education sciences

MDPI

Article

Reader Burden and Linguistic Register in AI-Generated Text: Implications for Higher Education

Michelle Paula Haguma [1], Ikram Ur Rehman [1,*] and Aamir Anwar [2]

[1] Faculty of Science and Technology, Middlesex University Dubai, Dubai Knowledge Park, Dubai P.O. Box 500697, United Arab Emirates; michellehagumapaula@gmail.com
[2] School of Computing, University of Portsmouth London, Juniper House, Hoe Street, Waltham Forest, London E17 9PH, UK; aamir.anwar@port.ac.uk
* Correspondence: i.rehman@mdx.ac.ae

Abstract

Generative AI is increasingly used to prepare feedback, explanations and learning materials in higher education, raising a practical question beyond authorship: how suitable are these texts for the learners who read them? Four linked exploratory studies examine this question through perceived reader burden, defined as immediate feelings of fatigue and effort, alongside linguistic differences, source classification and computational rewriting. In a comparison involving 110 university students, an AI-generated news passage received a higher mean score than a human-written passage on an exploratory six-item reading-burden composite (3.20 vs. 2.05). To understand how these passages differed linguistically, the news pair was analysed alongside matched academic and blog pairs. Across all three, AI-generated passages showed higher lexical diversity, whereas readability and sentence-length variation followed less consistent patterns. The investigation then examined computational methods: logistic regression and an RBF-SVM achieved high accuracy in classifying text sources within their evaluation settings, whereas the RoBERTa-based detector achieved lower accuracy. Rewriting reduced heuristic scores but did not consistently produce coherent prose. Finally, an independent analysis of 60 cybersecurity passages found mixed linguistic patterns and limited success in recovering individual passages' intended registers. Together, these findings suggest that reader burden, linguistic characteristics and source classification offer distinct forms of evidence rather than interchangeable indicators of educational suitability.

Keywords: generative artificial intelligence; artificial intelligence in education; higher education; digital education; AI in pedagogy; ethical AI in education; reader burden; AI fatigue; linguistic register; AI-generated text

Check for updates

Academic Editor: Shirley O'Neill

Received: 4 September 2026
Revised: 13 October 2026
Accepted: 5 November 2026
Published: 20 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Generative Artificial Intelligence (Gen AI) has become part of higher-education practice, bringing with it questions about when its use is appropriate and how that use should be governed. In the United Kingdom (UK), the Russell Group, an association of 24 research-intensive universities, promotes AI literacy, ethical use and equal access while maintaining academic rigour and integrity (Russell Group, 2023). The Quality Assurance Agency for Higher Education (QAA), the UK's independent quality body for higher education, reported in 2026 that student use of Gen AI, staff confidence, institutional policy and the implementation of that policy continue to vary across institutions, schools, modules and

tutors (Quality Assurance Agency for Higher Education, 2026). These variations in institutional policy and practice also raise questions about how students perceive the appropriate use of Gen AI. In a survey of 2555 University of Liverpool students, more than half had used or considered using Gen AI for academic purposes, while 70.4% were unsupportive or somewhat unsupportive of using ChatGPT to write an entire essay and 41.1% wanted a university-wide policy on appropriate use (Johnston et al., 2024). Jisc, a UK organisation supporting education and research through digital technology, similarly identified assessment and academic integrity as early sector concerns, while arguing that simple prohibition was unlikely to remain sustainable as Gen AI became more widely used (Webb, 2023). Together, these developments show a sector increasingly concerned with how AI should be used responsibly. They do not, however, answer a different question that arises when AI-generated material is placed directly in front of a learner.

The central question is whether the resulting text is suitable for its intended reader. If an educator uses Gen AI to prepare feedback, an explanation, guidance or a learning resource, its source may be known and its use permitted. In that situation, identifying its authorship does not establish how the intended reader will experience it. Reader responses may concern clarity, trust, processing effort and other aspects of the reading experience. This study focuses on a narrower aspect of this experience, termed *reader burden*: immediate fatigue- and effort-related responses to a particular passage. This is different from AI fatigue, which recent work describes as a broader response to sustained interaction with AI systems and which may include cognitive, emotional, physical and behavioural dimensions (Lau et al., 2026; Miranda et al., 2026). Reader burden, as used here, is an author-defined operational label for the immediate reported experience of processing the text, rather than a validated scale or a persistent state arising from repeated or prolonged AI use.

Authorship alone cannot account for processing difficulty. Cognitive load theory suggests that demands on the learner can arise from the material itself and from the way information is presented (Sweller, 1988). More recent evidence also shows that Gen AI may either reduce or increase cognitive load depending on factors such as learner knowledge, task design and scaffolding (Qian et al., 2026). An AI-generated passage cannot therefore be assumed to be more or less demanding simply because of its source. Prompt formulation, iterative refinement and subsequent human editing may also shape the linguistic complexity and pedagogical appropriateness of the resulting passage. A more experienced user may adapt prompts to the intended audience, although that possibility is not evaluated in the present data. Differences between the two Study 1 passages therefore cannot be attributed solely to authorship without considering the generation procedure and the particular texts selected. The linguistic characteristics of the passages consequently warrant closer examination.

Human-written and AI-generated texts have been reported to differ in readability, lexical diversity, sentence organisation and discourse structure, although these differences vary across models and datasets (Herbold et al., 2023; Muñoz-Ortiz et al., 2024; Reviriego et al., 2024). An important part of this variation is *register*, understood here as the pattern of language associated with a particular communicative purpose and writing context. Academic explanation, news reporting and professional blogging, for example, differ in their expected vocabulary, sentence structure, density, stance and discourse organisation. Recent research likewise shows that AI-generated text does not align with human-written text in the same way across genres and disciplines (Demir & Egbert, 2026; Reinhart et al., 2025). A linguistic feature that appears unusual in one register may therefore be entirely ordinary in another. Human-written and AI-generated texts need to be compared within appropriate writing contexts before particular linguistic differences are treated as meaningful indicators.

These linguistic differences can also become inputs to computational classification systems. Human–AI source classifiers use linguistic or learned patterns to estimate whether a text is more likely to be human-written or AI-generated. These systems can perform well under particular evaluation conditions, but their performance may change across generators, domains and modified texts (Dugan et al., 2024; Weber-Wulff et al., 2023). More importantly for educational use, successful source classification answers an authorship question. It does not establish that a passage is clear, accurate, easy to process or suitable for learning. This distinction becomes especially important when an educator already knows that AI has been used to produce the material.

A related issue arises when linguistic features are used to guide rewriting rather than classification. Rewriting systems can alter style, complexity or other textual properties while attempting to preserve the original meaning (Mukherjee & Dušek, 2024; Mukherjee et al., 2024). Nevertheless, making a passage appear closer to a preferred linguistic pattern does not necessarily make it better for a learner. Important qualifications, factual relations or other source information may be weakened or lost during rewriting. A text can therefore look linguistically improved while becoming less faithful to its original meaning. Source classification, linguistic similarity and rewriting each provide useful information for particular purposes, but none of them can automatically be treated as evidence of reader suitability.

Evaluating AI-generated educational text therefore requires several distinct forms of evidence. Research on reader experience, linguistic form, source classification and rewriting addresses related but different questions. A passage may be confidently classified as AI-generated and still be suitable for a learner. In contrast, a passage may resemble a familiar human register without being easy to process. A rewrite may also appear linguistically improved while losing information that should have been preserved. The issue is therefore not simply whether AI-generated text differs from human writing, but whether evidence about authorship, linguistic form, preservation of meaning and reader response can legitimately be used in place of one another.

In this paper, four linked exploratory studies address this problem through distinct but complementary forms of evidence. Study 1 compares immediate fatigue- and effort-related responses to two topic-matched news passages in 110 university students. Study 2 analyses the linguistic properties of the same news pair alongside matched academic and blog texts, examining whether human–AI differences are consistent across writing contexts. Study 3 examines the application of linguistic features to scoring, human–AI source classification and rewriting, including the limitations of these computational measures. Study 4 assesses the stability of register-relative comparisons in a separate 60-passage corpus comprising human-written, ChatGPT-generated and Claude-generated texts. The studies thus address different evaluation questions, rather than providing repeated tests of a single outcome. Studies 1–3 build on research documented in an earlier undergraduate dissertation (Haguma, 2026), with the original analyses re-examined in the present article alongside an independent fourth study.

To address the overall research aim, four research questions guide the analysis, focusing on students' reading experiences, linguistic differences across writing contexts, the capabilities and limitations of computational methods, and the stability of register-sensitive comparisons:

1. **RQ1 (Study 1).** How do students' reported levels of fatigue and reading effort differ after reading a human-written or an AI-generated passage on the same topic?
2. **RQ2 (Study 2).** How do the observed linguistic differences between matched human-written and AI-generated passages vary across academic, news and blog writing?
3. **RQ3 (Study 3).** How can computational methods be used to measure linguistic differences, distinguish human-written from AI-generated texts and guide rewriting,

and what are their limitations in assessing reader suitability and the preservation of meaning?

4. **RQ4 (Study 4).** How stable are register-relative linguistic comparisons when they are assessed in an independent, topic-matched corpus with explicit feature standardisation and sensitivity checks?

These questions are exploratory rather than tests of confirmatory hypotheses. In particular, Studies 2–4 do not measure reader burden directly and cannot validate the questionnaire outcomes in Study 1.

The four studies address complementary parts of the evaluation problem. Reader response cannot be inferred from linguistic form alone, linguistic differences do not establish educational quality, and computational performance does not demonstrate preservation of meaning or reader suitability. Taken together, the studies allow for these forms of evidence to be considered alongside one another without assuming that one can replace another. Rather than proposing a new detector or a single automatic quality score, this study investigates why source, linguistic fit, preservation of meaning and reader response require separate consideration when evaluating text intended for learners.

2. Literature Review

This review examines research on AI governance in higher education, readers' experiences, linguistic differences between human-written and AI-generated texts, source classification and computational rewriting. Although these topics are often investigated separately, they collectively help address a central question: how should AI-generated texts intended for learners be evaluated?

2.1. AI Governance and Reader Suitability

Institutional responses to Gen AI have developed rapidly since the public release of large language models. Early guidance focused particularly on assessment, academic integrity and appropriate use. Jisc argued that generative AI was likely to become sufficiently widespread that simple prohibition would not provide a sustainable long-term response (Webb, 2023). The Russell Group took a similar position in setting out principles concerned with AI literacy, staff capability, ethical use, equal access, academic rigour and integrity. Their guidance also highlights risks including privacy, intellectual property, bias, inaccurate information and plagiarism (Russell Group, 2023). Taken together, these sources show that the higher-education response has centred largely on how AI should be used responsibly rather than on whether it can simply be excluded.

In practice, however, these principles are applied differently across higher education. QAA reports differences in student use, staff confidence, institutional policy and implementation across institutions, schools, modules and tutors (Quality Assurance Agency for Higher Education, 2026). Similar variation can be seen in university-level guidance. Oxford requires those setting summative assessments to state whether and how AI may be used, whereas UCL categorises assessments according to the permitted use of generative AI, from tasks where its use is prohibited to those where it is integral (University College London, n.d.; University of Oxford, 2025). Variation is also evident outside the UK. Wang et al. studied highly ranked universities in the United States and found that many institutions gave individual instructors substantial responsibility for deciding whether Gen AI was permitted, while concerns about plagiarism, privacy and other limitations remained prominent (Wang et al., 2024).

These policies provide context for the present study, but they focus mainly on permission, disclosure, integrity, responsibility and student use of AI. They were not developed as standards for judging the quality of AI-generated prose that an educator or institution might

provide directly to learners. When AI use is permitted and disclosed, a further question remains: how suitable is the resulting text for the learner expected to read and learn from it? Addressing that question requires evidence about the reader as well as evidence about the source.

2.2. AI Fatigue and Reader Burden

Reader experience with AI sits within a broader literature on technostress, information overload and sustained digital interaction. Technology-related demands have long been associated with strain and reduced satisfaction (Ragu-Nathan et al., 2008), while excessive or poorly organised information can impair attention and decision-making (Shahrzadi et al., 2024). Several studies on AI use in higher education show a similarly mixed picture. AI can provide efficiency, access and support, but concerns have also been raised about dependence, uncertainty, isolation and stress (Klimova & Pikhart, 2025; Vieriu & Petrea, 2025). Negative responses to AI encompass distinct phenomena that require careful definition.

Recent research on AI fatigue helps distinguish sustained interaction-related responses from the experience of reading a single passage. Lau et al. developed and validated a multidimensional AI Fatigue Scale concerned with sustained human–AI interaction (Lau et al., 2026). Miranda et al. likewise described AI fatigue as an accumulating process associated with academic AI use and emphasised the need for continued development and validation of measurement instruments (Miranda et al., 2026). The construct is therefore broader than the immediate effort involved in reading one passage. Study 1 did not use a validated AI-fatigue instrument; instead, it examined immediate fatigue- and effort-related responses to reading a single passage. Its six-item exploratory composite combines mental tiredness, anticipated fatigue from further reading, sustained effort, perceived cognitive demand, perceived heaviness and perceived unnecessary length. Concentration difficulty and feeling drained were recorded separately in the questionnaire and are not components of that composite. The composite is interpreted as a broad self-reported reading-burden indicator, not a validated fatigue scale or evidence of a persistent AI-fatigue state.

The authors use *reader burden* as an operational term for immediate self-reported fatigue and effort associated with reading a particular passage. The term is informed by cognitive load theory, research on perceived mental effort, and studies of AI fatigue (Lau et al., 2026; Paas et al., 2003; Sweller, 1988), but it does not denote a validated psychometric construct. Textual density, organisation, repetition and presentation may contribute to that experience, although lower complexity is not necessarily desirable. Educational material still needs to retain necessary terminology, qualifications and conceptual relationships. The objective is not minimum complexity but an appropriate level of complexity for the text's purpose and intended readers.

Reader response also extends beyond fatigue and processing effort. AI-assisted communication can influence perceptions of authenticity, social presence and trust (Hohenstein et al., 2023; Wester et al., 2024). Research on human-centred AI further suggests that interaction design, socio-emotional cues and anthropomorphic features can shape how people perceive and work with AI systems (Akbulut et al., 2024; Alabed et al., 2022; Kolomaznik et al., 2024; Krakowski et al., 2025). Similar considerations appear in educational feedback. Recent studies indicate that the value of AI-supported feedback depends partly on how the feedback is designed, how learners interpret it and what role the AI system plays within the learning activity (Ates, 2026; Yilmaz et al., 2026). Reader experience therefore cannot safely be inferred from authorship or from a readability score alone. This raises the next question: whether characteristics of the text itself can provide a more informative account of why human-written and AI-generated material may be experienced differently.

2.3. Linguistic Form and Register

Research comparing human-written and AI-generated text has identified differences in a wide range of linguistic features, including readability, lexical diversity, sentence-length distribution, dependency structure, word choice and discourse organisation. In controlled news data, for example, human texts showed greater dispersion in sentence length and vocabulary variation than texts produced by several language models (Muñoz-Ortiz et al., 2024). Differences have also been reported in essays and scientific writing (Herbold et al., 2023; Reviriego et al., 2024; Terčon & Dobrovoljc, 2025). These findings are useful, but they also show an important limitation: individual linguistic features do not behave consistently across models, datasets and writing situations.

One possible explanation for these inconsistent findings is the influence of linguistic register. Register refers to the patterns of language typically associated with a particular communicative purpose and writing context. Research analysing writing styles shows that differences in writing contexts can influence linguistic patterns that might otherwise be attributed to individual authors (Grieve, 2023). Instruction-tuned language models have been found to show recurring patterns in grammar and the presentation of ideas across different writing contexts (Reinhart et al., 2025). Demir and Egbert, examining journal articles and textbooks in biology and history, similarly found mixed register alignment between human-written and AI-generated texts, with weaker functional alignment in some textbook contexts (Demir & Egbert, 2026). Human–AI differences should therefore not be expected to take the same form in every kind of writing.

This matters because an academic text, a news report and a professional blog are written for different purposes. They differ naturally in stance, sentence structure, information density, voice and discourse organisation. A feature that appears distinctive in one register may be common human usage in another. Fixed lists of supposedly AI-associated words or structures are therefore difficult to interpret without an appropriate comparison context. Register-sensitive analysis can help establish how texts differ, but it still does not tell us whether those differences identify authorship or whether they matter to readers. The distinction matters when linguistic patterns are used for source classification.

2.4. AI-Generated Text Detection and Its Limitations

AI-generated text detection addresses a different problem from reader suitability. It has received considerable attention because identifying source can matter in academic integrity, scientific publishing and information security. Yet detector performance depends on the conditions under which a system is developed and evaluated. RAID, for example, demonstrated sensitivity to unseen generators and text perturbations across a large benchmark (Dugan et al., 2024). Independent evaluations of open and commercial systems have likewise reported heterogeneous performance across tools, domains and text conditions, with continuing concerns about false positives and generalisation (Jabarian & Imas, 2025; Kalra et al., 2024; Weber-Wulff et al., 2023).

Source-classification accuracy has a restricted interpretation: it indicates how well a system distinguishes likely source classes under a specified evaluation distribution. It does not show whether the classified passage is accurate, understandable, tiring or useful for learning. The distinction between identifying a text's source and assessing its suitability becomes particularly important when educators knowingly use Gen AI to prepare learning materials. In such cases, the source is already known, but whether the resulting text is suitable for learners remains uncertain.

Linguistic features can, however, be used for more than classification. Once particular features are treated as desirable or undesirable, the same measures can also become targets

for rewriting. This creates a related but different problem: improving a linguistic proxy may not improve the text itself.

2.5. Rewriting and Proxy Optimisation

Text style transfer generally seeks to modify stylistic properties while preserving the underlying meaning (Mukherjee & Dušek, 2024; Mukherjee et al., 2024). Meaning preservation is particularly important for educational material. A rewrite that removes a qualification, changes a factual relationship or omits necessary information cannot reasonably be considered better simply because it is shorter or more closely resembles a preferred linguistic pattern.

A further difficulty arises when the measure guiding a rewrite also determines whether it is judged successful. Under those conditions, a system can optimise the measure without improving the reader's experience. A lower linguistic score may therefore reflect successful movement towards the chosen target while concealing loss of information or other undesirable changes.

A reader-oriented measure therefore requires evidence independent of the scoring procedure. It would need to relate to an external outcome such as reader effort or comprehension, remain sufficiently stable across the registers in which it is intended to operate, and contribute information beyond conventional readability measures and human editorial judgement. The present study examines several parts of this problem, but it does not assume that they can be reduced to a single validated score.

2.6. Research Gap and Contribution

The literature reviewed above addresses a common question from different perspectives: how can we determine whether AI-generated text is suitable for learners? Research on AI governance considers when and under what conditions generative AI may be used in higher education. Research on AI fatigue and reader experience examines how people respond to AI-supported interactions and reading experiences. Linguistic studies investigate differences between human-written and AI-generated texts, including how these differences vary across writing registers. Source-classification research examines whether a text can be identified as human-written or AI-generated, while rewriting research explores how linguistic characteristics can be modified while preserving the original meaning. Although these areas provide valuable insights, each addresses only part of the broader question of educational suitability.

This distinction is important because evidence about one characteristic of a text cannot automatically establish its other qualities. A classifier may accurately identify a passage as AI-generated without indicating whether students find it easy to read or understand. Similarly, a text may resemble the linguistic patterns of human writing without necessarily being suitable for its intended audience. A rewritten passage may also achieve a more favourable linguistic score without demonstrating that it remains coherent or preserves its original meaning. Therefore, source classification, linguistic characteristics, preservation of meaning and reader responses should be considered as distinct forms of evidence rather than equivalent measures of educational quality.

This study addresses this gap through four linked exploratory studies that examine different aspects of the same evaluation problem. Beginning with students' reported fatigue and reading effort, the investigation extends to linguistic differences across writing contexts, the capabilities and limitations of computational classification and rewriting, and the stability of register-related patterns in an independent corpus. These studies are not intended to measure the same outcome repeatedly. Instead, they provide complementary perspectives on what linguistic and computational analyses can reveal about AI-generated texts and

where direct evidence from readers remains necessary. The contribution is therefore not a new general-purpose detector or a validated measure of educational suitability, but a clearer understanding of how these different forms of evidence can inform the evaluation of AI-generated texts intended for learners.

3. Materials and Methods

3.1. Overall Design

This paper presents four linked exploratory studies that use different methods to investigate related aspects of evaluating AI-generated texts for educational use. A structured narrative review provides the conceptual background for the investigation. The four studies examine students' reading experiences, linguistic differences across writing contexts, computational scoring, source classification and rewriting, and register-sensitive linguistic patterns in an independent corpus.

The studies are connected by their research questions and, in some cases, by shared materials. Study 1 compares student responses to one human-written and one AI-generated news passage. Study 2 reuses the same news pair and examines its linguistic characteristics alongside matched academic and blog texts. This allows for the reader-response findings to be considered together with the linguistic properties of the texts and also shows whether similar human–AI differences appear across other writing contexts. Study 3 then considers what happens when linguistic features are used computationally for scoring, source classification and rewriting. Study 4 independently examines register-related patterns in a separate corpus of human-written, ChatGPT-generated and Claude-generated texts.

No study is treated as validating another. Instead, the analyses are interpreted together to establish the scope and limits of each form of evidence for AI-generated text intended for learners.

Table 1 summarises the design, purpose and evidence used in each study.

Table 1. Design and evidence used across the four studies.

Study	Design	Main Purpose	Evidence Used
Study 1	Reader-response comparison	Compare immediate fatigue- and effort-related responses to two topic-matched passages	110 university students, 55 per reading condition
Study 2	Descriptive matched-text linguistic analysis	Examine whether human–AI linguistic differences remain similar across different writing contexts	Three matched human–AI pairs across academic, news and blog writing, plus four additional human academic/reference texts
Study 3	Computational analysis	Examine the use of linguistic features for scoring, source classification and rewriting	19-feature Python analysis and a 2000-document same-source-corpus classifier evaluation
Study 4	Independent matched-register analysis	Assess register-related patterns using an independent corpus and sensitivity analysis	60 human, ChatGPT and Claude passages across formal technical/academic, news and blog writing

3.2. Structured Review Procedure

A structured narrative review was used to establish the conceptual background for the empirical studies. The review draws on several related areas, including AI fatigue, technostress, cognitive load, linguistic analysis and AI-generated text detection, rather than addressing a single intervention or outcome. Its purpose was to organise relevant evidence across these areas, not to estimate a pooled effect.

The literature search was conducted using several public databases and search platforms, including IEEE Xplore, Google Scholar, ScienceDirect and Frontiers. The search terms included AI fatigue, technostress, cognitive overload, humanising AI, emotional AI and AI detection. The initial search identified 537 records. The first screening stage excluded duplicates, non-English publications and records published before 2022, leaving 287 records.

The remaining records were screened for topical relevance, leaving 153. Of these, 66 did not proceed to full-text assessment, leaving 87 records for further review. Following full-text assessment, 54 studies were excluded because they did not meet the review's topical relevance criteria, leaving 33 studies for the final review.

The review process was informed by PRISMA 2020 principles (Page et al., 2021) and conducted as a structured narrative review. The reported screening counts describe the structured search; additional literature on AI fatigue, cognitive load and register alignment is discussed for contextual purposes and is not included in those counts.

The recorded screening process is shown in Figure 1.

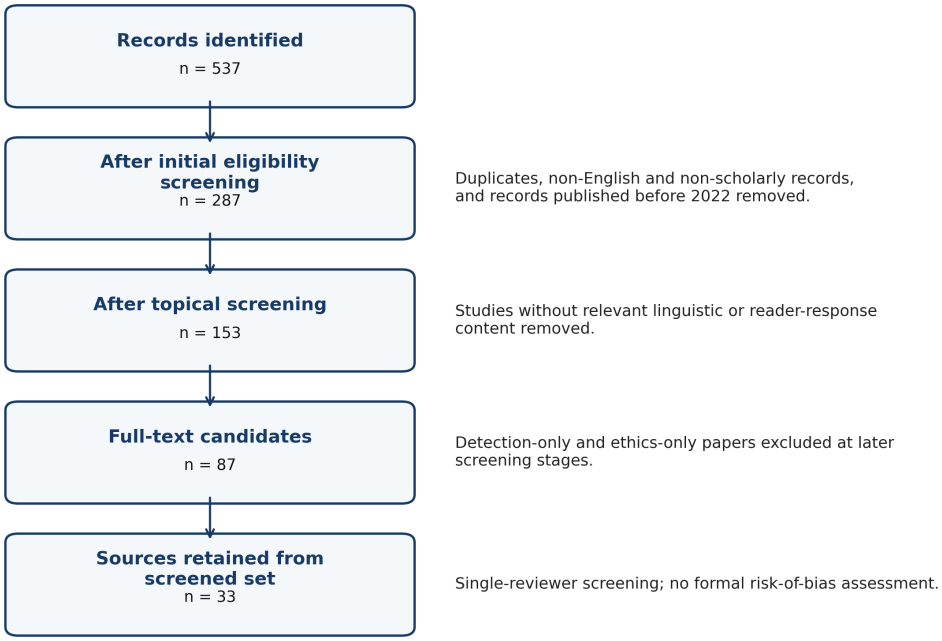


Figure 1. Recorded structured-review flow informed by PRISMA.

3.3. Study 1: Reader-Response Comparison

Study 1 investigated how students responded to AI-generated and human-written news passages on the same topic. The AI-generated passage was produced using ChatGPT. The comparison concerns reported fatigue and reading effort in response to these two passages. Subsequent linguistic analyses examine text characteristics but do not independently explain the observed reader responses.

Both passages concerned the Kursk submarine disaster, in which the Russian nuclear-powered submarine *Kursk* sank in the Barents Sea in August 2000. The human-written passage was extracted from a BBC News article published on 12 August 2010 (BBC News, 2010). The AI-generated passage covered the same topic and was created using ChatGPT with the prompt: "Analyse this BBC article and generate a similar writing piece on this exact same topic with the exact same number of words."

This was the generation instruction documented for the Study 1 passage. Prompting competence, alternative prompts, iterative prompt refinement and post-generation editorial changes were not compared as experimental factors. The two passages therefore cannot isolate the effect of model source from the particular generation procedure and resulting text.

University students were recruited through convenience sampling over a four-week period (Golzar et al., 2022). Recruitment was carried out through their lecturers, who helped circulate information about the study. Ethical approval was granted through the institutional research ethics process on 12 January 2026, before data collection began. Students who were invited to take part received a participant information sheet explaining the purpose of the study, what participation involved, how their data would be handled, and that participation was entirely voluntary. Those who agreed to participate were then asked to provide informed consent. They were also informed that they could withdraw from the study at any time and that no personally identifiable information would be collected.

A total of 110 students participated in the study. After recruitment, they were randomly assigned to one of two reading conditions: 55 students read the human-written BBC News passage and the remaining 55 read the AI-generated passage. Each participant read only one passage and was not told whether the text had been written by a human or generated using the free version of ChatGPT.

After reading the assigned passage, participants completed a questionnaire administered through Google Forms. The first set of questions checked how carefully they had read the passage and how long they had taken to read it. Of the 110 participants, 102 (92.7%) reported that they had read the complete passage carefully. Seventy participants (63.6%) reported taking less than three minutes, 34 (30.9%) reported taking between three and five minutes, five (4.5%) reported taking between six and ten minutes, and one (0.9%) reported taking more than ten minutes. Reading times are summarised for the full sample; they were not used to construct the reading-burden composite.

The next set of questions focused on participants' experience of reading the passage. A five-point agreement scale was used to assess fatigue and reading effort, together with broader perceptions of heaviness, length, repetition, structure, generic framing, emotional neutrality, reader awareness, quality, authenticity, clarity, coherence, trustworthiness and AI-likeness. The fatigue- and effort-related items received particular attention because they were most closely aligned with the operational definition of reader burden.

The exploratory perceived reading-burden composite was constructed from six questionnaire statements: "I felt mentally tired after reading this text", "I would feel fatigued if I had to read more text written in this style", "Reading this text required sustained effort", "The text felt cognitively demanding", "The text felt heavier than necessary to read", and "The text felt longer than it needed to be". Responses were coded from 1 (strongly disagree) to 5 (strongly agree), with higher values indicating greater perceived burden. Each item received equal weight. For participant i in condition g, the composite was $S_{ig} = \frac{1}{6}\sum_{j=1}^{6} x_{igj}$; the condition mean was $\overline{S}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} S_{ig}$. The participant-level survey responses and analytical records confirmed 55 valid responses per item in each condition ($n_g = 55$), with no missing responses across these six items; hence, no imputation or exclusion was required. The unrounded condition means were 3.203030303 (AI-generated, Text A) and 2.054545455 (human-written, Text B). The values 3.20 and 2.05 result from rounding the observed means to two decimal places.

For the mental-tiredness item, responses on the five-point scale were grouped as Disagree (categories 1–2), Neutral (category 3), and Agree (categories 4–5). This recoding was not used in the six-item composite. The composite combines reported tiredness and effort with anticipated fatigue and perceptions of unnecessary length and heaviness. It is therefore an exploratory indicator of perceived reading burden, not a validated psychometric

measure of AI fatigue or cognitive load. The concentration-difficulty and feeling-drained responses were separate questionnaire outcomes, not constituents of the reported six-item mean. The separate concentration and skimming responses were also coded from 1 to 5. For the skimming question, the ordered categories were Not at all, Slightly, Moderately, Strongly, and Very strongly. No responses to these two items were missing. Reliability and group dispersion for the six-item composite are reported descriptively below; internal consistency alone does not establish construct validity.

3.4. Study 2: Linguistic Analysis Across Three Writing Types

Study 2 investigated whether the linguistic differences between human-written and AI-generated text remained similar across three types of writing: academic, news and blog writing. The academic pair was based on the topic of a University of Leeds dissertation (University of Leeds Library, 2020). The news pair drew on the Kursk source material also considered in Study 1, allowing for the reader-response findings to be discussed alongside linguistic analysis of related news text. The blog pair focused on the 2018 royal wedding (Evans & Bacon, 2018). Within each pair, the human-written and AI-generated passages covered the same topic and were approximately matched in length. The passages were then analysed in R using the same set of linguistic measures so that the human-written and AI-generated texts could be compared within each writing type.

The linguistic measures described characteristics that might differ between human-written and AI-generated passages. The Flesch Reading Ease score was used to describe general readability (Flesch, 1948), while mean words per sentence and sentence-length variance captured differences in sentence length and structure. Type-token ratio (TTR) and moving-average type-token ratio (MATTR) were used to examine lexical diversity, with MATTR included because it is less affected by text length than ordinary TTR (Covington & McFall, 2010). Nominalisation, hedging and discourse-marker density were included to examine differences in writing style and organisation, while mean sentiment provided an indication of the overall tone of the text. The content ratio compared content-word categories with function words and was retained from the Study 2 analysis.

Four additional human-written texts were included to illustrate variation in academic and reference writing. These consisted of two university dissertations and two historical book extracts (Abdulla, 2017; Al Hammadi, 2017; Darwin, 1999; Healy, 1912). Together with the human-written academic text from the matched pair, this produced a set of five human texts for comparison. Because the texts were written in different periods and for different purposes, they were not treated as representing a single contemporary academic standard. They illustrate variation among the selected human-written texts.

Study 2 investigated linguistic differences between human-written and AI-generated texts across writing types. Study 3 examined related features in computational scoring, source classification and rewriting.

3.5. Study 3: Computational Analysis

Study 3 examined how 19 linguistic features covering readability, lexical diversity, sentence structure, writing style and sentiment could be used for computational scoring, source classification and rewriting. These areas were selected because similar linguistic characteristics have been used in previous studies to compare human-written and AI-generated text (Herbold et al., 2023; Muñoz-Ortiz et al., 2024). Together, these features characterised several dimensions of linguistic form rather than relying solely on readability. They were used as linguistic indicators and were not treated as direct or validated measures of reader burden.

The Python implementation was evaluated as a proof of concept. The 19 features were used to examine three computational tasks: scoring how far a text differed from a register reference, classifying whether a text was more likely to be human-written or AI-generated, and identifying possible changes through rewriting. Source classification was tested using logistic regression, a radial-basis-function support-vector machine (RBF-SVM), and the pretrained Hello-SimpleAI ChatGPT Detector RoBERTa model (Hello-SimpleAI, n.d.). Study 4 examined the register assumptions in a separate corpus using corrected feature counts, without adopting the Study 3 heuristic as its main outcome.

3.5.1. Feature Extraction and Deviation Score

The 19 linguistic features were grouped into five areas: readability, lexical diversity, sentence structure, stylistic markers and sentiment. Readability was represented by the Flesch Reading Ease score, the Gunning Fog index and SMOG (Flesch, 1948; Gunning, 1952; McLaughlin, 1969). Lexical diversity was described using TTR, 25-token MATTR and hapax density. Sentence structure was examined using mean sentence length, sentence-length variance, noun density, mean dependency depth and named-entity density. Together, these measures captured differences in readability, vocabulary use and sentence organisation that have also been examined in previous comparisons of human-written and AI-generated text (Herbold et al., 2023; Muñoz-Ortiz et al., 2024). They were used to describe linguistic characteristics of the texts rather than as direct measures of reader burden.

The remaining features described different aspects of writing style. These included the use of contractions, passive markers per sentence, discourse markers, hedging expressions, variation in sentence openings and the use of dashes. The feature set also included a manually specified list of words associated with AI-generated writing in the Study 3 scoring procedure. Sentiment was measured using the mean sentence-level VADER compound score (Hutto & Gilbert, 2014), which provided an overall indication of whether the language was more positive, negative or neutral.

Text preprocessing used lower-case alphabetic NLTK tokens, while the spaCy model `en_core_web_sm` was used for noun counts, named entities and dependency information. One feature in Study 3 was labelled `passive_ratio` and counted `nsubjpass` and `auxpass` dependency markers per sentence. Because more than one passive marker can occur in a sentence, a value of 1.0 represents an average of one marker per sentence and values above 1.0 are possible. The measure is therefore reported here as *passive markers per sentence* rather than as a ratio.

Several Study 3 features relied on exploratory counting rules. Contraction rate was estimated from words containing apostrophes, discourse markers were identified through basic word matching, and the hedge and AI-associated vocabulary lists were created manually. The feature definitions and preprocessing used in Study 4 are specified in Section 3.6.

For the heuristic score, each feature was compared with the reference value for the corresponding register. Relative deviation was capped at 2.0 before weighting, with a bounded rule used when the reference value was zero. Small non-zero reference values could still produce large relative differences. The Study 3 weights summed to 1.04; each was therefore divided by 1.04 to obtain weights summing to 1.00.

Table 2 summarises the feature groups and normalised weights in the Study 3 scoring procedure.

The ratio-based heuristic is sensitive to the scale of its reference values. Study 4 uses standardised feature distances, rather than that heuristic, to assess proximity to human register references.

Table 2. Feature families and normalised weights used in the Study 3 heuristic.

Feature Family	Implemented Features	Normalised Weights
Readability	Flesch; Gunning Fog; SMOG	0.0962; 0.0288; 0.0288
Lexical diversity	TTR; MATTR; hapax density	0.0481; 0.0481; 0.0192
Syntax and structure	Mean sentence length; sentence-length variance; mean dependency depth; noun density; named-entity density	0.0673; 0.0673; 0.0192; 0.0192; 0.0288
Stylistic markers	Contraction rate; passive markers/sentence; discourse-marker density; hedge density; opener diversity; dash frequency; AI-associated vocabulary density	0.0673; 0.0481; 0.0962; 0.0481; 0.0481; 0.0962; 0.0962
Sentiment	Mean sentence-level VADER compound score	0.0288

3.5.2. Human–AI Source Classifiers

The classification component of Study 3 addressed a question distinct from reader burden: whether the same linguistic features could be used to distinguish human-written from AI-generated text. The classification results were therefore treated as evidence about source identification rather than about readability or suitability for learners.

Two classical classifiers were trained using the same 19 linguistic features described above. Logistic regression was included as a simple linear classifier, while a radial-basis-function support-vector machine (RBF-SVM) was used to allow for more complex relationships between the features. Model-development data were taken from the public `artem9k/ai-text-detection-pile` aggregate dataset (artem9k, n.d.). This dataset was selected because it provided a sufficiently large collection of labelled human-written and AI-generated texts for developing and evaluating the feature-based classifiers. The aim was to test the computational use of the 19 linguistic features rather than to suggest that the dataset represented all writing domains or AI models.

The working sample contained 5000 human-written and 5000 AI-generated documents, each containing at least 100 characters. To use a consistent maximum input length, feature extraction was restricted to the first 3000 characters of each document. The data were divided into training and internal-test sets using an 80:20 stratified split with `random_state = 42`, producing 4000 training and 1000 internal-test documents for each source class.

`StandardScaler` was fitted using only the training data. Logistic regression used $C = 0.5$ and `max_iter = 2000`, while the RBF-SVM used $C = 2$ and `gamma = "scale"`. Predicted probabilities from both models were calibrated using five-fold isotonic calibration.

A pretrained `Hello-SimpleAI/chatgpt-detector-roberta` model was also included (Hello-SimpleAI, n.d.). Unlike the two classical classifiers, this model did not use the 19 manually extracted features but analysed the text directly. It had previously been trained on HC3 and was included as a separate transformer-based detector rather than as a directly comparable model. Each document was processed in consecutive 400-character sections up to a maximum of 1200 characters, and the resulting AI-class probabilities were averaged.

For the final evaluation, a separate set of 1000 human-written and 1000 AI-generated documents was selected from the same aggregate dataset after excluding the 5000 documents per class used for model development. The evaluation set was therefore separate from the training and internal-test data at the document level. However, because both the development and evaluation samples came from the same aggregate dataset, this remained a same-source-corpus evaluation rather than a fully independent cross-corpus test.

The three classifiers were not evaluated as a controlled head-to-head comparison. Logistic regression and the RBF-SVM were trained on the artem9k dataset, whereas the RoBERTa detector had previously been trained on HC3. The artem9k dataset also contains

HC3-derived material (artem9k, n.d.; Hello-SimpleAI, n.d.), and the retained metadata were not sufficient to determine whether any individual RoBERTa training documents also appeared in the 2000-document evaluation set. The RoBERTa results are therefore reported descriptively rather than as part of a controlled benchmark.

A separate AI-generated academic Test 2 passage was used as a diagnostic example because it was not part of the 2000-document evaluation set.

3.5.3. Sentence Triage and Rewriting Procedure

The system also used the heuristic score to identify sentences that differed substantially from the relevant register reference and might therefore be considered for rewriting. In the Study 3 procedure, 0.75 was used as a working development threshold: sentences reaching or exceeding this value were flagged for possible revision. This was an exploratory threshold rather than a validated linguistic or reader-burden cut-off.

The initial feature weights sum to 1.04; scores reported here use weights normalised to sum to 1.00. Let w_j denote an unnormalised non-negative feature weight ($\sum_j w_j = 1.04$), and let $d_j(x, r)$ denote the implemented, capped deviation or directional penalty for feature j in text x relative to register reference r, with $0 \leq d_j \leq 2$. The normalised score used in Study 3 was calculated using the following exploratory weighted-deviation formula. This scoring procedure is not an established or validated measure of reader burden:

$$S_{\text{norm}}(x, r) = \min\left(100,\ 100 \sum_j \frac{w_j}{1.04} d_j(x, r)\right). \quad (1)$$

Study 3 reported passage-level heuristic scores on a 0–100 scale, but used a 0–1 scale to identify sentences for possible rewriting. Because the feature weights were normalised by dividing by 1.04, the original sentence threshold of 0.75 corresponds to approximately 0.72 on the normalised scale. For passage scores below the ceiling of 100, the normalised scores were calculated by dividing the corresponding unnormalised scores by 1.04 (for example, $86.80/1.04 = 83.46$). Passage features were extracted from each complete passage, whereas sentence-level features were calculated separately; the passage score is not the average of its sentence scores. The sentence threshold is not a validated cut-off for reader burden. The threshold was examined using a small development set of 22 sentences: 8 from a human-written academic passage, 5 from the AI-generated academic Test 2 passage and 9 from an additional AI-generated passage written in a deliberately more human-like style.

The rewriting experiment compared reductions in the heuristic score with lexical retention and the quality of the resulting prose. Two pretrained text-generation models, PEGASUS (Zhang et al., 2020) and T5 (Raffel et al., 2020), were used to generate revised versions of the selected text. PEGASUS was tested using both a max-divergence procedure, which selected the least lexically similar candidate, and a guarded procedure, which rejected candidates that did not satisfy additional constraints. T5-base was tested for general paraphrasing and for rewriting directed towards particular linguistic features. The max-divergence variant was retained as a diagnostic failure case rather than treated as a viable rewriting intervention.

Each rewritten passage was processed again using the same feature-extraction and scoring procedure. Content-word retention measured lexical overlap with the source, while length ratio measured shortening or expansion. Neither measure was a test of semantic equivalence or factual preservation; a lower overlap may reflect paraphrasing rather than deletion of information.

As the heuristic was exploratory, a lower score was not interpreted as evidence of lower reader burden. Instead, the rewriting experiment examined whether improving the computational score could also lead to undesirable changes in the content of the passage.

Study 4 separately assessed the stability of register-related comparisons under explicit feature definitions and standardisation.

3.6. Study 4: Independent Matched-Register Analysis

Study 4 investigated register-related differences between human-written and AI-generated texts in computing and cybersecurity, using standardised linguistic features and a matched-topic design. Three writing registers were compared: formal technical/academic writing, news and professional blogs.

The corpus comprised 60 passages. Fifteen pre-2022 human passages, five from each register, were used only to create the three human reference baselines. A separate evaluation set was then constructed using five topics represented in each of the three registers, giving 15 human evaluation passages. Each of these 15 topic-register combinations was matched with one ChatGPT passage and one Claude Opus 5 passage. The evaluation set thus contained 45 passages (15 per source category), alongside the 15 human baseline passages.

The five topics were zero trust architecture; Internet of Things security; digital identity and multi-factor authentication; cybersecurity risk management and the NIST Cybersecurity Framework; and privacy risk management and the NIST Privacy Framework (Fagan et al., 2020; Grassi et al., 2017; National Institute of Standards and Technology, 2018, 2020; Rose et al., 2020). Human passages were drawn from publicly available NIST and NIST CSRC material published no later than 2021. NIST states that its publications are generally Federal Government works and provides guidance on their reuse and attribution (National Institute of Standards and Technology, 2026). In the tables and figures, the formal technical/academic register is labelled "academic", although it consists mainly of NIST Special Publications, NISTIR reports and related formal technical writing.

ChatGPT and Claude received the same topic and register instructions. Each prompt requested one paragraph without specifying a word count. ChatGPT passages were generated using GPT-5.6 Sol, while Claude passages were generated using Anthropic Claude Opus 5 at the High effort level. Neither model was instructed to imitate human writing or to evade AI detection. In the Study 4 figures, GPT is used as a short label for passages generated using ChatGPT.

The Study 4 corpus consists of 60 fixed passages, each with a recorded SHA-256 hash to support verification.

Feature extraction followed four preprocessing rules. Straight and curly apostrophes were treated consistently, obvious spacing errors around hyphenated words from PDF extraction were repaired, common contractions were counted rather than all words containing apostrophes, and discourse markers were matched as complete expressions rather than as substrings. These steps affected preprocessing only; the stored source passages were not changed.

The primary analysis compared 19 linguistic features across human, ChatGPT and Claude passages for each register and matched topic. Mean values were reported together with direction counts showing how often an AI value was higher or lower than its matched human value. Formal significance tests and confidence intervals were not used because each register contained only five matched topics.

A secondary analysis assessed each evaluation passage's proximity to the human reference profile for each register. A reference profile, or centroid, was calculated from the five baseline human passages in each register. The features were standardised using the 15 human baseline passages before calculating distance, to account for their different measurement scales. Lower distance indicated greater similarity to the corresponding human register reference. For feature j, let μ_j and s_j denote, respectively, its mean and sample standard deviation (denominator $N-1$) across the $N = 15$ human baseline passages, and let

$\bar{b}_{rj}$ denote its mean among the five human baseline passages for register r. With J denoting the features whose baseline standard deviations are non-zero, the distance was calculated using a root-mean-square form of standardised Euclidean distance. This approach accounts for differences in feature scales (Milligan & Cooper, 1988) and was implemented in this study as follows:

$$D(x, r) = \sqrt{\frac{1}{|J|} \sum_{j \in J} \left(\frac{x_j - \mu_j}{s_j} - \frac{\bar{b}_{rj} - \mu_j}{s_j} \right)^2}. \quad (2)$$

The SMOG feature was excluded, leaving 18 features in the full analysis. This root-mean-square formulation divides the squared Euclidean distance by the number of included features before taking the square root. No feature had zero standard deviation across the full set of 15 human baseline passages.

In the leave-one-baseline-out check, one human baseline passage was omitted at a time within the register being evaluated (five omissions per register). For every omission, the pooled human-baseline means and sample standard deviations were recalculated from the remaining 14 passages, as were the register centroids. Features with zero standard deviation under the reduced baseline were omitted from that iteration's distance calculation. This occurred for discourse-marker density when news baseline passage BN04 was omitted, reducing the feature count for that iteration from 18 to 17. SMOG remained in the full feature dataset for consistency with Study 3, but was not used in these distances because the passages contained too few sentences for stable estimation.

Two sensitivity checks assessed dependence on feature selection and the small reference sample. The analysis was repeated without discourse-marker density because of its unusually large academic-register difference. A leave-one-baseline-out analysis then recalculated the reference profiles and distances after each omission. Both checks assessed the stability of the register comparisons; neither was intended to develop a new classifier.

The distance analysis was used as an exploratory comparison of how closely human-written and AI-generated passages matched the linguistic profile of their intended register. It was used to examine the stability of the register-relative assumption rather than as a reader-burden score, authorship detector or validated register classifier. The analysis was conducted using Python 3.13.15, NLTK 3.9.1, spaCy 3.8.16 with `en_core_web_sm` 3.8.0, and textstat 0.7.13. The complete software environment was recorded using `pip freeze`.

4. Results

Results are reported in study order: reader responses, linguistic comparisons, computational analyses and an independent register comparison.

4.1. Study 1: Reader-Response Findings

Study 1 showed a descriptive difference between the two passages on several reading-response measures. The six-item perceived reading-burden composite was 3.20 (SD 1.19) for the AI-generated passage and 2.05 (SD 1.13) for the human-written passage (55 complete records per condition; rounded from 3.2030 and 2.0545). Mean concentration-effort ratings were 3.20 versus 2.42, while the mean reported temptation to skim was 3.33 versus 2.27. Cronbach's α for the exploratory composite was 0.93 and 0.95 in the respective conditions. The observed mean difference of 1.15 had an exploratory Welch 95% confidence interval of 0.71 to 1.59; this conditional interval reflects participant variability for the two particular passages, not a general effect of AI authorship.

The main Study 1 outcomes are summarised in Table 3.

Table 3. Study 1 reader-response outcomes for the AI-generated and human-written passages. Values are mean (sample standard deviation); the composite averages six equally weighted items ($n = 55$ per condition).

Outcome	AI-Generated	Human-Written
Six-item perceived reading-burden composite	3.20 (1.19)	2.05 (1.13)
Concentration effort	3.20 (1.22)	2.42 (1.20)
Temptation to skim	3.33 (1.29)	2.27 (1.38)

The reader-response summaries are shown in Figure 2.

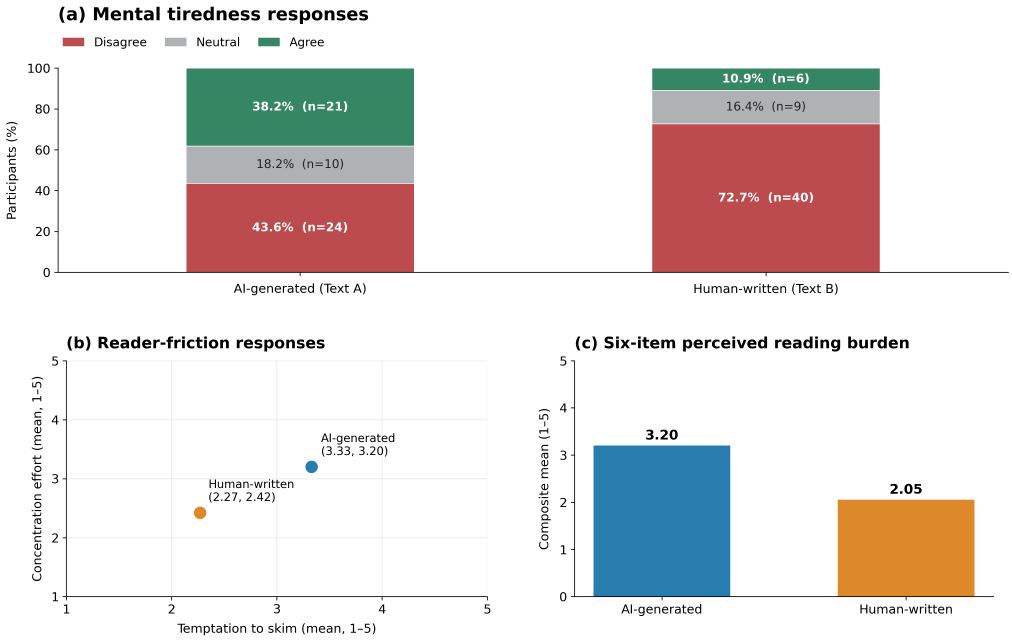


Figure 2. Study 1 reader-response summaries re-plotted from participant responses: (**a**) mental tiredness, grouped as Disagree (1–2), Neutral (3), and Agree (4–5); (**b**) mean concentration effort versus temptation to skim (each coded 1–5); and (**c**) mean exploratory six-item perceived reading-burden composite (1–5).

Among respondents assigned to the AI-generated passage, 38.2% selected Agree (4–5) for mental tiredness, compared with 10.9% of those assigned to the human-written passage. Eight participants (five in the AI condition and three in the human condition) reported not reading carefully. Excluding these responses as a sensitivity check gave six-item means of 3.12 (n = 50) and 1.96 (n = 52), respectively; the descriptive difference remained in the same direction. These findings concern the two specific passages and do not establish a causal effect of AI authorship across texts.

4.2. Study 2: Linguistic Findings Across Three Writing Types

Study 2 compared the linguistic characteristics of human-written and AI-generated passages in academic, news and blog writing. The analysis identified measurable differences in readability, lexical diversity and sentence structure within each writing type. TTR and MATTR were higher for the AI-generated passage in all three pairs, while the Flesch Reading Ease score and sentence-length variance also distinguished the passages, although their direction varied across registers. These results characterise the selected pairs and informed the subsequent computational investigation.

Table 4 summarises the selected linguistic measures for the three matched pairs. The table allows for the magnitude and direction of differences to be compared across registers.

Table 4. Selected linguistic measures for the three matched Study 2 pairs.

Register	Authorship	Flesch	Mean Words/Sentence	TTR	MATTR	Sentence-Length Variance
Academic	Human-written	19.98	25.30	0.55	0.71	138.68
Academic	AI-generated	3.72	24.60	0.62	0.76	72.84
News	Human-written	52.14	20.67	0.59	0.67	40.19
News	AI-generated	31.15	24.14	0.66	0.70	68.00
Blog	Human-written	62.33	20.26	0.53	0.73	149.80
Blog	AI-generated	70.93	15.92	0.59	0.76	88.24

Note: Sentence-length variance is expressed in words^2 because variance is calculated from squared deviations from the mean sentence length.

The additional academic and reference texts provided context for the matched academic pair. Across the five human-written texts, the Flesch Reading Ease score ranged from 7.20 to 43.09 and sentence-length variance ranged from 17.7 to 171.7. The AI-generated academic passage had a sentence-length variance of 72.84, which fell within this human range. The range demonstrates substantial variation among these human-written examples, in addition to the differences observed within the matched pair.

Because the five human texts differed in period and purpose, this range is presented as descriptive context rather than as a fixed academic standard.

4.3. Study 3: Computational Findings

Study 3 examined the correspondence between linguistic features and their reference values, source classification within an aggregate corpus, and the behaviour of score-guided rewriting. The results are reported separately for these three tasks.

4.3.1. Feature Definitions and Score Behaviour

Five feature abbreviations referred to different quantities in the Study 2 R analysis and the Study 3 Python procedure. Table 5 identifies these differences; the Python scoring procedure uses reference values calculated according to the Python feature definitions.

Table 5. Feature-mapping differences between the R (version 4.4.1) analysis and the Python implementation.

Transferred Key	R Meaning/Value	Python Feature/Value	Difference
GN	Group nouns, 18.0	Gunning Fog, 18.7042	Different measure despite similar value
CR	Content ratio, 0.61	Contraction rate, 0.0364	Different measure
ND	Nominalisation density, 0.02	Noun density, 0.2705	Different measure
HD	Hedging density, 0.04	Hapax density, 0.4088	Different measure
DD	Discourse-marker density, 0.004	Mean dependency depth, 3.6949	Different measure

The Python-based relative-deviation score was particularly sensitive to very small reference values. For example, AI-associated vocabulary density was 0.0225 for the AI-generated academic passage compared with an academic reference value of 0.0008, producing a raw relative deviation above 27 before the per-feature cap was applied. Normalising the feature weights from a total of 1.04 to 1.00 does not remove this sensitivity to small denominators.

The Flesch Reading Ease score can also take values below zero, making a ratio-based deviation unsuitable for this feature. This issue is examined further in the Study 4 sensitivity analysis.

4.3.2. Human–AI Source Classification

The source-classification analysis produced strong performance for the two feature-based classifiers. Logistic regression achieved an accuracy of 0.9865 and ROC AUC of 0.9968, while the RBF-SVM achieved an accuracy of 0.9915 and ROC AUC of 0.9977. The pretrained RoBERTa detector achieved an accuracy of 0.8315 and ROC AUC of 0.8564. Table 6 reports the classification results.

Table 6. Source-classification performance on the 2000-document evaluation set.

Model	Accuracy	Human Recall	AI Recall	ROC AUC	Brier Score
Logistic regression	0.9865	0.990	0.983	0.9968	0.0104
RBF-SVM	0.9915	0.989	0.994	0.9977	0.0068
RoBERTa detector	0.8315	0.968	0.695	0.8564	0.1432

Note: The Brier score measures the accuracy of predicted probabilities by comparing them with the observed class outcomes. Lower values indicate more accurate probability estimates.

The RBF-SVM produced the highest overall accuracy, AI recall and ROC AUC, while logistic regression produced the highest human recall. RoBERTa showed lower overall accuracy and substantially lower AI recall than the two feature-based classifiers.

The corresponding confusion matrices are shown in Figure 3.

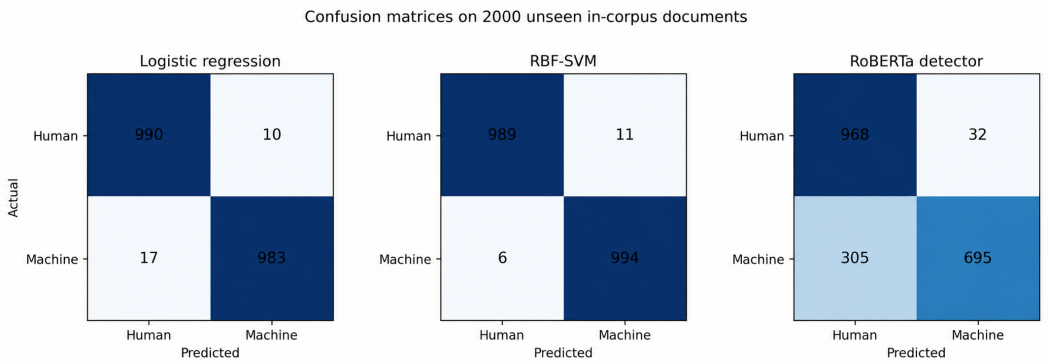


Figure 3. Confusion matrices for the three human–AI source classifiers on 1000 human-written and 1000 AI-generated documents. Rows represent actual classes and columns represent predicted classes. Darker blue cells indicate higher document counts, whereas lighter cells indicate lower counts.

The separate AI-generated academic Test 2 passage produced markedly different AI-class probabilities across the three models: 0.017 for logistic regression, 1.000 for the RBF-SVM and 0.064 for RoBERTa. The diagnostic passage therefore produced substantially greater disagreement between the classifiers than the main evaluation results.

4.3.3. Sentence Triage and Rewriting

Sentence triage and rewriting yielded two findings. The triage step distinguished between the AI-generated and human-written academic sentences in how frequently they were selected for revision, while the rewriting step revealed differences between score reduction and retention of original content words. Lexical retention does not establish preservation of meaning. At the 0.72 threshold, four of the five AI-generated academic sentences and one of the eight human-written academic sentences were flagged for possible revision. The variants differed in score reduction and lexical overlap; semantic preservation was not independently assessed.

Max-divergence PEGASUS reduced the heuristic score from 83.46 to 46.43 while retaining 18.5% of source content words, but the generated output contained repeated and incoherent sequences and was not a viable rewrite. This illustrates a failure of the score to

penalise unusable text. Generic T5 produced the largest numerical score reduction, from 83.46 to 44.43, with 64.1% content-word retention. Guarded PEGASUS had the highest content-word retention at 81.5%, but the score fell only to 73.38. Feature-directed T5 reduced the score to 49.49 with 63.0% content-word retention, and instruction leakage was observed in its output.

Table 7 summarises the rewriting results.

Table 7. Rewriting results for the AI-generated academic passage using the heuristic score.

Variant	Score Before	Score After	Content-Word Retention	Length Ratio	Observed Outcome
Max-divergence PEGASUS	83.46	46.43	0.185	0.49	Incoherent output; not a viable rewrite
Guarded PEGASUS	83.46	73.38	0.815	0.90	Higher content-word overlap; smaller score reduction
Generic T5	83.46	44.43	0.641	0.93	Largest score reduction; 64.1% lexical retention
Feature-directed T5	83.46	49.49	0.630	0.96	Instruction leakage observed

The relationship between score reduction and content-word retention is shown in Figure 4.

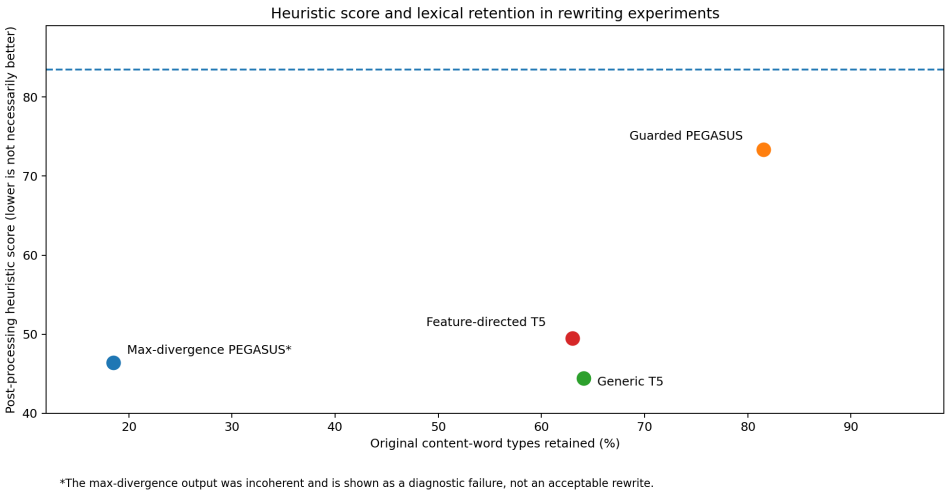


Figure 4. Relationship between the heuristic score after rewriting and content-word retention across the four rewriting variants. The blue dashed horizontal line indicates the original passage's heuristic score before rewriting (83.46), shown as a reference value.

4.4. Study 4: Independent Matched-Register Findings

Study 4 compared the three registers using standardised feature values and distances. The results showed differences in passage length, linguistic-feature patterns and proximity to the human register reference profiles across academic, news and blog writing.

4.4.1. Passage Length

Passage length differed across sources and registers. Claude produced the longest passages in all three registers, averaging 177.2 words in academic writing, 183.8 in news and 196.6 in blogs. ChatGPT passages were closer in length to the human-written passages, particularly in news, where the means were 127.0 and 124.6 words, respectively.

In the academic register, human-written passages averaged 113.0 words compared with 139.0 for ChatGPT. In blogs, human-written passages averaged 147.4 words compared with 139.6 for ChatGPT. The full means and ranges are reported in Table 8 and illustrated in Figure 5.

Table 8. Passage lengths across the three Study 4 registers.

Register	Human Mean Words	ChatGPT Mean Words	Claude Mean Words	Human Range	ChatGPT Range	Claude Range
Academic	113.0	139.0	177.2	105–132	132–144	171–183
News	124.6	127.0	183.8	106–161	124–131	169–200
Blog	147.4	139.6	196.6	144–154	135–142	187–201

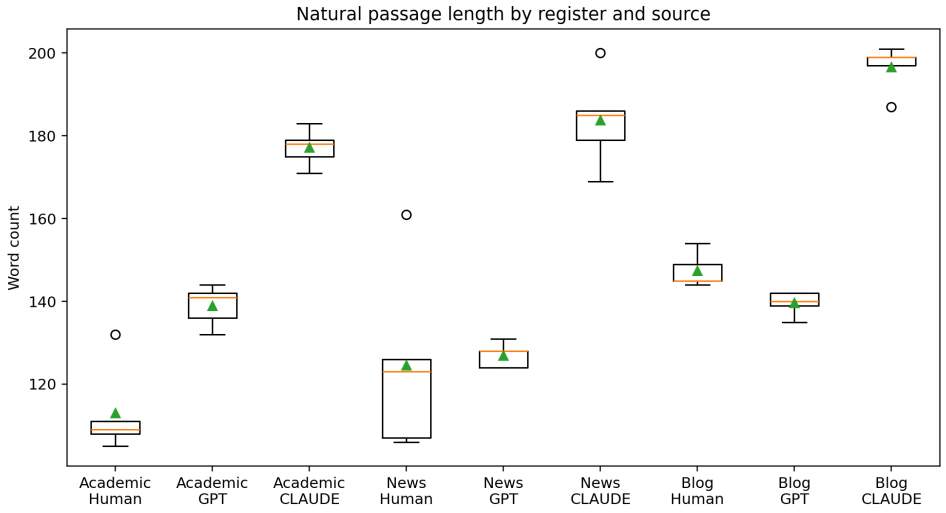


Figure 5. Word counts for the Study 4 evaluation passages across academic, news and blog registers. Boxes represent the interquartile range (25th–75th percentiles); orange horizontal lines indicate medians, green triangles indicate means, whiskers show the range of non-outlying observations, and open circles indicate outliers.

4.4.2. Feature-Level Comparisons

Feature-level comparisons showed differences between human-written, ChatGPT and Claude passages, but their direction varied across registers. Readability, lexical diversity, sentence structure and named-entity use all contributed to the observed patterns.

In the academic register, ChatGPT and Claude had a lower Flesch Reading Ease score than the human-written passages, with mean values of 8.41 for human text, −9.26 for ChatGPT and −14.38 for Claude. The Gunning Fog index was higher for both AI models, increasing from 20.48 for human-written passages to 24.35 for ChatGPT and 27.49 for Claude. Both models also showed higher TTR and MATTR and fewer named entities per sentence.

In news writing, ChatGPT and Claude again had a lower Flesch Reading Ease score and higher lexical-diversity measures than the human-written passages. Sentence-length variance differed more strongly between human and ChatGPT passages, at 80.67 and 23.72, respectively, whereas the Claude mean of 81.70 was close to the human value.

The pattern changed in blog writing. The Flesch Reading Ease score increased from 22.36 for human-written passages to 37.05 for ChatGPT and 53.24 for Claude, while the Gunning Fog index decreased from 19.06 to 16.07 and 14.74, respectively. ChatGPT and Claude also produced fewer named entities per sentence and fewer contractions than the human-written blog passages.

Table 9 reports selected feature means.

Table 9. Selected Study 4 feature means across human-written, ChatGPT and Claude passages.

Register	Source	Flesch	Fog	TTR	Sent. Variance	Entities/Sent.	Passive Markers/Sent.
Academic	Human	8.41	20.48	0.677	101.14	1.160	0.500
Academic	ChatGPT	−9.26	24.35	0.757	33.24	0.467	0.810
Academic	Claude	−14.38	27.49	0.710	65.76	0.147	1.120
News	Human	14.97	20.63	0.645	80.67	1.680	0.400
News	ChatGPT	6.98	20.51	0.758	23.72	0.330	1.081
News	Claude	0.56	25.54	0.701	81.70	1.160	0.760
Blog	Human	22.36	19.06	0.637	111.17	1.243	0.213
Blog	ChatGPT	37.05	16.07	0.739	59.12	0.275	0.731
Blog	Claude	53.24	14.74	0.649	91.54	0.464	0.486

Note: Values are descriptive means across five matched topics per register. Sentence-length variance is expressed in words^2; entity and passive-marker values are mean counts per sentence.

Across all 19 features, ChatGPT and Claude differed from the human-written passages in the same non-zero direction on 15 features in academic writing, 10 in news and 17 in blogs. Agreement between the two AI models was therefore highest in blog writing and lowest in news writing.

The standardised ChatGPT–human and Claude–human feature differences are shown in Figures 6 and 7.

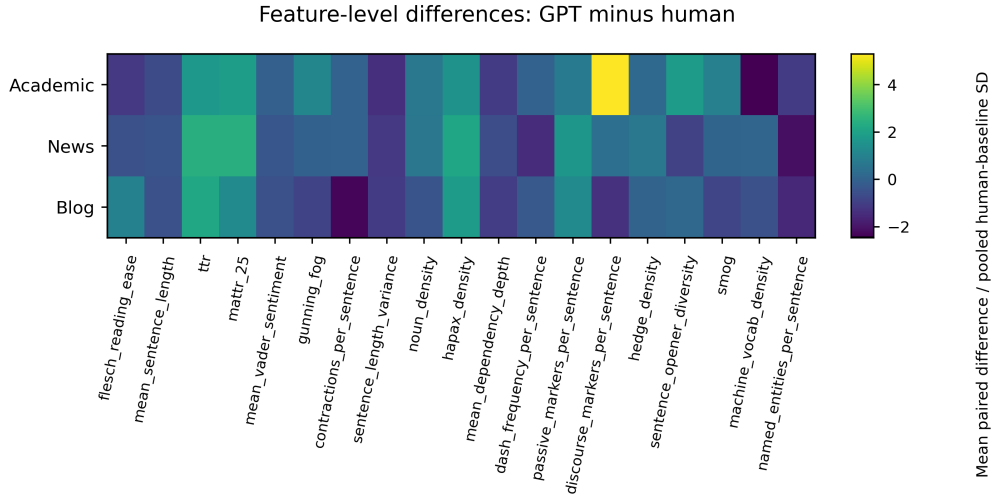


Figure 6. Standardised mean ChatGPT–human feature differences across the three Study 4 registers.

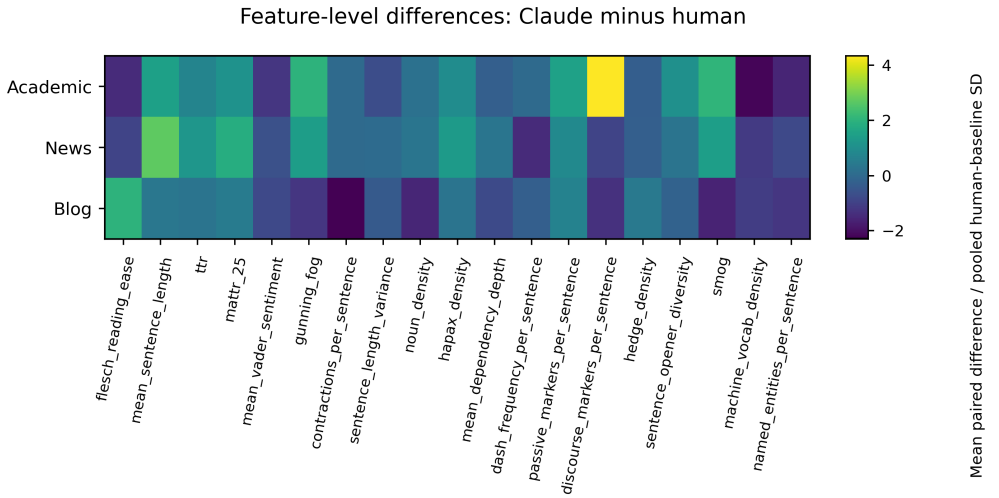


Figure 7. Standardised mean Claude–human feature differences across the three Study 4 registers.

4.4.3. Register-Distance Sensitivity Analysis

The register-distance results showed a clearer pattern at the group level than at the individual-passage level. For the human-written texts, the mean for each register was closest to its corresponding human register centroid. The academic mean distance was 1.310 compared with 1.410 and 1.418 for the other two centroids; the news mean was 1.173 compared with 1.196 and 1.262; and the blog mean was 1.445 compared with 1.534 and 1.474.

At the individual-passage level, however, the corresponding register was closest for seven of the fifteen human-written passages (46.7%): four of five academic, one of five news and two of five blog passages. Across all human-written passages, the mean distance to the corresponding register was 1.310 compared with 1.382 for the mean distance to the other two registers, giving a margin of 0.073.

For the AI-generated passages, the corresponding human register centroid was closest for nine of fifteen ChatGPT passages and ten of fifteen Claude passages. However, academic-register assignments were correct for none of the five ChatGPT passages and only one of the five Claude passages; news and blog passages accounted for most AI matches. These counts are descriptive and do not demonstrate reliable classification beyond this small corpus. The passage-level results are summarised in Table 10.

Table 10. Passage-level proximity to the corresponding human register centroid in Study 4.

Source	Register	Closest to Corresponding Register	Distance to Corresponding Register	Mean Distance to Other Registers	Margin
Human	Academic	4/5	1.310	1.414	0.105
Human	News	1/5	1.173	1.229	0.055
Human	Blog	2/5	1.445	1.504	0.059
Human	Overall	7/15	1.310	1.382	0.073
ChatGPT	Overall	9/15	1.358	1.472	0.114
Claude	Overall	10/15	1.182	1.372	0.189

Note: The margin is the mean distance to the other two registers minus the distance to the corresponding register. A positive value therefore indicates greater proximity to the corresponding register.

The human cross-register distance matrix is shown in Figure 8. Each register had the lowest group-mean distance to its corresponding centroid, although the separation from alternative centroids was smallest for news writing.

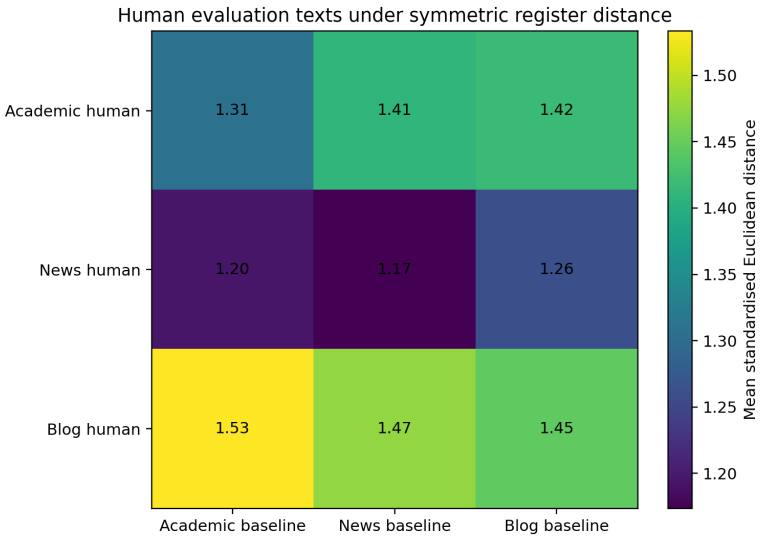


Figure 8. Mean standardised Euclidean distances of human evaluation passages from the three human register centroids. Lower values indicate greater proximity.

The first sensitivity check produced a marked change in the academic ChatGPT–human and Claude–human differences. When discourse-marker density was included, the mean academic difference in distance to the corresponding register was +0.494 for ChatGPT and +0.377 for Claude. After that feature was removed, the differences changed to −0.012 and −0.096, respectively. Human passage-level recovery changed from 7/15 to 8/15, while all three human group means remained closest to their corresponding register centroids.

The leave-one-baseline-out results were more consistent for the academic and blog registers. Across the five omissions, the academic ChatGPT–human difference ranged from +0.215 to +0.546 and the Claude–human difference from +0.104 to +0.481. In blogs, the corresponding ranges remained negative for ChatGPT (−0.552 to −0.407) and Claude (−0.641 to −0.398). Claude also remained negative in news (−0.250 to −0.082), whereas the ChatGPT news difference crossed zero, ranging from −0.030 to +0.306.

4.4.4. Assessment of the Study 3 Ratio-Based Heuristic

The assessment indicated that the Study 3 ratio-based score was unsuitable as the main Study 4 outcome. In the academic register, the five human baseline passages had a mean Flesch Reading Ease score of 8.05, whereas several ChatGPT and Claude passages had negative Flesch values. Because the Flesch Reading Ease score can fall below zero, it does not support meaningful ratio-based comparison in this form. As a result, the Flesch component alone added an average of +11.80 points to the ChatGPT–human academic difference and +13.36 points to the Claude–human academic difference. When Flesch was removed and the remaining weights were renormalised, the academic differences changed from +5.62 to −6.84 for ChatGPT and from +8.34 to −5.56 for Claude.

The assessment also identified zero-valued baselines for discourse-marker density in academic and blog writing, AI-associated vocabulary density in news, and the original contraction proxy in blogs. Zero baselines invoked a different scoring rule from small non-zero baselines. Taken together, these findings indicate that the ratio-based score was not a stable cross-register measure.

The numerical sensitivity of the score is summarised in Table 11 and Figure 9. These values characterise the sensitivity of the Study 3 heuristic and are separate from the standardised-distance outcomes in Study 4.

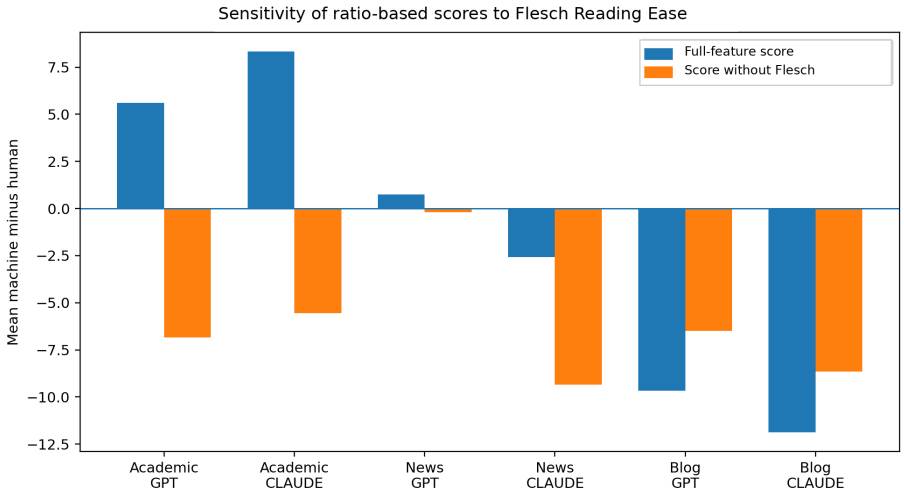


Figure 9. Sensitivity of the Study 3 ratio-based AI–human score to removal of the Flesch Reading Ease score.

Table 11. Sensitivity of the Study 3 ratio-based score to removal of the Flesch Reading Ease score.

Register	Family	Full-Feature AI–Human Score	Without Flesch	Flesch Contribution
Academic	ChatGPT	+5.62	−6.84	+11.80
Academic	Claude	+8.34	−5.56	+13.36
News	ChatGPT	+0.74	−0.21	+0.93
News	Claude	−2.58	−9.35	+5.87
Blog	ChatGPT	−9.66	−6.49	−3.80
Blog	Claude	−11.88	−8.65	−4.07

Note: After Flesch was removed, the remaining weights were renormalised. The values in the *Without Flesch* column are therefore recalculated scores and should not be interpreted as simple subtraction from the full-feature score.

5. Discussion

The four studies address related but distinct aspects of evaluating AI-generated texts intended for learners. Study 1 provides direct, passage-specific evidence of reader response, whereas Studies 2–4 examine linguistic characteristics, computational scoring, source classification, rewriting and register-sensitive comparisons. These later studies do not measure reader burden directly and cannot validate the Study 1 composite.

5.1. Reader Response and Linguistic Form Require Separate Evidence

Study 1 provides the only direct reader-response evidence. Students who read the AI-generated news passage gave higher mean perceived reading-burden ratings than those who read the human-written passage. Because only one passage represented each condition, this finding describes a difference between two texts rather than a general effect of AI authorship. This distinction is consistent with cognitive load research, which shows that processing demand can depend on the presentation of information, prior knowledge and task conditions rather than source alone (Qian et al., 2026; Sweller, 1988).

Study 2 adds a different form of evidence by analysing the linguistic characteristics of the same news pair alongside matched academic and blog passages. The human-written and AI-generated news texts differed in readability, lexical diversity and sentence structure. Across the three writing contexts, lexical-diversity differences were similar in direction, whereas readability and sentence-length variance did not follow a consistent pattern. This variation is consistent with previous research showing that linguistic differences between human-written and AI-generated text depend on model, dataset and register (Demir & Egbert, 2026; Herbold et al., 2023; Muñoz-Ortiz et al., 2024; Reinhart et al., 2025; Reviriego et al., 2024). Linguistic features can therefore describe how texts differ, but they cannot by themselves establish why readers responded differently.

This distinction matters in higher education. Knowing that a feedback comment, explanation or learning resource was generated with AI does not determine whether it is clear, appropriately detailed or suitable for a particular learner. When reader burden is the outcome of interest, evidence must therefore come from readers themselves rather than being inferred from authorship or linguistic form.

5.2. Linguistic Differences Depend on Register

Study 4 extends the register-related analysis from Study 2 using a separate corpus and two AI model families. Some feature-level human–AI differences changed direction across writing contexts, particularly readability, although several lexical measures showed the same directional difference across the selected registers. ChatGPT and Claude had a lower Flesch Reading Ease score than the human-written passages in academic and news writing, whereas the direction reversed in blogs. Other linguistic features also varied by register and model family. This pattern is consistent with previous register research showing that

AI-generated language does not reproduce human linguistic patterns uniformly across genres and disciplines (Demir & Egbert, 2026; Reinhart et al., 2025).

The register-distance analysis requires a more cautious interpretation. At the group level, each set of human-written passages was closest to its corresponding human register reference. At the individual level, however, only seven of the fifteen human-written passages were closest to the corresponding register centroid. ChatGPT and Claude reached the corresponding centroid in nine of fifteen and ten of fifteen passages, respectively, but this was concentrated in news and blog writing: academic-register recovery was zero of five for ChatGPT and one of five for Claude. The total counts should therefore not be interpreted as evidence of better overall register quality.

The sensitivity checks further showed that some of these distance differences depended on the features and baseline passages used. Removing discourse-marker density largely eliminated the academic ChatGPT–human distance difference (+0.494 to −0.012) and reversed the Claude–human difference (+0.377 to −0.096). In addition, excluding individual baseline passages caused the small ChatGPT–human difference in news writing to cross zero. The register-distance measure is therefore useful for examining linguistic similarity and sensitivity, but the present results do not support treating it as a validated measure of register quality or reader burden.

5.3. The Ratio-Based Heuristic Was Not a Stable Measure

The Study 4 assessment of the Study 3 ratio-based score identified further limitations beyond those associated with feature counting and preprocessing. The Flesch Reading Ease score can cross zero and therefore cannot be meaningfully compared through ratios to small register means. Features with exact-zero baseline values also followed a different scoring rule from features with small non-zero baselines. These properties made the resulting score sensitive to the choice and scale of individual features. Most notably, removing Flesch and renormalising the remaining weights reversed the academic ChatGPT–human and Claude–human differences. The Study 3 composite should therefore not be interpreted as a measure of reader burden or register distance.

The limitation is methodological rather than merely technical. Combining several plausible linguistic features does not by itself produce a meaningful measure of learner experience or linguistic suitability. Such a measure would require evidence that its components are combined appropriately, that its behaviour is stable across relevant contexts and, crucially, that it relates to the outcome it is intended to represent. Without that evidence, the score should be treated as an exploratory computational heuristic rather than as a measure of reader burden.

5.4. Source Classification and Educational Suitability

The source-classification results illustrate an important distinction between identifying authorship and evaluating educational suitability. Logistic regression and the RBF-SVM achieved high accuracy within the same aggregate source corpus used for model development, showing that the selected linguistic features could distinguish human-written and AI-generated documents effectively under those evaluation conditions. This success, however, does not show that the detected texts were easier to read, better suited to learners or lower in reader burden.

The three classifiers assigned markedly different AI probabilities to a separate AI-generated academic diagnostic passage. This disagreement illustrates why strong classification performance in one evaluation setting may not transfer consistently to other texts. Previous detector research similarly reports variation across generators, domains and text conditions (Dugan et al., 2024; Weber-Wulff et al., 2023).

For higher education, source classification may therefore be useful when authorship itself is the question, such as in some academic-integrity contexts. It should not be treated as a measure of the quality of feedback, explanations or learning materials. Whether a text is appropriate for learners requires separate evidence about its linguistic form, preservation of meaning and reader response.

5.5. Rewriting Requires Preservation of Meaning

The rewriting results show why improvement in an internal linguistic score should not be treated as evidence that a text has improved for learners. The max-divergence PEGASUS output was incoherent despite obtaining a substantially lower score, while the generic T5 condition showed a lower score alongside lower retention of original content words. This finding concerns lexical overlap and does not establish loss of essential information or meaning. Because the score itself was not a validated measure of reader burden, a lower value cannot be interpreted as evidence that the rewritten passage became easier or less burdensome to read.

For educational writing, revision must preserve the information that the learner needs, including relevant facts, qualifications, causal relationships and necessary terminology. The content-word retention measure used in Study 3 could flag differences in lexical overlap, but it could establish neither substantive information loss nor semantic equivalence. Linguistic features may therefore help identify passages that warrant review, but the quality of a rewrite requires separate evidence about preservation of meaning and, where reader burden is concerned, direct evidence from readers.

5.6. Prompting and Audience Adaptation

Prompt formulation is another potential source of variation that the present design cannot distinguish from passage source. Study 1 used one recorded prompt and one generated passage; it did not vary prompt specificity, users' AI competence, iterative refinement or editorial intervention. Study 4 likewise used fixed topic-and-register prompts to generate matched passages, rather than testing how different prompting strategies changed output quality. In teaching practice, prompts could be adjusted to a learner's prior knowledge, the required degree of technical detail or the intended genre. Controlled prompt comparisons with independent reader outcomes would be needed to evaluate this possibility. This limitation reinforces the distinction between comparing observed passages and attributing their properties to AI authorship alone.

5.7. Reader Burden and AI Fatigue

The findings concern immediate reading responses, not the broader construct of AI fatigue. Lau et al. describe AI fatigue as a multidimensional response associated with sustained human–AI interaction (Lau et al., 2026), while Miranda et al. describe a staged process of strain arising through repeated academic AI use (Miranda et al., 2026). Study 1, by contrast, measured fatigue- and effort-related responses after a single short passage and did not use a validated AI-fatigue instrument.

Reader burden is therefore the more appropriate term for the immediate responses examined here. The finding is that the two passages produced different fatigue- and effort-related ratings, not that participants developed AI fatigue. Whether repeated exposure to particular forms of AI-generated text contributes to broader AI fatigue remains a question for future reader studies.

5.8. Implications for Higher-Education Policy and Practice

The implications for policy should be interpreted cautiously. The four studies did not evaluate university policy, and the reader-response study did not use teaching material produced by a university. Guidance from the Russell Group, QAA, Oxford and UCL is therefore used here to establish the wider context of permitted AI use, academic integrity, disclosure, responsibility and reliable information (Quality Assurance Agency for Higher Education, 2026; Russell Group, 2023; University College London, n.d.; University of Oxford, 2025). Similar priorities appear in United States university guidance, including plagiarism, data privacy and the limitations of generative AI (Wang et al., 2024). The present study does not establish how widely such policies address reader suitability.

The findings identify an additional consideration when AI-generated material is provided directly to students. Evaluation can distinguish whether AI use is permitted and disclosed, whether the content is factually accurate, whether the language fits the intended register and learner, whether revision preserves meaning, and how readers respond to the resulting text. These are distinct evaluation questions, not policy criteria validated by this study, and they cannot be reduced to a single validated score on the present evidence.

This distinction also has practical implications because educational texts serve different purposes. A formal technical explanation, policy notice, short formative comment and conversational support message do not have the same linguistic expectations. Research on educational AI similarly shows that outcomes depend on how systems are implemented and on the needs of learners (Anwar et al., 2023; Ates, 2026; Dicheva et al., 2026; Sajja et al., 2024; Yilmaz et al., 2026). Reader suitability should therefore be considered in relation to the intended learner, purpose and writing context rather than inferred from AI authorship or a general linguistic threshold.

5.9. Synthesis Across the Research Questions

The answer to RQ1 concerns responses to the two passages used in Study 1; it is not an estimated population-wide difference attributable to AI authorship. RQ2 documents register-dependent linguistic differences, without identifying a causal explanation for those responses. For RQ3, successful source classification and reductions in a linguistic heuristic answer narrow computational questions, while lexical overlap during rewriting does not establish whether important meaning was retained. RQ4 demonstrates that the register-distance findings depend in part on feature choice and the reference corpus. Taken together, the findings suggest a sequence of evaluation questions: verify content accuracy and provenance, examine register and intended audience, check meaning after rewriting, and obtain reader evidence whenever claims about reader burden are made. This proposed sequence follows from the exploratory interpretation and has not been prospectively validated.

5.10. Future Research and Policy Development

Future research should investigate reader burden directly using larger and more varied sets of AI-generated educational texts. Studies should include multiple passages per condition, different disciplines and registers, several AI models and more diverse student populations. Self-reported effort measures should be combined with outcomes such as comprehension and reading time and, where appropriate, behavioural measures. Future pre-registered designs with multiple independent passages per condition and explicit outcome definitions would support stronger generalisation than this exploratory comparison.

Further linguistic work should test candidate features against reader outcomes rather than assuming that distance from a small human reference set indicates suitability. Register baselines should include broader institutional, disciplinary and publication contexts, with attention to factors such as organisational style and passage length. Rewriting should

likewise be evaluated using independent measures of preservation of meaning together with reader testing before any claim of reduced reader burden is made.

A systematic analysis of university AI guidance could examine whether and how institutions address readability, register appropriateness, cognitive demand and reader testing when AI is used to produce material for learners. Such evidence is needed before conclusions can be drawn about how widely reader suitability is addressed in current policy and could support future guidance based on measured educational outcomes.

5.11. Limitations

Study 1 provided direct reader-response evidence from 110 university students and used random assignment to compare the two passage conditions. However, only one human-written and one AI-generated passage represented the conditions. The findings therefore describe differences between these two stimuli rather than a general effect of AI authorship. The retained participant-level records allowed for direct verification of the six-item composite and its completeness. Nevertheless, the six items were selected as an exploratory aggregate rather than validated as a single psychological scale; they cover tiredness, effort, anticipated fatigue and perceptions of unnecessary text length. No measure of comprehension or sustained AI fatigue was collected, and the study cannot establish that the observed difference was caused by AI authorship rather than by the particular passages or their generation conditions.

Study 2 examined more than the shared news pair by comparing human-written and AI-generated texts across academic, news and blog writing and by considering additional variation within human academic/reference texts. However, three matched pairs are too few to characterise the full distribution of linguistic features within each register. The wider human comparison set also contains texts written in different periods and for different purposes, so it should be interpreted as descriptive evidence of human variation rather than as a contemporary academic benchmark.

Study 3 showed that the 19 linguistic features could support strong source classification within the evaluation corpus, while the rewriting results showed that reducing a heuristic score did not necessarily yield coherent or satisfactory prose. The ratio-based heuristic also showed sensitivity to small reference values and feature definitions. However, the classical classifiers were evaluated on documents drawn from the same aggregate source corpus used for model development. RoBERTa had different training conditions, and possible overlap between its HC3 training material and HC3-derived material in the evaluation source could not be ruled out. The heuristic also relied partly on manually specified feature rules and word lists, and neither the score nor the rewritten passages were directly validated against reader outcomes. These factors limit claims about generalisation beyond the settings examined here.

Study 4 strengthened the register analysis by using a separate 60-passage corpus, two AI model families, standardised feature processing and sensitivity checks. However, it remains an exploratory analysis based on five human baseline passages and five evaluation topics per register. The human passages were drawn from one institutional source family, NIST, so organisational style may partly overlap with register differences. The “academic” category also represents formal NIST technical/academic writing rather than peer-reviewed journal articles, and publication periods differed across the baseline registers.

These limitations restrict generalisation but do not erase the distinction between the types of evidence considered. Reader response, linguistic form, source classification and preservation of meaning provide different kinds of evidence and require separate evaluation.

6. Conclusions

This study examined AI-generated text in higher education through four complementary forms of evidence. Study 1 provided direct reader-response evidence, with the AI-generated news passage receiving higher fatigue- and effort-related ratings than the human-written passage, although the finding is limited to the two texts tested. Study 2 showed that linguistic differences between human-written and AI-generated passages varied across academic, news and blog writing. Study 3 demonstrated that linguistic features could support strong source classification within a defined evaluation setting, while also showing that score-guided rewriting could reduce an internal linguistic score alongside lower retention of original content words, without demonstrating whether meaning was preserved. Study 4 extended the register analysis using a separate corpus and two AI model families, finding both recurring and register-dependent human–AI feature differences.

The findings indicate why these forms of evidence should not be treated as interchangeable. Authorship classification addresses where a text is likely to have come from, linguistic features describe aspects of how it is written, rewriting requires preservation of meaning, and reader burden requires evidence from readers. The ratio-based heuristic examined in this study did not provide a stable measure that could combine these questions into a single score. Similarly, the Study 1 responses concern immediate reading effort and fatigue, not persistent AI fatigue.

For higher education, the main implication is that the suitability of AI-generated feedback, explanations and learning materials cannot be inferred from authorship or from a single linguistic measure. Evaluation should consider factual accuracy, linguistic and register fit, preservation of meaning and reader response according to the intended learner and educational purpose. Larger reader studies, broader register samples and direct links between linguistic features and learner outcomes are needed before automated text-quality measures can support educational or policy decisions.

Author Contributions: Conceptualisation, M.P.H. and I.U.R.; methodology, M.P.H. and I.U.R.; software, M.P.H.; validation, I.U.R. and A.A.; formal analysis, M.P.H.; investigation, M.P.H.; data curation, M.P.H.; writing—original draft preparation, M.P.H.; writing—review and editing, M.P.H., I.U.R. and A.A.; visualisation, M.P.H.; supervision, I.U.R.; project administration, M.P.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was approved as a low-risk student research project under the Middlesex University Dubai research ethics process on 12 January 2026. No protocol code is listed on the approval form.

Informed Consent Statement: All participants received an information sheet and consent form before taking part. Participation was voluntary, withdrawal was permitted, and no personally identifiable information was collected.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to applicable institutional policies, research ethics approval, participant consent and data-protection requirements.

Acknowledgments: Parts of Studies 1–3 were initially developed in M.P.H.'s undergraduate project portfolio at Middlesex University Dubai, which is cited in the manuscript. Generative AI systems were used to create the machine-generated text stimuli described in the Methods. Historical machine-generated stimuli were produced with ChatGPT, while Study 4 included passages generated with OpenAI GPT-5.6 Sol and Anthropic Claude Opus 5, with the High effort setting used for the Claude generation. During the write-up of this manuscript, the authors used OpenAI ChatGPT and Anthropic Claude for language revision and assistance with analysis code. The authors reviewed and edited all AI-assisted outputs, checked numerical results against the retained data, notebooks and computational

outputs, verified cited literature against source records, and take full responsibility for the content of the publication.

Conflicts of Interest: A.A. serves as a Guest Editor of the Special Issue to which this manuscript is submitted. In accordance with MDPI editorial policy, A.A. will have no role in the editorial handling, reviewer selection, peer-review process or editorial decision for this manuscript. The other authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
AUC	Area under the receiver operating characteristic curve
GPT	Generative pretrained transformer
HC3	Human ChatGPT Comparison Corpus
MATTR	Moving-average type-token ratio
NIST	National Institute of Standards and Technology
NLTK	Natural Language Toolkit
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QAA	Quality Assurance Agency for Higher Education
RBF	Radial basis function
ROC	Receiver operating characteristic
SMOG	Simple Measure of Gobbledygook
SVM	Support-vector machine
TTR	Type-token ratio
UCL	University College London
VADER	Valence Aware Dictionary and sEntiment Reasoner

References

Abdulla, H. S. E. M. (2017). *Determinants and consequences of knowledge sharing among employees in a UAE national oil and gas company* [Doctor's thesis, United Arab Emirates University]. Available online: https://scholarworks.uaeu.ac.ae/all_theses/639/ (accessed on 4 November 2026).

Akbulut, C., Weidinger, L., Manzini, A., Gabriel, I., & Rieser, V. (2024). All too human? Mapping and mitigating the risk from anthropomorphic AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 13–26. [CrossRef]

Alabed, A., Javornik, A., & Gregory-Smith, D. (2022). AI anthropomorphism and its effect on users' self-congruence and self-AI integration: A theoretical framework and research agenda. *Technological Forecasting and Social Change*, *182*, 121786. [CrossRef]

Al Hammadi, M. A. M. (2017). *The role of human resources in organizational restructuring: Perceptions of HR department, line managers, and employees* [Doctor's dissertation, United Arab Emirates University]. Available online: https://scholarworks.uaeu.ac.ae/all_dissertations/93/ (accessed on 4 November 2026).

Anwar, A., Rehman, I. U., Nasralla, M. M., Khattak, S. B. A., & Khilji, N. (2023). Emotions matter: A systematic review and meta-analysis of the detection and classification of students' emotions in STEM during online learning. *Education Sciences*, *13*, 914. [CrossRef]

artem9k. (n.d.). *AI text detection pile* [Data set]. Hugging Face. Available online: https://huggingface.co/datasets/artem9k/ai-text-detection-pile (accessed on 4 September 2026).

Ates, H. (2026). Human-centered GenAI feedback design in higher education: A multisite experiment on direct, reflective, and hybrid approaches to scientific argumentation. *International Journal of Educational Technology in Higher Education*, *23*, 38. [CrossRef]

BBC News. (2010, August 12). *Russians remember Kursk submarine disaster, 10 years on*. Available online: https://www.bbc.com/news/world-europe-10950701 (accessed on 4 November 2026).

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*, 94–100. [CrossRef]

Darwin, C. (1999). *The life and letters of Charles Darwin* (F. Darwin, Ed.). Project Gutenberg. Available online: https://www.gutenberg.org/cache/epub/2087/pg2087-images.html (accessed on 4 September 2026).

Demir, N. Y., & Egbert, J. (2026). Register alignment of ChatGPT-generated academic texts. *Applied Corpus Linguistics*, *6*(1), 100174. [CrossRef]

Dicheva, N. K., Rehman, I. U., Husamaldin, L., & Aleshaiker, S. (2026). Artificial intelligence in nursing education: A scoping review of intelligent tutoring systems—Evidence and implementation gaps. *Nurse Education Today*, *162*, 107056. [CrossRef]

Dugan, L., Hwang, A., Trhlík, F., Zhu, A., Ludan, J. M., Xu, H., Ippolito, D., & Callison-Burch, C. (2024, August 11–16). *RAID: A shared benchmark for robust evaluation of machine-generated text detectors*. The 62nd Annual Meeting of the Association for Computational Linguistics (pp. 12463–12492), Bangkok, Thailand. [CrossRef]

Evans, M., & Bacon, W. (2018). *When Harry married Meghan: Royal wedding*. Available online: https://www.whatwegandidnext.com/when-harry-married-meghan-royal-wedding/ (accessed on 4 November 2026).

Fagan, M., Megas, K. N., Scarfone, K., & Smith, M. (2020). *IoT device cybersecurity capability core baseline* (NISTIR 8259A). National Institute of Standards and Technology. [CrossRef]

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233. [CrossRef]

Golzar, J., Noor, S., & Tajik, O. (2022). Convenience sampling. *International Journal of Education and Language Studies*, *1*, 72–77. [CrossRef]

Grassi, P., Fenton, J., Newton, E., Perlner, R., Regenscheid, A., Burr, W., Richer, J., Lefkovitz, N., Danker, J., Choong, Y., Greene, K., & Theofanos, M. (2017). *Digital identity guidelines: Authentication and lifecycle management* (NIST Special Publication 800-63B). National Institute of Standards and Technology. [CrossRef]

Grieve, J. (2023). Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory*, *19*, 47–77. [CrossRef]

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill.

Haguma, M. P. (2026). *Beyond AI detection: Quantifying and mitigating AI fatigue through a reader-centric approach* [Unpublished undergraduate project portfolio, Middlesex University Dubai].

Healy, J. (1912). *Insula sanctorum et doctorum: Or, Ireland's ancient schools and scholars* (6th ed.). Sealy, Bryers and Walker. Available online: https://www.gutenberg.org/cache/epub/41073/pg41073-images.html (accessed on 4 September 2026).

Hello-SimpleAI. (n.d.). *ChatGPT detector RoBERTa* [Machine learning model]. Hugging Face. [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*, 18617. [CrossRef]

Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., & Jung, M. F. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, *13*, 5487. (Erratum in 2023, *Scientific Reports*, *13*, 16616). [CrossRef]

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*, 216–225. [CrossRef]

Jabarian, B., & Imas, A. (2025). *Artificial writing and automated detection* (NBER Working Paper No. 34223). National Bureau of Economic Research. [CrossRef]

Johnston, H., Wells, R. F., Shanks, E. M., Boey, T., & Parsons, B. N. (2024). Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity*, *20*, 2. [CrossRef]

Kalra, M. P., Mathur, A., & Patvardhan, C. (2024, July 27–28). *Detection of AI-generated text: An experimental study*. The 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (pp. 552–557), Gwalior, India. [CrossRef]

Klimova, B., & Pikhart, M. (2025). Exploring the effects of artificial intelligence on student and academic well-being in higher education: A mini-review. *Frontiers in Psychology*, *16*, 1498132. [CrossRef]

Kolomaznik, M., Petrik, V., Slama, M., & Jurik, V. (2024). The role of socio-emotional attributes in enhancing human-AI collaboration. *Frontiers in Psychology*, *15*, 1369957. [CrossRef]

Krakowski, S., Haftor, D., Luger, J., Pashkevich, N., & Raisch, S. (2025). Human-centered artificial intelligence: A field experiment. *Management Science*, *72*, 57–72. [CrossRef]

Lau, G. R., Kasturiratna, K. T. A. S., Goh, A. Y. H., Tong, E. M. W., & Hartanto, A. (2026). AI fatigue in human-AI interaction: Conceptual framework, scale development and validation, and associations with AI engagement. *Computers in Human Behavior Reports*, *23*, 101186. [CrossRef]

McLaughlin, G. H. (1969). SMOG grading—A new readability formula. *Journal of Reading*, *12*, 639–646.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*, 181–204. [CrossRef]

Miranda, J. P. P., Parreño, E. B., & Rivera, J. G. (2026). Defining AI fatigue in academic contexts: Dimensions, indicators, and a stage-based model using grounded theory. *International Journal of Learning, Teaching and Educational Research*, *25*(5), 91–107. [CrossRef]

Mukherjee, S., & Dušek, O. (2024). Text style transfer: An introductory overview. *arXiv*. [CrossRef]

Mukherjee, S., Lango, M., Kasner, Z., & Dušek, O. (2024). A survey of text style transfer: Applications and ethical implications. *arXiv*. [CrossRef]

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, *57*, 265. [CrossRef]

National Institute of Standards and Technology. (2018). *Framework for improving critical infrastructure cybersecurity, version 1.1* (NIST CSWP 6). National Institute of Standards and Technology. [CrossRef]

National Institute of Standards and Technology. (2020). *NIST privacy framework: A tool for improving privacy through enterprise risk management, version 1.0* (NIST CSWP 10). National Institute of Standards and Technology. [CrossRef]

National Institute of Standards and Technology. (2026). *Library FAQs: Reprint permission for NIST publications*. Available online: https://www.nist.gov/nist-research-library/library-faqs (accessed on 4 September 2026).

Paas, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71. [CrossRef]

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ...Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. [CrossRef]

Qian, W., Yang, F., Cao, Y., Yi, L., Gu, R., & Wang, Z. (2026). Generative AI and cognitive load in education: A systematic review of WoS/SSCI-indexed studies through the lens of cognitive load theory. *Frontiers in Psychology*, *17*, 1921504. [CrossRef]

Quality Assurance Agency for Higher Education. (2026). *The perfect storm: AI, assessment and a sector under pressure*. Available online: https://www.qaa.ac.uk/news-events/policy-and-leadership/state-of-the-nation (accessed on 4 September 2026).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67. Available online: https://jmlr.org/papers/v21/20-074.html (accessed on 4 November 2026).

Ragu-Nathan, T. S., Tarafdar, M., Ragu-Nathan, B. S., & Tu, Q. (2008). The consequences of technostress for end users in organizations: Conceptual development and empirical validation. *Information Systems Research*, *19*, 417–433. [CrossRef]

Reinhart, A., Markey, B., Laudenbach, M., Pantusen, K., Yurko, R., Weinberg, G., & Brown, D. W. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences of the United States of America*, *122*, e2422455122. [CrossRef]

Reviriego, P., Conde, J., Merino-Gómez, E., Martínez, G., & Hernández, J. A. (2024). Playing with words: Comparing the vocabulary and lexical diversity of ChatGPT and humans. *Machine Learning with Applications*, *18*, 100602. [CrossRef]

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero trust architecture* (NIST Special Publication 800-207). National Institute of Standards and Technology. [CrossRef]

Russell Group. (2023). *Russell Group principles on the use of generative AI tools in education*. Available online: https://russellgroup.ac.uk/media/6137/rg_ai_principles-final.pdf (accessed on 4 September 2026).

Sajja, R., Sermet, Y., Cikmaz, M., Cwiertny, D., & Demir, I. (2024). Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education. *Information*, *15*, 596. [CrossRef]

Shahrzadi, L., Mansouri, A., Alavi, M., & Shabani, A. (2024). Causes, consequences, and strategies to deal with information overload: A scoping review. *International Journal of Information Management Data Insights*, *4*, 100261. [CrossRef]

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257–285. [CrossRef]

Terčon, L., & Dobrovoljc, K. (2025). Linguistic characteristics of AI-generated text: A survey. *arXiv*. [CrossRef]

University College London. (n.d.). *Three categories of GenAI use in assessment*. Available online: https://www.ucl.ac.uk/teaching-learning/generative-ai-hub/three-categories-genai-use-assessment (accessed on 4 September 2026).

University of Leeds Library. (2020). *The rise of the snow dragon: Assessing the impact of China's increasing role in the arctic on the future development of its regional governance regimes*. University of Leeds. Available online: https://resources.library.leeds.ac.uk/final-chapter/dissertations/polis/pied3750_example3.pdf (accessed on 4 September 2026).

University of Oxford. (2025). *AI use in summative assessment*. Available online: https://www.ox.ac.uk/about/how-we-are-run/policies-and-statements/policy-hub/ai-use-in-summative-assessment (accessed on 4 September 2026).

Vieriu, A. M., & Petrea, G. (2025). The impact of artificial intelligence (AI) on students' academic development. *Education Sciences*, *15*, 343. [CrossRef]

Wang, H., Dang, A., Wu, Z., & Mac, S. (2024). Generative AI in higher education: Seeing ChatGPT through universities' policies, resources, and guidelines. *Computers and Education: Artificial Intelligence*, *7*, 100326. [CrossRef]

Webb, M. (2023). *Generative AI—A primer*. Jisc. Available online: https://www.jisc.ac.uk/reports/generative-ai-a-primer (accessed on 4 September 2026).

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*, 26. [CrossRef]

Wester, J., de Jong, S., Pohl, H., & van Berkel, N. (2024). Exploring people's perceptions of LLM-generated advice. *Computers in Human Behavior: Artificial Humans*, *2*, 100072. (Erratum in 2026, *Computers in Human Behavior: Artificial Humans*, *9*, 100352). [CrossRef]

Yilmaz, M., Temur, H. B., Emmungil, L., Çelik, E., Gauthier, A., & Cukurova, M. (2026). Supporting self-regulated learning through generative AI feedback in online higher education: The importance of student perceptions of the source of feedback. *International Journal of Educational Technology in Higher Education*, *23*, 16. [CrossRef]

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. *Proceedings of the 37th International Conference on Machine Learning*, *119*, 11328–11339. Available online: https://proceedings.mlr.press/v119/zhang20ae.html (accessed on 4 November 2026).

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.