

Article

Key Features to Distinguish Between Human- and AI-Generated Texts: Perspectives from University Professors

Georgios P. Georgiou ^{1,2} 

¹ Department of Languages and Literature, University of Nicosia, Nicosia 2417, Cyprus; georgiou.georg@unic.ac.cy

² Phonetic Lab, University of Nicosia, Nicosia 2417, Cyprus

Abstract

This study provides direct evidence from university professors' experiences regarding the key features they use to identify artificial intelligence (AI)-generated texts and ranks these features by their perceived importance. The research was conducted in two phases. In Phase 1, online interviews were used to identify the most salient features professors reported using to detect AI-generated texts. In Phase 2, an online survey asked professors to rate the extent to which each identified feature contributes to the successful detection of AI-generated text. The interview data yielded seven features that professors reported using when they suspected a text was AI-generated. Survey ratings varied across features, with hallucinated facts or explanations, nonexistent sources, and the absence of language errors receiving the highest mean ratings in this sample. The use of difficult words received the lowest mean rating. These results have important pedagogical implications, as they can inform the development of more effective detection tools and guide the design of academic integrity policies and instructional strategies to address the challenges posed by AI-generated content.

Keywords: artificial intelligence; identification; features; texts; university professors

1. Introduction

The rapid evolution of large language models (LLMs) has endowed them with great capabilities, allowing them to create texts that are nearly indistinguishable from those created by humans (Amirjalili et al., 2024; Brown et al., 2020; Ren, 2024). Such abilities might be useful in fields such as marketing, education, and research articles, but they also create several challenges, as the boundaries between machine- and human-authored output may be hard to discern. Researchers have characterized this phenomenon as a “conundrum” (Rashidi et al., 2023) since determining the true author of a text, which denotes authentic communication, has become even more complex. A growing body of empirical work echoes this concern, showing that even trained readers often perform only modestly above chance when asked to discriminate artificial intelligence (AI)-written from human-written prose and that accuracy can vary widely by genre and task (e.g., news, essays, or scientific abstracts) (Fiedler & Döpke, 2025; Jakesch et al., 2023).

This challenge holds a significant and widespread impact. Within academic settings, the irresponsible use of AI conflicts with basic notions of authorship and integrity (Weber-Wulff et al., 2023). Moreover, the capability of LLMs to generate fake content can lead to the rapid dissemination of false information, which could weaken public trust and disrupt constructive public discourse (Kreps et al., 2022; Solaiman et al., 2019). Spotlight



Academic Editor: Savvas
A. Chatzichristofis

Received: 29 September 2025

Revised: 14 January 2026

Accepted: 21 January 2026

Published: 2 February 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

has moved towards developing automated technologies that can identify texts created by AI aimed at reducing these concerns. However, current detection algorithms show some limitations in terms of accuracy, credibility, and adaptability in various LLMs and text types (Weber-Wulff et al., 2023). They are often prone to errors, such as misclassifying human-authored texts as AI-generated, which may lead to false accusations and undermine the credibility of human authors (Rashidi et al., 2023). Furthermore, detection systems require regular adaptation and refinement to keep pace with new advances, as LLMs are rapidly evolving (Pudasaini et al., 2025). Technical proposals span two broad families: provenance approaches (e.g., watermarking) and model-based detectors. Watermarking seeks to embed statistically detectable signals in generated text (Kirchenbauer et al., 2023), while model-based detectors estimate whether a passage lies in typical probability regions for a model's generations (e.g., DetectGPT) (Mitchell et al., 2023). Yet both face practical limits. More specifically, paraphrasing, even by another LLM, can sharply erode detector performance; in stress tests, paraphrase attacks reduced a leading detector's accuracy from strong to near-random levels, although retrieval-based provenance methods can partly recover signal (Krishna et al., 2023).

Rather than treating detection as purely a technical problem, the present study approaches it as a human-centered judgment task, asking which features professors report using and which they consider most diagnostic when evaluating student texts.

1.1. AI vs. Human Texts

In light of this context, it is crucial to understand the capacity of humans to distinguish between human-authored and AI-generated texts. In doing so, individuals may employ a range of linguistic, stylistic, and contextual features. Recent research has examined these elements, reporting that features such as lexical diversity, syntactic complexity, the use of emotional or subjective language, perceived coherence, redundancy, and even subtle inconsistencies or noticeable signs may impact human assessment (Doru et al., 2025; Georgiou, 2025; Herbold et al., 2023; Jakesch et al., 2023; Jones et al., 2025). Empirical work, however, suggests that human detection performance is often only modestly above chance. For example, Jakesch et al. (2023) showed that people's judgments of AI-generated language are guided by intuitive but flawed heuristics, such as associating first-person pronouns or informal topics with human authorship. Similarly, Fiedler and Döpke (2025) found that both human evaluators and AI detectors correctly classified AI-generated versus human-written academic text only slightly better than chance, with no clear advantage for either, while other studies report sizeable variability across genres and proficiency levels (Boutadjine et al., 2025). These findings indicate that, although readers rely on salient features, these features do not always map cleanly onto true authorship. This mismatch suggests that what readers perceive as AI-like may reflect surface impressions rather than stable features of machine authorship.

Complementing these perception-based findings, a parallel line of research has focused on measurable linguistic differences between human and AI-generated writing. Large-scale comparisons of student essays show that ChatGPT-produced texts can match or even exceed human essays on holistic quality ratings, yet differ systematically in style—for instance, exhibiting more nominalizations, higher lexical diversity, and more rigid organizational patterns, but fewer discourse and epistemic features (Herbold et al., 2023; Mizumoto et al., 2024). Survey and profiling studies similarly report that AI-generated texts tend to display distinct distributions of syntactic complexity, morphological features, and readability indices, suggesting an emerging AI register with its own characteristic patterns (Georgiou, 2025; Terčon & Dobrovoljc, 2025; Zannotto & Aroyehun, 2025). Markowitz et al. (2024), for example, show that inherently false AI-generated statements

can look superficially fluent while still carrying subtle linguistic features that differentiate them from human-authored content. Georgiou (2025) argues that AI-generated text includes a significantly higher number of nouns, adjectives, pronouns, and adjectival/prepositional modifiers, among others, compared to human-written texts. Together, this work indicates that even when AI output appears natural at first glance, it often diverges from human writing along multiple, quantifiable linguistic dimensions.

Beyond surface-level stylistic features, a particularly distinctive characteristic of AI-generated text is the phenomenon commonly termed hallucination. In the context of LLMs, hallucination refers to the production of fluent, plausible-sounding content that is factually incorrect, unverifiable, or entirely fabricated, despite being presented with high confidence. Recent surveys characterize hallucinations as a core reliability problem for LLMs, noting that they can range from minor distortions of known facts to fully invented entities or events (Huang et al., 2025; Sahoo et al., 2024). In academic and professional writing, this often manifests as fabricated references, non-existent journals, or incorrect bibliographic details; empirical evaluations of ChatGPT and related systems have repeatedly documented high rates of inaccurate or wholly invented citations in scientific and clinical domains (Chelli et al., 2024; Yoshida et al., 2025). In educational settings, such hallucinations pose specific risks, as students and instructors may accept erroneous content or bogus sources at face value, thereby undermining academic integrity and eroding trust in AI-assisted learning (Danyaro et al., 2025). Meanwhile, the very presence of confidently stated but verifiably false claims or references can serve as a powerful diagnostic signal for human evaluators, differentiating AI output from typical student errors, which are more likely to involve partial misunderstandings than wholly fabricated facts (Spearing et al., 2025).

Alongside computational analyses, several studies have examined which concrete features human experts actually rely on when judging authorship. Doru et al. (2025) demonstrated that both medical professionals and humanities scholars could reliably identify ChatGPT-generated medical texts, basing their decisions primarily on linguistic and stylistic attributes such as redundancy, repetition, and coherence rather than on content familiarity. Other work suggests that exposure and feedback can modestly improve human discrimination, but that people still lean on unstable heuristics and may overestimate their own accuracy (Jakesch et al., 2023; Milička et al., 2025). Overall, the literature points to a tension between (a) the existence of detectable linguistic differences between AI and human texts and (b) the fallibility of the human judgments that attempt to exploit those differences. Yet we still know relatively little about how university professors, specifically, prioritize these features when they suspect AI-generated student writing. These strands suggest that educators may notice salient cues but may differ in which cues they trust, an issue this study examines directly.

1.2. This Study

Despite growing work on distinguishing AI-generated from human-written texts, three gaps remain. First, relatively few studies focus specifically on university professors as evaluators, even though they play a central role in academic integrity decisions. Second, prior research has tended to catalogue potential features of AI authorship without systematically examining how educators rate each feature. Third, there is limited mixed-methods evidence that combines qualitative elicitation of features with quantitative weighting of their perceived importance.

This study aims to identify the most important features of AI use in text generation based on the experiences of university professors and to evaluate the relative importance of each in the overall judgment process. Addressing the gaps outlined above, the study is guided by the following research questions:

1. Which linguistic and contextual features do university professors report using when distinguishing between human- and AI-generated academic texts?
2. How do university professors evaluate the relative importance of these features when judging whether a text was generated by AI?

This study is expected to contribute to the development of more transparent, human-centered approaches for the detection of AI in academic contexts by examining how professors recognize AI-generated content. The findings can also inform institutional policy and promote digital literacy within an educational environment increasingly shaped by AI technologies. In sum, mapping the judgment criteria of academic experts offers a pragmatic complement to rapidly evolving technical solutions: it can (a) clarify which linguistic and contextual features are actionable in real classrooms, (b) surface equity risks (e.g., detector bias against non-native writers) so institutions can mitigate harm, and (c) provide an evidence base for policies that balance academic integrity with legitimate, disclosed uses of generative tools. The present study operationalizes these goals by eliciting and weighting professors' features, offering a scalable, human-centered framework to augment, or critically constrain, the role of automated detectors in higher education.

The remainder of this paper is structured as follows. Section 2 describes the methodology, including the participants, instruments, and procedures for data collection and analysis. Section 3 presents the qualitative and quantitative results regarding the features professors use to distinguish human-written from AI-generated texts. Section 4 discusses these findings in relation to prior work on AI-generated writing and academic integrity. Section 5 summarizes the main conclusions and implications for policy and practice in higher education, as well as the study's limitations.

2. Methodology

2.1. Participants

Ten university professors ($n_{\text{females}} = 5$; $M_{\text{age}} = 48.2$, $SD = 5.42$) participated in the qualitative study, while 38 university professors ($n_{\text{females}} = 20$; $M_{\text{age}} = 46.05$, $SD = 6.92$) participated in the quantitative study. All participants were in-service professors permanently residing in Cyprus, holding positions across public and private universities. All of them were PhD holders and were teaching either undergraduate or postgraduate courses, or both. Recruitment was conducted through purposive and snowball sampling: initial invitations were distributed via email to professional networks, and additional participants were referred by colleagues. Professors were selected based on their teaching experience (minimum of five years) and prior exposure to suspected AI-generated student work. Participants represented a diverse range of disciplines, including humanities, social sciences, and STEM fields, thereby ensuring variability in perspectives. All participants confirmed that they had previously encountered assignments they suspected of being generated with ChatGPT or similar LLMs. Prior to participation, they completed an online informed consent form in accordance with ethical research standards.

2.2. Instruments

We used a sequential mixed-methods design: interviews to elicit features, followed by a survey to quantify their perceived diagnosticity.

Qualitative data were collected through semi-structured interviews conducted via secure video-conferencing platforms. Each interview lasted approximately 15–20 min and followed a guiding protocol. Core prompts included "How do you differentiate between human writing and AI-generated content?" and "Can you provide concrete examples of features you have noticed?" Probing questions were used to elicit clarifications and

detailed descriptions. All interviews were audio-recorded with participants’ consent and subsequently transcribed verbatim for analysis.

Quantitative data were collected through an online survey, administered after the completion of the qualitative phase. The survey presented professors with a list of linguistic and stylistic features derived from the thematic analysis of interview data. Participants were asked to rate the extent to which they believed each feature could help them identify AI-generated text, using a 7-point Likert scale ranging from 1 (extremely unlikely) to 7 (extremely likely). Demographic information, including academic discipline, years of teaching experience, and prior familiarity with ChatGPT/LLMs, was also collected to contextualize responses.

2.3. Data Analysis

Qualitative data were analyzed using a focused thematic analysis approach. Transcripts were first read in their entirety to ensure familiarity. Relevant excerpts in which professors described suspected AI-related features were highlighted and coded. Codes were grouped inductively into broader themes based on conceptual similarity, resulting in a structured set of common features. The analysis focused on cataloging linguistic and stylistic features without extending into broader interpretative frameworks (e.g., attitudes or perceptions). Reliability was supported through iterative coding and discussion between the authors until consensus was achieved.

Quantitative data analysis was conducted in R (R Core Team, 2025). The dependent variable was professors’ ratings of the usefulness of each feature (Rating), while the independent variable was the set of identified features, seven in total (Features). A one-way ANOVA was performed to examine whether mean ratings differed significantly across features. When significant main effects were found, post hoc pairwise comparisons were conducted using the Tukey HSD method to control for multiple comparisons. Figure 1 illustrates an overview of the study’s methodology.

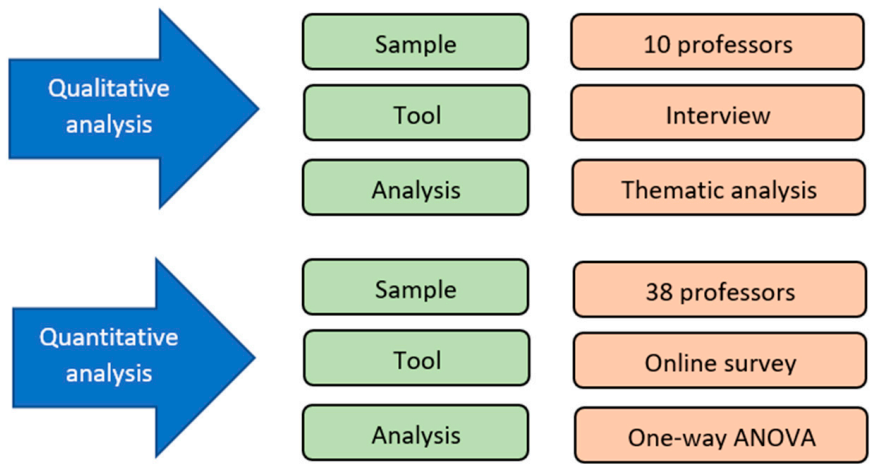


Figure 1. An overview of the study’s methodology.

3. Results

3.1. Qualitative Analysis

First, the seven features from the interviews are presented, and then professors’ ratings and statistical comparisons across features are reported.

The thematic analysis of the interview transcripts suggested seven recurrent features that professors commonly associate with AI-generated academic writing. These features reflect particular linguistic and stylistic patterns that diverge from typical student writing. Table 1 presents each feature along with the relevant explanation.

Table 1. Features useful for the identification of AI-generated texts based on professors' interviews.

Feature	Explanation
Absence of language errors	AI text is frequently grammatically perfect, unlike typical student writing.
Complex syntax/phrasing	AI often constructs sentences that are more complicated than needed, reducing clarity.
Difficult words	AI tends to use unnecessarily complex vocabulary to sound formal or academic.
Hallucinated Facts/Explanations	AI can generate false or unverifiable information presented confidently as facts.
Nonexistent sources	AI often fabricates citations that do not correspond to real sources.
Repetition of words/phrases	AI may unnaturally repeat the same words or phrases within short passages.
Specific text structure	AI-generated text can follow overly rigid or formulaic organizational patterns.

First, the absence of language errors was the most frequently cited feature. Professors were often suspicious of writing that contained no grammatical, typographic, or stylistic mistakes. As one participant noted, "Not a single comma out of place. That is not typical for undergraduates". Another remarked, "Even my best students make occasional errors; a weak one had none". Such texts were often perceived as unnaturally polished, particularly when the student had a documented history of language difficulties. Second, many interviewees pointed to complex syntax and phrasing as a red flag. Sentences featuring multiple subordinate clauses, nominalizations, and dense phrasal constructions were often considered inconsistent with a student's demonstrated ability. One professor observed, "The student barely spoke in class, yet their essay used nested clauses like a published academic". Another commented on "overly long sentences that seemed designed to sound impressive rather than convey ideas clearly".

The third theme involved the use of difficult words. Professors noticed the appearance of advanced or overly formal words, often deployed incorrectly or unnecessarily. Words such as "ameliorate", "epistemological", "juxtaposition", and "ontological" appeared in contexts where simpler or more precise terms would have sufficed. A participant shared, "It felt like the text was showing off, using impressive words that were not necessary". A particularly concerning theme was the presence of hallucinated facts or explanations. Several professors encountered confident-sounding yet false claims within assignments, such as attributing linguistic theories to the wrong scholars or inventing historical events. One participant explained, "It claimed a linguist had developed a theory I know for a fact they never did", while another remarked, "It described historical events that never happened". These errors were not typical misunderstandings but appeared entirely fabricated, despite being stated with academic authority. Closely related was the identification of nonexistent sources. Professors often attempted to verify citations and found that the referenced articles, authors, or journals did not exist. One stated, "The bibliography looked real, but when I searched for the articles, nothing came up." Another described discovering journals with plausible-sounding titles, only to realize they were fabricated. Common issues included fake DOIs, mismatched author-topic pairings, and incorrect publication years.

The sixth theme involved repetition of words or phrases, which many professors interpreted as a sign of artificial generation. Texts sometimes featured repeated sentence structures or discourse features in a way that felt mechanical. Examples included the repeated use of phrases like "It is important to note" or identical sentence openings such as "Another example is. . ." across multiple paragraphs. One participant commented, "The same sentence pattern repeated three times in a row, and it felt like it was filling space". Finally, professors noted specific text structure as another giveaway. Essays frequently adhered too closely to a formulaic structure, such as the classic five-paragraph format, with little variation in paragraph shape or rhetorical flow. One participant said, "It felt like a template since each paragraph started and ended the same way." Others described introductions and conclusions that mirrored each other almost verbatim, or arguments

that unfolded in overly neat ways, lacking the nuance and unpredictability of authentic student thinking.

3.2. Quantitative Analysis

To complement the qualitative insights, a follow-up survey was conducted in which professors were asked to evaluate the perceived strength of the seven features identified during the interview phase according to their perceived strength in signaling AI authorship. This step served to quantify the relative weight academics assign to each feature, moving beyond whether a feature is noticed at all to how strongly it influences evaluative judgments. By aggregating these ratings, the survey provided a clearer picture of which features professors view as most diagnostic, thereby translating individual observations into a collective assessment of detection priorities. Figure 2 illustrates professors' average rating of how likely each feature signals AI-generated text.

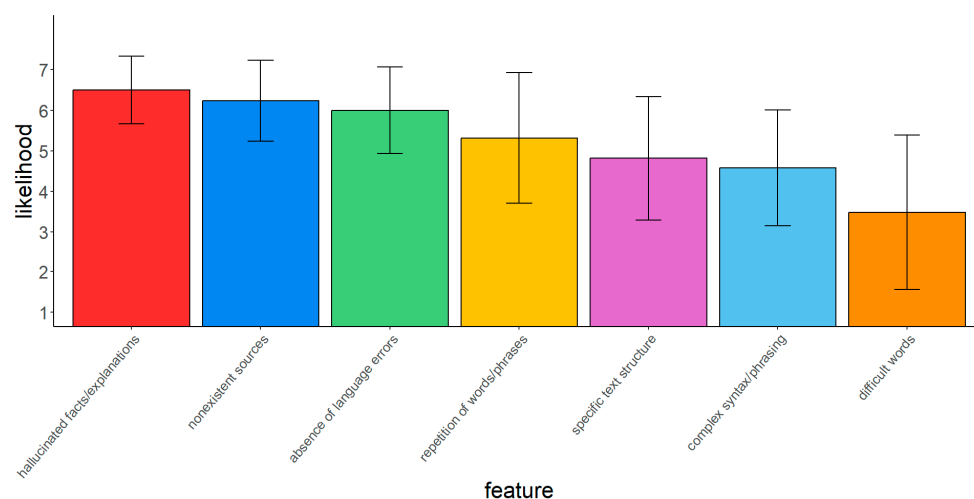


Figure 2. Likelihood that each feature constitutes a feature of AI-generated text.

To examine the perceived importance of each feature in detecting AI based on the evaluations of the professors, we used statistical analysis. The analysis showed a significant effect of Feature on Ratings [$F(6, 269) = 22.75, p < 0.001$]. The post hoc test indicated that mean ratings for hallucinated facts/explanations were significantly higher than ratings for repetition of words/phrases, specific text structure, complex syntax/phrasing, and difficult words. Nonexistent sources were rated significantly higher than specific text structure, complex syntax/phrasing, and difficult words; the absence of language errors displayed a similar pattern. Importantly, mean ratings for repetition of words/phrases did not significantly differ from specific text structure and complex syntax/phrasing, but they significantly differed from difficult words. Similarly, the ratings for specific text structure showed no significant difference with those of complex syntax/phrasing, but a notable difference with difficult words. Rating for complex syntax/phrasing also significantly differed from the ratings of the latter feature. Overall, the pattern of mean ratings suggests a gradient, where features related to factual verifiability received higher average ratings than stylistic features. This pattern implies that, within this sample, professors may prioritize features like hallucinations and source verification when forming suspicions about AI authorship, while placing less weight on stylistic elements such as difficult words. The results of the post hoc test are displayed in Figure 3.

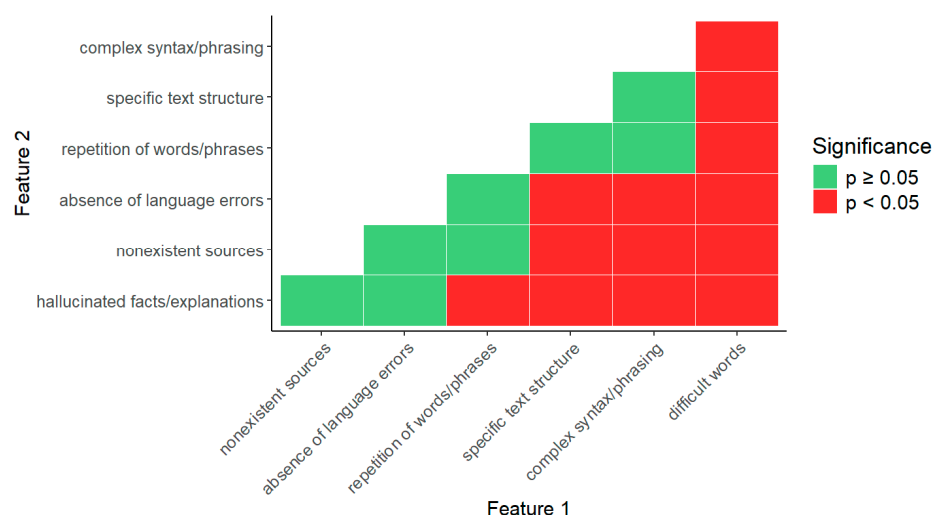


Figure 3. Significant differences between AI Indicators (Tukey post hoc).

4. Discussion

This study aimed to examine the linguistic and contextual features that university professors consider to distinguish between texts created by humans and those generated by AI in academic settings. The qualitative analysis indicated an array of features employed by professors during their assessments. These features include the absence of linguistic errors, the complexity of syntax and phrasing, the use of difficult vocabulary, the appearance of hallucinated text, the creation of fabricated sources, the repetition of words or phrases, and specific structural components of the text. Such features have been well-documented in previous work as markers indicative of AI-generated text (Doru et al., 2025; Mizumoto et al., 2024; Najjar et al., 2025). Recent synthesis papers similarly argue that these surface-level features are unstable because modern models can easily modulate style; therefore, educators should privilege verifiable content checks over stylistic diagnostics (Huang et al., 2025; Sahoo et al., 2024).

The qualitative data highlighted the presence of hallucinated facts or explanations as a frequently noted concern; this phenomenon refers to AI-generated material that conveys incorrect or unverifiable information with confidence and authority. This finding is consistent with prior research reporting factual inaccuracy as a persistent issue within the outputs of LLMs (Janéafik & Dusek, 2024; Kreps et al., 2022). In the quantitative survey, this feature received the highest average rating among participants, suggesting it may have a stronger influence on their evaluative judgments compared to other features, such as the use of complex syntax, repetition of words or phrases, or specific text structure. As this study shows, professors are especially sensitive to the veracity of content, and their disciplinary knowledge might be crucial in detecting AI-generated text in fields that require in-depth subject knowledge. Parallel evidence indicates that targeted verification workflows, such as structured fact logs, retrieval-augmented checking, and post hoc validation, can meaningfully reduce the risk of undetected hallucinations, supporting our emphasis on veracity over style. Large-scale surveys of LLM factuality concur that hallucinations persist across tasks and domains and recommend verification pipelines, including retrieval-augmented generation (RAG), claim decomposition, and post hoc fact checking, to mitigate risk (Sadi Rahman et al., 2025). Integrating RAG into domain workflows (e.g., biomedicine) improves answer veracity and auditability, aligning with our respondents' emphasis on content truthfulness (Li et al., 2025).

The second most highly rated feature was the creation of nonexistent sources, another known limitation of LLMs that often generate plausible-sounding but nonexistent

citations (Amirjalili et al., 2024; Salvagno et al., 2023). The fact that both hallucinated facts/explanations and nonexistent sources were rated high by professors as potential features of AI-generated text suggests that professors are sensitive to verifiability and the academic integrity of content. These features go beyond stylistic traits, pointing out whether the text adheres to core scholarly principles. Interestingly, professors considered the absence of language errors as a strong feature of AI authorship. This is because student writing, even at the university level, often contains minor errors, stylistic inconsistencies, or informal expressions that make a perfectly polished text suspect. This finding resonates with the notion that “too perfect” can itself be a red flag (Showemimo, 2025), echoing work by Alexander et al. (2023), who found that high levels of technical and grammatical accuracy and sophisticated language use were features of AI-generated text for English lecturers. Multiple audits across disciplines confirm that LLMs frequently produce fabricated or misattributed references that appear syntactically correct but are unverifiable, underscoring the need for routine citation verification in assessment (Zhang & Zhao, 2025). User-facing ecosystem studies also report fabricated information as a salient subtype of hallucination encountered by non-expert evaluators, reinforcing faculty perceptions documented here (Massenon et al., 2025).

Features such as repetition or specific text structure received lower average ratings. This suggests that, within the context of this survey, professors reported relying less on these stylistic features when forming an overall judgment. Notably, the application of difficult words has been recognized as one of the least effective diagnostic features for assessing written texts. Empirical research indicates that AI frequently utilizes a more advanced lexicon compared to human authors (Georgiou, 2025). This observation engenders the assumption among educators that university students possess the capability to produce written works characterized by sophisticated vocabulary. Furthermore, instructors are likely aware of the ability of AI to adapt and generate text employing more accessible language when appropriate (e.g., Sastre et al., 2024). Complementarily, evaluations of AI-output detectors show inconsistent performance and susceptibility to both false negatives (well-edited AI text) and false positives (certain L2 writers), suggesting that stylometric signals should not be used in isolation for high-stakes decisions (Erol et al., 2025).

The results of this study may have significant implications for educational institutions struggling with the challenges of generative AI. The development of detection policies that insist on human decision-making instead of completely relying on automated systems is important, as research has shown that automatic methods may be inadequate alone (Pan et al., 2024). In addition, professors should receive relevant training to identify academic misuse more easily and engage students in the moral use of AI technologies (Bibi et al., 2024). Finally, universities could abandon a potential punitive stance against AI by fostering digital literacy and promoting transparency to achieve better results; this could empower students to use generative tools responsibly and ethically. Expanding on these points, institutions should formally adopt a human-in-the-loop approach to any AI-writing detection, given persistent false positives, disparate impacts on multilingual writers, and unsettled accuracy claims across tools. Guidance from teaching and ed-tech units cautions that detectors can mislabel legitimate work and erode trust; faculty should instead pair any flags with conversational review of drafts and process evidence (e.g., notes, version history, or orals). Alongside this, staff development should emphasize AI literacy (conceptual knowledge, critical evaluation, and ethical use) so instructors can coach students on prompt transparency, citation of AI assistance, data privacy, and bias. Emerging frameworks in higher education provide practical scaffolds for curriculum and policy design (Kassorla et al., 2024). Policy should also pivot from prohibition to assessment redesign that makes productive, accountable AI use visible. Current scholarship

recommends stress-testing tasks, requiring reasoning traces or creation processes, and integrating oral or authentic assessments rather than defaulting to detection. Frameworks such as the AI Assessment Scale and sector advice echo this shift, especially as student AI use has become widespread. Finally, technical studies demonstrate why over-reliance on detectors is risky: distinguishing human from machine text remains non-trivial and error-prone, while over-reliance on dialogue systems can impair critical thinking, further reason to foreground pedagogy, not policing (Abdali et al., 2024; Zhai et al., 2024).

5. Conclusions

5.1. General Conclusions

The evolving presence of generative AI in academic contexts demands a better understanding of how human evaluators detect AI-generated content. This study contributes to that understanding by shedding light on the features university professors rely on when faced with the task of distinguishing human from machine authorship. We adopted a straightforward approach by identifying the key features used for AI detection and examining how professors assess the relative importance of each feature in their evaluations. The findings point to a pattern where, based on professor reports, the features perceived as most salient for raising suspicion are content-based—specifically, the presence of hallucinated facts/explanations and nonexistent sources—rather than purely stylistic or linguistic features. So, even in an era of technical fluency, factual precision, and the imperfections of human writing remain key features of authenticity.

These insights have direct implications for higher education. They suggest that the ability to verify information and cross-check sources could be considered a potentially important component of academic integrity practices in the age of AI, based on the perceived diagnosticity reported by educators in this study. Faculty may need to formalize verification strategies in their evaluation routines, while institutions could develop guidelines or training programs to help educators identify hallucinations more efficiently. At the same time, as generative AI systems improve, the boundary between tell-tale signs of machine authorship and legitimate student writing will likely blur, placing greater pressure on evaluators to combine textual scrutiny with contextual knowledge of students' prior work and performance.

5.2. Limitations and Future Work

Several limitations of the present study should be acknowledged. The relatively small sample, drawn entirely from universities in Cyprus, constrains the generalizability of the findings. Differences in academic conventions, linguistic norms, and institutional policies across regions or cultures may shape how educators interpret potential features of AI authorship. Likewise, although participants represented multiple disciplines, the sample size did not allow systematic comparisons, and the features identified here may not hold equally across fields with distinct writing traditions and epistemic norms. A further limitation is the reliance on self-reported detection strategies, which may reflect idealized judgments rather than the features faculty actually use when evaluating student work in real time.

These limitations naturally inform several directions for future research. Larger, cross-national studies involving a broader range of institutional types could test whether disciplinary, educational, or cultural contexts influence the weight assigned to different features. Comparative work could more closely examine whether instructors in technical fields (e.g., engineering or medicine) rely on different heuristics from those in the humanities or social sciences. Longitudinal designs would help trace how faculty evaluation strategies evolve as generative AI becomes more sophisticated and more routinely em-

bedded in student workflows. Finally, incorporating student perspectives—such as how students perceive the detectability of their own AI use or whether they intentionally introduce human-like imperfections—would provide a more complete picture of authorship dynamics in the AI era.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of the Department of Languages and Literature of the University of Nicosia.

Informed Consent Statement: Written informed consent has been obtained from the participants.

Data Availability Statement: Data are available on request.

Acknowledgments: This study was supported by the Phonetic Lab of the University of Nicosia.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Abdali, S., Anarfi, R., Barberan, C. J., & He, J. (2024, August 25–29). *Decoding the AI pen: Techniques and challenges in detecting AI-generated text*. 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 6428–6436), Barcelona, Spain.
- 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(2), 25–43. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- Bibi, A., Yamin, S., Natividad, L. R., Rafique, T., Akhter, N., Fernandez, S. F., & Samad, A. (2024). Navigating the ethical landscape: AI integration in education. *Educational Administration: Theory and Practice*, 30(6), 1579–1585. [\[CrossRef\]](#)
- Boutadjine, A., Harrag, F., & Shaalan, K. (2025). Human vs. machine: A comparative study on the detection of AI-generated content. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(2), 1–26. [\[CrossRef\]](#)
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J. L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26(1), e53164. [\[CrossRef\]](#) [\[PubMed\]](#)
- Danyaro, K. U., Abdullahi, S., Abdallah, A. S., & Chiroma, H. (2025). Hallucinations in large language models for education: Challenges and mitigation. *International Journal of Teaching, Learning and Education*, 4(6), 639993. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- Erol, G., Ergen, A., Gülşen Erol, B., Kaya Ergen, Ş., Bora, T. S., Çölgeçen, A. D., Araz, B., Şahin, C., Bostancı, G., Kılıç, İ., Macit, Z. B., Sevgi, U. T., & Güngör, A. (2025). Can we trust academic AI detective? Accuracy and limitations of AI-output detectors. *Acta Neurochirurgica*, 167(1), 214. [\[CrossRef\]](#)
- Fiedler, A., & Döpke, J. (2025). Do humans identify AI-generated text better than machines? Evidence based on excerpts from German theses. *International Review of Economics Education*, 49, 100321. [\[CrossRef\]](#)
- Georgiou, G. P. (2025). Differentiating between human-written and AI-generated texts using automatically extracted linguistic features. *Information*, 16(11), 979. [\[CrossRef\]](#)
- 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(1), 18617. [\[CrossRef\]](#)
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1–55. [\[CrossRef\]](#)
- Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11), e2208839120. [\[CrossRef\]](#)
- Janéafik, A., & Dusek, O. (2024). The problem of AI hallucination and how to solve it. In *European Conference on e-Learning* (pp. 122–128). Academic Conferences International Limited.

- Jones, E., Patrawala, A., & Steinhardt, J. (2025). Uncovering gaps in how humans and LLMs interpret subjective language. *arXiv*, arXiv:2503.04113. [CrossRef]
- Kassorla, M., Georgieva, M., & Papini, A. (2024, October 17). *AI literacy in teaching and learning: Executive summary* (An EDUCAUSE Working Group Paper). EDUCAUSE. Available online: <https://www.educause.edu/content/2024/ai-literacy-in-teaching-and-learning/executive-summary> (accessed on 29 September 2025).
- Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. In *International conference on machine learning* (pp. 17061–17084). PMLR.
- Kreps, S. E., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1), 104–117. [CrossRef]
- Krishna, K., Song, Y., Karpinska, M., Wieting, J., & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36, 27469–27500.
- Li, H., Huang, J., Ji, M., Yang, Y., & An, R. (2025). Use of retrieval-augmented large language model for COVID-19 fact-checking: Development and usability study. *Journal of Medical Internet Research*, 27, e66098. [CrossRef]
- Markowitz, D. M., Hancock, J. T., & Bailenson, J. N. (2024). Linguistic markers of inherently false AI communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, 43(1), 63–82. [CrossRef]
- Massenon, R., Gambo, I., Khan, J. A., Agbonkhese, C., & Alwadain, A. (2025). “My AI is lying to me”: User-reported LLM hallucinations in AI mobile apps reviews. *Scientific Reports*, 15(1), 30397. [CrossRef]
- Milička, J., Marklová, A., Drobil, O., & Pospíšilová, E. (2025). Humans can learn to detect AI-generated texts, or at least learn when they can't. *arXiv*, arXiv:2505.01877. [CrossRef]
- Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning* (pp. 24950–24962). PMLR.
- Mizumoto, A., Yasuda, S., & Tamura, Y. (2024). Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints. *Applied Corpus Linguistics*, 4(3), 100106. [CrossRef]
- Najjar, A. A., Ashqar, H. I., Darwish, O. A., & Hammad, E. (2025). Detecting AI-generated text in educational content: Leveraging machine learning and explainable AI for academic integrity. *arXiv*, arXiv:2501.03203. [CrossRef]
- Pan, W. H., Chok, M. J., Wong, J. L. S., Shin, Y. X., Poon, Y. S., Yang, Z., Chong, C. Y., Lo, D., & Lim, M. K. (2024, April 14–20). *Assessing AI detectors in identifying AI-generated code: Implications for education*. 46th International Conference on Software Engineering: Software Engineering Education and Training (pp. 1–11), Lisbon, Portugal.
- Pudasaini, S., Miralles, L., Lillis, D., & Salvador, M. L. (2025, January 19). *Benchmarking AI text detection: Assessing detectors against new datasets, evasion tactics, and enhanced LLMs*. 1st Workshop on GenAI Content Detection (GenAIDetect) (pp. 68–77), Abu Dhabi, United Arab Emirates.
- Rashidi, H. H., Fennell, B. D., Albahra, S., Hu, B., & Lazo, M. J. (2023). The ChatGPT conundrum: Human-generated scientific manuscripts misidentified as AI creations by AI text detection tool. *Cell Reports Medicine*, 4(12), 101331. [CrossRef]
- R Core Team. (2025). *R: A language and environment for statistical computing* (version 4.5.1) [Computer software]. R Foundation for Statistical Computing. Available online: <https://www.R-project.org/> (accessed on 29 September 2025).
- Ren, M. (2024). Advancements and applications of large language models in natural language processing: A comprehensive review. *Applied and Computational Engineering*, 97, 55–63. [CrossRef]
- Sadi Rahman, S., Islam, A., Alam, M., Zeba, M., Rahman, A., Sultana Chowdhury, S., Raiaan, M. A. K., & Azam, S. (2025). Hallucination to truth: A review of fact-checking and factuality evaluation in large language models. *arXiv*, arXiv:2508.03860. [CrossRef]
- Sahoo, P., Meharia, P., Ghosh, A., Saha, S., Jain, V., & Chadha, A. (2024). A comprehensive survey of hallucination in large language, image, video and audio foundation models. *arXiv*, arXiv:2405.09589. [CrossRef]
- Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Artificial intelligence hallucinations. *Critical Care*, 27(1), 180. [CrossRef]
- Sastre, A., Iglesias, A., Morato, J., & Sanchez-Cuadrado, S. (2024). Is ChatGPT able to generate texts that are easy to understand and read? In *World Conference on Information Systems and Technologies* (pp. 138–147). Springer Nature Switzerland.
- Showemimo, P. (2025). *The AI paradox in academia: When writing too well raises red flags*. Available online: <https://ssrn.com/abstract=5255976> (accessed on 20 January 2026).
- Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., & Wang, J. (2019). Release strategies and the social impacts of language models. *arXiv*, arXiv:1908.09203. [CrossRef]
- Spearing, E. R., Gile, C. I., Fogwill, A. L., Prike, T., Swire-Thompson, B., Lewandowsky, S., & Ecker, U. K. (2025). Countering AI-generated misinformation with pre-emptive source discreditation and debunking. *Royal Society Open Science*, 12(6), 242148. [CrossRef]
- Terčon, L., & Dobrovoljc, K. (2025). Linguistic characteristics of AI-generated text: A survey. *arXiv*, arXiv:2510.05136. [CrossRef]
- 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), 26. [CrossRef]

- Yoshida, K., Hasegawa, T., & Inoue, S. (2025). Hallucinations in medical literature searches using ChatGPT-5: The ongoing need for physician-led fact-checking. *Anesthesia & Analgesia*, 142, e28–e29. [\[CrossRef\]](#)
- Zanotto, S. E., & Aroyehun, S. (2025, November 4–9). *Linguistic and embedding-based profiling of texts generated by humans and large language models*. 2025 Conference on Empirical Methods in Natural Language Processing (pp. 22852–22869), Suzhou, China.
- Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11(1), 28. [\[CrossRef\]](#)
- Zhang, M., & Zhao, T. (2025). Citation accuracy challenges posed by large language models. *JMIR Medical Education*, 11, e72998. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.