Affective Bleaching in Out-of-Distribution Evaluation: Why Backtranslation Outperforms Paraphrasing

Anonymous ACL submission

Abstract

While emotion classifiers achieve high accuracy on static benchmarks, their Out-of-Distribution (OOD) robustness to semantic shifts remains a critical blind spot. Local Small Language Models (SLMs) can generate synthetic variations to stress-test these models. However, transforming emotional text risks *affective bleaching*, unintentionally stripping away original emotional intensity and invalidating test samples. We evaluate this vulnerability by comparing direct paraphrasing (PP) against backtranslation (BT) across 22 emotion categories. Evaluating four SLMs, 10 prompts, and two pivot languages, we show that while PP initially appears more stable with fewer invalid candidates, it ultimately fails by eroding the underlying emotional signal. Conversely, despite a higher number of invalid paraphrases, BT leverages a pivot language's constraints as an affective safeguard. With online data increasingly processed and homogenized by language models, classifiers face constant semantic shifts that may degrade their quality.

1 Introduction

Modern text-based emotion classifiers achieve impressive performance on standard, static benchmarks, such as the fine-grained 27 emotions + 1 neutral class GoEmotions (Demszky et al., 2020) dataset. However, this static nature obscures their Out-of-Distribution (OOD) robustness (Shen et al., 2021) to semantic and linguistic shifts. Dynamically rewriting test sets provides a scalable stress-testing mechanism. While commercial Large Language Models (LLMs) can execute this, relying on cloud-based APIs introduces severe methodological and ethical risks. Affective datasets frequently contain sensitive psychological disclosures, demanding zero-data-leakage environments for institutional compliance. Thus, creating machine learning models based on these data and evaluating their generalization abilities is often a challenge.

Furthermore, commercial models exhibit a pervasive sanitization bias driven by harmlessness training; when tasked with rewriting texts conveying intense negative emotions, they may return a more pacified text (or even reject to answer) as they follow human preferences of not generating harmful outputs.

Consequently, local, open-weight Small Language Models (SLMs) offer a critical alternative for generating unconstrained, privacy-preserving emotional perturbations. Yet, utilizing SLMs introduces a central methodological challenge: preserving affective nuance. A valid text transformation must alter linguistic structure without eroding the subtle emotional core. Failure results in affective bleaching. If an SLM inadvertently neutralizes a test sample's emotional intensity during rewriting, the augmented sample loses its original label integrity, invalidating the benchmark before evaluation even begins. In this paper, we demonstrate that backtranslation (BT) is a fundamentally superior evaluation methodology to direct paraphrasing (PP) when utilizing SLMs. Our results expose an Anchor Effect (Litwin, 2025; Tversky and Kahneman, 1974; Epley and Gilovich, 2010): the structural and semantic constraints of a pivot language may act as an affective safeguard, preventing the severe bleaching observed when models are granted the stylistic freedom of direct paraphrasing. By evaluating four SLMs: EuroLLM, Aya Expanse, Ministral 3, and Gemma 4, we show backtranslation consistently yields evaluation samples with significantly higher emotional preservation. By enforcing strict structural filters to remove trivial near-copies and length mismatches, we isolate true semantic robustness from basic generation failures. To establish a framework for OOD emotion evaluation, we address the following research questions: **RQ1: Semantic Robustness**: Which text transformation strategy (BT or PP) better preserves the measured emotional signal required to maintain

label integrity in OOD evaluation? **RQ2: Stress-Test Failure Modes**: What factors drive the generation of invalid test candidates across different prompts and SLMs, and what are the distinct failure characteristics of each rewriting strategy? **RQ3: Boundary of the Anchor Effect**: Finally, to determine if this preservation mechanism is strictly an evaluation-phase artifact or a general property of synthetic text, does the affective advantage of BT translate to training data augmentation?

2 Related work

Emotion Classification and OOD Generalization. Computational emotion modeling is grounded in categorical frameworks (Ekman, 1992; Plutchik, 1980) and dimensional spaces (Russell and Mehrabian, 1977). While early corpora operationalized these theories (Scherer and Wallbott, 1990; Mohammad et al., 2018; Buechel and Hahn, 2017), their limited scale drove the adoption of high-resolution benchmarks like GoEmotions (Demszky et al., 2020). Concurrently, Transformer encoders (Vaswani et al., 2017), evolving from BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) to architectures like DeBERTa (He et al., 2021) and ModernBERT (Warner et al., 2025), have become the standard for text classification. However, evaluating their Out-of-Distribution (OOD) robustness remains a critical challenge (Yang et al., 2023). While methods like synonym substitution (Miao et al., 2020; Zheng et al., 2020), counterfactual human generation (Wang and Culotta, 2020; Kaushik et al., 2021), and distributionally robust optimization (Sagawa et al., 2019) attempt to address distribution shifts, the rapid proliferation of Large Language Model (LLM) generated text (Wang et al., 2024; Xu et al., 2023) demands new, dynamic stress-testing frameworks.

Affective Preservation in Text Transformation. Generating synthetic OOD data via text transformation inherently risks eroding fine-grained affective nuance. Early cross-lingual studies noted that granular emotional details are frequently lost to lexical ambiguity (Kajava et al., 2020), with commercial systems failing to preserve original emotions in over 50% of translations (Qian et al., 2023). When applied to data augmentation, text transformations yield mixed results. While some approaches deliberately paraphrase texts to alter emotional intensity (Xie and Agrawal, 2023), attempts at strict preservation often fail; Troiano et al. (2020) observed significant emotional neutralization following traditional backtranslation (BT).

The Evaluation Gap: SLMs and Intrinsic Fidelity. Recent work by Radliński et al. (2025) demonstrated that backtranslation and paraphrasing via commercial APIs (e.g., ChatGPT (Ouyang et al., 2022)) yield comparable *extrinsic* improvements in downstream classifier training. However, this focus on large, closed-source models leaves local, open-weight Small Language Models (SLMs) under-researched. Closer to our evaluation setup, Wisniewski and Czudy (2026) found that ModernBERT demonstrates high efficacy on backtranslated GoEmotions data. Yet, their analysis of affective robustness was superficial; they failed to account for how frequently these models generate trivial, near-exact copies of the input text, conflating true semantic robustness with basic string similarity. Our work bridges this gap by enforcing strict structural filtering, allowing us to explicitly measure the *intrinsic* affective preservation of true synthetic perturbations generated entirely via local SLMs.

3 Methodology

To address our research questions, our methodology is structured into a shared generation and filtering pipeline, followed by two distinct experimental branches: evaluating a frozen classifier on augmented test sets (RQ1 & RQ2), and training new classifiers on augmented training sets (RQ3).

3.1 Shared Pipeline: Candidate Generation and Filtering

We selected the GoEmotions corpus (Demszky et al., 2020), a large-scale dataset of 57,732 Reddit comments with 27 emotion labels, as our foundational benchmark representing real-world, conversational text. For classification, we utilized ModernBERT. Its adoption of RoPE embeddings (Su et al., 2024) and a diverse 2-trillion-token pretraining corpus improves the context awareness as compared to older encoders (e.g., BERT, RoBERTa). Furthermore, Wisniewski and Czudy (2026) demonstrated ModernBERT's superior efficacy specifically for fine-grained emotion classification on this corpus. Following their research, we replicated their preprocessing pipeline: the original corpus was filtered to remove ambiguous examples, the neutral class, and extreme low-

frequency categories (pride, relief, grief, embarrassment, nervousness). Retaining only annotations with agreement from at least two experts yielded a high-confidence dataset of 36,930 examples across 22 fine-grained emotions. This was partitioned via iterative stratification (Sechidis et al., 2011) into a 29,544-sample training split and a 7,386-sample test split. We subsequently retrained the ModernBERT classifier on the training split, successfully replicating the baseline performance reported by Wisniewski and Czudy (2026). To generate the synthetic perturbations, we evaluated four popular open-weight SLMs (8–9B parameters). To ensure a comprehensive comparison, we selected two translation-optimized architectures (`AyaExpanse-8B` [1], `EuroLLM-9B-IT` [2]) alongside two general-purpose models (`Gemma-4-E4B-IT` [3], `Ministral-3-8B-IT` [4]). Because EuroLLM is focused on European languages, we restricted our backtranslation (BT) pivot languages – languages into which we translate before translating back – to French and Spanish. This deliberate constraint mitigates the risk of generating low-quality candidates due to SLM limitations with underrepresented languages. Finally, using Gemini (Team, 2023), we crafted 10 distinct prompt templates for both backtranslation (BT) and direct paraphrasing (PP) (listed in Table 1). These prompts were dynamically engineered to target various personas, stylistic constraints (e.g., text formality and length), and pragmatic perspectives, ensuring our stress-tests cover a broad semantic surface. We tried to keep the prompts aligned, where e.g., BT and PP prompts number 4 both target formal tone, and number 1 focus on text length.

Having a classifier trained on the original train split, we aimed at modifying test set to introduce distribution shifts. Each of the 7386 test examples were processed by each of the 4 SLMs $\times$ 10 prompts $\times$ two modes (backtranslation or direct paraphrasing). For backtranslation, each candidate was generated using French and Spanish separately.

Generations were executed via vLLM (Kwon et al., 2023) with FP8 quantization and greedy decoding (temperature = 0). Greedy decoding was chosen specifically to ensure reproducibility and to strictly isolate the impact of the prompt templates and the Anchor Effect, removing the confounding variable of random sampling variance.

[1] `CohereLabs/aya-expanse-8b`
[2] `utter-project/EuroLLM-9B-Instruct`
[3] `google/gemma-4-E4B-it`
[4] `mistralai/Ministral-3-8B-Instruct-2512`

3.1.1 Valid Paraphrase Identification

To ensure evaluation is conducted strictly on valid linguistic variations, every generated candidate was subjected to three structural and semantic checks. Candidates were rejected if they were: (i) **Near-copies:** Normalized Levenshtein Distance (NLD) ≤ 0.15; (ii) **Semantically unrelated:** BERTScore (Zhang et al., 2020) <0.8; or (iii) **Length-mismatched:** Word-count ratio (generated/reference) outside the [0.3,2.0] range.

3.1.2 Analyzing Rejection Drivers

To understand the factors driving invalid candidates generation (RQ2), we fit two logistic regression models (for PP and BT) using `statsmodels` (Seabold and Perktold, 2010). The binary outcome $Y_{rejected} \in \{0, 1\}$ was modeled using dummy-coded categorical variables.

For backtranslation, we modeled rejection probability using main effects and interaction terms for emotion, prompt, model, and pivot language.

$$\begin{aligned} logit(P(Y_{\text{rejected}} = 1)) = \beta_0 &+ \beta_1 X_{\text{pivot}} + \beta_2 X_{\text{emo}} + \beta_3 X_{\text{prompt}} \\ &+ \beta_4 X_{\text{model}} + \beta_5 (X_{\text{emo}} \times X_{\text{prompt}}) \\ &+ \beta_6 (X_{\text{emo}} \times X_{\text{model}}) \end{aligned} \tag{1}$$

For paraphrasing, we introduced an analogous model without the pivot language, which was used exclusively for backtranslation.

3.2 Branch A: OOD Evaluation (Test-Set Augmentation)

To determine whether BT or PP better preserves affective signals (RQ1), we evaluated the frozen ModernBERT classifier train on the train split exclusively on the subset of test-set candidates that passed our filtering pipeline. We computed Macro and Micro F_1 scores across the 22 unbalanced emotion classes. To provide a comprehensive view, we report score averages across all prompts, alongside the best, worst, and median prompt configurations.

3.3 Branch B: Training vs. Stress-Testing (Train-Set Augmentation)

To investigate whether the affective preservation advantages of BT persist during model training (RQ3), we reversed the experimental paradigm. Instead of augmenting the test set, we utilized synthetic data to enrich the training distribution to address long-tailed class imbalances, and evaluated

the resulting classifiers on the raw, not-augmented test split. To maintain computational feasibility, we scoped the generation engine for this branch to Gemma-4, executing all 10 prompts and fixing the BT pivot language to Spanish. Critically, while the initial dataset filtration by Wisniewski and Czudy (2026) excised extreme low-frequency outliers to remove unlearnable noise, the retained 22 emotion categories still exhibit the pronounced long-tailed distribution inherent to human affective expression in natural language (e.g., admiration – the most popular emotion has about 1,000 examples in the test set, while remorse, fear, desire or disgust have roughly 200 or less examples (Wisniewski and Czudy, 2026)). A substantial volume gap persists between the dominant majority class and the remaining minority categories. To determine if synthetic data can effectively bridge this remaining gap, we established an enrichment threshold, padding minority classes to a volume equal to $0.6\times$ the size of the the class with the highest number of examples (admiration). We compared two training pipelines: **(1) Baseline Oversampling:** A classifier trained on the raw dataset, utilizing standard random oversampling (duplicating raw examples) to reach the $0.6\times$ threshold. **(2) Augmented Enrichment (BT vs. PP):** A suite of 20 distinct classifiers (related to 10 prompts $\times$ 2 transformation modes) trained on data where minority classes were padded using valid generated text. If the synthetic candidate pool was insufficient, we applied sampling with replacement. In total, 18 out of 22 classes were augmented, with remorse emotion augmented with over 3 times the number of original instances.

3.4 Classification vs. real affective bleaching

Trusting only a classifier's results can lead to flawed judgments. A drop in classifier accuracy could indicate models' inability to learn the concept, or it could indicate that the generation pipeline destroyed the emotional signal (affective bleaching). To empirically disentangle this, we employed `Gemini-3.5-Flash` acting as a zero-shot, Chain-of-Thought proxy oracle (LLM-as-a-judge). The oracle evaluated the emotional intensity of the candidate text (PP or BT) against the original text and expected labels, isolating semantic degradation independently of the target classifier's predictions. While safety alignments (e.g., RLHF) may cause commercial LLMs, such as Gemini, to bleach affect during text generation, using the proxy oracle only for zero-shot classification bypasses the generative safety filters that trigger this bias.

To quantify the structural impact of our generation pipelines, specifically to measure affective bleaching, we structured our test data into evaluation triples consisting of an original reference text, its backtranslated version (BT), and its directly paraphrased version (PP). We then constructed two subsets for oracle evaluation: **The Base Rate Sample**: A random selection of 1,000 triples, drawn regardless of the classifier's labels. Because these examples were chosen without filtering for classifier success or failure, this sample establishes a baseline for how much affective bleaching naturally occurs in both BT and PP generation methods. **The Discrepancy Sample**: A targeted selection of 720 triples designed to isolate the mechanisms of classifier failure. This sample consists of exactly 360 cases where the classifier succeeded on the BT text but failed on the PP text, and 360 cases where it succeeded on the PP text but failed on the BT text. By contrastively analyzing these triples, we evaluate whether affective bleaching is the primary driver of classifier failure when one generation method succeeds and the other does not. Both samples are perfectly balanced across the 4 generation models, the 10 prompt variations, and the pivot languages used for backtranslation.

4 Experiment

4.1 Candidate Validity and Failure Modes

Table 2 reports the proportion of structurally valid candidates (passing all structural and semantic filters) generated by each model, contrasting the overall average across all configurations with the single highest-performing setup.

As anticipated, direct paraphrasing (PP) consistently yields a higher baseline of valid candidates compared to backtranslation (BT). A fine-grained failure analysis (Figure 1) reveals that the primary driver of BT rejection is the generation of near-copies (NLD $\leq$ 0.15). This error accounts for 15.8% of BT failures versus only 3.0% for PP, suggesting models often default to literal, word-for-word translation paths in BT, bypassing structural alteration.

Conversely, a frequent failure mode for PP is excessive verbosity, with 8.0% of PP candidates exceeding the maximum length threshold, compared to just 3.6% for BT. Other structural errors, such as generating excessively short or semantically unrelated texts, were marginal across both

Table 1: Comparison of semantically aligned prompts used for Backtranslation and Paraphrasing.

ID	Backtranslation (Pivots: ES / FR)	Paraphrasing
0	**System:** You are a professional translator. **User:** Translate the following text into [X], preserving all nuances.	**System:** You are an expert editor. **User:** Paraphrase the following text in English, ensuring the meaning remains identical.
1	**System:** You are a precise, structurally focused translator. **User:** Translate the following text into [X]. You must maintain the approximate length, phrasing, and structural rhythm of the original text.	**System:** You are a linguistic assistant. **User:** Rewrite this sentence using different words, but keep it approximately the same length as the original.
2	**System:** You are a translator specializing in literature and dialogue. **User:** Translate this to [X], paying special attention to the emotional intensity and tone.	**System:** You are an expert in affective linguistics. **User:** Paraphrase this text while strictly preserving its original emotional label and intensity.
3	**System:** You are a Reddit user from Europe. **User:** How would you say this Reddit post in [X] language?	**System:** You are a Reddit user. **User:** Rewrite this post in your own words, but keep the exact same feeling and "vibe" of the original.
4	**System:** You are a formal translator. **User:** Provide a precise formal translation of this text into [X].	**System:** You are a formal editor. **User:** Rewrite this text in a slightly more formal English style while keeping the core sentiment.
5	**System:** You are a literary translator specializing in high lexical diversity. **User:** Translate the following text into [X]. Avoid the most basic, literal word choices. Instead, use a rich, diverse vocabulary and creative phrasing while strictly maintaining the underlying emotional state of the original text.	**System:** You are a creative writer. **User:** Paraphrase this text by using as many synonyms as possible without changing the underlying emotion.
6	**System:** You are a professional and objective translator. **User:** Translate this internet comment into [X] exactly as it was written, preserving all of its original emotional undertones.	**System:** You are a helpful and polite AI assistant. **User:** Please rephrase this internet comment in English. Ensure the core information remains the same.
7	**System:** You are a context-aware translator. **User:** Translate this text into [X]. Ensure that the implicit feelings and subtext of the original speaker are maintained in your translation.	**System:** You are an expert in reading comprehension and semantics. **User:** Paraphrase this text in English. Ensure that the implicit feelings and subtext of the original speaker are strictly preserved in your rewrite.
8	**System:** You are an expert in translating informal internet dialogue. **User:** Translate this conversational text into a natural, conversational equivalent in [X], keeping its specific emotional weight.	**System:** You are an expert in internet communication and linguistics. **User:** Rewrite this text in English using different words, but maintain the exact conversational tone and emotional weight of the original.
9	**System:** You are a localization expert. **User:** Provide a slightly more concise translation of this text into [X] without losing any of its original emotional impact.	**System:** You are an expert copywriter. **User:** Rewrite this text to be slightly more concise in English, but you must strictly retain its original emotional impact.

Table 2: Percentage of valid paraphrases generated by SLMs across augmentation modes: overall average (with std. dev.) vs. the single best configuration with prompt ids and ES/FR pivot for BT in square brackets.

	Overall Average		Best Configuration	
Model	**BT**	**PP**	**BT**	**PP**
EuroLLM-9b-Instruct	0.78 (0.06)	**0.87 (0.07)**	0.91 [Prompt 5 / FR]	**0.95 [Prompt 4]**
aya-expanse-8b	0.83 (0.05)	**0.88 (0.07)**	0.92 [Prompt 5 / ES]	**0.97 [Prompt 0]**
gemma-4-E4B-it	0.8 (0.09)	**0.97 (0.03)**	0.93 [Prompt 5 / ES]	**1.0 [Prompt 1]**
Ministral-3-8b-Instruct-2512	0.81 (0.2)	**0.82 (0.15)**	0.94 [Prompt 9 / FR]	**0.98 [Prompt 1]**

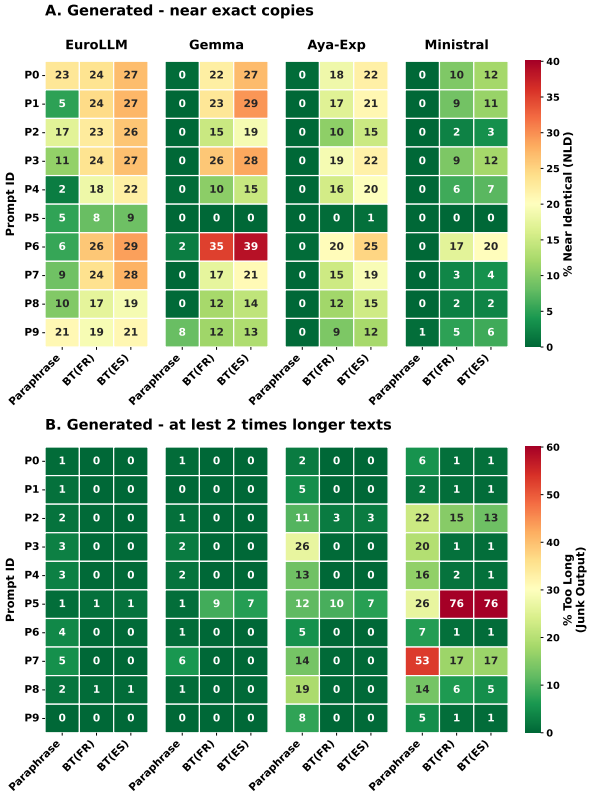
modes (<0.3% combined). Crucially, despite BT's higher average rejection rate, careful configuration selection significantly narrows this performance gap. Under optimal prompt and pivot pairings (e.g., Gemma-4 with Prompt 5 / Spanish pivot), BT candidate validity consistently exceeds 90%, demonstrating that the structural overhead of pivoting is manageable when prompting is optimized.

4.2 Branch A, OOD evaluation

Affective Preservation in OOD Evaluation To isolate the phenomenon of affective bleaching from basic generation errors, this evaluation was conducted exclusively on the subset of synthetic candidates that passed all structural filters. As detailed in Table 3, from the ModernBERT's perspective, Backtranslation (BT) consistently and significantly outperforms direct Paraphrasing (PP) across all SLMs in both Macro- and Micro-F_1 scores (RQ1).

Transformation Mode & Pivot Language Switching from PP to BT yields substantial preservation of the emotional signal. For example, EuroLLM's Macro-F_1 jumps from 0.418 in PP to 0.519 in BT-Spanish. Interestingly, while PP sometimes exhibits lower prompt-to-prompt variance (e.g., Aya's ±0.020 Macro-F_1), its performance ceiling remains drastically lower than BT. Between the two pivot languages, Spanish demonstrates a marginal but consistent advantage over French, achieving the highest mean Macro-F_1 scores across all models without increasing variance.

Figure 1: Heatmap showing per prompt/model and mode percentage of invalid paraphrase reasons.



Model & Prompt Robustness EuroLLM emerges as the most robust architecture for preserving emotion, when considering all prompts, achieving the highest overall Mean Macro-F_1 (0.519) and Micro-F_1 (0.571) via BT-Spanish, alongside the lowest standard deviation. Ministral struggled the most, establishing the lowest performance floor across both modes. Crucially, prompt selection matters. In BT modes, Prompt 5 was overwhelmingly responsible for the worst scores across almost all models. Conversely, Prompts 3 and 6 showed high synergy with BT. In PP mode, Prompt 9 proved highly effective, while Prompts 4 and 7 consistently triggered the worst affective bleaching.

Drivers of Candidate Rejection To analyze the factors driving rejection during generation (RQ2), our logistic regression models isolated distinct trade-offs between configurations and emotional categories. For the paraphrasing (PP) pipeline, Prompt 1 achieved the highest overall yield of valid candidates ($\beta = -1.217, p < .001$), though its rejection rate spiked significantly when processing high-arousal emotions such as Fear and Anger. Conversely, the Gemma-4-E4B-IT architecture demonstrated exceptional stability ($\beta = -1.652, p < .001$), consistently maintaining high validity rates across all emotional distributions without localized anomalies. Within the backtranslation (BT) pipeline, Prompt 9 maximized the overall valid text yield ($\beta = -0.634, p < .001$) but suffered a statistically significant drop in validity when handling Anger. While Prompt 8 yielded slightly fewer valid candidates overall ($\beta = -0.493, p < .001$), it proved highly stable across all emotion categories. Finally, Aya-Expanse emerged as the most structurally robust model for the BT pipeline ($\beta = -0.341, p < .001$), demonstrating consistent validity across all tested inputs regardless of the intermediary pivot language.

4.3 Branch B: Training vs. Stress-Testing

As the aggregate metrics for the train-set augmentation experiment demonstrate (Table 4), enriching the training set with synthetically generated data yields no meaningful improvement over baseline random oversampling (RQ3). Across all configurations, both BT and PP generally resulted in slight performance degradations on the raw test set. Even the optimal synthetic configuration (Best BT via Prompt 5) offered only a negligible +0.003 Macro-F_1 gain over basic oversampling, while failing to

Table 3: Emotion Classification Results (Macro-F1 and Micro-F1) per Model and Mode. **The baseline (trained on unprocessed test set) is: 0.619 F_1-micro, 0.561 F_1-macro**

Model	Mode	Macro-F1				Micro-F1			
		Mean±Std	Worst (P)	Median (P)	Best (P)	Mean±Std	Worst (P)	Median (P)	Best (P)
EuroLLM-9B-Instruct	PP	0.418 ± 0.031	0.359 (4)	0.416 (5)	0.463 (9)	0.478 ± 0.025	0.426 (4)	0.479 (5)	0.513 (0)
	BT-French	0.511 ± 0.016	0.470 (5)	0.521 (1)	0.525 (9)	0.564 ± 0.016	0.525 (5)	0.573 (7)	0.576 (1)
	BT-Spanish	**0.519** ± 0.019	**0.475** (5)	**0.525** (4)	**0.536** (7)	**0.571** ± 0.017	**0.529** (5)	**0.579** (9)	**0.586** (7)
Ministral-3-8B-Instruct-2512	PP	0.335 ± 0.039	0.261 (7)	0.339 (3)	0.400 (9)	0.393 ± 0.044	0.289 (7)	0.396 (3)	0.455 (9)
	BT-French	0.452 ± 0.064	0.283 (5)	0.469 (7)	0.511 (6)	**0.511** ± 0.062	**0.355** (5)	0.519 (7)	**0.573** (6)
	BT-Spanish	**0.455** ± 0.064	**0.286** (5)	**0.479** (4)	0.502 (6)	0.510 ± 0.062	0.349 (5)	**0.529** (4)	0.565 (6)
aya-expanse-8b	PP	0.374 ± 0.020	0.351 (4)	0.369 (5)	0.404 (6)	0.437 ± 0.015	0.416 (1)	0.429 (5)	0.461 (6)
	BT-French	0.503 ± 0.040	**0.392** (5)	0.512 (1)	**0.532** (6)	0.558 ± 0.034	**0.467** (5)	0.568 (1)	**0.586** (3)
	BT-Spanish	**0.505** ± 0.041	0.390 (5)	**0.521** (7)	0.527 (3)	**0.559** ± 0.036	0.459 (5)	**0.570** (4)	0.584 (3)
gemma-4-E4B-it	PP	0.391 ± 0.040	**0.346** (5)	0.373 (1)	0.467 (9)	0.452 ± 0.036	**0.415** (7)	0.433 (2)	0.519 (9)
	BT-French	0.488 ± 0.067	0.300 (5)	0.513 (1)	0.528 (3)	0.548 ± 0.055	0.395 (5)	0.565 (1)	0.583 (3)
	BT-Spanish	**0.499** ± 0.068	0.309 (5)	**0.518** (2)	**0.538** (3)	**0.560** ± 0.056	0.404 (5)	**0.578** (2)	**0.594** (3)

Table 4: Comparison of trainset augmentation methods. Basic oversampling represents oversampling directly from trainset (with returning). From 20 SLM-augmented classifers (Gemma + 10 prompts x 2 modes - BT and PP), we report best, worst, and median scores.

Type	Micro-F1	Macro-F1
Basic oversampl.	0.616	0.571
Best BT	0.614 (P5)	0.574 (P5)
Best PP	0.61 (P5)	0.57 (P5)
Worst BT	0.604 (P8)	0.563 (P1)
Worst PP	0.595 (P4)	0.553 (P0)
Median BT	0.61	0.567
Median PP	0.6	0.56

match its Micro-F_1 baseline (0.616 vs. 0.614). This confirms our hypothesis for RQ3: the affective preservation superiority of backtranslation is exclusively an evaluation-phase phenomenon, not a mechanism for training enhancement. Moving beyond aggregate metrics, a granular, per-emotion analysis corroborates the overarching trend: synthetic enrichment yields negligible improvements across the majority of classes. However, isolating the extremes reveals significant, highly localized performance swings (Table 5). Interestingly, the most substantial performance gains for both augmentation strategies are concentrated in the exact same subset of emotions (*Disgust*, *Excitement*, and *Disappointment*). Conversely, the most severe degradations exhibit structural variance between the two methods. For example, *Annoyance* suffers heavily under BT (dropping 13.98 pp), whereas *Curiosity* and *Desire* see their sharpest declines under PP. Notably, the specific prompts driving these extreme fluctuations vary considerably across different emotions.

4.4 Classification vs. real affective bleaching

Stability of backtranslation

Evaluating BT and PP texts using LLM-as-a-judge on our Base Rate Sample reveals a severe vulnerability in PP mode. The `Gemini-3.5-Flash` model confirmed that 50.2% of random Direct Paraphrases suffered from affective bleaching. In contrast, the Backtranslation pipeline reduced this base rate to 28.6%. This 21.6% absolute reduction provides quantitative validation of the *Anchor Effect*, the structural constraints of the pivot language strictly bound semantic drift.

Disentangling True Brittleness from Generation Failure: Analyzing the 720-item Discrepancy Sample allows us to isolate why the classifier fails. When the classifier failed on PP, the oracle confirmed the text was truly bleached 61.7% of the time. For BT failures, this dropped to 50.8%. However, the inverse of these statistics is equally critical: in 38.3% of PP failures and 49.2% of BT failures, the synthetic text preserved the emotional intensity, yet the classifier still failed. This proves that both generation methods successfully yield true Out-of-Distribution instances that expose genuine brittleness in ModernBERT.

The Masking Effect of Binary Thresholds: Our analysis reveals a critical vulnerability in static classifier evaluation: models frequently succeed for the wrong reasons. In instances where the classifier successfully predicted the target emotion on paraphrased text, the proxy oracle (Gemini) determined the text was actually bleached 38.3% of the time (20.3% for BT). This discrepancy may stem from a misalignment between categorical classification thresholds and continuous emotional intensity. Because binary metrics (like Macro-F_1) evaluate outputs against a threshold (usually 0.5), models can output a correct prediction by latching onto residual lexical artifacts, masking a steep drop in

Table 5: Highest observed performance swings (Δ F1 in percentage points) per emotion relative to the baseline oversampling. Dashes indicate the emotion was not among the top three degradations for the augmentation method.

	Backtranslation (BT)		**Paraphrasing (PP)**	
Emotion (Baseline F_1)	**Prompt**	**Δ F_1 (pp)**	**Prompt**	**Δ F_1 (pp)**
Highest F_1 Gains				
Disgust (0.426)	P2	+7.65	P9	+4.94
Excitement (0.39)	P2	+6.05	P8	+4.88
Disappointment (0.279)	P2	+5.99	P8	+6.69
Highest F_1 Degradations				
Annoyance (0.403)	P2	-13.98	–	–
Remorse (0.672)	P1	-7.69	P0	-7.78
Sadness (0.582)	P2	-5.61	–	–
Curiosity (68.3)	–	–	P9	-8.33
Desire (57.5)	–	–	P4	-7.08

underlying predictive confidence.

5 Discussion

The Illusion of Paraphrasing Stability Our findings show that direct paraphrasing offers an illusion of stability, producing structurally sound, fluent texts that easily pass basic validation filters. However, this fluency comes at the cost of affective bleaching. If researchers rely on direct paraphrasing to stress-test emotion classifiers, they risk optimizing for a sanitized, homogenized representation of human emotion rather than semantic robustness.

The Value of the Anchor Effect By forcing the model to decode through an intermediary linguistic space, backtranslation seem to disrupt this sanitization process, preserving the core affective intensity. While this process is yielding a higher rate of near-copies, these errors can be easily filtered. As online discourse is increasingly processed, translated, and rewritten by LLMs, such synthetic stress-tests are vital to ensure classifiers maintain their accuracy.

The Boundaries of Synthetic Augmentation Counterintuitively, the affective fidelity that makes backtranslation an exceptional evaluation tool does not translate into a training advantage (RQ3). The localized, extreme fluctuations in performance across specific emotion categories suggest that while synthetic data preserves the macro-label, it simultaneously introduces subtle, prompt-specific artifacts. During evaluation, a robust classifier can generalize through these artifacts. During training, however, the model risks overfitting to these synthetic boundaries rather than learning the underlying emotional taxonomy, explaining why basic oversampling remains the superior methodology for handling class imbalance.

The Anchor Effect The superiority of BT stems from the mechanical constraints of the *Anchor Effect*. Translation forces a rigid, phrase-to-phrase mapping, compelling the model to retain original affective valence. Conversely, paraphrasing operates as abstract semantic summarization, granting excessive stylistic freedom. This allows pre-trained safety alignments to aggressively regress the text toward a sterilized, affectively bleached mean.

6 Conclusions

In this study, we established that backtranslation (BT) is fundamentally superior to direct paraphrasing (PP) for generating synthetic Out-of-Distribution (OOD) emotionally charged data for fine-grained emotion classifiers stress-testing. Addressing RQ1, we demonstrated that the constraints of a pivot language introduce an *Anchor Effect*, acting as a necessary safeguard against the severe affective bleaching inherent to direct paraphrasing. Regarding RQ2, our failure analysis revealed a critical trade-off: while PP produces fewer structurally rejected candidates, it compromises benchmark validity by eroding emotional intensity. Conversely, BT ensures that surviving candidates retain strict label integrity. Finally (RQ3), the negligible training improvements observed from BT augmentation align with the structural constraints of the Anchor Effect. Because the pivot language heavily restricts stylistic deviation, the resulting synthetic samples likely lack the broad lexical diversity required to meaningfully expand the classifier's decision boundaries. These results suggest that BT's primary utility for local SLMs lies in diagnostic stress-testing rather than training enhancement.

Limitations

Gemini prompt selection To evaluate affective bleaching in LLM-as-a-judge scenario, we used a single prompt to obtain results due to high costs of processing 1720 $\times$ 2 (BT / PP) sentences[5]. However, the prompt was a detailed instruction, always applied to both PP candidate and BT candidate for a given model, and one of the 10 generation prompts. As generation prompts were semantically aligned between PP and BT, a comparison between PP and BT candidates could be possible. For each analyzed triple (reference text, PP text, BT text), and expected as well as generated labels, we used the prompt described below to grade both PP text and PP text in the context of reference text and labels.

System: `You are an expert linguist and annotator specializing in Affective Computing and emotion classification. Your task is to detect "affective bleaching" in text transformations. Affective bleaching occurs when a generated text (the candidate) loses, significantly dilutes, or neutralizes the specific target emotion present in the original text. Rules for annotation: 1. Focus STRICTLY on the provided Target Emotion. Ignore other emotions that might be present. 2. A text is "Bleached" (True) if the candidate text sounds notably less intense, more generic, or completely devoid of the Target Emotion compared to the original. 3. A text is NOT "Bleached" (False) if the candidate text maintains or increases the intensity of the Target Emotion. 4. Minor vocabulary changes do not count as bleaching unless they reduce the emotional impact.You must format your response as a valid JSON object with two keys: - "reasoning": A brief 1-2 sentence explanation comparing the emotional intensity of both texts. - "is_bleached": A boolean (true or false).`

User: `Target Emotions: {target_emotion} Original Text: "{original_text}" Candidate Text: "{candidate_text}". Evaluate if the Candidate Text suffers from affective bleaching regarding the Target Emotion. Output your evaluation in the required JSON format.`

[5]The cost of this experiment was roughly 20 USD. For other experiments in this paper, we used 1xNVIDIA RTX 4090 for about 12 hours.

Candidate Survivorship Bias Our experimental design introduces a survivorship bias by evaluating exclusively the synthetic candidates that pass the structural and semantic filters. Because direct paraphrasing (PP) and backtranslation (BT) exhibit different drop-out rates, comparing their surviving subsets does not account for the total *yield cost* of each method. However, this filtering is a deliberate and practically necessary methodological constraint. Because unconstrained SLMs inherently generate structural failures, such as hallucinations, repetitive loops, or severe length mismatches, aggressive post-hoc filtering is a prerequisite for any viable augmentation pipeline. By restricting our evaluation to structurally valid texts, we explicitly decouple baseline generation errors from our core variable of interest. This ensures that the severe affective bleaching observed in PP is definitively a property of the transformation itself, rather than a confounding artifact of poor text generation. While future work might explore end-to-end metrics that penalize initial generation failures (akin to an intention-to-treat analysis), isolating the valid candidates remains essential for accurately measuring intrinsic emotional preservation.

Methodological Scope (Models, Pivots, and Prompts) To maintain computational feasibility, our study is explicitly constrained to local Small Language Models ($\leq$ 10B parameters). Intermediate-scale architectures (e.g., 14B–70B parameters) and large commercial APIs remain unexplored in this specific research. Furthermore, our backtranslation pivot languages were limited to French and Spanish, a constraint dictated by the official multilingual capabilities of the selected SLMs (for instance, outside of European languages, Ministral primarily supports only Chinese and Korean). Because the evaluated models are natively competent in specific high-resource languages, the choice of using French and Spanish limits the chance that affective bleaching stems from translation errors or hallucination common when forcing SLMs into unsupported or weakly supported linguistic spaces. Finally, while we engineered 10 diverse prompts to capture a wide array of stylistic variations, the prompt space is theoretically infinite, and alternative instructional paradigms could yield different generation behaviors.

Annoyance quality drop in BT While backtranslation acts as a robust safeguard for most emotions, it is limited when evaluating classes heavily reliant on language-specific pragmatics. For instance, we observed significant degradation in the *Annoyance* category. *Annoyance* frequently utilizes subtle, passive-aggressive English phrasing and syntactic sarcasm that lack direct, one-to-one structural equivalents in French or Spanish. Consequently, the rigid mapping enforced by the pivot language strips away the pragmatic cues essential for classifier activation. Future work must consider how cross-lingual stress tests handle these highly culture-bound emotional nuances.

References

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Nicholas Epley and Thomas Gilovich. 2010. Anchoring unbound. *Journal of Consumer Psychology*, 20(1):20–24.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Kaisla Kajava, Emily Öhman, Piao Hui, and Jörg Tiedemann. 2020. Emotion preservation in translation: Evaluating datasets for annotation projection. In *Digital Humanities in the Nordic Countries*, pages 38–50. CEUR.

Divyansh Kaushik, Amrith Setlur, Eduard H. Hovy, and Zachary Chase Lipton. 2021. Explaining the efficacy of counterfactually augmented data. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Maciej Litwin. 2025. Translation and anchoring bias: A study of equivalence in english-polish student translation tasks. *Linguodidactica*, 29:129–142.

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

Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan. 2020. Snippext: Semi-supervised opinion mining with augmented data. In *Proceedings of The Web Conference 2020*, WWW '20, page 617–628, New York, NY, USA. Association for Computing Machinery.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Shenbin Qian, Constantin Orasan, Felix Do Carmo, Qiuliang Li, and Diptesh Kanojia. 2023. Evaluation of Chinese-English machine translation of emotion-loaded microblog texts: A human annotated dataset for the quality assessment of emotion translation. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 125–135, Tampere, Finland. European Association for Machine Translation.

Łukasz Radliński, Mateusz Guściora, and Jan Kocoń. 2025. Backtranslation and paraphrasing in the llm era? comparing data augmentation methods for emotion classification. In *International Conference on Computational Science*, pages 3–17. Springer.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731.

KR Scherer and H Wallbott. 1990. International survey on emotion antecedents and reactions (isear).

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *SciPy*.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis P. Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III*, Lecture Notes in Computer Science, pages 145–158. Springer.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *CoRR*, abs/2108.13624.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2020. Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th international conference on computational linguistics*, pages 4340–4354.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xing Xie. 2024. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *IEEE Data Eng. Bull.*, 48(1):48–62.

Zhao Wang and Aron Culotta. 2020. Identifying spurious correlations for robust text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, Findings of ACL, pages 3431–3440. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547.

Dawid Wisniewski and Igor Czudy. 2026. Beyond semantics: Measuring fine-grained emotion preservation in small language model-based machine translation. *Preprint*, arXiv:2604.27920.

Justin J. Xie and Ameeta Agrawal. 2023. Emotion and sentiment guided paraphrasing. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, WASSA@ACL 2023, Toronto, Canada, July 14, 2023*, pages 58–70. Association for Computational Linguistics.

Albert Xu, Xiang Ren, and Robin Jia. 2023. Contrastive novelty-augmented learning: Anticipating outliers with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11778–11801, Toronto, Canada. Association for Computational Linguistics.

Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. 2023. Out-of-distribution generalization in natural language processing: Past, present, and future. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4533–4559, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-domain detection for natural language understanding in dialog systems. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:1198–1209.