No Innocence in Styling: Discovery of Privacy Protection Capabilities and Security Risks in Consumer Generative AI Writing Assistants

Mohd. Farhan Israk Soumik[1], Syed Mhamudul Hasan[1], Malithi Mithsara Wanniarachchi Kankanamge[1], Ahmed Imteaj[2], Abdur R. Shahid[1]
[1]Southern Illinois University Carbondale, [2] Florida Atlantic University
mohdfarhanisrak.soumik@siu.edu, shahid@cs.siu.edu

Abstract

Generative AI writing assistants are now integrated into consumer platforms such as Apple Intelligence and Microsoft Copilot, enabling millions of users to automatically rewrite and stylize their text. While positioned as productivity tools, their deployment at scale introduces important and underexplored implications for privacy and platform safety. This paper examines the dual-use nature of platform-level text stylization. Stylization can enhance privacy by suppressing stylistic signals used for profiling and personal data inference. However, the same transformations can be leveraged to evade automated safeguards, including misinformation detection systems. We conduct empirical case studies on emotion inference and misinformation detection across benchmark datasets using deployed stylization modes. We evaluate downstream impact with fine-tuned open-source models and GPT-4o in a zero-shot setting. Our results show that stylization reduces emotion inference accuracy, lowering profiling risk, while increasing error rates in misinformation detection. This discovery reveal a measurable trade-off among privacy protection, moderation robustness, and stylization, highlighting new design and governance challenges for industry deployment.

1 Introduction

The overarching goal of this paper is to examine the real-world viability and implications of large language models deployed within widely adopted consumer platforms, specifically Apple Intelligence within Apple's device ecosystem and Microsoft M365 Copilot across Microsoft's productivity infrastructure. These platforms operate at unprecedented scale, placing text stylization capabilities directly in the hands of millions of users and amplifying both their potential impact and associated risks. On one hand, such tools may be leveraged for benign purposes, including the obfuscation of sensitive personal information or the attenuation of emotional signals to mitigate intrusive social media-based profiling and emotion inference. On the other hand, the same stylization mechanisms may be exploited for malicious objectives, such as manipulating textual content to evade misinformation detection systems, thereby influencing large-scale platforms and public discourse.

In this paper, our study focus on understanding this dual-use paradox through two critical cybersecurity lenses, as outlined in Figure 1: **(1) preservation of user's sensitive information through emotional privacy and (2) exploitation of text stylization for malicious intents such as inference attack**. Currently, a growing privacy concern is adversarial emotion inference, where malicious actors or automated systems extract latent psychological states from user text without consent to enable profiling, workplace discrimination and behavioral manipulation(Zhang et al., 2023; Kqiku et al., 2022; Yuan et al., 2024; Klüwer et al., 2025). The integration of consumer LLMs into NLP workflows has further expanded this risk (Zhang et al., 2025; Bucher and Martini, 2024; Liu et al., 2024b). Consequently, regulatory efforts such as the EU AI Act require providers of high-impact AI systems to mitigate systemic risks to public discourse and individual rights (Neuwirth, 2023; Durovic and Corno, 2024). However, while the law regulates the deployer of these systems, it offers no technical shield to the user because as text leaves a personal device, it enters a grey market of surveillance where privacy is a policy, not a guarantee. Conversely, as stylization becomes a native capability in operating systems and enterprise tools, it opens a new avenue for adversarial linguistic manipulation. Prior work shows that safety classifiers for fake news and toxic content rely heavily on stylistic signals, including emotional intensity and semantic patterns (Rashkin et al., 2017; Potthast et al., 2018). By enabling users to remove these

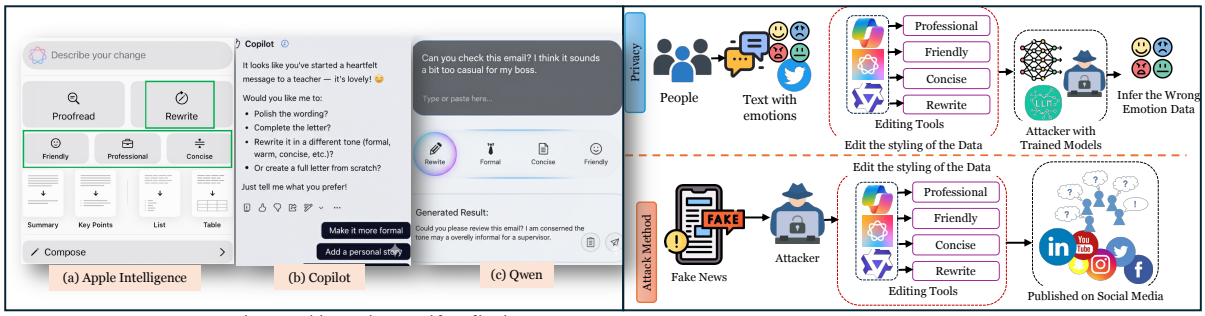

Figure 1: Overview of text stylization, privacy, misinformation, and editing features in deployed platforms operating at scale. Our discovery identifies key privacy protection capabilities and security risks arising from text style change of these tools. The Qwen interface, shown as a representative open-source models' example, was generated using Gemini Banana Pro for illustrative purposes.

signals while preserving meaning, industrial-scale LLMs may unintentionally lower the barrier for bypassing established moderation systems (Wu et al., 2024; Przybyła et al., 2025).

Contributions. This paper makes the following contributions. *(Research Contributions)* First, we present a discovery-driven study demonstrating that platform-integrated text stylization simultaneously enhances user privacy and weakens moderation robustness at deployment scale. Second, we formalize a unified dual-use threat model in which stylization suppresses affective signals used for automated profiling while enabling moderation evasion against misinformation detection systems. Third, we conduct a cross-model evaluation using Apple Intelligence and Microsoft Copilot across common rewriting modes, measuring downstream effects with fine-tuned open-source models and GPT-4o in a zero-shot setting. *(Industry Impact)* Our findings show that writing assistants act as large-scale transformation layers that shape how platforms profile and moderate content. Because they require no technical expertise and operate in black-box settings, they can be leveraged for both privacy protection and moderation evasion by millions of users. This dual-use dynamic introduces new design and safety challenges for real-world AI deployment.

2 Related Work

LLM-Based Text Stylization. Earlier work on text stylization attempted to disentangle content and style in high-dimensional latent spaces using encoder–decoder models and VAEs (Shen et al., 2017; Hu et al., 2017). Limitations of these methods motivated alternative strategies, including prototype editing and self-attention–based transformers that modify style-carrying tokens while maintaining semantic coherence (Li et al., 2018; Dai et al., 2019). More recently, stylization has evolved toward controlled generation with large language models (LLMs), leveraging in-context learning and instruction tuning to guide transformer attention through prompts (Liang et al., 2024; Toshevska et al., 2025). Subsequent research explores manipulating internal representations by identifying style-specific neurons and modeling activation differences, enabling zero-shot transfer while preserving semantics and reducing copying (Lai et al., 2024; Kong et al., 2025). While early approaches primarily targeted sentence-level attributes such as sentiment or formality, ZeroStylus introduces a hierarchical framework for long-text stylization with structural coherence, crucial for professional writing assistants (Wu and Deng, 2025). LLM based Stylization, now a days, is increasingly being treated as an alignment task which utilizes feedback loops to refine output quality and enforces strict content constraints to preserve factual accuracy (Liu and May, 2025).

Platform Scale Integration. The integration of LLMs into consumer operating systems coincides with a shift from cloud-centric inference to hybrid edge architectures with on-device models. To enable such deployment, prior work has relied on model distillation. Zhang et al. (Zhang et al., 2024) used LLM-generated self-explanations to improve stylistic fidelity in compact models, while subsequent efforts incorporated rationale-augmented distillation and federated synthetic training (Liu et al., 2024a; A. Fondekar et al., 2024). Building on these advances, platforms such as Apple Intelligence deploy specialized 3B-parameter on-device foundation models integrated with a Private Cloud Compute architecture to provide writing assistance

across devices (Gunter et al., 2024). Similarly, Microsoft M365 Copilot embeds stylization into productivity suites, making AI-driven text transformation accessible to hundreds of millions of users (Microsoft, 2023). This broad device coverage effectively normalizes AI-based stylization as a default communication interface (Maslej et al., 2025).

Security and Privacy Implications for Text Stylization. The widespread adoption of stylization tools creates a dual-use tension between privacy and safety. Defensively, stylization can act as a data obfuscation mechanism, removing stylistic or emotional signals to reduce profiling and authorship attribution risks (Staab et al., 2023). Conversely, adversarial studies show that similar transformations can bypass safety guardrails (Wang et al., 2025b). Adversarial rephrasing can neutralize stylistic and emotional markers relied upon by misinformation detectors (Przybyła et al., 2025; Potthast et al., 2018; Cheng et al., 2025). Moreover, even robust LLM-based classifiers degrade significantly under such attacks (Cheng et al., 2025; Loth et al., 2026), highlighting the need for safety systems resilient to commercial stylization tools.

Distinction from Prior Work. Prior work on text stylization primarily focuses on improving generation quality, semantic preservation, or internal control mechanisms within models. In contrast, we treat stylization as a deployed, platform-level transformation layer and evaluate its downstream impact on profiling and moderation systems. While existing research optimizes stylistic control or alignment, we quantify the systemic privacy and security implications of stylization in real-world consumer ecosystems. Our contribution lies in revealing and measuring the dual-use trade-off between privacy protection and moderation robustness at deployment scale.

3 Threat Model

Target System. We consider online platforms such as social networks where users generate and share text content. These platforms deploy automated inference and moderation systems, including emotion profiling and misinformation detection models. Users may employ built-in generative AI writing assistants, such as platform-integrated stylization tools, to transform their text before posting. The threat model is illustrated in Figure 1.

Scenario 1: Privacy Protection Against Profiling. From a privacy perspective, the adversary is the platform itself or third-party entities with access to user content. (*Attacker Goal*) Infer sensitive personal attributes such as emotional state or behavioral traits from user-generated text for profiling or downstream personalization. (*Attacker Capability*) Access to large-scale textual data and the ability to train or fine-tune powerful inference models on this data.

Scenario 2: Moderation Evasion. From a security perspective, the user becomes the adversary. (*Attacker Goal*) Bypass automated moderation systems, such as misinformation or fake news detection models, while preserving the semantic intent of the original content. (*Attacker Capability*) Access to generative AI stylization tools that can rewrite content in different tones or styles. In this scenario, no technical expertise is required, lowering the barrier to attack and increasing potential scale.

4 Methodology

To evaluate the dual-use implications defined in our threat model, we design the methodology to simulate two realistic deployment scenarios: (1) emotion and attribute inference used for profiling, and (2) misinformation detection used for automated content moderation. The core components of the methodology are as follows. The methodology is summarized in Table 1.

Stylization Systems. We consider two representative ecosystems reflecting current industrial practice: (i) Apple Intelligence writing tools (on-device consumer deployment) and (ii) Microsoft Copilot (cloud-based enterprise assistant). For comparison, we also include open-source on-device stylization frameworks built on Qwen-30B-3AE, LLaMa3.1-8B, and Mistral-7B. This enables comparison across proprietary on-device, proprietary cloud, and open-source small language model paradigms.

Stylization Modes. For controlled comparison, we align stylistic transformations across systems using four common modes: *Rewrite*, *Professional/Formal*, *Friendly*, and *Concise*. Each original text instance is transformed under all available modes, producing parallel stylized corpora for downstream evaluation.

Scenario	Adversary	Datasets	Classifier Training	LLM Config.	Evaluation Pipeline	Metrics	Measured Effect
Emotional Privacy	Platform / Third-party	Dair-AI, Daily-Dialog, ISEAR	BERT, RoBERTa, DistilBERT, DeBERTa-v3, Flan-T5 (emotion)	GPT-4o (zero-shot)	Input → Stylization → Emotion Inference	Entailment Score (E), BODEGA Score (S_{BODEGA}), sBLEU, Privacy Preservation Rate (PPR)	Reduced inference accuracy
Moderation Evasion	Malicious User	ISOT Fake News	BERT, RoBERTa, DistilBERT, DeBERTa-v3, Flan-T5 (fake news)	GPT-4o (zero-shot)	Input → Stylization → Fake News Detection	Entailment Score (E), BODEGA Score (S_{BODEGA}), sBLEU, Label Flipping Rate(LFR)	Increased misclassification

**Stylization:* Rewrite, Professional, Friendly, Concise using Apple Intelligence, Copilot,Qwen-30B-3AE, LLaMa3-8B, Mistral-7B
**Dataset Statistics:* Dair-AI (20K), DailyDialog (80K), ISEAR (76K), ISOT (45K); 4× stylized variants per instance. Microsoft Copilot and open source models (Qwen-30B-3AE, LLama3-8B, Mistral-7B) were applied to the full datasets, while Apple Intelligence used a balanced 10K-instance subset due to API constraints.

Table 1: Dual-use evaluation framework for privacy and moderation under stylized text transformation.

Datasets. We evaluate three emotion classification benchmarks and one misinformation dataset. For emotional privacy analysis, we use Dair-AI (Saravia et al., 2018) (20K tweets representing short-form text), DailyDialog (Li et al., 2017) (80K human-written conversational utterances reflecting everyday personal communication), and ISEAR (Scherer and Wallbott, 1994) (approximately 76K longer narrative texts describing emotional experiences). For moderation robustness, we use the ISOT fake news dataset (Ahmed et al., 2018), which contains approximately 45K news articles labeled as real or fake. Due to the absence of a public API for Apple Intelligence, we manually constructed balanced subsets comprising 10K original instances across emotion and news categories. Each instance was transformed using four stylization modes (rewrite, professional, friendly, concise). For Microsoft Copilot and open source models, we applied the same four stylization modes to the full datasets, generating four stylized variants per original text.

Downstream Inference/Moderation Models. To simulate realistic inference and moderation pipelines for our scenarios, we evaluate stylized and original texts using both encoder-based and large language model classifiers. Encoder-based models include BERT, RoBERTa, DistilBERT, and DeBERTa-v3 (Radford and Narasimhan, 2018; Radford et al., 2019; Liu et al., 2019). We additionally include an encoder–decoder model (Flan-T5) and a frontier large language model (GPT-4o) in zero-shot settings (Qiu et al., 2020; Raffel et al., 2020). This multi-model evaluation ensures that observed effects are not architecture-specific and reflect broader ecosystem-level impacts.

Evaluation Metrics. In the evaluation, we consider Entailment Score (E) (Galimzianova et al., 2025), sBLEU (Papineni et al., 2002), and BODEGA Score (S_{BODEGA}) (Pawlick et al., 2019) metrics to evaluate the models in both scenarios. We also consider Privacy Preservation Rate (PPR) (Pawlick et al., 2019) to quantify privacy improvement in the scenario of privacy preservation using styling and consider Label Flipping Rate (LFR) (Wang et al., 2025a) to measure the adversarial effect of styling on fake news detection models in the second scenario.

5 Evaluation Results

5.1 Results Aggregation.

To generate the results for both scenarios, we utilized the following aggregation strategy. For the **emotional privacy scenario**, we evaluate each platform (Apple Intelligence, Microsoft Copilot, and open-sourced models) under four stylization modes: *Rewrite*, *Professional*, *Friendly*, and *Concise*. After stylization, we run all downstream models. For each stylization mode, we compute Entailment Score (E), sBLEU, BODEGA Score (S_{BODEGA}), and Privacy Preservation Rate (PPR), and report the average across models in Figure 2. For the **moderation evasion scenario**, we apply the same four stylization modes for each platform and evaluate all downstream models. For each stylization condition, we compute the Entailment Score (E), sBLEU, BODEGA Score (S_{BODEGA}), and Label Flipping Rate (LFR), and again report the average across models in Figure 3. Figure 4(a-c) and Figure 4(d) demonstrates the effect of stylization augmentation on downstream models's privacy preservation and moderation evasion.

5.2 Analysis of Privacy Protection Capability

Figure 2 summarizes the privacy-utility trade-off induced by stylization across three datasets and five platforms. Privacy is measured by PPR (higher is better), while semantic fidelity is captured through entailment (E) and sBLEU, with BODEGA summarizing the joint trade-off. Across all settings, stylization exposes a clear tension: methods that

Platform	Dataset	Feature	PPR (%) ↑	E ↑	sBLEU ↑	Bodega ↑
Apple Intelligence	Dair-AI	Rewrite	29.65	0.93	**0.60**	0.150
		Professional	**49.79**	0.91	0.40	**0.160**
		Friendly	48.91	0.71	0.52	0.150
		Concise	21.31	**0.98**	0.48	0.100
	DailyDialog	Rewrite	14.87	**0.92**	**0.85**	0.217
		Professional	**47.28**	0.88	0.75	0.069
		Friendly	46.93	0.86	0.79	0.138
		Concise	12.49	0.79	0.61	**0.245**
	ISEAR	Rewrite	16.00	**0.89**	**0.81**	0.115
		Professional	**42.00**	0.86	0.72	**0.260**
		Friendly	40.00	0.84	0.76	0.255
		Concise	14.00	0.76	0.58	0.062

Platform	Dataset	Feature	PPR (%) ↑	E ↑	sBLEU ↑	Bodega ↑
Microsoft Copilot	Dair-AI	Rewrite	32.00	**0.95**	**0.89**	0.271
		Professional	**54.00**	0.93	0.81	**0.407**
		Friendly	52.00	0.91	0.84	0.398
		Concise	22.00	0.86	0.67	0.127
	DailyDialog	Rewrite	13.50	**0.93**	**0.86**	0.108
		Professional	**44.00**	0.90	0.79	**0.313**
		Friendly	42.00	0.88	0.82	0.303
		Concise	11.00	0.82	0.63	0.057
	ISEAR	Rewrite	18.00	**0.90**	**0.82**	0.133
		Professional	**40.00**	0.89	0.74	**0.263**
		Friendly	38.00	0.86	0.77	0.252
		Concise	16.00	0.78	0.60	0.075

Platform	Dataset	Feature	PPR (%) ↑	E ↑	sBLEU ↑	Bodega ↑
Mistral	Dair-AI	Rewrite	46.20	0.90	**0.32**	**0.132**
		Professional	**55.10**	0.78	0.26	0.115
		Friendly	52.30	0.68	0.28	0.102
		Concise	37.40	**0.94**	0.20	0.071
	DailyDialog	Rewrite	41.20	**0.91**	**0.46**	**0.129**
		Professional	**54.60**	0.74	0.33	0.102
		Friendly	49.80	0.66	0.37	0.095
		Concise	35.10	0.92	0.22	0.065
	ISEAR	Rewrite	38.20	**0.77**	**0.47**	0.125
		Professional	**46.50**	0.70	0.45	**0.143**
		Friendly	44.10	0.57	0.41	0.112
		Concise	32.60	0.64	0.20	0.052

Platform	Dataset	Feature	PPR (%) ↑	E ↑	sBLEU ↑	Bodega ↑
Qwen	Dair-AI	Rewrite	24.00	**0.92**	**0.87**	0.192
		Professional	**39.00**	0.87	0.74	**0.251**
		Friendly	36.00	0.85	0.80	0.245
		Concise	14.00	0.80	0.62	0.069
	DailyDialog	Rewrite	12.00	**0.90**	**0.84**	0.091
		Professional	**37.00**	0.82	0.68	**0.206**
		Friendly	33.00	0.80	0.76	0.201
		Concise	9.00	0.75	0.57	0.038
	ISEAR	Rewrite	14.00	**0.86**	**0.80**	0.096
		Professional	**33.00**	0.78	0.65	**0.167**
		Friendly	29.00	0.76	0.72	0.159
		Concise	12.00	0.70	0.54	0.045

Platform	Dataset	Feature	PPR (%) ↑	E ↑	sBLEU ↑	Bodega ↑
LLaMA	Dair-AI	Rewrite	48.50	0.86	**0.28**	**0.115**
		Professional	**57.20**	0.74	0.22	0.102
		Friendly	54.10	0.66	0.25	0.095
		Concise	36.80	**0.92**	0.18	0.068
	DailyDialog	Rewrite	39.80	**0.88**	**0.42**	**0.112**
		Professional	**52.40**	0.71	0.31	0.090
		Friendly	47.60	0.67	0.36	0.088
		Concise	31.50	0.90	0.20	0.058
	ISEAR	Rewrite	36.80	**0.74**	**0.45**	0.112
		Professional	**45.80**	0.66	0.42	**0.126**
		Friendly	42.90	0.55	0.40	0.103
		Concise	30.80	0.62	0.19	0.047

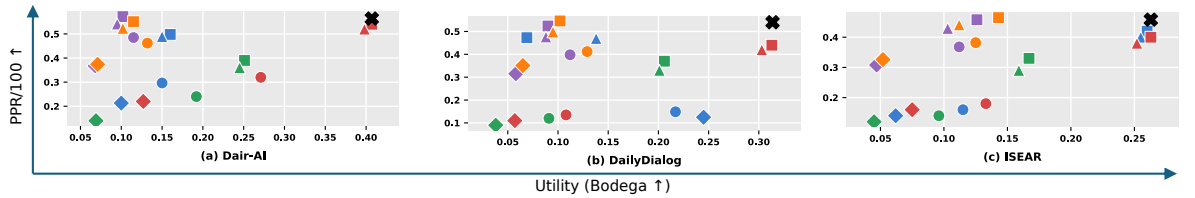

Figure 2: Privacy–utility trade-off across Dair-AI, DailyDialog, and ISEAR datasets. Values are averaged across stylization modes and platforms. Higher values indicate stronger privacy (PPR) or better utility (BODEGA). BODEGA combines entailment and self-BLEU; we therefore analyze the trade-off between PPR and BODEGA. × symbol denotes the maximum value.

preserve text more literally achieve higher E and sBLEU but yield smaller privacy gains, whereas more transformative styles increas PPR at the risk of semantic drift. Among the four modes, Professional and Friendly overall provide strong privacy gains, often achieving the highest PPR values (frequently in the 40-57% range) with moderate-to-high entailment. This indicates that richer stylistic rewrites can effectively weaken affective cues exploited by emotion classifiers while largely preserving communicative intent. However, this scenario significantly shifts under the joint privacy-utility trade-off captured by BODEGA, where the dominant mode often becomes dependent on stylization models. In cloud and controlled systems such as Copilot and Apple Intelligence and comparatively larger open source model like Qwen-30B-3AE, Professional and Friendly frequently occupy the upper region of the Pareto frontier across datasets. In constrast to this, comparatively smaller open source models like LLaMA3-8B and Mistral-7B undescore a more variable behavior, where Rewrite occasionally achieves competitive BODEGA scores owing to its stronger semantic fidelity, closely followed by Professional and Friendly. Among all the modes, Rewrite in general preserves the highest E and sBLEU but offers limited privacy benefit, suggesting that surface-level paraphrasing leaves emotional signals largely intact. While Concise, across most of the settings, produces weaker overall performance, reducing entailment and lexical similarity without delivering commensurate privacy gains. Although there exists isolated exceptions such as DailyDialog under Apple Intelligence where compression effects yield competitive BODEGA scores. Platform differences further highlight the role of underlying model architectures and capability. Microsoft Copilot, with it's cloud based GPT-5.2 model, overall exhibits the strongest privacy-utility frontier, followed by Apple Intelligence with its black-box PCC architecture. While open-source and smaller models like Qwen-30B-3AE, LLaMA3-8B, and Mistral-7B demonstrates similar qualitative trends but with reduced or more variable performance. The most important aspect is that although there lies quantitative differences between platforms but most of them follow the very same foundamental trade-off pattern across stylization modes. Beyond model-level differences, dataset characteristics also shape the observed privacy gains. Stylization is more effective on Dair-AI, where emotions are often lexically explicit, and more moderate on DailyDialog and ISEAR, where affect is expressed more implicitly. Overall, Professional remains the most stable privacy-oriented mode, although its advantage varies across datasets.

Platform	Feature	LFR (%) ↑	E ↑	sBLEU ↑	BODEGA ↑
Apple Intelligence	Rewrite	34.0	**0.94**	**0.88**	0.281
	Professional	**62.0**	0.90	0.74	**0.413**
	Friendly	48.0	0.88	0.80	0.338
	Concise	40.0	0.83	0.62	0.206
Microsoft Copilot	Rewrite	38.0	**0.95**	**0.89**	0.321
	Professional	**68.0**	0.92	0.78	**0.488**
	Friendly	52.0	0.90	0.83	0.389
	Concise	44.0	0.85	0.67	0.251
Qwen-30B-3AE	Rewrite	28.0	**0.92**	**0.87**	0.224
	Professional	**55.0**	0.86	0.70	**0.331**
	Friendly	42.0	0.84	0.78	0.275
	Concise	32.0	0.78	0.60	0.150
LLaMA3.1-8B	Rewrite	42.3	**0.47**	0.40	**0.078**
	Professional	44.2	0.40	**0.43**	0.076
	Friendly	**47.4**	0.22	0.39	0.040
	Concise	42.2	0.37	0.13	0.020
Mistral-7B	Rewrite	39.0	**0.71**	0.46	0.127
	Professional	42.2	0.67	**0.47**	**0.135**
	Friendly	**47.0**	0.35	0.40	0.066
	Concise	42.4	0.50	0.15	0.033

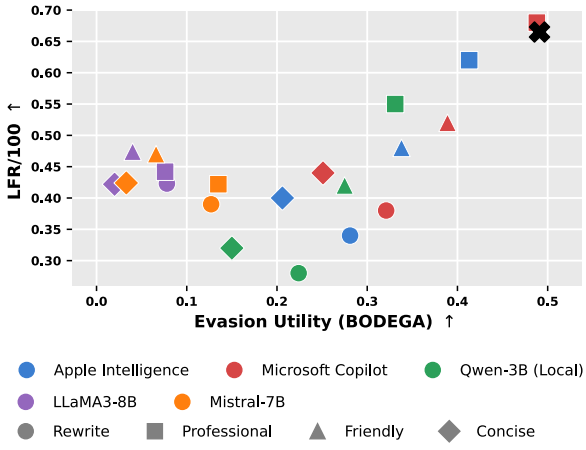

Figure 3: Evasion-Utility frontier for ISOT dataset. Values represent the mean performance of stylization modes across platforms; higher values signify a more effective balance in adversarial evasion.

Takeaway: Stylization provides effective but conditional privacy benefits, with Professional favoring higher privacy while the overall trade-off remains dependent on model capability and dataset characteristics.

5.3 Analysis on the Security Issues In Terms of Moderation Evasion

Figure 3 shows how stylization affects adversarial evasion on the ISOT dataset, where safety degradation is measured by the Label Flipping Rate (LFR; fake→real) and fidelity by entailment (E) and sBLEU, with BODEGA capturing evasion utility. Professional consistently produces the highest LFR on cloud and controlled platforms, reaching 62.0% on Apple Intelligence, 68.0% on Copilot, and 55.0% on Qwen-30B-3AE, alongside the highest BODEGA scores on these systems (0.413, 0.488, and 0.331). This indicates the most effective laundering of fake content into semantically plausible text. Friendly ranks behind Professional on the same platforms, trading lower LFR for higher surface similarity. The picture shifts on LLaMA3.1-8B and Mistral-7B, where Friendly attains the highest LFR (47.4% and 47.0%) but at a steep entailment cost (E of 0.22 and 0.35), so its evasion gains do not translate into competitive BODEGA. Rewrite preserves the highest entailment across all five platforms again but achieves substantially lower LFR on the cloud and controlled platforms, showing that surface-level paraphrasing retains the cues exploited by detectors. Concise as observed in privacy study, generally performs worst, degrading fidelity without commensurate evasion gains. Across platforms, Copilot exhibits the strongest evasion-utility frontier, followed by Apple Intelligence and Qwen-30B-3AE, while LLaMA3.1-8B and Mistral-7B remain considerably lower across all modes.

Takeaway: The same stylization modes that best preserve utility also most strongly undermine misinformation detection, reinforcing the dual-use risk of consumer stylization tools.

5.4 Downstream Effects of Stylization-Based Data Augmentation

We further examine the downstream effects of stylization as a scalable data augmentation tool for both privacy protection and fake news evasion. We augment the original datasets with stylized samples generated by multiple systems (Apple Intelligence, Copilot, Qwen-30B-3AE, LLaMA3-8B, and Mistral-7B) using the same rewriting categories as original experiments. Most datasets receive approximately 24K additional stylized samples (10K for the smaller ISEAR dataset), followed by the same 80-20 train-test evaluation approach. Results show that augmentation weakens privacy protection (Figure 4(a-c)) while substantially reducing fake news evasion (Figure 4(d)).

6 Actionable Insights for Consumer AI Writing Assistants

Our findings show that platform-integrated text stylization substantially alters stylistic and affective signals, directly impacting downstream inference systems. On one hand, stylization reduces the effectiveness of emotion inference models, indicating

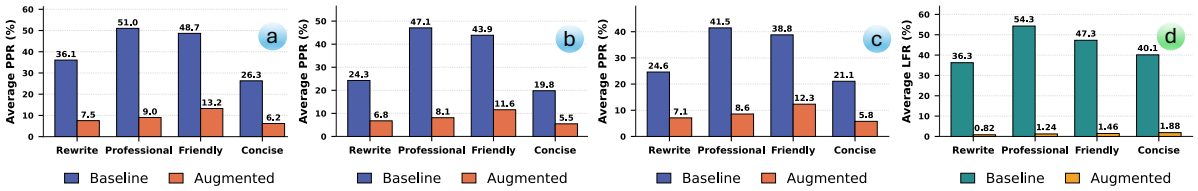

Figure 4: Baseline vs. augmented results for stylization-based data augmentation across four datasets [(a-c) DAIR-AI, DailyDialog, ISEAR, and (d) ISOT]. Augmentation with stylized samples generally lowers emotional privacy protection (PPR) while substantially reducing fake news evasion (LFR), highlighting the downstream dual-use effects of stylization.

its potential as a privacy-enhancing transformation against unwanted profiling. However, the same transformations increase vulnerability in misinformation detection systems, revealing a moderation robustness gap. These results suggest that stylization features should be treated as system-level transformation layers rather than neutral productivity tools. At the deployment scale, writing assistants can simultaneously protect users from unintended personal data inference and enable moderation evasion. This dual-use property requires careful design considerations.

From a privacy perspective, platforms could provide user-configurable stylization controls that make explicit the trade-off between expressiveness and profiling resistance. The empirical results presented in this paper can guide the design of robust, user-centric, on-device privacy-preserving systems. From a safety perspective, moderation systems must be designed to remain robust under stylistic transformations, potentially through stylization-invariant training or adversarial augmentation. The augmentation results further show that the stylized text can provide an scalable approach to augment training data, improving detector robustness while potentially strengthening models that infer sensitive user attributes.

Embedding such safeguards at the platform level is critical, as these tools operate in largely black-box settings and require no technical expertise, making their impact scalable across millions of users.

7 Conclusion

This study focuses on formalizing and validating the dual-use paradox of industrially deployed LLM-based writing assistants, demonstrating that the same semantic rewriting mechanisms that protect users from emotion inference attacks can simultaneously undermine downstream safety systems. Across Apple Intelligence, Microsoft Copilot, and open source on-device language model frameworks, stylization systematically suppresses emotional and stylistic signals relied upon by emotional profiling and moderation models. This attenuation of cues improves individual privacy by degrading emotion inference accuracy while simultaneously enabling semantic laundering, allowing toxic or deceptive content to bypass lexical and style-dependent downstream safety classifiers. These findings suggest that among the current industrial safety pipelines, which largely rely on lexical and emotional markers, are insufficient for the post-LLM era. For industry, this reframes writing assistants as dual-use system components that must be explicitly integrated into safety design. Future safety architectures should emphasize style-invariant modeling, provenance awareness, and resilience to semantically preserved but stylistically transformed adversarial content.

8 Acknowledgement

This material is partly based upon work supported by the U.S. National Science Foundation (NSF) under Grant No. CRII-IIS-RI-2553868. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

Ashweta A. Fondekar, Milind M. Shivolkar, and Jyoti D. Pawar. 2024. Unpacking faux-hate: Addressing faux-hate detection and severity prediction in code-mixed Hinglish text with HingRoBERTa and class weighting techniques. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)*, pages 6–11, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018.

Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned'small'llms (still) significantly outperform zero-shot generative ai models in text classification. *arXiv preprint arXiv:2406.08660*.

Yize Cheng, Vinu Sankar Sadasivan, Mehrdad Saberi, Shoumik Saha, and Soheil Feizi. 2025. Adversarial paraphrasing: A universal attack for humanizing ai-generated text. *arXiv preprint arXiv:2506.07001*.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. Association for Computational Linguistics.

Mateja Durovic and Tommaso Corno. 2024. The privacy of emotions: From the gdpr to the ai act, an overview of emotional ai regulation and the protection of privacy and personal data. *Privacy, Data Protection and Data-driven Technologies*, pages 368–404.

Daria Galimzianova, Aleksandr Boriskin, and Grigory Arshinov. 2025. From rag to reality: Coarse-grained hallucination detection via nli fine-tuning. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pages 353–359.

Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, Aonan Zhang, Bowen Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, et al. 2024. Apple intelligence foundation language models. *arXiv preprint arXiv:2407.21075*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

Nils Klüwer, Irina Nalis, and Julia Neidhardt. 2025. Context over categories: Implementing the theory of constructed emotion with llm-guided user analysis. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Chaona Kong, Jianyi Liu, Yifan Tang, and Ru Zhang. 2025. Neuron activation modulation for text style transfer: Guiding large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7735–7747, Vienna, Austria. Association for Computational Linguistics.

Lindrit Kqiku, Marvin Kühn, and Delphine Reinhardt. 2022. From sentiment to sensitivity: The role of emotions on privacy exposure in twitter. In *Proceedings of the 2022 Workshop on Open Challenges in Online Social Networks*, pages 10–15.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443, Miami, Florida, USA. Association for Computational Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Xun Liang, Hanyu Wang, Shichao Song, Mengting Hu, Xunzhi Wang, Zhiyu Li, Feiyu Xiong, and Bo Tang. 2024. Controlled text generation for large language model with dynamic attribute graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5797–5814, Bangkok, Thailand. Association for Computational Linguistics.

Pusheng Liu, Lianwei Wu, Linyong Wang, Sensen Guo, and Yang Liu. 2024a. Step-by-step: Controlling arbitrary style in text with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15285–15295, Torino, Italia. ELRA and ICCL.

Shuai Liu and Jonathan May. 2025. Style transfer with multi-iteration preference optimization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2663–2681, Albuquerque, New Mexico. 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*.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024b. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5487–5496.

Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Industrialized deception: The collateral effects of llm-generated misinformation on digital ecosystems. *arXiv preprint arXiv:2601.21963*.

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily

Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, et al. 2025. Artificial intelligence index report 2025. *arXiv preprint arXiv:2504.07139*.

Microsoft. 2023. Introducing copilot for microsoft 365. https://www.microsoft.com/en-us/microsoft-365/copilot.

Rostam J Neuwirth. 2023. Prohibited artificial intelligence practices in the proposed eu artificial intelligence act (aia). *Computer Law & Security Review*, 48:105798.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pawlick, Edward Colbert, and Quanyan Zhu. 2019. A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy. *ACM Computing Surveys (CSUR)*, 52(4):1–28.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 231–240.

Piotr Przybyła, Euan McGill, and Horacio Saggion. 2025. Attacking misinformation detection using adversarial examples generated by language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27614–27630.

Piotr Przybyła, Alexander Shvets, and Horacio Saggion. 2024. Verifying the robustness of automatic credibility assessment. *Natural Language Processing*, pages 1–29.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China technological sciences*, 63(10):1872–1897.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

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.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Martina Toshevska, Slobodan Kalajdziski, and Sonja Gievska. 2025. Style knowledge graph: Augmenting text style transfer with knowledge graphs. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, pages 123–135, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Qianli Wang, Nils Feldhus, Luis Felipe Villa-Arenas, Christin Seifert, Sebastian Möller, Vera Schmitt, et al. 2025a. Truth or twist? optimal model selection for reliable label flipping evaluation in llm-based counterfactuals. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 80–97.

Xinyu Wang, Wenbo Zhang, Sai Koneru, Hangzhi Guo, Bonam Mingole, S Shyam Sundar, Sarah Rajtmajer, and Amulya Yadav. 2025b. Have llms reopened the pandora's box of ai-generated fake news? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2795–2811.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3367–3378.

Yusen Wu and Xiaotie Deng. 2025. Implementing long text style transfer with llms through dual-layered sentence and paragraph structure extraction and mapping. *arXiv preprint arXiv:2505.07888*.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036, Miami, Florida, USA. Association for Computational Linguistics.

Chiyu Zhang, Honglong Cai, Yuezhang Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. 2024. Distilling text style transfer with self-explanation from llms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 200–211.

Ting Zhang, Ivana Clairine Irsan, Ferdian Thung, and David Lo. 2025. Revisiting sentiment analysis for software engineering in the era of large language models. *ACM Transactions on Software Engineering and Methodology*, 34(3):1–30.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

9 Appendix

9.1 Evaluation Metrics

We evaluate stylization with four complementary metrics to capture *attack success* (security), *semantic fidelity* (meaning preservation), *privacy gain* (inference suppression), and an *overall utility trade-off*. No single metric reflects all four dimensions.

Label Flipping Rate (LFR): Measures adversarial efficacy as the fraction of samples whose predicted label changes after stylization (Wang et al., 2025a):

$$LFR = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\big(C(x_n^{\text{styl}}) \neq y_n\big)\,. \quad (1)$$

Entailment Score (E): Uses DeBERTa-v3-NLI to verify that the stylized text preserves the original meaning by computing entailment probability (Galimzianova et al., 2025):

$$E(x, y) = P(\text{entailment} \mid x, y). \quad (2)$$

Privacy Preservation Rate (PPR): Quantifies privacy improvement as the relative drop in an inference attacker's accuracy after stylization (Pawlick et al., 2019):

$$PPR = 1 - \frac{Acc_{\text{stylized}}}{Acc_{\text{original}}} \quad (3)$$

Self-BLEU (sBLEU): Measures lexical overlap between the original text x and stylized text x_{styl} to quantify the degree of stylistic transformation (Papineni et al., 2002). A lower score indicates higher diversity and a more significant rewrite:

$$sBLEU(x, x_{styl}) = \exp\left(\sum_{n=1}^{N} w_n \ln p_n\right) \quad (4)$$

BODEGA Score (S_{BODEGA}): Aggregates attack success with semantic and lexical fidelity into one utility score (Przybyła et al., 2024):

$$S_{\text{BODEGA}} = succ \cdot sim_{\text{sem}} \cdot sim_{\text{lex}} \quad (5)$$

where $succ \in \{0, 1\}$ indicates a successful evasion/flip, sim_{sem} is semantic similarity, and sim_{lex} is lexical similarity (normalized Levenshtein).

9.2 Experimental Setup

The experimental environment is distributed across specialized hardware to ensure the reproducibility of both cloud-based and local inferences. On-device evaluations for Apple Intelligence were conducted on a MacBook Pro equipped with an Apple M2 Silicon chip and 32 GB of unified RAM, while the Qwen3-4B with 4-bit quantization and M365 Copilot on a Windows workstations featuring an NVIDIA RTX 4500 Ada GPU, A100 GPU respectively.

9.3 Models

9.3.1 BERT

BERT is a deep learning model based on the transformer architecture. It is developed to learn contextual representations bidirectionally using multi-head self attention along with feedforward layers.

Configuration	Value
Pretrained Model	bert-base-uncased
Learning Rate	2e-4
Dropout Rate	0.2

Table 2: BERT Configuration

9.3.2 RoBERTa

RoBERTa, developed by Facebook AI is an optimized variant of BERT. Compared to BERT, it has been trained on much larger dataset and it uses masked language modeling instead of next token generation.

Configuration	Value
Pretrained Model	twitter-roberta-base
Learning Rate	3e-4

Table 3: RoBERTa Configuration

9.3.3 DeBERTa

DeBERTa is a transformer-based language model developed by Microsoft that improves upon BERT and RoBERTa by introducing two key innovations: disentangled attention and an enhanced decoding mechanism. Unlike traditional models that combine word content and position embeddings before feeding them into the attention mechanism, DeBERTa keeps them separate, allowing the model to better capture the relationships between words based on both their content and position independently.

Configuration	Value
Pretrained Model	deberta-base-uncased
Learning Rate	2e-4
Training Method	Full Fine Tuning

Table 4: DeBERTa Configuration

9.3.4 DistilBERT

DistilBERT is a lighter version of BERT developed by Hugging Face through knowledge distillation. It was trained to mimic the behavior of the larger BERT model by learning from its outputs, effectively compressing the knowledge without significant performance loss.

Configuration	Value
Pretrained Model	distilbert-base-uncased
Learning Rate	3e-4
Training Method	Full Fine Tuning

Table 5: DistilBERT Configuration

9.3.5 Flan T5

Flan-T5 is an advanced version of Google's T5 model. It was fine-tuned using instruction tuning on a variety of tasks to enhance its ability to follow natural language instructions. Flan-T5 significantly improves zero-shot and few-shot learning performance across multiple benchmarks.

9.3.6 GPT-4o

GPT-4o is OpenAI's latest AI model. It was designed to handle text, image, and audio inputs, offering a multi-modal experience. GPT-4o uses a unified architecture to integrate and understand information across modalities in real time. It maintains the strong language capabilities of GPT-4 while significantly improving performance on vision and audio tasks, such as interpreting images, recognizing emotions in speech, or holding fluid voice conversations.

Configuration	Value
Pretrained Model	google/flan-t5-base
Learning Rate	5e-4
Training Method	PEFT-LoRa
Lora Rank Matrix	16

Table 6: Flan T5 Configuration

9.4 Datasets

We use two datasets to demonstrate the results:

9.4.1 Dair-AI Emotion Dataset

The Dair-AI Emotion Dataset is a collection of English Twitter messages labeled with six basic emotions: anger, fear, joy, love, sadness, and surprise. This dataset is designed for emotion recognition research and has been preprocessed for ease of use in NLP pipelines. It contains 20,000 text instances which are divided into training (16,000 instances), validation (2,000 instances), and test (2,000 instances) sets.

Emotion	Count
Anger	4666
Fear	5362
Joy	1304
Love	2159
Sadness	1937
Surprise	572

Table 7: Emotion distribution in the Dair-AI Emotion dataset

9.4.2 DailyDialog Dataset

The DailyDialog dataset is a multi-turn, open-domain English dialog collection. It comprises 13,118 dialogues, reflecting daily communication and covering various topics. The data set is divided into training sets (11,118 dialogues), validation sets (1,000 dialogues), and test sets (1,000 dialogues). On average, each dialogue consists of approximately 8 speaker turns, with around 15 tokens per turn. The conversations are manually crafted, ensuring high-quality and natural language interactions and encompass a wide range of daily life

topics, providing a rich resource for open-domain conversation modeling. Unlike conversations, dialogues are also manually labeled with communication intentions and emotion information, facilitating research in dialogue systems, emotion recognition, and natural language understanding.

Category	Count
Total Dialogues	13,118
Training Set	11,118
Validation Set	1,000
Test Set	1,000
Average Turns per Dialogue	8
Average Tokens per Turn	15

Table 8: Distribution of the DailyDialog Dataset

9.4.3 ISEAR Dataset

The International Survey on Emotion Antecedents and Reactions (ISEAR) dataset is a prominent resource for emotion classification, consisting of 7,666 first-person emotional reports. Collected from nearly 3,000 participants across diverse cultural backgrounds, the dataset captures personal experiences linked to seven distinct emotions: anger, disgust, fear, guilt, joy, sadness, and shame. Unlike social media-based datasets, ISEAR provides formal, structured narratives that offer deep psychological insights into how individuals describe emotional triggers.

Emotion	Count
Anger	1,096
Disgust	1,096
Fear	1,095
Guilt	1,093
Joy	1,094
Sadness	1,096
Shame	1,096

Table 9: Emotion distribution in the ISEAR dataset

9.4.4 ISOT Fake News Dataset

The ISOT Fake News dataset is a comprehensive collection of real and fabricated news articles specifically curated for misinformation research. It contains over 44,000 articles, where truthful content was obtained from Reuters.com and fake content was sourced from unreliable websites flagged by fact-checking organizations like PolitiFact. The dataset primarily focuses on political and world news topics from 2016 to 2017, providing a robust benchmark for binary classification tasks in the domain of fake news detection.

Category	Count
Total Articles	44,898
Real News (Reuters)	21,417
Fake News	23,481
Average Article Length	500 words
Primary Topics	Politics, World

Table 10: Distribution of the ISOT Fake News Dataset

9.5 Detailed Results

Table 11: Emotion Inference Models Accuracy (%) Comparison on Dair-AI on Apple AI

Emotion Category	**Inference Model**	**Original Text**	**Apple Rewrite**	**Apple Professional**	**Apple Friendly**	**Apple Concise**
Anger	BERT	95.12	62.50	35.32	45.00	87.50
	RoBERTa	95.20	65.00	40.00	40.00	87.50
	DeBERTa	97.56	67.50	37.50	35.00	87.50
	DistilBERT	100.00	67.50	60.00	40.00	92.50
	Flan T5	97.56	75.00	32.00	65.00	78.00
	GPT-4o	98.00	96.50	88.00	90.00	98.00
Sadness	BERT	100.00	85.00	49.20	62.50	82.50
	RoBERTa	97.50	75.00	56.41	55.00	87.50
	DeBERTa	97.50	70.00	52.45	47.50	77.50
	DistilBERT	97.50	62.50	28.21	42.50	75.00
	Flan T5	100.00	72.50	64.10	60.00	67.50
	GPT-4o	100.00	100.00	92.30	98.00	98.50
Love	BERT	87.50	55.00	50.00	42.00	70.00
	RoBERTa	100.00	65.00	50.00	47.50	82.50
	DeBERTa	100.00	50.00	35.00	37.50	70.00
	DistilBERT	100.00	55.00	35.00	32.50	65.00
	Flan T5	100.00	30.00	25.00	22.50	45.00
	GPT-4o	65.00	37.50	37.50	37.50	37.50
Joy	BERT	100.00	90.00	82.50	75.00	92.50
	RoBERTa	92.50	72.50	73.17	80.00	80.00
	DeBERTa	92.50	82.50	77.50	80.00	85.00
	DistilBERT	92.50	82.50	90.00	77.50	85.00
	Flan T5	92.50	65.00	72.50	57.50	65.00
	GPT-4o	65.00	37.50	37.50	37.50	37.50
Fear	BERT	100.00	82.50	67.50	50.00	87.50
	RoBERTa	100.00	82.50	67.50	50.00	87.50
	DeBERTa	97.50	82.50	10.00	52.50	82.50
	DistilBERT	95.12	85.37	70.00	45.00	85.00
	Flan T5	95.12	70.00	62.50	60.00	80.00
	GPT-4o	79.79	67.50	67.50	67.50	67.50
Surprise	BERT	100.00	47.50	15.00	32.50	50.00
	RoBERTa	100.00	60.00	20.00	45.00	62.50
	DeBERTa	100.00	60.00	20.00	45.00	62.50
	DistilBERT	100.00	60.00	37.50	10.00	55.00
	Flan T5	100.00	42.50	37.50	50.00	50.00
	GPT-4o	67.00	42.50	42.50	45.00	45.00

Table 12: Emotion Inference Models Accuracy (%) on Dair-AI under Microsoft Copilot Stylization

Emotion Category	**Inference Model**	**Original Text**	**Copilot Rewrite**	**Copilot Profes-sional**	**Copilot Friendly**	**Copilot Concise**
Anger	BERT	95.12	60.50	33.32	43.00	85.50
	RoBERTa	95.20	63.20	38.20	38.20	85.70
	DeBERTa	97.56	65.70	35.70	33.20	85.70
	DistilBERT	100.00	63.00	55.50	35.50	88.00
	Flan T5	97.56	72.50	29.50	62.50	75.50
	GPT-4o	45.50	11.50	11.50	11.50	11.50
Sadness	BERT	100.00	83.00	47.20	60.50	80.50
	RoBERTa	97.50	73.20	54.61	53.20	85.70
	DeBERTa	97.50	68.20	50.65	45.70	75.70
	DistilBERT	97.50	58.00	23.71	38.00	70.50
	Flan T5	100.00	70.00	61.60	57.50	65.00
	GPT-4o	65.00	48.00	48.00	48.00	48.00
Love	BERT	87.50	53.00	48.00	40.00	68.00
	RoBERTa	100.00	63.20	48.20	45.70	80.70
	DeBERTa	100.00	48.20	33.20	35.70	68.20
	DistilBERT	100.00	50.50	30.50	28.00	60.50
	Flan T5	100.00	27.50	22.50	20.00	42.50
	GPT-4o	65.00	35.50	35.50	35.50	35.50
Joy	BERT	100.00	88.00	80.50	73.00	90.50
	RoBERTa	92.50	70.70	71.37	78.20	78.20
	DeBERTa	92.50	80.70	75.70	78.20	83.20
	DistilBERT	92.50	78.00	85.50	73.00	80.50
	Flan T5	92.50	62.50	70.00	55.00	62.50
	GPT-4o	65.00	35.50	35.50	35.50	35.50
Fear	BERT	100.00	80.50	65.50	48.00	85.50
	RoBERTa	100.00	80.70	65.70	48.20	85.70
	DeBERTa	97.50	80.70	8.20	50.70	80.70
	DistilBERT	95.12	80.87	65.50	40.50	80.50
	Flan T5	95.12	67.50	60.00	57.50	77.50
	GPT-4o	79.79	65.50	65.50	65.50	65.50
Surprise	BERT	100.00	45.50	13.00	30.50	48.00
	RoBERTa	100.00	58.20	18.20	43.20	60.70
	DeBERTa	100.00	58.20	18.20	43.20	60.70
	DistilBERT	100.00	55.50	33.00	5.50	50.50
	Flan T5	100.00	40.00	35.00	47.50	47.50
	GPT-4o	67.00	40.50	40.50	43.00	43.00

Table 13: Emotion Inference Models Accuracy (%) on Dair-AI under Qwen Stylization

Emotion Category	**Inference Model**	**Original Text**	**Qwen Rewrite**	**Qwen Professional**	**Qwen Friendly**	**Qwen Concise**
Anger	BERT	95.12	62.00	34.82	44.50	87.00
	RoBERTa	95.20	64.60	39.60	39.60	87.10
	DeBERTa	97.56	67.10	37.10	34.60	87.10
	DistilBERT	100.00	66.00	58.50	38.50	91.00
	Flan T5	97.56	74.20	31.20	64.20	77.20
	GPT-4o	45.50	13.00	13.00	13.00	13.00
Sadness	BERT	100.00	84.50	48.70	62.00	82.00
	RoBERTa	97.50	74.60	56.01	54.60	87.10
	DeBERTa	97.50	69.60	52.05	47.10	77.10
	DistilBERT	97.50	61.00	26.71	41.00	73.50
	Flan T5	100.00	71.70	63.30	59.20	66.70
	GPT-4o	65.00	49.50	49.50	49.50	49.50
Love	BERT	87.50	54.50	49.50	41.50	69.50
	RoBERTa	100.00	64.60	49.60	47.10	82.10
	DeBERTa	100.00	49.60	34.60	37.10	69.60
	DistilBERT	100.00	53.50	33.50	31.00	63.50
	Flan T5	100.00	29.20	24.20	21.70	44.20
	GPT-4o	65.00	37.00	37.00	37.00	37.00
Joy	BERT	100.00	89.50	82.00	74.50	92.00
	RoBERTa	92.50	72.10	72.77	79.60	79.60
	DeBERTa	92.50	82.10	77.10	79.60	84.60
	DistilBERT	92.50	81.00	88.50	76.00	83.50
	Flan T5	92.50	64.20	71.70	56.70	64.20
	GPT-4o	65.00	37.00	37.00	37.00	37.00
Fear	BERT	100.00	82.00	67.00	49.50	87.00
	RoBERTa	100.00	82.10	67.10	49.60	87.10
	DeBERTa	97.50	82.10	9.60	52.10	82.10
	DistilBERT	95.12	83.87	68.50	43.50	83.50
	Flan T5	95.12	69.20	61.70	59.20	79.20
	GPT-4o	79.79	67.00	67.00	67.00	67.00
Surprise	BERT	100.00	47.00	14.50	32.00	49.50
	RoBERTa	100.00	59.60	19.60	44.60	62.10
	DeBERTa	100.00	59.60	19.60	44.60	62.10
	DistilBERT	100.00	58.50	36.00	8.50	53.50
	Flan T5	100.00	41.70	36.70	49.20	49.20
	GPT-4o	67.00	42.00	42.00	44.50	44.50

Table 14: Emotion Inference Models'Accuracy (%) Comparison on DailyDialog under Apple AI Stylization

Emotion Category	Inference Model	Original	Apple Rewrite	Apple Professional	Apple Friendly	Apple Concise
Anger	BERT	86.50	77.14	13.9	14.23	70.00
	RoBERTa	86.40	65.00	33.33	23.33	73.33
	DeBERTa	86.04	67.50	37.50	35.00	88.89
	DistilBERT	86.23	66.67	16.67	20.00	63.33
	Flan T5	89.56	68.00	16.67	25.00	75.00
	GPT-4o	77.34	53.33	53.33	20.00	53.33
Disgust	BERT	76.63	62.30	73.33	9.26	67.30
	RoBERTa	79.10	66.67	68.00	16.67	66.67
	DeBERTa	79.30	66.67	66.53	20.00	63.33
	DistilBERT	77.64	60.00	60.00	10.00	63.33
	Flan T5	82.30	67.50	70.00	22.50	68.00
	GPT-4o	45.30	16.67	16.67	16.67	16.67
Fear	BERT	85.71	73.80	21.43	40.00	83.33
	RoBERTa	85.71	84.00	36.66	46.67	84.50
	DeBERTa	85.71	83.33	40.00	43.33	86.67
	DistilBERT	83.33	76.67	23.33	36.67	83.33
	Flan T5	87.30	80.00	35.00	45.00	86.67
	GPT-4o	100.00	93.33	90.00	86.67	93.33
Happiness	BERT	85.71	85.71	40.47	81.95	71.42
	RoBERTa	88.10	87.27	46.67	83.87	76.67
	DeBERTa	88.10	86.67	66.63	93.54	86.67
	DistilBERT	90.47	80.00	43.33	87.50	83.33
	Flan T5	89.00	85.32	35.54	87.50	86.67
	GPT-4o	95.45	70.00	70.00	70.00	63.33
Surprise	BERT	88.37	67.00	9.34	30.00	61.67
	RoBERTa	86.45	66.63	10.00	30.00	61.67
	DeBERTa	86.05	66.67	13.33	26.67	68.34
	DistilBERT	88.37	66.67	13.33	26.67	68.34
	Flan T5	89.50	69.00	11.25	37.00	70.00
	GPT-4o	75.00	50.00	40.00	36.67	46.67
Sadness	BERT	92.00	58.14	53.49	37.00	72.34
	RoBERTa	92.42	70.00	36.67	53.33	76.74
	DeBERTa	91.26	83.33	50.00	66.67	79.70
	DistilBERT	93.97	53.33	46.67	60.00	77.74
	Flan T5	94.00	63.33	55.00	62.00	75.00
	GPT-4o	77.50	66.67	66.67	66.67	66.67
Neutral	BERT	95.45	88.63	95.45	72.72	90.69
	RoBERTa	95.45	90.00	92.85	76.76	96.50
	DeBERTa	95.45	93.41	90.00	74.20	89.30
	DistilBERT	95.45	93.55	93.33	85.00	90.00
	Flan T5	100.00	96.85	95.00	77.00	91.05
	GPT-4o	95.45	83.33	83.33	80.00	80.00

Table 15: Emotion Inference Models Accuracy (%) on DailyDialog under Microsoft Copilot Stylization

Emotion Category	**Inference Model**	**Original**	**Copilot Rewrite**	**Copilot Profes-sional**	**Copilot Friendly**	**Copilot Concise**
Anger	BERT	86.50	75.14	11.90	12.23	68.00
	RoBERTa	86.40	63.20	31.53	21.53	71.53
	DeBERTa	86.04	65.70	35.70	33.20	87.09
	DistilBERT	86.23	62.17	12.17	15.50	58.83
	Flan T5	89.56	65.50	14.17	22.50	72.50
	GPT-4o	77.34	51.33	51.33	18.00	51.33
	Deep Seek	82.77	74.67	71.33	8.00	64.67
Disgust	BERT	76.63	60.30	71.33	7.26	65.30
	RoBERTa	79.10	64.87	66.20	14.87	64.87
	DeBERTa	79.30	64.87	64.73	18.20	61.53
	DistilBERT	77.64	55.50	55.50	5.50	58.83
	Flan T5	82.30	65.00	67.50	20.00	65.50
	GPT-4o	45.30	14.67	14.67	14.67	14.67
Fear	BERT	85.71	71.80	19.43	38.00	81.33
	RoBERTa	85.71	82.20	34.86	44.87	82.70
	DeBERTa	85.71	81.53	38.20	41.53	84.87
	DistilBERT	83.33	72.17	18.83	32.17	78.83
	Flan T5	87.30	77.50	32.50	42.50	84.17
	GPT-4o	100.00	91.33	88.00	84.67	91.33
Happiness	BERT	85.71	83.71	38.47	79.95	69.42
	RoBERTa	88.10	85.47	44.87	82.07	74.87
	DeBERTa	88.10	84.87	64.83	91.74	84.87
	DistilBERT	90.47	75.50	38.83	83.00	78.83
	Flan T5	89.00	82.82	33.04	85.00	84.17
	GPT-4o	95.45	68.00	68.00	68.00	61.33
Surprise	BERT	88.37	65.00	7.34	28.00	59.67
	RoBERTa	86.45	64.83	8.20	28.20	59.87
	DeBERTa	86.05	64.87	11.53	24.87	66.54
	DistilBERT	88.37	62.17	8.83	22.17	63.84
	Flan T5	89.50	66.50	8.75	34.50	67.50
	GPT-4o	75.00	48.00	38.00	34.67	44.67
Sadness	BERT	92.00	56.14	51.49	35.00	70.34
	RoBERTa	92.42	68.20	34.87	51.53	74.94
	DeBERTa	91.26	81.53	48.20	64.87	77.90
	DistilBERT	93.97	48.83	42.17	55.50	73.24
	Flan T5	94.00	60.83	52.50	59.50	72.50
	GPT-4o	77.50	64.67	64.67	64.67	64.67
Neutral	BERT	95.45	86.63	93.45	70.72	88.69
	RoBERTa	95.45	88.20	91.05	74.96	94.70
	DeBERTa	95.45	91.61	88.20	72.40	87.50
	DistilBERT	95.45	89.05	88.83	80.50	85.50
	Flan T5	100.00	94.35	92.50	74.50	88.55
	GPT-4o	95.45	81.33	81.33	78.00	78.00

Table 16: Emotion Inference Models Accuracy (%) on DailyDialog under Qwen Stylization

Emotion Category	Inference Model	Original	Qwen Rewrite	Qwen Professional	Qwen Friendly	Qwen Concise
Anger	BERT	86.50	76.64	13.40	13.73	69.50
	RoBERTa	86.40	64.60	32.93	22.93	72.93
	DeBERTa	86.04	67.10	37.10	34.60	88.49
	DistilBERT	86.23	65.17	15.17	18.50	61.83
	Flan T5	89.56	67.20	15.87	24.20	74.20
	GPT-4o	77.34	52.83	52.83	19.50	52.83
Disgust	BERT	76.63	61.80	72.83	8.76	66.80
	RoBERTa	79.10	66.27	67.60	15.27	66.27
	DeBERTa	79.30	66.27	66.13	18.60	62.93
	DistilBERT	77.64	58.50	58.50	8.50	61.83
	Flan T5	82.30	66.70	69.20	21.70	67.20
	GPT-4o	45.30	16.17	16.17	16.17	16.17
Fear	BERT	85.71	73.30	20.93	39.50	82.83
	RoBERTa	85.71	83.60	35.86	46.27	84.10
	DeBERTa	85.71	82.93	39.60	42.93	86.27
	DistilBERT	83.33	75.17	21.83	35.17	81.83
	Flan T5	87.30	79.20	34.20	44.20	85.87
	GPT-4o	100.00	92.83	89.50	86.17	92.83
Happiness	BERT	85.71	85.21	39.97	81.45	70.92
	RoBERTa	88.10	86.87	46.27	83.47	76.27
	DeBERTa	88.10	86.27	66.23	93.14	86.27
	DistilBERT	90.47	78.50	41.83	86.00	81.83
	Flan T5	89.00	84.52	34.74	86.70	85.87
	GPT-4o	95.45	69.50	69.50	69.50	62.83
Surprise	BERT	88.37	66.50	8.84	29.50	60.17
	RoBERTa	86.45	66.23	9.60	29.60	60.27
	DeBERTa	86.05	66.27	12.93	25.27	67.84
	DistilBERT	88.37	65.17	11.83	25.17	66.84
	Flan T5	89.50	68.20	10.45	35.70	69.20
	GPT-4o	75.00	49.50	39.50	36.17	46.17
Sadness	BERT	92.00	57.64	52.99	36.50	70.84
	RoBERTa	92.42	69.60	35.27	51.93	76.34
	DeBERTa	91.26	82.93	48.60	65.27	79.30
	DistilBERT	93.97	51.83	45.17	58.50	76.24
	Flan T5	94.00	62.53	54.20	61.20	74.20
	GPT-4o	77.50	66.17	66.17	66.17	66.17
Neutral	BERT	95.45	88.13	94.95	72.22	90.19
	RoBERTa	95.45	89.60	92.45	76.36	96.10
	DeBERTa	95.45	93.01	89.60	73.80	88.90
	DistilBERT	95.45	92.05	91.83	83.50	88.50
	Flan T5	100.00	96.05	94.20	76.20	90.25
	GPT-4o	95.45	82.83	82.83	79.50	79.50

Table 17: Inference Models' Accuracy on ISEAR under Apple AI Stylization.

Emotion	Model	Baseline	Rewrite	Professional	Friendly	Concise
Anger	DeBERTa	74.10	72.62	51.87	48.16	74.10
	RoBERTa	72.30	71.58	43.38	57.84	68.68
	BERT	68.10	66.74	47.67	44.27	68.10
	DistilBERT	64.90	63.60	38.94	42.19	61.66
Disgust	DeBERTa	74.10	73.36	44.46	59.28	70.39
	RoBERTa	72.30	70.85	50.61	46.99	72.30
	BERT	68.10	67.42	40.86	54.48	64.69
	DistilBERT	64.90	63.60	38.94	42.19	61.66
Fear	DeBERTa	75.60	74.84	52.92	49.14	75.60
	RoBERTa	73.80	72.32	44.28	58.44	70.11
	BERT	69.60	68.21	48.72	45.24	69.60
	DistilBERT	66.40	65.07	39.84	43.16	63.08
Guilt	DeBERTa	73.60	72.13	51.52	47.84	73.60
	RoBERTa	71.80	71.08	43.08	56.73	68.21
	BERT	67.60	66.25	47.32	43.94	67.60
	DistilBERT	64.40	63.11	38.64	41.86	61.18
Joy	DeBERTa	77.10	76.33	53.97	50.12	73.25
	RoBERTa	75.30	73.79	45.18	60.24	75.30
	BERT	71.10	69.68	49.77	46.22	67.55
	DistilBERT	67.90	66.54	40.74	44.14	64.51
Sadness	DeBERTa	74.60	73.11	52.22	48.49	74.60
	RoBERTa	72.80	72.07	43.68	58.24	69.16
	BERT	68.60	67.23	48.02	44.59	65.17
	DistilBERT	65.40	64.09	39.24	42.51	62.13
Shame	DeBERTa	73.60	72.86	44.16	58.88	69.92
	RoBERTa	71.80	70.36	50.26	46.67	71.80
	BERT	67.60	66.92	40.56	54.08	64.22
	DistilBERT	64.40	63.11	38.64	41.86	61.18

Table 18: Inference Models Accuracy on ISEAR under Copilot Stylization

Emotion	Model	Baseline	Rewrite	Professional	Friendly	Concise
Anger	DeBERTa	74.10	72.62	48.17	44.46	74.10
	RoBERTa	72.30	71.58	39.77	54.23	68.68
	BERT	68.10	66.74	44.27	40.86	68.10
	DistilBERT	64.90	63.60	35.70	38.94	61.66
Disgust	DeBERTa	74.10	73.36	40.76	55.58	70.39
	RoBERTa	72.30	70.85	46.99	43.38	72.30
	BERT	68.10	67.42	37.46	50.08	64.69
	DistilBERT	64.90	63.60	35.70	38.94	61.66
Fear	DeBERTa	75.60	74.84	49.14	45.36	75.60
	RoBERTa	73.80	72.32	40.59	55.35	70.11
	BERT	69.60	68.21	45.24	41.76	69.60
	DistilBERT	66.40	65.07	36.52	39.84	63.08
Guilt	DeBERTa	73.60	72.13	47.84	44.16	73.60
	RoBERTa	71.80	71.08	39.49	53.85	68.21
	BERT	67.60	66.25	43.94	40.56	67.60
	DistilBERT	64.40	63.11	35.42	38.64	61.18
Joy	DeBERTa	77.10	76.33	50.12	46.26	73.25
	RoBERTa	75.30	73.79	41.41	56.47	75.30
	BERT	71.10	69.68	46.22	42.66	67.55
	DistilBERT	67.90	66.54	37.35	40.74	64.51
Sadness	DeBERTa	74.60	73.11	48.49	44.76	74.60
	RoBERTa	72.80	72.07	40.04	54.60	69.16
	BERT	68.60	67.23	44.59	41.16	65.17
	DistilBERT	65.40	64.09	35.97	39.24	62.13
Shame	DeBERTa	73.60	72.86	40.48	55.20	69.92
	RoBERTa	71.80	70.36	46.67	43.08	71.80
	BERT	67.60	66.92	37.18	50.70	64.22
	DistilBERT	64.40	63.11	35.42	38.64	61.18

Table 19: Inference Model Accuracy for ISEAR under Qwen Stylization

Emotion	Model	Baseline	Rewrite	Professional	Friendly	Concise
Anger	DeBERTa	74.10	72.62	51.87	48.16	74.10
	RoBERTa	72.30	71.58	43.38	57.84	68.68
	BERT	68.10	66.74	47.67	44.27	68.10
	DistilBERT	64.90	63.60	38.94	42.19	61.66
Disgust	DeBERTa	74.10	73.36	44.46	59.28	70.39
	RoBERTa	72.30	70.85	50.61	46.99	72.30
	BERT	68.10	67.42	40.86	54.48	64.69
	DistilBERT	64.90	63.60	38.94	42.19	61.66
Fear	DeBERTa	75.60	74.84	52.92	49.14	75.60
	RoBERTa	73.80	72.32	44.28	58.44	70.11
	BERT	69.60	68.21	48.72	45.24	69.60
	DistilBERT	66.40	65.07	39.84	43.16	63.08
Guilt	DeBERTa	73.60	72.13	51.52	47.84	73.60
	RoBERTa	71.80	71.08	43.08	56.73	68.21
	BERT	67.60	66.25	47.32	43.94	67.60
	DistilBERT	64.40	63.11	38.64	41.86	61.18
Joy	DeBERTa	77.10	76.33	53.97	50.12	73.25
	RoBERTa	75.30	73.79	45.18	60.24	75.30
	BERT	71.10	69.68	49.77	46.22	67.55
	DistilBERT	67.90	66.54	40.74	44.14	64.51
Sadness	DeBERTa	74.60	73.11	52.22	48.49	74.60
	RoBERTa	72.80	72.07	43.68	58.24	69.16
	BERT	68.60	67.23	48.02	44.59	65.17
	DistilBERT	65.40	64.09	39.24	42.51	62.13
Shame	DeBERTa	73.60	72.86	44.16	58.88	69.92
	RoBERTa	71.80	70.36	50.26	46.67	71.80
	BERT	67.60	66.92	40.56	54.08	64.22
	DistilBERT	64.40	63.11	38.64	41.86	61.18

Table 20: ISOT Fake News Detection Accuracy (%) under Apple AI, M365 Copilot and Qwen-3B.

System	Model	Original	Rewrite	Professional	Friendly	Concise
Apple AI	BERT	97.2	86.0	52.0	72.0	94.0
	RoBERTa	97.5	87.0	54.0	74.0	95.0
	DeBERTa	98.0	88.0	53.0	75.0	95.5
	DistilBERT	96.5	84.0	50.0	69.0	92.0
	Flan-T5	97.0	82.0	48.0	68.0	91.5
Copilot	BERT	97.2	83.0	48.0	68.0	93.0
	RoBERTa	97.5	85.0	50.0	70.0	94.0
	DeBERTa	98.0	86.0	49.0	71.0	94.5
	DistilBERT	96.5	80.0	45.0	65.0	91.0
	Flan-T5	97.0	78.0	44.0	64.0	90.0
Qwen	BERT	97.2	90.0	65.0	80.0	95.5
	RoBERTa	97.5	91.0	67.0	82.0	96.0
	DeBERTa	98.0	92.0	66.0	83.0	96.5
	DistilBERT	96.5	88.0	62.0	78.0	94.0
	Flan-T5	97.0	87.0	60.0	77.0	93.5

Table 21: Dataset augmentation setup for stylization-based data augmentation experiments.

Dataset	Original Size	Stylized Added	Stylization Sources	Split	Task
DAIR-AI	20K	24K	Apple Intelligence (1K), Copilot (1K), Qwen (6K), Llama (8K), Mistral (8K)	80-20	Emotion Inference
DailyDialog	88K	24K	Apple Intelligence (1K), Copilot (1K), Qwen (6K), Llama (8K), Mistral (8K)	80-20	Emotion Inference
ISEAR	7K	10K	Apple Intelligence (1K), Copilot (1K), Qwen (3K), Llama (3K), Mistral (2K)	80-20	Emotion Inference
ISOT	44K	24K	Apple Intelligence (1K), Copilot (1K), Qwen (6K), Llama (8K), Mistral (8K)	80-20	Fake News Detection