Are More Emotion Theories Better?
Joint Prompting Improves Emotion Categorization Robustness

Albina Sarymsakova[1,2] **and Roman Klinger**[2]
[1]Instituto de Estudios Gallegos Padre Sarmiento (IEGPS-CSIC), Santiago de Compostela, Spain
[2]Fundamentals of Natural Language Processing, University of Bamberg, Germany
albina.s@iegps.csic.es, roman.klinger@uni-bamberg.de

Abstract

Emotion analysis systems are expected by users to provide reliable results. Optimizing their robustness, in addition to accuracy, is particularly challenging in noisy environments, such as social media data or other forms of naturally expressed language. Our hypothesis is that robustness can be improved by guiding a language model to pay attention to multiple emotion signals in the input text. Therefore, requesting labels according to multiple emotion theories may improve robustness. To study this hypothesis, we run perturbation experiments on input instances to test the robustness of LLM configurations based on basic emotions only, and jointly with valence–arousal, and cognitive appraisals. We find that adding valence–arousal and cognitive appraisal along with basic emotion instructions to the emotion classification improves robustness, with the effect being most pronounced for adding appraisals. In a further analysis, we find that this improvement is less susceptible to word order perturbations and more to random word substitutions and deletions. These differences hold across multiple LLMs studied in our work.

1 Introduction

Text-based emotion detection connects affective computing with opinion mining, emphasizing the analysis, interpretation, and extraction of emotion cues from subjective text (Canales and Martínez-Barco, 2014; Muhammad et al., 2025). While traditional sentiment analysis is often confined to labeling texts as positive, negative, or neutral (Zhao et al., 2016), text-based emotion detection allows for the recognition of a wider spectrum of emotions (Troiano et al., 2023b; Greschner and Klinger, 2025), by relying on various psychological theories, such as Ekman's discrete model (Ekman and Friesen, 1978; Ekman, 1992), Russell's dimensional model of valence and arousal (Russell, 1980) and appraisal theories (Sherer et al., 1982; Smith and Ellsworth, 1985).

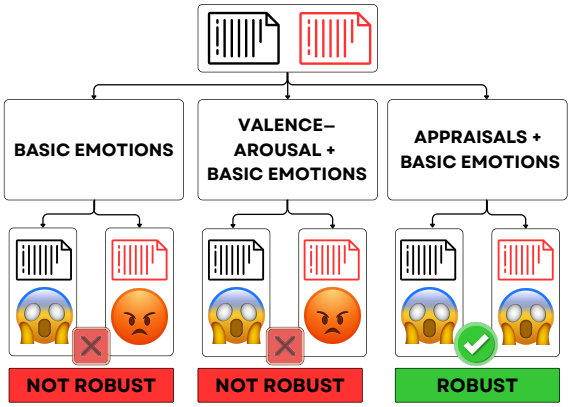

Figure 1: Our goal is to identify which LLM configuration, whether containing basic emotions, valence–arousal + basic, or appraisals + basic, yields more robust emotion-detection outputs between the original text (black) and the perturbed text (red).

In some cases, downstream tasks, such as measuring the potential attention of a car driver through their arousal (Cevher et al., 2019) or interpreting if a piece of art is perceived to be pleasant (Haider et al., 2020) dictate the choice of a psychological theory. Nonetheless, when no concrete downstream task drives the selection, the choice needs to be made based on other signals.

We take the stance that emotion analysis is inherently subjective, and that variances in the output of models or model configurations might appear and be reasonable (often referred to as the phenomenon of human label variation, Plank, 2022). This aspect of affective computing does, however, not change the expectation of users that an automatic system works reliable and produces comparable results across similar texts. Nevertheless, we know that large language model outputs are brittle and sensitive to small meaningless changes to the input texts (as humans are as well, Li et al., 2025a).

Proceedings of the 11th Workshop on Natural User-generated Text, pages 10–25
October 28, 2026 ©2026 Association for Computational Linguistics

We therefore aim at the optimization of the robustness and therefore reliability of emotion classification decisions of prompted large language models. While text-based emotion detection has shown promising results across various textual domains (Acheampong et al., 2021; Al Maruf et al., 2024), including LLMs (Muhammad et al., 2025; Zhang et al., 2025) tests of their robustness (Khurana et al., 2021; Jeon et al., 2025; Sabour et al., 2024; Li et al., 2025a) showed that they exhibit a substantial performance gap relative to humans, particularly due to their sensitivity to noisy texts and contextual variation (Li et al., 2025b). Furthermore, even advanced prompting strategies, such as chain-of-thought approaches, only show limited improvements (Jeon et al., 2025).

The main idea of our work is that prompting large language models according to multiple emotion theories improves their robustness in emotion categorization. The reasoning behind that is that varying emotion theories guide the language model towards different parts of the text to make their classification decision. Theories of affect (Russell, 1980) might, for instance, make the model focus more on adjectives, while appraisal theories (Smith and Ellsworth, 1985) move the focus to events and verbs.

We test this hypothesis by measuring the pair-wise accuracy (Jeon et al., 2025) between predictions on the original and perturbed texts across three LLM-based configurations (basic emotions, valence–arousal+basic, appraisals+basic). We stress that in our work the pair-wise accuracy metric does not quantify correctness: we do not evaluate the labels of perturbed texts against gold emotion labels. Rather, the pair-wise accuracy metric measures how consistent the labels of perturbed texts are with those of the original ones.

This experiment allows us to address the following research questions:

RQ1: Which emotion detection configuration (basic emotions, valence–arousal+basic, appraisals+basic) contributes most to LLMs' predictions robustness under textual perturbation?

RQ2: Which text perturbation method provokes more unrobust labels in these configurations when analyzing their dependency on word, syntax, or broader semantic context changes while predicting emotions?

2 Related Work

2.1 Emotion Analysis and Models

Emotion analysis in natural language processing (NLP) encompasses two major theoretical frameworks: theories of emotion categories and theories of emotion dimensions (Cavicchio, 2025). Each has resulted in numerous available resources for emotion detection (Bostan and Klinger, 2018).

Basic emotion theories conceptualize emotions as discrete categories, defined by how the experiencer displays them. Ekman and Friesen (1978) identified six basic emotions (happiness, fear, sadness, surprise, disgust, and anger) from cross-cultural facial expression studies, which Plutchik (1980) extended to eight primary emotions, adding trust and anticipation, and organized into opposing pairs (e.g., joy vs. sadness, anger vs. fear).

Dimensional theories, in contrast, represent emotions along continuous dimensions that characterize the experiencer's attitude toward an event. Mehrabian and Russell (1974) proposed a three-dimensional model of valence, arousal, and dominance, while Russell (1980) introduced the two-dimensional Circumplex Model of Affect, defined by valence (pleasant–unpleasant) and arousal (calm–activated) alone.

Within the dimensional perspective, cognitive appraisals offer a distinct approach, proposing that emotions arise from subjective assessment of situations (Scherer et al., 2001). It links emotions to appraisal dimensions (Troiano, 2024; Troiano et al., 2023a). Appraisals are dynamic, continuously updated assessments, modifiable at any stage of the emotional response. In contrast to valence–arousal and basic emotion theories, appraisals center on evaluating multiple factors of an event as they relate to the experiencer, such as its suddenness or the experiencer's familiarity with it.

All three theories differ in their perspectives and application to text-based emotion detection. Basic emotion theory addresses this task by tracing emotion lexicons or other explicit lexical cues. The valence–arousal approach rates the experiencer's state along valence and arousal, deriving an emotion from it. In contrast, appraisals evaluate the experiencer–event relationship across a broader set of dimensions (Roseman, 1979; Sherer et al., 1982; Smith and Ellsworth, 1985), demanding the most context reconstruction of the three theories for emotion analysis. This motivated notable progress in appraisal-based approaches for emotion detection

across textual domains, including coping strategies (Troiano et al., 2024), social media (Stranisci et al., 2022), and emotion event self-reports (Hofmann et al., 2020). Troiano et al. (2023b) introduced Crowd-enVent, the most comprehensive appraisal-annotated dataset for event emotion analysis, and Greschner et al. (2026b) developed ContArgA, the largest corpus of arguments annotated for emotions and appraisals.

These studies reflect a growing shift toward the cognitive appraisal approach, which offers a more context- and human-oriented perspective on emotion analysis. This suggests that the appraisal-based setting may yield more robust emotion categorization, since it requires analyzing broader context and properties, part of the hypothesis we explore.

2.2 Robustness Testing

The *robustness* in our study is defined as the evaluation of models' capacity to maintain consistent emotion predictions when the instance's text is perturbed (Pandia and Ettinger, 2021). We emphasize that the notion of robustness in the present study refers to how constant a configuration's prediction on the original and on the perturbed text is, and is therefore orthogonal to correctness against gold emotion labels. A configuration can be robust while consistently reproducing an incorrect label, or accurate on clean text yet volatile under perturbation. We adopt robustness rather than gold-label accuracy as our primary target because our motivating concern is deployment reliability, i.e., whether a system's output changes for noisy or perturbed inputs a user would consider equivalent, a failure mode that persists independently of how accurate a configuration is on non-noisy text, and that has not previously been compared across joint emotion analysis configurations. Existing methods for LLM output robustness testing vary across tasks. Within this context, Liu et al. (2021) explore the language understanding capabilities of LLMs by introducing real-world dialogue perturbations, including linguistic variety, speech characteristics, and noise. For textual perturbations, the authors employed a synonym replacement, random insertion, random swapping, and deletion.

Li et al. (2021) propose a set of perturbation methods, including random substitutions and word insertions, to assess the robustness of pretrained models within the framework of textual adversarial attacks. These perturbation methods are further corroborated by Pandia and Ettinger (2021).

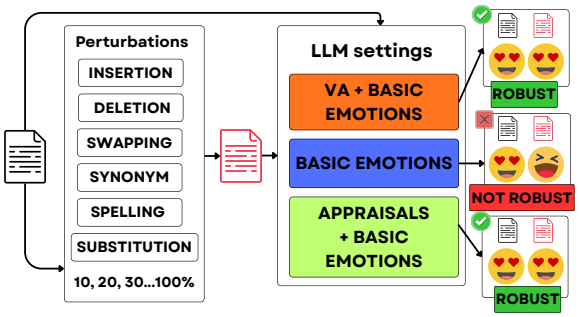

Figure 2: Experimental setup for the robustness study of emotion detection outputs of three configurations: B, VA+B, App+B.

In the context of testing LLMs' sensitivity to emotion context, Jeon et al. (2025) analyze their ability to understand emotions under textual perturbations, such as inserting or substituting words referring to time, place, or agent to create misleading or distracting content. Furthermore, Khurana et al. (2021) evaluated BERT-based models' robustness in sentiment analysis by introducing part-of-speech substitutions, spelling errors (e.g., swapping adjacent characters), and random word swaps. Spelling perturbations are also used for LLM prompt brittleness testing by Li et al. (2025a), across offensiveness and politeness rating, irony detection, and emotion classification tasks.

Based on the aforementioned studies, several strategies are systematically used for robustness testing of LMs, including in emotion analysis to assess sensitivity to textual manipulations and noise. The common techniques include synonym replacement, word swapping, deletion, insertion, substitution, and spelling errors. To the best of our knowledge, these methods have been applied to LMs and LLMs, but none has yet been used to test the robustness of predictions of discrete and dimensional theories that complement one another.

3 Experimental Setup

In the present study, we evaluate the robustness of predictions of three emotion detection configurations. We prompt LLMs for emotion labels under textual perturbations using basic emotions (B), valence–arousal+basic emotions (VA+B), and appraisal+basic emotions (App+B) settings, in order to test their robustness and identify which perturbation method (word-, syntax- or semantic-level) most exposes their sensitivity. The emotion analy-

Configuration	Basic emotions	Appraisals+Basic	VA+Basic
Role & Task	You are an expert in text-based emotion classification. Process the statements below.		
Output format	`emotion`	`22 appraisals, . . . emotion`	`valence, arousal, emotion`
Task description	Single-label classification. Emotion {anger, disgust, fear, guilt, joy, sadness, shame, surprise, trust}.	Task 1: Rate each of 22 appraisals. (1-5 scale). Task 2: *idem* Non-joint.	Task 1: Rate valence & arousal (1-9 scale). Task 2: *idem* Non-joint.
Constraints	Never write "unknown", "neutral", "none", "n/a", or "other". If unsure, pick closest emotion. One line per statement; no blank lines, headers, or explanations. Outputs not conforming to the expected format are marked `INVALID` after 3 retries.		
Example	"I studied all night and still failed the exam." → `sadness`; "Blessed with a baby boy today . . ." → `joy`	"I opened my email and found out that I had been awarded a fully funded research grant I applied for last year." → `4,3,5,5,5,3,1,2,4,4,4,4,3,4,5,4,4,4,2,5,1,1,joy`	"Blessed with a baby boy today . . ." → `7.5,2,joy`; "I can't believe they lied to me." → `2,7,anger`
Input	Now classify: {text1}, {text2}, . . .		

Table 1: Prompt structure.

sis scripts used in this study are publicly available.[1]

Figure 2 provides a visual overview of the experimental workflow. Our experiments utilize the robustness testing strategies detailed in Section 2.2 to perturb selected dataset instances. We perform emotion analysis using LLMs on both the original and perturbed instances. To evaluate the stability of the outputs from both discrete and dimensional models, described in Section 2.1, we employ joint and non-joint prompts to obtain predictions. Finally, we assess robustness by calculating the pair-wise accuracy scores between the emotion labels generated for the original instances versus those for their perturbed counterparts.

3.1 Prompting Method

The prompt structure was adapted from Bagdon et al. (2024) and is summarized in Table 1. The prompts differ across the three emotion detection configurations. Each prompt contains up to six components. The *role and task* establish the model's roles and provide a general instruction on what to do. The *output format* specifies the expected output structure, which varies across configurations: the B outputs only a categorical label; the App+B outputs 22 numerical appraisal ratings (Troiano et al., 2022) followed by a categorical label; and the VA+B outputs integer or decimal ratings for valence and arousal alongside a categorical label. The *task description* provides detailed instructions on how to perform the emotion detection task, and differs across three configurations. The *constraints* define the formatting rules the model must follow and how outputs are parsed: predictions outside the expected format or containing an emotion label outside the predefined set are marked *INVALID* after three unsuccessful retries. Parsing is rule-based: each model's raw response is split into lines, each line is stripped and lower-cased, and, depending on configuration, either matched exactly against the fixed emotion-label set (B) or split on commas and accepted only if it has the expected number of fields with a valid emotion label in the last position (VA+B and App+B); the first line that matches is kept, and the whole request is retried (up to three times) if no line matches. This rule-based parser is identical across all three models. Instance pairs where either the original or perturbed-text label is INVALID are dropped from the pair-wise accuracy computation. Overall dropped percentages per model are: Gemma – 8.05, Llama – 9.45, Qwen – 16.91.[2] The *example* provides labeled in-context instances[3] that illustrate the expected output for each configuration. Finally, the *input* contains the statements the model is asked to process.

We adopted the prompt configuration proposed by Greschner et al. (2026a). It consists of a non-joint setting, in which emotion categories and appraisal dimensions are predicted within a single task, and a joint setting, in which both are predicted together, allowing the two outputs to mutually inform each other. We adopted and adapted this framework, for (i) *Non-joint prompt*, to establish a baseline, using the B configuration emotion detections; and for (ii) *Joint prompt*, to test if VA+B and App+B configuration may benefit from insights from the other emotion theories, while out-

[1] https://www.uni-bamberg.de/en/nlproc/resources/

[2] Because the parsing procedure does not differ across models, we suppose that Qwen shows a higher drop rate in its outputs more often due to the required format under the strict parsing rule described above, most likely on the 23-field App+B task, which is the most constrained of the three output formats (see Table 1).

[3] We use in-context examples since multiple studies (Brown et al., 2020; Bareiß et al., 2024; Creanga et al., 2025; Barfi et al., 2025) highlight that one- and few-shot prompting brings improvements for emotion classification in LLM-based tasks.

Emotion theory	Corpus	Subset
Appraisals	crowd-enVENT	200
	ContArgA	200
Basic emotions	SSEC	200
	ISEAR	200
Valence–Arousal	EmoBank	200
	Facebook VA	200
Total		1200

Table 2: Overview of evaluation corpora, grouped by emotion theory. We select two corpora per theory to test whether results generalize across datasets that share a theoretical grounding but differ in domain and collection method, without implying identical label sets within each group. Each corpus contributes a balanced subset of 200 instances.

put valence–arousal and appraisals rates alongside categorical labels.

3.2 LLMs

We select three LLMs adequate for local deployment for evaluation, based on their established utility and proven performance in text-based emotion classification tasks (Muhammad et al., 2025; Soligo et al., 2026; Yacoubi et al., 2025; Huang et al., 2025; Zhang and Zhong, 2025): Gemma-3-4b (Kamath et al., 2025), Llama-3.3-70b (Grattafiori et al., 2024), and Qwen2.5-7b (Hui et al., 2024). All models are accessed and executed using the ollama (version 0.12.9) library[4]. All models were run at Ollama's default quantization level for each model release (Q4_K_M, 4-bit), without any custom quantization flags.

The decoding configuration was selected to match the model's real-world use, with temperature: 1, top_k: 64, top_p: 0.95. These settings correspond to the default configuration used for the Gemma-3-4b model in the ollama library. We also used the default decoding configuration for Llama-3.3-70b and Qwen2.5-7b with temperature 0.8, top_k 40, and top_p 0.9. The experiments are run on Nvidia A100 GPUs, with a total estimation of 646.35 GPU hours (166.16 h for B, 170.17 h for VA+B, 310.02 h for App+B).

3.3 Emotion Benchmarks and Corpora

This study examines whether B, VA+B, or App+B configuration produces more robust predictions under different textual perturbation methods, and which of these methods acts as more destabilizing at word, syntax, or broader context levels. We select benchmarks and corpora, summarised in Table 2, that are psychologically grounded, empirically well-established, and compatible in terms of emotion categories in order to achieve this goal. We select a subset of corpora that has been used with each theory before.

For the cognitive appraisals benchmark, we use crowd-enVENT (Troiano et al., 2022), which provides ratings across 22 appraisal dimensions (1–5 scale) grounded in componential appraisal theories (Sherer et al., 1982; Smith and Ellsworth, 1985). It covers nine emotion categories (anger, disgust, fear, guilt, joy, sadness, shame, surprise, trust), extending the Ekman set with self-conscious and social emotions closely tied to specific appraisals. We also apply ContArgA (Greschner et al., 2026b), which shares these categories and grounding.

For the basic emotions benchmark, we select SSEC (Schuff et al., 2017) and ISEAR (Scherer and Wallbott, 1994) as the most compatible benchmarks with the appraisal corpora in terms of emotion categories, noting that neither covers the full emotion set: SSEC lacks guilt and shame, while ISEAR omits trust and surprise.

For the valence–arousal benchmark, we use EmoBank (Buechel and Hahn, 2017)[5] and Facebook VA (Preoţiuc-Pietro et al., 2016), restricting the dimensional representation to valence and arousal, the two most consistently identified dimensions of affective studies (Russell, 1980; Posner et al., 2005). Prior work shows that VA scores enable predicting categorical emotion labels with reasonable accuracy (Buechel and Hahn, 2017), which also enables us to evaluate whether the VA-inclusive configuration shows greater robustness in emotion predictions. We use a combination of all corpora for all experiments.

3.4 Perturbation Methods

We use nlpaug library[6] to apply six perturbation methods (see Table 3 for examples): synonym replacement (SR), random swap (RS), random deletion (RD), misspelling augmentation (MA), Conditional Bidirectional Encoder Representations from Transformers random substitution (CBERTRS), and Conditional Bidirectional Encoder Representations from Transformers random insertions (CBER-

[4] https://github.com/ollama/ollama/releases/tag/v0.12.9

[5] While EmoBank uses a 5-point scale, we revert to the original 9-point SAM scale (Bradley and Lang, 1994) in line with standard practice in affective psychology.

[6] https://github.com/makcedward/nlpaug

Method	Text
SR	My biggest is loosing you so delight don river ' thyroxine earn maine ever so present that fearfulness. You cause no absolutely no idea how much you mean to maine.
RS	My biggest so is loosing you don please ' t me ever make face fear. that You no have no absolutely how you idea to much mean me.
RD	My [...] you [...] [...] ' [...] [...] face. [...] have absolutely no how much you to.
MA	Ma biggest is losing you soo please don ' it moke My eve faice that fear. You hvae not absolutely now ideia how much yu meat to ne.
CBERTRS	every secret fear if you only please try ' f cause me ever face that fear. will have an absolutely zero problem how much you enjoy losing me.
CBERTRI	and my biggest is loosing precious you dead so please don ' t again make me have ever to face that fear. may you have my no secrets absolutely scary no fucking idea just how much love you really mean me to please me.

Table 3: Outputs for 50% of the perturbed context for an original instance *My biggest is loosing you so please don't make me ever face that fear. You have no absolutely no idea how much you mean to me*.

TRI). The methods are drawn from various benchmarks, as described below.

Easy Data Augmentation (EDA) (Wei and Zou, 2019) presents SR, which substitutes randomly selected words with synonyms retrieved from a lexical database (Miller, 1994), and it is part-of-speech-aware, meaning a verb is replaced with a verb synonym and a noun with a noun synonym; RS, which randomly selects word positions and swaps each chosen word with its immediate neighbor; and RD, randomly selects words and removes them from the sentence entirely.

MA (Pruthi et al., 2019) replaces randomly selected words with common misspellings from a pre-compiled dictionary of real human spelling errors. The words are chosen randomly, as are the misspellings from the available candidates for that word.

Conditional Bidirectional Encoder Representations from Transformers (CBERT) (Wu et al., 2019; Anaby-Tavor et al., 2019) uses BERT-base-uncased (Devlin et al., 2019) to perform context-aware random substitution (CBERTRS), masking and replacing randomly selected words, and random insertions (CBERTRI), inserting new words at randomly chosen positions between existing tokens. These substitutions and insertions are sampled randomly from the top-k predictions.

3.5 Implementation

We perform experiments to test the robustness of outputs of three emotion detection configurations (B, VA+B, App+B) under textual perturbation. Holding each configuration fixed, we also ablate over the six perturbation methods to determine which type of textual change most destabilizes its predictions to understand whether these configurations are more sensitive to word-level, syntactic, or semantic-level changes.

To perform the experiments, we follow the steps outlined next. We start with an initial dataset of 1.200 instances (see Table 2). We apply textual perturbations using the methods described in Section 3.4. Each of the 1.200 original instances is run 10 times per fraction, yielding 12.000 instances per fraction. Multiplied across 10 incremental fractions of modified content (10%, 20%, ..., 100%), this gives 120.000 instances per perturbation method, and multiplied across 6 perturbation methods, a total of 720.000 perturbed instances. We run emotion predictions on both original and perturbed instances using the prompting methods from Section 3.1 and the LLMs from Section 3.2. We compute pair-wise accuracy between predictions on original and perturbed instances to measure the robustness of the three configurations (B, VA+B, App+B). Finally, we analyze predictions across perturbation methods to evaluate each configuration's sensitivity at three levels: word (SR, MA), syntactic (RS, RD), and semantic (CBERTRS, CBERTRI).

4 Results

In this section, we present the results and analysis following the implementation in Section 3.5, and answer the proposed research questions. We evaluate the robustness of three emotion detection settings (B, VA+B, App+B) by measuring pair-wise accuracy between the original and perturbed texts' emotion outputs. We also analyze which perturbation method, word-level (SR, MA), syntactic-level (RS, RD), or semantic-level (CBERTRS, CBERTRI), most affects robustness, exposing each configurations' sensitivity.

4.1 RQ1: Which emotion detection configuration (basic emotions, valence–arousal + basic, appraisals + basic) contributes most to LLMs' predictions robustness under textual perturbation?

Figure 3 illustrates the average variation in LLM configurations' prediction robustness across perturbation fractions. All three settings (B, VA+B, App+B) show decreasing robustness as the per-

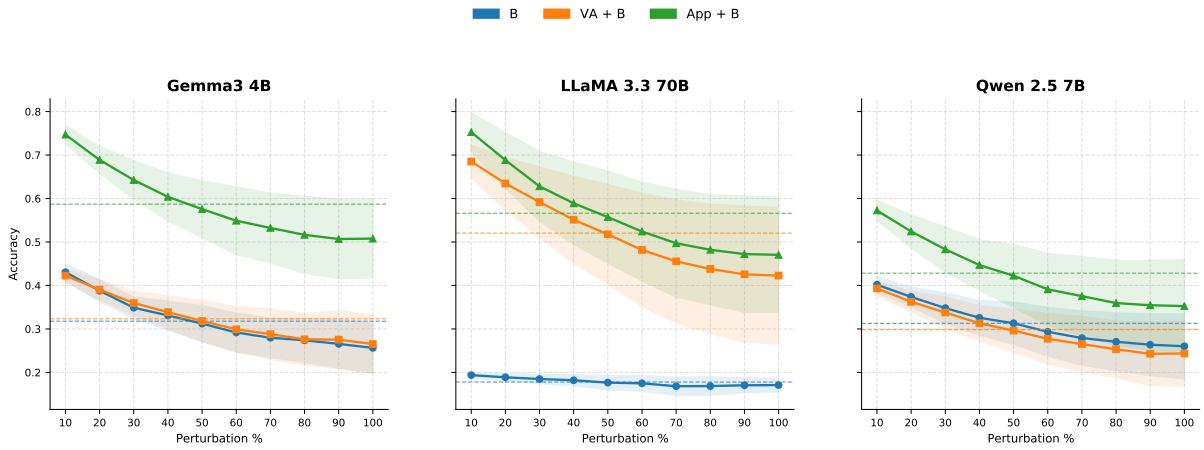

Figure 3: Pair-wise accuracy average curves for emotion predictions between original and perturbed texts. Each panel corresponds to one model (Gemma, LLama, Qwen), with curves comparing three settings (B, VA+B, App+B). Each point is an averaged pair-wise accuracy. For a given perturbation method and perturbation fraction (10%, 20%, 30%, etc.), 10 runs are averaged into one mean per fraction. Repeating this across all 6 methods gives 6 method-level averages per fraction, which are averaged again into the plotted value. The shaded band shows the standard deviation across those 6 method-level averages, not the 10 runs.

turbation fraction increases from 10% to 100%. Nonetheless, notable differences emerge across the three configurations.

The B setting (blue curve), our baseline, yields the lowest accuracy overall, starting below 0.5 for Gemma and Qwen and below 0.2 for Llama at 10% perturbation, and dropping to less than 0.2 at full perturbation. The VA+B setting (orange curve) shows higher accuracy throughout, suggesting that the additional valence–arousal rating task in a joint prompt stabilizes robustness. Within the App+B setting (green curve), all models achieve substantially higher pair-wise accuracy across all perturbation fractions, remaining above roughly 0.3 even at maximum perturbation and approaching 0.8 at 10% fraction for Gemma and Llama. This shows that the appraisal-inclusive configuration produces more robust predictions than either other setting, even as the perturbation fraction increases. It is further validated by the CheckList Invariance test (Ribeiro et al., 2020), specifically by mean fail rates, as Figure 5 shows in Appendix A.

Regarding overall LLMs performance, Llama consistently achieves the highest pair-wise accuracy values across joint configurations and perturbation fractions, while Gemma and Qwen perform comparably, especially for the App+B setting exhibiting the best results. Despite the differences in model sizes, the overall improvement is consistently observed in the App+B setting across all models, suggesting it is more robust even for the smaller models. As for the explanation of the shaded bands representing one standard deviation, the answer to RQ2 provides it.

4.2 RQ2: Which text perturbation method provokes more unrobust labels in these configurations when analyzing their dependency on word, syntax, or broader semantic context changes while predicting emotions?

We ablate across the six perturbation methods to identify which kind of textual change most destabilizes the configuration's (B, VA+B, App+B) predictions, revealing whether a given setting is more sensitive to word-level, syntactic, or semantic changes.

Figure 4 presents the prediction robustness, measured in pair-wise accuracy, of three LLMs across six perturbation types, synonym replacement (SR), random swap (RS), random deletion (RD), misspelling augmentation (MA), and CBERT-based random substitution (CBERTRS) and insertion (CBERTRI), for each emotion detection configuration. Differences in robustness emerge across both perturbation types and configurations. CBERTRS and RD consistently yield the lowest accuracy across all settings, proving the most destabilizing, while RS and SR yield the highest.

The B setting shows the lowest accuracy values overall across all perturbation types. The predictions from the CBERTRS and RD methods in this

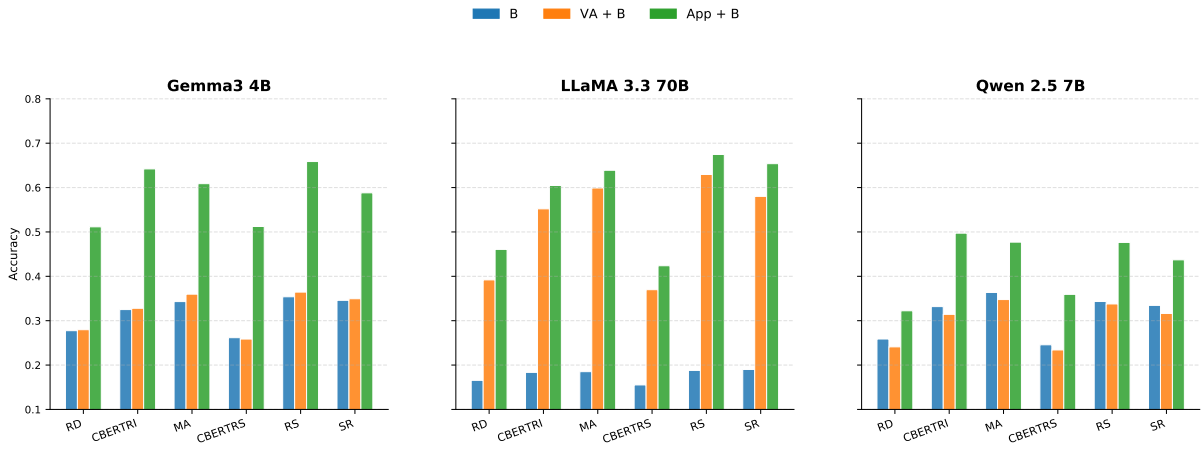

Figure 4: Average pair-wise accuracy by perturbation method and setting. Each bar height is a single average. For one method (e.g., CBERTRI), setting (e.g., VA+B), and model, the data already contains one average pair-wise accuracy per perturbation fraction (10%, 20%, ..., 100%), each itself the average of the 10 runs at that fraction. The bar chart averages those 10 fraction-level values once more into the bar height. A bar thus answers: on average across all fractions from 10% to 100%, how robust were this model's predictions under this perturbation method and setting?

Text	Perturbed	B	B pert.	App+B	App+B pert.
I asked my partner to do a task for me	Asked partner do for me	sadness	trust	**trust**	**trust**
I put someone down without thinking to make myself feel better	put someone without make myself better	surprise	guilt	**guilt**	**guilt**
I failed my driving test that I really wanted to pass	Failed driving I really pass	fear	anger	**sadness**	**sadness**
You'll also find bargain clothes for sale at the markets and on push-carts.	'll bargain clothes markets on-carts.	surprise	trust	**joy**	**surprise**
thinks that for girls boys are easily found. But GIRLS don't realize they need MEN!!	that for girls found. But ' they MEN!!	surprise	joy	**disgust**	**surprise**
Someone vomited on the street.	Vomited on the.	anger	disgust	**disgust**	**disgust**
A mother who shouts at her child for nothing.	A mother for nothing.	disgust	sadness	**anger**	**anger**
Finding out that I passed my final masters exams with a first (distinction)	Finding I final masters (distinction)	anger	sadness	**joy**	**joy**
I felt shame when I ate a lot of food because I was sad rather than hungry.	I felt shame I ate food was sad.	fear	anger	**shame**	**shame**
when i failed my driving test	i failed my driving	fear	sadness	**sadness**	**sadness**

Table 4: Llama predictions for the original and perturbed text of RD of 50% fraction across B and App+B configurations. Labels in bold indicate the predictions of the App+B configuration, which are more robust than those of the B setting.

setting are the least robust. In contrast, RS, CBERTRI, MA and SR perturbations exhibit higher accuracy values, indicating that these perturbation methods enable the B setting to make more robust predictions. The VA+B setting exhibits a similar pattern across perturbation types but maintains slightly higher accuracy values throughout. This configuration also suffers most under the CBERTRS method, followed by the RD, which produce the least robust results.

In the App+B setting, all models achieve substantially higher pair-wise accuracy across all perturbation methods. RS, CBERTRI, MA, and SR remain the least destabilizing, while RD and CBERTRS continue to produce the greatest robustness drops, though considerably smaller than in the other two configurations.

This tendency suggests that the additional context provided in the App+B setting helps the models anchor their predictions more robustly, reducing their sensitivity to surface-level lexical and structural changes. Notably, the most disruptive perturbations are those that alter or remove textual content rather than merely paraphrase it, indicating that the instability stems from genuine shifts in meaning rather than superficial noise. This is further supported by the CheckList Invariance test (Ribeiro et al., 2020), specifically the mean failure rates in Figure 6 (Appendix A).

5 Discussion

Table 4 presents examples and outputs for the B and App+B configurations with Llama, the best-performing model in this study, illustrating how one of the most destabilizing perturbation methods, RD, affects emotion detection. It reveals a consistent pattern for the B setting, our baseline, which proves uniformly unstable across all ten instances. This may stem from the B setting's reliance on specific textual traits that, when disrupted, lose the emotion hints needed for correct identification. Consequently, any perturbation of these elements yields less robust predictions. The B configuration thus appears particularly vulnerable when both syntactic and lexical structure are disturbed, as by random deletion.

The App+B configuration, by contrast, exhibits greater stability, preserving the emotion label in eight of ten cases.[7] Unlike the B setting, which proves to depend more on textual cues, App+B draws on deeper contextual features not strictly tied to lexical-syntactic structure. This broader contextual grounding makes App+B more resilient.

Taken together, these observations indicate that the joint App+B setting owes its greater robustness to drawing on broader contextual cues beyond the lexical-syntactic features the non-joint B setting depends on.

6 Conclusions

We studied whether the robustness of LLM-based emotion detection improves by jointly requesting labels from multiple emotion theories (basic emotions, valence–arousal, or appraisal-based). Our findings show that LLMs are more robust when instructed to predict emotions following more than one theory. Further analysis reveals that the App+B setting considerably stabilizes emotion detection. This effect is most pronounced for word order and least for random word deletions and substitutions.

The divergence between joint (App+B) and non-joint (B) outputs further underscores this effect. While both suffer from random word substitutions and deletions, the non-joint setting relies more on lexical and syntactic cues, yielding less robust predictions, whereas the joint setting captures emotions from a broader contextual basis, yielding more robust outputs. This suggests that combining appraisal and basic emotion theories not only improves robustness but also captures a richer set of emotion cues from the input.

Our research reveals that appraisal and basic emotion theories jointly make LLM predictions more robust. This raises two questions: how LLMs encode appraisals and basic emotions, and which contextual elements this configuration focuses on when predicting. Future research should address both. The first calls for systematic probing of internal model representations. The second demands investigating how the App+B configuration relies on lexical-syntactic context, particularly given that word deletions and substitutions produced the strongest destabilizing effects.

Limitations

Before our emotion-detection configurations can be considered fully robust for text-based emotion analysis, they must be evaluated on a broader scale. Despite promising results in emotion detection robustness, several potential limitations warrant further discussion.

The results we present are based on a limited number of LLMs. Our study is restricted to three of them, Gemma, Llama, and Qwen. Different model families, sizes, and training parameters may respond differently to textual perturbations while predicting emotions, and conclusions drawn from this selection may not be universal. Our model selection also leaves a size gap between the largest model (Llama) and the smaller ones (Gemma and Qwen); adding a mid-size model would help clarify whether the observed robustness gains scale with model size or are specific to the emotion theories tested.

The prompts we used in this study follow a relatively constrained structure, and neither the prompt parameters nor the overall prompt design were systematically varied or optimized. Alternative prompting approaches, such as chain-of-thought or instruction-tuned prompt formats, may yield different robustness patterns for all three configurations.

A further limitation concerns the scope of our robustness assessment. We evaluated the robustness of the categorical emotion labels alone, and did not assess whether the numerical outputs of the VA and APP dimensions within the joint settings are robust in the same manner. This reflects the focus of the present work: our primary interest lies in the categorical emotion labels robustness than in

[7] We also performed a statistical significance test to validate these results, as shown in Appendix B.

the VA and APP dimensional outputs. We therefore leave a systematic evaluation of the robustness of these numerical predictions to future work.

Regarding the text perturbation methods, only six were employed in this study. While these represent a reasonable selection of state-of-the-art approaches, a broader range of perturbation methods could extend and refine the conclusions about which types of textual modifications are most destabilizing across the three configurations.

As noted in Section 2.2, our study evaluates how constant a model's prediction is on the original and on the perturbed text, not correctness against the gold emotion labels that the underlying benchmarks already provide. A configuration could in principle score well on every metric in this paper while being consistently wrong under noisiness. We chose robustness as our primary motivation because it targets deployment reliability rather than clean-text accuracy. An evaluation of robustness alongside gold-label accuracy across configurations is left to future work.

Additionally, all perturbations applied in this study are synthetically generated on selected benchmark texts. Among these benchmarks, there is naturally occurring noisy user-generated text, such as organic typos, slang, or code-switching, since at least one of these benchmarks originates from social media. Nonetheless, we consider evaluating robustness on larger, naturally noisy real-world user-generated corpora an important direction for future work.

Finally, the analysis is constrained to joint and non-joint prompt configurations, namely B, VA+B, App+B. Non-joint configurations, in which basic emotions, valence–arousal, and appraisals are prompted independently, could not be included because there is currently no established method to map their outputs in an all-to-all manner, making direct robustness comparisons across them infeasible.

Ethical Considerations

This work is intended to serve the research community by demonstrating that certain LLM configurations, specifically those in which multiple psychological emotion theories operate in conjunction, can improve the robustness of emotion detection outputs. Our findings aim to illuminate principled criteria for selecting robust configurations for the emotion detection task, thereby reducing unnecessary expenditure of computational resources and time during the configuration search process.

We do not seek to establish the superiority of any emotion theory over the others. Such a claim would be scientifically inappropriate, given that all theories referenced in this work are empirically grounded, theoretically validated, and widely adopted within the research. Rather, our goal is to demonstrate how emotion theories can be mutually reinforcing when applied jointly within LLM-based emotion detection. Similarly, we make no claim that any particular LLM is inherently superior or inferior to others; instead, we examine the extent to which different models may benefit from configurations grounded in combined emotion theories.

Furthermore, emotion analysis carries inherent ethical risks, particularly when applied to manipulate human behavior. These risks may be amplified as LLMs become increasingly capable and convincing in human-computer interaction, potentially enabling more sophisticated forms of emotional manipulation. In our study, we remained mindful of these implications and prioritized transparency and safeguards against misuse when deploying emotion analysis configurations.

All datasets used in this study are publicly available, serve as established benchmarks for emotion analysis in the research community, and contain no personally identifiable information. Their use in this work is strictly limited to advancing research in emotion analysis and does not involve any direct interaction with human participants. All datasets used in this study were employed in accordance with their respective copyright standards, properly cited, and used in compliance with the Creative Commons (CC) license terms.

Finally, AI assistance was used in the preparation of this paper. Specifically, we used the Claude Code extension (version 2.1.143) for Visual Studio Code for programming support, and Claude Sonnet 4.6 for assistance with LaTeX tables formatting and the rephrasing of text for clarity and grammar.

Acknowledgments

This study is funded by CSIC Momentum project (ref. MMT24–IEGPS–01), financed by European Union's Recovery and Resilience Facility-Next Generation, in the framework of the General Invitation of the Spanish Government's public business entity Red.es to participate in talent attraction and retention programmes within Investment 4 of

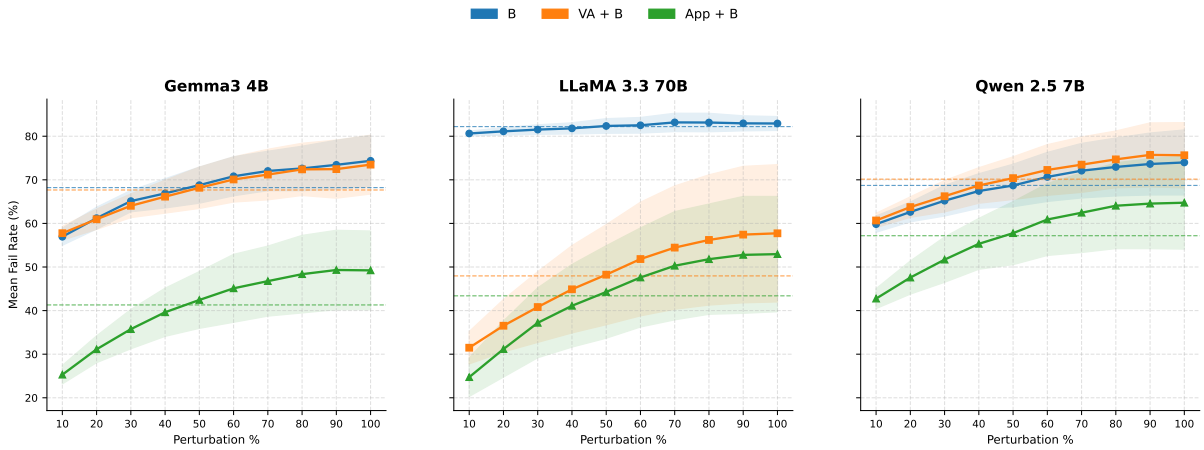

Figure 5: Robustness degradation curves under increasing perturbation fractions per model and setting.

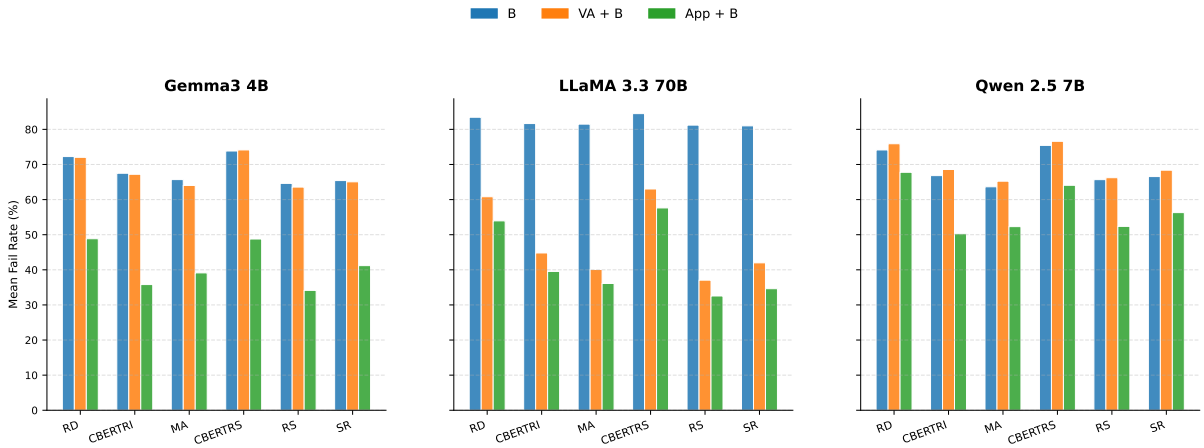

Figure 6: Robustness by perturbation method per model and setting.

Component 19 of the Recovery, Transformation and Resilience Plan. This study was also supported by the "Ministerio de Ciencia, Innovación y Universidades" and "Ayudas para estancias de movilidad en el extranjero José Castillejo para jóvenes doctores 2024" postdoctoral mobility fellowship (ref. CAS24/00327).

A Appendix: CheckList Mean Fail Rate

Figure 5 quantifies emotion predictions robustness using the CheckList framework's Invariance test (Ribeiro et al., 2020), which treats each (original text, perturbed text) pair as a test case and flags a failure whenever the perturbed prediction diverges from the original one. For every model and emotion detection setting (B, VA + B, and App + B), the six perturbation methods (SR, RS, RD, MA, CBERTRS, CBERTRI) were each applied at ten perturbation fractions (10-100%), with ten repeated runs per perturbation fraction. For each (method, fraction) pair, the fail rate was first averaged across the ten runs; these six method-level means were then averaged again within each fraction to produce the plotted point, with the standard deviation across methods rendered as the shaded band. Figure 5 cross-validates that the App + B setting tends to produce more robust emotion labels under perturbations, reported by the pair-wise accuracy metrics in Section 4.1.

Figure 6 summarizes the same INV test results to show which perturbation method degrades each model's label robustness the most. For every (model, emotion detection setting) combination, we obtain a per-run fail rates across all ten fractions and all runs for each method into a single mean. The resulting bars validate previously observed patterns in Section 4.2 that RD and CBERTRS are consistently the most disruptive perturbation methods across all three settings and models. This convergence between two independently computed

Model	Setting	Agree. (A)	Agree. (B)	Diff. (A−B)	Significant
Gemma3	B vs. App+B	0.320	0.583	-0.263	Yes (< 0.0001)
Gemma3	B vs. VA+B	0.320	0.328	-0.007	No (p=0.1614)
Gemma3	App+B vs. VA+B	0.586	0.326	+0.260	Yes (< 0.0001)
Llama	B vs. App+B	0.178	0.552	-0.374	Yes (< 0.0001)
Llama	B vs. VA+B	0.178	0.523	-0.345	Yes (< 0.0001)
Llama	App+B vs. VA+B	0.551	0.522	+0.030	Yes (p=0.0013)
Qwen	B vs. App+B	0.308	0.428	-0.120	Yes (< 0.0001)
Qwen	B vs. VA+B	0.315	0.301	+0.014	Yes (p=0.0343)
Qwen	App+B vs. VA+B	0.428	0.292	+0.137	Yes (< 0.0001)

Table 5: Paired bootstrap significance test resutls. Acc A and Acc B report the pairwise accuracy of the two settings listed in the Setting column: Acc A for the first setting, Acc B for the second. Results with $p < 0.05$ are treated as significant.

metrics (pair-wise accuracy and Checklist mean fail rates), one measuring prediction robustness (pair-wise accuracy) and the other measuring prediction invariance under perturbation, validates the fragility observed across perturbation methods.

B Appendix: Significance Test Results

We performed a bootstrap test (Koehn, 2004) to make sure our results are statistically significant. Table 5 summarizes these results. For each model, we first compute a per-instance accuracy for each setting (B, App+B, VA+B) between original and perturbed texts, aggregating across all perturbation methods, fractions, and runs. We then compare two settings using a bootstrap test: from the original instances with a valid label and at least one valid prediction in both settings, we draw 10,000 resamples with replacement. The p-value is the fraction of resamples in which the accuracy ranking of the two settings flips relative to the ranking actually observed.

Under this test, the App+B joint setting is significantly more accurate than the B non-joint setting across all three models (Gemma3: $\Delta = +0.263$, $p < 0.0001$; Llama: $\Delta = +0.374$, $p < 0.0001$; Qwen: $\Delta = +0.120$, $p < 0.0001$), indicating that incorporating appraisal theory consistently improves accuracy regardless of perturbation method or underlying model.

References

Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.

Abdullah Al Maruf, Fahima Khanam, Md Mahmudul Haque, Zakaria Masud Jiyad, Muhammad Firoz Mridha, and Zeyar Aung. 2024. Challenges and opportunities of text-based emotion detection: A survey. *IEEE access*, 12:18416–18450.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, N. Tepper, and Naama Zwerdling. 2019. Do not have enough data? deep learning to the rescue! In *AAAI Conference on Artificial Intelligence*.

Christopher Bagdon, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. 2024. "you are an expert annotator": Automatic best–worst-scaling annotations for emotion intensity modeling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7924–7936, Mexico City, Mexico. Association for Computational Linguistics.

Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English prompts are better for nli-based zero-shot emotion classification than target-language prompts. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 1318–1326, New York, NY, USA. Association for Computing Machinery.

Mobin Barfi, Sajjad Mehrpeyma, and Nasser Mozayani. 2025. UncleLM at SemEval-2025 task 11: RAG-based few-shot learning and fine-tuned encoders for multilingual emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1563–1569, Vienna, Austria. Association for Computational Linguistics.

Laura Ana Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*,

pages 2104–2119. Association for Computational Linguistics.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25 1:49–59.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Lea Canales and Patricio Martínez-Barco. 2014. Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pages 37–43, Quito, Ecuador. Association for Computational Linguistics.

Federica Cavicchio. 2025. *Emotion Detection in Natural Language Processing*. Springer.

Deniz Cevher, Sebastian Zepf, and Roman Klinger. 2019. Towards multimodal emotion recognition in german speech events in cars using transfer learning. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers*, pages 79–90, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Claudiu Creanga, Teodor George Marchitan, and Liviu Dinu. 2025. Team Unibuc - NLP at SemEval-2025 task 11: Few-shot text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 468–475, Vienna, Austria. Association for Computational Linguistics.

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 and Emotion*, 6(3-4):169–200.

Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Lynn Greschner, Meike Bauer, Sabine Weber, and Roman Klinger. 2026a. Categorical emotions or appraisals - which emotion model explains argument convincingness better? In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8190–8203, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Lynn Greschner and Roman Klinger. 2025. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 628–646, Albuquerque, USA. Association for Computational Linguistics.

Lynn Greschner, Sabine Weber, and Roman Klinger. 2026b. Trust me, i can convince you: The contextualized argument appraisal framework and the contarga corpus. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8327–8346, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. 2020. POEMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1652–1663, Marseille, France. European Language Resources Association.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dawei Huang, Qing Li, Chuan Yan, Zebang Cheng, Zihao Han, Yurong Huang, Xiang Li, Bin Li, Xiaohui Wang, Zheng Lian, and 1 others. 2025. Emotion-qwen: A unified framework for emotion and vision understanding. *arXiv preprint arXiv:2505.06685*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2409.12186*.

Eojin Jeon, Mingyu Lee, Sangyun Kim, Junho Kim, Wanzee Cho, Tae-Eui Kam, and SangKeun Lee. 2025. "going to a trap house" conveys more fear than "going to a mall": Benchmarking emotion context sensitivity for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14848–14869, Suzhou, China. Association for Computational Linguistics.

Gemma Team Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram'e, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, and 191 others. 2025. Gemma 3 technical report. *ArXiv*, abs/2503.19786.

Urja Khurana, Eric Nalisnick, and Antske Fokkens. 2021. How emotionally stable is ALBERT? testing robustness with stochastic weight averaging on a sentiment analysis task. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 16–31, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069, Online. Association for Computational Linguistics.

Jiahui Li, Sean Papay, and Roman Klinger. 2025a. Are humans as brittle as large language models? In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2130–2155, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Jiayi Li, Yingfan Zhou, Pranav Narayanan Venkit, Halima Binte Islam, Sneha Arya, Shomir Wilson, and Sarah Rajtmajer. 2025b. Can third parties read our emotions? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21478–21499, Vienna, Austria. Association for Computational Linguistics.

Jiexi Liu, Ryuichi Takanobu, Jiaxin Wen, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, and Minlie Huang. 2021. Robustness testing of language understanding in task-oriented dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2467–2480, Online. Association for Computational Linguistics.

Albert Mehrabian and James A Russell. 1974. *An approach to environmental psychology.* MIT Press.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Lima Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Dario Mario Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, and 2 others. 2025. SemEval-2025 task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2558–2569, Vienna, Austria. Association for Computational Linguistics.

Lalchand Pandia and Allyson Ettinger. 2021. Sorting through the noise: Testing robustness of information processing in pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1583–1596, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press.

Jonathan Posner, James A. Russell, and Bradley S. Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17:715 – 734.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th*

Annual Meeting of the Association for Computational Linguistics, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Ira J Roseman. 1979. Cognitive aspects of emotion and emotional behavior. In *87th Annual Convention of the American Psychological Association, New York*, volume 58.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Klaus R Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

Klaus R. Scherer and Harald G. Wallbott. 1994. Evidence for the Universality and Cultural Variation of Differential Emotion response Patterning. *Journal of Personality and Social Psychology*, 66(2):310–328.

Hendrik Schuff, Jeremy Barnes, Julian Mohme, Sebastian Padó, and Roman Klinger. 2017. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

Mark Sherer, James E Maddux, Blaise Mercandante, Steven Prentice-Dunn, Beth Jacobs, and Ronald W Rogers. 1982. The self-efficacy scale: Construction and validation. *Psychological reports*, 51(2):663–671.

Craig A Smith and Phoebe C Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813.

Anna Soligo, Vladimir Mikulik, and William Saunders. 2026. Gemma needs help: Investigating and mitigating emotional instability in llms. *Preprint*, arXiv:2603.10011.

Marco Antonio Stranisci, Simona Frenda, Eleonora Ceccaldi, Valerio Basile, Rossana Damiano, and Viviana Patti. 2022. APPReddit: a corpus of Reddit posts annotated for appraisal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3809–3818, Marseille, France. European Language Resources Association.

Enrica Troiano. 2024. Where are emotions in text? a human-based and computational investigation of emotion recognition and generation. *Journal for Language Technology and Computational Linguistics*, 37(1):15–26.

Enrica Troiano, Roman Klinger, and Sebastian Padó. 2023a. On the relationship between frames and emotionality in text. *Northern European Journal of Language Technology*, 9.

Enrica Troiano, Sofie Labat, Marco Antonio Stranisci, Rossana Damiano, Viviana Patti, and Roman Klinger. 2024. Dealing with controversy: An emotion and coping strategy corpus based on role playing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1634–1658, Miami, Florida, USA. Association for Computational Linguistics.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023b. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Enrica Troiano, Laura Oberländer, Maximilian Wegge, and Roman Klinger. 2022. x-enVENT: A corpus of event descriptions with experiencer-specific emotion and appraisal annotations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1365–1375, Marseille, France. European Language Resources Association.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science – ICCS 2019*, pages 84–95, Cham. Springer International Publishing.

Ibtissen Yacoubi, Radhia Ferjaoui, Warith Eddine Djeddi, and Anouar Ben Khalifa. 2025. Advancing emotion recognition through llama3 and lora fine-tuning. In *2025 IEEE 22nd International Multi-Conference on Systems, Signals and Devices (SSD)*, pages 348–353.

Jingxiang Zhang and Lujia Zhong. 2025. Decoding emotion in the deep: A systematic study of how llms represent, retain, and express emotion. *arXiv preprint arXiv:2510.04064*.

Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2025. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Neural Networks*, 192:107901.

Jun Zhao, Kang Liu, and Liheng Xu. 2016. *Sentiment analysis: Mining opinions, sentiments, and emotions*. MIT Press.