Thesis Proposal: Comparing Human and Model Perception of Writing Style under Controlled Perturbations

Ewelina Księżniak
Poznań University of Business and Economics
ewelina.ksiezniak@ue.poznan.pl

Abstract

Writing style functions both as a vehicle of expression and as a marker of authorial identity. Stylometric methods enable automatic recognition of authors based on linguistic regularities, while recent advances in adversarial learning — demonstrate how data can be intentionally modified to prevent models from learning usable representations. Yet it remains unclear whether such perturbations, designed to disrupt machine learning processes, also influence human perception of style.

This thesis investigates how humans and models perceive writing style under controlled perturbations and whether manipulations that reduce algorithmic recognition likewise obscure stylistic identity for human readers. The study combines computational and behavioral approaches: constructing semantically controlled yet stylistically diverse text datasets, and conducting human evaluation experiments to compare recognition accuracy between models and readers.

The results are expected to clarify how linguistic cues contribute differently to human and algorithmic perception of style and to inform broader applications in authorship analysis, privacy-preserving text transformation, and creative expression. By situating writing style as a dimension of information quality, the research contributes to understanding how authenticity, anonymity, and expressivity interact in digital communication.

1 Introduction

Writing style is not merely a vehicle for communication but also a trace of authorial identity. Stylometric analysis can uncover authorship based on subtle linguistic cues, which has both empowering and threatening implications. In contexts such as political repression or investigative journalism, automatic authorship recognition may endanger anonymity and freedom of expression. Conversely, in creative contexts, style forms an integral part of personal and artistic expression.

Recent advances in text anonymization and adversarial methods may protect authors from machine-based identification by perturbing text representations in ways that induce models to rely on spurious stylistic cues, thereby misaligning learned representations with natural authorial style (Wang, 2023). However, little is known about whether such perturbations also alter human perception of style.

This research aims to explore how humans and models perceive writing style and to determine whether perturbations that degrade the generalization of automatically learned style representations to natural, unperturbed text likewise impair human recognition of stylistic identity. To provide a structured overview, this thesis proposal is organized as follows. Section 2 presents the related works, outlining the key research areas that ground this study, including authorship attribution, verification, and user profiling; approaches to unlearnable examples and data poisoning; and the theoretical debate on style–content disentanglement. Section 3 introduces the thesis idea, detailing the motivation, guiding hypothesis, and main research questions. Section 4 describes the proposed methodology, including dataset design, perturbation strategies. Section 5 discusses preliminary works. Section 6 discusses potential application areas where the findings may have practical impact, such as privacy protection and creative expression. Finally, the paper also outlines the risks and limitations of the study.

2 Related works

2.1 Authorship attribution, verification, and user profiling

Authorship-related tasks aim to determine who wrote a text or whether two texts share the same author. In authorship attribution, a text is assigned to one of several candidate authors based on stylis-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 831–839

tic features learned from known samples (Bevendorff et al., 2025). Authorship verification instead focuses on deciding whether two unseen texts originate from the same author, without explicit knowledge of alternatives (Bevendorff et al., 2025). A related line of work, user profiling, extends this idea by inferring demographic (Deutsch and Paraboni, 2023) or psychological traits (e.g., age, gender, personality) from linguistic and stylistic patterns.

These problems have been extensively studied within the PAN@CLEF shared tasks, which provide standardized benchmarks and evaluation frameworks for stylometric research. Early PAN systems relied on interpretable, handcrafted features such as character n-grams, function word frequencies, and syntactic patterns (Potthast et al., 2017), (Stamatatos et al., 2018). More recent editions have incorporated deep representation learning techniques such as fine-tuned transformers (Lin et al., 2025), contrastive models (Chen et al., 2023), and LLM-based approaches (Chen et al., 2025).

2.2 Unlearnable examples and data poisoning

In parallel with advances in classification and generation tasks, a line of research has focused on unlearnable examples—training data intentionally modified so that machine learning models fail to acquire useful representations of the underlying signal. In this setting, perturbations are optimized to directly obstruct the learning process, often preventing convergence or causing models to perform poorly even on the perturbed training data itself (Huang et al., 2021). Such approaches aim to make the target signal effectively unlearnable, rather than merely difficult to generalize.

Closely related, but conceptually distinct, are data poisoning approaches. Instead of blocking learning altogether, poisoning methods inject carefully designed artifacts into the training data that cause models to learn misleading or spurious correlations (Cinà et al., 2023). As a result, models may achieve high training performance while internalizing representations that do not align with the true underlying structure of the data and fail to generalize beyond the poisoned distribution.

A representative example of this latter paradigm is Glaze, which protects artists from style imitation by subtly modifying images at training time. Rather than preventing models from learning altogether, Glaze induces generative models to internalize a distorted version of an artist's style that does not transfer to unperturbed artworks, effectively poisoning the learned style representation (Shan et al., 2023).

Transferring these ideas to text remains substantially more challenging. Language is discrete and semantically fragile: even small perturbations, such as word substitutions or syntactic rearrangements, can alter meaning, tone, or grammaticality and may become perceptually salient to human readers. Designing perturbations that either block learning or poison stylistic representations while preserving semantic content and human-perceived style therefore remains an open technical and conceptual problem in textual authorship analysis.

2.3 Style vs. content

A major theoretical issue complicating this challenge is the (in)separability of style and content. Early work in textual style transfer (TST) assumed that these two dimensions could be cleanly disentangled—that one could modify a text's style while preserving its semantic content (Shen et al., 2017), (Fu et al., 2018). However, as Jafaritazehjani (2023) demonstrated, this assumption oversimplifies the problem: style and content are inherently entangled, and the degree of this entanglement varies across stylistic domains. By analyzing the latent representations of several text style transfer models, she showed that the extent of this entanglement depends on the stylistic dimension: sentiment is closely tied to content, whereas formality can be more readily separated from it (Jafaritazehjani, 2023).

Conceptually, style can be viewed as the way content is expressed rather than an independent layer applied to it. Consequently, attempts to modify or anonymize style must recognize that complete separation from content is not achievable in practice.

2.4 Human perception of style

Writing style is a composite construct encompassing multiple linguistic dimensions. It manifests through lexical preferences (e.g., word choice, frequency of function words), syntactic organization (e.g., sentence length, clause structure, word order), semantic–pragmatic features (e.g., topic framing, tone, and formality) and may also include rhythmic and coherence-related cues, as well as many other features that are informative yet difficult to operationalize and quantify.

An important question in stylistic research is which linguistic cues humans use to judge whether

two texts share the same author or stylistic pattern. Early psychological work provides a foundation for studying stylistic perception. Gardner (1971) investigated how individuals recognize and reproduce distinctive modes of written expression. In his experiments, participants completed short stories written in contrasting stylistic registers ("fairy-tale" vs. "jivy") and generally extended each text in a way consistent with its tone and rhythm. These findings suggest that people can internalize stylistic regularities and apply them in production even without explicit awareness of the underlying rules. Gardner further argued that sensitivity to style develops gradually, reflecting a shift from attention to surface linguistic cues toward an implicit grasp of expressive intent.

Linguistic studies on formality provide a complementary perspective on style perception. Pavlick and Tetreault (2016) conducted two experiments in which participants rated sentence formality and rewrote informal sentences in a formal register. Results were consistent across both experiments, indicating a shared cognitive representation of formal style among participants. Research on authorship attribution further demonstrates the complexity of stylistic perception. Rexha et al. (2018) showed that human participants can identify or group texts by author above chance level, but their accuracy declines sharply when overt lexical markers are removed or topical hints are neutralized. Taken together, these studies suggest that humans perceive writing style as a coherent construct that integrates multiple linguistic cues and supports consistent judgments across diverse stylistic domains.

3 Thesis idea

The review above highlights three intertwined challenges at the intersection of stylometry, adversarial learning, and human perception. First, while unlearnable and poisoning-based approaches in computer vision often rely on subtle, human-imperceptible perturbations, their effectiveness lies not in blocking learning but in inducing misaligned representations. Extending this paradigm to text is nontrivial, as language is discrete and semantically coupled, meaning that even minimal changes may affect readability and meaning. Second, style and content cannot be cleanly separated—stylistic expression inherently shapes how content is conveyed. Third, human perception of style integrates lexical, structural, and affective cues in a holistic manner, which may diverge from the statistical representations used by neural models. Together, these findings motivate a shift in focus: instead of pursuing fully imperceptible perturbations, this research aims to systematically analyze how manipulations along different stylistic dimensions affect recognizability for humans and models. Crucially, the focus is not on rendering style unlearnable, but on understanding how certain perturbations may redirect model learning toward spurious stylistic cues that fail to generalize to natural, unperturbed text. Accordingly, this work focuses on the perception of stylistic variation in texts generated by large language models, motivated by the increasing prevalence of writing practices in which authorship is no longer defined by direct human composition, but by the selection, editing, and adaptation of model-generated text.

Research gap: Despite rapid advances in both stylometric modeling and research on unlearnable examples and data poisoning, there is currently no systematic understanding of how algorithmic and human perceptions of writing style diverge. Most stylometric research emphasizes improving automatic authorship recognition, whereas adversarial and privacy-oriented studies focus on obscuring or distorting author identity representations learned by such models —both generally adopting a model-centered view of style. In contrast, little is known about how these same perturbations influence human perception of style, coherence, or authorial voice. Furthermore, prior work rarely disentangles the relative roles of different stylistic dimensions—lexical, syntactic, and rhythmic—in shaping recognizability across humans and models. As a result, there is little empirical evidence on which aspects of style remain robust or become fragile under adversarial or privacy-preserving transformations. This research addresses this gap by directly comparing human and model sensitivity to targeted stylistic manipulations, thereby bridging computational and perceptual perspectives on textual style.

Hypothesis: There exist stylistic perturbations that substantially degrade models' ability to generalize from perturbed training data to natural, unperturbed instances of the same stylistic category, while leaving human judgments statistically indistinguishable from baseline.

Research questions

- ***RQ1:*** *How sensitive are humans and*

transformer-based classifiers to controlled manipulation of specific surface-level stylistic cues?

Addressed through the controlled dataset construction, model-based experiments, and human evaluation described in Section 4. This question examines whether controlled manipulations of low-level stylistic dimensions differentially affect human and model recognition of high-level stylistic categories, and whether lexical–semantic or syntactic–distributional cues play a dominant role in each case.

- ***RQ2:*** *Do the same stylistic perturbations produce asymmetric effects on human and model style recognition accuracy?*

  Addressed through the same experimental framework described in Section 4. This question evaluates whether identical stylistic perturbations lead to asymmetric changes in recognition accuracy for humans and transformer-based classifiers, and whether perturbations that substantially degrade model generalization leave human style judgments statistically indistinguishable from baseline.

Main contributions: This work makes the following contributions:

- It provides a diagnostic comparison of human and transformer-based model sensitivity to controlled stylistic perturbations, establishing an empirical framework for analyzing how writing style is differently perceived, represented, and disrupted across cognitive and computational systems. Rather than proposing an operational anonymization method, the study offers foundational insight into the conditions under which stylistic cues remain robust for human readers while becoming fragile for automated models.

- It introduces controlled, semantically equivalent datasets with systematically varied stylistic profiles, enabling fine-grained analysis of stylistic perturbations across both model-based and human evaluation settings. These datasets are designed to support reproducible investigation of stylistic robustness, generalization failure, and perception under controlled manipulation.

Although the proposed perturbations are not themselves adversarial or privacy-preserving transformations, they serve as controlled proxies for studying which stylistic dimensions can be modified without affecting human-perceived style while disrupting model representations. Identifying such dimensions is a necessary prerequisite for the design and evaluation of future anonymization or style-obfuscation methods.

4 Proposed methodology

The study adopts a multi-stage methodology combining controlled text generation, computational modeling, and human evaluation. In this work, writing style is not treated as a fully quantifiable construct, but is operationalized for experimental purposes through a limited set of measurable linguistic features. Each selected feature corresponds to a predefined stylistic dimension and serves as a proxy for specific aspects of stylistic variation.

4.1 Controlled dataset construction

The first stage of the study involves constructing datasets composed of semantically equivalent texts that instantiate distinct stylistic profiles. Two texts x and x_1 are considered semantically equivalent if they satisfy all of the following conditions:

- bidirectional NLI predicts mutual entailment with probability ≥ 0.9 in both directions. Bidirectional NLI will be computed using a pretrained NLI model by evaluating both sentence orderings.

- Question-answering–based evaluation yields an answer overlap of at least 80% across predefined question sets.

- Manual validation on a randomly sampled subset yields substantial inter-annotator agreement (Cohen's $\kappa \geq 0.6$).

To ensure semantic equivalence and reduce topic leakage, texts will be generated as LLM-based paraphrases. The source texts used for paraphrasing will consist of short narrative fragments from publicly available, human-authored literary works (e.g., fairy tales from Wikisource), selected to ensure stylistic flexibility and to avoid strong topic–register mismatches that could make certain stylistic realizations unnatural. Semantic equivalence will be verified using the criteria defined above; any generated paraphrase pair that fails to

meet these requirements will be discarded and regenerated rather than included in the dataset.

Stylistic variation is then introduced by systematically controlling a predefined set of stylistic dimensions while keeping propositional content fixed. Each generated text instantiates a *stylistic profile*, defined operationally as a concrete assignment of intensity levels to these dimensions, corresponding to a specific point in the resulting stylistic feature space. In this work, stylistic dimensions are organized into two levels of abstraction, reflecting differences in semantic salience and perceptual accessibility.

- First, **high-level stylistic dimensions** capture stylistic contrasts that are readily accessible to human readers and commonly used in explicit style judgments. We restrict this level to a small set of categories, including *formal vs. informal writing* and *imitation of selected literary authors*.
- Second, **low-level surface dimensions** correspond to measurable linguistic cues that have limited impact on propositional meaning but are highly informative for computational models and plausibly processed less consciously by human readers. We focus on a restricted set of such cues, including *function-word frequency*, *sentence length*, *punctuation patterns* and *word order inversions* restricted to syntactic alternations that are likely to preserve surface naturalness and fluency (e.g., infinitival verb–auxiliary alternations).

The selected high- and low-level stylistic dimensions were chosen based on prior stylometric research and on their suitability for controlled manipulation under semantic equivalence constraints. In particular, we focus on dimensions that are both informative for stylistic modeling and sufficiently weakly coupled with propositional content to allow paraphrasing-based variation without altering meaning. To mitigate generation artifacts and prompt-induced confounds, all stylistic variants will be generated using a shared base prompt.

For each low-level stylistic dimension, we define discrete *intensity levels* based on the empirical distribution of the underlying linguistic feature f_i in the dataset. Specifically, low, medium, and high intensity levels correspond to the 25th, 50th, and 75th percentiles of the observed feature distribution, respectively. Variations in intensity levels are treated as controlled stylistic perturbations relative to a baseline (medium-intensity) condition.

Validation is performed using dimension-specific criteria that are aligned with this operationalization. For distributional dimensions (e.g., sentence length, function-word frequency), each variant is validated by verifying that its measured feature values fall within the target percentile range for the intended intensity level. For structurally defined dimensions (e.g., word order inversions), intensity is operationalized in terms of the presence and relative frequency of the targeted syntactic patterns, which are verified using rule-based or parser-based checks. Samples that fail to meet the corresponding validation criteria are excluded.

Dataset sizes are balanced across stylistic conditions, with approximately 1,000–2,000 base instances per high-level stylistic contrast, each realized under all defined low-level intensity configurations.

4.2 Model experiments

In this stage, we evaluate transformer-based classifiers under controlled stylistic perturbations applied to low-level stylistic dimensions. We will evaluate a minimum of five transformer architectures that achieve competitive performance on widely used text representation and classification benchmarks (e.g., MTEB (Muennighoff et al., 2023)); illustrative candidates include RoBERTa-base (Liu et al., 2019) and DeBERTa-v3-base (He et al., 2021) models.

Models will be trained to distinguish between predefined high-level stylistic categories, such as formal vs. informal writing or imitation of Author A vs. Author B. As a baseline condition, models will be trained and evaluated on data in which all low-level stylistic dimensions are realized at their baseline (medium-intensity) levels. Performance is measured using accuracy and area under the ROC curve (AUC) on a held-out test set.

In perturbed conditions, models will be trained on data in which one or more low-level stylistic dimensions deviate from baseline intensity levels, while evaluation is always performed on unperturbed (baseline) test data. Differences between baseline and perturbed conditions are quantified using: (i) absolute and relative changes in accuracy, (ii) changes in AUC, (iii) the generalization gap between performance on perturbed training data and unperturbed test data, and (iv) model calibration, assessed via expected calibration error (ECE) or

Brier score. All experiments will be conducted using a minimum of 5 random seeds. Reported results reflect mean performance across seeds, and statistical comparisons between baseline and perturbed conditions will be performed across seed-level estimates.

4.3 Human evaluation

In this stage, we will conduct survey-based behavioral experiments with human participants to examine how stylistic distinctions are perceived under controlled conditions. We plan to recruit approximately 180 participants who are native speakers of the language in which the dataset is generated. This target sample size was selected to ensure sufficient statistical power for the planned equivalence testing. Participants will be presented with generated text samples instantiating different stylistic profiles and will be provided with a small set of reference examples for each high-level stylistic category to familiarize them with the target distinctions.

The evaluation adopts a within-subject design. Each participant will complete approximately 80 trials in total, comprising an equal number of baseline and perturbed items. On each trial, participants will assign a text fragment to one of the predefined high-level stylistic categories (e.g., formal vs. informal writing or Author A vs. Author B). Trial order will be fully randomized, such that baseline and perturbed items are interleaved, preventing participants from inferring condition structure or relying on task order cues.

Baseline items correspond to texts in which all low-level stylistic dimensions are realized at their baseline (medium-intensity) levels and serve to establish individual performance in the absence of stylistic perturbations. Perturbed items follow the same classification procedure but incorporate controlled stylistic perturbations introduced through systematic variations in the intensity levels of selected low-level stylistic dimensions. To avoid content-based confounds, participants will never be exposed to multiple stylistic variants of the same underlying content.

The set of stylistic perturbations presented to human participants will be informed by the model experiments and will focus on configurations that produced the largest performance differences relative to the baseline condition. Human classification performance under perturbed conditions will be compared to baseline performance to assess whether stylistic perturbations that substantially degrade model generalization also affect human recognition of high-level stylistic categories.

In line with the central hypothesis of this study, human performance will be evaluated using Two One-Sided Tests (TOST) (Lakens, 2017) to determine whether recognition accuracy under stylistic perturbations is statistically equivalent to baseline performance within a predefined equivalence margin of 5 percentage points. The planned sample size and number of trials are selected to provide sufficient power to detect or confirm equivalence for differences in accuracy smaller than the specified equivalence threshold.

In addition to categorical style judgments, response times and self-reported confidence scores will be collected for each decision and analyzed as secondary measures of potential changes in cognitive effort or uncertainty. Participants will also rate the perceived naturalness of each text, which will be used as a control measure to monitor potential fluency degradation introduced by stylistic perturbations.

The study protocol will follow standard ethical guidelines for human-subject research, and informed consent will be obtained from all participants.

5 Preliminary works

To assess the feasibility of the proposed experimental framework and to validate the core assumptions concerning model sensitivity to controlled stylistic perturbations, we conducted a set of preliminary experiments combining controlled dataset construction, model-based classification, and a small-scale human evaluation. We constructed a controlled dataset consisting of 1,500 semantically equivalent sentences in polish language, generated as paraphrases of short narrative fragments written by Ignacy Krasicki and Hans Christian Andersen, and realized in two high-level stylistic variants corresponding to the styles of Henryk Sienkiewicz and Adam Mickiewicz. The paraphrases were generated using the GPT-4-turbo (Hurst et al., 2024) language model, guided by a strictly constrained prompt that required the production of two stylistic variants while enforcing exact preservation of propositional content and prohibiting any semantic additions, omissions, or factual modifications. A randomly sampled subset of generated pairs was additionally inspected manually to verify that semantic equivalence was preserved. Importantly,

the attribution to "style" refers to stylistic imitation, not authorship attribution to the original texts.

Each stylistic variant was further parameterized along a small set of explicitly controlled stylistic dimensions. In addition to the high-level stylistic contrast introduced above, we applied controlled perturbations to two predefined low-level stylistic dimensions that are known to be informative for stylometric models but are typically less accessible to conscious human judgment: (i) function-word distribution and frequency, and (ii) syntactic word order patterns, operationalized through fixed ordering constraints on selected dependency relations (noun–adjective, noun–possessive modifier, verb–adverb).

As a baseline condition, a binary classifier based on the mDeBERTa-v3-base (He et al., 2021) architecture was trained to discriminate between the two high-level stylistic variants. Under the unperturbed (baseline) condition, the model achieved an accuracy of 97.87%, when trained for three epochs with a learning rate of 2×10^{-5}, a training batch size of 16 and an evaluation batch size of 32, weight decay of 0.01, and model selection based on validation loss. To assess the effect of controlled perturbations of low-level stylistic dimensions on stylistic learnability, we evaluated two perturbed conditions: function-word reduction implemented via constrained LLM-based paraphrasing, and syntactic order normalization enforced through rule-based dependency parsing, which imposed fixed ordering constraints such that adjectival or possessive pronominal modifiers followed the noun and adverbs followed the verb. A third setting combined both perturbations. Under the three perturbation settings—function-word reduction, syntactic order normalization, and their combination—the retrained classifier achieved accuracies of 93,97 %, 97,87%, and 89,24 %, respectively. For all classification settings, the input length was capped at 200 tokens, and identical model architectures and training hyperparameters were used for both the baseline and perturbed training conditions.

To examine whether the same perturbations affect human perception of style, we conducted a small-scale human evaluation. Participants (36 native Polish-speaking university students) completed a short questionnaire comprising 15 multiple-choice items, in which they were shown short text samples and asked to decide whether each sentence was stylistically closer to Sienkiewicz or Mickiewicz. Before the task, participants were given reference samples for both styles (up to 25 sentences), which remained available during the questionnaire. The questionnaire included 10 unperturbed items (baseline), 2 items with function-word reduction, and 3 items with combined function-word reduction and syntactic inversion. Mean accuracy was 76.11% on unperturbed sentences, 65.3% under function-word reduction, and 67.6% under the combined perturbation. This human evaluation was conducted solely as a feasibility and pilot study, intended to validate the practical viability of the proposed experimental framework and to obtain preliminary insight into whether humans are able to recognize high-level stylistic distinctions under controlled perturbations. Given the small sample size and limited number of items per condition, the results are not intended to support inferential claims or generalization, but rather to inform the design choices of the full-scale human evaluation described in Section 4.3. All participants provided informed consent prior to taking part in the study.

6 Potential applications

While the primary contribution of this work is conceptual and diagnostic—to uncover how humans and models differ in perceiving writing style — the findings may also inform several practical domains where style recognition and manipulation are consequential. Understanding the specific cues that drive human versus algorithmic sensitivity to stylistic variation could reveal which dimensions of expression are most robust, most fragile, or most easily obfuscated.

First, in the domain of privacy and security, more accurate knowledge of which stylistic cues remain recognizable to humans but invisible to models (and vice versa) could guide the development of effective text anonymization strategies. Techniques derived from this research could help protect authors in politically sensitive or high-risk environments—such as investigative journalists, whistleblowers, or activists operating under authoritarian regimes.

Second, the findings may contribute to the protection of creative expression. Authors, poets, and screenwriters often rely on distinctive stylistic signatures that can be imitated or appropriated by large language models trained on public text. While this thesis does not aim to deliver a complete technical framework for authorship protection, it may serve as an initial step toward understanding

whether style-level perturbations—analogous to Glaze in visual art — can be meaningfully applied to text. If certain stylistic subspaces can be selectively masked or randomized without degrading meaning or altering human-perceived style, this could open pathways for safeguarding literary or journalistic voice in the era of generative models.

Limitations

Several limitations and potential risks should be acknowledged when interpreting the scope and implications of this project. First, the study relies primarily on synthetically generated texts rather than naturally authored materials. While controlled generation enables precise manipulation of stylistic dimensions, it also limits ecological validity, as the stylistic signatures of large language models may differ systematically from those of human writers and introduce generation-specific artifacts that influence both model and human perception. At the same time, this limitation reflects a deliberate scoping decision. Contemporary writing practices increasingly involve a hybrid mode of authorship in which texts are produced through interaction with large language models and subsequently edited, curated, or adapted by human authors. In this emerging setting, stylistic identity does not correspond to a purely human writing style, but to a composite, model-mediated form of expression. From this perspective, investigating the perception and perturbability of style in LLM-generated text remains informative for the scope of the present study, as it captures a growing and practically relevant class of stylistic phenomena, even if these do not fully align with traditional notions of individual authorial style.

Second, a potential risk concerns participant fatigue in survey-based behavioral experiments, particularly given the within-subject design and the number of trials required to compare baseline and perturbed conditions. Fatigue may affect attention, response times, and decision consistency, thereby introducing noise into the behavioral data. To mitigate this risk, the experimental design limits the total number of trials per participant, randomizes trial order, and includes short instructions and practice items to stabilize task understanding. In addition, response times and confidence ratings are monitored to identify patterns indicative of reduced engagement, and participants exhibiting consistently implausible response behavior can be excluded from analysis.

A further challenge concerns generalization. Results derived from small, controlled datasets—constructed under fixed stylistic profiles and limited topical diversity—may not extend to heterogeneous real-world corpora. In practice, natural texts exhibit far more intricate overlaps between content, genre, and authorial intent than those captured in laboratory-style experiments. The findings of this thesis should therefore be viewed as diagnostic rather than predictive: they reveal structural tendencies rather than definitive behavioral laws.

References

Janek Bevendorff, Matti Wiegmann, Emmelie Richter, Martin Potthast, and Benno Stein. 2025. The two paradigms of llm detection: Authorship attribution vs authorship verification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3762–3787.

Dongjie Chen, Jijie Li, and Haoliang Qi. 2025. Llama-3 with 4-bit quantization and ia^3 tuning for multi-author writing style analysis. In *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece. CEUR-WS.org.

Haoyang Chen, Zhongyuan Han, Zengyao Li, and Yong Han. 2023. A writing style embedding based on contrastive learning for multi-author writing style analysis. In *CLEF (Working Notes)*, pages 2562–2567.

Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. 2023. Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Computing Surveys*, 55(13s):1–39.

Caio Deutsch and Ivandré Paraboni. 2023. Authorship attribution using author profiling classifiers. *Natural Language Engineering*, 29(1):110–137.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Howard Gardner. 1971. The development of sensitivity to artistic styles. *The Journal of Aesthetics and Art Criticism*, 29(4):515–527.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. 2021. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv:2101.04898*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Somayeh Jafaritazehjani. 2023. *Towards an Improved Understanding of the Concept of Style and Its Implications for Textual Style Transfer*. Ph.D. thesis, Technological University Dublin, Dublin, Ireland. Ph.D. Thesis.

Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362.

Xiaocan Lin, Chang Liu, Xianbing Duan, and Zhongyuan Han. 2025. Team wqdatstyle change detection in multi-author writing: A deep learning approach based on deberta. In *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece. CEUR-WS.org.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Martin Potthast, Francisco Rangel, Michael Tschuggnall, Efstathios Stamatatos, Paolo Rosso, and Benno Stein. 2017. Overview of pan'17: author identification, author profiling, and author obfuscation. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 275–290. Springer.

Andi Rexha, Mark Kröll, Hermann Ziak, and Roman Kern. 2018. Authorship identification of documents with high content similarity. *Scientometrics*, 115(1):223–237.

Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. 2023. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Efstathios Stamatatos, Francisco Rangel, Michael Tschuggnall, Benno Stein, Mike Kestemont, Paolo Rosso, and Martin Potthast. 2018. Overview of pan 2018: Author identification, author profiling, and author obfuscation. In *International conference of the cross-language evaluation forum for european languages*, pages 267–285. Springer.

Haining Wang. 2023. Defending against authorship identification attacks. *arXiv preprint arXiv:2310.01568*.