade (Essays)	Human	1717	269.93	13.58	18-60-22
	AI	3788	280.38	13.71	18-56-26

Table 1: Key characteristics of human and AI texts across domains. Word and sentence counts per document are comparable between the two. I-B-C denotes the ratio of introduction (I), body (B), and conclusion (C) lengths in the original setting (Setting *E*), with the body segment consistently longer than the others.

3.2 Text segmentation

Segmenting text into introduction, body, and conclusion is inherently subjective (Hearst, 1994; Aumiller et al., 2021), as these sections often lack clear boundaries and vary significantly across writing styles, domains, and contexts. Manually annotating a large dataset would be prohibitively expensive and time-consuming. However, recent advances in LLMs have demonstrated strong performance in natural language understanding tasks, often achieving human-level performance (Thapa et al., 2023; Michelmann et al., 2025; Sun et al., 2024). Therefore, we employ *gemini-1.5-flash* (excluded from our authorship analysis to mitigate bias) to segment texts in our original setting (*E*).

Dataset/Source	**S**	**Judgement criteria**	**S**
Persuade	0.96	Gemini vs GPT4	0.93
Enron	0.90	Gemini vs Human 1	0.91
Reuter	0.87	Gemini vs Human 2	0.92
Human	0.87	GPT4 vs Human 1	0.92
ChatGPT	0.91	GPT4 vs Human 2	0.91
PaLM	0.93	Human 1 vs Human 2	0.94
Llama-2	0.93	Gemini vs Finetuned BERT	0.92
Mistral	0.96		

Table 2: Segmentation Similarity Score (S) across datasets, LLMs, and criteria. Scores are higher for essays and emails with clearer structure, but lower for news. AI texts are easier to segment than human texts. The similarity scores across humans, LLMs, human–LLM pairs, and LLM–computational methods are nearly identical, with no statistically significant differences.

Since body segments are typically longer (Henry and Roseberry, 1997; Raharjo and Nirmala, 2016), we also explore length-controlled segmentation: in *C1*, dividing texts into three equal parts, and in *C2*, sampling a body portion matching the average length of the introduction and conclusion. In all settings, we ensure that the segments contain complete sentences to preserve semantic coherence and readability (Van Dijk, 1980; Graesser, 2003).

Given the subjective nature of text segmentation, we show that our LLM-based approach is robust and well-aligned with alternative methods. We use the Segmentation Similarity Score (Fournier and Inkpen, 2012) (0 to 1, where 1 indicates identical segmentation) to evaluate text segmentation based on sentence counts. To validate our method, we segment a subset of 300 samples across all domains. Two human annotators (authors of this paper) provide manual segmentations to assess alignment with human perception, and we use *GPT-4* to evaluate consistency between LLMs. Additionally, we fine-tune a BERT model on the human-segmented data to compare with standard computational techniques. As shown in Table 2, all comparisons yield segmentation similarity scores above 90%, with no statistically significant differences ($\alpha = 0.05$) among human-human, LLM-LLM, and LLM-human pairings. These results confirm that our LLM-based method, though not exact, reliably captures the structure of segmented text.

3.3 Feature extractions

We extract traditional stylometric feature sets such as LIWC (Linguistic Inquiry and Word Count), which provides psycholinguistic characteristics

Criteria	Description
White vs Black	Human as white (53.08%), AI as white (46.92%)
AI win % as white	Win (71.36%), Draw (5.29%), Loss (23.35%)
AI win % as black	Win (67.23%), Draw (4.79%), Loss (27.98%)
Elo ratings	Human (1503-2433), AI (1557-2761)
Game category	Blitz (29.71%), Lighting (29.29%), Standard (41%)
Move category	Opening (28.31%), Middle (29.23%), End (42.46%)
Top 4 ECO codes	A00(4.54%), A45(4.09%), D00(3.23%), C50(2.37%)

Table 3: Overview of the chess games analyzed in our study. The AI players generally have higher Elo ratings and win percentages compared to their human counterparts in the dataset.

(Boyd et al., 2022), and Writeprint features, which capture an author's distinctive stylometric patterns (Abbasi and Chen, 2008). Additionally, we examine how specific features vary across different segments and sources. Therefore, we include several individual lexical (vocabulary richness, readability), syntactic (part-of-speech tags, named entity tags, stopwords distributions) opinion (formality, sentiment, subjectivity), contextual (text embedding), and text perplexity-related features, offering a comprehensive analysis of the text's stylistic and structural attributes (details in Appendix D).

To use these features for **segment** and **source** comparison using statistical significance tests, we first define a difference measure, denoted as Δ, between two feature values. Features are categorized as either scalar (e.g., vocabulary richness, readability, sentiment score) or distributional (e.g., POS-tag, stopwords, and LIWC distributions). For scalar features, we use absolute difference. For distributional features, we apply Jensen–Shannon Divergence (JSD) (Lin, 1991), a symmetric, bounded metric well-suited for comparing discrete probability distributions (Endres and Schindelin, 2003). For vector-based features not summing to one, such as perplexity scores and contextual embeddings, we use correlation distance and cosine distance, respectively. These capture relational and angular differences, making them appropriate for high-dimensional comparisons (Ruppert, 2004; Huang et al., 2008; Turney and Pantel, 2010).

3.4 Statistical significance test

As we are interested in understanding how linguistic features differ between human and AI texts across textual segments (Introduction, Body, Conclusion), we design statistical significance tests with feature values extracted from each segment. Specifically, We conduct separate statistical tests for each linguistic feature. Given two text sources **(Sources, H: Human, A: AI)** and three segments from each text **(Segments, I: Introduction, B: Body, C: Conclusion)**, we define Z_x as an individual feature from segment x for source Z.

For **source comparison** tests, we consider pairwise segments, $x, y \in \{I, B, C\}$, compute their differences for human and AI texts, $\Delta(H_x, H_y)$ and $\Delta(A_x, A_y)$, respectively. We evaluate whether human cross-segment differences $\Delta(H_x, H_y)$ are statistically greater than (>), less than (<), or comparable ($\sim$) to AI cross-segment differences $\Delta(A_x, A_y)$, for specific pair of segments. Similarly, for **segment comparison**, we compute the difference between human and AI texts for all three segments, $\Delta(H_I, A_I)$, $\Delta(H_B, A_B)$, and $\Delta(H_C, A_C)$ to determine whether human-AI differences are statistically similar across segments. Details of the tests are mentioned in the Appendix B.

3.5 Chess dataset creation

Since our study was motivated by the chess middlegame analogy, we conduct a concise yet systematic analysis of chess games to computationally explore whether these differences vary by phases. Using games from the Free Internet Chess Server (FICS) database[3], we compile a dataset of ranked human vs AI games played between 2018 and 2020, selected due to the rise of AlphaZero (Silver et al., 2018) and the emergence of open-source AI chess bots (McIlroy-Young et al., 2020). We include only games between 30 and 100 moves, excluding short (due to early blunders or resignations) or excessively long games (repetitive moves). Table 3 summarizes the final dataset of 166,738 games. We then segment each game into opening, middlegame, and endgame phases and extract features from chess moves in each segment (see Appendix C for details).

4 Results

We present our findings on **segment** and **source** comparisons across different experimental settings, identify which text segment contributes most to AI text detection, and explore whether similar segmental differences exist between human and AI chess players.

4.1 Segment and source comparison results

We conduct a comprehensive analysis of individual features across all possible combinations to evaluate both **segment** and **source** comparisons, with

[3] https://www.ficsgames.org/download.html

Feature	Dataset	Source comparison			Segment comparison
		Δ(I, B)	Δ(I, C)	Δ(B, C)	Δ(H,A)
Vocabulary Richness	Reuter	H>A	∼	H>A	B>C>I
	Enron	∼	H>A		∼ († ‡)
	Persuade	∼	H>A		B>C>I
Readability Score	Reuter	H>A			C>I>B † ‡
	Enron	A>H	H>A		I>C>B †
	Persuade	∼			C>I>B † ‡
Sentiment Score	Reuter	∼			I∼C>B ◇
	Enron	A>H			C>B>I ‡
	Persuade	∼			I>C>B ◇
Formality Score & Content Similarity (same results)	Reuter	H>A			I∼C>B † ‡
	Enron	H>A			I∼C>B †
	Persuade	H>A			I∼C>B † ‡
Perplexity Scores	Reuter	∼			B>I>C
	Enron	H>A	A>H	H>A	C>I>B ◇
	Persuade	∼			B>I∼C
Parts of Speech Tags Distribution	Reuter	H>A			I>C>B ◇
	Enron	H>A			I∼C>B ◇
	Persuade	H>A			I>C>B ◇
Named Entity Tags Distribution	Reuter	∼	H>A		C>I>B ◇
	Enron	∼	H>A	∼	∼
	Persuade	∼	H>A		∼

Table 4: Statistical significance test results in the original experimental setting *(E)*. **Source Comparison:** Δ(I, B) represents the difference in a given feature between the Introduction (I) and Body (B) for both human and AI texts. Violet (H > A) indicates that this difference is significantly greater in human texts, while orange (A > H) denotes the opposite and (∼) indicates no statistically significant difference. **Segment Comparison:** Δ(H, A) captures the feature difference between human and AI texts within a specific segment (I, B, or C). Green highlights cases where the body segment shows a significantly greater difference than the introduction or conclusion, while red marks the opposite. (∼) indicates no significant difference across segments. The symbols (†) and (‡) denote cases where the body segment shows higher differences in the length-controlled settings *C1* and *C2*, respectively. The ◇ symbol indicates no significant segmental difference in both *C1* and *C2*. Cells without symbols represent cases where the original setting (E) aligns with both length-controlled settings.

key findings summarized in Table 4. In the original experimental setting (*E*), **segment comparison** reveals that the body segment exhibits less distinction between human and AI texts compared to the introduction and conclusion. However, this lower contrast is due to the body's greater length, which can dilute syntactic features like POS-tag or named entity distributions and flatten opinion-based features such as sentiment or formality through averaging. The extended length also allows AI text to align more easily with human content in the body segment. Nevertheless, length-independent features like vocabulary richness and perplexity indicated higher differences in the body.

In the length-controlled experiments (*C1* and *C2*) settings, stylometric (e.g., LIWC, Writeprints) and linguistic features (e.g., vocabulary richness, readability, sentiment) show higher differences in the body/middle segment. When segment lengths are normalized, several features show no statistically significant differences across segments. Given that the body segment typically hosts the core arguments, elaboration, and creativity in writing (Medvid and Podolkova, 2019), our findings suggest that while LLMs may mimic surface-level structure, they struggle to replicate the nuanced, adaptive strategies humans employ in this more demanding segment, as validated through human vs. AI text detection in the following subsection.

In the **source comparison**, our findings show human texts exhibit higher cross-segment variation than AI text, offering an innovative lens to differentiate between the two. While prior studies (Guo et al., 2023; Muñoz-Ortiz et al., 2024) have shown that AI texts tend to be more structurally consistent and formal, our analysis uncovers how this consistency manifests across segments. LLMs inherently prefer structured text generation, often incorporating a distinct introduction, body, and conclusion boundary, leading to smoother transitions and uniform distribution of content, named entities, and POS tags across segments. In contrast, human writers tend to modulate their linguistic fingerprints between segments, a trait not yet replicated by AI. Additional analysis on individual LLM behavior can be found in the Appendix E.

4.2 Checkmating AI text: which segment reveals its origins?

To explore how different text segments contribute to AI text detection, we evaluate a suite of prominent detectors: GPT-Zero (Tian, 2023), MAGE (Li et al., 2024), Radar (Hu et al., 2023), Binocular (Hans et al., 2024b), GPT-Who (Venkatraman et al., 2024), and a fine-tuned BERT classifier (detailed in the Appendix F). Our primary goal is to

Dataset	Criteria	GPT Zero	MAGE	RADAR	Binoculars	GPT-Who	Finetuned Bert
Reuter (News)	**Total text**	0.84	0.75	0.77	0.91	0.82	0.96
	Voting	0.83 (↓1.19%)	0.51 (↓45.74%)	0.72 (↓6.49%)	0.85 (↓6.59%)	0.82 (↓1.2%)	0.97 (↑1.04%)
	Body only	0.85 (↑1.19%)	0.76 (↑1.33%)	0.81 (↑5.19%)	0.84 (↓7.69%)	0.84 (↑2.44%)	0.94 (↓2.08%)
	Intro+conclusion	0.76 (↓9.52%)	0.62 (↓17.33%)	0.77 (↓0%)	0.77 (↓15.38%)	0.79 (↓3.66%)	0.93 (↓3.12%)
Enron (Emails)	**Total text**	0.62	0.78	0.82	0.73	0.77	0.98
	Voting	0.61 (↓1.61%)	0.78 (↑0.0%)	0.73 (↓10.98%)	0.71 (↓2.74%)	0.85 (↑10.39%)	0.96 (↓2.04%)
	Body only	0.71 (↑14.52%)	0.72 (↓7.69%)	0.79 (↓3.36%)	0.74 (↑1.37%)	0.78 (↑1.3%)	0.93 (↓5.1%)
	Intro+conclusion	0.55 (↓11.29%)	0.7 (↓10.26%)	0.74 (↓9.76%)	0.68 (↓6.85%)	0.75 (↓2.6%)	0.96 (↓2.04%)
Persuade (Essay)	**Total text**	0.94	0.94	0.79	0.82	0.83	0.99
	Voting	0.9 (↓4.26%)	0.75 (↓20.21%)	0.64 (↓18.99%)	0.86 (↑4.88%)	0.8 (↓3.61%)	0.97 (↓2.02%)
	Body only	0.88 (↓6.38%)	0.82 (↓12.77%)	0.67 (↓15.19%)	0.84 (↑2.44%)	0.82 (↓1.2%)	0.96 (↓3.03%)
	Intro+conclusion	0.89 (↓5.32%)	0.73 (↓22.34%)	0.61 (↓22.78%)	0.78 (↓4.88%)	0.75 (↓9.64%)	0.96 (↓3.03%)

Table 5: AI text detection results (original setting *E*). Each cell value represents the F1 score of various detection methods, with higher scores indicating better performance. The results are presented across multiple datasets and evaluated using different criteria to assess how different segments can contribute to AI text detection.

understand the relative importance of the introduction, body, and conclusion in distinguishing human and AI text. Accordingly, we apply each detector to the total text, individual segments, and a combined introduction & conclusion segment. We also test a simple voting mechanism across the three segments. Results are summarized in Table 5.

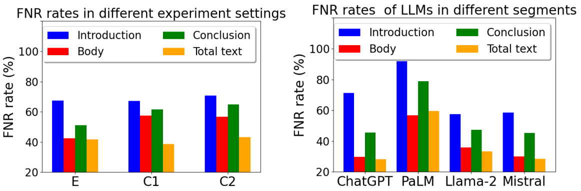
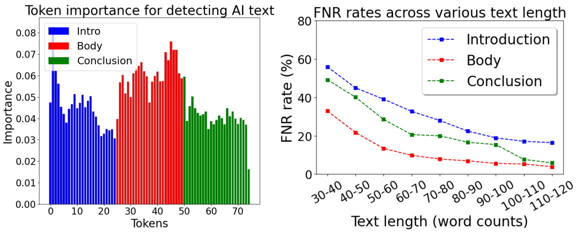


Figure 4: Comparison of False Negative Rates (FNR) in different experimental settings & datasets. Lower value indicates this segment contributes more in detection.

Overall, using the entire text yields the highest detection performance across most domains, except for the email. It aligns with the nature of email writing: introductions and conclusions often include formulaic greetings or closing remarks, while the body contains the most meaningful content. Across all domains, the body consistently plays a dominant role in AI text detection, outperforming both the introduction and conclusion, even when combined. Interestingly, the voting mechanism across segments fails to improve performance, likely due to redundancy or the overwhelming influence of the body segment. Notably, fine-tuned classifiers consistently benefit from analyzing the complete text, as they leverage more data during training. Appendix F provides the AI text detection results for other experimental settings.



(a) (b)

Figure 5: **(a)** Average importance from each token for identifying AI-generated text, showing higher contributions from tokens in the body segment. **(b)** False Negative Rate (FNR) decreases as the word count of a given segment increases. Even when the introduction, body, and conclusion are around the same length, the body segment has a lower FNR, making it stand out from the introduction and conclusion.

To account for the body segment's longer length in the original setting *(E)*, we assess detection performance using *False Negative Rate* (FNR), the proportion of AI text misclassified as human, across all settings & datasets (Figure 4). A lower FNR indicates better detector performance, as the text is more easily identified as LLM-generated, making it more distinguishable from human text. Conversely, a higher FNR suggests that the text closely resembles human writing, causing the detector to struggle to label it as AI text. Consistently, the body segment yields the lowest FNR, suggesting that it is more distinguishable from human text than the introduction or conclusion. Prior work (Huang et al., 2024; Wu et al., 2024) shows that longer texts generally improve detection, a trend we confirm in (Figure 5), where FNR declines as text length increases. Yet, within comparable length ranges, the body segment still exhibits the lowest FNR. To quantify which parts of the text contribute most to being flagged as AI text, we use Integrated Gradients (Sundararajan et al., 2017) to estimate token-level importance in our fine-tuned BERT classifier. For each correctly predicted sample, we compute the gradient of the model's output with respect to each token and normalize the resulting attribution scores to obtain a list of token importances. We then divide each sample into three equal-length segments: start, middle, and end (mirroring our $C1$ segmentation strategy to minimize length confounds), and average normalized importance scores within each segment. These scores are then aggregated across all samples to produce a final token importance profile for each segment. As shown in Figure 5 (left), we find that the middle segment consistently receives higher attribution, suggesting that it plays a more decisive role in distinguishing AI-generated text from human-written text.

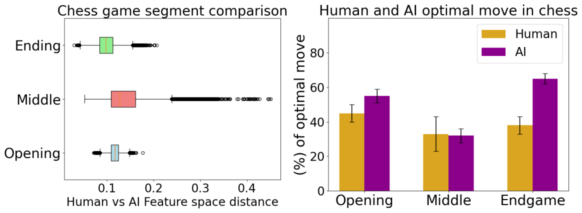
Dataset	MAGE	MAGE+	RADAR	RADAR+	Binocular	Binocular+
Reuter	0.85	0.87	0.69	0.87	0.68	0.91
Persuade	0.86	0.88	0.84	0.85	0.89	0.90
Enron	0.88	0.81	0.82	0.7	0.57	0.65

Table 6: Cross-segment feature differences enhance the performance of base detectors in identifying AI text from human-AI text pairs. Green cells indicate improved performance when using cross-segment variation instead of detector confidence scores, while Red cells indicate decreased performance.

Finally, cross-segment variation between human and AI texts (**source comparison** results) prompts us to explore its utility in AI text detection. We frame the task as identifying the AI text from a given (human, AI) pair. When existing detectors assign the same label to both texts, rather than relying solely on their confidence scores (denoted as *detector_name*), we use the cross-segment variation (based on the C1 setting, which splits text into three equal parts and is more practical for real-world use) as the deciding factor (*detector_name*+). This simple yet effective strategy improves detection accuracy across most detectors and datasets (Table 6), demonstrating that cross-segment variation offers a promising new lens for AI text detection.

4.3 Human and AI chess moves comparison

As our study was inspired by the chess middlegame analogy, We also investigate whether the differences between human and AI players emerge most noticeably in the middlegame. To quantify these differences, we calculate the JSD distance between the feature sets of human and AI moves across the opening, middlegame, and endgame phases. As shown in Figure 6, the middlegame exhibits a statistically significant ($\alpha = 0.05$) increase in JSD, indicating higher divergence during this phase. Moreover, the middlegame shows a broader spread of JSD values, reflecting higher variability in how humans and AI play diverges. We further compute Jaccard similarity over unique move patterns, represented by distinct Standard Algebraic Notation (SAN) moves exhibited by a player and observe lower overlap in the middlegame compared to the opening and endgame, reinforcing that this phase carries the most distinction. These findings echo our **text segment comparison** results, where the body or "middlegame" segment also reveals the highest differences between humans and AI.



(a) (b)

Figure 6: **(a)** The middlegame exhibits the most significant divergence between human and AI players. **(b)** AI players outperform humans in optimal move percentage during the opening and endgame, but the difference is not statistically significant in the middlegame.

Finally, we analyze the percentage of optimal moves and win probability using the Stockfish game engine (Romstad et al., 2008) for each move. As expected, AI players achieve higher optimal move rates and win probabilities, particularly in the endgame phase. AI engines often perform exceptionally well in endgames due to their access to precomputed endgame scenarios, which provide exact move sequences for optimal play to ensure victory. These tablebases (Thompson, 1986) are derived from exhaustive analysis rather than historical data and offer perfect information, giving AI engines a decisive advantage in such positions, an advantage human players, regardless of skill level, typically do not possess. Additionally, we observe that the percentage of optimal moves generally increases with Elo rating, but at a steeper rate for AI players than for humans (Figure 7). Within the same Elo range, AI players also make more optimal moves in the endgame compared to the middlegame, whereas for humans, performance remains relatively stable across these phases. While differences in play style between humans and AI across game segments (Christian, 2011; McIlroy-Young et al., 2021) motivated part of our analysis, this was not the primary focus of the study, and we therefore refrained from a deeper investigation.

Nonetheless, these findings align with and further validate our broader observations on segment-level distinctions, and we believe this direction merits dedicated exploration in future work.

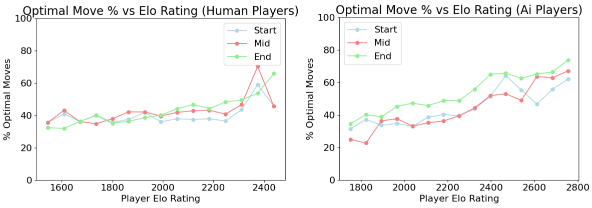


Figure 7: Percentage of optimal moves across game positions for human and AI players, plotted against player Elo rating. Each point represents the mean percentage of optimal moves within the corresponding Elo bin.

5 Discussion

In this section, we highlight key findings that reinforce our central claim, offer valuable insights into human creativity, and demonstrate the broader applicability of our results.

Text length matters We find that LLMs' ability to replicate human stylometry and linguistic features is influenced by text length. Initially, the body segment appears more similar to human text due to its greater length. Longer texts also yield higher AI text detection accuracy, aligning with prior studies (Liu et al., 2020; Liu, 2024; Baillargeon and Lamontagne, 2024; Jeon and Strube, 2021), which show improved classification and higher similarity scores in lengthier samples (Klaussner et al., 2015; Päpcke et al., 2023). Therefore, LLMs can better approximate human writing when given the chance to generate more tokens, as they have more room to establish consistent stylistic patterns, an insight critical to understanding and detecting AI text.

Distribution vs. textual divergence Our study offers a comprehensive view of how well LLMs replicate different linguistic features. LLMs consistently excelled at replicating the features that do not rely on word orders in sentences but instead depend on overall word choices, such as pos-tags, stopword distributions, or readability scores, showing no observable statistical differences with humans across experiments. In contrast, for features that capture the continuous flow of text, such as token-level perplexity or content change through that text, human and AI texts exhibited significant differences across experimental conditions. These insights can assist platforms like Turnitin, Grammarly, or Originality to integrate flow-based stylometric checks for AI text detection.

Body segment: more interesting for Human-AI text distinction While a longer body segment makes human and AI texts appear more stylistically similar for that segment, body/middle consistently shows higher divergence in length-controlled settings. Additionally, AI-generated introductions and conclusions yield higher false negative rates, suggesting detectors perceive them as more human-like. Token importance further confirms the body segment's superior discriminatory power. Thus, when distinguishing between human and AI texts, the body segment offers the most revealing starting point.

Cross-segment variation as a signal for AI text detection Our **source comparison** shows that cross-segment linguistic and contextual differences are consistently more pronounced in human texts than in AI-generated ones. It suggests that LLMs maintain a uniform writing style across segments, while humans naturally vary their linguistic patterns throughout a text. Importantly, we find that leveraging these cross-segment stylometric differences as a secondary signal can enhance the performance of existing AI text detectors, highlighting a promising new direction for detection strategies.

6 Conclusion

Our paper offers a novel perspective by identifying subtle differences between human and AI texts across specific text segments, an area that has remained largely overlooked. Drawing parallels from chess game phases, we conduct a thorough evaluation of linguistic features, analogous to chess "chokepoints" and explore how they vary in each segment between AI and human text. Our experimental design and detailed segment-wise analysis offer robust insights into LLMs' strengths and limitations in mimicking human text. Overall, our findings highlight the pivotal role of the body segment in distinguishing AI from human text and propose that cross-segment feature differences may serve as a novel and valuable characteristic for AI text detection. In future, we aim to extend our findings to other domains and contribute to responsible LLM usage to ensure accurate outputs across all text segments.

Limitations

While this study presents new findings in differentiating between human and AI text, inspired by chess game dynamics, there are some limitations to acknowledge. First, the scope of our analysis is restricted to three domains and texts from four LLMs. Additionally, the AI texts are collected from existing datasets that used generic prompts, which may affect the generalization of our findings to other domains, models, or prompting techniques. Secondly, dividing a text into introduction, body, and conclusion is inherently subjective, and while we show that an LLM can perform this segmentation, demonstrating alignment with human judgment, alternative approaches may yield different results. Moreover, not all domains, such as creative writing or social media posts, naturally follow a tripartite structure. Thus, applying our framework to such cases will require special attention. Despite these constraints, our study makes a substantial contribution by exploring human-AI text distinctions from a novel angle and can inform ongoing AI text detection research.

Ethical Considerations

Our study raises important ethical considerations regarding the responsible development, evaluation, and deployment of Large Language Models (LLMs). By analyzing segment-level distinctions between human and AI-generated texts, our goal is not to stigmatize AI use in writing but to promote transparency and accountability in its application. The insights from this research intend to strengthen detection mechanisms that help prevent misuse, such as academic dishonesty, misinformation, or deceptive authorship, while also informing the development of more interpretable and aligned LLMs. All AI-generated texts used in this study were created under controlled, non-deceptive conditions or collected from existing public datasets, and no personal, sensitive, or private human data was used. As detection technologies advance, it remains crucial to balance innovation with privacy, avoid over-surveillance, and ensure that such tools are not misused to unjustly penalize legitimate human writing.

Acknowledgments

This work was in part supported by U.S. National Science Foundation (NSF) awards #1934782 and #2114824. Some of the research results were obtained using computational resources provided by National Artificial Intelligence Research Resource (NAIRR) award #240336 and complimentary API support by the GPTZero. D. Lee has a financial interest in the GPTZero, which is being appropriately managed by Penn State University.

References

Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ta, Raj Tomar, Bimarsha Adhikari, Saad Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, et al. 2024. Llm-detectaive: a tool for fine-grained machine-generated text detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 336–343.

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):1–29.

Jokha Al Hosni. 2024. Stylometric analysis of ai chatbot-generated emails: Are students losing their linguistic fingerprint? *Journal of English Language Teaching and Applied Linguistics*, 6(3):33–42.

Dennis Aumiller, Satya Almasian, Sebastian Lackner, and Michael Gertz. 2021. Structural text segmentation of legal documents. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 2–11.

Nikolay Babakov, David Dale, Ilya Gusev, Irina Krotova, and Alexander Panchenko. 2023. Don't lose the message while paraphrasing: A study on content preserving style transfer. In *Natural Language Processing and Information Systems*, pages 47–61, Cham. Springer Nature Switzerland.

Jean-Thomas Baillargeon and Luc Lamontagne. 2024. Assessing the impact of sequence length learning on classification tasks for transformer encoder models. In *The International FLAIRS Conference Proceedings*, volume 37.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Lamia Berriche and Souad Larabi-Marie-Sainte. 2024. Unveiling chatgpt text using writing style. *Heliyon*, 10(12).

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10.

James A Brown, Alfredo Cuzzocrea, Michael Kresta, Korbin DL Kristjanson, Carson K Leung, and Timothy W Tebinka. 2017. A machine learning tool for supporting advanced knowledge discovery from chess game data. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 649–654. IEEE.

Étienne Brunet et al. 1978. *Le vocabulaire de Jean Giraudoux structure et évolution*. Slatkine, Genève, Switzerland.

Murray Campbell, A Joseph Hoane Jr, and Feng-hsiung Hsu. 2002. Deep blue. *Artificial intelligence*, 134(1-2):57–83.

J Elliott Casal and Matthew Kessler. 2023. Can linguists distinguish between chatgpt/ai and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, 2(3):100068.

Shirish Chinchalkar. 1996. An upper bound for the number of reachable positions. *ICGA Journal*, 19(3):181–183.

Brian Christian. 2011. *The most human human: What talking with computers teaches us about what it means to be alive*. Anchor, NewYork.

Scott A Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, and Ulrich Boser. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (persuade) corpus 1.0. *Assessing Writing*, 54:100667.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.

Mark Dvoretsky. 2020. *Dvoretsky's endgame manual*. SCB Distributors, Gardena, CA.

Dominik Maria Endres and Johannes E Schindelin. 2003. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860.

Michele Filannino. 2011. Dbworld e-mails.

Ronald Aylmer Fisher. 1970. Statistical methods for research workers. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 66–70. Springer, New York, NY.

Chris Fournier and Diana Inkpen. 2012. Segmentation similarity and agreement. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 152–161.

AC Graesser. 2003. What do readers need to learn in order to process coherence relations in narrative and expository text. *Rethinking Reading Comprehension/Guilford Publications*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2024. Benchmarking linguistic diversity of large language models. *arXiv preprint arXiv:2412.10271*.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024a. Spotting llms with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, pages 17519–17537.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024b. Spotting llms with binoculars: Zero-shot detection of machine-generated text. In *International Conference on Machine Learning*, pages 17519–17537. PMLR.

Marti A Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 9–16.

Ernst A Heinz. 1999. Endgame databases and efficient index schemes for chess. *ICGA Journal*, 22(1):22–32.

Felix Helfenstein, Jannis Blüml, Johannes Czech, and Kristian Kersting. 2024. Checkmating one, by using many: Combining mixture of experts with mcts to improve in chess. *arXiv preprint arXiv:2401.16852*.

Alex Henry and Robert L Roseberry. 1997. An investigation of the functions, strategies and linguistic features of the introductions and conclusions of essays. *System*, 25(4):479–495.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific reports*, 13(1):18617.

M Hollander. 2013. *Nonparametric statistical methods*. John Wiley & Sons Inc, New York.

Israel Albert Horowitz. 1986. *How to Win in the Chess Openings*. Simon and Schuster, New York.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Radar: Robust ai-text detection via adversarial learning. *Advances in neural information processing systems*, 36:15077–15095.

Anna Huang et al. 2008. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *SIGKDD explorations*.

Sungho Jeon and Michael Strube. 2021. Countering the influence of essay length in neural essay scoring. In *Proceedings of the second workshop on simple and efficient natural language processing*, pages 32–38.

Cameron R Jones and Benjamin K Bergen. 2025. Large language models pass the turing test. *arXiv preprint arXiv:2503.23674*.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Jules King, Perpetual Baffour, Scott Crossley, Ryan Holbrook, and Maggie Demkin. 2023. LLM - Detect AI Generated Text. Kaggle. https://kaggle.com/competitions/llm-detect-ai-generated-text.

Carmen Klaussner, John Nerbonne, and Çağrı Çöltekin. 2015. Finding characteristic features in stylometric analysis. *Digital Scholarship in the Humanities*, 30(suppl_1):i114–i129.

Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pages 217–226. Springer.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023. Deepfake text detection in the wild. *arXiv preprint arXiv:2305.13242*.

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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Tingyue Liu, Yunji Liang, and Zhiwen Yu. 2020. The influence of text length on text classification model. In *Green, Pervasive, and Cloud Computing–GPC 2020 Workshops: 15th International Conference, GPC 2020, Xi'an, China, November 13–15, 2020, Proceedings 15*, pages 79–90. Springer.

Xiaoyu Liu. 2024. Study for text length impact on text classification accuracy based on the transformer method. In *IET Conference Proceedings CP895*, volume 2024, pages 174–178. IET.

Erika Matruglio. 2020. Beyond'introduction, body, conclusion': Purpose and form in senior history essays. *Teaching History*, 54(2):5–11.

Reid McIlroy-Young, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. 2020. Aligning superhuman ai with human behavior: Chess as a model system. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1677–1687.

Reid McIlroy-Young, Yu Wang, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. 2021. Detecting individual decision-making style: Exploring behavioral stylometry in chess. *Advances in Neural Information Processing Systems*, 34:24482–24497.

Olena Medvid and Svitlana Podolkova. 2019. Essay as a form of academic writing. *Edukacyjna Analiza Transakcyjna*, (8):215–225.

Sebastian Michelmann, Manoj Kumar, Kenneth A Norman, and Mariya Toneva. 2025. Large language models can segment narrative events similarly to humans. *Behavior Research Methods*, 57(1):1–13.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume PMLR 202, pages 24950–24962.

Karsten Müller and Jonathan Schaeffer. 2018. *Man vs. Machine: Challenging Human Supremacy at Chess*. SCB Distributors, Gardena, CA.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, 50(6):1–36.

Aron Nimzowitsch. 1925. *My system*. Open Road Media, New York.

Andrea Nini, Oren Halvani, Lukas Graner, Valerio Gherardi, and Shunichi Ishihara. 2024. Authorship verification based on the likelihood ratio of grammar models. *arXiv preprint arXiv:2403.08462*.

Barak Oshri and Nishith Khandwala. 2016. Predicting moves in chess using convolutional neural networks. *ConvChess. pdf*.

Hitanshu Panchal, Siddhant Mishra, and Varsha Shrivastava. 2021. Chess moves prediction using deep learning neural networks. In *2021 International Conference on Advances in Computing and Communications (ICACC)*, pages 1–6. IEEE.

Bruce Pandolfini. 1997. *Kasparov and Deep Blue: The historic chess match between man and machine*. Simon and Schuster.

Simon Päpcke, Thomas Weitin, Katharina Herget, Anastasia Glawion, and Ulrik Brandes. 2023. Stylometric similarity in literary corpora: Non-authorship clustering and deutscher novellenschatz. *Digital Scholarship in the Humanities*, 38(1):277–295.

Suko Raharjo and Deli Nirmala. 2016. Generic structure and cohesive devices: A study on the final project report presentation of the accounting students of polines semarang. *Parole: Journal of Linguistics and Education*, 6(2):27–40.

Diego Rasskin-Gutman. 2009. *Chess metaphors: Artificial intelligence and the human mind*. MIT Press, Cambridge, MA.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Pál Révész. 2014. *The laws of large numbers*, volume 4. Academic Press, New York.

T. Romstad, M. Costalba, J. Kiiski, G. Linscott, Y. Nasu, M. Isozaki, H. Noda, and et al. 2008. Stockfish. URL https://stockfishchess.org.

Ariel Rosenfeld and Teddy Lazebnik. 2024. Whose llm is it anyway? linguistic comparison and llm attribution for gpt-3.5, gpt-4 and bard. *arXiv preprint arXiv:2402.14533*.

Anian Ruoss, Grégoire Delétang, Sourabh Medapati, Jordi Grau-Moya, Li Kevin Wenliang, Elliot Catt, John Reid, and Tim Genewein. 2024. Grandmaster-level chess without search. *arXiv preprint arXiv:2402.04494*.

David Ruppert. 2004. The elements of statistical learning: Data mining, inference, and prediction. *Journal of the American Statistical Association*, 99(466):567.

Scott Shane and Michael S Schmidt. 2015. Hillary clinton emails take long path to controversy. *The New York Times*.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12395–12412.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (llms)? aka will llms replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 311–325.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*.

Ken Thompson. 1986. Retrograde analysis of certain endgames. *ICGA Journal*, 9(3):131–139.

Edward Tian. 2023. Gptzero. Online; accessed 23-Mar-2023.

Martina Toshevska and Sonja Gievska. 2025. Llm-based text style transfer: Have we taken a step forward? *IEEE Access*.

Nafis Irtiza Tripto, Adaku Uchendu, Thai Le, Mattia Setzu, Fosca Giannotti, and Dongwon Lee. 2023. Hansen: Human and ai spoken text benchmark for authorship analysis. In *Findings of Conf. on Empirical Methods in Natural Language Processing (EMNLP-Findings)*.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. 2022. On the state of the art in authorship attribution and authorship verification. *arXiv e-prints*, pages arXiv–2209.

Teun A Van Dijk. 1980. An interdisciplinary study of global structures in discourse, interaction, and cognition. *MacrostructuresErlbaum, Hillsdale, NJ.*

MH Van Emden. 1982. Chess endgame advice: A case study in computer utilization of knowledge. *Introductory Readings in Expert Systems*, 1:113.

Yuli Vasiliev. 2020. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press, San Francisco, CA.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2023. Gpt-who: An information density-based machine-generated text detector. *arXiv preprint arXiv:2310.06202*.

Saranya Venkatraman, Adaku Uchendu, and Dongwon Lee. 2024. Gpt-who: An information density-based machine-generated text detector. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 103–115.

Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717.

Junchao Wu, Runzhe Zhan, Derek F Wong, Shu Yang, Xinyi Yang, Yulin Yuan, and Lidia S Chao. 2024. Detectrl: Benchmarking llm-generated text detection in real-world scenarios. *arXiv preprint arXiv:2410.23746*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Evgeni Aleksandrovich Znosko-Borovski. 1922. *The middle game in chess*. G. Bell and Sons, Limited, London.

A Prompt engineering

While we primarily use human and AI text in various domains from existing datasets, we also employ LLMs for missing data generation and text segmentation. As mentioned, we select *GPT-3.5* (OpenAI), PaLM *text-bison-001* (Google), *LLaMA 2-Chat-7B* (Meta), and *Mistral-7B* (Mistral AI) as our LLMs. Several data were missing in the original datasets collected from (Verma et al., 2024) or (King et al., 2023). For example, Reuters news articles from any Google model were unavailable in the original *Ghostbuster* dataset (Verma et al., 2024). So, we generated them using *text-bison-001* using identical prompts from the original paper (Verma et al., 2024). Similarly, for the email dataset, we generate AI text from all four LLMs, as only human-written emails are available in the Enron corpus (Klimt and Yang, 2004). For segmentation, we use *Gemini-1.5-Flash* (Google) and *GPT-4* (OpenAI), which are distinct from the models used for text generation in our study. Proprietary models from Google and OpenAI are accessed via their official APIs, while open-source models from Meta and Mistral are sourced from their stable weights on Hugging Face. Across all settings, we use **top_p** = 0.95 and **temperature** = 0.9 to maintain consistency. However, it is important to note that even with identical prompts and hyperparameters, LLM outputs are not entirely deterministic.

Prompt for news data

```
Suppose You are <reporter_name>, a news reporter
in Reuter. Write a news article in
<original_word_count> words with the following
headline (output news text only, do not include
headline):
<original_headline>
```

Prompt for email data

```
Create an email (only the email body) as an
Enron employee <sender_name> to <receiver_name>
around <original_word_count> words based on the
subject: <original_email_header>. The summary of
the original email is as follows.
<original_email_summary>
```

Prompt for text segmentation

```
You are advanced in essay understanding and
writing. Given the following text you need to
divide it into three parts: introduction, main
body and conclusion. For each part, only copy
relevant portion from the original text. Do not
use any other formatting.
{Introduction}:the intro goes here
{Body}:the main body goes here
{Conclusion}:the conclusion goes here
The text is as follows:
<original_text>
```

B Statistical test details

As mentioned in Subsection 3.4, we have two text sources **(Sources, H: Human, A: AI)** and three segments from each text **(Segments, I: Introduction, B: Body, C: Conclusion)**. Z_x is an individual feature extracted from segment x for source Z.

For **source comparison** tests, we consider pairwise segments, $x, y \in \{I, B, C\}$, compute their differences for human and AI texts, $\Delta(H_x, H_y)$ and $\Delta(A_x, A_y)$, respectively. Then, we address the key question, whether $\Delta(H_x, H_y)$ differs significantly from $\Delta(A_x, A_y)$ for any segment pair. We conduct a two-way ANOVA test ($\alpha = 0.05$) (Fisher, 1970) focusing on the interaction effect of source (H vs. A) and cross-segment differences. If the interaction effect is significant, we proceed with post-hoc pairwise comparisons using the Wilcoxon signed-rank test. We opted for Wilcoxon signed-rank tests instead of t-tests due to the robustness to non-normal distributions (Hollander, 2013). These pairwise tests reveal whether human cross-segment differences $\Delta(H_x, H_y)$ are statistically greater than (>), less than (<), or comparable ($\sim$) to AI cross-segment differences $\Delta(A_x, A_y)$, for specific segment pairs. If no significant interaction effect is found in the ANOVA test, we infer that cross-segment differences between human and AI texts are not statistically meaningful.

Similarly, for **segment comparison**, we compute the difference between human and AI texts for all three segments, $\Delta(H_I, A_I)$, $\Delta(H_B, A_B)$, and $\Delta(H_C, A_C)$. Then, we conduct a one-way ANOVA test ($\alpha = 0.05$) with the three measures. If the result is statistically significant, we perform post-hoc pairwise comparisons between $\Delta(H_x, A_x)$ and $\Delta(H_y, A_y)$ for all segment pairs $x, y \in \{I, B, C\}$. The post-hoc tests determine whether the human-AI feature difference is more pronounced in a specific segment or whether the differences are statistically indistinguishable across segments. If the ANOVA test shows no significant effects, we con-

Opening conditions	**reasonings**	**Mid game**	**End game conditions**	**reasonings**
# of moves <= 16 **OR**	All classic chess openings are done in mostly 16 moves (Horowitz, 1986)	All other moves that are not classified as opening or end game moves	If total # moves<=50 then end game consist 35% of last moves else 45% of last moves **OR**	Overall distribution of moves in different phases and general ideas(Van Emden, 1982)
# of pieces exchanged<=8 **OR**	Initial exchanges have taken place and game has moved to mid game (Chinchalkar, 1996)		Less then 12 pieces remain **OR**	Board is simplified and both players aim for strategic checkmate (Dvoretsky, 2020; Heinz, 1999)
Both castling are available	If both players have done castling, game has moved to mid game (Nimzowitsch, 1925)		# of legal moves for both kings>=8 and both kings are in third row (row 3 or 6)	King has taken a more active role in the game (Dvoretsky, 2020; Heinz, 1999)

Table 7: Criteria used for categorizing chess moves into opening, midgame, or endgame phases. The rationale for each criterion is provided in separate columns for clarity.

clude that the differences between human and AI texts for the analyzed feature do not vary meaningfully across segments.

C Chess features extractions

Similar to segmenting text, dividing chess moves into opening, middlegame, and endgame can be subjective, as there are no strict rules for defining these transitions (Helfenstein et al., 2024). While openings are identified by ECO codes, the middle game does not always begin immediately after these moves, nor can the start of the endgame be consistently determined by board conditions alone. Therefore, we draw on reasoning from existing studies (Horowitz, 1986; Van Emden, 1982; Chinchalkar, 1996; Dvoretsky, 2020; Heinz, 1999; Nimzowitsch, 1925), using factors such as piece counts, board conditions, and castling status to segment the games (Table 7). To validate our rule-based method, we employ an LLM (*GPT-4*) to segment a subset of 2000 games, achieving a segmentation similarity score of 0.94, indicating its effectiveness in approximating chess move segmentation.

Prompt for chess game segmentation

```
You are an expert in chess game understanding and
moves. From the given list of moves you need
to divide them into chess start, middle and end
game moves. Your output should be strictly in the
following format:
{Start}: <list of start game moves in comman
seperated format>
{Middle}: <list of mid game moves in comman
seperated format>
{End}: <list of mid game moves in comman seperated
format>
moves list: <original_move_list>
```

Our next step involves creating a feature list from chess moves to computationally assess the differences between human and AI across game segments. While prior works have focused on cognitive aspects of chess play (e.g., memory, decision-making (Rasskin-Gutman, 2009)) or expert-driven analysis of key moments (Müller and Schaeffer, 2018), recent advances in deep learning have enabled computational feature extraction in chess for tasks like next optimal move prediction, game outcome projection, and game clustering (Oshri and Khandwala, 2016; Brown et al., 2017; Panchal et al., 2021). Drawing on these studies, we extract 72 features related to board conditions, piece movements, positions, and captures. We also incorporate the optimal move and the corresponding player's win probability, as determined by the Stockfish engine (Romstad et al., 2008) (with $time\_limit = 0.1$ second) for each position.

D Text Features Extraction Details

In this section, we discuss the details of extracting linguistic features from text that are essential to our analysis. For vocabulary richness, we consider the Brunét Index (Brunet et al., 1978), as it is less sensitive to text length than the type-token ratio (TTR), making it more suitable for segments of varying lengths. For readability, we compute the Flesch Reading Ease score and employ the Python Textdescriptive library for additional linguistic insights.

Syntactic features include part-of-speech (POS) tags, named entity recognition (NER), and stopword distributions extracted using SpaCy (Vasiliev, 2020). We further assess affective and stylistic elements through average sentiment and subjectivity scores using the VADER sentiment library, and formality scores via a pre-trained classifier (Babakov et al., 2023).

For content analysis, we use OpenAI text embeddings (*text-embedding-ada-002*) to capture the content within segments and measure the variation in embeddings between consecutive sentences or evaluate text predictability, we utilize GPT-2 to calculate both average perplexity and token-level perplexity scores, alongside burstiness, a metric

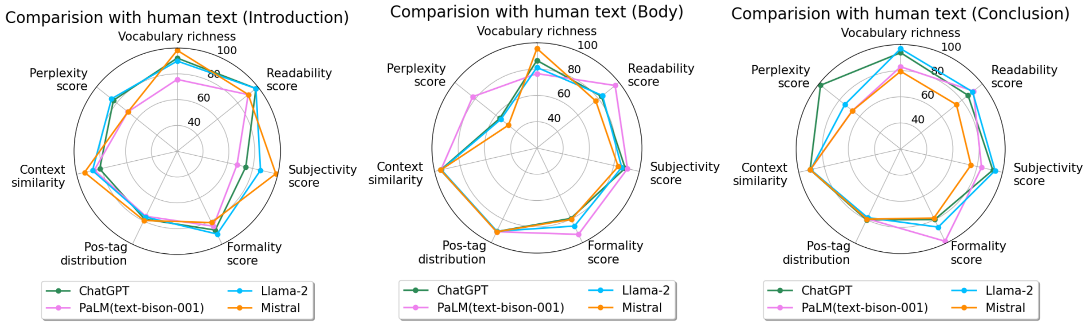


Figure 8: Comparison of individual LLMs to human text across segments. Values represent the similarity percentages (0-100) between AI text and human text for specific features, illustrating the extent to which individual LLMs can replicate human feature distributions.

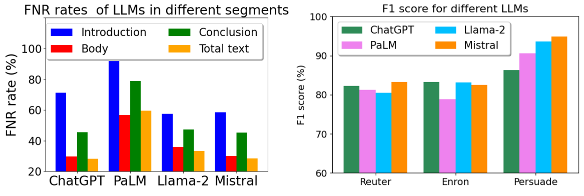
that captures shifts in sentence structure and word choice. These features, shown to be impactful in recent AI text detection efforts (Tian, 2023; Venkatraman et al., 2023; Mitchell et al., 2023), provide a comprehensive lens through which to explore the nuanced differences between human and AI-generated writing.

E Results for individual LLMs

We also analyze how individual LLMs replicate human feature distributions across different text segments (Figure 8 and 9). Overall, LLMs effectively mimic linguistic features, with the highest similarity observed in the body segments for most features, except for perplexity scores. ChatGPT demonstrates relatively balanced performance, while PaLM exhibits higher variability across segments. However, higher similarity scores do not necessarily imply that these LLMs are more challenging to detect, as detailed in the following subsection. Our analysis shows consistent performance across the three datasets. However, due to the shorter length of emails, they often lack clear structural distinctions. This results in some statistically insignificant findings in source and segment comparisons when contrasted with the other datasets.

F AI text detection methods

GPTZero: To determine whether a text is LLM-generated, GPTZero (Tian, 2023) uses perplexity to measure the text's complexity and burstiness to evaluate sentence variants for providing the final output. We utilize the official API of GPT-Zero in our experiments.



(a)

(b)

Figure 9: **(a)** Analysis of False Negative Rates (FNR) across segments. Body shows lower FNR than introduction and conclusion for all LLMs. **(b)** F1 score comparison across datasets (using Finetuned BERT method). Essays demonstrate the highest F1 score, indicating more differences between human and AI-generated texts in this domain. Notably, F1 scores show minimal variation across different LLMs.

MAGE: MAGE (Machine-generated Text Detection in the Wild) is a Longformer model (Li et al., 2024), finetuned on the entire Deepfakedetect (Li et al., 2023) dataset (comprising 447,674 human-written and AI texts). By effectively managing more than 512 tokens, Longformer (Beltagy et al., 2020), a modified Transformer architecture, gets around the drawbacks of conventional transformer models. Longer documents can be processed more easily because of their attention pattern, which scales linearly with sequence length. We also access the model from the HuggingFace repository[4].

RADAR: RADAR is a robust AI text detection framework that leverages adversarial learning by jointly training a paraphraser and a detector (Hu et al., 2023). The paraphraser aims to generate realistic, human-like text that can evade detection, while the detector learns to identify such paraphrased AI-generated content. In our study, we utilize the hosted version of RADAR available on Hugging Face[5].

[4] https://huggingface.co/yaful/MAGE

[5] https://huggingface.co/spaces/TrustSafeAI/RADAR-AI-Text-Detector

Binocular: Binoculars is a zero-shot, domain-agnostic method for AI text detection that operates without the need for training data (Hans et al., 2024a). It relies on cross-perplexity, computed as the cross-entropy between two language models that sharing the same tokenizer and vocabulary, when evaluated on a given text. Following the original implementation, we use the *Falcon-7B* and *Falcon-7B-Instruct* models for cross-perplexity computation in our experiments.

GPT-who: GPT-who (Venkatraman et al., 2023) is a domain-agnostic statistical AI text detector that uses UID-based characteristics to capture unique statistical signatures. UID features are created via GPT2 inference and trained with a logistic regression model.

Finetuned-BERT: We fine-tuned BERT (*bert-base-cased*) on each dataset training set and evaluated it on the test set, as fine-tuned language models have been state-of-the-art in a lot of text classification and authorship tasks (Tyo et al., 2022).

G AI text detection results for controlled settings

As noted earlier, we conduct a length-controlled analysis to examine whether the middle portion of a text is more distinctive and contributes more to AI text detection. Tables 8 and 9 present the results for settings $C1$ (equal segmentation) and $C2$ (subsampled body matched to the length of the introduction and conclusion), respectively. In these experiments, detection is performed using only a specific segment of the text, along with voting across segment-level predictions. We exclude results using the total text, as they replicate the outcomes already reported for the original setting (E) in Table 5.

Across most datasets and detectors, the body (middle) segment consistently achieves higher F1 scores than the introduction or conclusion, reinforcing our findings from the original setting. It suggests that, even when length is controlled, the middle segment conveys stronger signals for distinguishing AI-generated text from human-written text. Moreover, voting shows improved performance compared to the original setting, highlighting its robustness. Finally, we note that results from $C1$ generally surpass those of $C2$ when using the body segment, since $C2$ preserves intact text segments rather than artificially truncated ones.

Dataset	**Criteria**	**GPT Zero**	**MAGE**	**RADAR**	**Binoculars**	**GPT-Who**
Reuters	**Intro**	0.7309	0.7985	0.8186	0.7902	0.7201
	Body	**0.7574**	**0.8271**	0.8401	0.8576	**0.7278**
	Conclusion	0.7117	0.8263	0.8526	0.8572	0.7102
	Voting	0.7512	0.8026	**0.8583**	**0.8965**	0.7196
Enron	**Intro**	0.3772	0.8545	0.8290	0.1974	0.8127
	Body	**0.6295**	0.8193	**0.8231**	**0.7052**	**0.8100**
	Conclusion	0.5070	0.8612	0.8157	0.3552	0.8086
	Voting	0.4667	**0.8760**	0.8180	0.3237	0.8078
Persuade	**Intro**	0.7972	0.7209	0.5021	0.7903	0.7862
	Body	**0.8106**	**0.7423**	**0.5221**	0.7864	0.7872
	Conclusion	0.7320	0.7210	0.4908	0.8061	0.7857
	Voting	0.7990	0.7231	0.4821	**0.8436**	**0.8336**

Table 8: F1 scores of AI text detectors in the length-controlled setting ($C1$). Each value corresponds to the F1 score using the specified segment for a given dataset. **Bold** values highlight the segment or criterion achieving the highest F1 for each detector. Overall, the Body segment or Voting generally yields the best performance.

Dataset	**Criteria**	**GPT Zero**	**MAGE**	**RADAR**	**Binoculars**	**GPT-Who**
Reuters	**Intro**	0.6167	0.7270	0.7222	0.7831	0.7186
	Body	**0.7566**	**0.7475**	**0.7295**	0.7608	**0.7251**
	Conclusion	0.6943	0.7531	0.7292	0.8093	0.7237
	Voting	0.7158	0.7280	0.7213	**0.8407**	0.7186
Enron	**Intro**	0.3281	0.8660	0.8021	0.2021	0.8070
	Body	0.5962	0.7956	0.8030	0.3453	**0.8076**
	Conclusion	**0.6048**	0.8580	**0.8032**	**0.6489**	0.8031
	Voting	0.4673	**0.8713**	0.8021	0.3066	0.8021
Persuade	**Intro**	0.7092	0.5821	0.4410	0.7557	0.7604
	Body	**0.8334**	**0.6512**	**0.4408**	0.7571	0.7637
	Conclusion	0.7809	0.6247	0.4398	0.8146	0.7756
	Voting	0.8085	0.5854	0.4398	**0.8406**	**0.8142**

Table 9: F1 scores of AI text detectors in the length-controlled setting ($C2$). We observe similar results like setting $C1$, Table 8