Differentially-Private Text Rewriting reshapes Linguistic Style

Stefan Arnold
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany
stefan.st.arnold@fau.de

Abstract

Differential Privacy (DP) for text matured from disjointed word-level substitutions to contiguous sentence-level rewriting by leveraging the generative capacity of language models. While this form of text privatization is best suited for balancing formal privacy guarantees with grammatical coherence, its impact on the register identity of text remains largely unexplored. By conducting a multidimensional stylistic profiling of differentially-private rewriting, we demonstrate that the cost of privacy extends far beyond lexical variation. Specifically, we find that rewriting under privacy constraints induces a systematic functional mutation of the text's communicative signature. This shift is characterized by the severe attrition of interactive markers, contextual references, and complex subordination. By comparing autoregressive paraphrasing against bidirectional substitution across a spectrum of privacy budgets, we observe that both architectures force convergence toward a *non*-involved and *non*-persuasive register. This register-blind sanitization effectively preserves semantic content but structurally homogenizes the nuanced stylistic markers that define human-authored discourse.

1 Introduction

Language Models (LMs) (Radford et al., 2018; Brown et al., 2020) have demonstrated a remarkable capacity to capture and reproduce properties of their training data. While this emergent capability underpins their success across a wide range of natural language processing tasks, it has also been shown to expose sensitive information. Prior work has documented that LMs can reveal authorship (Song and Shmatikov, 2019) and disclose personally identifiable information (Carlini et al., 2021; Nasr et al., 2023), largely as a consequence of unintended memorization (Carlini et al., 2019). This raised substantial privacy concerns and led to the development of mechanisms for text privatization.

To mitigate the potential risk of information leakage, researchers have adopted *Differential Privacy* (DP) (Dwork et al., 2006). Originally developed for structured databases, DP formalizes privacy through a rigorous notion of indistinguishability obtained by injecting noise so that the inclusion or exclusion of any individual sample does not substantially affect the outcome distribution.

Applying DP to text presents unique challenges due to the discrete and highly structured nature of language. Feyisetan et al. (2020) addressed this challenge at the word-level by perturbing words in the embedding space (Mikolov et al., 2013) with semantically proximate substitutions. By framing this process as a randomized response mechanism (Warner, 1965), this approach provides plausible deniability (Bindschaedler et al., 2017). However, by operating on isolated lexical units, this approach tends to degrade grammatical coherence and produce disjointed sentences (Arnold et al., 2023).

To overcome this limitation, the field has transitioned from *disjointed* word-level perturbations toward more *fluent* sentence-level rewriting by leveraging the generative capacity of LMs (Mattern et al., 2022; Igamberdiev and Habernal, 2023; Meisenbacher et al., 2024). This paradigm shift toward rewriting entire sequences rather than perturbing words in isolation allows for the generation of privatized text that maintains grammatical integrity while satisfying privacy constraints.

While DP rewriting powered by LMs retains semantic content, its impact on *linguistic style* remains insufficiently understood (Çano and Habernal, 2025). This limitation is critical. Since linguistic variation encodes communicative intent, stylistic features determine how meaning is conveyed within a social and functional context (Halliday and Matthiessen, 2013). By shaping distinctions such as a personal opinion and a formal discourse, style governs how texts are perceived (Biber, 1991). Consequently, when rewriting text under privacy

Proceedings of the Seventh Workshop on Privacy in Natural Language Processing, pages 96–106

constraints, DP mechanisms may not only alter stylistic fingerprints but fundamentally distort the communicative intent.

Contribution. We conduct a profiling of stylistic variation of texts obtained from rewriting under constraints of DP. By contrasting autoregressive and bidirectional architectures, we isolate to which extent stylistic distortion arises from the interaction between privacy budget and model design. Specifically, we employ autoregressive rewriting from a fine-tuned LM (Mattern et al., 2022) and bidirectional rewriting from a pre-trained LM (Meisenbacher et al., 2024), and evaluate them for 67 lexico-grammatical features derived from Biber (1991) using the *Corpus of Online Registers of English* (CORE) (Egbert et al., 2015). This corpus comprises 47 diverse registers that range from personal blogs to formal documents. By tracing the degradation of linguistic style down to grammatical features, we can dissect the extent to which DP distorts communicative intent at multiple levels of abstraction. We make three primary contributions showing that DP rewriting does not merely introduce noise at the level of lexemes, but systematically reshapes linguistic style:

1. We establish the stylistic preservation by measuring the distributional distance between human and rewritten style features, as a function of the privacy budget. Our results indicate that bidirectional rewriting captures and reproduces human style more effectively than autoregressive paraphrasing across all privacy budgets. Notably, while the style of the bidirectional approach converges toward the human style as privacy guarantees are relaxed, the autoregressive approach reaches a distinct convergence plateau, indicating persistent stylistic bias. We attribute this to stylistic artifacts stemming from its training history on paraphrasing corpora with limited styles.

2. We pinpoint deviations in feature distributions to show that bidirectional rewriting generally maintains a closer proximity to human baselines than autoregressive rewriting. Both stylometric profiles are characterized by diminished interactive and contextual markers such as personal pronouns and place/time adverbials along with simplified subordinator structure. Our style profiling reveals a sanitization process that under-represents causative (e.g., *because*) and concessive (e.g., *although*) structures but over-generates temporal (e.g., *since*) and adversative subordination (e.g., *whereas*) to maintain a veneer of logical coherence while losing diverse dependencies.

3. We project these distributions of features onto the multidimensional framework proposed by Biber (1991) to interpret the functional transformation of communicative intent. Our analysis reveals that while bidirectional rewriting remains more faithful to the communicative intent compared to autoregressive rewriting, both architectures trigger a fundamental functional shift that often transforms involved and persuasive discourse into sterile and informational reports. This characterizes the register-blind nature of current DP mechanisms which effectively preserve topic and content but fall short of maintaining stylistic identity.

2 Stylometric Analysis

To isolate and measure the stylistic shift introduced by DP rewriting, we contrast the paraphrases of two diametrical model architectures against a human-authored baseline. By leveraging a high-dimensional stylometric framework, we move beyond superficial string-similarity metrics to quantify the functional and communicative distortions that occur under varying privacy constraints.

2.1 Model Selection and Architecture

To ensure a controlled comparison of potential architectural impacts under DP constraints, we selected two foundational paradigms representing the primary approaches to sentence-level rewriting: DP-PARAPHRASE (Mattern et al., 2022) and DP-MLM (Meisenbacher et al., 2024). We implement DP-PARAPHRASE utilizing a `GPT-2` (Radford et al., 2019) backbone *fine-tuned* for paraphrasing (Dolan et al., 2004). This represents an autoregressive approach where privatized text is generated sequentially. DP-MLM employs a *pre-trained* `RoBERTa` (Liu et al., 2019) architecture. This represents a bidirectional approach where privatized text is generated by substituting tokens within the full context of the surrounding sentence. We evaluate both approaches across a privacy budget spectrum of $\varepsilon \in \{10, 25, 50, 100, 250\}$, aligning with the benchmarks established by Meisenbacher et al. (2024) to ensure comparability.

Table 1: Examples from the CORE corpus (Egbert et al., 2015), categorized according to four *archetypal* dimensions presented by Biber (1991). Scores denote the aggregated factor values calculated from the summation of standardized feature frequencies relative to the corpus mean. Representative features are *italicized*.

Dimension	Pole	Example	Score
Focus	Involved	Do *you think not* going to University is a barrier in life, and *you are* unable to be successful? *I used to think this* was a huge barrier. *I didn't* go because *I didn't* have enough money, and *I really don't want to* take student loans, because *I see* [...]	+41.12
	Informational	The *weekly programme* of this *ongoing series* addressing *cutting-edge research* will include *talks* by *invited guest speakers* as well as *original papers* from those teaching across *heritage studies* at *UCL*. [...]	−19.37
Discourse	Narrative	An heiress *found* dead in *her* 70 million home in London *had been* due to enter a rehabilitation clinic in California with *her* husband just weeks before, *her* mother *said*. Nancy Kemeny *said* the family *had maintained* "high hopes" that [...]	+17.25
	Expository	Here at Taumarunui High School, opportunities for learning abound with a range of subjects to suit even the most exacting study requirements. Courses and subjects offered are as diverse as farming and painting, computing and music, [...]	−3.55
References	Situational	[...] The After care report published by Roger Morgan looked at views from 308 care leavers both who had *recently* left care and those *still* in care. The main messages *this year* were that *nearly* half felt they left *too early* and were *not* prepared *well enough*.	+11.82
	Elaborated	The *aim* of this Special *Issue* is to further our *understanding* of the *manner* in *which* [...] impact [...]. Our *intent* is to stimulate critical *debate* and *analyses*.	−13.25
Argumentation	Persuasive	[...] You *may think* you *know how to do* everything, but *if* you're a real man, then you *can actually admit* that you probably don't *know how to change* a light bulb. Since every man *should know how to do* this, here are things you *should be* familiar with in order *to be* a man.	+18.32
	Neutral	[...] Quebec is the largest province in Canada by area and borders Ontario, New Brunswick and Newfoundland [...]. The territory of Quebec represents [...]	−4.33

Crucially, both approaches for rewriting rely on temperate sampling (Holtzman et al., 2020) to select from the vocabulary distribution. By setting the temperature as a function of the privacy budget, this process operates as an instantiation of the exponential mechanism (McSherry and Talwar, 2007). This shared mathematical formulation establishes rigorous experimental control, ensuring that any observed stylistic differences stem from the generative architecture and training history rather than being confounded by discrepancies in the underlying privacy formalization.

We note that we deliberately omit mechanisms at word-level (e.g., Chen et al., 2023; Tian et al., 2026) because they frequently fail to preserve basic grammatical coherence (Arnold et al., 2023). This limitation confounds stylometric analysis as it is linguistically invalid to meaningfully measure the functional style of a disjointed string.

2.2 Dataset and Feature Extraction

We evaluate our selected rewriting mechanisms on a subset of 9,691 documents from the *Corpus of Online Registers of English* (CORE) (Egbert et al., 2015). Table 1 provides concrete examples from the CORE corpus. Ranging from personal blogs or discussion forums to legal terms and technical reports, this corpus is uniquely suited for stylometric stress testing due to its unparalleled register diversity found on the open web.

To ensure linguistic interpretability, we eschew opaque style embeddings (Rivera-Soto et al., 2021; Patel et al., 2023) in favor of a curated set of lexicogrammatical features (Biber, 1995). While style embeddings proved effective for authorship attribution, they aggregate stylistic variance into often uninterpretable high-dimensional representations. Such representations obscure the specific grammatical shifts induced by DP rewriting, making it im-

possible to determine whether a shift in the embedding space stems from lexical distortion, syntactic simplification, or a loss of grammatical markers.

In contrast, the framework compiled by Biber (1995) enables the precise quantification of 67 distinct linguistic features, organized into 16 grammatical and functional categories. This comprehensive set spans a wide range of linguistic aspects, including *tense and aspect markers*, *place and time adverbials*, *nominal forms*, *passive voice*, *modality words*, *negation types*, and *subordination features*. Table 2, deferred to the Appendix, provides a full inventory of these grammatical features and functional categories. By tracking the frequencies of these grammatical constructs, we can pinpoint exactly what private rewriting fails to preserve, decomposing an abstract measure of stylistic drift into a concrete inventory of grammatical shifts.

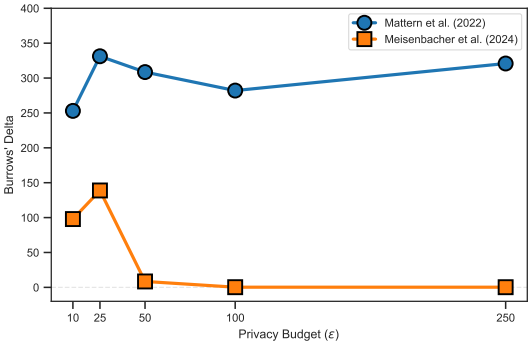

Figure 1: Stylistic deviation of privatized text compared to the human-authored text, measured using *Burrows' Delta*. Both approaches exhibit erratic behavior under tight privacy. DP-MLM consistently demonstrates significantly lower deviation across the entire privacy spectrum and converges toward the human style. DP-PARAPHRASE reaches a distinct convergence plateau, failing to recover the human style.

3 Style Variance

Aimed at dissecting the impact of private rewriting on linguistic style, we adopt a hierarchical analysis across three levels of granularity.

3.1 Stylistic Fidelity

To characterize the stylistic shift induced by DP rewriting, we quantify the divergence between original and privatized text using *Burrows' Delta* (Burrows, 2002). This measure captures the distance between standardized relative frequencies of linguistic items and serves as a proxy for the stylistic fingerprint of a text. A low value indicates close adherence to style properties, whereas a high value reflects increased stylistic deviation.

Figure 1 presents the relationship between the privacy budget and stylistic preservation for both autoregressive and bidirectional rewriting, revealing a significant architectural disparity. Across the entire spectrum of privacy budgets, bidirectional privatization enables DP-MLM to better retain distributional properties than the autoregressive privatization process. This is reflected by consistently lower delta values for DP-MLM compared to DP-PARAPHRASE, indicating a systematically higher degree of stylistic preservation.

Contrary to a monotonic decay in stylistic distortion given the privacy budgets, we observe that the relationship between the privacy budget and stylistic distance is not strictly monotonic. In the regime of tight privacy, both techniques exhibit fluctuations, showing a slight rise in delta followed by a subsequent dip. This pattern reflects the interaction between the noise injection of the exponential mechanism and the sampling distribution, where extremely high noise levels can occasionally lead to erratic stylistic texts. As privacy constraints are progressively relaxed, both architectures diverge markedly in their convergence behavior. While DP-MLM demonstrates a trajectory toward the original stylistic signature, DP-PARAPHRASE reaches a distinct plateau and fails to recover the original style even at high privacy budgets.

The persistent divergence in stylistic convergence exhibited by DP-PARAPHRASE points to an underlying architectural limitation. A plausible explanation for this lack of convergence resides in its training history. Specifically, DP-PARAPHRASE relies on a `GPT-2` backbone (Radford et al., 2019) fine-tuned for paraphrasing, thereby inheriting the stylistic regularities of the paraphrase corpus (Dolan et al., 2004). Rather than gradually aligning with the stylistic signature as noise diminishes, the model instead converges toward the paraphrasing artifacts, even with almost no privacy guarantee. This dynamic effectively traps the model within its paraphrasing style. Collectively, these findings establish that stylistic degradation is not solely governed by the privacy budget, but is strongly mediated by architectural priors and training history.

3.2 Feature Deviations

To elucidate the linguistic drivers of stylistic divergence, we examine the specific lexico-grammatical

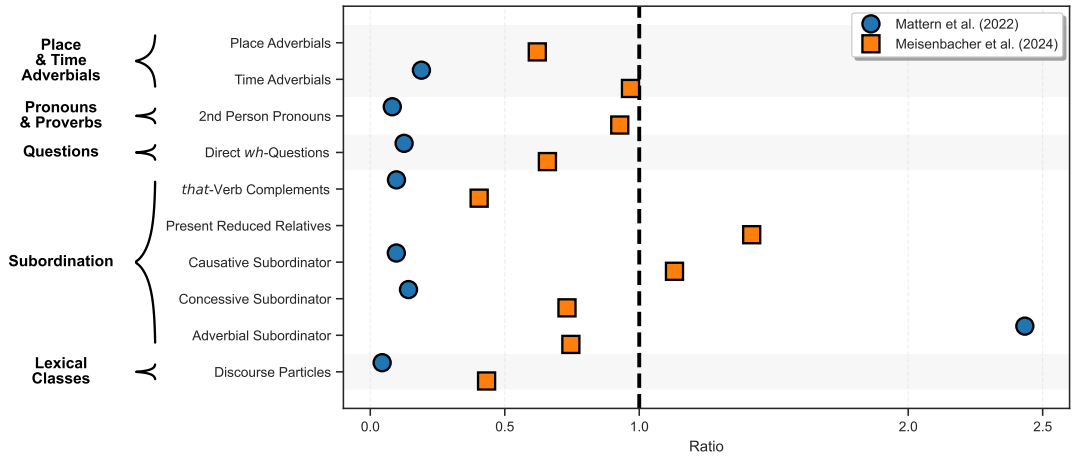

Figure 2: Relative usage ratios of key lexico-grammatical features in privatized text compared to the human-authored CORE baseline (dashed line). Ratios below unity indicate under-representation, while those above signify over-generation. Across all of the most deviating features, DP-MLM (Meisenbacher et al., 2024) maintains usage significantly closer to the human baseline than DP-PARAPHRASE (Mattern et al., 2022).

features that exhibit the highest degree of deviation from human-authored style. This effectively transforms an abstract measure of stylistic decline into a concrete inventory of linguistic drift. We focus on the ten features exhibiting the largest deviations. By isolating these most deviating features, we can identify which grammatical markers are systematically suppressed or amplified by the privatized models. We quantify feature shifts using the usage ratio $R = f_{\text{model}}/f_{\text{human}}$, where $R = 1$ denotes a perfect preservation of human-like frequency, $R < 1$ indicates under-representation, and $R > 1$ signifies over-generation. This formulation allows for a direct comparison of functional feature usage between original and privatized text.

Figure 2 reveals a pronounced asymmetry. Most of the features that deviate are substantially underrepresented. We notice a substantial loss of contextual and interactive markers, where *place adverbials*, *time adverbials*, *personal pronouns*, and *direct questions* all exhibit ratios well below unity. This trend suggests that DP rewriting exerts a sanitizing effect on text, stripping away the specific spatial, temporal, and interpersonal anchors that characterize original communicative intent.

A similar trend emerges for complex syntactic constructions. DP-PARAPHRASE flattens subordinators associated with causation and concession and replaces them with a narrow set of temporal and adversative subordinators to maintain logical coherence. Compared to causative and concessive subordinators, DP-PARAPHRASE over-generates temporal and adversative subordinators with usage frequencies approaching 2.5 times the human baseline, representing the sole feature that is significantly over-generated. In contrast, DP-MLM trades concessive, temporal and adversative subordinators with causative sentence structures. This simplification of syntactic variance suggests that, under privacy constraints, models default to limited repertoire of sentence constructions.

Despite notable deviation from human usage, DP-MLM largely maintains feature usage within a bounded range. For almost all deviating features, its ratios remain approximately within a ± 0.5 interval around the human baseline, indicating that stylistic variation remains within a tolerable margin. This pattern implies that bidirectional substitution preserves a substantial degree of stylistic fidelity, maintaining nuanced stylistic characteristics despite the constraints imposed by the privacy mechanism. In contrast, DP-PARAPHRASE frequently exceeds the range of tolerable margin. This architectural comparison reinforces that DP-MLM is more stylistically robust than DP-PARAPHRASE, which tends to prioritize semantic equivalence over the preservation of stylistic properties.

3.3 Functional Shifts

While lexico-grammatical usage ratios identified the specific linguistic casualties, they do not inherently explain how these isolated changes aggregate to alter the broader communicative intent of the text. To interpret how grammatical variation affects communicative intent, we project the stylistic profiles onto the multidimensional factor

(a) Production (b) Discourse (c) References (d) Argumentation

Figure 3: Functional deviation of privatized text along the four canonical dimensions interpreted by Biber (1991), where the zero-axis represents the human-authored communicative intent. Positive and negative poles correspond to distinct functional registers. Across most dimensions, DP-MLM (Meisenbacher et al., 2024) clusters significantly closer to the original communicative intent than DP-PARAPHRASE (Mattern et al., 2022). However, both approaches tend to favor informational, expository, and neutral modes of communication.

analysis introduced by Biber (1991). Building on the assumption that cooccurrence patterns reflect underlying communicative functions, factor interpretations dictate which features contribute to the positive or negative pole of a given dimension. We adopt the functional dimensions established by Biber (1991) and reuse their canonical groupings of positively and negatively associated features — rather than deriving factor loadings strictly from CORE. These factor loadings were derived from extensive and highly representative corpora spanning a wide array of spoken and written English registers, rendering these dimensions as a robust and widely accepted standard in sociolinguistics. By reusing the established positive and negative feature groupings and projecting our scores onto his canonical dimensions, we anchor observed stylistic shifts within this validated framework.

We calculate the dimension score D_k for each text by summing the standardized frequencies of features with positive loadings $z^{(+)}$ and subtracting those with negative loadings $z^{(-)}$:

$$D_k = \sum z_i^{(+)} - \sum z_j^{(-)}.$$

This procedure yields interpretable scores that reflect shifts in communicative function along established axes of variation. We center the human-authored baseline at zero, enabling comparison of communicative shifts induced by privatization.

Figure 3 illustrates the relative deviation of the privatized text from the original communicative intent across the four canonical dimensions. Across all dimensions, privatized texts exhibit a systematic functional drift away from the distinct functional norms of the original text.

Involved vs. Informational Production: This dimension distinguishes between interactive, affective discourse and abstract, information-dense prose. We observe the most pronounced shift along this dimension. Relative to human discourse, DP-PARAPHRASE exhibits a marked drift toward an informational style, signaling a systematic loss of interactive and affective markers. This shift manifests as a more sterile, impersonal mode of expression, driven by the suppression of personal pronouns and direct questions. While DP-MLM likewise strips away the involved character of a text, it maintains substantially higher production fidelity, with deviations of roughly half the magnitude observed for DP-PARAPHRASE. Overall, this transformation recasts dynamic, person-oriented discourse as impersonal, informational reporting.

Narrative vs. Expository Discourse: This dimension separates narrative discourse, characterized by past-tense construction, from expository styles marked by present-tense usage. DP-MLM remains closely aligned with the narrative degree of human discourse, suggesting the perseveration of discourse structure, whereas DP-PARAPHRASE drops sharply into an expository style. This pattern is reflected by structural erosion of temporal sequencing and storytelling devices. Consequently, the rewriting process homogenizes narrative arcs, effectively converting narrative storytelling into a series of detached expository utterances.

Situational vs. Elaborated Reference: This dimension contrasts explicit reference, typically realized through relative clauses, with references that depend heavily on the immediate temporal or physical context. Whereas elaborated reference renders a text autonomous and interpretable without external cues, situational reference is characterized by a high degree of deixis, anchoring communicative intent in expressions such as *here*, *there*, and *now*. DP-PARAPHRASE largely maintains the level of referential density. DP-MLM elaborates the reference structure by replacing deictic placeholders with more context-independent descriptors. This divergence underscores how architectural constraints shape distinct referential strategies.

Persuasive vs. Neutral Argumentation: This dimension captures the extent to which a human explicitly marks a point of view through the use of infinitives, modals, and suasive verbs. DP-PARAPHRASE departs markedly from the human expression of persuasion, exhibiting a significantly more neutral tone. This reduction is mirrored by a systematic attrition of the rhetorical and persuasive force, effectively neutralizing the original argumentative intent. In contrast, this functional shift is far less pronounced for DP-MLM. This divergence suggests that bidirectional substitution is more robust in preserving argumentative structure, whereas autoregressive paraphrasing tends to flatten subjective stances into a more neutral register.

The dimensional analysis reveals that DP rewriting induces systematic transformations to the communicative signature of text. Although both architectures demonstrated to successfully preserve semantic content, DP-PARAPHRASE and DP-MLM fall short of maintaining the nuanced stylistic variances characteristic of human-authored registers. Owing to its reliance on autoregressive paraphrasing, DP-PARAPHRASE tends to structurally mutate the functional identity, whereas DP-MLM, through bidirectional substitution, remains comparatively more faithful. Nevertheless, both processes manifest into a forced convergence toward a less interactive and less persuasive mode of communication, confirming that formal privacy extends beyond surface variation to fundamentally reshape the text's communicative function.

4 Related Work

Since the inception of DP for text at word-level (Feyisetan et al., 2020), significant progress has been made in enhancing privacy guarantees (Xu et al., 2021) and task utility (Carvalho et al., 2021). Recent advancements have further optimized this trade-off through mapping schemes (Yue et al., 2021; Chen et al., 2023) and selective mechanisms (Tian et al., 2026) that distribute privacy budgets more intelligently. Parallel to these refinements, the field has transitioned toward generating human-readable paraphrases at sentence-level. Early efforts utilized variational autoencoders to perturb latent vectors rather than discrete tokens (Weggenmann et al., 2022), while more recent work takes advantage of the vocabulary space of conditional (Igamberdiev and Habernal, 2023), causal (Mattern et al., 2022; Utpala et al., 2023), and masked (Meisenbacher et al., 2024) language models.

We extend existing analyses into the language quality of differentially-private text. Mattern et al. (2022) and Arnold et al. (2023) examine the retention of morphosyntactic integrity under word-level substitution, whereas Çano and Habernal (2025) assess text at sentence-level using surface-oriented metrics such as lexical diversity and grammatical correctness. We shift the analytical focus from surface measures to the preservation of stylistic and functional identity in differentially-private text.

5 Conclusion

The evolution of differentially-private text rewriting has successfully bridged the gap between formal privacy and grammatical fluency, yet this technical maturation has come at a significant cost to the functional identity of the text. Through a stylometric profiling of lexico-grammatical features and their projection onto broader dimensions of communicative intent, we demonstrate that privacy-constrained rewriting induces systematic stylistic homogenization which is characterized by stripping away the interactive and persuasive markers that anchor human-authored intent.

Beyond this functional shift of communicative intent, our analysis further reveals a critical architectural disparity in how models navigate the privacy-utility trade-off. While the bidirectional substitution of DP-MLM maintains a trajectory of alignment with human stylistic norms as privacy constraints are relaxed, the autoregressive generation of DP-PARAPHRASE reaches a distinct convergence plateau. This persistent deviation, even without formal privacy guarantees, suggests that the model remains trapped by the stylistic priors of its

training history, prioritizing formulaic paraphrasing over the nuanced preservation of grammatical and functional integrity. Advances in DP rewriting depend on moving beyond fluency as the primary objective and addressing stylistic erosion, so that privatized text retains both formal guarantees and the diverse functional integrity of discourse.

Limitations. We note that the primary limitation of this stylometric analysis lies in the evaluation of linguistic style in isolation from adversarial authorship attribution (Huang et al., 2024). This limitation stems from a fundamental entanglement of register and identity: linguistic variation operates along a continuum from functional registers (e.g., persuasive opinion) to the idiosyncratic fingerprint unique to an author (e.g., punctuation densities). Although demonstrating that DP rewriting sanitizes the communicative intent, inducing a convergence toward information-dense, neutral prose, it remains unclear whether this sanitization is a necessary prerequisite for thwarting deanonymization.

We therefore plan to extend this line of linguistic inspection by assessing the *Pareto frontier* between the successful retention of register-level functional identity and author-level privacy risks.

References

Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023. Guiding text-to-text privatization by syntax. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 151–162, Toronto, Canada. Association for Computational Linguistics.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge university press.

Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Vincent Bindschaedler, Reza Shokri, and Carl A Gunter. 2017. Plausible deniability for privacy-preserving data synthesis. *arXiv preprint arXiv:1708.07975*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

John Burrows. 2002. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287.

Erion Çano and Ivan Habernal. 2025. Differentially-private text generation degrades output language quality. *arXiv preprint arXiv:2509.11176*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Ricardo Silva Carvalho, Theodore Vasiloudis, and Oluwaseyi Feyisetan. 2021. Tem: High utility metric differential privacy on text. *arXiv preprint arXiv:2107.07928*.

Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. A customized text sanitization mechanism with differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5747–5758, Toronto, Canada. Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, Geneva, Switzerland. COLING.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Jesse Egbert, Douglas Biber, and Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, 66(9):1817–1831.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 178–186.

Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2013. *Halliday's introduction to functional grammar*. Routledge.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Baixiang Huang, Canyu Chen, and Kai Shu. 2024. Can large language models identify authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 445–460, Miami, Florida, USA. Association for Computational Linguistics.

Timour Igamberdiev and Ivan Habernal. 2023. DP-BART for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934, Toronto, Canada. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 867–881, Seattle, United States. Association for Computational Linguistics.

Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE.

Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. 2024. DP-MLM: Differentially private text rewriting using masked language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9314–9328, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. 2023. Learning interpretable style embeddings via prompting LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15270–15290, Singapore. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–206.

Fengwei Tian, Payel Bhattacharjee, Heidi Hanson, Geoffrey D Rubin, Joseph Y Lo, and Ravi Tandon. 2026. Stamp: Selective task-aware mechanism for text privacy. *arXiv preprint arXiv:2603.12237*.

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. 2023. Locally differentially private document generation using zero shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8442–8457, Singapore. Association for Computational Linguistics.

Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. 2022. Dp-vae: Human-readable text anonymization for online reviews with differentially private variational autoencoders. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 721–731, New York, NY, USA. Association for Computing Machinery.

Nan Xu, Oluwaseyi Feyisetan, Abhinav Aggarwal, Zekun Xu, and Nathanael Teissier. 2021. Density-aware differentially private textual perturbations using truncated gumbel noise. In *The International FLAIRS Conference Proceedings*, volume 34.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online. Association for Computational Linguistics.

A Appendix

Table 2: Lexico-grammatical features, adapted from (Biber, 1991), describing stylistic variance in English and organized based on grammatical and functional categories. Examples *italicized*.

A. Tense and aspect markers	
1	Past tense (e.g., *walked, saw*)
2	Perfect aspect (e.g., *walked, seen*)
3	Present tense (e.g., *walks, sees*)
B. Place and time adverbials	
4	Place adverbials (e.g., *above, beside*)
5	Time adverbials (e.g., *early, soon*)
C. Pronouns and proverbs	
6	First-person pronouns (e.g., *I, we, us*)
7	Second-person pronouns (e.g., *you, yours*)
8	Third-person personal pronouns (e.g., *he, she*, excluding *it*)
9	Pronoun *it*
10	Demonstrative pronouns (*that, this, these, those* as pronouns)
11	Indefinite pronouns (e.g., *anybody, nothing, someone*)
12	Proverb *do*
D. Questions	
13	Direct WH questions (e.g., *What did he see?*)
E. Nominal forms	
14	Nominalizations (ending in *-tion, -ment, -ness, -ity*)
15	Gerunds (participial forms functioning as nouns)
16	Total other nouns (e.g., *house, dog, idea*)
F. Passives	
17	Agentless passives (e.g., *the work was done*)
18	*by*-passives (e.g., *the work was done by ...*)
G. Stative forms	
19	*be* as main verb (e.g., *he is happy*)
20	Existential *there* (e.g., *there is a chance*)
H. Subordination features	
21	*that* verb complements (e.g., *I said that he went*)
22	*that* adjective complements (e.g., *I'm glad that you like it*)
23	WH-clauses (e.g., *I believed what he told me*)
24	Infinitives (e.g., *to walk, to see*)
25	Present participial adverbial clauses (e.g., *Smiling, Joe left.*)
26	Past participial adverbial clauses (e.g., *Built well, the house would stand for fifty years*)
27	Past participial postnominal clauses (e.g., *the solution produced by this process*)
28	Present participial postnominal clauses (e.g., *The event causing this decline was ...*)
29	*that* relative clauses on subject position (e.g., *the dog that bit me*)
30	*that* relative clauses on object position (e.g., *the dog that I saw*)
31	WH relatives on subject position (e.g., *the man who likes popcorn*)
32	WH relatives on object position (e.g., *the man who Sally likes*)

Continued on next page

Table 2: Lexico-grammatical features, adapted from (Biber, 1991), describing stylistic variance in English and organized based on grammatical and functional categories. Examples *italicized*. (Continued)

33	Pied-piping relative clauses (e.g., *the manner in which he was told*)
34	Sentence relatives (e.g., *Bob likes fried mangoes, which is most disgusting.*)
35	Causative adverbial subordinator (*because*)
36	Concessive adverbial subordinators (*although, though*)
37	Conditional adverbial subordinators (*if, unless*)
38	Other adverbial subordinators (*since, while, whereas*)
I. Prepositional phrases, adjectives, and adverbs	
39	Total prepositional phrases (e.g., *in the garden, at the office*)
40	Attributive adjectives (e.g., *the big horse*)
41	Predicative adjectives (e.g., *the horse is big*)
42	Total adverbs (e.g., *quickly, very, happily*)
J. Lexical specificity	
43	Type-token ratio (i.e., *vocabulary diversity*)
44	Mean word length (i.e., *word complexity*)
K. Lexical classes	
45	Conjuncts (e.g., *consequently, furthermore, however*)
46	Downtoners (e.g., *barely, nearly, slightly*)
47	Hedges (e.g., *at about, something like, almost*)
48	Amplifiers (e.g., *absolutely, extremely, perfectly*)
49	Emphatics (e.g., *a lot, for sure, really*)
50	Discourse particles (e.g., sentence-initial *well, now, anyway*)
51	Demonstratives (e.g., *this, that, these, those*)
L. Modals	
52	Possibility modals (e.g., *can, may, might, could*)
53	Necessity modals (e.g., *ought, should, must*)
54	Predictive modals (e.g., *will, would, shall*)
M. Verb classes	
55	Public verbs (e.g., *assert, declare, mention*)
56	Private verbs (e.g., *assume, believe, doubt, know*)
57	Suasive verbs (e.g., *command, insist, propose*)
58	Speculative verbs (e.g., *seems, appear*)
N. Reduced forms and dispreferred structures	
59	Contractions (e.g., *don't, it's, won't*)
60	Subordinator *that* deletion (e.g., *I think [that] he went*)
61	Stranded prepositions (e.g., *the candidate that I was thinking of*)
62	Split infinitives (e.g., *He wants to convincingly prove that...*)
63	Split auxiliaries (e.g., *They were apparently shown to ...*)
O. Coordination	
64	Phrasal coordination (e.g., cats *and* dogs; hot *and* cold)
65	Independent clause co-ordination (e.g., clause-initial *and*)
P. Negation	
66	Synthetic negation (e.g., *No answer is good enough*)
67	Analytic negation (e.g., *That's not likely*)