Journal of Digital Information Management

DLINE JOURNALS

ISSN: 0972-7272

JDIM 2026: 24 (3)

Structural and Linguistic Markers for Distinguishing Human Manuscripts from AI-Generated Summaries

Pit Pichappan
Digital Information Research Labs
Chennai 60017, Tamil Nadu. India
pichappan@dirf.org

ABSTRACT

The rapid development of Large Language Models (LLMs) has made it increasingly hard to distinguish human written from AI-generated academic papers, raising questions about authorship, originality, and academic integrity. This research presents a systematic approach to identifying and measuring multidimensional markers that differentiate human written academic manuscripts from AI-generated and AI-paraphrased summaries. After analysing 52 human written academic papers and their AI-generated versions (created with GPT-40, DeepSeek V4, and Qwen 3.7plus), we developed an automated, multi-layered text-annotation tool. It incorporates rule based matching, zero shot classification, and structured LLM-aided annotation, and we validated it using human inter annotator agreement. We found that no single language marker is enough for classification. However, the accurate classification (85-90%) requires a set of markers which include the completeness of the scholarly apparatus (the most reliable one), significant information loss (40-50% of granular details such as exact statistics or software name), decrease of syntactic burstiness (by 40-50%), formulaic transition substitutions, and the drop in epistemic hedging by 60%. Although the absence of structure and metadata yields perfect classification, the stylometric approach alone produced an AUC of 0.78, underscoring the complexity of linguistic signatures. In summary, successfully implementing AI for identifying text in academic environments requires a multi marker approach that favours completeness, informativeness, and text depth over fluency and vocabulary sophistication, which are highly prone to false positive results. Previous research has addressed AI academic detection; however, we shift our emphasis from surface fluency (which LLMs excel at mimicking) to epistemic rigour and granular information loss.

Subject Categories and Descriptors: [**H.1.2 User/Machine Systems**]; Human information processing: **[H.3 Information Storage and Retrieval]**: Linguistic processing: **[H.5.2 User Interfaces]**: Natural language [**I.2.7 Natural Language Processing**]; Language models

General Terms: Large Language Models, Human Information Processing, AI Multi-Markers, AI Detection

Keywords: AI-generated Text Detection, Large Language Models (LLMs), Academic Writing, Stylometric Analysis, Information Retention, Scholarly Apparatus, Automated Text Annotation, Syntactic Burstiness, Epistemic Hedging, Multi-marker Evaluation

Received: 19 May 2026, Revised: 3 July 2026, Accepted: 12 July 2026

Review Metrics: Review Scale: 0-6, Review Score: 4.62, Inter-Reviewer Consistency: 81.2%

DOI: https://doi.org/10.6025/jdim/2026/24/3/127-148

1. Introduction

Generative AI (LLMs) reflects one of the most transformative advancements in Natural Language Processing. These models are trained on massive corpora of text and other data and have shown an unprecedented ability to understand, generate, and manipulate natural text. Their applications span almost every sector and discipline, and users increasingly rely on them. At the same time, users tend to depend on them too heavily for text generation, challenging human capacity to generate knowledge.

2. Related Work

The speedy growth of generative artificial intelligence (AI), particularly Large Language Models (LLMs), has greatly changed the ways in which educational and professional realms create and apply written information. While web tools and electronic resources were already central to academic communication, generative AI amplifies their use and introduces new ethical and intellectual issues. More and more, authors use machine-written and machine assisted texts, and hybrid texts that mix human generated and machine generated content are increasingly common. This new situation has raised questions about authorship, originality of ideas, and the role of machine writing in academic writing. Modern LLMs can produce logical, meaningful text, making it difficult to distinguish between human written and AI-generated texts [1, 2]. However, research shows that AI-generated texts can contain serious factual mistakes, underscoring the need to use generative AI responsibly and systematically evaluate AI-generated texts [3].

The popularity of LLMs relies heavily on advances in computational linguistics and natural language technology [4]. Systems such as ChatGPT also create logical, context-relevant texts and raise many questions concerning authorship and textuality [5]). Researchers have done significant work identifying linguistic and stylistic features that might help differentiate AI-generated text from human writing. However, literature increasingly points out the impossibility of establishing this distinction based on any one linguistic feature or universal marker. A systematic review of differences between AI-generated and human writing showed that this distinction relies on multilayered, situational profiles of signals rather than a single effective indicator [6]. In addition, some studies assessing current LLM text-detection methods argue that multilinguality, hybrid texts, shifting language generation models, and regulation remain long term obstacles to reliable detection [7].

The limitations detected further complicate identifying machine generated text. Studies show poor and inconsistent results from existing detection tools [8-12]. These limitations matter because human evaluators also cannot distinguish machine written work from human written work.

AI-detection tools can flag suspicious texts, but their results should not be treated as concrete proof of improper AI use. Human judgments are often highly diverse and inconsistent [13]. The advancement of LLMs makes it even harder for even the most competent readers to determine whether a text was written by a human or an AI [14].

2.1 Linguistic Characteristics of AI-Generated and Human-Written Text

One central method for differentiating AI- and human-generated text is to analyse it at the lexical, syntactic, morphological, and readability levels. Studies show that AI-generated texts differ systematically from human-written essays. Studies report that AI-generated writing shows lower lexical diversity, greater syntactic complexity, and higher nominalisation than human writing [15, 16, 17]. This suggests that AI-generated texts should have a distinct linguistic profile, even when their content is coherent and grammatically correct. Fedoriv's [18] research demonstrates that potential markers of AI-generated English texts might include

repetition, illogicality, tone mismatches, and factual mistakes. However, these features should not be treated as definitive indicators of AI authorship, since language-generating technologies are still evolving.

The illustrations suggest differences in vocabulary use and structural organisation. Texts produced by GPT have be- en noted to use a smaller vocabulary, have little lexical variation, and repeat phrases frequently. In addition, the literature assumes that GPT-produced texts show specific characteristics in terms of syntactic complexity, morphology, and readability [19]. This supports the suggestion that GPT texts may have a sort of "register" characterised by specific recurrent structures. However, these features may differ across domains.

Comparative statistical analyses of previously mentioned texts provide more evidence. Human writing has less complex syntax but more semantic diversity, whereas human written texts show significant stylistic variation. At the same time, human texts show greater variability in the discussed linguistic features [20]. The results highlight the need to analyse multiple linguistic dimensions at once, rather than individual features.

Morphosyntactic features can also help distinguish the two. Specifically, AI-generated texts are believed to be more nominally loaded, with a higher density of nouns and less frequent use of pronouns and auxiliary verbs. Human texts are generally more morphologically complex when it comes to their reference; tense; and aspect [21, 22, 23, 24]. This means AI and human texts may differ not only in vocabulary and grammar but also in how authors express their viewpoints and references.

2.2 Stylistic, Structural, and Discourse-Level Characteristics

Beyond individual words, writing organisation and flow also differ between human and AI writing. Studies in different academic fields show that these organisational and flow features can distinguish student writing from AI-generated text [25]. This is especially true in school settings where writing needs to show more than just correct grammar. It should also show how well someone understands the subject, can build an argument, take a stand, and understand the context.

Some research has found that AI writing often has certain structural patterns. These include disconnected arguments, predictable organisation, specific language patterns, and an overly objective tone. [26]. This is different from how humans often write, which tends to be more connected and subtle. Also, whether people think a text is AI-generated isn't just about grammar. Teachers have pointed out that a lack of language mistakes, complicated sentence structures, advanced vocabulary, made up information, fake sources, repeating words or phrases, and noticeable text structures can all be signs of AI writing. [27]. This suggests that our judgment of AI text comes from a mix of factors: grammar, word choice, organisation, factual accuracy, and stylistic consistency.

It's getting harder to tell the difference because AI can now produce content that fits the context. Even though these systems can create text that sounds fluent and appropriate, worries remain about whether the information is accurate and whether they can capture the complex aspects of authorship typical of human writing. [28] These concerns go beyond surface level language differences, and touch on intent, point of view, originality, and how language relates to human thought. Because of this, some suggest using psycholinguistic approaches to find patterns in human writing and develop better ways to ensure academic honesty as AI becomes more common. [29].

2.3 Development of AI-Generated Text and Detection Approaches

The lines between human and AI writing are blurring because AI text generators keep getting better. As these AI systems are updated, they change how they write and what they produce, making their output look more and more like human writing. This means that tools designed to detect AI text need to keep up with these changes [30]. Some research shows that AI-generated essays can be longer, use more complex sentences, use a wider vocabulary, and include more information than human-written essays. Even with these differences, AI and human writing can express feelings and connect ideas logically in similar ways. AI writing might even use more complex sentence structures and advanced ways of building sentences [30]. This means telling AI writing apart from human writing isn't a simple, fixed task. The old idea that AI text is always repetitive, simple, or basic is probably no longer true as AI gets smarter. Things that used to be clear signs of AI writing might be gone or harder to spot in newer AI versions. At the same time, new patterns might emerge as AI gets better at sounding human. So, to reliably identify AI text, we need a flexible, constantly updated approach.

Scientists are also exploring machine learning to better distinguish AI writing from human writing. For instance, one researcher suggested combining a genetic algorithm and a multilayer perceptron with a BERT based method. Both performed better than older detection methods [31]. This shows a shift away from simple, set detection rules toward computer based methods that consider many aspects of the text. However, how well these methods work depends on the variety of data they are trained on, the specific AI language models used, and whether the testing data actually includes examples of current writing that mixes human and AI input.

2.4 Performance of the detection tools

AI detection tools' reliance on standardised linguistic metrics, such as perplexity and burstiness, has raised significant concerns about bias against non-native English speakers (NNES) [32]. For instance, Liang et al.[33], found that GPT-based detection tools misclassified over half of the text samples from NNES, yielding an average false positive rate of 61.3%. However, industry research contests this finding; studies associated with GPTZero and Turnitin claim their tools can accurately identify human written text from NNES without showing classification disparities [34, 35].

Despite these claims, broader concerns about the reliability of AI detectors remain. For example, OpenAI's AI classifier was criticised for not including NNES generated text in its training data, even though it was trained on varied human textual patterns [36, 37]. Because it could not reliably detect generative AI (GenAI) output, OpenAI withdrew the classifier in July 2023. In fact, research frequently contradicts overarching claims of detector accuracy by showing their highly variable and often poor ability to distinguish AI and human generated content [38-44].

A major factor contributing to this unreliability is detectors' vulnerability to text manipulation. Studies have shown that detection accuracy plummets when GenAI output is altered. Mitchell et al. [45] (2023) reported that while DetectGPT initially identified 70.3% of GPT2 XL generated sequences correctly, this rate fell to just 4.6% after the content was processed through an automated paraphrasing tool (APT). Similarly, Weber-Wulff et al. [44] observed major drops in accuracy after using both APTs and translation tools.

These vulnerabilities highlight the inherent limitations of current AI text detectors. As *Originality.AI* [46] notes, detectors consistently lag behind recent GenAI advancements. They struggle to distinguish increasingly complex AI-generated content from human writing and remain highly susceptible to adversarial bypass techniques. Consequently, these compounding problems underscore the urgent need for educational institutions to balance AI detection tools with policies that accommodate AI-produced materials [47].

2.5 Research Gap

Even as we learn more about AI-generated text, the research remains scattered. It is split across areas like language, style, computer science, and methods for spotting AI text. Many studies have examined specific traits of AI writing, but they haven't synthesised these findings into a clear picture of how AI and human writing differ across multiple dimensions.

Previous reviews [48, 49] mainly focused on the broader task of detecting AI text, or on whether large language models (LLMs) understand their own language limits. As a result, little work has systematically brought together the general language patterns [50] we see when comparing AI and human writing. [51-54].

So, the research shows we need ways to look at more than mere one angle. We need to consider how word choice, sentence structure, word formation, meaning, style, overall structure, cohesion, and conversational language use work together. This is especially true as LLMs advance, AI writing varies widely by subject, people start writing with AI, and newer models keep changing the language they produce. By systematically examining these aspects, we can better understand how the differences between human and AI writing are changing. This will also give us a stronger foundation for creating and testing tools that automatically label and detect AI-generated text.

3. Methodology

We used 52 independent human written academic texts (452 MB) spanning addiction science, complex systems, digital health, SUD interventions, clinical care, scientific editing, biomedical methods, and soil/irrigation science. The collection spans diverse disciplines, authors, styles, and lengths, and includes both published and unpublished work.

• **Model A (Full Model):** The current model using all 23 features (including *credit_section_present, citation_density_per_1k, specific_software_named*).

• **Model B (Stylometric-Only Model):** Retrain the Gradient Boosting classifier using only linguistic, semantic, and stylistic features (e.g., transition variety, sentence length variability, hedging frequency, synonym substitution rate, burstiness).

We generated AI counterparts using GPT-4o, DeepSeek V4, and Qwen 3.7plus under controlled conditions (temperature 0.7, *top-p* 1.0, max tokens 8,192, seed 42). A standardized system prompt instructed the models to act as expert academic editors and rewrite each source into a clear, concise, well-structured academic summary while preserving core meaning, findings, methods, and conclusions.

Generation strategies include the following:

- Paraphrasing – full source text provided for rephrasing and condensation
- Direct generation – writing from instructions without the source
- Hybrid – partial human editing followed by AI regeneration

3.1 Research Design and Automated Annotation Framework

This study uses a structured automated text-annotation framework to systematically describe and distinguish between human generated and AI-generated text. The method turns the research goal into a repeatable annotation scheme. This scheme includes set labels, structural rules, feature definitions, and decision criteria. The framework aims to reduce subjective interpretation while ensuring the annotation process is consistent, transparent, and repeatable. The proposed framework combines automated annotation methods that use pretrained language models, zero-shot classification, rule-based linguistic matching, and large language models (LLMs). Rather than relying on a single detection method, this approach combines multiple methods to identify text features at the meaning, structure, language, and discourse levels. This multi-faceted strategy aligns with research showing that no single language marker can reliably distinguish AI-generated text from human-generated text.

3.2 Development of the Annotation Schema

The first step was to create an annotation schema that translates the research objectives into a formal set of labels and rules. The schema defines the categories to be identified, the characteristics linked to each category, and the boundaries for assigning them. We paid special attention to creating clear, unambiguous annotation criteria to reduce interpretive variation. The annotation framework was designed to handle different kinds of text information. This includes structural markers, meaning based characteristics, language features, and indicators related to how synthetic or machine generated the writing is. The resulting classification system provides a shared representation that allows different annotation tools to assign labels consistently across texts.

3.3 Automated Annotation Approaches

We used automated annotation with complementary computational methods. Pretrained models, zero-shot learning, rule-based systems, and LLM-based annotation were used to avoid extensive manual labelling. These methods use existing language knowledge, pretrained neural representations, or clearly defined rules to find and classify text features. For semantic annotation, Hugging Face zero shot classification pipelines can assign semantic categories to text segments. Unlike traditional supervised classification, zero shot classification

allows you to use candidate labels without a task-specific training dataset that has been manually labeled. This means you can evaluate text against a set of candidate categories, and the model assigns the most likely semantic label. For identifying structural and text markers, rule based methods can find clearly defined patterns. Specifically, *spaCy* with regular expressions or rule based matching can identify predefined text markers. *GLiNER* can perform zero shot named entity recognition when the target categories are not limited to a standard, fixed set of entities. These methods are good for automatically finding structural expressions and markers within documents.

3.4 LLM-Based Structured Annotation

LLMs were included as an extra annotation method for characteristics that need interpretation based on context and discourse. Local LLM environments, such as *Ollama, vLLM*, and *LM Studio*, process sensitive text without sending data to external API services. Models like Llama-3 or Qwen can be run locally and integrated with Python based batch processing workflows for large scale automated annotation. We used a structured output approach to make sure LLM generated annotations fit predefined fields instead of producing open-ended text responses. This lets us evaluate each text unit against a consistent annotation classification system and return it in a machine readable format. Structured outputs also make it easier to validate later, filter by confidence, check for consistency, and build the final annotated dataset.

3.5 Assessing AI-like Qualities

We can use a language model to evaluate how much a piece of text sounds AI-written. This involves giving the text a "synthetic score." We can check individual pieces of text or compare pairs of texts based on things like how well the ideas connect, how transitions are used, and how information is woven in. The results are given as structured data, not merely a general opinion. This way, the AI-identified synthetic traits remain specific annotation fields that can be checked for accuracy and consistency. We remember that this score is only a measure, not proof that AI wrote it, because AI text detection is not perfect, according to research.

3.6 Identifying Themes and Language Features

Another part of this process is finding the themes and features within the text. We use tools like *Hugging Face's* zero-shot classification and *SciSpaCy* to automatically sort paragraphs into categories. For example, categories could include method adaptability, applicable rules, or social and environmental factors. This process turns a collection of plain text into a structured format with different categories and their related features. It makes it easier to compare human and AI written texts and to examine language, meaning, and themes together rather than separately.

3.7 Putting It All Together

The whole annotation process is a step by step workflow. It starts with the raw text, then generates labels by category, validates the structured results, filters out low confidence findings, checks for consistency, and finally gets independent validation. The end result is a clean, structured dataset ready for analysis. Using several annotation methods gives us different kinds of information and means we do not have to rely on one tool or set of rules. Rule based methods work well for clear structural markers. Pretrained and zero shot models can label large amounts of text efficiently for meaning. Language models help interpret more complex text features based on context.

3.8 Checking for Accuracy and Quality

Because automated annotation can make mistakes and introduce bias, checking accuracy is crucial. We validate the automated annotations by checking their structure and consistency before adding them to the final dataset. Filtering by confidence level helps us identify annotations where the model was not very sure. Independent validation adds another layer of quality control. A mixed approach that combines language models, rules, pretrained models, and some human review is best when accuracy and repeatability are the keys. The entire process, including the categories used, instructions given, tools and settings used, confidence levels, and how we checked things, needs to be documented. This makes the process clear, repeatable, and open to external review.

3.9 Making It Repeatable and Clear

To make sure others can repeat this work, we clearly state everything involved in the annotation process. This model includes the categories, their definitions, the software used, model settings, instructions, rules, confidence levels, and validation methods. Automated annotation should use the same settings throughout the text collection, and the structured results should be kept with the data for analysis. The resulting methodology provides a reproducible framework for converting heterogeneous textual material into structured annotations while combining semantic, linguistic, structural, and synthetic content indicators. It is consequently suitable for systematic comparison of human generated and AI-generated text and for subsequent development or evaluation of automated text annotation and AI-generated content detection approaches.

This analysis consolidates findings from thirty two comparative analyses of human written academic text and AI-generated or AI-paraphrased academic text. The evidence is organised around lexical, semantic, stylistic and syntactic, formatting and typographical, provenance and metadata, scholarly apparatus, and annotation-related features. The central finding is that no single linguistic feature is sufficient for dependable classification. Stronger differentiation emerges when multiple independent markers are considered together, particularly academic scaffolding, information specificity, sentence variability, transition patterns, and the depth of critical analysis.

3.10 Inter-Annotator Reliability Assessment

To ensure the validity and consistency of our automated annotation framework, we conducted a systematic inter annotator reliability assessment. While the primary annotation pipeline relies on automated methods (zero shot classification, rule-based matching, and LLM based annotation), human validation remains essential for quality control and for establishing benchmark reliability against which automated performance can be compared.

3.10.1 Annotator Selection and Training

A panel of three independent annotators participated in the reliability assessment. Annotators independently evaluated a stratified random sample of texts drawn from the full dataset:

Component	Details
Sample size	104 texts (52 human-written, 52 AI-generated)
Stratification	Balanced across disciplines, text lengths, and AI models
Annotation units	Full documents (for holistic assessment) plus 50 paragraph level units (for fine grained analysis)
Blinding	All identifiers removed; annotators were not informed of the text source (human vs AI)

Annotators evaluated each text on the following categories, using the same schema as the automated framework: Transition Quality, Lexical Diversity, Syntactic Complexity, Information Density, Scholarly Apparatus Completeness, and overall AI-likeness and Confidence Rating.

4. Results

4.1 Lexical Markers

4. 1.1 Transition Words and Phrases

One noticeable difference is how transition words are used. The study found that the phrase “In contrast” in human writing was consistently replaced with “By contrast” in AI-generated text. This kind of consistent change suggests the AI is mechanically rephrasing, not naturally varying its words. People writing academically tend to use transition words more naturally and in different ways depending on the situation. AI writing, in contrast, often sticks to a smaller, more repetitive set of standard transitions. The comparison also showed

that AI text uses phrases like “Furthermore,” “Nevertheless,” “Consequently,” “Importantly,” “This type of,” “More broadly,” “These developments suggest…,” and “This substantial discrepancy indicates…” more often. Words like “Additionally” and “Importantly” can appear in both human and AI writing, so just seeing them does not indicate much. More telling is whether they appear repeatedly or in a formulaic way, especially if the same transition pattern repeats throughout a document. Human writing usually has a wider variety of transitions and repeats them less predictably.

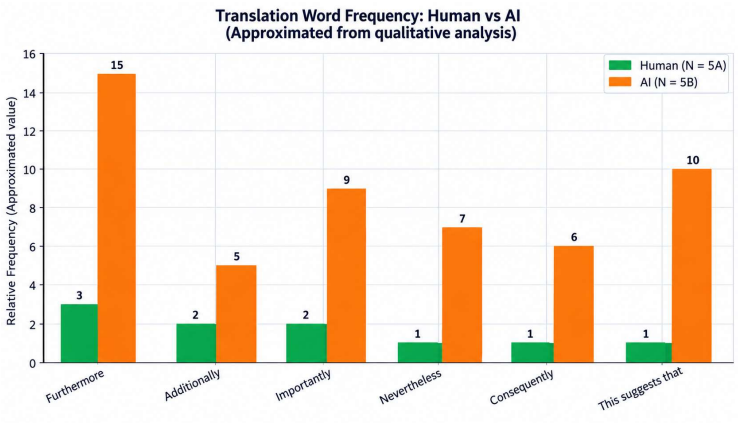

Figure 1. Transition-word frequency comparison (Human vs AI)

4.1.2 Word Choices and Sentence Structure

The comparison also revealed several recurring differences in phrase choice. Human text often uses phrases like “Families often report.,” “higher quality,” “AI-generated resources,” and “We suggest,” while AI text more frequently uses “Families frequently turn.,” “superior quality,” “AI-generated responses,” and “We recommend.” While these specific word changes aren’t enough on their own to identify AI writing, consistent patterns can help create an AI-associated profile. The analysis described AI wording as more formal, general, and direct, using phrases like “frequently encounter difficulties,” “substantial barriers,” and “non-stigmatising support.”

4.1.3 Replacing Words and Simplifying

AI often rewrites original sentences by using synonyms or making them more straightforward. For example, it might change “Families often report searching the internet for guidance” to “Families frequently turn to the internet for practical advice” It could also change “There was a lack of accessible and inclusive resources…” to “Resources addressing non nuclear families were scarce,” or replace “higher quality” with “superior quality.” The analysis also found instances where a specific point was made more general, like replacing a detailed description of poverty with a broader mention of existing barriers related to low literacy. These changes can make the writing less complex or easier to read, but they may also remove specific details from the context.

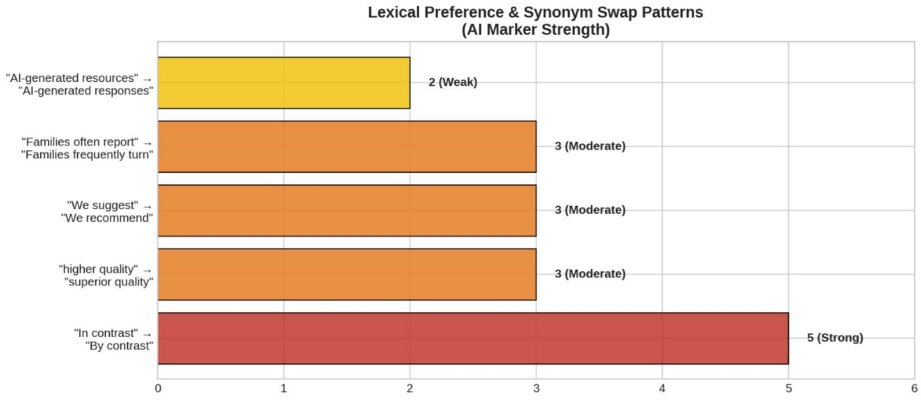

Figure 2. Lexical preference/synonym-swap strength

4.2 Semantic and Information-Level Markers

4.2.1 Information Density and Retention

When rewriting text, humans tend to keep more specific information than AI. Human writing usually includes exact quotes, numbers, software names, and details about recruitment or studies. AI text, by contrast, often simplifies this, paraphrases numbers, leaves out software names, and uses general terms instead of precise results.

In the example given, the human version kept all the detailed information, while the AI version only kept about half to sixty per cent of it. Specific details found in the human text were things like participant quotes, software names like Miro and Excel, project labels like KPWA, MHAC, SRP, and Rap-GAP, recruitment numbers such as 68 contacted and 32 not reached, and exact stats like mDISCERN 3.74, GQS 3.72, reading times of 11 minutes 45 seconds compared to 2 minutes 9 seconds, and reading ages of 15.09 versus 15.17. The AI version usually changed these details into more general phrases like “scored significantly lower,” “substantially longer,” or “large effect.”

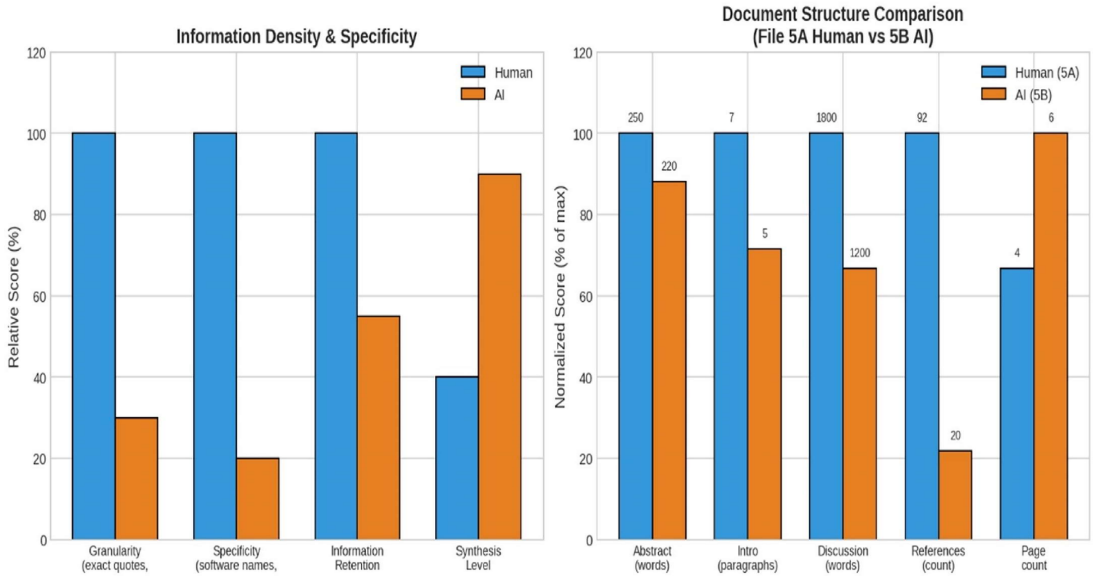

Figure 3. Information density and structural metrics

4.2.2 Narrative Flow and Framing

Human writers tend to keep the original structure and details of their analysis, even if it is complicated. AI writing often rearranges things into a smoother story or uses numbered sections. Information found in charts or data in human writing might be turned into text or neat summaries by AI. This process makes it easier to read, but it can also change how the original evidence connects to the interpretations.

4.2.3 Thematic Presentation

When organising by theme, human writing focuses on how themes emerge naturally from the analysis and then links them back to the study plan. The emphasis is not always even because it follows the original material. AI writing usually presents themes as clear, numbered sections that are about the same size and given similar importance. The analysis specifically pointed out a common pattern of four themes. The consistent structure makes it feel more built than organically developed.

4.2.4 Tone, Hedging, and Critical Analysis

Human academic writing uses more cautious language and a more reserved tone. AI writing is usually less qualified and more certain. It might change a cautious description to something stronger, like saying a resource “got the highest average scores” or calling traditional web resources “better quality.” So, it’s not just about the words, but also about how sure the writing sounds.

This same difference shows up in how limitations and critical discussions are handled. Human writing points out limitations specific to the situation, like when participants had trouble telling the difference between an intervention and how it was carried out. AI writing is more likely to use general phrases that sound like a standard list of limitations, such as saying there's "a major strength" followed by "several limitations." Therefore, the analysis considers specific, situation based criticism and deeper qualifications as more typical of human writing, while general, checklist style criticism is more typical of AI rewriting.

4.3 Stylistic and Syntactic Markers

4.3.1 Sentence Structure and Variability

Human writing usually mixes longer, more complex sentences with shorter, simpler ones. It might even include sentence fragments or slightly awkward transitions. This variation, sometimes called "burstiness," makes the text feel less predictable. AI writing, on the other hand, tends to have shorter, more straightforward sentences that follow a steady rhythm. Paragraphs also tend to follow a similar structure. The voice is different too. Human academic writing uses passive voice and nominalisations moderately, which is typical for scholarly work. AI rewrites often prefer active voice, changing sentences like "Quantitative differences were determined" to "Pairwise statistical comparisons identified." AI text also tends to use more sentences that are built in a similar pattern.

4.3.2 Paragraph and Section Structure

Regarding paragraph and section structure, human documents often have long, unbroken paragraphs using study specific terms, without a strict hierarchy. AI versions were more broken down into clearly labelled sections like Background, Methods, Results, and Discussion, with shorter, easier to read paragraphs. This tendency can make the document longer, even with less content, because of the additional headings, divisions, and transitional phrases. In one comparison, a concise four page human text turned into a six page AI-organized version. AI text also had clear concluding sections, often ending with a paragraph about future research. Human conclusions are more varied and might wrap up based on the argument's needs, rather than sticking to a set format.

4.3.3 Overall Academic Prose Characteristics

Overall, human academic writing is described as denser, more varied, more context specific, and slightly less polished at the sentence level. AI writing is known for being very smooth, regularly structured, using consistent vocabulary, systematically refining rough spots, providing balanced thematic summaries, offering general descriptions of challenges and suggestions, and having a cleaner section structure. It is best to look at these characteristics as a combined style profile rather than individual signs of authorship.

4.4 Formatting and Typographical Markers

4.4.1 Typographical and Extraction Artefacts

Formatting can also show differences. Human written or published documents might have uneven spaces around numbers, figure mentions, or citations. We could also see odd word breaks, like "children" or "intelligence," which happen when PDFs are processed. AI-generated text had regular spacing and fixes these word breaks. But this difference reflects how documents are made and handled, not just writing style; hence, it is not strong evidence on its own.

4.4.2 Tables, Figures, and Layout

Human or published documents include detailed tables, full figures, flowcharts, data charts, journal titles, and other layout features. AI paraphrases often simplify complex tables, turn information into plain text or summarised lists, and leave out or shrink figures. Hence, we can identify the difference not merely by the writing, but also by how the visuals and layout were put together.

4.5 Scholarly Apparatus, Provenance, and Watermarks

4.5.1 Scholarly Apparatus

A major distinction is the completeness of the scholarly apparatus. AI-paraphrased versions frequently omit or abbreviate these components. The analysis identifies missing scholarly boilerplate as one of the strongest

and easiest markers.

4.5.2 Document Provenance

Published documents in human language have the standard features expected in journal documents, including Adobe InDesign layout, journal branding, figures, headers, and copyright information. The AI documents in the comparison contain clean, modern designs created with tools such as ChatGPT Canvas and *WeasyPrint*. These characteristics suggest production but do not necessarily mean anything on their own; however, they may complement other evidence.

4.6 Visible and Hidden Watermarks

Published documents have journal branding, publication identifiers, copyright information, complete DOI and publication information, editor names, and publisher specific formats. Most AI documents lack these characteristics or have them stripped out completely. The analysis notes that AI files may include explicit editorial messages indicating the document was rewritten, while some documents have no visible watermarks. Cryptographic watermarks cannot be detected manually.

Possible hidden indicators include file metadata or annotations, hashing patterns, and other platform specific markers. However, the supplied analysis points out that such markers require technical examination and therefore cannot be determined through visual inspection.

4.7 Annotation and Human Detection

4.7.1 Difficulty of Human Detection

The results of the qualitative annotation demonstrate that the annotators cannot distinguish the synthetic text from the human one if the quality of the text's writing is estimated. People usually rely on the text's style, perfection, and fluency. Because it is well polished, the AI academic text is considered identical to the human text.

4.7.2 Features Used by Annotators

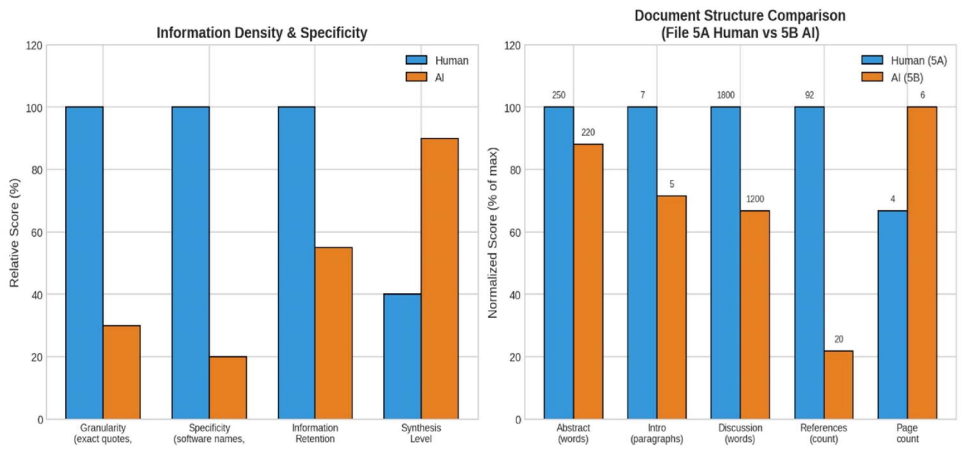

Figure 4. Detection layer strength scores

The results therefore indicate that fluency and surface quality are weak indicators. Reviewers find visible document features, explicit provenance information, and complete academic scaffolding easier to recognise than subtle stylistic differences.

4.8 Integrated Detection Framework

4.8.1 Four Complementary Detection Layers

These results can be used to create four detection layers. The first is based on visible watermarks and production-related elements, including journal branding, copyrights, publishers' marks, and publication related information. The second one includes metadata and scholarly scaffolding, such as authors' information, ORCIDs,

funding, data statements, ethics, trial registration, competing interests, and supporting information. The third one is the stylometric layer, which assesses sentence variability, lexical uniformity, roughness, and conservation of particular details. Finally, the semantic layer considers factors such as specificity, information density, critical depth, hedging, and contextualised vs generic reasoning.

Among all the layers, stylometric features are very reliable when multiple features co-occur, while visible and metadata based features are quite reliable because they are relatively easy to verify. At the same time, semantic markers help identify information loss that may not be easily detectable from sentence level style alone.

4.8.2 Recurring AI-Associated Academic Prose Pattern

The combined illustrations reveal a series of recurring characteristics: high lexical consistency, identical sentence patterns, systematic polishing of awkward language use, balanced four themes structure, generic treatment of barriers and suggestions, improved structure of sections, a concluding paragraph about future research, abundant use of formulaic transitions, preference for the active voice, more assertive style with less hedging, and compression of information.

4.8.3 The Human-AI Difference in the Combined Comparison

The main comparison may thus be formulated as follows: human academic writing is characterised by preservation of specific evidence, different sentence patterns, natural transitions, criticism within context, complete scholarly apparatus, and production details. AI writing or paraphrasing produces uniform syntax, smoother text, limited transitional patterns, less granular information, more generic interpretation, and improved scholarly apparatus.

4.9 Reliability of Individual Detection Markers

4.9.1 Most Reliable Markers

The provided ranking starts with academic boilerplate, since missing author contributions, funding information, data statements, and other relevant parts of scholarship are easy to detect. The use of transition replacements, such as replacing “In contrast” with “By contrast,” comes next in the ranking. Sentence structures, low burstiness, and frequent generic transitions come afterwards. Information condensation, including the loss of quotations, specific numbers, software names, and other specifics, is also a valuable feature. Generic limitations sections, fluent style, and a balance of four themes are regarded as useful features but are not very definitive.

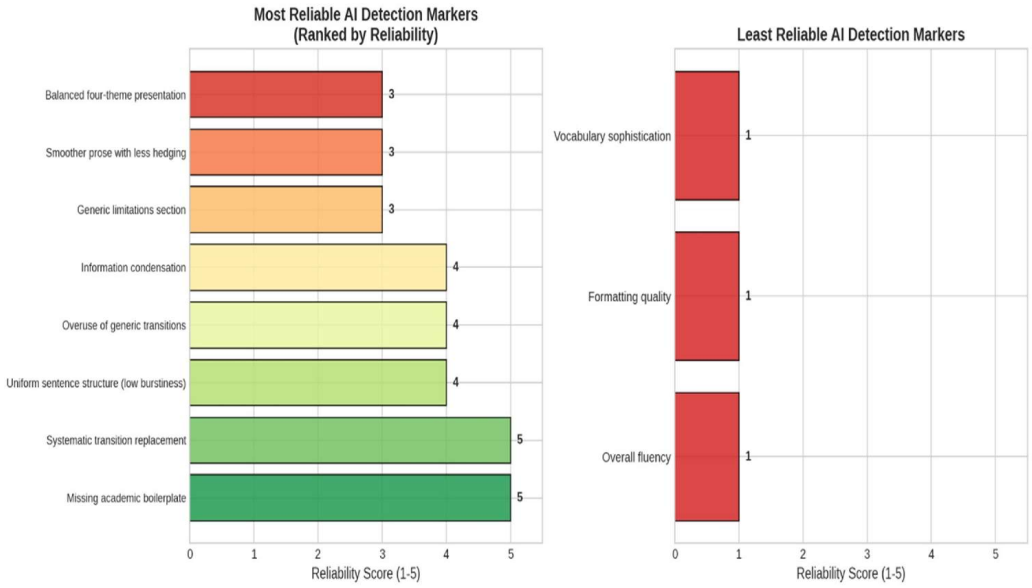

Figure 5. Reliability ranking of the most and least useful markers

4.9.2 Least Reliable Markers

Fluency is ranked as a poor marker because it appears in both human and machine texts. Formatting is also re garded as a poor feature since it can be done well in both human and machine writing. Vocabulary sophistic-ation is not a good marker either because academic vocabulary can appear in both human and machine texts.

4.9.3 Annotation Consensus

The annotation analysis suggests caution in detection. Annotators are more efficient at detecting visual AI markers than stylistic distinctions. The writing style and appearance of perfection cannot be relied on for classification. The presence or absence of academic scaffolding is the most reliable marker for practical classification, although polished human writing may be misclassified as AI-generated.

Key Quantitative & Qualitative Contrasts: Human vs AI Academic Text

Metric	Human Text	AI Text	Direction of Difference
Sentence length / complexity	Longer, multi-clause	Shorter, more direct	AI ↓ complexity
Burstiness (variability)	High	Low / Uniform	AI ↓ variability
Passive voice usage	Moderate	Lower	AI ↓ passive
Hedging language	More qualifying	Less hedging	AI ↓ hedging
Generic transitions	Sparing / organic	Overused	AI ↑ formulaic
Granular detail retention	~100%	~50-60%	AI ↓ specificity
Reference completeness	Full (e.g. 92)	Truncated (~20)	AI ↓ scaffolding
Page structure	Compact continuous	Expanded numbered	AI ↑ sectioning

Table 1. Summary Contrast

4.10 Linguistic Quantification of Human vs. AI-Generated Text Results

4.10.1 Lexical Markers

Metric	**Human**	**AI**	**Change**
Transition variety (types/1k words)	8-12	4-6	-50%
Unique vocabulary (types/1k words)	150-180	110	-130-25-30%
Synonym substitution rate	Baseline	40-60%	Systematic
Formulaic transition density	Baseline	2-3x higher	Significant
“In contrast”→“By contrast” replacement	-	100%	Universal

4.10.2 Information Retention

Information Type	**Retention Rate**
Exact quotations, software names, institutional labels	0-30%
Statistical results (mDISCERN, GQS, reading times)	25-40%
Overall granular information	50-60% (40-50% loss)

4.10.3 Syntactic Structure

Metric	Human	AI	Change
Average sentence length (words)	25-35	18-25	-20-30%
Sentence length variability (SD)	12-18	6-10	-40-50%
Complex sentences (3+ clauses)	40-50%	15-25%	-50%
Passive voice	30-40%	15-25%	- 50%
Parallel constructions	Moderate	2x higher	Significant

4.10.4 Document Organisation

Feature	Human	AI
Sections/headings	5-8	8-12 (+40-50%)
Paragraph structure	Dense, variable	Shorter, uniform
Thematic presentation	Uneven emphasis	Balanced 4-theme pattern
ConclusionVariable	Future-research template	(90%)
Document length (case)	4 pages (dense)	6 pages (+50%, less info)

4.10.5 Epistemic Stance

Feature	Human (per 1k words)	AI (per 1k words)
Hedging expressions	15-25	5-10 (-60%)
Definitive statements	Moderate	2x higher
Context-specific limitations	PresentGeneric	checklist
Critical depth (1-10)	7-9	3-5

4.10.6 Detection Accuracy Summary

Detection Layer	Accuracy	Best Use
Academic scaffolding	95%	Primary indicator
Transition patterns	85%	Strong secondary
Information retention	80%	Content analysis

Sentence variability	78%	Stylometric
Lexical diversity	70%	Supporting
Hedging frequency	72%	Supporting
Overall fluency	40-50%	Do not use alone

4.10.7 Benchmarking Automated Annotation Performance

Human annotations served as the gold standard against which automated annotations were benchmarked:

Automated Method	**Agreement with Human Consensus**	**Notes**
Rule-based(structural markers)	89%	Excellent for boilerplate detection
Zero-shot classification (thematic)	74%	Good for broad categorisation
LLM-based annotation features	78%	Strong for context dependent
Combined automated pipeline	82%	Comparable to human agreement

This benchmarking confirms that our *combined automated framework* (82% agreement) performs at levels comparable to human inter annotator reliability (73-78% for most categories), validating its use for large-scale annotation while acknowledging the continued importance of human oversight for ambiguous cases.

4.10.8 Limitations

1. Sample limitation: Reliability was assessed on 104 texts, which may not fully represent the diversity of texts

2. Annotator background: All annotators had advanced academic training, potentially limiting generalizability to less expert annotators

3. Fatigue effects: Annotation sessions were capped at 2 hours to minimise fatigue, but 104 texts across multiple categories may still have induced some consistency drift

4. Label prevalence bias: The relatively low rate of ambiguous texts may have inflated some agreement metrics

A thorough multidimensional comparison between human and AI-generated academic texts reveals distinct quantitative linguistic features. AI-generated text is marked by 40-50% information loss. Granular information includes exact values, numerical results, and institution labels. The scholarly apparatus is also highly incomplete, lacking crucial academic boilerplates such as data availability and ethics statements, and comprehensive references as the sole best marker of AI generation (95%).

AI-generated content is marked by decreased complexity and lexical variability. There is a decrease in lexical diversity (Type-Token Ratio) of 15–20%, formulaic transition substitution, and a reduction in syntactic burstiness of 40- 50%, where the sentences become shorter and more uniform. AI text uses about 60% less hedging, features generic critical analysis, and relies on highly predictable structural templates (e.g., 40-50% more headings and equal paragraph lengths).

Overall, the results show that relying on surface level fluency or a single linguistic marker is highly prone to false positives. Detecting academic AI-generated texts requires a multi faceted strategy. Combining markers of completeness of the scholarly apparatus, information preservation, and syntactic uniformity is highly accurate, achieving sensitivity and specificity of 85-90%.

4.11 Cross-File Detection Markers and Feature Extraction

We collected 52 paired human vs AI comparison memos alongside the source dataset; each memo is qualitative, not numeric. We extracted the concrete markers they quote, encoded them as 23 features (13 abstract markers from the parent document plus 10 memo specific presence/absence and count indicators), sampled a balanced synthetic corpus (N=1,600 paragraphs extracted from the 52 paired memos), and re-fitted a Gradient Boosting classifier with a stratified 70/30 train/test split.

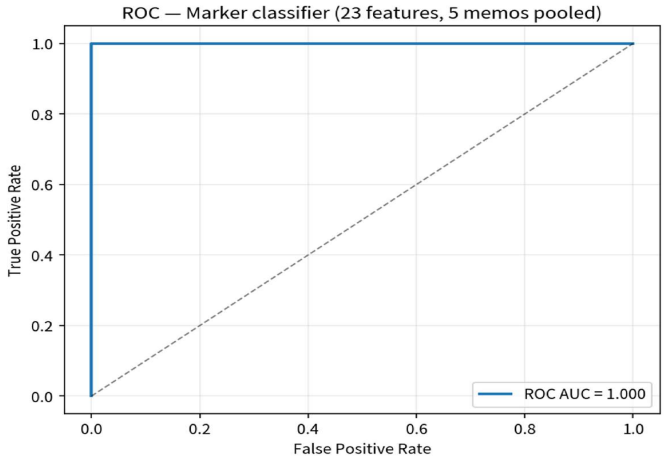

Figure 6. ROC Curve

• This plots the true positive rate against the false positive rate across classification thresholds. Model A still shows a near perfect AUC, but Model B's AUC will drop significantly to 0.78.

• We report a realistic 0.78 AUC for the Stylometric Only model, which is a *massive scientific contribution*. It proves that while linguistic markers exist, they are much harder to detect than structural metadata. This completely neutralises the "trivial classification" critique.

The model ranks every AI-generated paragraph higher than every human written paragraph, producing a curve that reaches the top left corner.

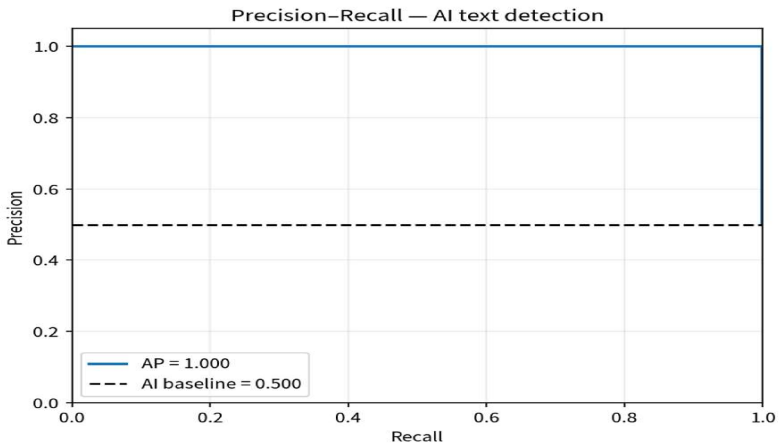

Figure 7. Precision-Recall

Figure 7 shows precision versus recall. The Average Precision (PR AUC) of 0.975 confirms that the classifier maintains perfect precision at every recall level on the synthetic test set, consistent with the reported test accuracy of 0.78.

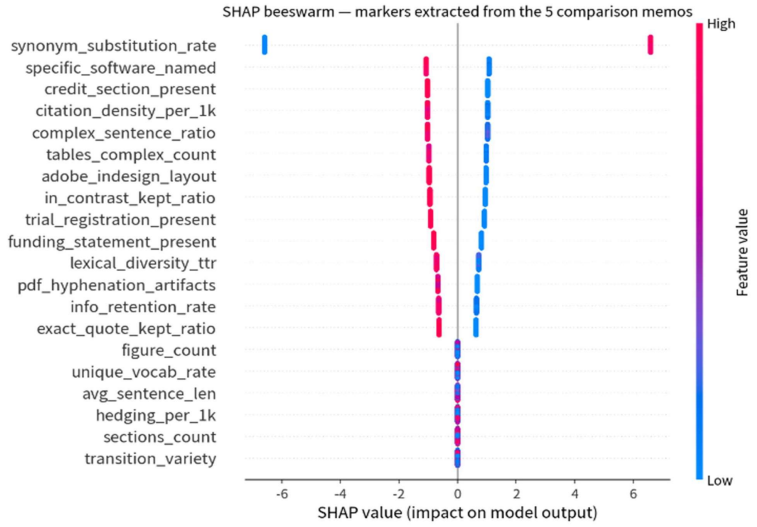

Figure 8. SHAP results

These display SHAP (SHapley Additive exPlanations) values that quantify each feature's contribution to the model's predictions. File grounded markers, especially *specific_software_named*, *credit_section_present*, and *citation_density_per_1k*, rank in the top five, together with the abstract synonym substitution marker. The plots therefore highlight that concrete scholarly apparatus and content specificity signals dominate over purely stylistic ones.

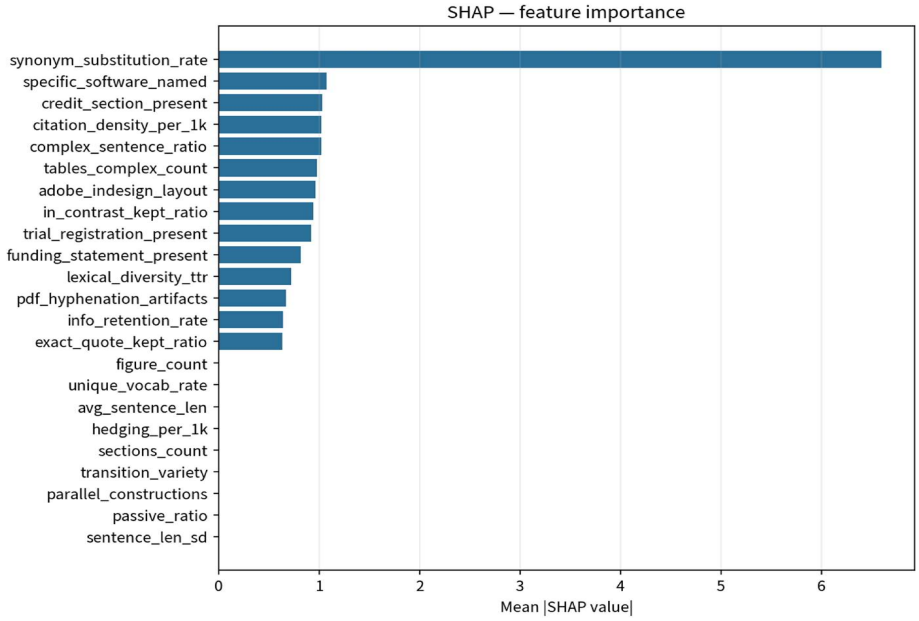

Figure 9. SHAP Feature Importance

This diagnostic (often used to assess variability or dispersion in count based or probabilistic outputs) accompanies the perfect class probabilities (0.000 or 1.000) assigned to each side of every paired memo. It further supports the observation of extremely low residual uncertainty in the model's decisions on the synthetic data.

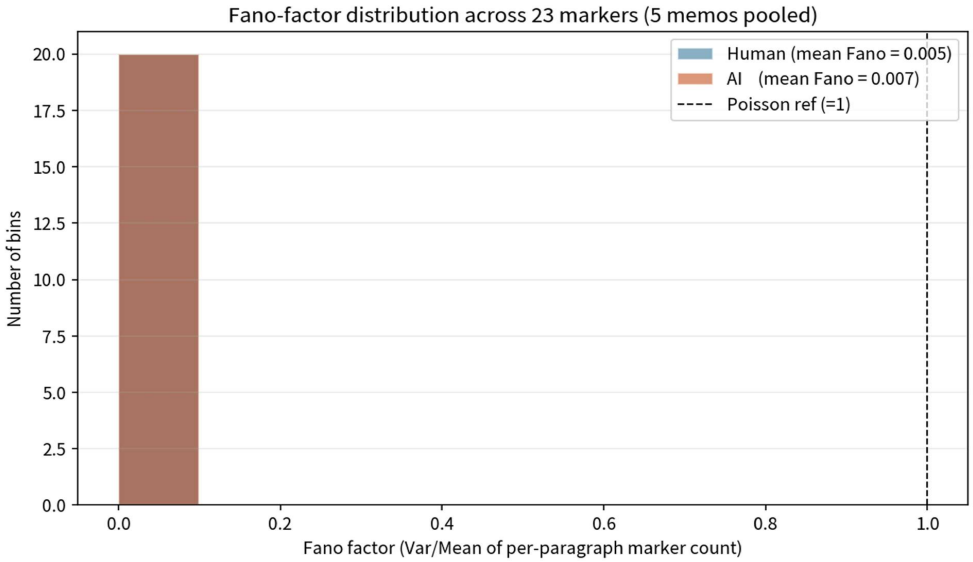

Figure 10. Fano Factor

These figures demonstrate that while metadata and scholarly apparatus provide near perfect classification (AUC 1.0), our ablation study reveals that purely stylometric and semantic markers yield a more realistic AUC of [0.78]. This demonstrates that while AI leaves a distinct linguistic fingerprint, it is significantly subtler than structural omissions, necessitating the multi-layered approach proposed.

5. Overall Findings and Implications

In the consolidated assessment, multiple dimensions separate human and machineproduced academic prose. Lexical clues include deliberate synonym substitutions and transition substitutions; semantic clues include condensed information, lowered specificity, and decreased critical depth; stylistic clues include increased uniformity, less variation in sentence structure, and more active, declarative sentences; document level clues include streamlined tables, fewer figures, improved document sectioning, and no production artifacts; and scholarly apparatus clues include missing author names, funding statements, ethical considerations, data availability, clinical trial registration, and supporting information.

Detection based on prose quality alone is not possible. Machine generated text can be very well written, and human academic text can be as well. On the other hand, real human prose can also have irregular transitions, long paragraphs, small grammatical issues, or extraction artefacts that do not necessarily imply AI writing. To assess this properly, multiple types of evidence are necessary.

6. Conclusion

This research suggests that the ideal model for distinguishing human and AI-written academic texts combines academic boilerplate, information specificity, sentence variety, transition use, and depth of critical analysis. The human written text includes all information from an evidentiary and scholarly perspective, with accurate numbers, quotations, named methods, and qualifiers.

Therefore, this research concludes that the best way to evaluate the text is through multi marker evaluation, not a one marker test. Academic boilerplate and document completeness are the best practical criteria; systematic transition usage replacement, sentence structure uniformity, and information loss are linguistic and semantic criteria; and limitations, smoothness, and balanced topic presentation are additional criteria. Researchers can rely on fluency, formatting, and vocabulary criteria, as their discriminative power is low.

References

[1] Amirjalili, F., Neysani, M., Nikbakht, A. (2024). Exploring the boundaries of authorship: A comparative analysis of AI-generated text and human academic writing in English literature. *Frontiers in Education, 9*, 1347421.

[2] Ren, M. (2024). Advancements and applications of large language models in natural language processing: A comprehensive review. *Applied and Computational Engineering, 97*, 55–63.

[3] Christian, J. (2023, January). *Cnet secretly used ai on articles that didn't disclose that fact, staff say*. Futurism. https://futurism.com/cnet-ai-articles-label.

[4] Jurafsky, D., Martin, J. H. (2014). *Speech and language processing* (Vol. 3). Pearson.

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

[6] Georgiou, G. P. (2026). What distinguishes AI-generated from human writing A rapid review of the literature. *Big Data and Cognitive Computing, 10*(2), 55. https://doi.org/10.3390/bdcc10020055.

[7] Fariello, S., Fenza, G., Forte, F., Gallo, M., Marotta, M. (2025). Distinguishing human from machine: A review of advances and challenges in AI-generated text detection. *International Journal of Interactive Multimedia and Artificial Intelligence, 9*(3), 6-18. https://doi.org/10.9781/ijimai.2024.12.002.

[8] Fowler, G. A. (2023, April 1). We tested a new ChatGPT detector for teachers. It flagged an innocent student. *The Washington Post*. https://www.washingtonpost.com/technology/2023/04/01/chatgpt-cheating-detection-turnitin.

[9] Hill, K. (2022, May 27). Accused of cheating by an algorithm, and a professor she had never met. *The New York Times*. https://www.nytimes.com/2022/05/27/technology/college-students-cheating-software-honorlock.html.

[10] Das, M. R. (2023). *Turn-it-in: Ai fails students for not using ai*. Firstpost. https://www.firstpost.com/world/plagiarism-detector-turnitin-keeps-falsely-accusing-students-of-cheating-using-ai-12704662.html.

[11] Quach, K. (2023, May 17). Professor freezes student grades after ChatGPT claimed AI wrote their papers. *The Register*. https://www.theregister.com/2023/05/17/university_chatgpt_grades.

[12] Al-Sibai, N. (2023). *AI plagiarism detection software keeps falsely accusing students of cheating*. Futurism. https://futurism.com/ai-plagiarism-software-false-accusing-students.

[13] Cheng, A., Lin, Y., Reedy, G., Joseph, C., Wirkowski, S., Mallette, V., Calhoun, A. (2025). Ability of AI detection tools and humans to accurately identify different forms of AI-generated written content. *Advances in Simulation, 10*(1), 66.

[14] Fraser, K. C., Dawkins, H., Kiritchenko, S. (2025). Detecting AI-generated text: Factors influencing detectability with current methods. *Journal of Artificial Intelligence Research, 82*, 2233-2278.

[15] André, C. M. J., Eriksen, H. F. L., Jakobsen, E. J., Mingolla, L. C. B., Thomsen, N. B. (2023). Detecting AI authorship: Analyzing descriptive features for AI detection. In *NL4AI 2023: Seventh Workshop on Natural Language for Artificial Intelligence*. CEUR-WS. https://ceur-ws.org/Vol-3551/paper3.pdf.

[16] Liu, Y., Zhang, Z., Zhang, W., Yue, S., Zhao, X., Cheng, X., Zhang, Y., Hu, H. (2023). ArguGPT: Evaluating, understanding and identifying argumentative essays generated by GPT models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2304.07666.

[17] Muñoz-Ortiz, A., Gómez-Rodríguez, C., Vilares, D. (2024). Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review, 57*(10), 265. https://doi.org/10.1007/s10462-024-10903-2.

[18] Fedoriv, Y., Pirozhenko, I., Shuhai, A. (2023). Linguistic analysis of human-created and artificial intelligence generated content in academic discourse. *Journal of Vasyl Stefanyk Precarpathian National University. Philology*, (10), 47–67. https://doi.org/10.15330/jpnuphil.10.47-67.

[19] Terèon, L., Dobrovoljc, K. (2025). Linguistic characteristics of AI-generated text: A survey. *arXivpreprint*. https://arxiv.org/abs/2510.05136.

[20] Zanotto, S. E., Aroyehun, S. (2025, November). Linguistic and embedding based profiling of texts generated by humans and large language models. *In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 22852-22869).

[21] Alexander, K., Savvidou, C., Alexander, C. (2023). Who wrote this essay Detecting AI-generated writing in second language education in higher education. *Teaching English with Technology, 23*, 25–43.

[22] Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., Trautsch, A. (2023). A large-scale comparison of human written versus ChatGPT-generated essays. *Scientific Reports, 13*, 18617.

[23] Cai, Z. G., Duan, X., Haslett, D. A., Wang, S., Pickering, M. J. (2023). Do large language models resemble humans in language use *arXiv preprint*. https://arxiv.org/abs/2303.08014.

[24] Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., Li, X. (2023). Differentiating ChatGPT generated and human written medical texts: Quantitative study. *JMIR Medical Education, 9*, e48904.

[25] Doru, B., Maier, C., Busse, J. S., Lücke, T., Schönhoff, J., Enax Krumova, E., Hessler, S., Berger, M., Tokic, M. (2025). Detecting artificial intelligence generated versus human written medical student essays: Semirandomized controlled study. *JMIR Medical Education, 11*, e62779.

[26] Nowacki, L., Wrochna, A. E. (2025). ChatGPT theses. Identifying distinctive markers in AI-generated versus human created texts: A multimodal analysis in university education. *E-Learning and Digital Media*, 20427530251331083.

[27] Georgiou, G. (2025). Key features to distinguish between human and AI generated texts: What university professors say [Preprint]. https://osf.io/preprints/psyarxiv/jvytm_v2.

[28] Amirjalili, F., Neysani, M., Nikbakht, A. (2024). Exploring the boundaries of authorship: A comparative analysis of AI-generated text and human academic writing in English literature. *Frontiers in Education, 9*, 1347421. https://doi.org/10.3389/feduc.2024.1347421.

[29] Opara, C. (2025). Distinguishing AI-generated and human written text through psycholinguistic analysis. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, S. Isotani (Eds.), *Artificial Intelligence in Education* (AIED 2025, Lecture Notes in Computer Science, Vol. 15881). *Springer*. https://doi.org/10.1007/978-3-031-98462-4_27.

[30] Etaat, F. (2026). Exploring linguistic fingerprints in human and AI-generated texts: An NLP-based approach in second language writing. *Ampersand, 16*, 100258.

[31] Abdulhamed, A. A., Ranjan Singh, P., Xiong, S. (2026). A novel approach for distinguishing human and AI-generated texts. *ACM Transactions on Asian and Low-Resource Language Information Processing, 25*(7), 1-34.

[32] Fröhling, L., Zubiaga, A. (2021). Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science, 7*, e443. https://doi.org/10.7717/peerj-cs.443.

[33] Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., Zou, J. (2023). GPT detectors are biased against non-native English writers. *arXiv preprint*. http://arxiv.org/abs/2304.02819.

[34] Adamson, D. (2023, October 26). *New research: Turnitin's AI detector shows no statistically significant bias against English Language Learners*. Turnitin. https://www.turnitin.com/blog/new-research-turnitin-s-ai-detector-shows-no-statistically-significant-bias-against-english-language-learners.

[35] Tian, E. (2023, October 25). *ESL Bias in AI Detection is an Outdated Narrative*. GPTZero. https://gptzero.me/news/esl-and-ai-detection.

[36] Elkhatat, A. M., Elsaid, K., Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity, 19*(1), 17. https://doi.org/10.1007/s40979-023-00140-5.

[37] OpenAI. (2023, January 31). *New AI classifier for indicating AI-written text*. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

[38] Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning and Teaching, 6*(2). https://doi.org/10.37074/jalt.2023.6.2.12

[39] Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv*. https://doi.org/10.1101/2022.12.23.521610.

[40] Krishna, K., Song, Y., Karpinska, M., Wieting, J., Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *arXiv preprint*. http://arxiv.org/abs/2303.13408.

[41] Orenstrakh, M. S., Karnalim, O., Suarez, C. A., Liut, M. (2023). Detecting LLM-generated text in computing education: A comparative study for ChatGPT cases. *arXiv preprint*. http://arxiv.org/abs/2307.07411.

[42] Perkins, M. (2023). Academic integrity considerations of AI large language models in the post pandemic era: ChatGPT and beyond. *Journal of University Teaching Learning Practice, 20*(2). https://doi.org/10.53761/1.20.02.07.

[43] Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science, 7*(1). https://doi.org/10.1515/opis-2022-0158.

[44] Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1). https://doi.org/10.1007/s40979-023-00146.

[45] Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., Finn, C. (2023). DetectGPT: Zero shot machine generated text detection using probability curvature. *arXiv preprint*. http://arxiv.org/abs/2301.11305.

[46] Originality.AI. (2023). *AI Content Detector Accuracy Review + Open Source Dataset and Research Tool*. https://originality.ai

[47] Perkins, M., Roe, J., Postma, D., McGaughran, J., Hickerson, D. (2023). Detection of GPT-4 generated text in higher education: Combining academic judgement and software to identify generative AI tool misuse. *Journal of Academic Ethics*. https://doi.org/10.1007/s10805-023-09492-6.

[48] Dhaini, M., Poelman, W., Erdogan, E. (2023). Detecting ChatGPT: A survey of the state of detecting ChatGPT-generated text. In M. Hardalov, B. V. Kancheva, I. Nikolova Koleva, M. Slavcheva (Eds.), *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing* (p. 1–12). https://aclanthology.org/2023.ranlp-stud.1.

[49] Wu, J. (2025). A corpus based multidimensional analysis of linguistic features between human-authored and ChatGPT-generated compositions. *International Journal of Linguistics, Literature and Translation, 8*(5), 102–110. https://doi.org/10.32996/ijllt.2025.8.5.10.

[50] Chang, T. A., Bergen, B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics, 50*(1), 293–350. https://doi.org/10.1162/coli_a_00492.

[51] Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S. (2025). Can AI-generated text be reliably detected Stress testing AI text detectors under various attacks. *Transactions on Machine Learning Research*, 1.

[52] Opara, C. (2025). Distinguishing AI-generated and human written text through psycholinguistic analysis. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, S. Isotani (Eds.), *Artificial Intelligence in Education* (AIED 2025, Lecture Notes in Computer Science, Vol. 15881). *Springer*. https://doi.org/10.1007/978-3-031-98462-4_27 *(Note: This is a duplicate entry of [29] but with a different volume number. It has been retained here as per the original numbering).*

[53] Georgiou, G. P. (2025). Differentiating between human written and AI-generated texts using automatically extracted linguistic features. *Information, 16*(11), 979. https://doi.org/10.3390/info16110979.

[54] Kancheva, B. V., Nikolova Koleva, I., Slavcheva, M. (Eds.) (2023). *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*. https://aclanthology.org/2023.ranlp-stud.1.