Perceiving the humanness: reader evaluation of AI-generated versus human-written romantic texts

Student Name: Mingjue Liu
Student Number: 696495

Supervisor: Dr. Marlen Komorowski-Albert

Master Media Studies - Media & Creative Industries
Erasmus School of History, Culture and Communication
Erasmus University Rotterdam

Master's Thesis
June 2025

Word Count: 12690

Perceiving the humanness: reader evaluation of AI-generated versus human-written romantic texts

ABSTRACT

The rapid advancement of generative AI (GenAI), particularly large language models (LLMs) like GPT-4o, has revolutionized content creation, enabling efficient eBook production. However, this raises concerns about authorship ambiguity, quality, and the proliferation of low-quality AI-generated content on Amazon Kindle Direct Publishing. Prior research on AI-text detection has focused on logical, academic, or short texts (e.g., AI-generated messages and hotel reviews), leaving a gap in understanding AI's capability to mimic human writing in emotionally rich genres like romance, which relies on nuanced emotional expression, authenticity, and tone of voice.

While AI can generate content quickly in massive amounts with high-cost efficiency, its ability to replicate the emotional depth and authenticity of human-written romance narratives remains uncertain. This creates challenges for readers, platforms, and regulators in assessing authorship and ensuring content quality. Which leads to the research question: What are readers' perceptions of AI-generated and human-written romantic narratives based on (a)linguistic naturalness, (b)coherence, and (c) emotional tone?

A mixed-methods approach was employed, combining quantitative surveys (5-point Likert scales) and qualitative thematic analysis. Participants (n = 85) evaluated two 170-word romantic reunion passages: *Text A* (Chapter 2 from Nicholas Sparks' *The Notebook*) and *Text B* (GPT-4o-generated). Metrics included linguistic naturalness (vocabulary simplicity, sentence structure), coherence (logical flow, consistency), and emotional tone (authenticity, intensity). Text order was randomized to minimize bias. More specifically, quantitative analysis involves paired-sample t-tests and MANOVA tests, coupled with Shapiro-Wilk and Wilcoxon tests analyze survey responses. And qualitatively, thematic analysis of open-ended explanations is used.

The results showed that only 51.8% correctly identified the human text (*Text A*), with a marginal advantage (11.8%) over AI text attribution, indicating difficulty in distinction. In terms of linguistic naturalness, AI text used significantly *more* diverse vocabulary ($d = 1.32$, $p < .001$), *longer* sentences ($d = 1.58$, $p = .002$), and *more descriptive details* ($d = -0.35$, $p = .033$). Human text was qualitatively associated with simpler vocabulary and sentences. With coherence metrics, no significant differences emerged in ease of reading, logical consistency, or transitions ($p > .05$), which reinforces the importance of developing genre-specific frameworks. The measurement of emotional tone was quantitatively indistinguishable, but qualitative analysis revealed that AI text showed more emotional intensity, while human text was more authentic. Lastly, em-dashes emerged as a prominent indicator of AI-writing by participants, serving as a novel detection cue.

In conclusion, genre is a critical aspect in AI-text detection: romance narratives demand metrics beyond lexical or coherence markers, but more emotional and micro-punctuation measures. GPT-4o narrows the human-AI gap in romantic narration but struggles to replicate emotional authenticity. Em-dashes emerged as a reliable, non-theoretical AI indicator, showing systematic biases in training data. A mixed-method approach is essential as both measures provide a comprehensive insight into this study. Practical implications demand that writers prioritize authentic storytelling, educators teach critical reading, and developers refine LLMs by suppressing robotic markers.

KEYWORDS: *AI, Generative AI, LLM, Romance, eBooks*

Table of contents

1. Introduction

The rapid development of generative artificial intelligence (GenAI), particularly large language models (LLMs) like GPT-4o, has profoundly revolutionized content creation, enabling individuals to produce text-based outputs efficiently. As Flower and Hayes (1981) proposed in the *Cognitive Process Theory of Writing*, the three key elements in the thinking process are planning, translation, and reviewing (p. 369). In detail, planning refers to idea generation, organization, and setting specific goals for the final writing product (p. 373). Translating refers to transforming ideas into texts that are comprehensible to readers (p. 374). And reviewing involves evaluating and revising with the written passages (p. 374). As Chakrabarty et al. (2024) observed, LLM (GPT-3.5) was most helpful in the translation and reviewing stages by providing functions such as content elaboration, drafting feedback, and critiques, as well as spotting inconsistencies in logic (pp.16, 19, 22). These functions enable the writers to bring their creative ideas to fruition more easily.

One of the prominent implications of LLMs is in the ebook creation, which is also a form of AIGC. Artificial Intelligence-Generated Content (AIGC) refers to a content creation process that uses AI to assist or replace humans in generating digital content (Wu et al., 2023, p.1). The content generation includes a vast range of outputs such as text, images, audio, video, code, and even interactive 3D (Wang et al., 2023, p.280). By simply providing prompts, AI will start producing desired content based on the understanding of prompts and specific requirements (Cao et al., 2023, p.1).

Since LLMs can produce massive amounts of content in an extremely short time, their output volume and speed far exceed those of human writers (Wang et al., 2023, p.286; Cao et al., p.8). This capability allows individuals to generate content at a much faster rate and in significantly greater quantities. Furthermore, as the barrier for content generation has been drastically lowered by simply inputting prompts, cost efficiency is largely enhanced because the need for human labor was largely reduced, making content generation much more affordable (Wang et al., 2023, p.286). As a result, several creators have already successfully monetized through AI-generated children's books published on Amazon Kindle Direct Publishing (KDP) (Bensinger, 2023).

However, the advancement of LLMs has also raised serious ethical concerns regarding their utilization and implications in content creation. As central issue was the ambiguity of authorship of AI-generated works. Scholars (Hosseini, Resnik & Holmes, 2023) argued that AI, lacking free will and thus incapable of holding any moral or legal accountability, cannot be viewed as authors (p. 462). This unaccountability of LLMs for their outputs underscores the critical necessity for human supervision, particularly when content is intended for public distribution. The potential for harm is evident, as a prominent model of GPT, GPT-3, has demonstrated cases of generating inaccurate and biased information due to its inability to make critical judgments and distinctions among all data they were trained on (Brown et al., 2020, pp. 26, 36).

The consequences of publishing unsupervised content can lead to the spread of inaccurate information or misinformation to the public. In the case of fake news generation, scholars have suggested LLMs were likely be a new way to foster fake news generation, which would negatively affect societal trust and enhanced polarization among people in different societal groups (Sallami, Chang & Aïmeu, 2024, Introduction, para.2). These findings collectively emphasized on the prominent moral and legal risks in utilizing Generative AI for content generation and demanding proper regulation from platform.

The challenges posed by these ethical considerations and the urgency for content regulations are important on large self-publishing platforms. For example, the consistent monitoring and regulation of the ebooks quality within the Amazon Kindle Direct Publishing represents a long-standing challenge, even before the recent advancement of AI. Bad actors, exemplified by people like Luca de Stefani and Mikkelsen twins, have already demonstrated the capacity to monetize through the sale of low-quality ebooks. Their methods included manipulating rankings of these books by leaving fake five-star reviews, using extensive keyword optimizations, and even selling online courses on how to exploit Amazon's algorithm loopholes (AI generated ebooks changing publishing on Amazon, 2024).

The rapid development of generative AI has critically amplified this enduring problem by adding additional elements to the complexity of the problem. The ambiguous authorship, coupled with the speed and cost-efficiency of LLM models, has created an optimal ground for intellectual property violations and platform exploitation. For instance, there was a significant lag between the upload of low-quality AI-generated and fraudulent ebooks, and the systematic procedure of detection and removal. This asymmetry in time enabled individuals to have opportunities for unlawful monetization, like selling unauthorized summary books and fake biographies that are generated based on the published books of real authors before the system could detect and took down those works (The Authors Guild, 2024).

To regulate the oversupply of ebooks with unclear quality, Amazon has implemented certain measures. These include setting a limitation on the number of e-book submissions to three per day (Edwards, 2023) and demanding that the authors indicate if the work is AI-generated or AI-assisted before submission (Amazon, n.d.-b). However, Amazon's current approach falls short of efficiency. The platform does not specify any robust method for AI text detection, especially when there are numerous AI humanization tools that can easily help those books to bypass the check. Furthermore, Amazon does not publicly label ebooks as AI-generated, AI-assisted, or human-written on its website. The lack of process transparency and inefficiency in authorship labels have caused trouble for readers to make an immediate distinction regarding the content they consume.

Although the advancement of generative AI has resulted vast amount of low-quality works, it has simultaneously proposed a critical evaluation of the capacity of LLMs in creating high-quality ebooks, and their potential to truly mimic or even surpass human authors in writing complex content. Prior studies focus on human perception of text generated by GPT-2 and GPT-3, particularly in the linguistic contexts of formal writing such as English newspapers and generative messages (Muñoz-Ortiz, Gómez-Rodríguez & Vilares, 2024, p. 265; Hohenstein & Jung, 2020, p. 4). Given that AIGC has been shown to lack emotional depth in music production and revealed inadequacy in creativity because it was trained by pre-existing information (Wu et al., 2023, pp.8-9). There is a compelling need to investigate the performance of LLMs in generating emotionally rich text genres such as romance novels.

Having identified the gap in acknowledging AI's capacity in creating emotionally nuanced content, this study first investigates the most recent top ten best-selling ebooks on Amazon in the United States market, concluding that thrillers, romance, and fiction are the most popular genres (Amazon, n.d.-a). This aligns with previous research indicating that romance is one of the most popular genres, especially among female readers (Fletcher, Driscoll, & Wilkins, 2018, p. 997; Thelwall, 2019, p. 403). Romantic novels are rich in emotional narration, often utilizing intricate techniques such as internal monologues and perspective shifts to convey deep feelings and relational dynamics (Carter, 2018, p. 90). These stylistic elements are closely relevant to the constructs of Appraisal Theory (White, 2015), which provides a framework for understanding how emotions and stances are conveyed within narratives (p.2). These characteristics make romance novels a suitable material for examining the subtle emotional cues readers use to differentiate between human and AI writing.

While prior research has extensively investigated textual differences between human and AI-generated content, focusing on aspects like complexity and variety of the vocabulary, presence of structural inconsistency (Yildiz Durak, Eğin & Onan, 2025, p. 7; Yang et al, 2024, p. 13). This study

aims to further explore areas of distinction. Particularly, this paper seeks to explore whether readers can distinguish between AI-generated and human-written romantic narratives based on their emotional nuance, tone of voice, and overall logical flow. The investigation seeks to potentially bridge the research gap in studying the capability of advanced large language models like GPT-4o to mimic human writing in emotionally rich genres like romance.

Building upon the identified phenomena of potential oversupply of low-quality ebooks created by LLMs and the documented limitations of AI-generated content across various genres. This research aims to investigate readers' perceptions and evaluation of AI-generated versus human-written romantic text. To achieve this, a mixed-methods approach was employed, utilizing a survey with multiple-choice questionnaires and a qualitative open-ended question. To systematically understand and analyze the distinctions readers perceive between human and AI-generated narratives, especially in the case of romance narratives, a solid theoretical framework is essential. The following section outlines the existing theories that help establish the investigation for textual and emotional characteristics between texts.

2. Theoretical Framework

Previous research has extensively investigated the key differences between AI-generated texts and human-written texts, broadly categorizing these distinctions into textual and emotional aspects. These researchers studied texts generated by various LLMs such as GPT-3, GPT-4, Gemini, and BingAI. Various study methods were employed, ranging from quantitative linguistic analysis to comparative studies, as well as initial comparison (eg, Muñoz-Ortiz & Gómez-Rodríguez & Vilares, 2024; Yang et al., 2024; Reviriego et al., 2024). For instance, Muñoz-Ortiz et al. (2024) utilized a linguistic analysis of English news texts by comparing human-written news with news generated by six different LLMs. The results showed that human writers tend to demonstrate more diversity in lexical choices, sentence lengths, and reveal more negative emotions than LLMs (p. 264).

In terms of linguistic patterns, several scholars have diverse points of view. Markowitz, Hancock and Bailenson (2024) conducted a quantitative approach that studied the hotel reviews created by GPT-3.5 and human written ones on TripAdvisor, the results suggested that AI-generated texts not only appear to use more descriptive, in terms of the frequency in using adjectives, but also slightly less readable than human-generated text, meaning AI-generated text is slightly higher in number and complexity of words (pp. 66, 68, 72). However, contrasting findings emerge from research using a different research method and text form. Specifically, Yildiz Durak et al. (2025) employed a qualitative case study approach comparing 30 discussion articles written by students, ChatGPT 3.5, Gemini, and BingAI based on the same question (pp. 3- 4). And the result shows that human-produced content uses more singular words on average and longer sentences, suggesting a potentially higher level of linguistic complexity (p. 7). Therefore, Markowitz et al. (2024) suggested that human-written texts use simpler words, while Yildiz Durak et al. (2025) believed humans used more difficult words in their writings.

In terms of lexical diversity and sentence length, Muñoz-Ortiz et al. (2024) have similar findings to Yildiz Durak et al. (2025). By comparing English news articles generated by six different LLMs versus human-written news articles with a qualitative approach, Muñoz-Ortiz et al. (2024) suggested that humans used fewer restricted words and longer sentences compared to LLMs (p. 10). Moreover, in the quantitative research conducted by Reviriego et al. (2024), comparing 126 TOEFL essays written by GPT-3.5 and humans across topics from computer science, medicine, to finance (pp. 3- 4). Their result also suggested that humans use more diverse vocabulary in all categories of the materials (p. 4). These findings collectively reinforce the pattern observed by Yildiz Durak et al. (2025, p.7) that human-written texts show greater lexical diversity and use longer sentences than output from LLMs.

The contrast in findings in lexical simplicity is likely due to the difference in the genre of text materials. Hotel reviews tend to focus on experience sharing, in other words, utilizing more colloquial and easier-to-understand words. Whereas English news, TOEFL, and discussion essays need to use formal academic words in writing. Since the genre for this research study is romantic narratives, which focus on storytelling rather than making logical arguments, it is assumed that human writing will tend to use simpler words in romantic texts.

Additionally, the research of Muñoz-Ortiz et al. (2024) highlights that AI tends to have more repetitive structures in expressions, such as repetition of bigrams (p. 8).

In the case of coherence, all the findings suggested similar insights. Mündler et al. (2023) conducted an experimental evaluation on the text descriptions for Wikipedia entities across multiple LLMs, including GPT-4 and GPT-3.5. (p. 6). The results showed that although self-contradictions regularly occurred across LLMs, the more advanced the model is, the fewer errors were discovered (p.7). Both Ma et al. (2023) and Yang et al. (2024) used a mixed-method approach. While Ma et al. (2023) studied academic abstracts written by humans and mainly GPT-3 (p. 2), Yang et al. (2024) compared 50 argumentative essays composed by native English speakers and 50 generated by GPT-3.5 (p. 5).

Both papers suggested that GPTs suffer from hallucination problems, meaning they create information that does not exist, and they include inaccurate and incomplete information (Ma et al., 2023, p. 2; Yang et al., 2024, p. 2). This similarity of the results is likely due to the utilization of similar versions of LLMs (GPT-3 and GPT-3.5) and study material (argumentative essays and academic abstracts). However, the significant deficiency in inconsistency and occurrence of self-contradictions are frequently found in logic-centric texts, and the application of this measurement has not been studied in the context of romantic narration yet.

In the emotional aspect, there was less research done on this metric compared to linguistic features and consistency. Nevertheless, there is still several insightful information from relevant studies. Matthews and Volpe (2023) conducted semi-structured interviews among 16 participants who were experienced with assessing academic works (p. 87). These participants were asked to evaluate two human-authored texts and two texts generated by GPT-3.5 (p. 96). The results show that “voice” (41.8%) was a prominent differentiator between the two, emphasizing the uniqueness of human tone of voice and a direct association with humanness (p. 89).

Besides tone of voice, positive emotion was strongly associated with AI. Hohenstein and Jung (2020) conducted an experimental design that studied people’s perception of message replies by humans and AI-generated replies by Google Allo (p. 4). Since text replies from AI were perceived with higher trust among participants, the researchers assumed that it likely resulted from priming

effects, meaning people tend to have more trust over messages with greater positive emotions (p. 8). In other words, AI-generated messages have more positive emotions than human-composed messages. In addition, AI-written newspaper tends to have less toxic and aggressive emotions, such as anger and fear (Muñoz-Ortiz et al.,2024, p.1). And LLMs are designed to only generate positive hotel reviews (Markowitz et al., 2024, p.76).

Moreover, the emotional authenticity of humans was perceived to be higher than AI. Kirk and Givi (2025) conducted an experimental design on the perceived authenticity of emails produced by salespeople and ChatGPT (p. 5). And the results show that since authenticity is strongly correlated with internal state, which is absent in AI, resulting emails written by salespeople have more perceived emotional authenticity (p. 9).

These results suggested that although AI and human-generated content show both positive and negative emotions, AI-generated content tends to have more positive emotions but lacks in tone of voice and emotional authenticity.

However, there are several limitations to the generalizability and validity of these studies. Firstly, most of the text materials utilized in the studies are academic and formal texts, such as news, TOEFL essays, academic abstracts, and customer service emails (Muñoz-Ortiz et al., 2024; Reviriego et al., 2024; Yang et al., 2024; Kirk & Givi, 2025). These text forms often emphasize logical flow, formality of structures and languages, resulting in the limitation of the generalizability of the frameworks. For instance, when studying texts that are richer in emotional rather than logical aspects, the existing frameworks are likely to be unsuitable or insufficient.

Secondly, the performance of different LLMs largely depends on the systematic design and the limitations of their training datasets (Ma et al., 2023, p.4). However, there have not been sufficient investigations on the influence of their training datasets in terms of cultural and geographic bias, the existence of a specific tone of voices or syntactic structures (Muñoz-Ortiz et al., 2024, p.265). These results variation in the performance among types of LLMs, which further questions the generalizability of the existing findings and theoretical frameworks.

Lastly, the rapid evolution of LLM models suggests that the current findings may easily be invalid due to the advancement of new models. Since the validity of current results may diminish over time, it poses a constant challenge for any subject regarding the study of GenAI.

Building upon these critical insights and addressing the limitations within the existing literature, this research paper investigates how readers perceive AI-generated versus human-written texts within the emotionally rich romance genre. This investigation will focus on three key dimensions: linguistic naturalness, cohesion, and emotional tone. A comprehensive theoretical framework would be established to study the perceived quality of romantic texts. To provide a solid

foundation for analyzing the key dimensions of linguistic naturalness, cohesion, and emotional tone within this specific genre, this research integrate four complementary theoretical perspectives: Media Naturalness Theory (Kock, 2005) for understanding the essence of perceived naturalness in language; Processing Fluency Theory (Alter & Oppenheimer, 2009) for linking linguistic features to cognitive effort of processing; Cohesion Theory (Halliday and Hasan, 1976) for analyzing textual coherence; and Appraisal Theory (Martin & White, 2005) for understanding emotional expression, attitude and tone of voices. The following subsections will explain the relevance of these theories to this research topic.

2.1 Media naturalness theory

The media naturalness theory proposed by Kock (2005) argued that media naturalness is determined by five factors: communication occurred in the same location, instant feedback, presence of facial expressions and body language, as well as the ability to hear the speech. And the suppression or absence of these factors, due to the nature of online communication, leads to a decline in the perceived naturalness in communication (p. 121). To compensate for the absence of some factors, such as body language and facial expression, people need to take more cognitive effort, defined as mental activity, during online communication than in a face-to-face setting (pp. 121-122). Mental activity is indirectly measured by the time it requires for people to convey the same idea in online versus offline communication settings, which is referred to as "fluency" by Kock (p. 122). And the lack of media naturalness induced communication ambiguity because individuals tend to have different personal interpretations of the absent information (pp. 122-123). Moreover, less physiological arousal will be induced due to the absence of factors like facial expression, body language and voice (pp. 123-124). Thus, the content with low naturalness leads to higher cognitive effort in processing and delivering information, higher communication ambiguity, and less physiological arousal (p. 124).

2.2 Processing fluency theory

Aligned with the findings of Kock (2005), Alter and Oppenheimer (2009) suggested that the more difficult it is to process information, the more cognitive effort is required, thus leading to a lower level of fluency (pp. 220-222). In the processing fluency theory, they stated that when people process information, texts that are easier to process have a greater degree of fluency, leading to greater perceived truth, preference, and confidence when judged by people (pp.219, 227-229).

A prior experiment conducted by Oppenheimer (2006) shows that when replacing original words with their longer and more complex synonyms, in college admission essays, it enhances the difficulty in comprehension, resulting in a lower degree of perceived intelligence of the authors (pp.

140-143). On the other hand, when replacing difficult words (words composed by nine or more letters) from dissertations with their simpler version, the level of perceived intelligence of the writers increased (pp. 146-147). These results, coupled with an additional fluency manipulation experiment, suggested that fluency is a key determinant in the judgment of the text; people tend to have more positive judgments over the content that requires lower effort in comprehension (p.151). These findings (Kock, 2005; Alter & Oppenheimer, 2009) suggested that fluency and naturalness are positively correlated. People have more positive perceptions of easy-to-understand texts because they require less mental effort in comprehension and are more like natural face-to-face interaction.

To define the metrics for measuring linguistic naturalness, Alter and Oppenheimer (2009) suggested that linguistic fluency is determined by lexical, syntactic, and orthographic fluencies. In other words, the utilization of simple and familiar words, simple grammatical structure, and letters instead of letter-like symbols contributes to a higher level of fluency, respectively (p. 225).
Based on the prior studies on the differences of AI-generated versus human-written texts (Matthews & Volpe, 2023, p. 86; Yildiz Durak et al., 2025, p. 2; Markowitz et al., 2024, p.72), the human-produced text appears to use longer sentences, more diverse vocabulary, and simpler words. In combination with media naturalness theory (Kock, 2005) and process fluency theory (Alter & Oppenheimer, 2009), it suggests that text that requires less cognitive effort to understand tends to show more linguistic naturalness.

2.3 Cohesion Theory

After discovering the fluency and linguistic naturalness of words and sentences, this research intends to study the overall content quality by analyzing the cohesion and logical flow of both passages. As a prior study (Mündler et al., 2023) showcased that large language models (LLMs) tend to generate notable self-contradictory information by showing inconsistency between sentences (pp. 1,7). However, this incident has been reduced significantly by the advancement of the language model; for instance, GPT-4 and ChatGPT outperformed other language models in their capability in generating consistent information (p.8). Therefore, it is crucial to build up the metrics of consistency based on the cohesion theory proposed by Halliday and Hasan (1976).

They argued that cohesion has a semantic property, as “it refers to relations of meaning that exist within the text, and that define it as a text “(p. 4). Moreover, they proposed that there are five kinds of “ties”, referring to “the occurrence of a pair of cohesively related items” (p. 3). And those cohesive ties are reference, substitution, ellipsis, conjunction, and lexical cohesion (p.4). By utilizing those cohesive ties properly, the writer gets to achieve the “continuity that exists between one part of the text and another” (p. 299).

In other words, by using correct cohesive ties, one can ensure the logical flow throughout the sentences and the overall text, which eventually conveys a meaningful message to its readers. Therefore, the consistency can be broken down into two aspects: using proper cohesive ties to build logical information and the capability to convey meanings without breaking coherence. These two aspects are measured by the smoothness in sentence transitions and the continuity in conveying logical ideas.

2.4 Appraisal theory

Appraisal theory is a framework for analyzing how texts convey emotional assessments, the intensity of these assessments, and the way writers engage with their viewpoints (White, 2015, p.2). The theory consists of three components: attitude, referring to the emotional reaction; engagement, writer's reaction to other perspectives; and graduation, the intensity of the tone (pp. 2-5).

There are three dimensions of attitude (Martin & White, 2005): affect, judgement, and appreciation (pp. 42-43). They argued that affect refers to the assessment of positive and negative feelings (p. 42), judgment refers to "attitudes towards behavior" (p. 42), and appreciation means the evaluation of the aesthetics of phenomena (pp. 43-44). While affect measures the polarity of the emotions, judgement and appreciation assessed the "institutionalized feelings" (p. 45). By analyzing these three attitudinal meanings collectively, the emotional attitude of the author can be assessed.

Engagement refers to "speakers/ writers adopt a stance towards the value positions being referenced by the text and for those they address" (Martin & White, 2005, p. 92). In other words, engagement discusses how authors position themselves with respect to other voices and positions; thus, this concept helps measure the tone of the voices of each text.

As prior research suggested (Muñoz-Ortiz et al., 2024, p. 264; Hohenstein & Jung, 2020, p. 8; Markowitz et al., 2024, p. 65), AI-generated messages tend to have a more positive attitude and utilize more affective and descriptive terms in texts. The tendency to generate positive text is likely due to the responsibility and ethical constraints of AI. As Markowitz et al. (2024) described, when using AI to generate hotel reviews, they were informed that AI was not allowed to generate negative reviews due to its programming restriction (p. 68).

Therefore, this research aims to test the affect and appreciation factors in emotion narratives by measuring the occurrence of positive and negative emotions, as well as tone of voice. Since the prior finding of Matthews and Volpe (2023) suggested that human-written context exhibits a unique "voice" (p. 86), it is assumed that the human-written text would likely have a distinct personal tone of voice and clear emotional attitude.

2.5 Hypotheses

Linguistic naturalness

Based upon Media Naturalness Theory (Kock, 2005) and Processing Fluency Theory (Alter & Oppenheimer, 2009), which emphasize the role of simpler, diverse, and naturally flowing language in reducing cognitive effort and enhancing naturalness, in couple with prior findings on linguistic differences (Muñoz-Ortiz et al., 2024, p.10; Markowitz et al., 2024, p.72; Reviriego, 2024, p.4), the following hypothesis regarding linguistic naturalness is proposed:

H1: It is assumed that human-written text will (a) use simpler, (b) more diverse vocabulary, (c) more adjectives and descriptive details, (d) longer, and (e) less repetitive sentence structures than AI-generated text.

Coherence

Guided by Cohesion Theory (Halliday & Hasan, 1976), which suggested proper lexical ties ensure logical continuity and the empirical findings on the occurrence of inconsistency and self-contradictory information (Mündler et al., 2023, p. 7; Yang et al.,2024, p. 2), the following hypothesis is proposed:

H2: It is assumed that human-written text will (a) be perceived as easier to follow, (b) exhibit greater coherence through consistent ideas and (c)smoother transitions between sentences and paragraphs, and (d) present less contradictory information compared to AI-generated text.

Emotional Tone & Expressiveness

Rooted in Appraisal Theory (Martin & White, 2005), which provides a framework for analyzing emotional attitude and voice, and considering prior evidence on emotional differences (Hohenstein & Jung, 2020, p. 8; Kirk & Givi, 2025, p. 9; Matthews et al., 2023, p. 89), the following hypothesis is predicted regarding emotional tone and expressiveness:

H3: It is assumed that human-written text will (a) express a stronger emotional attitude, (b) more authentic emotions, (c) adopt a more distinct personal tone of voice, and (d) exhibit less distinctly positive emotions than AI-generated text.

Attribution Accuracy

Overall, by combining the predictions from H1 to H3, where human texts are predicted (Kock, 2005; Alter & Oppenheimer, 2009; Martin & White, 2005) to exhibit superior naturalness, coherence, and emotional expressiveness. The readers will be expected to show the following trait:

H4: Participants are more likely to assign text A as human-written text after reading both texts.

3. Research Design

3.1 Research Approach and Design

A mixed-method approach is employed to investigate readers' perceptions of human and AI-generated romance narratives. This method combines quantitative data analysis from a 5-point Likert-scale survey with qualitative thematic analysis of open-ended responses. This approach was utilized because it merges the benefits of numerical measurements with meaningful interpretation. As stated by Babbie (2017), the quantitative approach enables explicit observation by using statistical data analyses (p. 25), hypothesis testing, and the discovery of causal relationships (p. 234). The survey, being one of the prominent forms of quantitative approaches, particularly demonstrates strong reliability (p. 287), generalizability (p. 286), feasibility in interpreting findings from large samples, especially with the "self-administered ones" (p. 286), and flexibility in data analysis (p. 286). These traits allow research studies to show consistent results across repeated measures, reduce the chance of unreliable observations made by researchers, and make the observations and results applicable to larger populations and groups.

Additionally, Babbie (2017) suggested that a qualitative approach, such as thematic analysis, allows researchers to gain deep insights into social phenomena and develop theories grounded in empirical observations (p. 297). By examining the proposed hypothesis and generating new insights through observations, the integration of qualitative and quantitative methods fosters a more comprehensive and robust understanding and interpretation of results (pp.26, 310). In this instance, not only will quantitative analysis assess the linguistic naturalness, coherence, and emotional expressiveness features, but qualitative measurements will also facilitate the discovery of new observations.

While the research is inspired by the concern about the quality of AI-generated romantic novels and their increasing presence in self-publishing platforms, evaluating entire novels is not feasible within the scope of this study due to time constraints and the cognitive overload for participants. Thus, this study adopted a paragraph-level comparison to maintain participant engagement and ensure manageable data collection. This approach is consistent with prior research that analyzes short-form text to assess linguistic and emotional features (e.g., Matthews & Volpe, 2023, p. 86; Hohenstein & Jung, 2020, p. 8). By utilizing emotionally rich passages from romance novels and generated paragraphs by GPT-4o, this study can still effectively investigate the core textual and affective elements that affect readers' perceptions of human and AI-generated texts.

3.2 sample and sampling method

Participants were recruited through purposive sampling since this method ensure

participants meet certain knowledge level and criteria for this study (Babbie, 2017,p.196). In which the participants must be at least 18 years old, have sufficient proficiency in English, and have a basic understanding of texts. The surveys were distributed through purposive sampling by sending survey links via social media networks, such as friends, family, group chats of classmates from junior high school, high school, and bachelor programs. And some of the participants were gathered through the online platform called Surveyswap, this platform automatically gathered respondents for this survey after the researchers helped complete the survey of fellow students and researchers. The respondents from Surveyswap were also required to fulfill the age and English proficiency requirements.

3.3 Materials and Text Selection

By looking at the common scope of romance novels, it was clear that emotional engagements were achieved through detailed portraits of the characters and the building up of crucial moments for relationship development between the characters (Fletcher et al., 2018, p.1011; Thelwall, 2019, p.428). Also, direct declaration of love from heroes and internal dialogues made by heroines, as well as shifting in point of view, are the unique patterns of romance novels (Carter, 2018, pp. 90,91). Therefore, the romantic paragraphs were selected based on those criteria: they consist of internal dialogues or direct declarations and describe a crucial moment between the main characters, which can evoke the audience's emotions. Aligned with those criteria, the first few paragraphs in Chapter 3 of *The Notebook* by Nicholas Sparks (1999) were chosen for this research.

Firstly, these paragraphs depict a reunion moment of protagonists Allie and Noah after seven years apart. This reunion marked a pivotal moment in their relationship development because it built up the tension by exploring the possibility of reigniting their love after seven years. Making these paragraphs a highlight moment that fully conveys the complexity of human emotions.

Secondly, Sparks used both internal dialogues and direct conversations in the selected paragraphs to portray emotional nuances, vulnerability, and the tension between two characters. The utilizations of short, descriptive sentences align with the characteristics of natural language. Meanwhile, using internal dialogues and conversations is a commonly seen tactic in classic romance novels, making this an ideal sample for this research study.

Lastly, the novel was published in 1999, long before the widespread use of LLMs; therefore, it eliminates any possibility of AI influence.

The following paragraphs, written by Sparks (1999, "Chapter 3-Reunion," para. 3-8), were marked as *Text A*:

“Thoughts of the summer they’d shared came back to her, and as she stared at him, she noticed how little he’d changed since she’d last seen him. He looked good, she thought. With his shirt tucked loosely into old faded jeans, she could see the same broad shoulders she remembered, tapering down to narrow hips and a flat stomach. He was tan, too, as if he’d worked outside all summer, and though his hair was a little thinner and lighter than she remembered, he looked the same as he had when she’d known him last.

When she was finally ready, she took a deep breath and smiled.“Hello, Noah. It’s good to see you again.”Her comment startled him, and he looked at her with amazement in his eyes. Then, after shaking his head slightly, he slowly began to smile.“You too . . . ,” he stammered. He brought his hand to his chin, and she noticed he hadn’t shaved. “It’s really you, isn’t it? I can’t believe it. . . .”

After deciding on Text A, GPT-4o then needed to match Text A closely in content and style to ensure a fair comparison. To do so, the prompt followed these instructions: 1. match the number of words in Text A; 2. match the names of protagonists; 3. match the point of view used in Text A;4. use both internal dialogues and conversations; 5. Include physical and behavioral descriptions.

Then Text B was generated by following the prompt:

Write me two paragraphs for a reunion scene in a romance novel, around 170 words, third-person perspective, with dialogues, internal thoughts, descriptions of behavior, physical features, and attire for Allie and Noah.

And Text B was generated by GPT-4o as follows:

“Noah glanced up from his untouched coffee, and for a second, the air between them stilled. Allie—her auburn hair in soft waves now, her lips slightly parted in hesitation—was as striking as ever. The warm lighting caught the gold in her earrings, making her seem almost unreal. "I wasn’t sure you’d come," he said, his voice even, though his grip on the ceramic mug tightened. She swallowed, stepping closer. "Neither was I," she admitted, tucking a loose strand of hair behind her ear. A habit he still remembered.

She slid into the chair across from him, their knees grazing beneath the table. A rush of warmth spread through Noah at the contact, but he forced himself to stay still. "It’s been a long time," she said, studying him. He nodded, exhaling slowly. "Yeah. Too long." The silence stretched, heavy with things unsaid. And yet, in that quiet space, something familiar still lingered.”

3.4 Survey Design and Operationalization

The survey comprises the following sections: a consent form, demographic information, comprehension check questions, 17 Likert-scale questions for each text (evaluating language

naturalness, cohesion, and emotional context, immediate assessment), one Likert-scale question immediate authorship attribution after each text, one final authorship attribution question, followed by an open-ended question for providing a short explanation of final attribution. To reduce the perception bias, the appearance order of the text is randomized.

The questionnaire began with demographic questions, including age, gender, and level of education, followed by a self-evaluation question on English proficiency. Then the participants were assigned to read the first text, answer comprehension check questions, and evaluate cohesion, naturalness, and emotional content, as well as an immediate assessment of authorship attribution of the text. This process was repeated for the second text. After reading both texts, the participants were asked to select which text was written by a human and provide a detailed explanation with a minimum of 10 words.

Text cohesion was measured using three questions regards the idea consistency, smoothness in sentence transitions, and a reverse-coded question about self-contradictory ideas. Participants responded on a 5-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). The higher scores indicate greater cohesion and humanness. Higher scores for self-contradictory statements are linked to AI writing.

Linguistic naturalness was assessed through five questions in the following aspects: simplicity of words, vocabulary diversity, use of descriptive details, sentence length, and occurrence of sentence repetition. For simplicity of words, vocabulary diversity, higher scores indicate more linguistic naturalness, hence humanness. And higher score in descriptive details and sentence repetition indicates less naturalness.

And emotion and person voices were assed with five questions regarding the positive and negative tendency of the emotions, emotional authenticity, and the degree of personal tone of voice and emotional attitude. Higher scores in positive emotional tendency were strongly associated with AI-generated texts, whereas higher scores in emotional authenticity, personal tone, and emotional attitude were linked to human-written texts. The detailed questionnaire is attached in Appendix A

Table 1: Key metric measured in survey

Dimension	Factor	Sub factor	Questions
Coherence	Flow		The text is easy to read and follow.
	Cohesive ties		Transitions between sentences and paragraphs are smooth.

	Logic continuity		The text presents its ideas in a consistent way
			The text contains self-contradictory information
Linguistic Naturalness	Lexical	Lexical simplicity	The text uses simple and easy-to-understand words
		Lexical diversity	The text uses a diverse range of vocabulary
	Prior findings		The text uses many adjectives and descriptive details
			The sentences in the text are generally long
			The sentence structures are repetitive
Emotional Tone	Attitude	Affect	The text expresses clear negative emotions
			The text expresses clear positive emotions
		Judgement	The author's emotional attitude toward the topic is clear
	Engagement		The text conveys a personal tone of voice.
	Prior findings		The emotions expressed feel authentic

3.5 Data Collection Procedure

The survey was designed, and the data were collected through Qualtrics, an online survey platform. Upon accessing the survey, participants were given the options to either consent and continue or quit the survey.

As both AI-generated and human-written texts were presented for participants to review, the order of appearance may induce potential bias in viewers' perceptions. To minimize this potential bias, the order of appearance was randomized in Qualtrics. Also, there were immediate attribution assessments following each passage and a final assessment after reading both texts. This

ensures participants' constant engagement throughout the process and helps them navigate which text to choose in the final attribution question.

A total of 85 valid responses were collected, 69 of which were collected from internal social networks, such as WeChat and WhatsApp group chats, as well as individual responses from friends, family members, classmates from junior high school, high school, and bachelor programs. And 16 of which were collected from an external platform, Surveyswap. It is a public platform for exchanging surveys among students and researchers who are looking for ideal participants for their study.

3.6 Data Processing and Analysis

For this research, both quantitative data analysis and thematic analysis were employed.

First, frequency tables were conducted for demographic information to gather the patterns and distributions among age, gender, level of education, and English proficiency.

Then, a reliability analysis was conducted for items in each criterion (naturalness, coherence, emotional tone); if the $\alpha \geq .70$, then the items can be computed together as a collective unit. If not, an investigation on "Cronbach's Alpha if item deleted" will be employed. After deleting the necessary item(s), the rest of the items were computed into a collective unit. If deleting none of the items will result in $\alpha \geq .70$, the internal consistency is proven to be insufficient. And these items will be compared individually in pairs.

After conducting the reliability test, a paired sample t-test was employed to compare individual items from Text A and Text B. The analysis focused on differences in mean scores with a significance determined at $p < .05$. To ensure the validity of these parametric tests, the assumptions of normality were then tested using the Shapiro-Wilk test. For any item whose scores significantly deviated from normality ($p < .05$), non-parametric Wilcoxon signed-rank tests were then conducted as an alternative.

After analyzing each criterion (naturalness, coherence, emotional tone), a frequency analysis was employed to gather the distribution for final authorship. A crosstab analysis of immediate attribution to A and B with final attribution to A was employed, so that it will be helpful to interpret how many participants changed their ideas and if there is any significance in the Chi-Square measure.

Lastly, qualitative responses were analyzed through thematic analysis. The response was initially categorized by three key criteria: naturalness, coherence, and emotional tone. Any recurring pattern will be labeled with other names. The detailed coding book for thematic analysis is attached in Appendix B.

During the data analysis phase, certain items were intentionally excluded from the final

analysis to maintain conceptual clarity and theoretical alignment. First, the question assessing negative emotion was omitted because while it was initially included as a balance to the question regards positive emotions, it lacked support from the theoretical framework and introduced redundancy if it was treated as a reverse-coded item.

Also, all items related to participant confidence (e.g., confidence in their judgments or comprehension) were excluded. These items were designed to sustain engagement and ensure attentive participation but were not directly relevant to the study's core research objectives. Including them in the statistical analysis would not provide insights for this research topic, while omitting these items helps ensure the analytical focus remains theoretically grounded.

3.6 Ethical considerations

The following ethical practices will be employed throughout the research phase: voluntary participation, informed consent, and confidentiality. Specifically, participants are free to join or withdraw from the study at any time, their identities will not be recorded, and they will be fully informed about the study's purpose, benefits, and potential risks.

4. Results

Among the 85 participants, 61 (71.8%) identified as female, 19 (22.4%) as male, and 5 (5.9%) chose not to disclose their gender. The average age was 27.81 years (SD=6.2). A majority of participants had completed a bachelor's degree (50.6%), and most self-assessed their English proficiency as fluent (60.0%).

Given the number of comparisons (n = 13), the risk of Type I error increases. Therefore, it is essential to interpret the results in conjunction with effect sizes. As a result, an interpretive threshold of $d \geq 0.4$ is used as the indicator of practical significance, following Cohen's (1988) convention.

4.1 Comprehension Checks

For Text A, 51 participants (60.0%) successfully passed the comprehension check by selecting the most accurate summary (A). An additional 10 participants (11.8%) selected summary C, which reflected a more surface-level understanding of the content without acknowledging the emotional nuances beyond the literal text. Similarly, 62 participants (72.9%) passed the comprehension check for Text B by choosing summary B. An additional 9 participants (10.6%) selected summary C, also suggesting a partial understanding with a lack of deeper emotional comprehension.

Demographic analysis of those who passed the comprehension check revealed consistent patterns across both texts. Among those who passed the comprehension check for Text A, the average age was 28.37 years. Most of them had either a bachelor's degree (n = 27) or a master's degree (n = 18). In terms of English proficiency, the majority reported being fluent (n = 33) and advanced (n = 8). For those who passed the comprehension check for Text B, the average age was 27.55 years. Similarly, most of them also held a bachelor's degree (n = 31) or a master's degree (n = 24). Their language proficiency levels were also quite high, with 33 of them self-assessing as fluent and 10 reporting advanced English skills.

The similarity in comprehension rates among different demographics suggests that both texts were equally accessible to participants from diverse educational and linguistic backgrounds. This supports the assumption that evaluations of the texts were not affected by differences in participants' age and their ability to access the content.

4.2 Linguistic naturalness

4.2.1 Quantitative findings

As stated in H1, human-written text will (a) use simpler, (b) more diverse vocabulary, (c) fewer adjectives and descriptive details, (d) longer, and (e) less repetitive sentence structures than

AI-generated text. To assess linguistic naturalness, these five dimensions were measured.

After conducting the reliability analysis, the results indicated insufficient internal consistency for both texts (Text A: α = .47, Text B: α = .20), suggesting these items do not reliably measure a collective unit. Therefore, each item was analyzed individually using paired-sample t-tests (α = .05).

In terms of vocabulary simplicity, Text A (M = 4.24, SD = 0.92) was rated significantly higher than Text B (M = 3.59, SD = 1.12), t (84) = 3.90, p < .001, d = 0.46, 95% CI [0.24, 0.68], indicating a moderate effect, showing that human-written text used easier words. However, the Shapiro-Wilk test revealed a violation of normality (W = 0.93, p < .001). Moreover, a Wilcoxon signed-rank test also showed a non-significant difference, Z = -3.96, p < .001. Therefore, despite the significant differences found with the paired-sample t-test, the difference between the vocabulary simplicity of Text A and Text B was not considered to be significant.

Conversely, vocabulary diversity of Text A (M = 3.24, SD = 1.00) was rated lower than Text B (M = 3.76, SD = 0.89), t (84) = - 3.71, p < .001, d = 1.32, 95% CI [1.03, 1.61], showing a robust effect in this item. However, the assumption of normality was violated according to the Shapiro-Wilk test (W = 0.95, p < .001). A Wilcoxon signed-rank test revealed a significant difference with Z = -0.35, p < .001. Confirming the effectiveness of this finding. This indicates that AI-generated text used significantly more diverse words than human-produced work, with a large and reliable effect size.

Similarly, the sentence lengths of Text A (M=2.85, SD=1.13) were lower than Text B (M=3.25, SD=1.03), t (84) = - 2.34, p = .011, d = 1.58, 95% CI [1.26, 1.90], suggesting a large effect. However, the assumption of normality was also violated as the Shapiro-Wilk test revealed to be significant (W = 0.94, p < .001). But the Wilcoxon signed-rank test revealed a significant difference in text (Z = -0.31, p = .002). Therefore, AI tends to use significantly longer sentences than human writing.

In terms of the appearance of adjectives, while the pair-sample t-test suggested that Text B (M = 4.09, SD = 1.03) has significantly higher utilization of adjectives and descriptive details than Text A (M = 3.59, SD = 1.03), t (84) = -3.23, p < .001, d = - 0.35. 95% CI [−0.57, −0.13], with a small effect. This normality was not confirmed by the Shapiro-Wilk test, W = 0.95, p = .003. But the Wilcoxon signed-rank test provides a significant difference (Z = -2.135, p = .033). It is concluded that AI tends to use extensively more adjectives than human writing.

Additionally, the p-value of repetitive sentence structures is 0.437, which is greater than 0.001; therefore, the differences in sentence repetitiveness between the two texts were non-significant.

4.2.2 Qualitative findings

The results from thematic analysis revealed that use of adjectives (n = 12), simplicity of words (n = 10), and sentence structure (n = 6) were the most frequently mentioned indicators of linguistic naturalness. Some considered AI writing to be less descriptive and unable to utilize metaphors, while others believed AI-written text should engage more descriptive details. This illustrates that participants have opposite opinions on whether human-written or AI-produced text should utilize more descriptive details. Which resulted in seven participants misattributed adjectives and descriptive language to the incorrect source (Text A) in the final attribution. Thus, although statistically AI was proven to use more adjectives and detailed descriptions, this characteristic was mistaken as a human writing trait by several respondents.

Among participants who mentioned simple and natural vocabulary, most of them (n=6) attributed that to human-produced text, whereas the rest attributed words that are "close to spoken English" and "used in daily life" as the characteristics of AI writing. Moreover, almost all the participants (n=5) who mentioned simple sentence structure attributed this trait to human-written text.

In conclusion, statistical findings supported (c) but rejected (a) due to a violation of the normality assumption, (b) and (d) were also rejected because the results show opposite results from what H1 proposed, and (e) was rejected due to non-significance in the difference. Therefore, human-written text used fewer adjectives, more restricted vocabulary, and shorter sentences than AI-generated texts. The thematic analysis showed that the simplicity of both vocabulary and sentence structures were the key determinants of human-written text; the use of adjectives is a rather confusing factor since participants have mixed opinions on the attribution of this characteristic. These results aligned with the findings of Markowitz et al. (2024, p.72) but opposed those of Yildiz Durak (2025, p.7), Muñoz-Ortiz et al. (2024, p.10), and Reviriego et al. (2024, p.4).

4.3 Coherence

4.3.1 Quantitative findings

As stated in H2, human-written text will (a) be perceived as easier to follow, (b) exhibit greater coherence through consistent ideas and (c) smoother transitions between sentences and paragraphs, and (d) present less contradictory information compared to AI-generated text.

The initial reliability test was conducted with all four items, which revealed $\alpha < .70$. Following the investigation on "Cronbach's Alpha if item deleted", the reverse-coded question was taken out of the analysis, after re-conducting the reliability test with three items: (a)ease of reading,

(b)idea consistency, and (c) smooth transitions. The result confirmed an acceptable internal consistency with $\alpha = .70$ and $\alpha = .75$ for Text A and B, respectively.

However, a deeper investigation of standard deviations among these three items suggests that (a) ease of reading likely drove the overall variability (Text A: SD=0.92, Text B: SD=1.01), while (b) idea consistency shows limited variation (Text A: SD=0.82, Text B: SD=0.80). Therefore, (a), (b), and (c) are not ideal to be computed into a collective unit. As a result, all four items will be measured individually using paired-sample t-tests ($\alpha = .05$).

After conducting the pair-sample t-tests on individual items, the results show that none of the items have statistically significant differences. The p-values yield at p=.080, p=.364, p=.180, and p=.350 for (a) ease of reading, (b) idea consistency, (c) smooth transitions, and (d) contradictory information, respectively.

And since pair-sample t-tests have relatively low power for detecting small but consistent effects across multiple metrics. A one-way repeated measures MANOVA was conducted to further investigate the pattern of coherence across all four metrics. However, similar results were found in the MANOVA test, where all the metrics lack statistical significance.

Overall, there was no statistically significant multivariate effect of text condition on the combined coherence metrics, Pillai's Trace = .035, $F (4, 80) = 0.723$, $p = .579$, partial $\eta^2 = .035$. Follow-up univariate ANOVAs also revealed no significant differences for any individual coherence metric because p-values for (a) ease of reading (p= .134) (b) idea consistency (p= .728), (c) smooth transitions (p= .320), and (d) contradictory information (p=.700) are all greater than .05. Moreover, descriptive statistics further show that the means for each condition were similar across all metrics. Therefore, non-significant results from both tests reinforce the statement that all the items within the coherence measure do not differ much from Text A to Text B.

This result contradicts findings from Yang et al. (2024, p.12) and those of Mündler et al. (2023, p. 7). However, it supports their prediction on the potential of improved coherence with advanced LLMs (Mündler et al., 2023, p. 8).

4.3.2 Qualitative findings

Among all the comments regarding coherence measures, ease of reading is a prominent indicator, which was mentioned by 6 participants. However, half of the participants think Text A is easier to read and follow, whereas the other half believes Text B is more fluent. This aligns with the result from the paired-sample t-test, further proving that human-produced romantic narration is not necessarily considered any easier to read or follow.

In conclusion, coherence was not a distinguishing factor between AI and human-written romantic narratives. There is no significant difference between ease of reading, idea consistency, transitions' smoothness, and the occurrence of self-contradictory information between human-written and AI-produced romantic narrations. The qualitative feedback also reveals that judgments of coherence may also be shaped by personal perception rather than strictly logical structure.

4.4 Emotional Tone

4.4.1 Quantitative findings

As stated in H3, human-written text will (a) express a stronger emotional attitude, (b) more authentic emotions, (c) adopt a more distinct personal tone of voice, and (d) exhibit less distinctly positive emotions than AI-generated text.

Therefore, emotional tone was assessed through four items: emotional attitude, emotional authenticity, positive emotionality, and personal tone of voice.

The result of reliability analysis revealed there was sufficient internal consistency among the four items (Text A: $\alpha = .71$; Text B: $\alpha = .80$). Thus, the emotional tone will be tested as a collective unit between the two texts with paired-sample t-tests ($\alpha = 0.05$).

In terms of emotional tone, the result shows that the difference between human-written and AI-generated is not significant ($p = .325$). This indicates that human-written romantic text does not show a more distinct emotional attitude, personal tone of voice, more emotional authenticity, or fewer occurrences of positive emotions than AI-generated romantic texts. The result not only contradicts with several prior research (Muñoz-Ortiz et al., 2024, p. 264; Hohenstein & Jung, 2020, p. 8; Markowitz et al., 2024, p. 65; Matthews & Volpe, 2023, p. 86), but also suggested emotional expressiveness and tone of voice may not be an ideal detector for AI-produced texts. Further suggested that GPT-4o can mimic personal tone in romance narration, making the two texts difficult to distinguish.

4.4.2 Qualitative findings

Among those who commented on emotional tone, emotional intensity (n=7), and emotional authenticity (n=5) were the key indicators for the final attribution of authorship. However, more than half (n=4) of those who addressed on the emotional intensity indicated that AI-generated text shows stronger emotion, with comments include "The emotion from text B could be more obvious to tell" and "Text A seemed very out of touch, like there was no emotion in the text", suggesting AI can generate romantic narration with stronger emotional cues. As for authenticity, only one of the participants perceived the emotion in Text B as "more raw and real", the rest of the participants (n=4) linked authentic and realistic emotion with Text A.

Overall, while quantitative analysis failed to detect differences in emotional tone between the two texts, qualitative responses indicated that participants considered AI writing to have strong emotional intensity and human writing to be more emotionally authentic. This qualitative finding supports that of Hohenstein and Jung (2020, p. 8), indicating emotional tone has the potential to be a key differentiator between the two.

4.5. Summary tables on findings

Table 2: Summary of quantitative findings of linguistic naturalness

Dimension	**M(A)**	**SD**	**M(B)**	**SD**	**t**	**df**	**p (t-test)**	**d**	**95% CI Low**	**95% CI High**	**Shapiro-Wilk W**	**Shapiro-Wilk p**	**Wilcoxon Z**	**p (Wilcoxon)**	**Conclusion**
Vocabulary Simplicity	4.24	0.92	3.59	1.12	3.9	84	< .001	0.5	0.24	0.68	0.93	< .001	-3.96	< .001	Non-significant
Vocabulary Simplicity	3.24	1	3.76	0.89	3.7	84	< .001	1.3	1.03	1.61	0.95	< .001	-0.35	< .001	AI > Human
Adjectives/Descriptive Details	3.59	1.03	4.09	1.03	3.2	84	< .001	-0.4	-0.57	-0.13	0.95	0.003	-2.14	0.033	AI > Human
Sentence Length	2.85	1.13	3.25	1.03	2.3	84	0.011	1.6	1.26	1.9	0.94	< .001	-0.31	0.002	AI > Human
Repetitive Structures	-	-	-	-											

Table 3: Summary of quantitative findings of coherence

Dimension	**M(A)**	**SD**	**M(B)**	**SD**	**p (t-test)**	**MANOVA F**	**MANOVA p**	**Partial η^2**	**Conclusion**

Ease of Reading	-	0.92	-	1.01	0.08	-	0.13	0.035	Non-significant
Idea Consistency	-	0.82	-	0.8	0.36	-	0.73	0.035	Non-significant
Smooth Transitions	-	-	-	-	0.18	-	0.32	0.035	Non-significant
Contradictory	-	-	-	-	0.35	-	0.7	0.035	Non-significant

Table 3: Summary of quantitative findings of emotional tone

Dimension	Text A α	Text B α	p (t-test)	Conclusion
Emotional Tone	0.71	0.8	0.325	Non-significant

Table 4: Summary of qualitative findings

Dimensions	Themes	Frequency	Human Attribution	AI Attribution	Contradictions
Linguistic Naturalness	Adjectives/Descriptive Details	12	7	5	Split on attribution
	Simplicity of Vocabulary	10	6	4	-
	Simple Sentence Structures	6	5	1	-
Coherence	Ease of Reading	6	3	3	50/50 split on perceived fluency
Emotional Tone	Emotional Intensity	7	3	4	AI perceived as more intense
	Emotional Authenticity	5	4	1	Humans perceived as more authentic

4.6 Final attribution

As stated in the hypothesis section, since human-written text was assumed to have more linguistic naturalness, better cohesion and consistency, as well as more distinctive emotional tone and personal voice, H4 was stated as participants are more likely to assign text A as human-written text after reading both texts.

4.6.1 Frequency of Attribution

After examining the frequency, 44 (51.8%) of the participants assigned Text A as human-written text, 34 (40.0%) assigned Text B, and 7 (8.2%) of them were not sure of which text was human-written. Therefore, Text A was more frequently assigned as human-written. However, there are only 10 more people (11.8%) who chose Text A over Text B for human-written text, suggesting there is only a slight difference in the attribution of authorship.

4.6.2 Attribution Consistency

By cross-examining the immediate assessment of attribution after reading each text and the final assessment. Among participants who initially attributed Text A to a human (n = 41), 56.1% maintained this attribution in their final assessment, suggesting moderate consistency in authorship judgment.

Cross-tabulation results indicated a stronger relationship between immediate attribution of Text B and the final attribution decision ($\chi^2 = 27.47$, $p < .001$), compared to Text A ($\chi^2 = 1.60$, $p = .809$). This suggests that participants relied more on their perception of Text B for their final attribution. Since the final assessment asked the participants to determine the authorship of human-written text. This can be an elimination strategy, where participants can decide which text is more artificial, and proceed to choose the other option.

4.7 Punctuation: Em-dash

Besides the metrics of linguistic naturalness, coherence, and emotional tone, it is revealed that one of the most prominent indicators of AI-generated text is the usage of dashes (—) in between sentences.

12 of the participants mentioned that the appearance of the punctuation of em-dash is a prominent clue for their decisions in authorship attribution. Besides, one person assumed that “hyphen ‘- ‘couldn’t be written by AI”, the other participants (n=11) strongly associated them with AI-generated writing. In detail, some suggested that em-dash is a rarely used punctuation in

everyday life, some directly addressed it as “very AI coded”, others suggested that they “haven’t seen that in a book (yet)”.

This observation suggested that although GPT-4o can imitate human-written romantic narration in natural language, coherent and consistent, and even creates equally emotional context, the use of certain punctuation still gives away the clue. And this can be a new area of exploration in human and AI authorship detection, focusing on nuances that may escape traditional textual and emotional analysis.

5. Discussion

5.1 Overview

This research aims to examine people's perceptions and evaluations of AI-generated and human-written romantic narratives. Based on the prior findings of key indicators between AI and human-written content (e.g., Markowitz et al., 2024, p. 65; Mündler et al., 2023, p.7; Reviriego et al., 2024, p.7), this research is built upon three main dimensions: linguistic naturalness, coherence, and emotional tone.

From the result of text attribution, only 10 more people (11.8%) successfully attributed Text A as human-produced romantic text. This slight difference suggests that it is quite difficult for people to distinguish the authorship. In terms of linguistic naturalness, the use of adjectives and descriptive details is the only prominent indicator of AI-writing. Moreover, GPT-4o used significantly longer sentences and more diverse vocabulary, as opposed to what H1 stated. And there was no significant statistical difference in sentence repetition and vocabulary simplicity. In contrast, the thematic analysis suggested that the simplicity of vocabulary and sentence structures were the crucial characteristics of human-written text. In terms of coherence, there is no significant difference found between the two texts in both quantitative and qualitative measures. And thematic analysis suggests that AI narration tends to have stronger emotion while human human-written novel shows more authentic emotion. There is no significant difference between the tone of voices found within the texts.

Besides those elements mentioned, the use of dashes has surprisingly appeared as a key differentiator between AI and human-produced romantic context. Suggesting that the frequent use of dashes is not commonly seen in romance novels.

5.2 Interpretations

5.2.1. Linguistic naturalness

Initial quantitative analysis using pair-sample t-tests suggested that human-written text used significantly simpler vocabulary. However, statistical validation, as indicated by Shapiro-Wilk test ($p < .001$) and Wilcoxon signed-rank test ($p < .05$), suggested a violation of normality assumptions for vocabulary simplicity, making this finding non-significant. Moreover, AI-generated text employed more adjectives and descriptive details, showed greater lexical diversity, and used longer sentences. No significant difference was found in sentence repetition.

To complement these quantitative insights and gain a deeper understanding of linguistic features, a thematic analysis was conducted. The qualitative analysis identified simplicity of vocabulary and sentence structures to be the key characteristics of human writing. However, this finding likely has limitations in generalizability as the Shapiro-Wilk test suggested a lack of normality

assumption in sample size. Moreover, the use of richer adjectives is also frequently mentioned, but often linked to human writing. Although descriptiveness emerged as a prominent indicator, it is rather a misleading one, as over half of the participants believed AI writing was less descriptive.

Therefore, these findings collectively suggested that the utilization of descriptive details was the only reliable indicator of text difference, as both quantitative and qualitative data supported this hypothesis. Aligning with the findings of Markowitz et al. (2024, p.66), even though the attribution of this characteristic is rather confusing and unclear. Surprisingly, AI tend to use significantly more diverse vocabulary and longer sentence lengths than human authors, as opposed to the findings proposed by Yildiz Durak (2025, p.7), Muñoz-Ortiz et al. (2024, p.10), Reviriego et al. (2024, p.4). This finding suggested that genre is a crucial aspect in learning the human perception of texts, as the findings from academic and argumentative articles did not apply to the romantic narrative study. Thus, it was essential to develop and utilize a genre-specific framework when analyzing different texts.

Also, this result highlights the importance of verifying statistical assumptions, particularly when working with relatively small to moderate sample sizes (n=85), with most of the participants being female (71.8%) and an average age of 27.81 years, which limits the stability and generalizability of quantitative results.

In conclusion, the mixed-method approach was proven to be essential as quantitative analysis revealed insights into the existing metrics and showcased the limitations in generalizability. In combination, the open-ended questions not only cross-examined quantitative findings but also introduced new dimensions (simplicity of sentence structures) to the study of linguistic naturalness.

5.2.2. Coherence

Quantitative analysis revealed no significant differences in ease of reading, idea consistency, smoothness in transition, or the occurrence of self-contradictory information between human and AI texts. From the thematic analysis, participants split evenly on perceived coherence: 50% described Text A (human) as "easier to follow," while the other 50% believed the opposite. This reinforces that coherence judgments were highly subjective in romantic texts, thus not making it an ideal metric to be quantifiable.

Moreover, the non-significant difference between two texts suggests that although the self-contradictory and inconsistency are more frequently seen in abstracts, Wikipedia entities and argumentative essays, where logic flow is a crucial factor to be examined (Yang et al, 2024, p. 13; Ma et al., 2023, p. 3; Mündler et al., 2023, p.7). Whereas romance novels focus more on emotional narrations (Carter, 2018, p. 90) instead of composing logical arguments. Therefore, a genre-specific

framework is demanded as the measurements for academic material do not apply to emotionally rich genres, such as romance.

In addition, the non-significance of results can stem from the advancement of GPT-4o in matching human-like coherence in romantic narratives. Prior studies used less advanced LLM models, such as GPT-3, GPT-3.5 (Ma et al., 2023, p. 3; Markowitz et al., 2024, p.72), while this research utilized the most advanced model, GPT-4o, in text generation. The improvement of LLM, in this case ChatGPT, can also be a crucial factor for making two texts indistinguishable.

5.2.3. Emotional tone

From quantitative analysis, emotional tone, including attitude, authenticity, personal voice, and positivity, showed no significant difference between texts. Qualitatively, however, a drastic difference in intensity and authenticity of emotion was found. AI text was perceived to have stronger emotional intensity. For instance, 4 out of 7 participants cited Text B as “more obvious emotion" and “stronger emotions”. Meanwhile, human text was associated with greater authenticity. 4 out of 5 participants noted that the emotion is more "raw and real" and “more authentic”. This further proved the importance of conducting a mixed method as statistical measures failed to distinguish the texts, while qualitative results revealed that participants associated intensity with AI and authenticity with humans. Overall, the emotional dimension is the most reliable differentiator among linguistic naturalness, coherence, and emotional tone.

Although AI can generate more descriptive details, which can contribute to the emotional intensity perceived by people, participants still rely on their intuition in perceiving the authenticity. One of the participants who attributed Text A as human-written stated that it was based on the “gut feeling”.

Therefore, this further suggests that the quantitative method may not be an ideal method for studying the emotion and tone of voice, since those qualities do not show significant differences in quantitative measures. Also, this proves that even LLMs are developing at an extremely fast pace, but it is still challenging for them to compose authentic emotions.

5.2.4. Unexpected indicator: Em Dash

In the thematic analysis, the frequent use of em dashes emerged as an unexpected trait of AI. 11 of the participants explicitly cited them as "unnatural" and "AI-coded," highlighting a reliable detection cue beyond the metrics of linguistic naturalness, coherence, and emotional tone.

There have not been many academic papers that studied the reason behind the frequent utilization of em dashes in content generation, especially in the case of ChatGPT. However,

discussions of this phenomenon are emerging on the Internet. For instance, there are plenty of Reddit discussions, such as “Why do ChatGPT love the em dash so much?”, and ways to stop GPT from generating em dash (Longjumping-Speed511, 2025; reliablepayperhead, 2025). Because ChatGPT uses em dash so frequently, people are intentionally avoid using this punctuation in their writing, so their writings would not be mistaken as output from LLMs (Csutoras, 2025).

According to Csutoras (2025), the frequent occurrence of em dashes was embedded in its training data, where em dashes appeared in all sorts of text materials, and it was not flagged as something to be avoided during the model training process. Therefore, for ChatGPT, using em dashes as sentence connections aligns with its rule of creating realistic human writing.

This unexpected indicator further proves that romantic content, unlike academic essays, hotel reviews, and interactive messages, is not suitable to be studied with a purely quantitative method. Since it conveys nuances in emotion, tone of voice, and other subjective factors that cannot be measured in quantitative units.

Overall, this research demonstrates that participants have difficulty in distinguishing AI-generated (GPT-4o) from human-written romantic narratives, with only 11.8% more people choosing the correct authorship attribution. Linguistic analysis revealed that human texts used simpler vocabulary and sentence structures, while AI used more descriptive details, though participants often mismatched richer adjectives to humans. Coherence showed no significant differences in ease of reading, logical flow, or contradictions across both methods, placing less emphasis on logical flow among romance contexts. Emotional tone was statistically indistinguishable, yet participants perceived AI texts as more emotionally intense and human texts as more authentic, highlighting the necessity of incorporating qualitative measures. Surprisingly, em-dashes (—) emerged as a strong AI indicator, with 11 participants identifying them as "unnatural" and "AI-coded." The findings challenge prior findings on the AI detection found in academic and short messages, and potentially encourage the development of a new framework to study the AI-generated content in the romance genre.

5.3 Theoretical Implications

This study encouraged a re-evaluation of current AI-text detection frameworks by demonstrating that genre plays a crucial role in determining the optimal metrics for authorship attribution. Although prior research emphasized the coherence markers, like logical flow and self-contradictory information, as reliable indicators in argumentative texts, these are not directly applicable in romantic narratives. Instead, emotion, tone of voice, and other nuances are more suitable metrics for affective storytelling.

It is essential to incorporate a mixed-methods approach when studying an exploratory topic.

Quantitative analysis tested the items that were based on a theoretical framework and existing findings. Also, tests such as the Shapiro-Wilk test, Wilcoxon signed-rank test, and pair-sample t-tests not only showcased the degree of differences but also warned of the limitations in generalizability. And qualitative approach cross-examined the findings from quantitative measures and enriched the study by introducing new metrics, such as sentence simplicity and utilization of Em dashes, providing insights for genre-specific frameworks. Therefore, the mixed method provided a comprehensive understanding of the human perception of emotionally rich texts.

Moreover, the misattribution of linguistic features, specifically, attributing rich descriptiveness to humans, exposes biases in authorship perception. Although this item can be quantifiable, it is not an objective measure due to the difference in participants' personal interpretations.

Lastly, em-dashes emerged to be a distinct detector for AI writing, revealing certain loopholes in AI model training. This phenomenon demands more academic attention and theoretical investigations. More studies are needed to address the reasons, mechanisms, and results of overusing em dashes in text generation. For instance, some people are actively avoiding using em dashes due to the influence of AI programming (Csutoras, 2025), while others are advocating for people to not change their writing habits just because AI uses em dashes redundantly (Gillett, 2025). It would be interesting to see how this phenomenon will impact people's creativity, perceptions of AI, and other domains.

5.4 Practical Implications

On a practical level, this research provides guidance and valuable insights for writers and editors, people who work in educational sectors, and AI developers.

For writers and editors, this research provides valuable insight to them that AI would not be able to replace human creativity because it is difficult for them to generate texts with emotional authenticity. During these years, there has been an emerging discussion on the probability of LLMs replacing human authors and writers. Some believe this is not likely to happen because AI cannot convey authentic emotions, while others believe the rapid development of LLMs has the potential to exceed the creativity of human authors (All Business, 2025; Trigg, 2024). This research supports the former statement, especially in the romance genre; however, LLMs can potentially create efficient short content that requires less effort (Rochi, 2025). Besides, the excessive use of em dashes in AI-generated texts was explicitly flagged by participants as "AI-coded", revealing that there is a limitation in the capabilities of LLMs in replacing human writing. Therefore, writers should prioritize creating authentic and engaging storytelling instead of worrying about being replaced by LLMs.

For educators, these results emphasize the need to train students in deeper reading skills.

Especially during this digital age, where social media is so commonly used and the number of books read by youngsters has drastically declined (Twenge, 2024). By training the students to read and evaluate texts critically, recognize narrative tone, and establish deep understandings of the context, it helps them in differentiating AI-generated content, even with the rapid development of this technology.

Lastly, for AI and LLM developers, these insights can help them in model refinement. If they want to achieve a more natural or human-like writing, they should suppress some robotic features, such as the excessive use of em dashes in creative writing. Also, they need to adjust the number of descriptive expressions since some participants indicate that AI-generated content “contains too many details” and they are “descriptions of facts, not emotions”. Because too many descriptive details, especially about facts, can also reduce the perceived authenticity and naturalness.

5.5 Limitations

Despite these insights, the study also has certain limitations. First, there were constraints in survey design. Although the randomized text presentation (Text A/B) was initially introduced to reduce perception bias, this design still appeared to confuse participants during the final attribution questions. By analyzing the responses from open-ended questions, contradictory patterns were discovered. For instance, one participant described Text A as "more natural" while attributing humanity to Text B. Another participant pointed out “long dash ‘- ‘is very AI coded,” while also attributing humanity to Text B. This pattern may be caused by task misinterpretation or the assumption that the first text is A. This mismatch in final attribution and responses from open-ended questions suggests there are likely some inaccurate attributions of authorship, which affects the result of perception accuracy.

Second, this research suggests a rather narrow demographic profile and a skewed sample that limit the generalizability. From the normality violation in Section *4.2.1*, it is revealed that the sample was skewed in gender and age (71.8% female, average age 27.81). Moreover, the educational background and English proficiency also show a similar pattern, as 50.6% of the participants held bachelor’s degrees and 60.0% were fluent in English, suggesting the results of the research can potentially focus on the perception of adults with a Western educational background.

Third, this research is a mixed-methods study, in which qualitative analysis, more specifically thematic analysis, inherently involves researcher interpretation. While open-ended questions successfully uncovered unanticipated insights, such as the overuse of certain punctuation, variation in emotional measures, and the perceived simplicity in sentence structures, thematic analysis still potentially leads to interpretation bias. For example, "raw emotion" was labeled as authenticity rather than intensity based on the researcher’s interpretation, suggesting the

involvement of subjective judgment during the coding process.

Lastly, although the two-way mixed ANOVA was initially considered to be conducted between the order of the text appearance and item means to test if the order influences participants' perceptions in all the metrics. However, it is later found that the order of text appearance data was not recorded due to a missing step Qualtrics setting. Therefore, the effect of text order remained unstudied for this research.

5.6 Future Research

Despite the limitations, this study sheds light on several potential pathways for future research. First, genre should be a prominent element in AI-text detection research instead of a background variable. The theoretical framework of this research was mostly built on the results found from argumentative and scientific articles, which focus on metrics such as logical flow and consistency. However, romantic narratives were proven to demand distinct detection criteria apart from those elements, like emotional depth and subjective narration are better measurements. For genres such as detective, science fiction, and horror stories, the metrics and results might vary depending on the genre.

The disconnect of qualitative and quantitative results, especially in linguistic and emotional tone evaluations, where statistical analysis shows no significant difference, but thematic analysis shows otherwise, suggests that qualitative measure is a better approach to evaluate emotion-centric content. Based on the result of thematic analysis, the construct of emotional authenticity requires deeper theoretical grounding. Since participants associated human texts with "raw" or "gut feeling", the operational frameworks are required for such measures. Future studies could even employ non-textual measures, such as eye-tracking and facial expression analysis tools, to better understand and correlate the subjective opinions, such as perceived authenticity and intensity, to measurable units. This also suggested the need to further explore the measurements of elements like emotional authenticity and intensity.

Lastly, future work should explore more on the effects and implications of micro-markers, such as em dashes, punctuation flow, and sentence rhythm, in people's perception of AI-produced and human-written texts. These subtle cues may turn out to be more effective than lexical or grammatical indicators, especially when LLMs continue to improve at a fast pace.

6. Conclusion

In reflecting on this research, the crucial difference of AI from human writing is not simply relying on the surface-level measurements, such as the number of various vocabularies, frequency of using adjectives, or scoring sentence length. It is about how language makes us feel and how much we can engage and relate to the emotion the text conveys. While the gap between AI and human writing may be narrowing, AI cannot replace the authenticity and emotional engagement of human authors.

As AI-generated content becomes more prevalent, it is more important to understand the advantages and drawbacks of each method than to distinguish which one is better. Since human writing is more emotionally engaging, while AI is better at composing descriptive details, the co-creation between the two is highly encouraged. This approach enables us to effectively use LLMs, enhancing both the quality and productivity of creative content.

References

AI-generated ebooks: Changing publishing on Amazon. (2024, February 21). elearncollege.com. Retrieved June 18, 2025, from https://elearncollege.com/business-and-management/ai-generated-ebooks-changing-publishing-on-amazon/

All Business. (2025, March 3). *Will AI replace writers? Here's why it's not happening anytime soon.* Forbes. Retrieved June 13, 2025, from https://www.forbes.com/sites/allbusiness/2025/03/03/will-ai-replace-writers-heres-why-its-not-happening-anytime-soon/

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the Tribes of Fluency to Form a Metacognitive Nation. *Personality and Social Psychology Review, 13*(3), 219-235. https://doi.org/10.1177/1088868309341564

Amazon. (n.d.-a). *Amazon best sellers – Kindle eBooks.* https://www.amazon.com/gp/bestsellers/2025/digital-text/154606011/ref=zg_bsar_cal_ye

Amazon. (n.d.-b). *Kindle Direct Publishing: Help. Amazon KDP*. https://kdp.amazon.com/en_US/help/topic/G200672390

Babbie, E. R. (2017). *The basics of social research* (Seventh edition). Cengage Learning.

Bensinger, G. (2023, February 21). *Focus: ChatGPT launches boom in AI-written e-books on Amazon*. Reuters. Retrieved June 18, 2025, from https://www.reuters.com/technology/chatgpt-launches-boom-ai-written-e-books-amazon-2023-02-21/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2018). *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT.* https://doi.org/10.48550/arXiv.2303.04226

Carter, E. (2018). Expressing Emotion in Romance Fiction: How Readers Like Their Heroes and Heroines to Communicate Passion. *Recherches anglaises et nord-américaines*. pp. 87-96. https://doi.org/10.3406/ranam.2018.1566

Chakrabarty, T., Padmakumar, V., Brahman, F., & Muresan, S. (2024). *Creativity Support in the Age of Large Language Models: An Empirical Study Involving Professional Writers*. https://doi.org/10.1145/3635636.3656201

Csutoras, B. (2025, April 29). *The em dash dilemma: How a punctuation mark became AI's stubborn signature.* Medium. Retrieved June 10, 2025, from https://medium.com/@brentcsutoras/the-em-dash-dilemma-how-a-punctuation-mark-became-ais-stubborn-signature-684fbcc9f559

Edwards, B. (2023, September 21). AI-generated books force Amazon to cap e-book publications to 3 per day. Arstechnica. Retrieved May 29, 2025, from https://arstechnica.com/information-technology/2023/09/ai-generated-books-force-amazon-to-cap-ebook-publications-to-3-per-day/

Fletcher, L.M., Driscoll, B., & Wilkins, K. (2018). Genre Worlds and Popular Fiction: The Case of Twenty-First-Century Australian Romance. The Journal of Popular Culture. https://doi.org/10.1111/jpcu.12706

Flower, L., & Hayes, J. R. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication*, *32*(4), 365–387. https://doi.org/10.2307/356600

Gillett, C. (2025, April). Do not remove your em dashes because you're afraid people will think you're ChatGPT. (Please.). If I learned [Post]. LinkedIn. Retrieved June 13, 2025, from https://www.linkedin.com/posts/acgillett_do-not-remove-your-em-dashes-because-you-activity-7308861380047081473-TfXJ/

Halliday, M.A.K., & Hasan, R. (1976). Cohesion in English (1st ed.). Routledge.

Hohenstein, J., & Jung, M. (2020). AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior, 106*. https://doi.org/10.1016/j.chb.2019.106190

Hosseini, M., Resnik, D. B., & Holmes, K. (2023). The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. *Research Ethics Review*. https://doi.org/10.1177/17470161231180449

Kirk, C. P., & Givi, J. (2025). The AI-authorship effect: Understanding authenticity, moral disgust, and consumer responses to AI-generated marketing communications. *Journal of Business Research, 186*. https://doi.org/10.1016/j.jbusres.2024.114984

Kock, N. (2005). Media richness or media naturalness? The evolution of our biological communication apparatus and its influence on our behavior toward E-communication tools. *IEEE Transactions on Professional Communication, 48*(2). https://doi.org/10.1109/TPC.2005.849649

Longjumping-Speed511. [ChatGPT]. (2025, May 19). *Why does ChatGPT love the em dash so much?* [Online forum post]. Reddit. Retrieved June 8, 2025, from https://www.reddit.com/r/ChatGPT/comments/1kqi0bf/why_does_chatgpt_love_the_em_dash_so_much/

Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X. (2023). *AI vs. Human -- Differentiation Analysis of Scientific Content Generation.* https://doi.org/10.48550/arXiv.2301.10416

Markowitz, D. M., Hancock, J. T., & Bailenson, J. N. (2024). Linguistic Markers of Inherently False AI Communication and Intentionally False Human Communication: Evidence From Hotel Reviews. *Journal of Language and Social Psychology, 43*(1), 63–82. https://doi.org/10.1177/0261927X231200201

Martin, J., & White, P. R. (2005). *The Language of Evaluation*. New York: Palgrave Macmillan. https://doi.org/10.1057/9780230511910

Matthews, J., & Volpe, C. R. (2023). Academics' perceptions of ChatGPT-generated written outputs: A practical application of Turing's Imitation Game. *Australasian Journal of Educational Technology, 39*(5), 82–100. https://doi.org/10.14742/ajet.8896

Mündler, N., He, J., Jenko, S. & Vechev, M. (2023). *Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation*. https://doi.org/10.48550/arXiv.2305.15852

Muñoz-Ortiz, A. & Gómez-Rodríguez, C. & Vilares, D. (2024). Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review*. https://doi.org/57. 10.1007/s10462-024-10903-2.

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly. *Applied Cognitive Psychology*, *20*(2), 139–156. https://doi.org/10.1002/acp.1178

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*. https://doi.org/10.1016/j.mlwa.2024.100602

Rochi. (2025, October 30). Can AI tools replace content writers? Retrieved May 2, 2025, from https://plusai.com/blog/can-ai-replace-writers.

reliablepayperhead. [ChatGPT]. (2025, March 4). *Every single day I tell GPT to never use em dashes, but GPT doesn't care what I want.* [Online forum post]. Reddit. Retrieved June 1, 2025, from https://www.reddit.com/r/ChatGPT/comments/1j8vq4z/every_single_day_i_tell_gpt_to_never_use_em/

Sallami, D., Chang, Y. K., & Aïmeur, E. (2024). *From Deception to Detection: The Dual Roles of Large Language Models in Fake News*. https://doi.org/10.48550/arxiv.2409.17416

Sparks, N. (1999). *The Notebook*. Grand Central Publishing. Retrieved April 3, 2025, from https://www.ebook.nl/ebook/9789000325269-the-notebook-het-dagboek-nicholas-sparks/

The Authors Guild. (2024, March 15). *AI is driving a new surge of sham "Books" on Amazon*. Retrieved June 18, 2025, from https://authorsguild.org/news/ai-driving-new-surge-of-sham-books-on amazon/#:~:text=Enter%20AI.,Amazon%20almost%20immediately%20after%20publication

Thelwall, M. (2019). Reader and author gender and genre in Goodreads. *Journal of Librarianship and Information Science*, *51*(2), 403–430. https://doi.org/10.1177/0961000617709061

Trigg, M. (2024, April 17). *Think AI is bad for authors? The worst is yet to come*. Writer'S Digest. Retrieved June 9, 2025, from https://www.writersdigest.com/be-inspired/think-ai-is-bad-for-authors-the-worst-is-yet-to-come

Twenge, J. M. (2024). *Are books dead? Why Gen Z doesn't read*. Generation Tech. Retrieved June 8, 2025, fromhttps://www.generationtechblog.com/p/are-books-dead-why-gen-z-doesnt-read

Vodolazka, S.., Krainikova, T., Ryzhko, O., & Sokolova, K. (2024). Authors versus AI: Approaches and Challenges. *Current Issues of Mass Communication*, *35*, 73-89. https://doi.org/10.17721/CIMC.2024.35.73-89

Wang, Y., Pan, Y., Yan, M., Su Z. & Luan, T. H. (2023) A Survey on ChatGPT: AI–Generated Contents, Challenges, and Solutions, *IEEE* Open Journal of the Computer Society*, 4.* https://doi.org/10.1109/OJCS.2023.3300321.

White, P. R. (2015). Appraisal Theory. In K. Tracy, C. Ilie, & T. Sandel (Eds.), *The International Encyclopedia of Language and Social Interaction (pp. 1-7*). Wiley.

Wu, J., Gan, W., Chen, Z., S. Wan & Lin. H. (2023), *AI-generated content (AIGC): A survey*, https://doi.org/10.48550/arXiv.2304.06632

Yang, S., Chen, S., Zhu, H., Lin, J., & Wang, X. (2024). A comparative study of thematic choices and thematic progression patterns in human-written and AI-generated texts. *System, 126*. https://doi.org/10.1016/j.system.2024.103494

Yildiz Durak, H., Eğin, F., & Onan, A. (2025). A Comparison of Human-Written Versus AI-Generated Text in Discussions at Educational Settings: Investigating Features for ChatGPT, Gemini and BingAI. *European Journal of Education, 60*(1). https://doi.org/10.1111/ejed.70014

Appendices:

Appendix A: Suvery Questionnaire

Section 1: Informed Consent
Welcome and Thank You for Participating!

You are invited to take part in a research study examining how people perceive and evaluate (romantic) texts written by humans versus generated by AI. This study is part of a master's thesis project.

Participation Details:

- You will read two short paragraphs
- You will answer questions related to your perception of each text and a final assessment after reading two texts.
- The survey should take approximately 10–15 minutes.
- Your responses will remain anonymous and confidential.

Voluntary Participation:

- Participation is entirely voluntary.
- You can withdraw at any time without penalty.
- By clicking "Consent", you consent to participate in this study.

If you have questions about this research, please contact Mingjue Liu (696495ml@eur.nl).

[ ] Consent. [ ] I do not consent.Te

Section 2: Demographic Information

1. **What is your age?**
   (Open response -numbers)
2. **What is your gender?**

- Male
- Female
- Non-binary / Other
- Prefer not to say

3. **What is your highest level of education completed?**

- High school diploma or equivalent
- Bachelor's degree
- Master's degree
- Doctorate or equivalent
- Other (please specify): _____

4. **How would you describe your proficiency in English?**

- Beginner
- Intermediate
- Advanced
- Fluent
- Native

Section 3: Text Evaluation (Completed Separately for Text A and Text B)

Text A / Text B: *(Randomized order)*

Please read the following passage carefully. This is **Text A**.
(Text A)

Thoughts of the summer they'd shared came back to her, and as she stared at him, she noticed how little he'd changed since she'd last seen him. He looked good, she thought. With his shirt tucked loosely into old faded jeans, she could see the same broad shoulders she remembered, tapering down to narrow hips and a flat stomach. He was tan, too, as if he'd worked outside all summer, and though his hair was a little thinner and lighter than she remembered, he looked the same as he had when she'd known him last.

When she was finally ready, she took a deep breath and smiled."Hello, Noah. It's good to see you again."Her comment startled him, and he looked at her with amazement in his eyes. Then, after shaking his head slightly, he slowly began to smile."You too . . . ," he stammered. He brought his hand to his chin, and she noticed he hadn't shaved. "It's really you, isn't it? I can't believe it. . . ."

Please read the following passage carefully. This is **Text B**.
(Text B)

Noah glanced up from his untouched coffee, and for a second, the air between them stilled. Allie—her auburn hair in soft waves now, her lips slightly parted in hesitation—was as striking as ever. The warm lighting caught the gold in her earrings, making her seem almost unreal. "I wasn't sure you'd come," he said, his voice even, though his grip on the ceramic mug tightened. She swallowed, stepping closer. "Neither was I," she admitted, tucking a loose strand of hair behind her ear. A habit he still remembered.

She slid into the chair across from him, their knees grazing beneath the table. A rush of warmth spread through Noah at the contact, but he forced himself to stay still. "It's been a long time," she said, studying him. He nodded, exhaling slowly. "Yeah. Too long." The silence stretched, heavy with things unsaid. And yet, in that quiet space, something familiar still lingered.

Comprehension Check

5. **Which option best summarizes the main message of the text? (A-1, B-2)**

- A warm reunion sparks memories of a past summer romance
- Two former lovers reconnect in a café
- Two friends catch up without addressing their emotional history
- I'm not sure

6. **How confident are you in your understanding of the text?**

- Not at all confident
- Slightly confident

- Moderately confident
- Very confident
- Extremely confident

Perception Measures (1 = Strongly Disagree, 5 = Strongly Agree)

Coherence & Cohesion
7. The text is easy to read and follow.
8. The text presents its ideas in a consistent way.
9. Transitions between sentences and paragraphs are smooth.
10. The text contains self-contradictory information. *(Reverse-coded)*

Linguistic Naturalness
11. The text uses simple and easy-to-understand words.
12. The text uses a diverse range of vocabulary.
13. The text uses many adjectives and descriptive details.
14. The sentences in the text are generally long.
15. The sentence structures are repetitive. *(Reverse-coded).*

Emotional Tone & Expressiveness
16. The text expresses clear negative emotions (e.g., fear, sadness, anger).
17. The text expresses clear positive emotions (e.g., joy, hope, excitement).
18. The emotions expressed feel authentic.

Tone of voice & Human Perspective
19. The author's emotional attitude toward the topic is clear.
20. The text conveys a personal tone of voice.

Overall Impression & Source Attribution.
21. I believe this text was written by:

- A human
- AI
- Not sure

Final Comparison

After reading both texts, which one do you think is more likely to be written by human?

- Text A
- Text B
- I'm not sure

How confident are you of your final assessment?

- Not at all confident
- Slightly confident
- Moderately confident
- Very confident
- Extremely confident

What cues led you to your decision about the text? (Minimum 10 words)
(Open-ended)

Appendix B: Thematic analysis

Assessment	Reasons	Theme
Not sure	well, i think text A flows better.	flow
A	The structure is more familiar for me. It feels like the common structure that American writers use.	lingustic naturalness: structure
B	bebebebe bebbee bebbee bbebe bbebebbeeb	
A	11 q w w w. We. W w. W w w w	
A	Text A feels more authentic and awkward, which is exactly what humans would feel in that situation.	emotions: authentic,awkward
B	The emotion from text B could be more obvious to tell and more like a story.	emotion: obvious
A	Story feels more personal, less, fanfictiony	authentic & personal
A	The portrayal of the character's inner thoughts and emotions feels very authentic.	emotions: authentic
B	Text A has a lot of works I don't know	
B	Text A seemed very out of touch, like there was no emotion in the text.	emotion: absence
A	AIs tend to like to use a lot of adjectives in their writing to fill up the word count	lingustic naturalness: adj
B	the way the sentences were structured and also the frequent use of ‘-‘ mid sentence.	lingustic naturalness: structure/ punctuation: dash
Not sure	Sentences are too long in the first text	
A	During the reading process, I felt that text B's grammar was more sophisticated.	lingustic naturalness: grammar
B	The emotions are more raw and real	emotion: raw & real
A	Make the details more realistic.	realistic
A	simple language	lingustic naturalness: simple
A	morerealistic interms of language	lingustic naturalness: realistic
B	more realistic	realistic
A	Crisp precise sentences by AI I'm Text B	lingustic naturalness: sophisticated sentences
A	A give me a feeling of strong emotion,quite easy to understand the emotions and get into it.	emotion: realistic
B	Text B is easy to read, and vocabulary is usually used in daily life.	lingustic naturalness: simple words
A	I could resonate with Text 1 more	emotion: aligning with personal experience
A	In B, quotes & emotional descriptions switch too frequently. Harder to follow. A is more natural.	coherence: jumping around, difficult to follow
Not sure	Not sure but basically depends on insights from previous experience	
B	.....	

B	The sentence which used hyphen "-" couldn't be written by AI I thought.	punctuation: hyphen
B	The latter one conveys more emotional feelings.	emotion: strong
B	每一句的字数变化比较大 不像 ai 的风格(The number of words in each sentence varies a lot, unlike the AI's style.)	lingustic naturalness: variety in sentence lengths
B	AI writing and communication habits.	
A	Stronger emotions and clearly attitude	emotion: strong/ attitude: clear
A	First text had some repetitive sentence structures and focused a lot on physical appearance description	sentence structures: repetitive /lingustic naturalness: detailed description
A	Text A is more clearly and used some emotial word to express personal attitude	emotion: strong and clear /personal attitude: clear
A	My instinct and gut feeling	
A	A sounds more natural B too many big words	lingustic naturalness: natural, simple words
B	The words text B are using is more close to spoken English, plus the grammar is more like a casual essay.	lingustic naturalness: simple language, simple gramma
A	The tone of the contents, and the way of using the words.	tone of voice
A	Texts written by human are generally shorter and structure of the sentence is more easier.	lingustic naturalness: short & simple sentence structure
A	A is more clear and easy to read	coherence: easy to read
B	With the AI able to replicate styles, it is hard to determine whether it is written by human or AI. However, if a person writes like text A I would say it is very redundant	lingustic naturalness: redundant sentence
A	The use of “—“ in the second one and some words that I usually don’t use in everyday life	punctuation: dash
A	No special thoughts from nd	
A	Text B had a few long dashes, which is usually a telltale sign of its AI origin. Furthermore, it used more periods and less run on sentences than Text A.	punctuation: dash/ more period /less run on
B	The use of this long dash "-" is very AI coded, and the first text was clearly an amateur writer, the kind of writing AI can't produce	punctuation: dash
A	The language being more simple and the way the sentences are structured.	lingustic naturalness: simple words /simple sentence structure
A	The choice of words and how it's written	lingustic naturalness:word usage
A	The first text felt more personal and descriptive. The second text feels like the AI has been told to sound like a human.	tone of voice: personal/ lingustic: descriptive

A	Text B is ai because the "--" marks in between the sentences gave it away and the text was less emotional charching to me. The first text gave me a rush of that something was going on.	punctuation: dash /emotion: strong
B	The way the story was phrased seemed more AI like in text A. Also, text B reminded me a lot of wattpad stories which made me believe it was written by humans	similar to wattpad stories
A	Text B used this — and I haven't seen that in a book (yet)	punctuation: dash
A	It felt more natural. The text was easy to read and the story seemed more realistic, like something that actually happened.	coherence: easy to read
A	Text B had a great number of adjectives, was very vague, the use of punctuation, and overall it felt not authentic. Text A felt way more real.	lingustic naturalness: many adj/ use of punctuation / emotional: not authentic
Not sure	Feelings from a human being	
A	Text B uses dashs when writing, and the emotional expression is richer.	punctuation: dash/ emotion: richer
Not sure	B is better, but it depends on if the writing skills of author is better than AI, or both of them written by AI	
B	Through the application of punctuation marks, and sentences.	punctuation
B	Text B just feels more fluent and easier to read	coherence: easy to read, fluent
B	Text B reads more fluently and the emotional interactions between the characters are more sophisticated. More like human.	coherence: fluent / emotion: stronger & more sophisticated
A	The words in the first texts seemed like words humans don't use often	lingustic naturalness: word usage
Not sure	cannot figure out which text is written by human	
B	More Description about the movements and emotions	lingustic naturalness: adj (descriptive)
A	AI writing usually less descriptive and more straightforward.	lingustic naturalness: less descriptive
B	The second text feels more with character and uses more adjectives to portrait the situation.	lingustic naturalness: more adjectives
A	I feel like text B contains too much details that are not relevant to the main topic.	lingustic naturalness: too descrptive/ details
B	The text looks a lot more shorter than text A because text A has a lot more words	lingustic naturalness: wordy
B	It is easy to read and feel like more eloquent	coherence: easy to read/ lingustic: colloquial
A	Similar to my writing style	lingustic naturalness: human writing style

A	The atmosphere built through very detailed descriptions such as tucking hair behind the ear, a touch of the knees	lingustic naturalness: detailed descriptions
Not sure	Everything in the text is pretty descriptive and verbally expressive	lingustic naturalness: descriptive and expressive
B	The second text conveyed more emotion, the sentences were shorter and overall it sounded more human	emotion: stronger
A	It seems like text B had more colloquial words used in it.	natural word usage
B	the use of m-dashes and just the writing	punctuation: dash
A	It's more of a guess, but text A felt a bit less standard and a bit more personal.	less structured /personal tone of voice
B	Text B was filled with metaphors that might be too complicated for AI to come up with	lingustic naturalness: descriptive (metaphors)
A	The fact that the sentences seemed authentic and the structure wasn't repetitive.	lingustic naturalness: non-repetitive structure
A	I think use of various en-dashes in the paragraph was one if the cues	punctuation: dash
B	slightly more immersive than the first one	immersive
A	The -- dashes in the AI text.	punctuation: dash
B	Text B felt more authentic with grammar widely used by individuals text A was more expressive with advanced words	lingustic naturalness: simple (used by human) grammar
A	Text B is more like just description of facts, not emotions	emotion: weak/ descriptive facts
B	the use of ''-'', the structure of the sentences, the text A seemed more natural,more emotional, more vague like a real author could make it whereas when reding text, I really feel less the emotions.	punctuation: dash / emotion:less emotion
B	I tried to look at illogical things or something that seemed more "human", but it was hard	
B	Text B was more exagerated	
B	It seemed quite original, something AI can't do	
B	The adjectives were suitable for the situation and relatable.	lingustic naturalness: suitable adjectives