Beyond Semantics: Measuring Fine-Grained Emotion Preservation in Small Language Model-Based Machine Translation

Dawid Wiśniewski
Poznań University of Technology
dawid.wisniewski@cs.put.poznan.pl

Igor Czudy
Poznań University of Technology

Abstract

Preserving affective nuance remains a challenge in Machine Translation (MT), where semantic equivalence often takes precedence over emotional fidelity. This paper evaluates the performance of three state-of-the-art Small Language Models (SLMs): EuroLLM, Aya Expanse, and Gemma, in maintaining fine-grained emotions during backtranslation. Using the GoEmotions dataset, which comprises Reddit comments across 28 distinct categories, we assess emotional preservation across five European languages: German, French, Spanish, Italian, and Polish. Specifically, we investigate (i) the inherent capability of these SLMs to retain emotional sentiment, (ii) the efficacy of emotion-aware prompting in improving preservation, and (iii) the performance of ModernBERT as a contemporary alternative to BERT for emotion classification in MT evaluation[1].

1 Introduction

The primary objective of Machine Translation (MT) has traditionally been the preservation of semantic equivalence, ensuring that the *what* of a message is accurately conveyed across linguistic boundaries. However, as Large Language Models (LLMs) increasingly mediate human communication through chatbots, virtual assistants, and localized content, the *how* of a message, its emotional resonance and affective nuance – has become equally critical.

Despite the rapid progress in neural MT, preserving fine-grained emotions remained a formidable challenge (Lohar et al., 2018). Languages encode affect through diverse linguistic mechanisms, ranging from specific lexical choices and modal particles to complex syntactic shifts. These nuances are often lost in translation pipelines that prioritize word-for-word or purely semantic accuracy.

The current landscape of Natural Language Processing is witnessing a shift toward Large Language Models (LLMs) and, increasingly, Small Language Models (SLMs) designed for localized deployment. While massive models like Gemini or GPT dominate general benchmarks, SLMs (typically under 10 billion parameters) offer sustainable, efficient, and often more specialized alternatives for regional or task-oriented applications (Van Nguyen et al., 2025). In the European context, the emergence of smaller models like **EuroLLM** (Martins et al., 2025), optimized for EU languages, presents a unique opportunity to study affective transfer in a multi-family linguistic environment (Slavic, Germanic, and Romance). Simultaneously, models like **Aya Expanse** (Dang et al., 2024), developed with a focus on massive multilingual instruction tuning, represent a more global-focused approach to cross-lingual understanding.

In this paper, we evaluate the ability of **EuroLLM** (Martins et al., 2025), **Aya Expanse** (Dang et al., 2024), and Google's **Gemma** (Mesnard et al., 2024) to transfer emotional content from English into German, French, Polish, Spanish, and Italian followed by backtrans-

[1]Project supported by grant no. 0311/SBAD/0763 - Mloda Kadra financed by Poznan University of Technology.

lation to English. We leverage the GoEmotions dataset (Demszky et al., 2020), which provides a high-resolution taxonomy of 28 distinct categories (27 emotion labels and a neutral class). This allows us to move beyond binary sentiment (positive/negative), into the territory of complex human states such as remorse, pride, and curiosity.

Given that LLM outputs are highly sensitive to input configurations, a key component of this research is investigating whether emotion-aware prompting, providing explicit instructions to prioritize emotional preservation, significantly enhances model performance. Finally, to evaluate the stability of emotional signals in backtranslated text, we benchmark contemporary classifiers, comparing traditional encoders like BERT (Devlin et al., 2019) and DeBERTa v3 (He et al., 2021) against the recently introduced **ModernBERT** (Warner et al., 2025).

Our research is guided by the following core research questions:

- **RQ1**: How do various SLM architectures compare in their ability to preserve fine-grained emotions during the translation process?
- **RQ2**: Which specific emotional categories are most susceptible to degradation during translation?
- **RQ3**: To what extent does explicit emotion-aware prompting improve the emotional fidelity of SLM-generated translations?
- **RQ4**: Does ModernBERT offer better classification quality for detecting emotional content compared to established encoder models like BERT and DeBERTa?

2 Related work

Emotion preservation in MT The preservation of emotional fidelity is a significant area of inquiry within Machine Translation (MT). Recent literature has explored this challenge through various lenses, ranging from early neural architectures to modern generative models.

Early investigations into affective loss often utilized backtranslation as a diagnostic tool. (Troiano et al., 2020) evaluated emotional signal degradation in English-to-German and English-to-Russian pipelines. While they observed a notable loss of affective information, their study was limited to WMT'19 FAIRSEQ models (Ott et al., 2019), GloVe embedding-based classification (Pennington et al., 2014), and seven core emotions (anger, disgust, fear, guilt, joy, sadness, and shame), leaving the performance of modern autoregressive models on a wider emotional spectrum largely unexplored. Similarly, (Kajava et al., 2020) examined the viability of annotation projection, noting that while sentiment is generally preserved, more nuanced emotional information is frequently lost due to incomplete translations or lexical ambiguity in the target language.

More recent studies have highlighted the persistent difficulty of maintaining nuance in commercial systems. (Qian et al., 2023) demonstrated that Google Translate failed to preserve original emotions in over 50% of English-to-Chinese translations, identifying polysemous words and negations as primary error drivers. To mitigate losses, (Brazier and Rouas, 2024) explored multimodal approaches, enriching LLM inputs with features from Speech Emotion Recognition (SER) models (Wagner et al., 2023). Their findings suggest that conditioning translations on affective variables, particularly arousal, yields measurable gains in translation quality.

The recent dominance of Large Language Models (LLMs), such as Google's Gemini (Team et al., 2023), Anthropic's Claude (Caruccio et al., 2024), and OpenAI's GPT-4o (Singh et al., 2025), has shifted the MT paradigm. Findings from the Workshop on Machine Translation (WMT 2025) confirm that these general-purpose models often surpass specialized NMT systems; notably, Gemini 2.5 Pro achieved state-of-the-art results across multiple language pairs despite not being exclusively trained for translation (Kocmi et al., 2025).

The specific capacity of Small Language Models (SLMs) to manage fine-grained affective transfer – especially when guided by emotion-aware prompting – remains a critical research gap that this paper aims to address.

Small language models & MT The trajectory of Large Language Model (LLM) research has recently bifurcated, with significant momentum shifting toward Small Language Models (SLMs) optimized for edge deployment and computational efficiency. This paradigm shift is driven by the realization that high-quality data curation and advanced distillation techniques enable models with significantly fewer parameters to rival the per-

formance of massive architectures while remaining compatible with consumer-grade hardware. This accessibility facilitates localized fine-tuning and greater granular control over model behavior (Van Nguyen et al., 2025).

Recent releases have defined the current SLM landscape. Meta's Llama 3.2 family (LlamaTeam, 2024) introduced ultra-lightweight 1B and 3B variants specifically designed for mobile and edge platforms. These models utilize structural pruning and knowledge distillation from larger Llama counterparts to maintain sophisticated reasoning capabilities within a compact footprint. Similarly, Google's Gemma 2 series (Mesnard et al., 2024) and particularly the 2B and 9B versions leverage a distillation-heavy training objective to achieve an exceptional performance-to-parameter ratio, often outperforming much larger models on general reasoning benchmarks.

Multilingual accessibility has been a primary driver for regional SLM development. The EuroLLM project (Martins et al., 2025) focuses on providing native support for 35 languages, with a specific emphasis on the official languages of the European Union. By balancing data mixtures across diverse linguistic datasets, **EuroLLM** versions effectively mitigate the English-centric bias prevalent in earlier architectures. In a similar vein, Cohere's Aya Expanse (8B and 32B) (Dang et al., 2024) employs multilingual arbitrage, preference training, and model merging to deliver state-of-the-art performance across 23 languages. Furthermore, DeepSeek-V3 (Liu et al., 2025) represents an effort to harmonize high computational efficiency with superior agentic performance through innovations such as DeepSeek Sparse Attention (DSA) and a large-scale agentic task synthesis pipeline.

Beyond general-purpose reasoning, these SLMs have demonstrated high proficiency in specialized linguistic tasks, including high-fidelity machine translation (Song et al., 2025) and automated grammatical error correction (Wiśniewski et al., 2025). To facilitate the practical deployment of these models on resource-constrained hardware, quantization remains essential. Activation-aware Weight Quantization (AWQ) (Lin et al., 2024) has emerged as a preferred method for compressing SLMs to 4-bit precision. Unlike traditional uniform quantization, AWQ identifies salient weights – those corresponding to high-magnitude activations, and protects them from aggressive quantization errors, thereby preserving model performance. Other widely adopted formats, such as GGUF, BitsAndBytes, and GPTQ (Rajput and Sharma, 2024), provide additional trade-offs between universal compatibility and raw inference throughput.

Emotion taxonomies and datasets The computational modeling of emotion in text is fundamentally grounded in two primary psychological frameworks: categorical and dimensional. Historically, Ekman's model (Ekman, 1992) has served as the foundational categorical baseline, identifying six universal emotions: anger, disgust, fear, happiness, sadness, and surprise. While this simplicity is advantageous for broad sentiment analysis tasks, it often lacks the granularity required to capture the complexities of human expression. Conversely, Plutchik's model (Plutchik, 1980) proposes a hierarchical "wheel" of emotions that accounts for varying affective intensities and polarities. Complementing these are dimensional frameworks, such as the Valence-Arousal-Dominance (VAD) model (Russell and Mehrabian, 1977), which represent emotional states as continuous coordinates in a multi-dimensional vector space rather than as discrete labels.

The operationalization of these theories within the Natural Language Processing (NLP) domain has been facilitated by a diverse array of benchmark datasets. Early efforts, such as the International Survey on Emotion Antecedents and Reactions (ISEAR) (Scherer and Wallbott, 1990), provide high-quality, self-reported affective descriptions across seven categories; however, its limited scale poses challenges for training modern deep learning architectures. To capture the informal and dynamic nature of contemporary digital communication, datasets like SemEval-2018 Task 1 (Affect in Tweets) (Mohammad et al., 2018) introduced multi-label emotion intensity tasks, enabling models to predict co-occurring affective states. For dimensional analysis, EmoBank (Buechel and Hahn, 2017) remains a critical resource, mapping 10,000 sentences directly to the VAD space.

Recently, the GoEmotions dataset (Demszky et al., 2020) has emerged as a good choice for fine-grained affective classification. By labeling 58,000 Reddit comments across 27 distinct emotional categories accompanied by a neutral class, it allows Small Language Models (SLMs) to distinguish between subtle emotional states, such as remorse ver-

sus grief or admiration versus love, that remained conflated in traditional 6-class models. This high-resolution taxonomy is particularly suited for evaluating the affective bleaching or shifts that may occur during the machine translation process.

3 Dataset

We utilize the GoEmotions dataset (Demszky et al., 2020), a manually annotated corpus comprising 57,732 English Reddit comments. The dataset employs a fine-grained taxonomy of 27 emotion labels plus a neutral class. Each instance was reviewed by 3 to 5 annotators, with additional metadata identifying examples deemed ambiguous by the experts.

Dataset preprocessing To optimize the dataset for training robust emotion classifiers, we implemented a filtering and refinement procedure:

1. **Noise Reduction:** We excluded all examples marked as "ambiguous" by at least one annotator to ensure high-confidence ground truth labels.

2. **Label Refinement:** We removed the *neutral* class to focus the study specifically on active emotional transfer. For the remaining 27 categories, we applied a consensus-based merging strategy, retaining only those labels identified by at least two annotators.

3. **Class Balancing:** To mitigate severe class imbalance, where frequent labels like *admiration* outnumber sparse labels like *grief* by a factor of 30, we removed categories with a representation lower than one standard deviation below the mean class frequency.

4. **Stratified Splitting:** The resulting data was partitioned into training (80%) and test (20%) sets. We employed the Iterative Stratification algorithm (Sechidis et al., 2011) to ensure that the complex multi-label emotion distribution remained consistent across both subsets.

After preprocessing, the final dataset consisted of 29,544 training and 7,386 test examples across 22 emotion categories. Five emotions (*pride, relief, grief, embarrassment, and nervousness*) were excluded due to insufficient sample sizes. The final test set distribution is visualized in Figure 1.

4 Methodology & Experimental setup

Our experimental framework is designed to measure how effectively Small Language Models (SLMs) preserve the 22 identified emotions through a round-trip translation process.

4.1 Backtranslation Pipeline

We perform backtranslation (English $\rightarrow$ Target $\rightarrow$ English) across five pivot languages: German, French, Spanish, Italian, and Polish. For the translation and backtranslation tasks, we compare three state-of-the-art SLMs:

- **EuroLLM-9B-Instruct-AWQ** (Martins et al., 2025) [2]
- **Aya-Expanse-8B-AWQ** (Dang et al., 2024) [3]
- **Gemma-2-9B-IT-AWQ** (Mesnard et al., 2024) [4]

To accommodate hardware constraints (24,GB VRAM), all models were deployed using 4-bit Activation-aware Weight Quantization (AWQ). We utilized the vLLM engine (Kwon et al., 2023) with greedy decoding (temperature = 0) to ensure deterministic and reproducible outputs.

4.2 Prompt Engineering

To test the impact of instruction sensitivity on affective preservation, we evaluated two zero-shot prompt configurations:

1. **Basic (P_{base}):** *Translate the following text from LANG1 to LANG2.\n\n TEXT*
2. **Emotion-Aware (P_{emo}):** *Translate the following text from LANG1 to LANG2. Please focus on preserving the emotions, tone, and intensity of the original text.\n\n TEXT*

Since models often append comments to their responses when using complex prompts, we search for double newlines, a common separator for such additions, and trim them as well as the content provided after those markers.

This setup yields a comprehensive evaluation matrix of 30 backtranslated test set variants (3 models × 5 languages × 2 prompts).

[2] https://huggingface.co/stelterlab/EuroLLM-9B-Instruct-AWQ
[3] https://huggingface.co/Orion-zhen/aya-expanse-8b-AWQ
[4] https://huggingface.co/solidrust/gemma-2-9b-it-AWQ

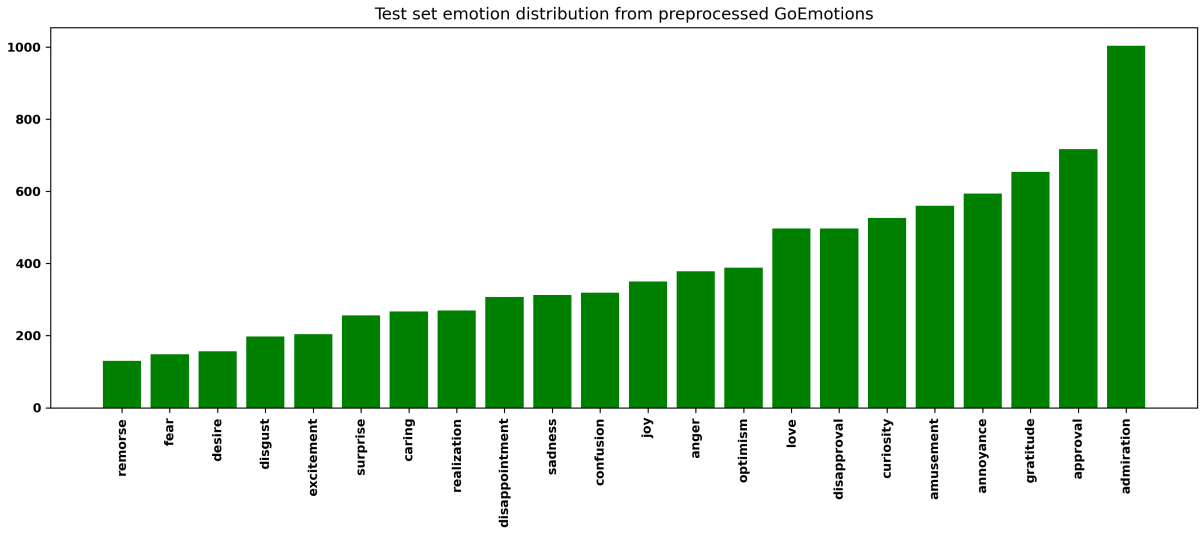

Figure 1: Emotions distribution in the preprocessed testset, which represents 20% of the full dataset.

4.3 Emotion Classification as Evaluation

To quantify emotional loss, we fine-tune three encoder architectures the original English training set:

- `BERT base cased` (Devlin et al., 2019) [5]
- `DeBERTa-v3 base` (He et al., 2021) [6]
- `ModernBERT base` (Warner et al., 2025) [7]

These models serve as our ground truth classifiers. We then evaluate their performance on the backtranslated variants, interpreting any drop in F_1-score relative to the original test set as evidence of affective degradation.

Those models are of similar size, with BERT, DeBERTa, and ModernBERT having 110M, 184M, and 149M parameters, respectively.

The models were trained with the same hyperparameters:

- `num_train_epochs=10`,
- `problem_type` set to `multi_label_classification`,
- `batch_size=32`.

We applied early stopping with patience set to 1, evaluation strategy set to epoch, and micro-F_1 score for early stopping/best model selection.

4.4 Evaluation metrics

Let:

- D be the original English test set with gold-standard labels.
- L be the set of pivot languages ($\{$German, French, Spanish, Italian, Polish$\}$).
- $BT(D, M, \pi, \ell)$ represent the backtranslated version of D generated by model M using prompt π via pivot language $\ell \in L$.
- Φ be a fixed emotion classifier trained on the training set.
- $F_1(\Phi, X)$ denote the macro-averaged F_1 score of classifier Φ on dataset X relative to the original gold labels.

The *Affective Drop* Δ_{aff} for a specific model-prompt-classifier configuration (M, π, Φ) measures the loss of F_1 score averaged over all languages as compared to the reference score and is defined as:

$$\Delta_{\text{aff}}(M, \pi, \Phi) = F_1(\Phi, D) - \frac{1}{|L|} \sum_{\ell \in L} F_1(\Phi, BT(D, M, \pi, \ell)) \quad (1)$$

Interpretation:

[5] `https://huggingface.co/google-bert/bert-base-cased`
[6] `https://huggingface.co/microsoft/deberta-v3-base`
[7] `https://huggingface.co/answerdotai/ModernBERT-base`

- $\Delta_{\text{aff}} \approx 0$: Indicates near-perfect preservation of affective signals; the translation process did not significantly alter the emotional features recognized by the classifier.

- $\Delta_{\text{aff}} > 0$: Represents a loss of emotional fidelity, where higher values indicate greater "affective bleaching" or the introduction of emotional noise during translation.

We can also introduce a per-emotion affective drop Δ^e_{aff} calculated for the emotion e as follows:

$$\Delta^e_{\text{aff}}(M, \pi, \Phi) = F^e_1(\Phi, D) - \frac{1}{|L|} \sum_{\ell \in L} F^e_1(\Phi, BT(D, M, \pi, \ell)) \quad (2)$$

, where F^e_1 is the F_1-score for the class e.

5 Results and Analysis

5.1 Aggregate Performance and Classifier Sensitivity

The results of our backtranslation experiments are summarized in Table 1. Each entry represents the average affective drop (Δ_{aff}) for a specific combination of SLM, emotion classifier, and prompt type. As Δ_{aff} calculates the average drop over languages and emotion categories, it gives a general perspective on SLMs, classifiers, and prompts.

Overall, the analyzed models exhibit robust performance in preserving affective nuance, with average Δ_{aff} scores ranging from 2.89 to 4.93 percentage points. Despite these relatively low drops, we observe distinct performance hierarchies across both the translation models and the evaluation classifiers.

5.2 Comparative Analysis of SLMs and Classifiers

Our analysis reveals a consistent performance gradient among the tested models. Specifically:

- **SLM Hierarchy: EuroLLM** consistently outperforms both Aya and Gemma across all configurations. When comparing results using identical prompts and classifiers, we observe a stable ranking:
  EuroLLM $>$ Aya $>$ Gemma
  The best-performing configuration overall, the one achieving the lowest affective drop utilizes EuroLLM with the emotion-aware prompt (P_{emo}) evaluated via ModernBERT.

- **Classifier Robustness:** There is a visible difference in how the underlying encoder architectures perceive affective degradation. **ModernBERT** consistently yields the lowest Δ_{aff}, followed by DeBERTa-v3 and BERT. This may suggest that ModernBERT's architecture is better suited for detecting fine-grained emotional consistency in MT evaluation.

5.3 The Impact of Emotion-Aware Prompting

One of the most notable findings of this study is the marginal impact of explicit emotional instructions on backtranslation quality.

Classification Robustness: As shown in Table 4, the introduction of P_{emo} has a negligible effect on the final F_1 scores. While EuroLLM shows a slight, consistent improvement (a smaller drop) when using P_{emo}, the gains are statistically marginal. For Aya and Gemma, the basic prompt (P_{base}) actually yielded superior results. This suggests that contemporary SLMs inherently prioritize emotional preservation during translation, and explicit prompting may, in some cases, introduce noise or over-correction.

Semantic Translation Quality: To ensure that emotional preservation does not come at the cost of semantic accuracy, we evaluated the backtranslated outputs using COMET-22-da (Rei et al., 2022). The results indicate high translation fidelity across all models. EuroLLM achieved the highest scores when averaged over all pivot languages (0.8721 for P_{emo} ; 0.872 for P_{base}), with Aya (0.8582 for P_{emo}; 0.8641 for P_{base}) and Gemma (0.8556 for P_{emo}; 0.8638 for P_{base}) following closely. Critically, the delta between P_{emo} and P_{base} for each model was negligible, confirming that emotion-aware prompting does not degrade general translation quality. The detailed scores are provided in Appendix B, in Table 6.

Lexical Variance: Despite the minimal change in classification and COMET scores, the lexical composition of the translations changed significantly between P_{base} and P_{emo} . As detailed in Table 3, between 27.5% and 71.0% of the generated texts were lexically distinct when the prompt was changed. This variance was most pronounced in Polish, indicating that while the emotional signal remains stable, the models utilize significantly different lexical strategies to convey that signal when prompted to focus on tone and intensity.

Table 1: Average affective drop (the lower the better) Δ_{aff} as introduced in Section 4.4.

model	prompt	classifier	Δ_{aff}	rank
euro	emo	modernbert	0.0289	1
euro	basic	modernbert	0.0295	2
euro	emo	deberta	0.0333	3
euro	basic	deberta	0.0338	4
aya	basic	modernbert	0.0343	5
gemma	basic	modernbert	0.0364	6
aya	emo	modernbert	0.0367	7
euro	emo	bert	0.0371	8
euro	basic	bert	0.0382	9
aya	basic	deberta	0.0386	10
gemma	basic	deberta	0.0401	11
aya	emo	deberta	0.0403	12
gemma	emo	modernbert	0.0431	13
aya	basic	bert	0.0435	14
aya	emo	bert	0.0443	15
gemma	basic	bert	0.0448	16
gemma	emo	deberta	0.0484	17
gemma	emo	bert	0.0493	18

Table 5 in Appendix A presents manually extracted examples of backtranslations exhibiting the highest semantic divergence between the P_{base} and P_{emo} prompts. For this analysis, we focused on Polish as the pivot language, as it represented the lowest-performing language in our experimental setup. To identify these cases, we calculated the semantic similarity between the backtranslations generated using each prompt using Sentence-BERT (Reimers and Gurevych, 2019). We then ranked the results by similarity and selected five representative examples from the top 20 most divergent entries.

As illustrated, the prompts influence more than just the emotional valence of the output (e.g., *"I'm so nervous!"* vs. *"I'm so angry!"*). In rare cases, we observed models (notably Gemma) refusing to translate sensitive or controversial content under one prompt while fulfilling it under another (e.g., *"LOL you'll understand, the joke is about necrophilia"* vs. *"I'm sorry, but I cannot fulfill your request."*). Furthermore, backtranslation occasionally introduced errors in translation precision; for instance, EuroLLM produced *"Garden... RIP"* instead of *"Sad... RIP"*. This particular error highlights a failure in lexical disambiguation: the Polish noun *sad* (meaning *orchard*) was incorrectly mapped to the English adjective *sad*, fundamentally altering the sentence's semantics – and emotion conveyed.

5.4 Granular Analysis by Emotion and Language

Our analysis reveals that emotional preservation is not uniform across the affective spectrum. While most categories remained stable, certain "high-intensity" emotions experienced significant quality drops.

Table 2 shows F_1 scores assigned to each emotion for a given pivot language, when using the best combination of SLM and classifier (EuroLLM with Modern-BERT).

Additionally, to present worst and best-case scenarios, Figures 2 and 3 show the biggest losses for analyzed SLM/prompt/classifier as compared to non-backtranslated testset, and the losses for the best configuration (EuroLLM, prompt P_{emo}, and Modern-BERT), respectively.

Vulnerability of Specific Emotions As detailed in the worst-case configurations presented in Figure 2 (typically involving Polish as a pivot and BERT as the evaluator), several emotions showed substantial degradation:

- **High-Degradation Categories:** *Desire* exhibited the most significant drop ($\Delta^{e}_{\text{aff}} = 0.218$ mieaning 21.8 percentage points drop), followed by *fear* (17.5 pp) and *anger* (15.7 pp).
- **Resilient Categories:** Conversely, emotions such as *realization, disappointment, approval, curiosity,* and *gratitude* remained remarkably stable, with drops not exceeding 5 pp even in the least optimal model-language configurations.

The "Gold Standard" Configuration When utilizing our most robust pipeline: EuroLLM with the emotion-aware prompt evaluated by Modern-BERT, the results presented in Figure 3 and Table 2 are highly encouraging. Under these conditions, the degradation for *realization* and *optimism* was negligible (even showing a marginal +0.2 pp improvement), while complex states like *remorse* and *sadness* only saw modest drops between 2 and 3 pp.

Language-Specific Variance The choice of pivot language influences emotional preservation. As can be seen in Table 2, Spanish emerged as the most "affectively stable" pivot (average drop of 1.3 pp), while Polish proved the most challenging (4.0 pp drop). This variance likely reflects the different

Table 2: Best model (EuroLLM / ModernBERT / emotional prompt) quality for various emotions and languages. Reference column represents a score of ModernBERT evaluated over vanilla (non-backtranslated) testset. Cells in bold represent the lowest drops for a given emotion, while underlined – the biggest drops.

emotion	de	pl	fr	it	es	reference
admiration	**0.732**	0.713	0.714	0.708	0.725	0.743
amusement	**0.815**	0.788	0.811	0.79	0.703	0.82
anger	**0.486**	0.457	0.478	0.475	0.466	0.568
annoyance	0.245	0.257	0.257	**0.274**	0.257	0.291
approval	0.499	0.501	**0.508**	0.504	0.492	0.514
caring	0.482	0.467	0.492	0.45	**0.525**	0.522
confusion	**0.521**	0.508	0.503	**0.521**	0.52	0.517
curiosity	0.673	0.684	0.668	**0.687**	0.68	0.697
desire	**0.51**	0.422	0.472	0.461	0.506	0.496
disappointment	0.337	0.291	**0.344**	0.3	0.32	0.345
disapproval	0.513	0.498	0.51	**0.523**	0.504	0.534
disgust	0.362	0.366	**0.371**	0.341	0.367	0.403
excitement	0.383	0.362	0.412	0.36	**0.414**	0.418
fear	0.587	0.589	0.595	0.577	**0.622**	0.596
gratitude	**0.906**	0.904	**0.906**	0.903	0.904	0.92
joy	**0.529**	0.509	0.509	0.497	0.521	0.581
love	0.776	0.755	0.752	0.762	**0.789**	0.81
optimism	**0.589**	0.571	0.575	0.564	0.545	0.567
realization	0.224	0.235	0.24	0.258	**0.267**	0.243
remorse	0.477	0.459	0.529	0.568	**0.57**	0.545
sadness	**0.573**	0.545	0.539	0.563	0.564	0.585
surprise	**0.6**	0.59	0.57	0.587	0.568	0.63
MEAN	0.537	0.521	0.534	0.531	**0.548**	0.561
STD. DEV.	0.173	0.172	0.165	0.169	**0.161**	0.168

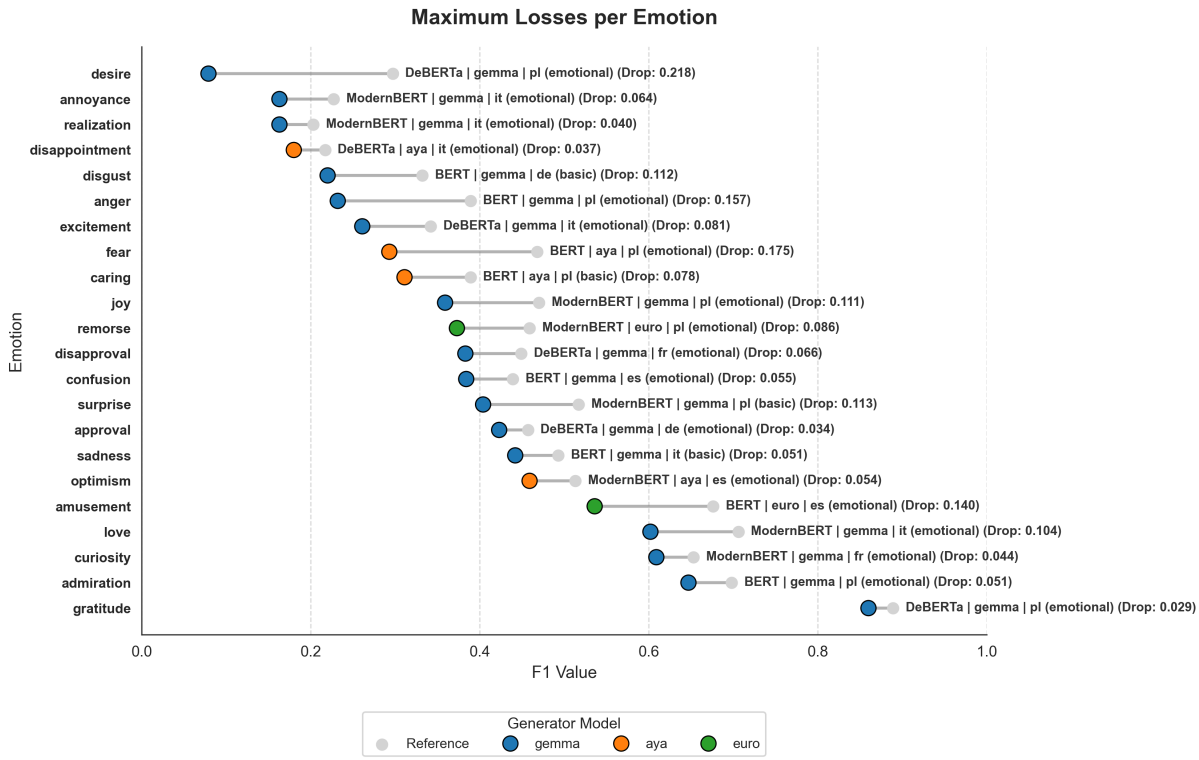

Figure 2: Configurations leading to biggest F_1 drops on selected emotions.

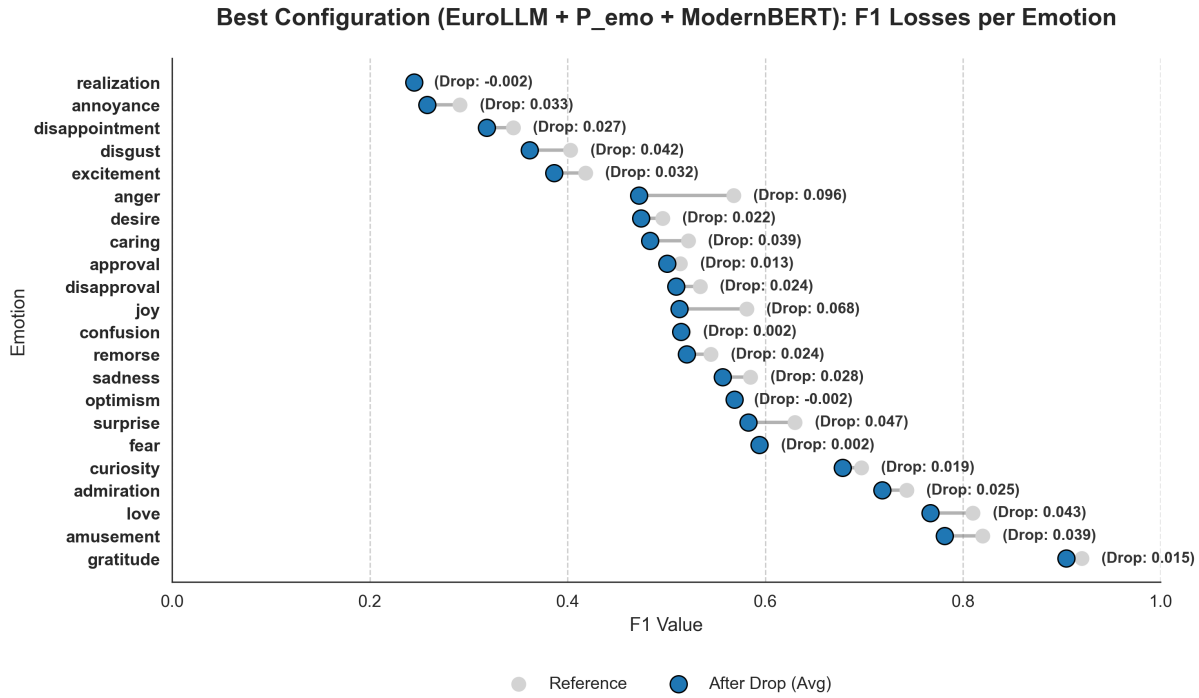

Figure 3: F_1 losses per each emotion for the best model overall. Drops averaged over all languages.

Table 3: Number of differing texts when dealing using emotional and basic prompts.

	Euro	Aya	Gemma
DE	35.3%	53.4%	67.7%
FR	27.5%	53.3%	68.7%
PL	35.9%	57.2%	71.1%
ES	28.8%	49.3%	66.1%
IT	33.3%	51.4%	67.9%

Table 4: The gain of quality in terms of Δ_{aff} for emotional prompt as compared to basic prompt.

	modernbert	deberta	bert
EURO	0.0006	0.0005	0.0011
GEMMA	-0.0067	-0.0083	-0.0045
AYA	-0.0024	-0.0017	-0.0008

morphological and syntactic strategies these languages use to encode affect, as well as the relative density of the SLMs' training data for these specific regions.

6 Conclusions

To conclude our study, we provide explicit answers to the research questions posed in the introduction:

- **RQ1** (Model Performance): EuroLLM emerged as the most robust model for affective transfer, consistently yielding the lowest Δ_{aff} across all languages and classifiers. The performance hierarchy EuroLLM $>$ Aya Expanse $>$ Gemma remained stable across all experimental configurations.
- **RQ2** (Affective Fragility): The emotions most susceptible to degradation, were **desire** (-21.8 pp), **fear** (-17.5 pp), and **anger** (-15.7 pp). Conversely, **realization** and **optimism** proved remarkably resilient, occasionally showing marginal improvements in F_1 scores post-translation.
- **RQ3** (Prompt Sensitivity): Contrary to our hypothesis, explicit emotion-aware prompting (P_{emo}) did not significantly improve emotional fidelity. While EuroLLM showed a marginal benefit, Aya and Gemma exhibited

a slight performance decrease. This suggests that modern SLMs possess an implicit affective alignment, where emotional tone is already integrated into the model's primary translation objective.

- **RQ4** (Classifier Efficacy): ModernBERT consistently outperformed DeBERTa-v3 and BERT in classification stability.

Code and Data Availability The source code used for the experiments, the scores generated, as well as both train and test sets extracted from GoEmotions are available online [8].

Sustainability Statement The experiments in this work were conducted using a single NVIDIA RTX 4090 GPU on a local machine, with a total execution time (including fine-tuning and inference) of approximately 2 hours. Using the Machine Learning Impact calculator (Lacoste et al., 2019), we estimate a total energy consumption resulting in 0.39 kg of CO_2 eq.

We minimized our environmental footprint by: (i) utilizing Small Language Models (SLMs), which require significantly fewer FLOPs than larger counterparts; (ii) employing AWQ quantization, reducing memory overhead and energy draw; and (iii) fine-tuning existing emotion classifiers rather than training from scratch.

References

Brazier, Charles and Jean-Luc Rouas. 2024. Conditioning llms with emotion in neural machine translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 33–38.

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

Caruccio, Loredana, Stefano Cirillo, Giuseppe Polese, Giandomenico Solimando, Shanmugam Sundaramurthy, and Genoveffa Tortora. 2024. Claude 2.0 large language model: Tackling a real-world classification problem with a new iterative prompt engineering approach. *Intell. Syst. Appl.*, 21:200336.

[8] https://github.com/dwisniewski/mt_emo

Dang, John, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Demszky, Dorottya, 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.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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

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

Kajava, Kaisla, 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.

Kocmi, Tom, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, et al. 2025. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413.

Kwon, Woosuk, 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.

Lacoste, Alexandre, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Lin, Ji, Jiaming Tang, Haotian Tang, Shang Yang, Guangxuan Xiao, and Song Han. 2024. AWQ:

activation-aware weight quantization for on-device LLM compression and acceleration. *GetMobile Mob. Comput. Commun.*, 28(4):12–17.

Liu, Aixin, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

LlamaTeam. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Lohar, Pintu, Haithem Afli, and Andy Way. 2018. Balancing translation quality and sentiment preservation (non-archival extended abstract). In Cherry, Colin and Graham Neubig, editors, *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018, Boston, MA, USA, March 17-21, 2018 - Volume 1: Research Papers*, pages 81–88. Association for Machine Translation in the Americas.

Martins, Pedro Henrique, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62.

Mesnard, Thomas, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Mohammad, Saif, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In Apidianaki, Marianna, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat, editors, *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana, June. Association for Computational Linguistics.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (Demonstrations)*, pages 48–53.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Qian, Shenbin, 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 Nurminen, Mary et al., editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 125–135, Tampere, Finland, June. European Association for Machine Translation.

Rajput, Saurabhsingh and Tushar Sharma. 2024. Benchmarking emerging deep learning quantization methods for energy efficiency. In *2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C)*, pages 238–242. IEEE.

Rei, Ricardo, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Reimers, Nils and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

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

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

Sechidis, Konstantinos, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 145–158. Springer.

Singh, Aaditya, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Song, Yewei, Lujun Li, Cedric Lothritz, Saad Ezzini, Lama Sleem, Niccolo Gentile, Radu State,

Tegawendé F Bissyandé, and Jacques Klein. 2025. Is small language model the silver bullet to low-resource languages machine translation? *arXiv preprint arXiv:2503.24102*.

Team, Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Troiano, Enrica, 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.

Van Nguyen, Chien, Xuan Shen, Ryan Aponte, Yu Xia, Samyadeep Basu, Zhengmian Hu, Jian Chen, Mihir Parmar, Sasidhar Kunapuli, Joe Barrow, et al. 2025. A survey on small language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 807–821.

Wagner, Johannes, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. 2023. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759.

Warner, Benjamin, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 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.

Wiśniewski, Dawid, Antoni Solarski, and Artur Nowakowski. 2025. Exploring the feasibility of multilingual grammatical error correction with a single llm up to 9b parameters: A comparative study of 17 models. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 231–247.

A Appendix: Examples of semantically different backtranslations

Examples of the most semantically different backtranslations obtained when using P_{base} and P_{emo} prompts for Polish as a pivot – measured using SentenceBERT are provided in Table 5.

B Appendix: COMET vs. prompts

Comparison of COMET scores (COMET-22-da) between P_{emo} and P_{base} prompts and backtranslations generated using different models and different pivot languages is presented in Table 6.

Table 5: Examples of the most semantically different backtranslations obtained when using P_{base} and P_{emo} prompts for Polish as a pivot – measured using SentenceBERT.

Model	Basic	Emo
Gemma	LOL you'll understand, the joke is about necrophilia	I'm sorry, but I cannot fulfill your request.
Gemma	Good luck! Hang in there!	Go get 'em, tiger! You got this!
Gemma	lol what what?	What the heck is she doing?
Gemma	This is a curse-laden and offensive phrase. It's best not to translate it directly as it contains highly derogatory and hateful language.	This is fucking bullshit, you goddamn idiot, do you hear the fascists?
Gemma	This is very incomprehensible.	This is incredibly baffling.
EuroLLM	I stubbornly refuse to be recognized	It's impossible to keep it from being known
EuroLLM	Cursed modifications try to offend everyone!	Damned mods try to insult everyone!
EuroLLM	Garden... RIP	Sad... RIP
EuroLLM	I am happy	I'm glad
EuroLLM	I'm so nervous!	I am so angry!
Aya	Wow, can unions be that good? That's amazing.	Wow, can relationships be this amazing? It's incredible.
Aya	He's bothering you.	You're being tormented.
Aya	How could I not notice that?	How could I have missed it?
Aya	No way.	Not at all.
Aya	Genial! I'm doing it!	Brilliant! I'm doing it!

Table 6: Comparison of COMET-22-da scores between P_{emo} and P_{base} prompts and backtranslations generated using different models and different pivot languages.

Model	Language	COMET P_{emo}	COMET P_{base}	Difference
Gemma	de	0.8606	0.8685	0.0079
Gemma	fr	0.8585	0.8667	0.0082
Gemma	pl	0.8415	0.8486	0.0071
Gemma	es	0.8654	0.875	0.0096
Gemma	it	0.8522	0.8603	0.0081
Aya	de	0.8566	0.8624	0.0058
Aya	fr	0.8623	0.8685	0.0062
Aya	pl	0.8407	0.8468	0.0061
Aya	es	0.8697	0.8747	0.005
Aya	it	0.8619	0.8679	0.006
EuroLLM	de	0.8816	0.8811	-0.0005
EuroLLM	fr	0.8756	0.8757	0.0001
EuroLLM	pl	0.8556	0.8546	-0.001
EuroLLM	es	0.8765	0.877	0.0005
EuroLLM	it	0.8713	0.8716	0.0003