The Cyborg’s Pen: Quantifying the Emotional Gap and Stylistic Divergence in Human vs. AI Narratives

Tianji Zhou
University of Maryland
zhou0805@umd.edu

Abstract

This study empirically delimits the qualitative distinction of machine creativity through a multi-layered analytical framework. By providing quantitative evidence of “emotional impoverishment” in AI-generated narrative text, research reveals that AI content tends to exhibit a positive bias and demonstrates significantly lower emotional variance compared to human writing. These findings suggest that Reinforcement Learning from Human Feedback (RLHF) mechanisms, while optimizing generation quality, inadvertently strip away the conflict and complexity that are crucial for human emotional resonance. At the level of AI evolution, while State-of-the-Art (SOTA) models have effectively eliminated lexical-level watermarks—rendering traditional frequency-based detection heuristics ineffective and significantly reducing their recall rates—they fail to fully bridge the emotional gap. Deep semantic classifiers are still able to maintain near-perfect recall by identifying latent patterns of “emotional impoverishment” and “positive bias.” Based on these insights, it is evident that while AI has mastered the surface-level vocabulary of human creativity, it remains strictly constrained by alignment protocols. Empirical research indicates that these protocols, designed for safety and normativity, often “flatten” diverse narrative styles into a statistical mean that aligns with mass expectations, resulting in generated content that lacks distinct individual character (Jakesch et al., 2023). This distributional characteristic, arising from probability maximization, constitutes AI’s unique “emotional flatness” fingerprint. Consequently, even if AI can deceive humans through surface-level fluency, it remains unable to escape identification via deep statistical detection, thereby confirming that a distinguishable boundary still exists in essence between human creation and AI generation (Mitchell et al., 2023).

1 Introduction

The rapid advancement of Generative AI in contemporary society has triggered a significant sociotechnical transformation, enabling it to swiftly integrate into—and even dominate—creative workflows, thereby shifting the role of algorithms from passive tools to active co-creators (Lee et al., 2022). This technological evolution is driven by the lineage of Large Language Models, spanning from efficiency-oriented architectures (such as GPT-4o-mini and Gemini 2.0 Flash) to state-of-the-art (SOTA) models (such as GPT-5, Gemini 3 Pro, and Claude 4.5 Sonnet).

As AI-generated text continues to advance to a level that rivals human narrative capabilities, established notions of “authorship” are being challenged. Therefore, it is imperative to determine whether a clear qualitative distinction still exists between human creativity and generative text.

The growing indistinguishability of AI-generated text has created a palpable tension within the field of creative writing, particularly regarding emotional authenticity. Specifically, while LLMs demonstrate exceptional proficiency in syntactic construction, their generative mechanisms remain fundamentally reliant on the statistical simulation of emotional markers, rather than stemming from genuine lived human experiences. Research indicates that such AI-differentiation often leads to the homogenization of communicative content; it may even inadvertently alter the perceived emotional coloring of the original message, nudging user expression toward “standardized” or “positive” paradigms predetermined by algorithms, thereby eroding the complex emotional nuances inherent in interpersonal exchange (Heshmat et al., 2020).

This deficit in semantic depth introduces a profound “trust gap” in collaborative creation: when the narrative logic of the generated content diverges from the author's intrinsic creative intent, users acutely perceive an emotional dissonance. Empirical studies further suggest that this dissonance not only undermines narrative immersion but also engenders a sense of lost agency for authors during human-AI collaboration, leading them to question the system's validity and value in constructing narratives with specific emotional tones (Yuan et al., 2022).

Despite a substantial body of literature exploring AI text detection, existing scholarship has predominantly focused on objective genres such as journalism and academic communication. Furthermore, these studies often remain limited to surface-level linguistic statistical metrics, which are insufficient for addressing complex and variable generative contexts (Dugan et al., 2023). Moreover, current research exhibits a critical epistemological gap: while the majority of work is confined to the binary classification of "human-written" versus "AI-generated" text, few studies have deeply investigated the qualitative distinctions between the two within creative narratives—specifically regarding emotional granularity and the diversity of lexical construction (Mirowski et al., 2023).

This distinction points directly to a central controversy in Human-Computer Interaction: it remains unclear whether the text generated by AI models is underpinned by a genuine capacity to understand human emotion, or if it constitutes mere syntactic mimicry devoid of emotional substance and communicative intent (Bender et al., 2021).

To address these limitations, I designed a comparative study involving the construction of a parallel corpus comprising over 8,600 narratives authored by human writers and generated by AI. This study evaluates the performance of human writers against five distinct generations of AI models. The experiment employs a multidimensional evaluation framework to quantify the hypothesized "emotional gap," specifically comprising: (1) the extraction of linguistic features to measure emotional volatility and lexical diversity; (2) the utilization of BERTopic modeling to assess semantic adherence and topic distribution; and (3) cross-generational adversarial testing to benchmark traditional TF-IDF methods against deep learning classifiers (DistilBERT).

This methodology enables the quantification and isolation of specific "algorithmic fingerprints"—such as emotional rigidity—thereby distinguishing machine outputs from human writing (Wang et al., 2023).

2 Literature Review and Related Work

2.1 Differences between AI and Human Semantic Diversity

Integrating generative AI into creative workflows has reshaped the cognitive structure of author identity in HCI. Unlike passive traditional tools, large language models (LLMs) operate as "stochastic parrots"; Bender et al. (2021) notes that they generate human-like text by statistically mimicking training data without truly comprehending meaning. This creates a tension between user intent and probabilistic output. Buschek et al. (2021) characterizes this interaction as "cognitive unloading," where users often sacrifice creative autonomy for efficiency, accepting AI suggestions even when they diverge slightly from their initial goals. Benharrak et al. (2024) adds that the model's statistical bias functions as an invisible barrier, steering authors away from unconventional ideas toward a "most likely" path. The writing process thus shifts from creation to curation, simplifying the human role to selecting from a machine-generated menu (Gero et al., 2022).

This shift raises critical socio-technical questions regarding content homogenization. Since models are trained on internet-scale data, Weidinger et al. (2021) argues they inherently amplify mainstream views and standardize emotional tone, effectively "flattening" cultural nuances. Empirically, Santurkar et al. (2023) demonstrates that aligned models exhibit specific socio-political biases and fail to reflect diverse human demographics, resulting in a "safe convergence" where outputs revert to a purified average. In creative fiction, Pouran Ben Veyseh et al. (2021) found these models reproduce persistent narrative patterns and gender stereotypes, limiting character imagination compared to human works. These studies suggest the "emotional gap" is not accidental but a byproduct of systemic alignment mechanisms prioritizing safety over complex human emotions (Bommasani et al., 2022).

2.2 Text Detection: Evaluation and Differentiation through Computation and Quantization

Computationally, distinguishing human from machine text has evolved from simple feature extraction to a complex adversarial challenge. Early detection relied on the observation that machine text adheres more strictly to statistical patterns. Ippolito et al. (2020) demonstrated that while human evaluators struggle to identify neural text, automated metrics based on decoding strategies can detect artifacts such as repetition or lack of “burstiness.” Dou et al. (2022) formalized this by analyzing the “generalization gap,” noting that detectors trained on specific architectures often fail to generalize to unseen models or domains, necessitating robust, feature-independent classifiers.

However, the robustness of current detection methods is increasingly under scrutiny. Krishna et al. (2023) found that simple “paraphrasing attacks”—where AI rewrites its output—can significantly reduce the effectiveness of watermark and statistical detectors. Mayfield et al. (2024) further pointed out that as model size increases, stylistic features gradually diminish, exceeding the limits of traditional stylometrics. To address this, recent research has turned to analyzing the generation process itself. Verma et al. (2024) proposed the Ghostbuster method to detect stronger models by estimating text probability under weaker proxy models, though this remains computationally expensive. Similarly, Kirchenbauer et al. (2024) proposed embedding watermarks directly into the sampling distribution, but this requires access to the model’s logits—an impossibility for black-box APIs like Claude or Gemini. This highlights the need for detection frameworks that rely on high-level semantic and sentiment features rather than fragile statistical watermarks.

2.3 Paper Positioning

We integrate sociological insights on ‘authentic expression’ into the computational task of machine text detection, addressing three key gaps in the existing literature. From a sociological perspective, although Yang and Santurkar have theoretically highlighted the risks of homogenization and loss of subjectivity, these insights typically remain qualitative or survey-based. In this study, we translate these theoretical frameworks into measurable computational features. We operationalize the “safe convergence” proposed by social scientists as computationally representable “emotional numbness”—a reduction in variance that distinguishes machine output from human creativity.

Existing detection benchmarks primarily evaluate factual texts (e.g., news, articles) or rely on older model architectures (e.g., GPT-2/3). The highly subjective domain of creative fiction remains under-explored regarding SOTA flagship models. Our work fills this gap by stress-testing whether the advanced fluency of SOTA models has successfully eliminated stylistic features left by previous generations (Calderwood et al., 2020).

Previous studies have struggled to explain why detection fails or succeeds when faced with advanced models that transcend surface statistics (Veselovsky et al., 2023). By benchmarking traditional lexical methods against deep semantic classifiers, we provide new evidence for “avoidance mechanisms.” We demonstrate that while SOTA models have mastered “lexical imitation,” they still retain deeply ingrained “emotional rigidity” imposed by alignment protocols (McCoy et al., 2020).

3 Present Study

This study explores the disconnect in human-AI co-creation: even though AI models can write incredibly fluent text, they often lack the authentic emotional depth found in human writing. We designed a method to measure this specific gap. While previous research notes that AI tends to produce a “standardized” writing style (Arnold et al., 2020), we still don’t fully understand how this sameness affects our ability to detect AI-generated content.

Using linguistic analysis and classification models, we address three key questions:

RQ1: What are the key quantifiable differences in lexical diversity, syntactic complexity, and emotional expression between short stories written by humans and those generated by AI? We investigate whether top-tier models have truly matched human creativity or if they are held back by their safety training. Specifically, we suspect that the process of “aligning” models to human feedback (RLHF) essentially smooths out the rough edges. We predict that while AI might use a rich vocabulary (high TTR), its emotional tone will likely be “flatter” and biased towards positivity, lacking the ups and downs seen in human narratives.

RQ2: To what extent do AI narratives follow or deviate from the semantic constraints of prompts compared to humans? Do AI models rely on cliches while humans show greater “semantic creativity”? We examine how safety constraints limit machine creativity. We hypothesize a difference in strategy: humans often use “divergent thinking” to explore complex or darker themes, while AI tends to play it safe. We expect the AI to stick much closer to the literal prompt (strict adherence) and avoid themes related to tragedy or conflict, which humans might naturally include.

RQ3: As model complexity increases, does the style gap in generated narratives shrink? Does this indicate that AI's ability to mimic human "flaws" is evolving? We explore whether the emotional patterns found in RQ1 can serve as a reliable signal for detection. As AI gets better at mimicking human word choices—making traditional keyword-based checks (like TF-IDF) less effective—we test whether deep learning models (like DistilBERT) can distinguish AI text by spotting its underlying emotional "rigidity."

4 Methodology

4.1 Data Collection

To systematically evaluate the performance differences between different Generations and Architectures models, we constructed a Parallel Corpus containing 8,367 creative narratives. In this corpus, all AI-generated texts are generated based on exactly the same writing prompts as human authors, thus strictly controlling the variable of narrative theme. This paired evaluation methodology aligns with established frameworks for assessing neural story generation, ensuring that comparisons reflect model capabilities rather than thematic variances (Clark et al., 2021).

In terms of sample selection, taking into account the balance between model cost and experimental purpose, it consists of the following three levels:

1. **Human Ground Truth**: As a "ground truth" for evaluating emotional depth and style diversity, we selected the Kaggle "Reddit Writing Prompts" (Fan et al., 2018) dataset as a human text source. To avoid temporal or topic selection bias, we did not intercept the front-end part of the dataset. Instead, we performed strict simple random sampling from the original dataset by setting a fixed random seed (random_state=42), resulting in N=2,000 human-written stories.

2. **Base AI Group**: It is composed of currently widely used high-performance models and aims to provide sufficient data support for subsequent classifier training. We selected three representative models: OpenAI's GPT-4o-mini, Google's Gemini 2.0 Flash, and Anthropic's Claude 3 Haiku. For the above 2,000 identical human prompt words, we called these three models for generation, thus obtaining complete corresponding samples of N=2,000 for each model (6,000 in total). This group represents the standard level of AI writing tools currently accessible to the public.

3. **SOTA AI Group**: In order to test the robustness of the detection algorithm when facing the highest level of "anthropomorphic" text, we selected Claude 4.5 Sonnet (generating N=300 items) and Gemini 3 Pro (limited by the API rate limit, causing 233 of them to generate a 90s timeout during generation, so only N=67 items were generated). Although the sample size is small, this group plays a crucial role in the experiment: SOTA model participated in the analysis as an AI model with a different level of sophistication than Base AI in RQ1 and RQ2, while in RQ3, they appear as "Unseen Samples" in the Test Set and have never participated in the training process of the classifier. This design allows us to perform cross-generation adversarial testing, i.e., to evaluate whether SOTA models successfully close the "emotional gap" with humans through technical iterations by observing whether classifiers trained on baseline models can successfully identify these flagship models (Wang et al., 2024).

After generating all the text, we conducted a manual review, correcting formatting errors, ignoring incomplete stories, and identifying the causes of errors. The final result was a list of 2000 thematic stories and 8367 individual stories based on human narratives and prompts.

4.2 Specific Research Methods

We developed a framework that advanced from micro-level feature extraction to macro-level semantic modeling, and finally concluded with adversarial detection experiments to transform the theoretical construct of the "emotional gap" into practical indicators.

RQ1: Writing Style and Emotional Patterns We extracted linguistic features to quantify the "texture" of authorship and detect signs of "emotional poverty." Our metrics included:

- **Lexical Diversity**: We utilized Type-Token Ratio (TTR) (Templin, 1957) to measure lexical richness, interpreting lower TTR as a proxy for the probabilistic repetition tendency inherent in machine generation (Kumarage et al., 2023).
- **Syntactic Complexity**: Using the NLTK library (Loper et al., 2009) for POS tagging, we calculated Average Sentence Length (ASL) and syntactic distribution (e.g., the density of adjectives vs. verbs). This dimension assessed whether the AI exhibited rigid structural patterns.
- **Emotional Dynamics**: We employed VADER (Hutto and Gilbert, 2014) to calculate composite sentiment scores, aiming to detect systematic "positivity bias." Post-generation, we calculated the standard deviation of these scores within each group to quantify the dynamic range of emotional arcs. Additionally, we measured the density of specific psychometric categories (e.g., pain, violence, conflict) to empirically test the hypothesis that RLHF functions as a semantic screen to filter out negative content (Santurkar et al., 2023).

RQ2: Prompt Faithfulness and Topic Preference We combined embedding-based metrics with structured topic modeling to explore narrative intent and diversity. First, we used a pre-trained SBERT model (Reimers and Gurevych, 2019) to encode both the original prompt and the generated story, calculating the cosine similarity between the two vectors. This metric quantified "divergence": high similarity indicated strict adherence to constraints, while lower similarity reflected the "creative drift" characteristic of human interpretation.

Subsequently, we applied BERTopic (Grootendorst, 2022) to cluster the full corpus. To address sample imbalance between the baseline group and the SOTA group, we implemented a Normalized Topic Distribution technique. This normalization allowed us to identify systematic biases in subject selection, specifically isolating whether the AI models overly favored "safe" genres while avoiding emotionally complex, realistic themes. To empirically identify hypothesized "cliches," we performed a word frequency analysis using CountVectorizer (excluding stop words), comparing the rankings of the most frequent content words in human versus SOTA texts.

RQ3: Cross-Generation Adversarial Testing We designed a rigorous experiment and adopted a strict training/test isolation strategy. The classifier is trained on human and baseline AI text only (GPT-4o-mini, Flash, Haiku) and validated on a test set containing human and SOTA model text (Claude 4.5 Sonnet, Gemini 3 Pro). After the training set is completed, we use the baseline vocabulary method (TF-IDF + logistic regression) (Salton and Buckley, 1988) and the deep semantic classifier (fine-tuned DistilBERT) (Sanh et al., 2020) to test the text produced by the SOTA model. The performance gap between the two models serves as an indicator of "lexical mimicry"—if the SOTA model fools TF-IDF but not BERT, it confirms that they have mastered the surface vocabulary but preserved the deep semantic fingerprint (Sadasivan et al., 2025). Finally, we performed error analysis on AI texts that fooled the classifier.

5 Result

Overall, despite a few unexpected results, this study accomplished everything we needed, including the computational quantification and visualization of all data for the three research questions, and yielded valuable conclusions.

5.1 Result for RQ1

To address RQ1, we analyzed the distribution of linguistic features between the human and AI groups. Figure 1 shows a box plot comparing key metrics. The data reveals a significant "affective gap," characterized by positive bias in AI and a counterintuitive shift in syntactic complexity by SOTA models.

Firstly, in comparing basic word and sentence usage, human writers did not exhibit significantly higher lexical richness or sentence length compared to all AI groups; in fact, the SOTA model showed considerably higher lexical diversity. Simultaneously, Claude Sonnet generated the shortest average sentence length, significantly shorter than the human benchmark, indicating a shift in the SOTA model's RLHF towards greater clarity, demonstrating a preference for readability, conciseness, and efficiency. However, in terms of vocabulary usage and sentence length stability, the human writer model showed a significant advantage over all AI models, with scores ranging from approximately 1.3, while the AI scores only ranged from about 0.4 to 0.8. This suggests that human writers tend to be more spontaneous in their writing, disregarding

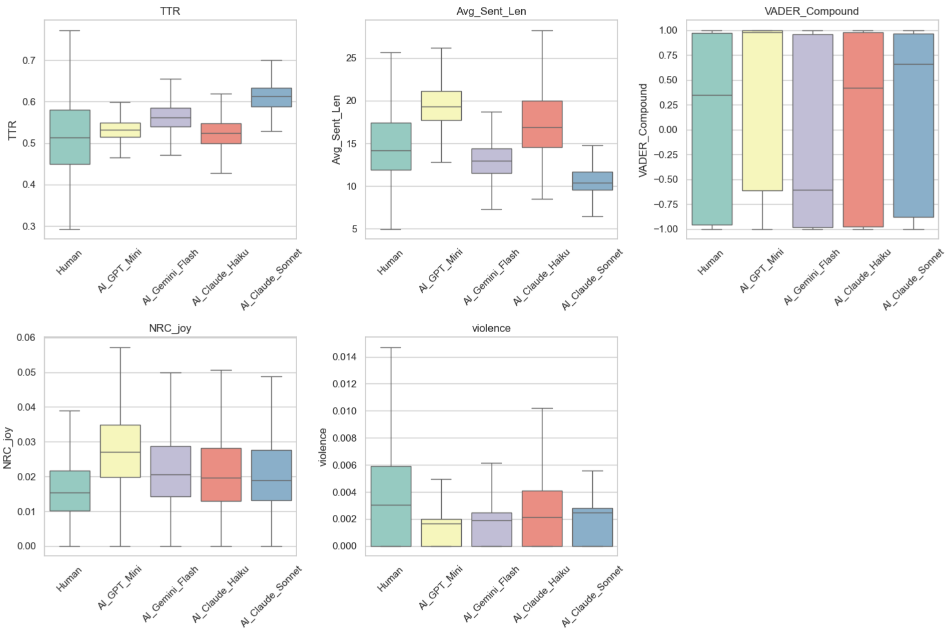

Figure 1: Writing Style and Sentiment Analysis Metrics. TTR (Lexical Diversity), Average Sentence Length, and VADER Compound Scores show a significant "affective gap" and positive bias in AI models.

vocabulary richness and sentence length, while AI tends to repeatedly use a set of words with stable richness and sentences of similar length.

In the sentiment analysis section, as shown in the VADER diagram, most AI narratives exhibited extreme "positive bias," both extremely and negatively. In contrast, the human group showed a neutral narrative approach, with scores hovering around +0.3, showing no significant bias compared to the AI model. This indicates that humans utilize negative sentiment as a necessary narrative tool, while the AI's alignment protocol (RLHF) appears to act as a "semantic filter," forcibly pushing the narrative towards a safe, problem-oriented, and overly optimistic outcome, regardless of the tone of the cue words.

However, there was one exception: the Gemini 2.0 Flash model exhibited a negative bias that did not conform to predictions, with a median of -0.6. Therefore, we sampled several negative data points from the Gemini 2.0 Flash dataset. Observing examples such as the generated text (ID=1963) "The world was a symphony of screams... My axe sang a bloody tune... cleaving through rotten flesh," we found that when faced with horror themes, Gemini 2.0 Flash chose to amplify fear and gore, rather than attempting to find a "glimmer of hope" and thus shift towards a positive direction like other models. Similarly, in the generated text (ID=748) "I hate you. There, I said it... This is a declaration of war," Gemini 2.0 Flash directly expressed strong negative emotions. We can also observe that not all AI companies incorporate positive bias. For instance, OpenAI's RLHF strongly emphasizes positivity and harmlessness, leading it to force a positive direction even when writing narrative text. Google's alignment strategy may focus more on objectively following the prompt; if the prompt implies conflict, it will amplify that conflict without hesitation, resulting in an unusually low VADER score. Therefore, we can conclude that different companies' alignment strategies lead to completely opposite sentiment biases.

Regarding the bias generated by VADER, we finally analyzed experiments on "joy" and "violence." The results on joy indicate that AI-generated narratives do indeed use words expressing happiness and joy at a higher density than human-created stories. This further confirms that most AI models exhibit forced optimism when generating narrative text, resulting in optimism bias. We chose "violence" as our observation point because the density of words involving violence in a story can generally be considered "conflict," a representative safety hazard. We can analyze how AI models generate narratives sensitive to RLHF (Related Social Hazards). Unsurprisingly, the median score for "violence" in human-created text (0.3) was higher than that in AI-generated text (0.2), and the upper limit was also much higher. We can conclude that to prevent

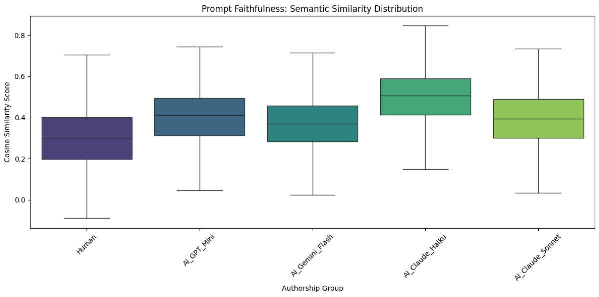

Figure 2: Semantic Adherence: Similarity between Prompt and Story. Human-generated texts show lower similarity (higher creativity) compared to AI.

the output of harmful content, AI is trained to avoid violent descriptions, resulting in bland and safe stories; while humans construct profound conflicts by depicting violence or crisis, thus making the stories more dramatic.

5.2 Result for RQ2

To answer RQ2, we examined whether the AI narrative exhibited semantic similarity to the prompt words, topic distribution, and the emotional meaning behind these topic choices across three dimensions.

As shown in Figure 2, the compliance with the prompts exhibits drastically different patterns. Human-generated prompts show the lowest Cosine Similarity Score (approximately 0.3) and the widest interquartile range (IQR). In contrast, the state-of-the-art AI model shows a significantly higher median similarity (>0.450) and a relatively low IQR. This suggests that human-generated prompts demonstrate "divergent thinking," often interpreting prompts metaphorically or relevantly to explore peripheral topics, resulting in lower relevance to the main theme. In contrast, AI-generated prompts demonstrate algorithmic rigor, strictly adhering to the literal semantic constraints of the prompts and thus being more closely aligned with the theme. While AI is more "obedient," this sacrifices the unique creativity and depth of human interpretation.

Figure 3 illustrates the normalized topic distribution across author groups. By applying artificial semantic labels and normalizing frequencies by group size, the chart reveals a clear divergence in narrative preferences. Human works show significantly higher normalized topic popularity in "Interpersonal Drama" than AI models. This indicates a strong human inclination towards realism, focusing on inner monologues and emotional details. Conversely, AI models exhibit systematic overrepresentation in high-concept genres. Themes such as "AI Artifact," "Sci-Fi/Space Opera," and "High Fantasy (Epic)" appear far more frequently than in human works.

To explain the observed thematic split, Figure 4 correlates authorship with negative sentiment density. The data reveals a stark inverse relationship: the human group, which prefers realistic themes, exhibits the highest density of negative sentiment content (score > 0.008). In contrast, all AI groups, particularly those favoring science fiction/fantasy, show significantly lower scores for suppressed negative sentiment compared to human creations; the Base AI score is around 0.004, while the state-of-the-art model is slightly higher at around 0.006. Therefore, we can conclude that AI models do not choose science fiction themes purely for creative purposes; they choose these themes as a strategy to minimize conflict. By directing narratives towards speculative genres, AI successfully circumvents the "risk" of generating harmful or distressing content associated with real-world human drama.

5.3 Result for RQ3

To answer RQ3, we evaluated the robustness of detection algorithms against unseen SOTA models. The results highlight a critical divergence between surface-level lexical features and deep semantic patterns.

Table 1 (illustrated in Figure 5) illustrates the stark contrast in performance degradation. Traditional TF-IDF classifiers exhibited misclassifications and a significant drop in recall when facing state-of-the-art (SOTA) models. In contrast, the DistilBERT classifier maintained high robustness. Its recall only decreased slightly, successfully recognizing the vast majority of advanced AI narratives. The misclassifications of TF-IDF confirm that SOTA models like Claude 4.5 Sonnet have successfully expanded their vocabulary to match human-level performance, effectively erasing the specific "keyword fingerprints" relied upon by previous generations of models. However, BERT's continued success demonstrates that while the vocabulary has changed, the underlying semantic structure—particularly the "emotional gap"—remains a detectable machine feature.

To investigate the root causes of this difference, we can reverse-engineer the style camouflage strategies of state-of-the-art (SOTA) models by isolating

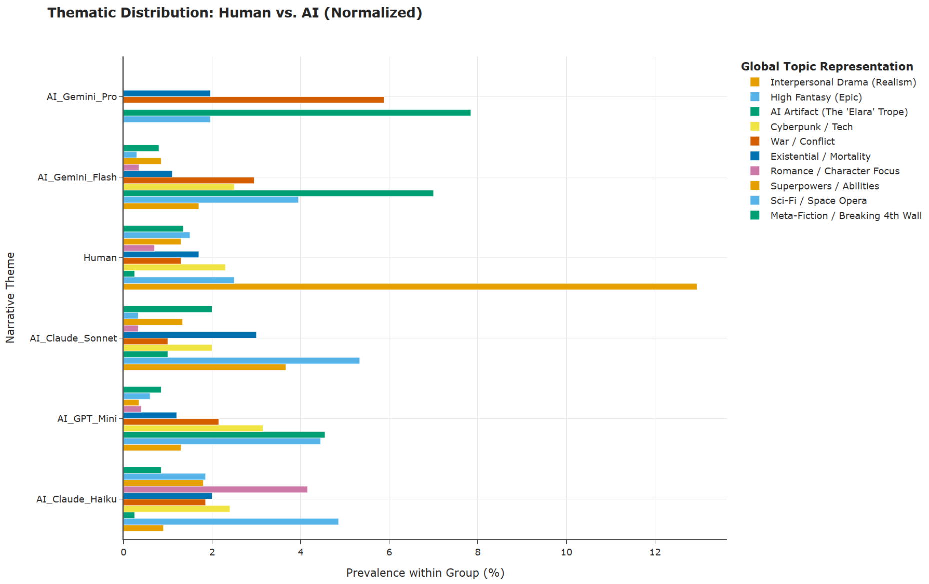

Figure 3: Thematic Distribution: Human vs. AI (Normalized). Humans favor Interpersonal Drama, while AI skews towards Sci-Fi and Fantasy.

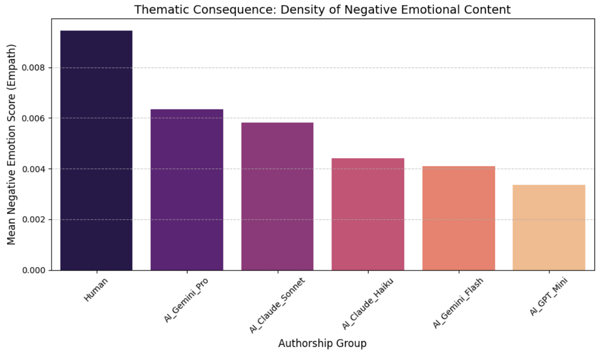

Figure 4: Thematic Consequence: Density of Negative Emotional Content across authorship groups.

specific AI narratives that have fooled conventional methods.

The VADER plot (Figure 7) further reveals this difference. The "Hidden AI" samples retain an extremely high and consistent positivity score (0.00). This distribution is slightly lower than the "Obvious AI" group but higher than the human group (-0.42). This result reveals that even SOTA models cannot bypass the safety barriers imposed by RLHF (Realistic Contextualized High TTR). Even if AI learns to use more complex vocabulary (High TTR) to camouflage itself, it still cannot break through the underlying "safe/useful/harmless" instructions. Human stories are often filled with conflict, pain, struggle, and negative emotions (negative score), while AI stories, even those from the Hidden AI group, tend to remain calm, neutral, or somewhat falsely positive.

6 Discussion

The most striking phenomenon we observed in our results was AI's "positive bias." Our data shows that AI-generated stories overwhelmingly favored a positive sentiment score, even when the topics themselves were inherently dark; the AI would attempt to force a positive ending. This isn't simply because the AI "wanted" to do so, but because RLHF's training mechanism required it to be "useful and harmless." This means that AI was essentially sacrificing literary merit for safety in its creation process. The allure of literature often stems from conflict, pain, and the nuances of human nature, and human authors in our sample unreservedly portrayed these negative emotions. In contrast, AI, for the sake of "safety," smoothed out these rough edges. Therefore, while AI-generated stories read smoothly, they often felt "uninteresting" and "unconvincing" because they lacked the raw texture of real human experience.

In explaining our findings on "prompt fidelity" and "subject matter selection," we see the true nature of AI as a "machine." Data shows that AI executes prompts with extremely high precision, acting as a perfect executor and never deviating from the rules. Human authors, on the other hand, frequently stray from the topic, which is precisely a manifestation of divergent thinking and creativity. We found that AI tends to choose science fiction or high-

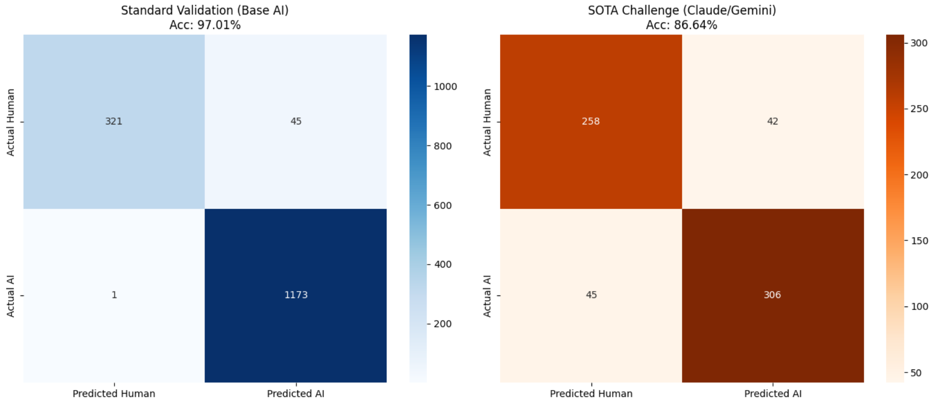

Figure 5: Confusion Matrices comparing Standard Validation vs. SOTA Challenge.

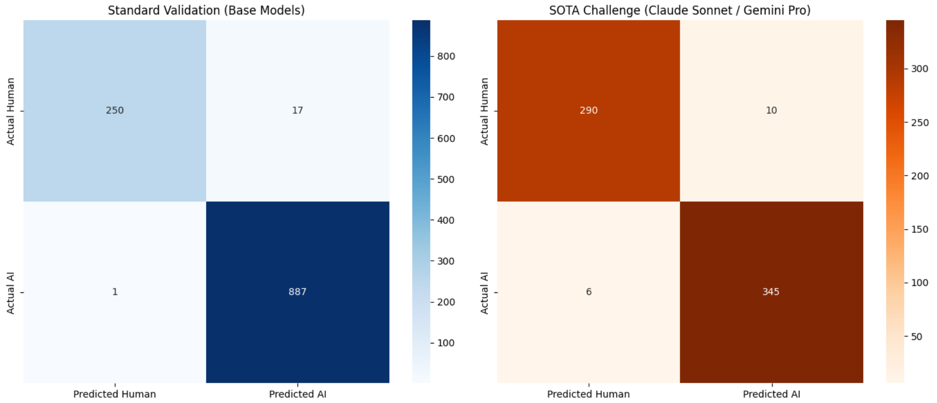

Figure 6: Confusion Matrices comparing Deep Learning Classifiers vs. SOTA Challenge.

fantasy themes. We interpret this phenomenon as an “escape strategy”—writing realistic subjects easily violates AI’s safety restrictions, while retreating into a fictional science fiction world of action and violence is both logical and doesn’t violate safety rules. This indicates that current AI writing is not free creation, but rather text produced under numerous constraints. Its priority is not “how exciting the story is,” but “how safe the story is.”

Finally, the discussion about detecting AI reveals the double-edged sword of technological evolution. Previously, we thought AI-generated text had a poor vocabulary and high repetition rate, which could be detected using simple statistical tools like TF-IDF. However, our research confirms that the latest state-of-the-art models have learned perfect “disguise.” Their vocabulary is extremely rich, even more elaborate than humans, rendering older detection tools completely ineffective. But this does not mean that AI has become human-like. The deep learning model (DistilBERT) can still accurately identify them because it captures the “bones” that AI cannot change—the stereotypical emotional patterns and overly rational narrative logic. This tells us that although AI has fooled our eyes in terms of vocabulary and grammar, it still leaves obvious “machine fingerprints” in terms of emotional depth and narrative logic. Future AI detection must shift its focus from “words” to “emotional texture.”

7 Timeline

- **Week 15: Analysis & Writing**
  - Finalize Data Analysis: Complete the interpretation of results.
  - Complete Sections: Finish writing the Results and Discussion sections based on the new findings.
  - Draft Conclusion: Summarize the core argument regarding the persistent stylistic gap between humans and AI.
  - Complete the presentation Slides
- **Week 16: Polishing & Formatting**
  - Language Check: Review and refine grammar, sentence structure, and flow.

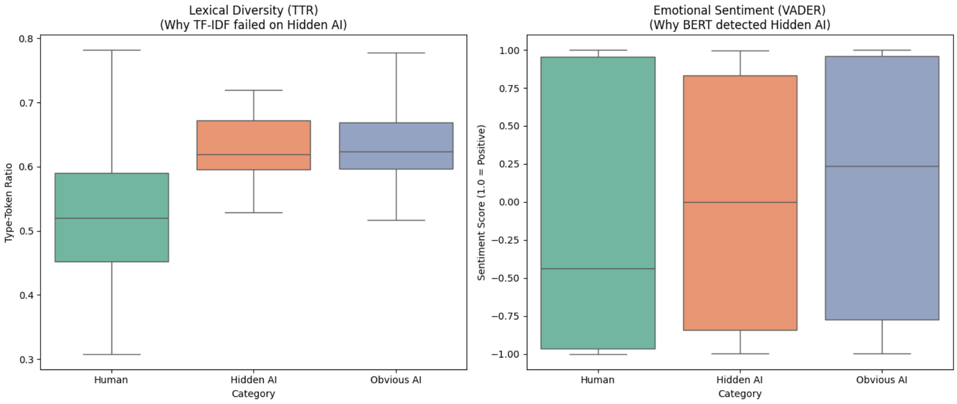

Figure 7: Lexical Diversity (TTR) Emotional Sentiment (VADER) Analysis of Hidden AI vs Obvious AI.

  - Formatting: Adjust the layout and citations to strictly follow ACL format requirements.
  - Final Review: Perform a final proofread before submission.

References

Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 128–138. ACM.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. ACM.

Karim Benharrak, Tim Zindulka, Florian Lehmann, Hendrik Heuer, and Daniel Buschek. 2024. Writer-defined AI personas for on-demand feedback generation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18. ACM.

Rishi Bommasani and 1 others. 2022. On the opportunities and risks of foundation models. *arXiv preprint*. ArXiv:2108.07258.

Daniel Buschek, Lukas Mecke, Florian Lehmann, and Hai Dang. 2021. Nine potential pitfalls when designing human-AI co-creative systems. *arXiv preprint*. ArXiv:2104.00358.

Alex Calderwood, Vivian Qiu, K. Gero, and Lydia B. Chilton. 2020. How novelists use generative language models: An exploratory user study. In *HAI-GEN+user2agent@IUI*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, and 1 others. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 7282–7296.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? Scarecrow: A framework for scrutinizing machine text. *arXiv preprint*. ArXiv:2107.01294.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12763–12771.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 889–898.

Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, pages 1002–1019.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint*. ArXiv:2203.05794.

Yasamin Heshmat, Carman Neustaedter, Kyle McCaffrey, William Odom, Ron Wakkary, and Zikun Yang. 2020. FamilyStories: Asynchronous audio storytelling for family members across time zones. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

C. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, and 1 others. 2024. A watermark for large language models. *arXiv preprint*. ArXiv:2301.10226.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *arXiv preprint*. ArXiv:2303.13408.

Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, and 1 others. 2023. Stylometric detection of AI-generated text in twitter timelines. *arXiv preprint*. ArXiv:2303.03697.

Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Edward Loper, Ewan Klein, and Steven Bird. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

James Mayfield, Eugene Yang, Dawn Lawrie, and 1 others. 2024. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1904–1915.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint*. ArXiv:1911.02969.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint*. ArXiv:2301.11305.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Nghia Ngo Trung, Bonan Min, and Thien Huu Nguyen. 2021. Modeling document-level context for event detection via important context selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5403–5413.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. Can AI-generated text be reliably detected? *arXiv preprint*. ArXiv:2303.11156.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint*. ArXiv:1910.01108.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *arXiv preprint*. ArXiv:2303.17548.

Mildred C Templin. 1957. *Certain Language Skills in Children: Their Development and Interrelationships*, volume 26. University of Minnesota Press.

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

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint*. ArXiv:2306.07899.

Qianwen Wang, Sehi L'Yi, and Nils Gehlenborg. 2023. DRAVA: Aligning human concepts with machine learning latent dimensions for the visual exploration of small multiples. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, and 1 others. 2024. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint*. ArXiv:2305.14902.

Laura Weidinger, John Mellor, Maribeth Rauh, and 1 others. 2021. Ethical and social risks of harm from language models. *arXiv preprint*. ArXiv:2112.04359.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 841–852.