Characterizing and Evaluating Working Emotion Vocabularies in Multilingual Large Language Models

Nicholas Deas[*1], **Iván Pérez Mejía**[*1], **Ellie Yang**[2], **Kathleen McKeown**[1]

[1]Department of Computer Science, Columbia University,
[2]Department of Computer Science, Amherst College
[ndeas,kathy]@cs.columbia.edu, iep2110@columbia.edu, eyang27@amherst.edu

Abstract

Prior work evaluating emotion and affective understanding in large language models (LLMs) typically rely on predetermined label sets or focus on a singular evaluation task (e.g., emotion detection). We consider affective states, referring to the much broader variety of terms people use to label their emotional experiences. We evaluate multilingual language models' understanding of affective states in English and Spanish through three different tasks: 1) *identification*, where models predict an affective state given text, 2) *expression*, where models generate text expressing a given affective state, and 3) *verification*, where models report whether a given term refers to an affective state. We show that performance on one task is not necessarily predictive of performance on another. Using these three tasks, we then begin to explore when and why models struggle to understand particular affective states compared to others. We examine systematic patterns in the affective state terms that are well and poorly understood by models, characterizing the working emotion vocabulary of LLMs. [1]

1 Introduction

Benchmarks and approaches for emotion detection typically prescribe basic or universal psychological models of emotions (e.g., Ekman basic emotions (Ekman, 1984); emotion sets found in Wang et al. (2020)). Recent work, however, has begun considering larger emotion sets (Demszky et al., 2020) as well as introduced benchmarks with greater coverage of world languages (Bianchi et al., 2022). Such studies are vital to ensuring that LLMs have a robust understanding of more nuanced and complex emotion expression in language.

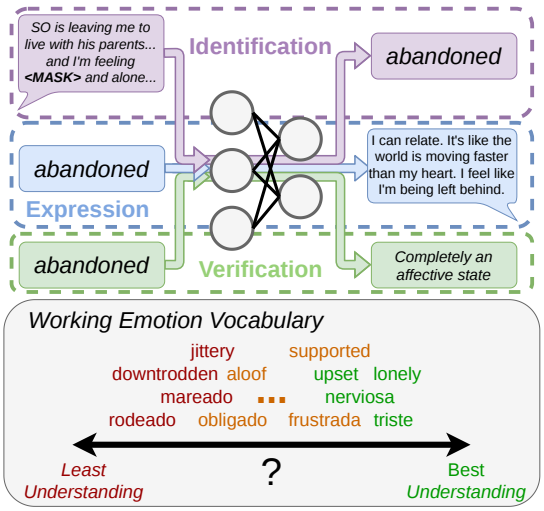


Figure 1: Through model performance on three tasks–affective state identification, expression, and verification–we use notions of *working emotion vocabularies* to study which affective states multilingual models least and best understand in English and Spanish.

Most recently, Deas et al. (2024) depart from traditional emotion detection and avoid a predetermined label set. Instead, they evaluate models' ability to predict any term used by an author to label their own feelings or emotions. The authors refer to these terms as *affective states*. In contrast to traditional emotion detection tasks, predicting affective states is shown to be difficult for recent large language models (LLMs) in both English and Spanish, warranting further investigation into *why* they exhibit poor performance in this setting. Furthermore, this and other prior work focus primarily on models' overall ability to detect emotion and affect in language, but we explore how jointly considering other tasks and evaluation metrics may reveal trends in model performance that were not otherwise apparent.

Accordingly, in this work, we investigate LLMs' understanding of nuanced affective expression with

[1]We make our code available at https://github.com/NickDeas/WorkingEmotionVocabulary/

[*]Equal contribution

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 47272–47294
July 2-7, 2026

multiple tasks to thoroughly analyze the affective states that models appear to least and best understand. We first follow Deas et al. (2024) in studying *affective state identification*, but evaluate a significantly larger set of recent models. Beyond this task, we additionally evaluate models on *affective state expression*, requiring models to generate text as if written by an author expressing a given affective state, and *affective state verification*, where models are expected to report the extent to which a given term in context refers to an affective state (e.g., compared to a purely physical sensation, "I feel sore").

Using these tasks, we seek to more comprehensively analyze where models succeed and fail in affective tasks by conducting a fine-grained analysis of model performance on particular affective states. Specifically, we draw on psycholinguistics studies of *working emotion vocabularies*, the set of readily accessible terms available to someone in labeling their emotional experience (Schrauf and Sanchez, 2004). Beyond language and culture-level differences in the conceptualization of emotions, individuals can further differ in the salient subset of a language's emotional lexicon that they tend to use at any given time. For example, one person may tend to use basic emotion labels (e.g., *happy*, *angry*, *sad*), while another may tend to use more precise terms (e.g., *ecstatic*, *enraged*, *discontent*) despite all terms being well understood by both individuals.

Similarly, despite being trained on text that likely includes a wide variety of emotional language, LLMs may show preferences or systematically superior understanding of particular affective states over others. These preferences may then hinder their understanding of affect in text written by diverse speakers. Therefore, we investigate LLMs' understanding of nuanced affective expression in language and aim to characterize LLMs' analogues to working emotion vocabularies. Due to the close relationship between language, cultural background, and emotion vocabularies, we specifically focus on multilingual LLMs and evaluate them across English and Spanish. We investigate the following research questions:

RQ1: *How well do multilingual LLMs understand nuanced affect in language across different tasks?* We benchmark 23 multilingual LLMs' affective understanding by evaluating their ability to identify nuanced affect in text, express a given affective state, and verify whether a term refers to an affective state. We show that generally, **models struggle to predict nuanced affect** (*identification*), **instruction-tuned models particularly struggle to express given affective states** (*expression*), and **few models align with humans in judging what constitutes a typical affective state** (*verification*).

RQ2: *How does understanding of affective states vary across tasks and models?* We evaluate whether understanding of different affective states is consistent across the three tasks for a given model, as well as whether there exist patterns across models. While we find some consistent patterns across models, we also find that **there is little to no agreement among tasks, suggesting individual tasks provide a limited perspective on models' affective understanding**.

Finally, **RQ3:** *What psycholinguistic factors characterize LLMs' working emotion vocabulary?* We explore the extent to which different psycholinguistic properties (e.g., valence, arousal, age of acquisition) of affective states may explain variation in LLM performance and behavior across tasks. We show that **frequency** of a given term in CommonCrawl appears to be a **significant predictor of model performance**, positively associated with model performance in Spanish, but negatively in English. We also see trends in psycholinguistic variables that are contrary to patterns in human studies, likely due to the composition of model pretraining corpora.

Overall, we show that relying solely on detection tasks provides an incomplete picture of models' affective understanding, and that future work should consider a variety of diverse tasks to thoroughly characterize understanding. Furthermore, we use performance across our tasks to investigate the working emotion vocabulary of LLMs, identifying systematic patterns that may shed light on when models struggle to understand affective states.

2 Background and Related Work.

Language and Emotions. Prior psycholinguistics work has distinguished terms that are considered *emotion-label words* (e.g., happy, afraid) which explicitly reference affective states, and *emotion-laden words* (e.g., celebration, tragedy) which express or elicit emotions without directly indicating a particular emotion (Pavlenko, 2008; Betancourt et al., 2024). Similarly, other work measures the prototypicality of emotion-label terms, considering how well a term represents the broader

category comprising emotions (Pérez-Sánchez et al., 2021; Wu et al., 2025). Such distinctions, however–and accordingly, differences in individuals' working emotion vocabulary–are realized differently according to cultural and linguistic background. For example, while distributions of positive, negative, and neutral emotion terms are largely similar across linguistic and cultural backgrounds (Bąk and Altarriba, 2024; Schrauf and Sanchez, 2004), *frustration* has been found to be a more salient *anger* term in English compared to other high-resource languages (Soriano and Ogarkova, 2025) and languages like Javanese can have notably diverse variation in terms for *sadness* (Suswandi and WS, 2017).

LLM Emotion Understanding. Traditionally, work in emotion analysis has focused on detection in language (e.g., Demszky et al. 2020; Bianchi et al. 2022; Li et al. 2017) including in the evaluation of LLMs (Liu et al., 2024b; Deas et al., 2024; Boutouta et al., 2025). Rather than predicting emotions in text, some studies have investigated producing text that expresses an indicated emotion through style transfer and controllable generation (MohammadiBaghmolaei and Ahmadi, 2023; Xie and Agrawal, 2023), emotional dialogue generation (Liu et al., 2024a; Song et al., 2019; Ma and Chang, 2024), and other task contexts (Zhang et al., 2019; Ishikawa and Yoshino, 2025). Rather than restrict evaluation to a single task, Sabour et al. (2024) aim to broadly evaluate emotional intelligence in LLMs. These researchers' study considers both emotion understanding (e.g., emotion detection) and emotional application (e.g., responses to emotion-laden dilemmas). In contrast to prior work, we focus on more comprehensive evaluation of models' linguistic understanding of particular affective states. We leverage affective state terms derived from natural language and structure our evaluation to study performance at the term-level.

3 Methods

3.1 Affective Understanding Tasks

We consider three tasks capturing different aspects of affective understanding: affective state identification, verification, and expression. **Affective state identification** frames emotion detection as a generative task (Deas et al., 2024). Models are provided with input text where affective state terms used by the texts' author are masked. Then, models are expected to predict these masked terms. Finally, **affective state verification** asks models to report whether an indicated term represents an affective state. Each term is provided with context to enable distinguishing cases that depend on the use of the term (e.g., "blue" as a color or as a descriptor of feeling). Finally, **affective state expression** requires models to generate a response to a given user prompt that expresses a target affective state. Illustrative examples of these tasks and their associated prompts are included in Figure 1 and Figure 2 respectively.

3.2 Data

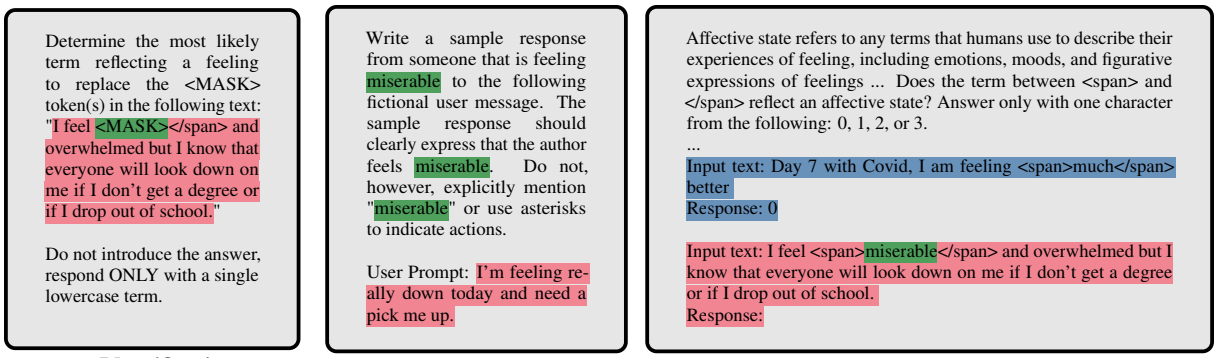
We draw data from the MASIVE dataset (Deas et al., 2024), consisting of English and Spanish Reddit posts that include explicit descriptions of the authors' emotional experiences (e.g., "I feel *frustrated*"). The researchers design an iterative procedure beginning with a seed set of affective state terms (e.g., the Ekman emotions (Ekman, 1984)) and a set of query templates (e.g., "I feel ___"). The templates used allow them to heuristically identify other affective state terms that appear in returned texts and include these terms in the next round of queries, resulting in a large set of unique affective states (∼1,600 in English, ∼1,000 in Spanish). The resulting data represents the discrete terms that people naturally use to label their own emotional experience and associated contexts.

Lang.	Split	Size	Unq. States	Len.	Score Dist. (%)			
					0	1	2	3
En	Exs	150	106	22.8	8.0	2.7	5.3	84.0
	Test	49	41	26.4	8.2	6.1	26.5	59.2
	Ovr	199	131	23.7	8.0	3.5	10.6	77.9
Es	Exs	150	113	63.8	27.3	3.3	8.0	61.3
	Test	50	45	68.5	20.0	2.0	22.0	56.0
	Ovr	200	143	65.0	25.5	3.0	11.5	60.0

Table 1: Summary data statistics with (rounded) percentage of words annotated within each score range by human annotators. *Exs* represents the single-coded samples used as few-shot examples, while *Test* represents the double-coded samples used for all evaluations. *Unq. States* represents count of unique affective states. [2]

A subset of text samples in this work were also annotated by 2 native English and Spanish speakers. These annotations aimed to identify whether the automatically detected terms represent affec-

[2]Length is measured as token count with the Llama-3.1-8B tokenizer (Grattafiori et al., 2024). Unique affective state counts for "Exs" and "Test" do not sum to that of "Ovr" due to partial overlap in affective states between the two subsets.



(a) **Identification** (b) **Expression** (c) **Verification**

Figure 2: Condensed prompt templates for the three tasks. Red text represents input texts drawn from MASIVE, green text represents a masked or target affective state term, and blue text represents in-context examples in the verification task.

tive states (e.g., rather than purely physical states, "I feel *sore*") in order to validate the data collection process. Judgments are assigned on a 4-point Likert scale, where 0 and 3 represent "*not an affective state*" and "*completely an affective state*" respectively. For all tasks, we leverage this annotated subset of MASIVE. Notably, these ratings are inherently subjective, and annotators may draw different meanings from how a particular term is used. Therefore, we avoid filtering the candidate affective states in this dataset given the ambiguity of selecting an appropriate threshold. To mitigate this variability, however, we extract those overlapping samples (41 and 45 affective states in English and Spanish respectively) and use average annotator judgments to construct a test set throughout experiments for RQ1 (Subsection 4.1) and RQ2 (Subsection 4.2). We instead leverage the non-overlapping texts coded by only one annotator as a pool of few-shot examples for the verification experiments. In later analyses that do not depend on annotator labels (RQ3 in Subsection 4.3), we use the full set of 131 and 143 affective states in English and Spanish respectively. Finally, in experiments involving binary model outputs, we treat human annotator ratings ≥ 2 as positive (i.e., judged to represent affective states) and < 2 as negative (i.e., judged to not represent affective states).

Summary statistics and the distributions of annotator ratings are shown in Table 1. Notably, a majority of terms included in the annotated subsets for English and Spanish are judged to represent affective states. Examples of texts, automatically-identified candidate affective state terms, and the associated annotators' ratings are shown in Table 2. See Deas et al. (2024) for additional details on data and human judgment collection.

Lang	Context	Score
En	I feel jittery waking up but its probably left over anxiety.	1
	I'm so anxious, hoping I'm paranoid, but pissed of that I feel manipulated into being fearful vs. having a normal relationship with someone.	3
Es	Tengo una hipoteca. Odio a los corruptos, la falta de meritocracia y la injusticia. Estoy rodeado de gente en paro, de empresas que no han sido pagadas por los servicios prestados y de hipotecados.	0
	He tratado de hacerlo muchas veces pero nunca ha dado resultado, incluso llendo a bailar con pololas o amigos el resultado es el mismo, nose donde poner mis brazos ni como moverme, me siento incomodo	3

Table 2: Example texts and affect ratings on a 4pt-Likert scale, higher being "more like an affective state."

Expression Prompt Data. To evaluate models in the expression task, we form a dataset of input prompts to elicit responses conditioned on an affective state that should be expressed. For the input prompts, we use texts from the same subset of MASIVE as used in other experiments in order to encourage models to provide emotion-laden responses. We consider all possible pairs of input prompts and target affective states, allowing us to generate a diverse set of model responses for each affective state.

Working Emotion Vocabulary Measures. To characterize LLMs' working emotion vocabularies, we consider a variety of word-level characteristics and psycholinguistic norms collected in prior work. Namely, we consider the following measures that have been previously studied in the context of affect and language: valence and arousal (Mohammad, 2018; Stadthagen-Gonzalez et al., 2016), human age of acquisition (AoA) (Kuperman et al., 2012; Alonso et al., 2014), and concreteness [3] (Brysbaert

[3]While concreteness norms for Spanish terms exist, we do

et al., 2013). We additionally consider frequency in the July 2025 CommonCrawl snapshot measured using infinigram-mini (Xu et al., 2025) as a proxy of how often each term may be seen during model pretraining. We use all of these variables as inputs to regression models predicting model performance and behavior in each task in order to characterize working emotion vocabularies. Additional details on these measures can be found in Appendix A.

3.3 Experimental Configuration

Identification. We use the same setup and prompt as described in Deas et al. (2024) and shown in Figure 2a, framing affective state identification as a masked-span prediction task. As input, models are provided with a text where mentions of affective states by the author of that text are masked. Models are then tasked with filling each mask with the terms most likely used by the original author to describe their feelings.

Expression. For affective state expression experiments, we instruct the model to generate a response to a provided user prompt and to ensure that the response clearly expresses a given affective state. [4] An example of the prompt format is shown in Figure 2b.

Verification. Following prior work on the evaluation of values and opinions in LLMs (Röttger et al., 2024), we evaluate LLMs on the verification task across a variety of evaluation formats. Specifically, we consider two output formats: binary (Yes/No) and ratings (Likert scores ranging from 0 to 3). An example of the rating format prompt is shown in Figure 2c. These formats reflect different levels of granularity in evaluating whether a given term is an affective state. We also explore two approaches to processing predictions: label classification, where we select the relevant token with the highest likelihood, and a probability-based approach, where we compute a continuous score by averaging over possible output tokens (i.e., either yes/no or Likert ratings) weighted by their associated probability. In summary, we test four variants: 1. **Binary (class)**: Most probable Yes/No response. 2. **Binary (prob)**: Weighted average over Yes/No responses. 3. **Rating (class)**: Most probable Likert-scale ratings. 4. **Rating (prob)**: Weighted average over Likert-scale ratings.

In addition to varying task formats, we experiment with few-shot prompting. We randomly sample texts and ratings coded by a single annotator from MASIVE (*Exs.* in Table 1) as in-context examples. We include k-examples in each prompt for $k \in \{0, 1, 3, 5\}$. The few-shot examples for each input are held constant across models. Additional details on the verification task setup are included in Appendix B.

3.4 Metrics.

Identification. We adopt the top-k similarity metric used in the original study (Deas et al., 2024) with $k \in \{1, 3, 5\}$. To account for affective state terms with context-dependent meanings (e.g., "blue"), we use a multilingual BERT model (Devlin et al., 2019) to produce contextualized embeddings, and extract the representation of each term within the context of the original text. Scores are then calculated as the maximum cosine similarity between the contextualized embeddings of the top-k predictions and ground truth terms.

Expression. To evaluate models in the expression task, we automatically detect the most likely affective state represented in each model-generated response and compare to the target affective state passed in the original prompt. Following prior work studying emotion expression and emotion-conditioned generation (Xie and Agrawal, 2023; Ishikawa and Yoshino, 2025) we detect affective states in model responses using a surrogate model. Specifically, given that they were shown to outperform larger LLMs in identifying affective states, we use two mT5 models, each finetuned separately on the English and Spanish subsets of MASIVE (Deas et al., 2024). To calculate scores, we first extract the top-5 predictions from the appropriate mT5 model. We then compare the surfaced affective states from mT5 to the target affective state using the top-k similarity metrics described above. We assess and discuss the validity of employing mT5 as part of our metric in Appendix C. Overall, this results in a proxy measure of an LLMs' ability to express a given target affective state.

Verification. We directly compare model predictions with human annotations and compute agreement using several metrics. First, we compute Spearman's ρ correlation between model predictions for each output format and the scores assigned by annotators. We additionally compute Cohen's κ

no include concreteness in the Spanish experiments due to the lack of coverage of the affective states we study.

[4] In earlier experiments, we found that passing target affective states into the system prompt is ignored by instruction-tuned models like OLMo 3. Therefore, we opt to include the instructions in a user prompt for all models.

for classification output formats. Throughout the presented results, we primarily focus on the correlation metric because it allows us to compare across all output formats, while Cohen's κ is limited to these classification formats.

Working Emotion Vocabularies. We analyze the linguistic and psycholinguistic characteristics that may explain variation in model understanding among affective states. In other words, we aim to identify classes of affective state terms that consistently result in high performance as well as those that result in low performance. To do so, we fit ordinal logistic regression models using affective state characteristics as predictors and a summary score of a model for each affective state across the three tasks as the response variable. [5] Again, in contrast to the experiments conducted in Subsection 4.1 and Subsection 4.2, we consider the full set of unique affective states (131 in English and 143 in Spanish) in the annotated subset of Deas et al. (2024), including both single and double-coded samples. To summarize model performance across tasks, we normalize each models' top-5 similarity scores in the identification task, weighted rating output in the verification task, and top-5 similarity scores in the expression task and take the average normalized score. Finally, we rank each affective state based on the aforementioned average score, with lower ranks indicating better performance. This allows us to identify high-level patterns that may explain variation in models' capabilities to identify, verify, and express each affective state.

Through this regression model, the coefficients provide insight into the relationship between each psycholinguistic measure and model understanding: negative coefficients indicate an increase in a variable is associated with a decrease in affective state rank (i.e., an increase in model performance), while positive coefficients indicate that a decrease in a variable is associated with an increase in affective state rank (i.e., a decrease in model performance). Additional details on the regression analyses are included in Appendix D.

3.5 Models

We evaluate 23 multilingual models in total with sizes between 4 and 8 billion parameters (a full list is included in Appendix E). Specifically, we evaluate Llama, Bloom, Mistral, Deepseek, OLMO, and Qwen model families. To measure the impact of different training strategies, we consider both base and instruction-tuned model variants, as well as model variants finetuned specifically for emotion analysis tasks (Liu et al., 2024b). For brevity throughout our main experiments, we present results for a subset of 8 models (base and instruction-tuned variants of 4 models) that show significant correlation with human judgments in the verification task while representing different model families. Full results for the remaining models are shown in the Appendices.

[5] We opt for an ordinal regression to avoid assuming that the averaged performance across tasks represents an interval scale–that is, that the absolute difference in average performance for two affective states is meaningful. Accordingly and as our primary aim focuses on broad patterns in model performance, we avoid making interval-scale claims.

4 Results

4.1 RQ1: Benchmarking Affective Understanding

	Model	Top-k Similarity ↑		
		1	3	5
English	OLMo 3 Base	37.1% ± 11.3	37.3% ± 11.4	37.3% ± 11.4
	OLMo 3	46.2% ± 19.4	58.6% ± 20.2	63.0% ± 19.9
	Llama 3.1 Base	30.3% ± 9.7	30.4% ± 9.6	30.4% ± 9.6
	Llama 3.1	44.8% ± 13.6	51.5% ± 15.2	56.0% ± 16.9
	Qwen 3 Base	34.2% ± 14.0	38.1% ± 17.2	38.3% ± 17.1
	Qwen 3	45.8% ± 13.9	55.8% ± 16.0	57.7% ± 16.4
	Ministral 3 Base	30.3% ± 8.4	30.8% ± 8.5	30.9% ± 8.5
	Ministral 3	43.3% ± 15.2	53.7% ± 16.9	58.4% ± 16.6
Spanish	OLMo 3 Base	30.7% ± 8.0	30.9% ± 8.1	30.9% ± 8.1
	OLMo 3	42.9% ± 14.2	49.4% ± 13.4	53.3% ± 14.4
	Llama 3.1 Base	39.8% ± 10.6	40.0% ± 10.7	40.1% ± 10.6
	Llama 3.1	44.3% ± 16.8	52.5% ± 17.2	56.8% ± 18.3
	Qwen 3 Base	37.7% ± 10.9	38.9% ± 11.0	40.2% ± 11.3
	Qwen 3	40.6% ± 15.5	49.5% ± 15.1	56.2% ± 17.4
	Ministral 3 Base	38.3% ± 10.8	38.4% ± 10.8	38.4% ± 10.8
	Ministral 3	45.7% ± 19.8	58.4% ± 20.3	60.9% ± 21.1

Table 3: Top-k similarity scores with ± 2SE's for a subset of models in the identification task. Scores are shaded relative to those of other models for each value of k (darker indicating better performance).

Identification. We first evaluate model performance on affective state identification to gauge models' capability to predict affect expressed in text. Top-k similarity results for the identification task are shown in Table 3. We reaffirm prior results for a much larger set of LLMs: LLMs underperform the significantly smaller, finetuned mT5 models (71.1% and 78.1% top-5 similarity for English and Spanish respectively; Deas et al. 2024). Among the models, instruction-tuning appears to consistently improve performance for both English and Spanish. Common instruction datasets (e.g., Flan (Wei et al., 2022)) explicitly include sentiment and emotion tasks, which likely offers improvements in affective state identification performance.

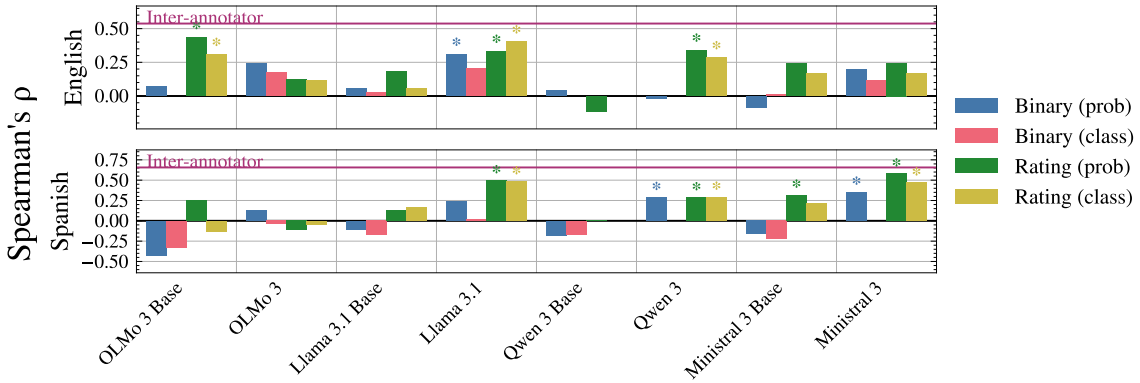


Figure 3: Spearman's correlations between model outputs and human annotations across verification task formats. Agreement between annotators is indicated by the horizontal line. * indicates significant correlation ($p \leq .05$)

Full results for the identification task are included in Appendix F.

	Model	Top-k Similarity↑		
		1	3	5
English	OLMo 3 Base	42.5% ± 0.7	51.4% ± 0.7	55.6% ± 0.7
	OLMo 3	45.4% ± 0.6	55.3% ± 0.6	60.3% ± 0.6
	Llama 3.1 Base	42.7% ± 0.6	51.5% ± 0.6	55.5% ± 0.6
	Llama 3.1	38.4% ± 0.7	49.1% ± 0.7	54.1% ± 0.7
	Qwen 3 Base	40.1% ± 0.9	50.4% ± 0.9	54.3% ± 0.9
	Qwen 3	43.3% ± 0.6	53.1% ± 0.7	57.0% ± 0.7
	Ministral 3 Base	39.4% ± 0.9	50.3% ± 0.9	54.9% ± 0.9
	Ministral 3	42.1% ± 0.6	52.2% ± 0.7	56.1% ± 0.7
Spanish	OLMo 3 Base	38.2% ± 0.9	48.5% ± 0.9	53.5% ± 0.9
	OLMo 3	40.2% ± 0.8	50.8% ± 0.8	56.7% ± 0.9
	Llama 3.1 Base	38.9% ± 0.7	47.9% ± 0.7	52.8% ± 0.8
	Llama 3.1	36.3% ± 0.7	46.3% ± 0.8	51.4% ± 0.8
	Qwen 3 Base	45.3% ± 1.2	56.0% ± 1.1	61.2% ± 1.1
	Qwen 3	36.7% ± 0.8	48.7% ± 0.9	55.2% ± 0.9
	Ministral 3 Base	46.8% ± 1.2	57.6% ± 1.2	62.5% ± 1.1
	Ministral 3	34.2% ± 0.7	44.8% ± 0.8	50.4% ± 0.8

Table 4: Top-k similarity scores with ± 2SE's for a subset of models in generating responses expressing a given affective state. Scores are shaded relative to those of other models for each value of k (darker indicating better performance).

Expression. As opposed to *only* identifying affective states given text, we additionally evaluate models' ability to do the opposite–to generate text expressing a given affective state. Table 4 presents the top-k similarity results for each model in English and Spanish. In contrast to the identification task, instruction-tuning does not consistently lead to improved performance in expressing affective states. Upon closer examination of model outputs, this may be due to instruction-tuning encouraging models to provide human-preferred responses to the given user prompt over expressing a given affective state. For example, the instruction-tuned models often provided empathetic or consoling responses to prompts mentioning negative emotions regardless of the affective state the model is instructed to express. Full results for the expression task are included in Appendix G.

Verification. To evaluate alignment between LLMs' reports with human annotations of whether terms represent affective states, we first consider zero-shot performance across all models. Figure 3 summarizes the model-annotator correlations across the four output formats. Few models (4-5 of those shown) exhibit significant correlations with human judgments with any output format in either language, likely due to lack of calibration of model predictions and the inherent subjectivity of the task. Similarly to the identification task, models with agreement closest to the inter-annotator agreement baseline tend to be instruction-tuned models. Among prompting formats, relying on binary ratings leads to low or even negative correlations across models, suggesting coarse-grained reports may not be robust. Instead, Likert-scale rating formats–and particularly those that leverage output probabilities–appear more robust and more closely approximate human judgments for some models. Due to this, we use models' rating probabilities for later analyses of behavior in the verification task. Full results including Cohen's κ for classification output formats are included in Appendix H.

In Figure 4, we present changes in models' performance across different numbers of in-context examples. For most models–namely, instruction-tuned models–few-shot examples tend to improve alignment with human judgments, and in some cases, even exceed inter-annotator agreement. This suggests that models have some ability to adapt to subjective human ratings in this setting and that few-shot examples are able to improve alignment. Given that our aim in this work is to characterize each models' inherent understanding of different

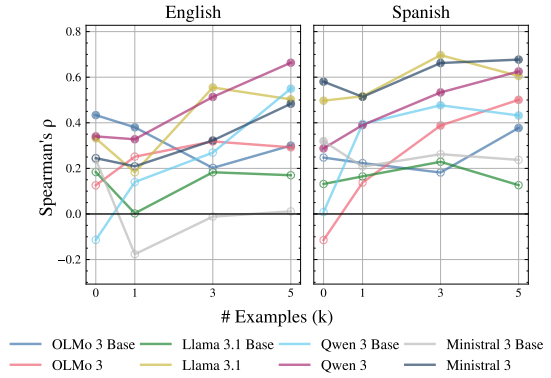


Figure 4: Spearman's correlations between models' predictions (rating probability) and human judgments in the verification task. Filled points ● indicate significant correlations ($p \leq .05$) compared to unfilled ○.

affective states, however, we focus on the 0-shot model predictions in later analyses.

4.2 RQ2: Model and Task Comparison

Table 5 presents the Spearman's correlations at the affective state level between task performance measures for select models. While we might expect that performance is correlated across tasks for any given model, we find only weak to moderate correlations in select cases. Among pairs of tasks, identification and expression task performance ($I{\times}E$ in Table 5) appear consistently positively correlated ($\rho = .09$ to .45). The verification task, however, shows inconsistent and rarely significant correlation with either the identification ($I{\times}V$, $\rho = -.06$ to .29) or expression ($V{\times}E$, $\rho = -.08$ to .47) tasks, supporting the lack of robustness of LLMs' reports on this task. Furthermore, consistency across tasks varies among models and between languages, with Llama 3.1, for example, exhibiting positive correlations among all pairs of tasks in Spanish but little correlation in English. Overall, the lack of strong or consistent correlations suggests that individual tasks do not provide a complete evaluation of models' affective understanding and that models should be evaluated from multiple perspectives.

We additionally qualitatively examine the affective states for which each model performs best and worst across tasks according to average rank (examples included in Appendix I.) Qualitatively, we see that models perform consistently well for select affective states (e.g., *abandoned* in English and *afortunada* in Spanish), even across languages (e.g., *nervous* and *nerviosa* appear highly ranked). In contrast, models perform consistently worse for terms that the original annotators judged to not represent affective states (e.g., *wide* and *jittery*). Given these qualitative trends observed across models, in the next section, we quantitatively analyze systematic patterns in affective state characteristics and model performance.

Model	I×V	I×E	V×E
English			
OLMo 3	0.11	0.39*	−0.01
Llama 3.1	0.04	0.12	−0.07
Qwen 3	0.08	0.40**	0.14
Ministral	−0.06	0.45**	−0.08
Spanish			
OLMo 3	0.10	0.15	0.18
Llama 3.1	0.29†	0.30*	0.47**
Qwen 3	0.06	0.23	0.13
Ministral	0.14	0.09	0.35*

Table 5: Affective state-level Spearman's correlations between performance measures in the **I**dentification, **V**erification, **E**xpression. Superscripts indicate significant correlations, $^{**}p \leq .01$, $^{*}p \leq .05$, $^{\dagger}p \leq .075$.

4.3 RQ3: Working Emotion Vocabularies

Regression coefficients for each affective state measure predicting model performance across tasks are shown in Table 6 for English and Spanish. The sign of notable coefficients is fairly consistent across models, suggesting these patterns may be largely independent of model architecture and training. Significant predictors, however, do vary across models, suggesting differences in the strength of relationships between predictors and performance.

Among predictors, term frequency is most consistently predictive of task performance across models. Interestingly, however, frequency is positively predictive of affective state rank (negatively predictive of performance) in English and negatively predictive of affective state rank (positively predictive of performance) in Spanish. This may suggest that for languages less-represented than English like Spanish, model understanding is most closely associated with representation in pretraining data, but for English as a heavily represented language, frequency alone is less important for distinguishing affective states. Among psycholinguistic characteristics, concreteness tends to have a positive coefficient (significantly for Qwen 3) in English, while valence (significantly for Qwen 3) and arousal (significantly for Qwen 3 Base and Ministral 3 Base)

		OLMo 3 Base	OLMo 3	Llama 3.1 Base	Llama 3.1	Qwen 3 Base	Qwen 3	Ministral 3 Base	Ministral 3
English	Freq	0.63^{*}	$0.42^{\dagger}$	0.67^{**}	$0.49^{\dagger}$	$0.43^{\dagger}$	0.57^{*}	0.11	$0.48^{\dagger}$
	Valence	-0.18	-0.04	-0.18	-0.09	-0.13	-0.00	-0.15	0.11
	Arousal	0.08	-0.03	-0.04	-0.07	0.15	-0.14	-0.14	-0.11
	Conc	0.19	0.21	0.07	0.09	0.15	0.35^{*}	-0.06	0.22
	AoA	-0.28	0.01	-0.51^{*}	-0.28	0.08	0.03	0.07	-0.35
Spanish	Freq	-0.51^{**}	-0.13	-0.53^{**}	-0.48^{*}	-0.90^{**}	-0.03	-0.94^{**}	0.18
	Valence	-0.10	0.29	0.02	0.30	0.07	$0.38^{\dagger}$	-0.01	0.05
	Arousal	0.24	-0.16	0.05	0.14	$0.36^{\dagger}$	-0.01	0.42^{*}	-0.01
	AoA	0.22	0.53^{**}	0.42^{*}	$0.35^{\dagger}$	0.05	$0.36^{\dagger}$	0.26	0.27

Table 6: Coefficients of the Ordinal Regression Model predicting affective state rank in English (top) and Spanish (bottom). Superscripts indicate significant coefficients, $^{**}p \leq .01$, $^{*}p \leq .05$, $^{\dagger}p \leq .075$.

both tend to have positive correlations in Spanish. These trends are counter to the relationships between psycholinguistic measures and emotion word processing in humans (e.g., higher valence is typically associated with faster word processing; Wu et al. 2023), suggesting models do not necessarily learn or understand emotion words as humans do.

Finally, AoA tends to have a negative coefficient (significantly for Llama 3.1 Base) in English, but a positive coefficient in Spanish (significantly for OLMo 3, Llama 3.1 Base and approaching significance for Llama 3.1 and Qwen 3).[6] The relationship between AoA and model performance in Spanish aligns with human studies that show a processing advantage for terms acquired earlier in life (Wu et al., 2023). The reverse relationship for English, however, may be due to the much higher representation of English data leading to more pronounced differences in human and model language acquisition; models are not necessarily trained on the typical style or amount of language seen during early human language acquisition (e.g., child-directed speech; Haga et al. 2024). Generally, we find a few initial trends among these results and believe considering a broader set of languages, affective states, and psycholinguistic characteristics to more comprehensively understand these relationships is a promising direction for future work.

5 Conclusion

In this work, we evaluate multilingual LLMs' understanding of affective states through three tasks: identifying affective states, generating text expressing a given affective state, and verifying whether a given term represents an affective state. We find that while instruction-tuning results in increased identification performance and alignment with human judgments in the verification task, instruction-tuned models struggle to generate text expressing a given affective state. Additionally, we find that model performance on one task is not consistently predictive of performance on others, suggesting that evaluations focusing on a singular task are inadequate to measure models' understanding of affective states. Through these tasks, we then begin to characterize the working emotion vocabulary of LLMs by studying psycholinguistic measures predictive of model performance and behavior. We find that term frequency is a consistently significant predictor of overall performance across models (negatively in English, positively in Spanish) as well as discusss trends among psycholinguistic variables that are contrary to patterns in human emotion word processing, likely due to model pretraining data composition.

Limitations

We note limitations accompanying our findings. First, we aim to more comprehensively evaluate models' understanding of affective states in natural language, but the tasks we consider are not exhaustive. We select three diverse tasks to capture different aspects of model understanding, although additional tasks and evaluation approaches may yield further insights into models' affective understanding. Additionally, we find few significant relationships between psycholinguistic characteristics and model performance across tasks. The lack of significant coefficients does not indicate that no relationship exists, but our findings on non-significant relationships are inconclusive. Our experiments are limited by the availability of annotated texts from the MASIVE dataset, and we hope future work will pursue larger scale studies of these potential phenomena.

[6] AoA and term frequency are moderately correlated ($r = -.63$ in English and $r = -.42$ in Spanish). Detailed results regression models excluding term frequency or AoA are included in Appendix J.

Furthermore, because it is publicly accessible, it is possible that the Reddit texts included in the dataset may also be included in pretraining datasets, but this appears to have limited impact on our results. Given many pretraining datasets are filtered for overall quality or similarity to sources deemed high-quality like Wikipedia (e.g., Li et al. 2024), the informal and social media-specific features included in many of the texts would likely lead to many of such texts being filtered out of pretraining corpora (see discussion in Deas et al. 2025). Furthermore, the low scores across models in the identification task (Table 11) suggest it is unlikely that models have memorized the texts included in the evaluation. Data from Reddit, however, may also not fully represent the linguistic patterns of naturalistic speech outside of social media. Emotion analysis work, however, has often focused on data drawn from social media (e.g., Reddit (Demszky et al., 2020); Twitter (Saravia et al., 2018)), and social media text enables collecting personal expressions of emotions representing a broad variety of backgrounds, making it well-suited to this work.

Finally, the expression task evaluation relies on affective states predicted by a finetuned mT5 model. Despite the finetuned mT5 model outperforming LLMs in this setting, this remains an imperfect measure, meaning that the results of the expression task may be influenced by the limitations of the mT5 model. To account for potential error propagating from mT5 predictions, we primarily focus on the top-5 similarity, as the top-5 predictions are more likely to capture the emotion expressed by a given text as opposed to relying on a single prediction alone (discussed in Appendix C). Additionally, we consider performance across tasks in our regression analyses, while mT5 only impacts expression task performance. Further investigation and improved evaluation of expression task performance is a promising area for future work.

Ethics Statement

Our results reveal difficulties in models' understanding of affective states in natural language. Given the importance of understanding affect and emotions in critical settings such as mental healthcare, models' misunderstandings may lead to negative impacts on individuals emotional health when deployed. We focus on evaluating and characterizing models understanding of nuanced affect in language in order to anticipate potential harms resulting from misunderstanding. Given that working emotion vocabularies are specific to individual speakers, however, increased understanding of nuanced affective states may also further enable mass surveillance of specific communities online if misused.

The anonymized data drawn from (Deas et al., 2024) is used for research purposes only. To maintain consistency with the original study, we do not conduct additional filtering of the data. Therefore, included texts may discuss the kinds of sensitive topics and offensive content that are typically be found online.

Acknowledgments

The first author is supported by National Science Foundation Graduate Research Fellowship DGE-2036197, the Columbia Provost Diversity Fellowship, and the Columbia School of Engineering and Applied Sciences Presidential Fellowship. Any opinion, findings, and conclusions or reccomendations expressed in this material are those of the authors and should not be interpreted as representing the official views or policies of the National Science Foundation. We thank the anonymous reviewers, John Hewitt, and Melody Ma for their helpful feedback on earlier drafts of this work.

References

María Angeles Alonso, Angel Fernandez, and Emiliano Díez. 2014. Subjective age-of-acquisition norms for 7, 039 spanish words. *Behavior Research Methods*, 47(1):268–274.

Ángel-Armando Betancourt, Marc Guasch, and Pilar Ferré. 2024. What distinguishes emotion-label words from emotion-laden words? the characterization of affective meaning from a multi-componential conception of emotions. *Frontiers in Psychology*, 15.

Federico Bianchi, Debora Nozza, and Dirk Hovy. 2022. XLM-EMO: Multilingual emotion prediction in social media text. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 195–203, Dublin, Ireland. Association for Computational Linguistics.

Hanane Boutouta, Abdelaziz Lakhfif, Ferial Senator, and Chahrazed Mediani. 2025. From context to emotion: Leveraging LLMs for recognizing implicit emotions. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 399–409, Southern Denmark

University, Odense, Denmark. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Halszka Bąk and Jeanette Altarriba. 2024. Similar, not universal: the cognitive dimensions of conceptual prototypes of basic emotions in english and in polish. *Cognition and Emotion*, 39(2):261–281.

Nicholas Deas, Elsbeth Turcan, Ivan Ernesto Perez Mejia, and Kathleen McKeown. 2024. MASIVE: Open-ended affective state identification in English and Spanish. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20467–20485, Miami, Florida, USA. Association for Computational Linguistics.

Nicholas Deas, Blake Vente, Amith Ananthram, Jessica A Grieser, Desmond U. Patton, Shana Kleiner, James R. Shepard Iii, and Kathleen McKeown. 2025. Data caricatures: On the representation of African American language in pretraining corpora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29192–29217, Vienna, Austria. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. 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. 1984. Expression and the nature of emotion. *Approaches to emotion*, 3(19):344.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2026. Olmo 3. *Preprint*, arXiv:2512.13961.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sa-

hana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew

Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Daya Guo et al. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Akari Haga, Akiyo Fukatsu, Miyu Oba, Arianna Bisazza, and Yohei Oseki. 2024. BabyLM challenge: Exploring the effect of variation sets on language model training efficiency. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 252–261, Miami, FL, USA. Association for Computational Linguistics.

Shin-nosuke Ishikawa and Atsushi Yoshino. 2025. AI with emotions: Exploring emotional expressions in large language models. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 614–627, Albuquerque, USA. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30, 000 english words. *Behavior Research Methods*, 44(4):978–990.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. 2024. Datacomp-lm: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 14200–14282. Curran Associates, Inc.

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

Chenxiao Liu, Zheyong Xie, Sirui Zhao, Jin Zhou, Tong Xu, Minglei Li, and Enhong Chen. 2024a. Speak from heart: An emotion-guided llm-based multimodal method for emotional dialogue generation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, ICMR '24, page 533–542, New York, NY, USA. Association for Computing Machinery.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Tianlin Zhang, and Sophia Ananiadou. 2024b. Emollms: A series of emotional large language models and annotation tools for comprehensive affective analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5487–5496, New York, NY, USA. Association for Computing Machinery.

Xiaoyang Ma and Weiqi Chang. 2024. A two-stage emotional dialogue generation model based on dialogpt. In *2024 5th International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, pages 703–706.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Rezvan MohammadiBaghmolaei and Ali Ahmadi. 2023. Tet: Text emotion transfer. *Knowledge-Based Systems*, 262:110236.

Aneta Pavlenko. 2008. Emotion and emotion-laden words in the bilingual lexicon. *Bilingualism: Language and Cognition*, 11(2):147–164.

Miguel Ángel Pérez-Sánchez, Hans Stadthagen-Gonzalez, Marc Guasch, José Antonio Hinojosa, Isabel Fraga, Javier Marín, and Pilar Ferré. 2021. Emopro – emotional prototypicality for 1286 spanish words: Relationships with affective and psycholinguistic variables. *Behavior Research Methods*, 53(5):1857–1875.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

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.

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

Robert W. Schrauf and Julia Sanchez. 2004. The preponderance of negative emotion words in the emotion lexicon: A cross-generational and cross-linguistic study. *Journal of Multilingual and Multicultural Development*, 25(2–3):266–284.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695, Florence, Italy. Association for Computational Linguistics.

Cristina Soriano and Anna Ogarkova. 2025. The meaning of 'frustration' across languages. *Language and Cognition*, 17:e16.

Hans Stadthagen-Gonzalez, Constance Imbault, Miguel A. Pérez Sánchez, and Marc Brysbaert. 2016. Norms of valence and arousal for 14, 031 spanish words. *Behavior Research Methods*, 49(1):111–123.

Irwan Suswandi and Afdol Tharik WS. 2017. Sad emotion in javanese language: An analysis of meaning component and relation. *People: International Journal of Social Sciences*, 3(2):2318–2336.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*.

Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. 2020. A review of emotion sensing: categorization models and algorithms. *Multimedia Tools and Applications*, 79(47–48):35553–35582.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

BigScience Workshop et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Chenggang Wu, Yiwen Shi, and Juan Zhang. 2023. Beyond valence and arousal: The role of age of acquisition in emotion word recognition. *Behavioral Sciences*, 13(7).

Chenggang Wu, Juan Zhang, and Yaxuan Meng. 2025. Determine emotion-label words: Quantifying emotional prototypicality of 1, 122 second-language english words. *Bilingualism: Language and Cognition*, page 1–8.

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

Hao Xu, Jiacheng Liu, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. Infini-gram mini: Exact n-gram search at the Internet scale with FM-index. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24944–24969, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,

Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025b. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rui Zhang, Zhenyu Wang, Kai Yin, and Zhenhua Huang. 2019. Emotional text generation based on cross-domain sentiment transfer. *IEEE Access*, 7:100081–100089.

A Working Emotion Vocabulary Measures

We consider 5 measures in total: term frequency, valence and arousal (Mohammad, 2018; Stadthagen-Gonzalez et al., 2016), concreteness (Brysbaert et al., 2013), and age of acquisition (Kuperman et al., 2012; Alonso et al., 2014). [7] Term frequency is measured through the count of each term in the July, 2025 CommonCrawl snapshot[8] training dataset, and is collected using infinigram-mini (Xu et al., 2025) through the API [9]. All other terms are drawn from lexicons collected in prior literature. In cases where exact matches do not exist, we search for all variants of each affective state term, including verb conjugations for past participles and noun forms (e.g., we search for *abandon* and *abandonment* to represent the affective state *abandoned*). In Spanish, we additionally consider both masculine and feminine forms of each term.

[7] Citations refer to sources for each set of measures in English and Spanish respectively.

[8] https://data.commoncrawl.org/crawl-data/CC-MAIN-2025-30/index.html

[9] https://infini-gram-mini.readthedocs.io/en/latest/api.html

Measure	Min	Max	μ	σ
Freq	4.51	9.02	6.34	1.13
Valence	-1.00	0.84	-0.43	0.47
Arousal	-0.96	0.90	0.11	0.47
Conc	1.46	4.07	2.52	0.57
AoA	3.60	13.41	8.24	2.76

Table 7: Summary statistics of psycholinguistic measures for the English affective states evaluated.

Measure	Min	Max	μ	σ
Freq	4.21	8.73	6.19	1.24
Valence	-0.87	0.79	-0.22	0.52
Arousal	-0.85	0.70	0.19	0.33
AoA	3.22	9.98	6.60	1.65

Table 8: Summary statistics of psycholinguistic measures for the Spanish affective states evaluated.

Summary statistics of the psycholinguistic measures associated with each affective state in our working emotion vocabulary experiments are shown in Table 7 and Table 8 for English and Spanish respectively. For Spanish data, valence and arousal are originally measured on a scale of 1 to 9. In these tables, we rescale the values to range from -1 to 1 for clarity.

B Prompt Details

Full Prompt templates for the binary and rating-based verification task formats are included in Figure 5 and Figure 6 respectively. For non-instruction-tuned models, the system prompt is not provided.

C Expression Task Metric Validation

To support the validity of the mT5-based metric used in the expression task evaluation, 2 native English-speaking graduate researchers in NLP and collaborators of the authors judge whether given affective state terms appropriately describe that which is expressed in a given model-generated text in English. Specifically, given a message and candidate affective state, the collaborating annotators were instructed: "*You will be provided with a short text intended and a candidate emotion that may be expressed by the text. You will then be asked to rate whether the candidate emotion appropriately captures the emotion expressed by the text from 1-4 (1 being not at all appropriate, 4 being completely*

Affective State Verification Prompt:

System: You are an expert in human emotions and feelings.

User: Affective state refers to any terms that humans use to describe their experiences of feeling, including emotions, moods, and figurative expressions of feelings (e.g. 'blue' as an expression of sadness instead of the color). Tell me if the word between and is an affective state or not. Output only Y if it is an affective state or N if it is not.
Do not explain or preface your answer.

Input text: Day 7 with Covid, I am feeling much better
Response: 0

Input text: I feel miserable and overwhelmed but I know that everyone will look down on me if I don't get a degree or if I drop out of school.
Response:

Figure 5: Full binary prompt template provided to models for affective state verification experiments.

Affective State Verification Prompt:

System: You are an expert in human emotions and feelings.

User: Affective state refers to any terms that humans use to describe their experiences of feeling, including emotions, moods, and figurative expressions of feelings (e.g. 'blue' as an expression of sadness instead of the color). Does the term between and reflect an affective state? Answer only with one character from the following: 0, 1, 2, or 3.
0 means Not an affective state: the term does not refer to an emotion, feeling, or internal state.
1 means Unlike an affective state: the term is referencing something that is not an emotion.
2 means Like an affective state: the term is likely referencing an emotion, feeling, or internal state.
3 means Completely an affective state: the term is definitely an emotion, feeling, or internal state.
Do not explain or preface your answer.

Input text: Day 7 with Covid, I am feeling much better
Response: 0

Input text: I feel miserable and overwhelmed but I know that everyone will look down on me if I don't get a degree or if I drop out of school.
Response:

Figure 6: Full rating prompt template provided to models for affective state verification experiments.

appropriate)." and provided a 4-point Likert scale. Annotators were also informed that the ratings may be shared for reproducibility purposes. The collaborating annotators digitally acknowledged that they understood the instructions and agreed to complete the task.

We randomly sample 150 texts generated in the expression task in total across the 8 models evaluated in the primary results as well. We consider both the top prediction output by mT5 (*mT5* in Table 9) and a baseline using randomly selected affective states (*Random*). We additionally collect ratings for the affective states input to the model being evaluated in the expression task to summarize model performance (*Input*). These yield 450 text and affective state pairs in total across the three evaluated label sources. Annotators shared 25 model-generated texts (75 text-label pairs). Given the subjective nature of what is considered an appropriate affective state label, we see slight positive correlations between annotators ($\rho = .14$). After filtering skipped samples, the below analysis is based on the 81 and 77 samples (243 and 231 text-label pairs respectively) completed by the two annotators.

Results of these judgments are shown in Table 9. We see that slightly above half of the randomly sample model predictions are considered appropriate, supporting that on average, models struggle on this task. The top mT5 prediction is judged as appropriate for nearly half of model-generated texts, substantially higher than the rate for randomly selected affective states. Regarding mT5, we evaluate only the top prediction to avoid large annotator burden, but notably, this percentage is strictly a lower bound on the appropriateness of the top-3 and top-5 mT5 predictions. In Deas et al. (2024), the performance of mT5 when considering the top-5 predictions substantially increases over the top prediction alone (20% to 35% for English, 25% to 41% for Spanish). Therefore, to account for noise in relying only on the top prediction from mT5, we focus primarily on the top-5 similarity results in the inter-task correlation and regression analyses. We note, however, that there is significant room for improvement in evaluating expressed affective states in this setting with a large set of possible affective state labels.

Label Source	Mean Rating	% ≥ 3
Input	2.50	56.8%
mT5	2.24	47.5%
Random	1.83	21.6%

Table 9: Appropriateness judgments of mt5-predicted and random affective states on model-generated texts. %≥3 indicates the percentage of predictions that are considered "somewhat" or "completely" appropriate.

D Working Emotion Vocabulary Regression Details

We fit Ordinal Logistic Regression models using the `statsmodels` package and the `bfgs` solver. All input measures except term frequency are normalized before fitting each model. For term frequency, we use the $\log_{10}$(frequency). As response variables, we first rescale model scores (top-5 similarity for identification, top-5 similarity for expression, and probability-based rating for verification) on each task to be between 0 and 1. We then group the average scaled performance across tasks into 4 quantiles. The ordered quantiles represent the response variable used in working emotion vocabulary regression models.

E Model Details

All evaluated models are listed in Table 10 including whether the model is instruction-tuned, and whether it has undergone EmoLLM training (Liu et al., 2024b). In contrast to instruction-tuning, we do not find consistent benefits in EmoLLM training across experiments.

For the identification task we use beam search to generate the top-5 predictions for all models. For the expression task, we generate responses for all models using a `temperature` of 0.7, `top-p` of 0.9, and a `no_repeat_ngram_size` of 4. For verification experiments, we examine the raw output probabilities produced by the model for each possible label (i.e., "Y"/"N" or an integer 1-4).

F Full Identification Results

Results for all 23 models in the identification task are shown in Table 11.

G Full Expression Results

Results for all 23 models in the expression task are shown in Table 12.

H Full Verification Results

Results for all 23 models in the verification task are shown in Figure 7. Cohen's κ metrics for classification formats are also included in Table 13.

I Affective State Rankings

The top and bottom 5 ranked affective states for 4 example models are shown in Table 14. we take the average ranking of each affective state compared to others across the three tasks. [12]

J Detailed Regression Analysis Results

As Age of Acquisition (AoA) and term frequency are moderately correlated ($r = -.63$, $p < .001$ for English and $r = -.42$, $p < .001$ for Spanish), we conduct two additional regression analyses based on Subsection 4.3 where one excludes term frequency (Table 15) and one excludes (Table 16). In either case, we see that coefficients for each variable generally trend in the same direction including term frequency and AoA when included in the model. The independent contribution of term frequency or age of acquisition is unclear from our results, although in either case, the trends for English contradict expectations; increasing term frequency in CommonCrawl is negatively predictive of model performance and increasing AoA is positively predictive of model performance.

[12] Note, based on the original text, "echo" in Table 14 appears to be a typo for "hecho".

Model	Checkpoint	Size	Training	
			Instruct	Emotion
OLMo (Groeneveld et al., 2024)	`OLMo-7B-Instruct-hf`	7B	✓	
OLMo-2 (Walsh et al., 2025)	`OLMo-2-1124-7B-Instruct`	7B	✓	
OLMo-3 (Ettinger et al., 2026)	`allenai/Olmo-3-1025-7B`	7B		
	`allenai/Olmo-3-7B-Instruct`	7B	✓	
Llama-2 (Touvron et al., 2023)	`Llama-2-7b-chat-hf`	7B	✓	
	`Emollama-chat-7b`	7B	✓	✓
Llama-3 (Grattafiori et al., 2024)	`Meta-Llama-3-8B`	8B		
	`Meta-Llama-3-8B-Instruct`	8B	✓	
Llama-3.1 (Grattafiori et al., 2024)	`Meta-Llama-3.1-8B`	8B		
	`Meta-Llama-3.1-8B-Instruct`	8B	✓	
BLOOM (Workshop et al., 2023)	`bloomz-7b1-mt`	7B	✓	
	`lzw1008/Emobloom-7b`	7B	✓	✓
Mistral (Jiang et al., 2023)	`Mistral-7B-v0.3`	7B		
	`Mistral-7B-Instruct-v0.3`	7B	✓	
Ministral[10]	`Ministral-8B-Instruct-2410`	8B	✓	
Ministral 3[11]	`Ministral-3-8B-Base-2512`	8B		
	`Ministral-3-8B-instruct-2512`	8B	✓	
Qwen-2.5 (Yang et al., 2025b)	`Qwen2.5-7B`	7B		
	`Qwen2.5-7B-Instruct-1M`	7B	✓	
Qwen 3 (Yang et al., 2025a)	`Qwen3-4B`	4B		
	`Qwen3-4B-instruct-2507`	4B	✓	
DeepSeek-R1 (Guo et al., 2025)	`DeepSeek-R1-Distill-Llama-8B`	8B	✓	
	`DeepSeek-R1-Distill-Qwen-7B`	7B	✓	

Table 10: Summary of models evaluated throughout experiments. Emotion finetuned models are based on models released by Liu et al. (2024b).

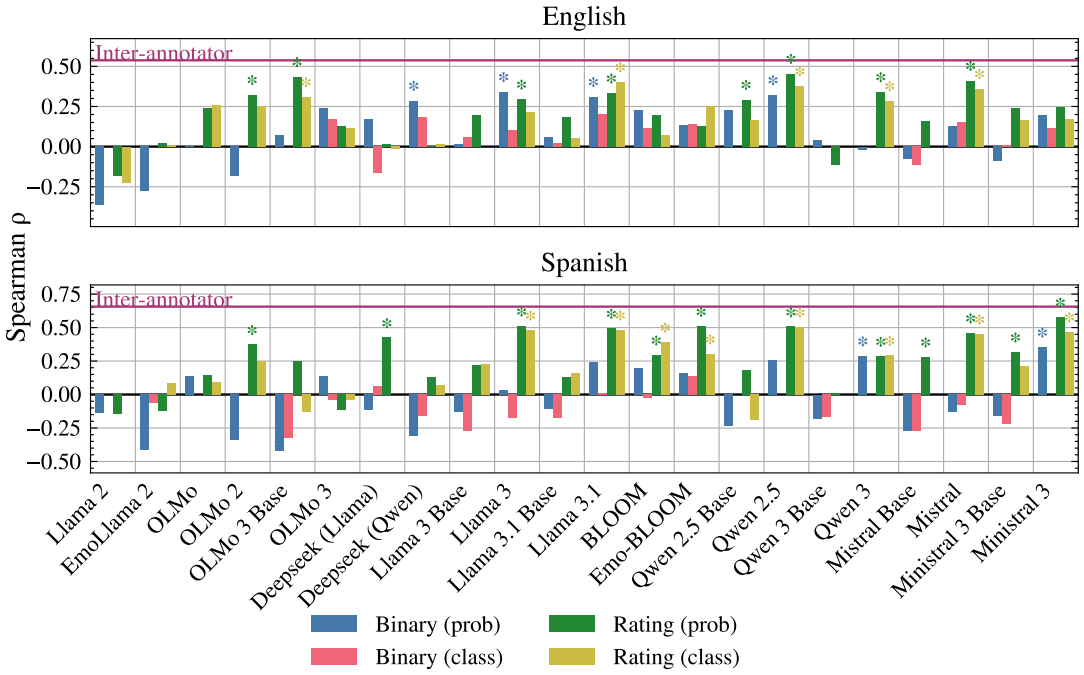


Figure 7: Full Spearman's correlations between model outputs and human annotations across verification task formats. $*$ indicates significant correlation ($p \leq .05$)

	Model	Top-k Similarity↑		
		1	3	5
English	Llama 2	37.0% ± 13.8	40.4% ± 14.5	40.8% ± 14.6
	EmoLlama 2	39.2% ± 12.8	45.1% ± 16.5	45.4% ± 16.4
	OLMo	31.1% ± 10.7	33.9% ± 10.5	35.9% ± 11.8
	OLMo 2	44.2% ± 14.4	52.0% ± 14.7	56.4% ± 13.7
	OLMo 3 Base	37.1% ± 11.3	37.3% ± 11.4	37.3% ± 11.4
	OLMo 3	46.2% ± 19.4	58.6% ± 20.2	63.0% ± 19.9
	Deepseek (Llama)	25.2% ± 12.1	29.2% ± 11.2	30.2% ± 11.0
	Deepseek (Qwen)	24.5% ± 6.3	26.8% ± 6.9	28.7% ± 7.6
	Llama 3 Base	31.9% ± 9.2	32.1% ± 9.4	32.1% ± 9.4
	Llama 3	49.1% ± 17.5	56.7% ± 18.0	60.2% ± 18.1
	Llama 3.1 Base	30.3% ± 9.7	30.4% ± 9.6	30.4% ± 9.6
	Llama 3.1	44.8% ± 13.6	51.5% ± 15.2	56.0% ± 16.9
	BLOOM	36.9% ± 12.4	47.1% ± 13.7	51.5% ± 15.2
	Emo-BLOOM	40.8% ± 12.2	47.0% ± 12.6	50.2% ± 12.5
	Qwen 2.5 Base	27.0% ± 9.2	43.1% ± 14.1	46.0% ± 13.7
	Qwen 2.5	45.9% ± 15.7	57.8% ± 15.3	61.8% ± 16.2
	Qwen 3 Base	34.2% ± 14.0	38.1% ± 17.2	38.3% ± 17.1
	Qwen 3	45.8% ± 13.9	55.8% ± 16.0	57.7% ± 16.4
	Mistral Base	31.5% ± 11.6	32.5% ± 11.6	32.9% ± 11.9
	Mistral	42.0% ± 13.3	48.3% ± 13.7	52.0% ± 14.8
	Ministral	45.7% ± 15.9	53.1% ± 15.0	57.2% ± 14.4
	Ministral 3 Base	30.3% ± 8.4	30.8% ± 8.5	30.9% ± 8.5
	Ministral 3	43.3% ± 15.2	53.7% ± 16.9	58.4% ± 16.6
Spanish	Llama 2	39.6% ± 11.0	43.0% ± 11.2	43.7% ± 11.1
	EmoLlama 2	38.6% ± 10.2	47.1% ± 12.3	48.9% ± 13.4
	OLMo	36.9% ± 11.0	38.3% ± 11.0	38.9% ± 11.0
	OLMo 2	41.4% ± 12.9	52.5% ± 14.0	54.4% ± 14.7
	OLMo 3 Base	30.7% ± 8.0	30.9% ± 8.1	30.9% ± 8.1
	OLMo 3	42.9% ± 14.2	49.4% ± 13.4	53.3% ± 14.4
	Deepseek (Llama)	27.2% ± 8.8	30.6% ± 8.1	31.5% ± 7.7
	Deepseek (Qwen)	28.7% ± 8.2	31.6% ± 9.4	33.1% ± 9.4
	Llama 3 Base	38.2% ± 9.8	38.5% ± 10.0	38.7% ± 10.2
	Llama 3	46.5% ± 18.1	52.2% ± 17.1	56.7% ± 18.4
	Llama 3.1 Base	39.8% ± 10.6	40.0% ± 10.7	40.1% ± 10.6
	Llama 3.1	44.3% ± 16.8	52.5% ± 17.2	56.8% ± 18.3
	BLOOM	34.3% ± 12.1	40.1% ± 13.2	43.7% ± 16.0
	Emo-BLOOM	34.8% ± 12.1	41.4% ± 11.2	43.3% ± 11.2
	Qwen 2.5 Base	37.5% ± 13.3	44.2% ± 12.4	47.7% ± 14.8
	Qwen 2.5	40.8% ± 14.0	52.0% ± 16.8	61.1% ± 18.7
	Qwen 3 Base	37.7% ± 10.9	38.9% ± 11.0	40.2% ± 11.3
	Qwen 3	40.6% ± 15.5	49.5% ± 15.1	56.2% ± 17.4
	Mistral Base	31.0% ± 7.0	31.5% ± 7.0	31.5% ± 7.0
	Mistral	43.2% ± 12.9	49.0% ± 14.6	50.6% ± 14.5
	Ministral	42.2% ± 13.3	50.0% ± 14.6	52.3% ± 13.8
	Ministral 3 Base	38.3% ± 10.8	38.4% ± 10.8	38.4% ± 10.8
	Ministral 3	45.7% ± 19.8	58.4% ± 20.3	60.9% ± 21.1

Table 11: Full top-k similarity scores with ± 2SE's for models in the identification task.

	Model	Top-k Similarity↑		
		1	3	5
English	Llama 2	42.0% ± 0.8	52.6% ± 0.8	57.3% ± 0.8
	EmoLlama 2	40.4% ± 0.7	51.0% ± 0.8	55.5% ± 0.8
	OLMo	41.9% ± 0.7	51.0% ± 0.7	55.1% ± 0.7
	OLMo 2	41.8% ± 0.7	52.3% ± 0.7	56.7% ± 0.7
	OLMo 3 Base	42.5% ± 0.7	51.4% ± 0.7	55.6% ± 0.7
	OLMo 3	45.4% ± 0.6	55.3% ± 0.6	60.3% ± 0.6
	Deepseek (Llama)	38.2% ± 0.8	48.7% ± 0.8	52.7% ± 0.8
	Deepseek (Qwen)	34.6% ± 0.7	45.3% ± 0.8	49.2% ± 0.8
	Llama 3 Base	42.7% ± 0.6	51.7% ± 0.6	55.5% ± 0.6
	Llama 3	39.7% ± 0.7	49.7% ± 0.7	54.5% ± 0.7
	Llama 3.1 Base	42.7% ± 0.6	51.5% ± 0.6	55.5% ± 0.6
	Llama 3.1	38.4% ± 0.7	49.1% ± 0.7	54.1% ± 0.7
	BLOOM	41.9% ± 0.5	49.9% ± 0.5	53.4% ± 0.5
	Emo-BLOOM	41.4% ± 0.4	49.1% ± 0.5	52.6% ± 0.5
	Qwen 2.5 Base	44.0% ± 0.6	52.6% ± 0.6	56.3% ± 0.6
	Qwen 2.5	42.3% ± 0.6	51.4% ± 0.6	55.4% ± 0.6
	Qwen 3 Base	40.1% ± 0.9	50.4% ± 0.9	54.3% ± 0.9
	Qwen 3	43.3% ± 0.6	53.1% ± 0.7	57.0% ± 0.7
	Mistral Base	40.3% ± 0.8	49.9% ± 0.9	54.4% ± 0.9
	Mistral	40.6% ± 0.7	49.5% ± 0.6	52.9% ± 0.6
	Ministral	42.0% ± 0.7	52.1% ± 0.7	56.6% ± 0.7
	Ministral 3 Base	39.4% ± 0.9	50.3% ± 0.9	54.9% ± 0.9
	Ministral 3	42.1% ± 0.6	52.2% ± 0.7	56.1% ± 0.7
Spanish	Llama 2	36.0% ± 0.6	44.7% ± 0.7	49.6% ± 0.7
	EmoLlama 2	36.9% ± 0.6	45.8% ± 0.7	50.6% ± 0.8
	OLMo	38.1% ± 0.8	48.4% ± 0.9	53.8% ± 0.9
	OLMo 2	38.2% ± 0.7	47.5% ± 0.7	52.5% ± 0.8
	OLMo 3 Base	38.2% ± 0.9	48.5% ± 0.9	53.5% ± 0.9
	OLMo 3	40.2% ± 0.8	50.8% ± 0.8	56.7% ± 0.9
	Deepseek (Llama)	43.9% ± 1.1	54.9% ± 1.1	59.9% ± 1.1
	Deepseek (Qwen)	35.7% ± 0.8	45.3% ± 0.9	49.6% ± 0.9
	Llama 3 Base	39.6% ± 0.7	48.0% ± 0.7	52.4% ± 0.8
	Llama 3	35.8% ± 0.7	46.2% ± 0.8	51.3% ± 0.8
	Llama 3.1 Base	38.9% ± 0.7	47.9% ± 0.7	52.8% ± 0.8
	Llama 3.1	36.3% ± 0.7	46.3% ± 0.8	51.4% ± 0.8
	BLOOM	40.1% ± 0.6	47.0% ± 0.6	51.0% ± 0.6
	Emo-BLOOM	39.7% ± 0.5	46.8% ± 0.6	50.5% ± 0.6
	Qwen 2.5 Base	38.3% ± 0.7	47.2% ± 0.7	52.2% ± 0.8
	Qwen 2.5	38.2% ± 0.7	46.8% ± 0.7	52.1% ± 0.8
	Qwen 3 Base	45.3% ± 1.2	56.0% ± 1.1	61.2% ± 1.1
	Qwen 3	36.7% ± 0.8	48.7% ± 0.9	55.2% ± 0.9
	Mistral Base	41.6% ± 0.9	50.8% ± 1.0	55.3% ± 1.0
	Mistral	38.7% ± 0.7	48.9% ± 0.8	54.0% ± 0.8
	Ministral	40.1% ± 0.7	48.5% ± 0.7	53.5% ± 0.8
	Ministral 3 Base	46.8% ± 1.2	57.6% ± 1.2	62.5% ± 1.1
	Ministral 3	34.2% ± 0.7	44.8% ± 0.8	50.4% ± 0.8

Table 12: Full top-k similarity scores with ± 2SE's for all models in generating responses expressing a given affective state.

Model	Format	English		Spanish	
		κ	κ_{bin}	κ	κ_{bin}
Llama 2	Binary	0.00		0.00	
	Rating	-0.05	-0.01	0.00	0.00
EmoLlama 2	Binary	0.00		0.01	
	Rating	-0.06	0.04	0.10	0.10
OLMo	Binary	0.00		0.00	
	Rating	0.07	0.00	-0.02	-0.07
OLMo 2	Binary	0.00		0.00	
	Rating	0.13	0.17	0.08	0.13
OLMo 3 Base	Binary	0.00		-0.13	
	Rating	0.22	-0.04	-0.07	0.00
OLMo 3	Binary	-0.12		-0.07	
	Rating	0.02	0.00	-0.03	0.00
Deepseek (Llama)	Binary	-0.07		0.01	
	Rating	-0.02	-0.08	0.00	0.00
Deepseek (Qwen)	Binary	-0.04		-0.16	
	Rating	0.05	0.06	0.03	0.10
Llama 3 Base	Binary	0.03		-0.08	
	Rating	0.00	0.00	0.03	0.13
Llama 3	Binary	-0.09		-0.07	
	Rating	0.25	0.00	0.22	0.00
Llama 3.1 Base	Binary	-0.06		-0.06	
	Rating	-0.04	-0.09	-0.02	0.06
Llama 3.1	Binary	0.04		-0.04	
	Rating	0.26	-0.09	0.22	0.42
BLOOM	Binary	0.01		-0.03	
	Rating	0.02	0.09	-0.03	0.38
Emo-BLOOM	Binary	0.01		0.00	
	Rating	-0.02	0.17	0.01	0.21
Qwen 2.5 Base	Binary	0.00		0.00	
	Rating	0.04	0.00	-0.06	0.00
Qwen 2.5	Binary	0.00		0.00	
	Rating	0.38	0.17	0.28	0.46
Qwen 3 Base	Binary	0.00		-0.04	
	Rating	0.00	0.00	0.00	0.00
Qwen 3	Binary	0.00		0.00	
	Rating	0.13	0.00	0.06	0.13
Mistral Base	Binary	-0.04		-0.12	
	Rating	0.00	0.00	0.00	0.00
Mistral	Binary	0.01		-0.10	
	Rating	0.23	0.00	0.18	0.00
Ministral	Binary	0.00		-0.08	
	Rating	0.00	0.00	0.00	0.00
Ministral 3 Base	Binary	0.01		-0.12	
	Rating	0.08	0.19	0.00	0.27
Ministral 3	Binary	0.01		0.00	
	Rating	0.19	0.00	0.22	0.47

Table 13: Cohen's κ scores for classification output formats in the verification task. For κ_{bin}, ratings output by the model are binarized before computing agreement.

OLMo 3			
English		Spanish	
Term	Rank	Term	Rank
abandoned	3.3	**frustrada**	5.7
dizzy	9.0	**afortunada**	8.3
panicked	9.0	curioso	8.3
lonely	9.3	*triste*	10.0
appalled	10.0	*deprimida*	11.0
reluctant	30.7	**rodeado**	34.7
aloof	31.3	contra	37.0
downtrodden	31.3	*convencido*	37.3
wide	33.3	obligado	38.3
jumpy	35.3	**trabado**	39.0

Llama 3.1			
English		Spanish	
Term	Rank	Term	Rank
lonely	5.0	**afortunada**	4.7
upset	6.3	**frustrada**	5.3
abandoned	7.7	nerviosa	9.0
paralyzed	10.0	devastado	9.0
unnoticed	13.7	*deprimida*	9.3
aloof	29.3	dormido	35.7
done	29.7	**rodeado**	35.7
wide	29.7	fea	36.7
jittery	31.0	**trabado**	38.0
unbalanced	36.3	echo	39.3

Qwen 3			
English		Spanish	
Term	Rank	Term	Rank
abandoned	3.0	**afortunada**	6.7
upset	4.7	*deprimida*	9.7
nervous	7.7	*ignorada*	10.3
aggressive	7.7	*triste*	10.3
lonely	7.7	**frustrada**	11.7
inexperienced	31.7	querido	34.3
downtrodden	32.0	contra	34.7
reluctant	32.3	*convencido*	35.7
jittery	33.3	**trabado**	38.0
aloof	35.3	**rodeado**	41.0

Ministral 3			
English		Spanish	
Term	Rank	Term	Rank
abandoned	4.0	**frustrada**	7.7
lonely	5.7	**afortunada**	9.7
supported	9.3	incomoda	11.3
deep	10.3	*ignorada*	12.3
nervous	10.7	mejor	12.3
aloof	30.3	contra	36.0
downtrodden	30.7	mareado	37.7
jittery	31.7	**trabado**	39.7
dizzy	31.7	*convencido*	40.0
rested	34.3	**rodeado**	40.0

Table 14: Top and bottom-5 affective states by average performance rank across the three tasks. **Bolded** terms appear in the top/bottom ranks for all four models, while *italicized* terms appear for two or three.

		OLMo 3 Base	OLMo 3	Llama 3.1 Base	Llama 3.1	Qwen 3 Base	Qwen 3	Ministral 3 Base	Ministral 3
English	Valence	-0.03	0.06	-0.00	0.02	-0.03	0.15	-0.12	0.22
	Arousal	-0.06	-0.12	-0.18	-0.18	0.07	-0.26	-0.17	-0.23
	Conc	0.19	0.21	0.08	0.10	0.16	0.35^{*}	-0.05	0.23
	AoA	-0.64^{**}	-0.23	-0.87^{**}	-0.55^{**}	-0.17	$-0.28^{\dagger}$	0.00	-0.62^{**}
Spanish	Valence	-0.23	0.24	-0.13	0.17	-0.21	$0.37^{\dagger}$	-0.26	0.11
	Arousal	0.27	-0.14	0.11	0.20	0.45^{*}	-0.01	0.45^{*}	-0.04
	AoA	0.42^{*}	0.58^{**}	0.60^{**}	0.50^{**}	$0.36^{\dagger}$	0.37^{*}	0.58^{**}	0.20

Table 15: Coefficients of the Ordinal Regression Model predicting affective state rank in English (top) and Spanish (bottom) excluding term frequency from the exogenous variables. Superscripts indicate significant coefficients, $^{**}p \leq .01$, $^{*}p \leq .05$, $^{\dagger}p \leq .075$.

		OLMo 3 Base	OLMo 3	Llama 3.1 Base	Llama 3.1	Qwen 3 Base	Qwen 3	Ministral 3 Base	Ministral 3
English	Freq	0.83^{**}	0.42^{*}	1.00^{**}	0.67^{**}	0.38^{*}	0.54^{**}	0.06	0.73^{**}
	Valence	-0.23	-0.03	-0.26	-0.13	-0.12	0.00	-0.14	0.06
	Arousal	0.11	-0.03	0.03	-0.04	0.15	-0.14	-0.15	-0.06
	Conc	0.23	0.21	0.13	0.14	0.14	0.35^{*}	-0.07	0.27
Spanish	Freq	-0.58^{**}	$-0.28^{\dagger}$	-0.66^{**}	-0.58^{**}	-0.92^{**}	-0.15	-1.03^{**}	0.09
	Valence	-0.10	0.31	0.03	0.29	0.07	$0.39^{\dagger}$	0.01	0.07
	Arousal	0.29	-0.04	0.13	0.22	$0.37^{\dagger}$	0.07	0.46^{*}	0.05

Table 16: Coefficients of the Ordinal Regression Model predicting affective state rank in English (top) and Spanish (bottom) excluding Age of Acquisition from the exogenous variables. Superscripts indicate significant coefficients, $^{**}p \leq .01$, $^{*}p \leq .05$, $^{\dagger}p \leq .075$.