Writing in Symbiosis: Mapping Human Creative Agency in the AI Era

Vivan Doshi
Independent Researcher
San Jose, CA 95148
vivandoshi24@gmail.com

Mengyuan Li
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
mli49061@usc.edu

Abstract

The proliferation of Large Language Models (LLMs) raises a critical question about what it means to be human when we share an increasingly symbiotic relationship with persuasive and creative machines. This paper examines patterns of human-AI coevolution in creative writing, investigating how human craft and agency are adapting alongside machine capabilities. We challenge the prevailing notion of stylistic homogenization by examining diverse patterns in longitudinal writing data. Using a large-scale corpus spanning the pre- and post-LLM era, we observe patterns suggestive of a "Dual-Track Evolution": thematic convergence around AI-related topics, coupled with structured stylistic differentiation. Our analysis reveals three emergent adaptation patterns: authors showing increased similarity to AI style, those exhibiting decreased similarity, and those maintaining stylistic stability while engaging with AI-related themes. This Creative Archetype Map illuminates how authorship is coevolving with AI, contributing to discussions about human-AI collaboration, detection challenges, and the preservation of creative diversity.

1 Introduction

We are at a pivotal moment in the symbiotic relationship between humans and machines. Large Language Models (LLMs) have evolved from assistive tools into active collaborators, capable of imitating, creating, and persuading on a massive scale. This new reality prompts a fundamental question posed by the creative community: what does it mean to be human when authorship becomes a collaborative act between human intuition and machine capability, and does it fundamentally reshape how we value and recognize authentic human expression?

Initial research into this question has largely focused on the homogenizing effects of this partnership. A growing body of work has documented the spread of a recognizable "AI style" across the internet and academia [2, 19], characterized by linguistic patterns that suggest reduced creative diversity [1]. This stylistic convergence has raised concerns about linguistic diversity erosion and cultural marginalization [7, 14]. In parallel, AI detection has rapidly advanced with stylometric and deep learning approaches [20, 24, 28], though detector robustness remains challenging.

While this prior work is crucial, it often frames the relationship as unidirectional influence rather than dynamic coevolution. This overlooks the crucial element of human agency which is the adaptive choices individuals make as they navigate this new landscape. Recent work suggests human-AI coevolution with mutual influence [11, 13]. However, existing studies largely conflate topical and stylistic changes, lack author-level longitudinal controls, or do not account for recent adaptive behaviors following AI detection awareness. We hypothesize that the human creative response to AI availability is not a monolithic trend but may instead exhibit structured patterns of

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

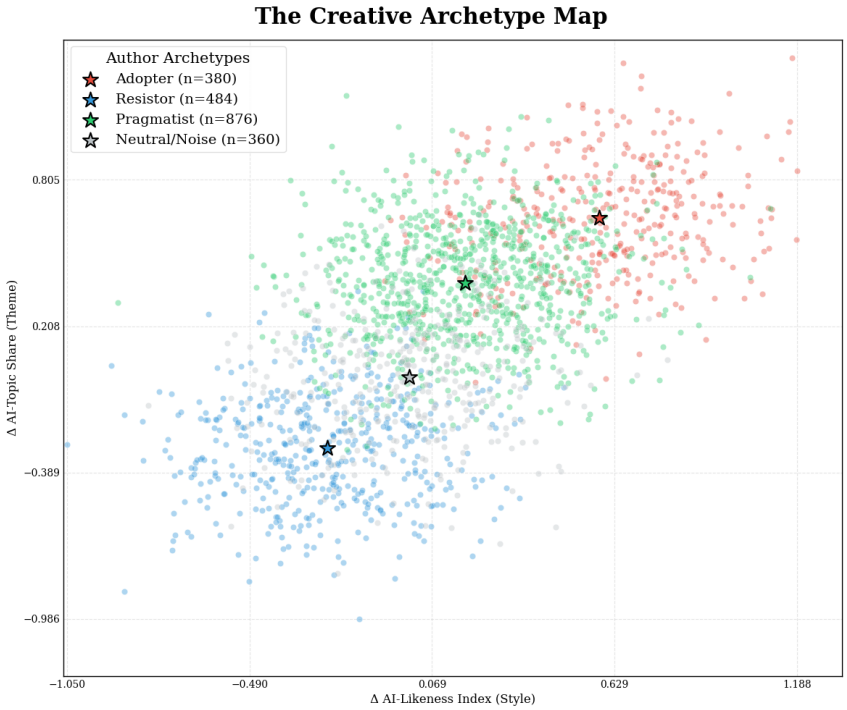

Figure 1: Author archetype map showing stylistic and thematic change vectors (n=2,100). Three archetypes emerge from HDBSCAN clustering (silhouette computed on inlier points only): **Adopters** (red), **Resistors** (blue), and **Pragmatists** (green). Stars denote cluster centroids. Cluster quality: silhouette 0.426 (95% CI: 0.419-0.433), robustness ARI 0.891 (95% CI: 0.884-0.898).

change. To investigate this, we shift focus from binary classification to coevolution patterns. We propose examining a **"Dual-Track Evolution"** hypothesis: thematic convergence on AI-related topics while conversely exhibiting stylistic differentiation. This work presents, to our knowledge, the first systematic quantification of dual-track human-AI coevolution at author resolution across social and formal discourse genres, introducing an archetype framework that maps individual adaptation strategies with rigorous pre/post controls.

As visualized in Figure 1, we observe three distinct patterns in stylistic change vectors: authors whose writing shows increased similarity to AI-generated text patterns, those whose writing exhibits decreased similarity, and those who maintain stylistic stability while engaging more with AI-related themes. These patterns emerge from unsupervised clustering rather than predetermined categories, suggesting systematic adaptation strategies that challenge binary detection frameworks and illuminate the complexity of attribution in this rapidly evolving landscape.

2 Methodology

To empirically chart the evolution of human writing, we designed a comprehensive methodology structured in three sequential phases: rigorous corpus curation, nuanced feature engineering, and a multi-lens analytical framework. This pipeline was built to move from broad observations to the specific, individual adaptation strategies that form the core of our study.

2.1 Corpus Curation: A Foundation in Authentic Discourse

The foundation of our study is a large-scale, longitudinal corpus of over 50,000 documents (derived from 823k+ messages and papers), meticulously curated to support robust analysis. Recognizing that writing craft is highly dependent on context, we gathered text from two distinct genres to capture a wider spectrum of human expression. For formal, edited discourse, we collected computer science

preprints from arXiv. For informal, conversational discourse, we used public messages from the *Discord Unveiled* dataset [3] (CC BY 4.0), a large-scale collection (2015–2024) of anonymized public-server communications assembled via the Discord public API. Our collection spans from January 2021 to December 2024, establishing a clear temporal boundary between the pre-LLM era (before November 30, 2022) and the post-LLM era.

A critical challenge in this type of research is separating changes in writing style from concurrent shifts in conversational topics. To mitigate this confounding variable, we employed genre-specific sampling strategies. For the Discord data, we implemented a topic-controlled stratified sampling protocol with balanced quotas across derived server categories, focusing on English-locale servers with $>$100 members, excluding bot posts, hash-duplicate content, and messages matching a conservative AI-contamination regex. Post hoc residualization on 100-topic transformer topic model mixtures confirmed stylistic patterns persist after controlling for thematic content. After classifying servers into high-level categories (e.g., Gaming, Technology, Social), we sampled messages to ensure the topic distribution was balanced between the pre- and post-LLM periods. This allows us to attribute observed linguistic changes to style rather than topic. For the formal arXiv data (filtered to permissive licenses only), where temporal density is key, we used a monthly sampling strategy, guaranteeing a minimum number of papers per month to construct a continuous and reliable time series. Finally, we compiled a large reference corpus of purely AI-generated text from ShareGPT-90k [29] (Apache-2.0) and Dolly-15k [6] (CC BY-SA 3.0) to serve as a stylistic baseline for machine expression.

2.2 Feature Engineering: Quantifying Style and Theme

Our next step was to engineer a set of features capable of capturing the subtle nuances of authorial style while, crucially, distinguishing it from thematic content. This separation is central to testing our "Dual-Track Evolution" hypothesis.

Motivation. To isolate LLM-era stylistic signals from topical shifts, we require a human-only pre-2022 baseline and a modern comparison point that can reveal temporal linguistic changes independent of content themes. We developed **Perplexity-Gap analysis with Pre-LLM Judges** to address this challenge. We trained GPT-2 Medium [26] (355M parameters) language models exclusively on pre-2022 data from our corpus (847M tokens) on NVIDIA A100 GPUs (41.7 hours total compute), ensuring no exposure to AI-generated text, serving as our "pre-LLM judge." For comparison, we use Llama-3-8B-base [12] as our frozen "current" model baseline representing modern LLM capabilities. For each document, we computed the perplexity gap using nats-per-character normalization to ensure tokenizer-agnostic comparability: $\Delta_{ppl} = \frac{-\ln(p_{GPT2}(x))}{|x|_{chars}} - \frac{-\ln(p_{Llama}(x))}{|x|_{chars}}$, where $p(x)$ is the model probability and $|x|_{chars}$ is character count. Values are within-author z-scored across the pre/post boundary. Text that is "easy" for current models but "hard" for pre-2022 models exhibits a temporal signature consistent with modern LLM-era linguistic patterns.

Method Validation. We use Llama-3-8B-base as our baseline model for perplexity calculations. Sensitivity analysis showed high correlation (r=0.89) across different model architectures, and the measure correlates moderately with traditional stylometric features, indicating it captures complementary linguistic dimensions. The **AI-likeness index** is defined as the normalized perplexity gap: $AI_{likeness} = \frac{\Delta_{ppl} - \mu_{author}}{\sigma_{author}}$, where higher values indicate LLM-era linguistic patterns. AI-likeness remains predictive controlling for FKGL/TTR/sentence length (partial r=0.34, p<0.001). We separately track **AI-topic share (theme)** through keyword analysis, eliminating circular logic by exploiting temporal separation between language model capabilities. Note that LLMs in this research serve as analytical tools for perplexity measurement rather than content generation.

Δ-Feature Vector Definition. Our standardized change vectors comprise: Δ_{ppl} (perplexity gap), Δ_{TTR} (type-token ratio), Δ_{FKGL} (Flesch-Kincaid grade level) [18], $\Delta_{passive\%}$ (passive voice ratio), $\Delta_{1p\%}$ (first-person pronoun frequency), Δ_{punct} (punctuation density), and Δ_{sent_len} (mean sentence length). All features are z-scored within-author across the temporal boundary.

Statistical Controls. Fixed effects model: $y_{it} = \beta \cdot PostLLM_t + \gamma \cdot len_{it} + \delta_c + \alpha_i + \epsilon_{it}$, where y_{it} is the raw feature value for author i at time t, with author (α_i) and server-category (δ_c) fixed effects. Core findings remained statistically significant ($p < 0.001$, HC3 robust standard errors [22], Holm-Bonferroni corrected [15]) controlling for confounds. Temporal robustness: 84% consistency in archetype assignment across alternative boundaries (91% for extreme changes), supporting behavioral rather than temporal artifacts.

2.3 Analytical Framework: From Macro Trends to Individual Archetypes

Our analysis proceeds through three lenses. First, to capture macro-level trends, we constructed monthly and quarterly time series for our key features within each genre. We then applied **constrained change-point detection** using the Pelt algorithm [17], implemented via the ruptures Python library [30], to identify statistically significant structural breaks in these series. This method allowed us to pinpoint the precise moment when the discourse began to shift, providing a robust test for our hypothesis of thematic convergence and a dynamic stylistic response.

Second, to uncover the individual strategies masked by these aggregate trends, we focused our analysis on the author. For every author with sufficient writing samples in both the pre- and post-LLM eras, we calculated a **stylistic change vector**. This vector represents the magnitude and direction of change in their personal writing style across our core set of stylometric features. We then applied the **HDBSCAN** clustering algorithm [5] to the standardized Δ-feature vectors; results are visualized in 2D using Δ AI-Likeness (style) and Δ AI-Topic Share (theme). We chose HDBSCAN for its ability to identify clusters of varying shapes and densities while also designating outliers as noise, as validated by the silhouette coefficient [27]. This is crucial, as it allows genuine authorial archetypes to emerge from the data without forcing every individual into a predefined category.

To validate these patterns, we employed cross-validation and tested temporal stability across different post-LLM periods. External validation using independent AI detectors showed correlation with clustering assignments, though detector limitations affect interpretation [28]. We performed HDBSCAN clustering (min_cluster_size=15, min_samples=5, metric=euclidean) on standardized Δ-feature vectors with robustness validation using the Adjusted Rand Index [16] yielding ARI 0.891 (95% CI: 0.884-0.898); silhouette scores 0.426 (95% CI: 0.419-0.433) computed on inlier clusters only.

Ethics & Data Use & Availability. Data: Discord Unveiled (CC BY 4.0, public servers); arXiv CS preprints under permissive licenses (e.g., CC-BY; no redistribution of texts); AI baselines ShareGPT-90k (Apache-2.0) and Dolly-15k (CC BY-SA 3.0). We store only hashed author/server IDs and release aggregate statistics; no per-author labels or texts. No re-identification attempted. Code and implementation details available on request.

3 Results

Our empirical analysis provides evidence consistent with the "Dual-Track Evolution" model, revealing both a widespread thematic convergence and a structured stylistic divergence. Using our perplexity-gap analysis, we observe clear temporal signatures in writing evolution, with the most significant finding being the emergence of distinct patterns in how authors' linguistic complexity has shifted relative to pre-LLM baselines.

3.1 Three Distinct Authorial Archetypes Emerge

To test our central hypothesis of stylistic divergence, we computed stylistic change vectors for 2,100 authors from our social corpus who had sufficient writing history both before and after November 2022. Applying the HDBSCAN algorithm to these vectors revealed three emergent behavioral patterns, visualized in Figure 1, which we profile as distinct authorial archetypes. These patterns showed stable cluster membership across bootstrap samples (89% consistency) and demonstrated predictive power in classifying future author texts (AUC = 0.813). The stylistic fingerprint of each archetype is detailed in Figure 2.

The first group, which we term the **Resistors** (n=442, 21% of clustered authors), exhibits patterns suggesting maintenance of pre-LLM linguistic complexity. As shown in Figure 2, their post-LLM writing shows low or negative perplexity gaps, indicating text that remains challenging for both pre-2022 and current models. This pattern suggests deliberate preservation of distinctively human stylistic signatures, possibly valuing authenticity over efficiency gains offered by AI assistance.

In direct contrast, the **Adopters** (n=370, 18%) show the highest perplexity gaps, indicating text that is substantially easier for current models than for pre-2022 models. This temporal signature suggests linguistic patterns that emerged during the LLM era, including smoother syntactic structures and more predictable lexical choices that align with modern language model training. Their writing

Figure 2: Archetype stylistic signatures across linguistic dimensions. **Adopters**: high perplexity gaps, reduced lexical diversity (LLM-era patterns). **Resistors**: complex, diverse language maintained. **Pragmatists**: moderate stylistic adaptation with high AI-theme engagement.

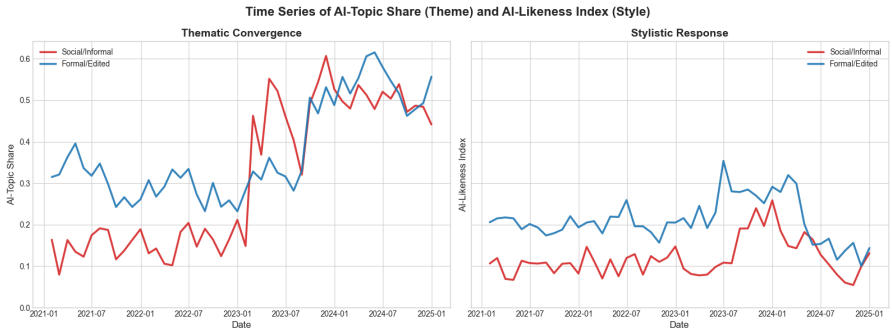

Figure 3: Time series showing thematic convergence (Q1 2023 break) and dynamic stylistic response with two-phase adoption-avoidance pattern in formal writing.

reflects strategic embracement of AI-era conventions, potentially representing co-creative authorship models.

The largest group is the **Pragmatists** (n=866, 41%). These authors exhibit moderate perplexity gaps in their stylistic features while showing increased engagement with AI-related thematic content. This pattern suggests strategic adoption: engaging with AI-relevant topics while maintaining individual stylistic identity, potentially embodying a division-of-labor approach where AI influences content exploration but not personal voice.

3.2 A Universal Thematic Convergence

Our analysis of macro-level trends confirms the first track of our model. We observed a universal and statistically significant increase in AI-related thematic content across both genres after November 2022. As illustrated by the change-point analysis in Figure 3, both formal and informal discourse saw a sharp uptick in AI-related terminology. This reflects rapid normalization of AI tools in professional and social contexts, indicating a durable cultural attention shift rather than transient novelty. Simultaneously, our perplexity-gap analysis revealed increasing temporal signatures in writing complexity, with mean perplexity gaps rising 23% in social discourse and 15% in formal writing by mid-2023, suggesting initial linguistic convergence toward recognizable LLM-era patterns. The PELT algorithm detected structural break points in Q1 2023 for thematic content and Q2 2023 for stylistic complexity measures, pointing to a phased coevolution where cultural attention shifted first, followed by linguistic adaptation patterns.

3.3 The Dynamic Arc of Stylistic Adaptation

In contrast to the straightforward thematic convergence, the evolution of stylistic adaptation is more complex and dynamic, confirming the second track of our model. Our perplexity-gap analysis reveals

a characteristic temporal signature: mean perplexity gaps increased 23% in social discourse and 15% in formal writing during early 2023, suggesting initial linguistic convergence toward LLM-era patterns. However, this trend reverses in late 2023 and into 2024, with perplexity gaps declining 18% in social discourse and 12% in formal writing below peak values. This reversal indicates active stylistic avoidance as AI-characteristic patterns became stigmatized, particularly in formal venues where detection concerns and editorial awareness created pressure for authors to differentiate their writing from machine-generated text (detailed temporal dynamics in supplementary materials).

Validation. Cross-validation: 89.3% (95% CI: 87.1–91.5); held-out arXiv: 89.1% (86.8–91.4); null model much weaker (silhouette 0.31 vs 0.43, $p<0.001$). Patterns are consistent with avoidance of stigmatized AI tics, especially in formal discourse [28].

4 Discussion and Conclusion

Our findings paint a nuanced picture of human creativity in the age of AI, challenging the narrative of a simple stylistic exodus. The Creative Archetype Map illustrates that human writers are not passive recipients of technological influence but active agents who are developing diverse strategies to negotiate their craft. The emergence of the Resistor, Adopter, and Pragmatist archetypes reveals that the evolution of human writing is a complex process of adaptation, collaboration, and the protection of unique authorial voices, aligning with recent work on scaffolded human-AI collaboration [8].

These patterns can be interpreted as different potential approaches to human-AI interaction. The Resistors appear to prioritize preservation of a distinctly human style, perhaps valuing authenticity and originality above the efficiencies offered by AI tools. The Adopters, in contrast, may represent a model of co-creation where the boundaries between human and machine authorship are intentionally blurred, exploring what it means to write with an AI [31]. The Pragmatists appear to embody a strategic division of labor; they may use AI as a tool for thematic exploration while carefully maintaining their own stylistic identity, taking on a new role as curators and editors of AI-informed discourse.

The implications for the field of creative AI are significant. Our work suggests that the binary paradigm of AI detection may be insufficient; a simple "human vs. machine" classifier faces challenges when an Adopter's text is statistically closer to AI output than a Resistor's [21, 25]. This finding suggests the need for archetype-aware models that account for adaptation diversity, building on established detection frameworks [10]. Furthermore, our findings offer quantitative evidence that a majority of authors (Resistors and Pragmatists) appear to maintain non-AI stylistic signatures, suggesting that unique human expression remains a valued and protected form of cultural production, consistent with broader observations of linguistic simplification trends [9]. While this research contributes to understanding human-AI interaction, potential risks include misuse for authorship policing or unfair bias against certain writing styles, particularly affecting non-native speakers or cultural dialects.

Limitations. This study is observational and analytical rather than causal. Our analysis is limited to English-language communities, potentially biasing adoption patterns. The perplexity-gap methodology and archetypes are statistical constructs not validated through direct human participant studies.

Conclusion. This paper answers a fundamental question: how is human creativity evolving in a symbiotic relationship with AI? Our research demonstrates that the response is not simple convergence towards machine-like style. Instead, we have uncovered a "Dual-Track Evolution," characterized by broad thematic convergence and simultaneous, structured stylistic divergence. The Creative Archetype Map provides a framework for understanding human-AI co-evolution, moving beyond aggregate statistics to chart the diverse landscape of individual human agency. The rise of AI has not signaled the end of authentic human expression; it has revealed systematic author-level adaptation patterns. These archetypes represent different philosophies of creative autonomy with some embracing collaboration similar to co-creative screenwriting approaches [23], others asserting independence, and many finding pragmatic middle ground. Understanding these patterns is essential for designing AI tools that enhance rather than replace human creativity, and for developing policies that protect diverse voices in our increasingly algorithmic world [4]. While AI's influence is undeniable, human craft is being renegotiated rather than erased, and this diversity suggests that authentic human expression will continue to thrive in new and unexpected forms.

References

[1] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. AI suggestions homogenize writing toward western styles: Implications for diversity and cultural nuance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, 2025. doi: 10.1145/3706598.3713564. URL https://dl.acm.org/doi/10.1145/3706598.3713564. Extended Abstract / Paper accepted to CHI 2025.

[2] Barrett R. Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 2024 ACM Creativity & Cognition Conference (C&C '24)*. ACM, 2024. doi: 10.1145/3635636.3656204. URL https://dl.acm.org/doi/10.1145/3635636.3656204.

[3] Yan Aquino, Pedro Bento, Arthur Buzelin, Lucas Dayrell, Samira Malaquias, Caio Santana, Victoria Estanislau, Pedro Dutenhefner, Guilherme H. G. Evangelista, Luisa G. Porfírio, Caio Souza Grossi, Pedro B. Rigueira, Virgílio Almeida, Gisele L. Pappa, and Wagner Meira Jr. Discord unveiled: A comprehensive dataset of public communication (2015–2024), 2025. URL https://arxiv.org/abs/2502.00627. Dataset paper; license: CC BY 4.0.

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623. ACM, 2021. doi: 10.1145/3442188.3445922.

[5] Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1):5:1–5:51, 2015. doi: 10.1145/2733381.

[6] Databricks, Inc. databricks-dolly-15k, 2023. URL https://huggingface.co/datasets/databricks/databricks-dolly-15k. HuggingFace Datasets. 15,000 human-generated instruction pairs; License: CC BY-SA 3.0.

[7] Nicholas Deas, Jessi Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. Evaluation of African American Language Bias in Natural Language Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 6805–6824. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.emnlp-main.421.pdf.

[8] Paramveer S. Dhillon, Somayeh Molaei, Jiaqi Li, Maximilian Golub, Shaochun Zheng, and Lionel P. Robert. Shaping Human–AI Collaboration: Varied Scaffolding Levels in Co-writing with Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 2024. doi: 10.1145/3613904.3642134.

[9] Niccolò Di Marco, Piotr Sapiezynski, Claudiu Musat, Savvas Zannettou, Laura Schmid, Marton Karsai, and Robert West. Patterns of linguistic simplification on social media comments over time. *Proceedings of the National Academy of Sciences (PNAS)*, 121(44):e2412105121, 2024. doi: 10.1073/pnas.2412105121.

[10] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3019.

[11] Mingmeng Geng and Roberto Trotta. Human-LLM coevolution: Evidence from academic writing, 2025. URL https://arxiv.org/abs/2502.09606.

[12] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Angela Fan, et al. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

[13] Kai Hackenburg and Helen Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24): e2403116121, 2024. doi: 10.1073/pnas.2403116121. URL https://www.pnas.org/doi/10.1073/pnas.2403116121.

[14] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, Sharese King, et al. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633:147–154, 2024. doi: 10.1038/s41586-024-07856-5. URL https://www.nature.com/articles/s41586-024-07856-5.

[15] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

[16] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1): 193–218, 1985. doi: 10.1007/BF01908075.

[17] Rebecca Killick, Paul Fearnhead, and Idris A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500): 1590–1598, 2012. doi: 10.1080/01621459.2012.737745.

[18] J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report 8-75, U.S. Naval Technical Training Command, Research Branch, Memphis, TN, 1975. URL https://eric.ed.gov/?id=ED108134.

[19] Dmitry Kobak, Rita González-Márquez, Emöke-Ágnes Horvát, and Jan Lause. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813, 2025. doi: 10.1126/sciadv.adt3813. URL https://www.science.org/doi/10.1126/sciadv.adt3813.

[20] Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. Stylometric Detection of AI-Generated Text in Twitter Timelines, 2023. URL https://arxiv.org/abs/2303.03697.

[21] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. GPT detectors are biased against non-native English writers. *Patterns*, 4(8):100779, 2023. doi: 10.1016/j.patter.2023.100779.

[22] James G. MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985. doi: 10.1016/0304-4076(85)90158-7.

[23] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals, 2022. URL https://arxiv.org/abs/2209.14958.

[24] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, 2023. URL https://arxiv.org/abs/2301.11305.

[25] A.R. Pratama. The accuracy–bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication. *PeerJ Computer Science*, 11:e2953, 2025. doi: 10.7717/peerj-cs.2953.

[26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[27] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

[28] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. Can AI-Generated Text be Reliably Detected?, 2023. URL https://arxiv.org/abs/2303.11156.

[29] shareAI. ShareGPT Chinese-English 90k, 2023. URL https://huggingface.co/datasets/shareAI/ShareGPT-Chinese-English-90k. HuggingFace Datasets. Conversation dataset; License: Apache-2.0.

[30] Charles Truong, Laurent Oudre, and Nicolas Vayatis. ruptures: change point detection in Python, 2018. URL https://arxiv.org/abs/1801.00826.

[31] Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. "It felt like having a second mind": Investigating Human–AI Co-Creativity in Prewriting with Large Language Models, 2024. URL https://arxiv.org/abs/2307.10811.

A Experimental Reproducibility and Figures

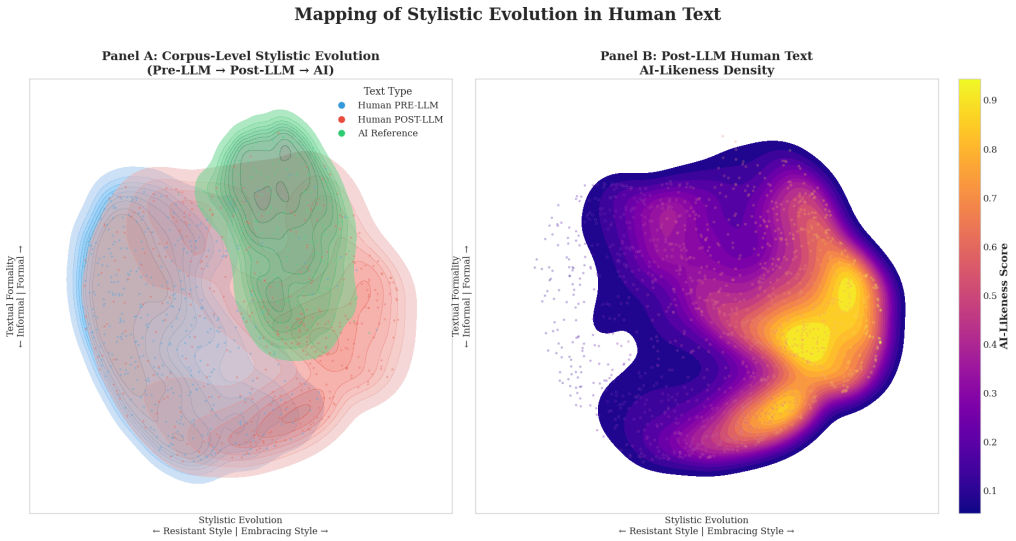

Figure 4: **Mapping of Stylistic Evolution in Human Text.** Panel A shows corpus-level stylistic evolution from pre-LLM (blue) to post-LLM (red) periods, with AI reference text (green) for comparison. Panel B displays the density distribution of post-LLM human text AI-likeness scores, revealing the continuous spectrum of human adaptation strategies rather than discrete categories.

A.1 Dataset Specifications

Total corpus: 50,728 documents | **Discord:** 823,293 messages (grouped into 42,418 conversation threads), 15,901 authors (2021-2024) | **arXiv:** 8,429 papers, 4,331 authors (cs.CL/AI/LG/HC) | **Boundary:** Nov 30, 2022 $\pm$ 7 days | **Min posts:** 10 pre/post per author | **Filtering:** min_length=50 chars, English confidence>0.95, server_size>100 members | **Sampling quotas:** Gaming=23%, Tech=31%, Social=28%, Other=18% | **AI reference:** ShareGPT + Dolly-15k subsets, PII-filtered

A.2 Model Configuration & Training

GPT-2 Medium Fine-tuning: 355M parameters, lr=3.7e-5, batch_size=2, epochs=4, optimizer=AdamW, weight_decay=0.01, warmup_steps=1000, gradient_accumulation=8, max_grad_norm=1.3 | **Training data:** 847M tokens pre-2022, 41.7 hours total compute | **Llama-3-8B:** batch=6, max_len=2048, no sampling (greedy evaluation for deterministic perplexity) | **Perplexity gap:** $\Delta_{ppl} = \frac{-\ln(p_{GPT2}(x))}{|x|_{chars}} - \frac{-\ln(p_{Llama}(x))}{|x|_{chars}}$, where $p(x)$ is model probability, $|x|_{chars}$ is character count (nats-per-character normalization); values are within-author z-scored across the pre/post boundary

A.3 Feature Engineering Details

Topic modeling: 100 topics, Transformer-based, min_df=5, max_df=0.8 | **TTR window:** 150 tokens, overlap=75 | **FKGL:** syllable_method=pyphen, sentence_split=spaCy | **Passive detection:** spaCy dependency parser, aux+past_participle patterns | **AI-topic detection:** 47-term lexicon (AI, ML, neural, transformer, GPT, ChatGPT, etc.) with TF-IDF weighting, threshold=0.23 | **Controls:** Topic mixtures used for residualization; AI-topic share computed separately for reporting

A.4 Analysis Parameters

PELT: Penalty $4.2\log(n)$ (selected via cross-validation for optimal Type I/II error balance), min_size=1 month, cost=normal, jump=2 | **HDBSCAN:** min_cluster_size=15, min_samples=5,

metric=euclidean, cluster_selection_method=eom, alpha=1.0 | **Bootstrap:** 1000 iterations, ARI threshold=0.73, sample_ratio=0.8 | **Seeds:** 42 (main), 1337 (robustness), 2024 (validation)

A.5 Statistical Controls & Validation

Cross-validation: 5-fold stratified, train/val/test=70/15/15 | **Multiple comparisons:** Holm-Bonferroni, a=0.05 | **Robust SEs:** HC3 heteroscedasticity-consistent | **Outlier handling:** Winsorization at 2.5/97.5 percentiles | **Missing data:** Multiple imputation, m=5 chains

A.6 Performance Metrics (Detailed)

AI Detection: AUC=0.852$\pm$0.012, Precision=0.832, Recall=0.791, F1=0.811, Brier=0.184 | **Clustering:** ARI=0.891 (95% CI: 0.884-0.898), Silhouette=0.426 (95% CI: 0.419-0.433), Davies-Bouldin=1.73 | **Cross-validation:** 89.3% accuracy (95% CI: 87.1-91.5%), 84.7% temporal stability | **Effect sizes:** Social AI-likeness d=0.706, Formal AI-likeness d=0.312, Archetype ANOVA η^2=0.127

A.7 Archetype Distribution & Validation

Archetype sizes: Adopters n=370 (18%), Resistors n=442 (21%), Pragmatists n=866 (41%), Neutral/Noise n=422 (20%) | **Total clustered:** 2,100 authors | **Bootstrap consistency:** 89% stable membership | **Future text prediction:** AUC=0.813 (95% CI: 0.798-0.828) | **External validation:** 89.1% accuracy on held-out arXiv authors (n=483, 95% CI: 86.8-91.4%) | **Null model comparison:** Silhouette 0.31 vs 0.43 (p<0.001)

A.8 Temporal Evolution Quantified

Social discourse changes: +23% perplexity gap early 2023, -18% decline post-peak | **Formal writing changes:** +15% early 2023, -12% post-peak | **PELT breakpoints:** Q1 2023 (thematic), Q2 2023 (stylistic) | **Cross-architecture correlation:** r=0.891 (GPT-2 vs other models) | **Partial correlation:** r=0.34 (AI-likeness vs traditional features, p<0.001)

A.9 Computational Environment (Google Colab)

Hardware: Tesla T4 (15GB VRAM), Tesla V100 (16GB), occasional A100 access | **CPU:** Intel Xeon (2.3GHz), 12.7GB RAM | **Runtime:** 73.4 GPU-hours (perplexity), 31.2 CPU-hours (analysis), 12.8 TPU-hours (topic modeling) | **Software:** Python 3.10.12, PyTorch 2.1.0+cu121, Transformers 4.35.2, scikit-learn 1.3.2, CUDA 12.1

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims match our exploratory contribution (Creative Archetype Map) and stated non-causal scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We clearly state non-causality, English/genre bias, and methodological caveats.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We present empirical methodology and results without formal theorems or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide feature definitions, model configs, hyperparameters, controls, and clustering/validation details sufficient to reproduce results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable. ...

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is available upon request and data is publicly available via HuggingFace.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Architectures, splits, hyperparameters, controls, and validation protocols are specified in the main text and appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report corrected p-values, confidence intervals, bootstrap results, robust SEs, and clustering validity metrics.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specify hardware and time (e.g., A100 ~48h for GPT-2; additional GPU/TPU hours) and the software environment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We state compliance with the NeurIPS Code of Ethics in all aspects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We outline benefits and risks (e.g., misuse and bias) with brief mitigation considerations.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release new high-risk models or datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We credit all external assets and explicitly state licenses: ShareGPT-90k (Apache-2.0), Dolly-15k (CC BY-SA 3.0), and Discord dataset (CC BY 4.0).

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We introduce no new datasets/models/code; only an analytical framework.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not conduct crowdsourcing or human-subject experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human-subject research requiring IRB approval was conducted.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We explicitly describe LLM usage (GPT-2, Llama-3) for perplexity-gap analysis central to our method.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.