EmoPair: A New Paradigm for Measuring Emotional Affect

Anonymous ACL submission

Abstract

The continuous Valence-Arousal-Dominance (VAD) framework maps emotions into a high-resolution, three-dimensional space, but expanding these datasets via manual annotation is costly and prone to subjective inconsistencies. To address this, we introduce EmoPair, which updates the EmoBank corpus by replacing direct scalar ratings with a scalable, LLM-driven pairwise comparison approach. This shift substantially improved manual inter-rater reliability, increasing Krippendorff's Alpha from 0.595 to 0.896 for Arousal and 0.570 to 0.865 for Dominance. Using Alternative Annotator Test (AltTest) validation to compare automated LLM annotations to a manual sample as well as Concept-Guided Chain-of-Thought (CGCoT) prompting, we generated reliable automated pairwise judgments, translating them into continuous scalar ratings via the Bradley-Terry model. Benchmarking with fine-tuned RoBERTa-large models (PERT and reward-based) showed EmoPair-trained models significantly outperformed EmoBank baselines on Arousal (84-85% vs. 73-74% accuracy) and Dominance (77% vs. 61-65%). These results demonstrate that pairwise comparisons provide superior, behaviorally aligned supervision data for emotional affect.

1 Introduction

The proliferation of digital text—from social media posts and consumer reviews to vast archives of literature and news—has created an unprecedented opportunity to study human expression at scale. A central challenge in this endeavor is the computational analysis of emotion. For years, the field has been dominated by a classification of text into positive, negative, or neutral polarities, with some additional work expanding categories (Demszky et al., 2020; Mouronte-López et al., 2023; Liu et al., 2019a). While useful for broad-stroke assessments, the categorical approach fails to capture the rich, nuanced, and complex nature of human affective states. For example, a review expressing serene contentment and another conveying frenetic excitement could both be labeled "positive," yet they represent profoundly different affective experiences. Likewise, texts expressing melancholy disappointment versus quiet despair differ substantially in emotional nuance despite sharing the "sad" category.

Prior research has used two broad frameworks for training emotion models: discrete models and continuous models. Discrete models conceptualize emotions as distinct categories and differ in the number of categories they propose. In contrast, continuous models characterize emotions along continuous dimensions. Among these, the Valence-Arousal-Dominance (VAD) framework proposed by Mehrabian and Russell (1974) is a widely adopted continuous model. It characterizes emotions along three independent dimensions: pleasure (Valence), activation (Arousal), and perceived control (Dominance).

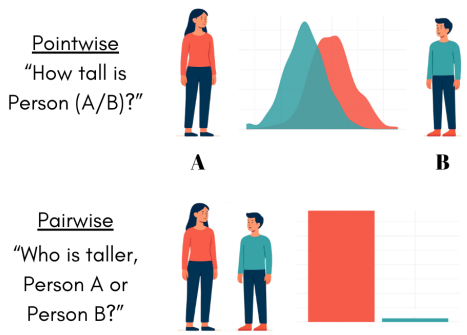

Figure 1: **Example comparison of pointwise and pairwise annotation modes.**

The dimensional approach provides a high-resolution map of the affective landscape, allowing any emotional state to be represented as a unique coordinate in a three-dimensional space. This ca-

pacity for nuance makes the VAD model exceptionally well-suited for computational applications, where some aspects of the complexity of feeling can be translated into analyzable data (Gamage et al., 2024; Cao and Cao, 2025; Hadimlioglu and Linga, 2025).

However, to date, there are very few data sources in this domain that can support model development for the measurement of multidimensional emotions. One exception is the EmoBank corpus, which provides one of the only large-scale dimensional emotion scorings of text annotated across multiple genres in English (Buechel and Hahn, 2017; Bradley and Lang, 1999; Redondo et al., 2007; Võ et al., 2009; Moors et al., 2013; Warriner et al., 2013; Mohammad, 2025). Developing and expanding such data traditionally relies on costly and labor-intensive human annotation, a process hindered by challenges such as resource constraints, human fatigue, and the nuances inherent in subjective judgments (Calderon et al., 2025; Nasution and Onan, 2024; Wu et al., 2024; Gehrmann et al., 2023). To overcome these limitations, the "LLM-as-a-judge" framework offers a potential scalable alternative, utilizing Large Language Models (LLMs) for their speed, efficiency, and advanced pattern recognition capabilities (Gu et al., 2024; Tan et al., 2024; Pan et al., 2024; Licht et al., 2025).

In this paper, we introduce a method for updating the EmoBank corpus using LLM-generated pairwise annotations and release an updated version of the dataset based on this method. Our process involves utilizing LLMs' ability to generate and then compare concept-specific breakdowns of the texts through Concept-Guided Chain of Thought (CGCoT) prompting to transform the judgment task into a tractable pattern recognition task (Wu et al., 2024). The resulting numerous pairwise preferences are subsequently translated into a continuous scale score for each breakdown using a robust probabilistic model—the Bradley-Terry (BT) model—which estimates the relative position of items along any latent dimension (Bradley and Terry, 1952; Licht et al., 2025; Fang et al., 2026).

Further, in order to establish LLMs as viable replacements for manual human labor in annotation tasks, researchers must implement statistically rigorous validation procedures. This is particularly vital as recent studies have questioned the reliability of LLMs, citing concerns over their inherent biases and shifting positionality (Thapa et al., 2023; Pangakis and Wolken, 2024; Nahum et al., 2025). To this end, we employ the Alternative Annotator Test (AltTest), a novel statistical procedure designed to specifically justify replacing human annotators with LLMs (Calderon et al., 2025).

We make four core contributions to the literature on computational modeling of dimensional emotion. First, we demonstrate the utility of pairwise comparisons relative to using discrete ratings for improving manual annotator reliability on this subjective construct, with improved inter-rater reliabilities (IRR) on all three emotional affect dimensions. Second, we release **EmoPair**, an updated version of the EmoBank corpus in which Arousal and Dominance scores are derived from LLM-generated pairwise comparisons using CG-CoT prompting. Third, we demonstrate a rigorous validation framework for LLM-as-annotator deployment: using the AltTest, we show that select LLMs meet a statistical bar for replacing manual annotators on Arousal and Dominance, but not Valence—a construct-specific failure that underscores the necessity of per-dimension validation. And finally, we benchmark four model–dataset combinations using fine-tuned RoBERTa-large architectures trained on both EmoBank and EmoPair, showing that pairwise-derived supervision yields consistent performance gains across Arousal and Dominance. Together, these findings establish pairwise comparison as a scalable, behaviorally grounded paradigm for dimensional textual emotion analysis and yield an open dataset for future work in affective computing.

2 Prior Work

2.1 Measuring Emotional Affect

The categorical approach to measuring emotional affect discussed in the Introduction has been heavily critiqued for its limitations (Deng and Ren, 2023; Nagase et al., 2026). Specifically, discrete frameworks struggle with "fuzzy emotional boundaries," failing to capture the nuance between closely related states or varying intensities of the same emotion (Acheampong et al., 2020). Furthermore, the reliance on incompatible, localized labeling schemes across different studies has led to fragmented datasets that are difficult to synthesize, making it challenging to capture the rich, multifaceted nature of human affective states (Dileep Kumar et al., 2025). The VAD framework

proposed by Russell and Mehrabian (1977) is notably less explored. *Valence* is defined as "the pleasantness of a stimulus", *Arousal* as "the intensity of emotion provoked by a stimulus", and *Dominance* as "the degree of control exerted by a stimulus" (Warriner et al., 2013). Across the studies that have explored this framework, Arousal and Dominance are known to be more difficult to measure (Mehrabian and Russell, 1974; Yani-de Soriano and Foxall, 2006; Foxall et al., 2006; Fontaine et al., 2007). Multiple lexicons have been created to measure the VAD placement of individual words and phrases across multiple languages (Bradley and Lang, 1999; Redondo et al., 2007; Võ et al., 2009; Moors et al., 2013; Warriner et al., 2013; Mohammad, 2025). However, only one such effort exists for larger units of text, EmoBank by Buechel and Hahn (2017).

2.2 Data Annotation and LLMs as Measurement Tools

The previous efforts to measure affect have all used similar practices for the data annotation process. Typically, researchers rely on pointwise, absolute rating systems such as Likert scales or continuous sliders, where multiple manual annotators read isolated texts and assign a numerical score to represent the text's emotional magnitude.

Many of these studies discuss the challenges entailed with this process (Metallinou and Narayanan, 2013; Niu et al., 2025). First, manual annotation is costly and labor-intensive, creating a bottleneck that limits the scale of emotional datasets. Second, assessing subjective constructs on absolute scales frequently results in low IRR. Annotators bring varying interpretations of the scale anchors, different cultural backgrounds, and human fatigue to the task, leading to inconsistencies in their ratings. These challenges are well-documented in the computational science literature, where high variance among raters often obscures the latent ground truth (Aroyo and Welty, 2015).

Pairwise annotation processes offer distinct advantages that directly address these limitations. Presenting annotators with two items and asking them to choose which exhibits a higher degree of a specific trait reduces cognitive load and removes the uncertainty of absolute scale calibration (see Figure 1 for an example). This approach can yield significant improvements in IRR (Kiritchenko and Mohammad, 2017; Burton et al., 2021). Consequently, this improvement in data quality provides a superior supervision signal, which is known to result in better performance for downstream modeling tasks (Ouyang et al., 2022).

Simultaneously, the advent of LLMs has opened potential new avenues for scalable data annotation. LLMs are increasingly deployed as zero-shot reasoning evaluators across various NLP tasks (Zheng et al., 2023). However, the impact of poor IRR on absolute scales extends to automated annotators, such as the "heaping" discussed previously (Licht et al., 2025). The choice of task framing is therefore critical when considering LLMs.

By utilizing techniques like CGCoT prompting, we can structure an LLM's reasoning process to extract relevant linguistic features before making a pairwise judgment. These relative preferences can then be translated into continuous ratings via probabilistic frameworks like the Bradley-Terry model. This pipeline generates highly reliable targets that are exceptionally well-suited for fine-tuning robust, reward-based emotion models. We demonstrate the utility of this approach in the present study.

3 Data

The EmoBank dataset published by Buechel and Hahn (2017) available on their GitHub repository contains 10,006 sentences that were human-annotated across the three dimensions of the VAD framework. EmoBank is comprised of sentences drawn from two sources: the American National Corpus and the SemEval-2007 Task 14. These sentences come from diverse contexts such as news headlines drawn from major newspapers, blogs, letters, travel guides, fiction, and essays from multiple genres. Each sentence was evaluated with five independent annotators who distinguished between writer perspective (emotion experienced during production, aided by linguistic cues) and reader perspective (emotion evoked in an average audience upon reception).

Since its introduction, the EmoBank corpus has been used to explore the relationship between emotional valence and phonemic content (Slavova and Andonov, 2021) and the use of LLMs for emotion annotation tasks (Niu et al., 2024). Here, we use the text that was annotated on the three emotional dimensions as well as their corresponding mean reader annotations as our ground truth given the reported high annotator agreement in the original data and their use in prior work (Bulla and Mongiovì, 2024). EmoBank is distributed under a CC

BY-SA 4.0 license and thus EmoPair is released under the same CC BY-SA 4.0 license in accordance with the share-alike terms of the original.

4 Methods

4.1 Pairwise Comparisons

4.1.1 Manual Annotations

To establish a high-quality ground truth for examining the potential of LLMs to match manual annotation performance, we first curated a subset of texts from the EmoBank corpus. A team of annotators consisting of university staff and students provided pairwise labels for the three dimensions. All annotators underwent a one-hour training session on the codebook to ensure a unified understanding of the VAD constructs before completing the task. Unlike traditional scalar rating tasks, which suffer from high subjectivity and cognitive load, annotators were presented with pairs of texts and asked to select which text exhibited a greater degree of the specific emotional dimension. Appendix B contains details on how we curated this subset and prompted manual annotators.

4.1.2 LLM Annotations

To avoid the poor calibration and "heaping" commonly observed when LLMs are asked for direct scalar ratings discussed above, we adopted CGCoT prompting (Wu et al., 2024). Specifically, for each text and each dimension, the LLM was first prompted to generate a concept-specific breakdown. This breakdown acts as a codebook for manual annotators, identifying and describing specific linguistic features relevant to the target dimension (e.g., "Indicate whether the speaker claims authority, issues a command, asks for permission, or expresses uncertainty in this text" for Dominance). Appendix B.2 contains the CGCoT prompts we designed for the task. Following the generation of breakdowns for two texts, the LLM was prompted to compare the breakdowns to decide which text exhibits more of the construct and to give a justification for its choice (Kojima et al., 2023; Wang et al., 2023; Wei et al., 2023; Chrzan, 2025).

4.2 AltTest

A critical component of our methodology is statistically justifying the use of LLMs as a replacement for human annotators. To this end, we employed the AltTest proposed by Calderon et al. (2025) and implemented in `pairadigm` (Chrzan, 2025). This procedure uses a leave-one-out approach to verify if an LLM's annotations align with the "ground truth" distribution (approximated by the collective human consensus) as well as, or better than, a single human annotator does.

The procedure results in two scores for each LLM tested, the Win Rate, (ω), and the Average Advantage Probability, (ρ). The Win Rate (ω) represents the proportion of human annotators the LLM statistically "beats" and is determined at a specific ϵ value, which is the difference in the probability the LLM needs to overtake the alignment the left-out human has with the other annotators. The Average Advantage Probability (ρ), however, represents the probability that the LLM is as good as or better than a randomly chosen human annotator. Following Calderon et al. (2025), we considered an LLM a valid alternative to our trained annotators only if $\omega \geq 0.5$ when $\epsilon = 0.10$. Then, among LLMs with $\omega \geq 0.5$, we chose the model with the highest ρ as our replacement annotator, if any exists.

4.3 Bradley-Terry Scoring

To convert the large-scale pairwise preferences into continuous ratings, we utilized the BT model. The BT model posits that the probability of text i being ranked higher than text j depends on their latent traits (objectivity, clarity, etc.), λ_i and λ_j, as shown in Equation 1.

$$P(i > j) = \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}} \quad (1)$$

This probabilistic approach allows us to estimate the latent intensity of Valence, Arousal, and Dominance for each text item on a continuous scale. By deriving scalar values from pairwise comparisons, we achieve a measure where differences are meaningful and robust, avoiding the calibration issues inherent in direct absolute scale prompting.

4.4 Model Training

To demonstrate the utility of our EmoPair dataset we fine-tuned RoBERTa-large models with regression heads using both the Prompting-based Emotion Regression from Transformers (PERT) methodology described by Bulla and Mongiovì (2024)—the best performing model on EmoBank as of our writing—as well as reward-based models and compared model performance from EmoBank and EmoPair. This 4-way modeling approach allows us to examine the efficacy of pairwise-derived

scalar scores against current state-of-the-art techniques.

4.4.1 Architecture 1 - PERT

As our baseline we retained the state-of-the-art emotion regression model as of our writing, PERT - Prompted Emotion Regression Transformer (Bulla and Mongiovì, 2024) using the EmoBank dataset. PERT leverages the latent knowledge of pre-trained language models (here, RoBERTa (Liu et al., 2019b)) by wrapping input texts in natural language prompts (e.g., "The level of arousal in the context is `<mask>`") during the fine-tuning process to take advantage of the format of text used in training BERT-like encoder models.

4.4.2 Architecture 2 - Reward Modeling

Following Licht et al. (2025), we examined the potential for further performance improvements by using a reward model architecture. We employ a unified reward model for predicting emotion dimensions across three complementary tasks. The architecture consists of a pre-trained RoBERTa-large encoder (355M parameters) followed by task-specific prediction heads. The encoder processes tokenized text with a maximum sequence length of 256 tokens, and its [CLS] token representation is passed through two heads:

- **Valence Head:** A linear layer with Xavier initialization that outputs a single continuous value for Valence regression on the 1–5 emotional scale.
- **Arousal-Dominance Head:** A linear layer that jointly predicts arousal and dominance as a two-dimensional output for pairwise decision tasks.

A dropout layer (p = 0.2) is applied between the encoder and prediction heads to reduce overfitting. The multi-task design allows the model to learn Valence through direct regression (see Section 5.1 for rationale) while simultaneously learning arousal and dominance through pairwise comparisons grounded in reference-based annotations. To isolate architecture from data effects, we also trained the reward model on EmoBank annotations recast as pairs (see Appendix C).

The training methodology for both model architectures can be found in Appendix D.

5 Results

5.1 Building EmoPair

We first assessed whether replacing direct scalar ratings with pairwise comparisons improves annotator reliability. As shown in Table 1, IRR increases across all three VAD dimensions under a paired comparison framework. For Valence, IRR rises from 0.738 in EmoBank to 0.929 in EmoPair, an improvement of 0.191 (26%). Improvements are even larger for Arousal and Dominance, which increase by 0.301 (51%) and 0.295 (52%), respectively. Collectively, these findings demonstrate that pairwise comparison yields higher agreement than Likert-style scoring, with particularly large gains for the more challenging constructs of Arousal and Dominance, as discussed in Section 2.1.

Table 1: Pairwise Comparison IRR Improvements

Construct	Number of Items	EmoBank IRR	EmoPair IRR	Improvement
Valence	170	0.738	0.929	0.191
Arousal	204	0.595	0.896	0.301
Dominance	208	0.57	0.865	0.295

Note: The Number of Items varies across dimensions due to the removal of items where one manual annotator chose ties. See Appendix F for more details.

We next evaluated whether LLMs can serve as statistically defensible substitutes for manual annotators under the AltTest. The results are strongly construct dependent, as seen in Figure A.1. For the Valence dimension, no model meets the substitution criterion of $\omega \geq 0.50$-indicating that none of the evaluated LLMs can reliably replace manual annotators for this dimension. We discuss potential rationale for this in the Section 8 and Appendix E. By contrast, both Arousal and Dominance admit viable alternatives.

When examining the advantage probabilities of each model depicted in Figure 2, we find that for Dominance, two models satisfy the win-rate criterion at or below the recommended $\epsilon = 0.10$ value, but Claude Sonnet 4.5 shows the strongest overall advantage probability ($\rho = 0.88$). For Arousal, GPT-5.1 provides the strongest evidence of substitution; the model achieves the highest average advantage probability ($\rho = 0.89$) on this dimension and surpasses the win-rate threshold with a negative ϵ value ($\epsilon = -0.01$), indicating that two of our annotators (the 50% threshold needed) required their probabilities of alignment to be lowered instead of the model.

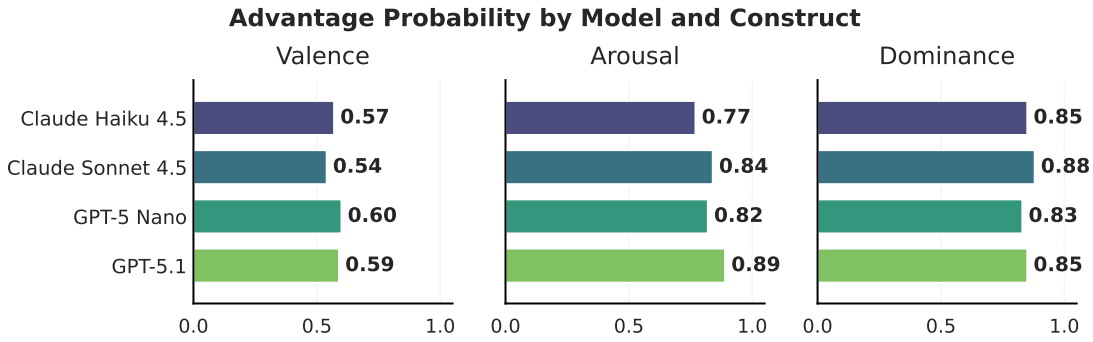

Figure 2: **Average Advantage Probability, ρ,** across four LLMs for the Valence, Arousal, and Dominance constructs. Higher values indicate a greater probability that a model performs as well as or better than a randomly selected human annotator.

Finally, we used these validated models to complete paired comparisons of EmoBank text and then we computed BT scores for Arousal and Dominance. These scores were then appended to the original EmoBank file for release. We also release the complete set of manual and LLM-generated pairwise judgments. As shown in Figure A.2, when compared to the EmoBank score distributions (rescaled), EmoPair BT scores exhibit greater dispersion and less clustering around the midpoint, suggesting that the pairwise framework also mitigates the potential centrality bias observed in the direct scalar ratings (Douven, 2018). BT scores also maintain significant rank-order alignment with original EmoBank scores (Figure A.2), suggesting the transformation increases differentiation without disrupting the underlying structure.

5.2 Modeling with EmoPair

We next examine whether these annotation differences yield downstream modeling gains. We compare four RoBERTa-large models on the EmoPair test set ($N = 1,000$ items; $8,484$ held-out pairs) that vary by the crossing training data (EmoPair vs. EmoBank) and by model architecture (PERT regression vs. reward-based pairwise modeling). This design isolates the relative contributions of the supervision source and model architecture. We evaluate model performance for each VAD dimension across three criteria: Pearson correlation with original EmoBank scalar scores (**orig r**), Spearman correlation with Bradley-Terry derived scores (**BT r**), and pairwise comparison accuracy (**Pair Acc**).

We find the two datasets diverge the least for Valence. As shown in the performance profiles in Figure 3, PERT-EmoBank and PERT-EmoPair models are effectively tied in their correlation with held-

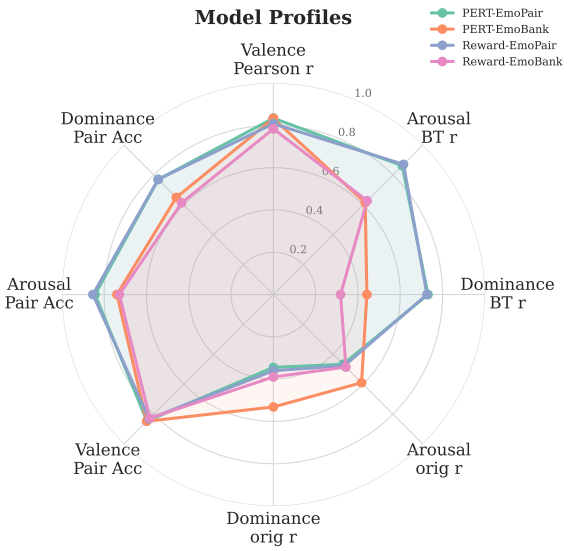

Figure 3: **Performance profiles of four model–dataset combinations across eight evaluation metrics.** Each axis represents a distinct criterion: Pearson correlation with EmoBank scores (orig r), Spearman correlation with BT scores (BT r), and pairwise accuracy (Pair Acc).

out Valence labels: PERT-EmoPair achieves a Pearson correlation of 0.8338, compared to 0.8326 for PERT-EmoBank, and the bootstrap confidence interval for their difference includes zero. In contrast, reward-based models perform slightly worse than both PERT variants on this dimension. Taken together, these results suggest that the original EmoBank labels remain a useful calibration target for Valence, but do not provide evidence of a meaningfully stronger underlying signal than EmoPair.

A very different pattern emerges for Arousal and Dominance when evaluating against EmoPair-defined BT targets. For Arousal, Reward-EmoPair and PERT-EmoPair achieve correlations of 0.87

and 0.86, respectively, while Reward-EmoBank and PERT-EmoBank reach only 0.63 and 0.61. A similar gap appears for Dominance: PERT-EmoPair and Reward-EmoPair achieve correlations of 0.73 and 0.73, compared to 0.44 and 0.32 for the EmoBank-trained models. This pattern is largely mirrored in pairwise accuracy, which more closely reflects the underlying annotation task. For Arousal, EmoPair-trained models reach 85.37% and 84.34% accuracy, versus 73.99% and 72.82% for EmoBank-trained models. For Dominance, the corresponding accuracies are 77.03% and 76.96%, compared to 64.93% and 61.33%. Overall, these results indicate that models trained on EmoPair generalize substantially better than those trained on EmoBank.

Notably, the two best-performing, EmoPair-trained models—Reward-EmoPair and PERT-EmoPair—are statistically indistinguishable on the EmoPair-defined BT targets. For Arousal, the bootstrap difference in BT correlation between Reward-EmoPair and PERT-EmoPair is 0.0090 (95% CI [-0.0032, 0.0216]). For Dominance, the corresponding difference (PERT-EmoPair minus Reward-EmoPair) is 0.0050 (95% CI [-0.0135, 0.0236]). Because both confidence intervals include zero and the effect sizes are minimal, these results suggest that the primary driver of improvement is not model architecture but training data. Once either architecture is trained on EmoPair labels—either regressed on BT scores or on reward signals from paired comparisons—performance increases substantially on the comparative targets the dataset was designed to capture.

When evaluation shifts from EmoPair-defined BT targets to the original EmoBank score averages for Arousal and Dominance, the ranking reverses. PERT-EmoBank achieves the strongest correlations with the legacy scales (r = 0.5898 for Arousal and r = 0.5317 for Dominance), and its advantage over Reward-EmoBank is statistically significant in both cases. This reversal does not indicate that EmoBank provides better supervision for comparative affect modeling. Rather, it reflects that models trained on EmoBank are better calibrated to reproduce the scalar scoring system that EmoBank itself defines. In other words, EmoBank is most applicable when the goal is compatibility with the legacy scale, whereas EmoPair is better suited for modeling comparative affect judgments and likely better reflects true placement on the construct scale.

6 Limitations

A primary limitation of our methodology is the bounds of its generalizability and the strict necessity for construct-specific validation. As demonstrated by the model's AltTest failure on Valence, an LLM that serves as a highly reliable annotator for one construct is not automatically reliable for another. This holds true even when the tasks share a corpus, prompt structure, annotation format, and unified theoretical relation. Consequently, expanding this methodology requires researchers to rigorously validate LLM capabilities on a construct-by-construct basis and remains an active area of research. Additionally, another limitation of this study is the size and demographic scope of the manual annotation team. The relatively small number of annotators may limit the representativeness of the emotional baseline across broader, more diverse populations.

Furthermore, the CGCoT prompts used to structure LLM reasoning were designed by the authors. These prompts could potentially be improved through collaboration with psychological experts on emotions to more accurately capture the nuanced linguistic features of Valence, Arousal, and Dominance. Finally, while our methodology relied on a forced-choice paradigm for LLM stability, the optimal handling of "ties"—where two texts are perceived as affectively equal—remains a critical area for future work to better align automated systems with the subtle complexities of human emotional perception (see Appendix F for more detail).

7 Ethics

The authors of this paper served as the annotators for the purpose of validation and had full knowledge of the study's aims, data sources, and intended use of their annotations prior to participation. The lead author designed the codebook and trained annotators. All other authors served as annotators. We provide the codebook used in this process in Appendix B. All annotators were staff or student employees at the same university research center. Annotators were recruited from within the center and participated as part of their standard responsibilities. As such, this study did not require review board approval; the annotation task used only existing, publicly available data (EmoBank)

and was conducted exclusively by compensated staff and student researchers affiliated with the institution. No sensitive populations, deceptive procedures, or collection of personal data from external participants were involved.

Further, the primary use of LLMs in this work is as annotation instruments rather than as deployed models; the fine-tuned RoBERTa-large models used for downstream evaluation are comparatively lightweight encoder architectures. This design choice substantially limits the environmental footprint of the pipeline relative to approaches that deploy frontier LLMs for inference. Nevertheless, practitioners should exercise caution when applying EmoPair-trained models outside of research contexts. Automated systems that misinterpret emotional affect can cause harm in sensitive applications such as mental health monitoring, content moderation, and affective human-computer interaction. We emphasize that **EmoPair is released as a research artifact and has not been validated for deployment in such settings**. Additionally, because both EmoBank and our annotator pool reflect English-language, Western cultural contexts, models trained on EmoPair may not generalize reliably.

More broadly, we note that the decision to deploy LLMs for measuring emotional affect must be grounded in rigorous evidence rather than assumptions. To mitigate these risks and promote accountability in NLP research, we emphasize the transparency of the failures of trying to automate the Valence dimensions. By publishing negative results we aim to surface construct-specific weaknesses that would otherwise remain invisible. This transparency ensures that practitioners do not blindly trust LLM outputs, underscoring the necessity of evidence-based deployment when relying on automated systems to interpret complex human emotions.

8 Conclusion

This work contributes to a broader validity argument for modeling affective dimensions in the VAD framework through comparative, rather than absolute annotation. Our findings demonstrate that humans exhibit higher agreement when making pairwise judgments than when assigning scalar scores, and that LLMs trained on these comparative criteria can match (or sufficiently approximate) this level of agreement. This suggests that reliable supervision for these dimensions does not require exhaustive manual annotation of absolute values. Instead, pairwise comparisons can provide a more stable and scalable foundation for capturing the underlying affective structure while also offering a path toward improving annotation quality and expanding datasets.

We posit that LLMs are unable to match human reasoning on the Valence dimension in part due to a failure to accurately represent the construct. As shown in Appendix E, manual annotators exhibit substantially higher reliability when evaluating Valence compared to the other two dimensions across both the EmoBank and EmoPair data, suggesting that Valence is a more accessible and less ambiguous affective construct for humans still. It is possible that, in such settings, the margin for improvement is minimal - even well-trained models can only approximate, but not exceed, an already high level of human consensus.

More broadly, this work introduces two complementary approaches for addressing challenging NLP tasks: improving annotation quality through pairwise judgments and leveraging reward-based learning over paired data. By modeling differences between instances rather than predicting absolute labels, our framework better reflects how many social and affective constructs are expressed in practice. At the same time, it highlights the ongoing challenge of data augmentation in NLP, further underscoring the need for careful integration of LLMs with rigorous evaluation protocols. Together, these contributions point toward a flexible and generalizable paradigm for scaling the study of abstract, hard-to-annotate dimensions in the social sciences.

8.1 AI Usage

The authors used AI assistants at several stages of this project beyond as the subject of study. Generative AI tools were used in manuscript preparation to assist with grammar checking, prose editing, and length reduction; in code development for debugging assistance; and in checklist evaluation during submission preparation. All AI-generated suggestions were reviewed, edited, and approved by the authors, who take full responsibility for the content of this paper. The use of LLMs as experimental annotation instruments is distinct from and described separately in the main body of the paper.

References

Suhaib Abdurahman, Alireza Salkhordeh Ziabari, Alexander K. Moore, Daniel M. Bartels, and Morteza Dehghani. 2025. A Primer for Evaluating Large Language Models in Social-Science Research. *Advances in Methods and Practices in Psychological Science*, 8(2):25152459251325174.

Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/eng2.12189.

Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Mag.*, 36(1):15–24.

Margaret M Bradley and Peter J Lang. 1999. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology ….

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

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.

Luana Bulla and Misael Mongiovì. 2024. Adequate Prompting Improves Performance of Regression Models of Emotional Content. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, GoodIT '24, pages 135–142, New York, NY, USA. Association for Computing Machinery.

Nichola Burton, Michael Burton, Carmen Fisher, Patricia González Peña, Gillian Rhodes, and Louise Ewing. 2021. Beyond Likert ratings: Improving the robustness of developmental research measurement using best–worst scaling. *Behavior Research Methods*, 53(5):2273–2279.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *arXiv preprint arXiv:2501.10970*.

Shixiong Cao and Nan Cao. 2025. Emotion-aware design: Modulating valence, arousal, and dominance in communication via design. *arXiv preprint arXiv:2502.16038*.

M.L. Chrzan. 2025. pairadigm: A python library for concept-guided chain-of-thought pairwise measurement of scalar constructs using large language models.

A. P. Dawid and A. M. Skene. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28. _eprint:https://rss.onlinelibrary.wiley.com/doi/pdf/10.2307/2346806.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Jiawen Deng and Fuji Ren. 2023. A Survey of Textual Emotion Recognition and Its Challenges. *IEEE Transactions on Affective Computing*, 14(01):49–67.

M. J. Dileep Kumar, M. Sukesh Rao, and K. C. Narendra. 2025. Multimodal Emotion Recognition: A Comprehensive Survey of Datasets, Methods, and Applications. *IEEE Access*, 13:201067–201097.

Igor Douven. 2018. A Bayesian perspective on Likert scales and central tendency. *Psychonomic Bulletin & Review*, 25(3):1203–1211.

Shuxing Fang, Ruijian Han, Yuanhang Luo, and Yiming Xu. 2026. Recent advances in the bradley–terry model: theory, algorithms, and applications. *arXiv preprint arXiv:2601.14727*.

Johnny R.J. Fontaine, Klaus R. Scherer, Etienne B. Roesch, and Phoebe C. Ellsworth. 2007. The World of Emotions is not Two-Dimensional. *Psychological Science*, 18(12):1050–1057.

Gordon R. Foxall, Jorge M. Oliveira-Castro, Victoria K. James, M. Mirella Yani-de Soriano, and Valdimar Sigurdsson. 2006. Consumer Behavior Analysis and Social Marketing: The Case of Environmental Conservation. *Behavior and Social Issues*, 15(1):101–125.

Gihan Gamage, Daswin De Silva, Nishan Mills, Damminda Alahakoon, and Milos Manic. 2024. Emotion aware: An artificial intelligence framework for adaptable, robust, explainable, and multi-granular emotion analysis. *Journal of Big Data*, 11(1):93.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Ismail Alihan Hadimlioglu and Siddharth Linga. 2025. Face2feel: Emotion-aware adaptive user interface. *arXiv preprint arXiv:2510.00489*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners. *arXiv preprint*. ArXiv:2205.11916 [cs].

Hauke Licht, Rupak Sarkar, Patrick Y Wu, Pranav Goel, Niklas Stoehr, Elliott Ash, and Alexander Miserlis Hoyle. 2025. Measuring scalar constructs in social science with llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32132–32159.

Chen Liu, Muhammad Osama, and Anderson De Andrade. 2019a. Dens: A dataset for multi-class emotion analysis. *arXiv preprint arXiv:1910.11769*.

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

Frank J. Massey. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253):68–78.

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

Angeliki Metallinou and Shrikanth Narayanan. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, Shanghai, China. IEEE.

Saif M Mohammad. 2025. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *arXiv preprint arXiv:2503.23547*.

Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, 45(1):169–177.

Mary Luz Mouronte-López, Juana Savall Ceres, and Aina Mora Columbrans. 2023. Analysing the sentiments about the education system trough twitter. *Education and Information Technologies*, 28(9):10965–10994.

Ryotaro Nagase, Ryoichi Takashima, and Yoichi Yamashita. 2026. Color-based Emotion Representation for Speech Emotion Recognition. *arXiv preprint*. ArXiv:2602.16256 [eess] version: 1.

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. 2025. Are llms better than reported? detecting label errors and mitigating their effect on model performance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26770–26797.

Arbi Haza Nasution and Aytuğ Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *Ieee Access*, 12:71876–71900.

Minxue Niu, Yara El-Tawil, Amrit Romana, and Emily Mower Provost. 2025. Rethinking Emotion Annotations in the Era of Large Language Models. *IEEE transactions on affective computing*, 16(4):2668–2679.

Minxue Niu, Mimansa Jaiswal, and Emily Mower Provost. 2024. From text to emotion: Unveiling the emotion annotation capabilities of llms. *arXiv preprint arXiv:2408.17026*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint*. ArXiv:2203.02155 [cs].

Qian Pan, Zahra Ashktorab, Michael Desmond, Martin Santillan Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. 2024. Human-centered design recommendations for llm-as-a-judge. *arXiv preprint arXiv:2407.03479*.

Nicholas Pangakis and Samuel Wolken. 2024. Keeping humans in the loop: Human-centered automated annotation with generative ai. *arXiv preprint arXiv:2409.09467*.

Jaime Redondo, Isabel Fraga, Isabel Padrón, and Montserrat Comesaña. 2007. The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods*, 39(3):600–605.

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

Velina Slavova and Filip Andonov. 2021. How deeply are emotions encoded in language communication and is this detectable in text. *International Journal "Information Theories and Applications"*, 28(3):271–299.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.

Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*. Association for the Advancement of Artificial Intelligence.

Melissa L. H. Võ, Markus Conrad, Lars Kuchinke, Karolina Urton, Markus J. Hofmann, and Arthur M. Jacobs. 2009. The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods*, 41(2):534–538.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint*. ArXiv:2201.11903 [cs].

Patrick Y Wu, Jonathan Nagler, Joshua A Tucker, and Solomon Messing. 2024. Concept-guided chain-of-thought prompting for pairwise comparison scoring of texts with large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pages 7232–7241. IEEE.

M. Mirella Yani-de Soriano and Gordon R. Foxall. 2006. The emotional power of place: The fall and rise of dominance in retail research. *Journal of Retailing and Consumer Services*, 13(6):403–416.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint*. ArXiv:2306.05685 [cs].

A Additional Figures

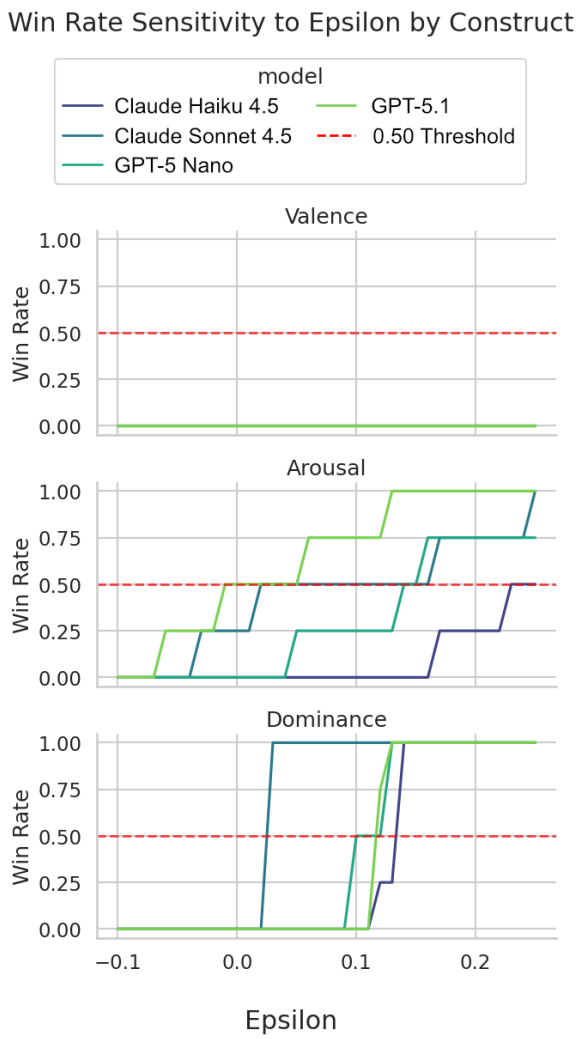

Figure A.1: **Win Rate, ω, sensitivity to the cost-benefit hyperparameter, ϵ**, across Valence, Arousal, and Dominance. Each curve shows which ϵ value models cross the 0.50 Win Rate threshold, indicating the quality of annotators the model is valid to replace. Given that the annotators here are trained staff, a model with $\omega \geq 0.50$ at $\epsilon = 0.10$ is considered good.

B Annotation Task

To ensure representative coverage of the EmoBank dataset within the subset chosen for pairs, we employed stratified sampling. We systematically evaluated 1,331 binning configurations (11 bin sizes per dimension: 5, 10, 20, 30, 40, 50, 100, 250, 500, 750, and 1,000 bins) to identify optimal stratification parameters that would yield a sample statistically representative of the full distribution. For each configuration, we created strata by discretizing the continuous VAD scores and allocated samples proportionally to stratum size (n=50 total). We assessed distributional similarity using two-sample Kolmogorov-Smirnov tests (Massey, 1951), requiring p > 0.05 for all three dimensions to ensure no significant deviation from the original distribution. Among valid configurations, we selected the one maximizing the minimum p-value across dimensions, which corresponded to V=250, A=40, and D=30 bins. This approach balanced granular stratification of the emotional space with sufficient samples per stratum, yielding a final sample of 50 text instances that faithfully represented the multidimensional emotional distribution of the source corpus (minimum KS test p-value: 0.22).

From these 50 items, we use the `pairadigm` package to generate pairings such that each item had 10 pairs, following Wu et al. (2024). This resulted in 293 pairs per dimension for annotators to evaluate. To ensure statistical connectivity and adequate comparison coverage, `pairadigm` employs a graph-theoretic pairing strategy that guarantees each item appears in a minimum number of comparisons while maintaining a connected comparison graph necessary for reliable scoring.

B.1 Human Annotations

Human annotators were the provided with three sets of the pairs - one for each dimension - and prompted with the following prompts for their choices for each dimension:

- Valence: Answer 1 or 2 (or 0 if tied) - Which sentence sounds more positive or pleasant? Choose the sentence that feels happier, more pleased, or more content.
- Arousal: Answer 1 or 2 (or 0 if tied) - Which sentence sounds more emotionally intense rather than calm or relaxed? Choose the sentence that feels more excited, more agitated, or more alert.
- Dominance: Answer 1 or 2 (or 0 if tied) - Which sentence implies that the speaker has more power over events or decisions in their situation? Choose the sentence that sounds more controlling, more commanding, or more empowered.

The full codebook is included in G.

B.2 CGCoT Prompts

The following figures display the prompts used to generate the breakdowns that were used for the pairwise comparisons via `pairadigm`. For the

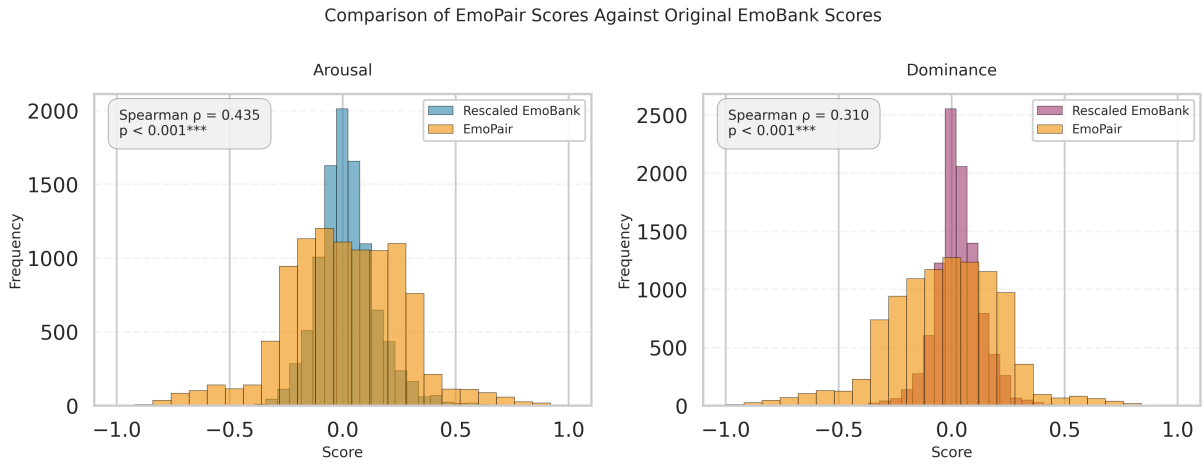

Figure A.2: **Score Distribution Comparison**. After linearly rescaling EmoBank scores to $[-1, 1]$, we see that the EmoPair scores contain much more variance. Despite differences in distribution shape, EmoPair scores maintain significant rank-order correlation with rescaled EmoBank scores for both Arousal ($\rho = 0.435$, $p < 0.001$) and Dominance ($\rho = 0.310$, $p < 0.001$), indicating consistent relative ordering of items across both annotation approaches.

pairwise annotation, we used the default comparison prompt in the package.

Figure B.1: CGCoT Prompts for Valence.

```
1. Summarize the text: {text}
2. Who or what is the primary target or
    focus of the following text (person,
    group, idea, object, or no clear target)
    ? Answer with a short label. Text: {text
    }
3. Describe the overall evaluative tone
    toward the target in the original text
    using concrete words (examples: praising
    , appreciative, approving,
    congratulatory; neutral, descriptive,
    matter-of-fact; critical, hostile,
    insulting, disparaging) and CONCISELY
    cite textual cues that justify this
    label. Text: {text}
4. Valence refers to how pleasant or
    unpleasant an emotion feels. High
    valence relates to positive, attractive
    feelings (e.g., joy), while low valence
    relates to negative, aversive feelings (
    e.g., sadness). Based on your previous
    analysis, provide a single categorical
    judgment: does the original text express
     Positive Valence, Neutral Valence, or
    Negative Valence toward the identified
    target? Answer with only one of those
    three words. Text: {text}
```

C Generating Paired Choices from EmoBank

For each unique text item, we collect the full distribution of reader-provided Likert scores on each affective dimension. We then construct all unordered

Figure B.2: CGCoT Prompts for Arousal.

```
1. Summarize the text: {text}
2. Identify the primary emotion or affective
     state conveyed in the original text if
    any (examples: anger, excitement,
    sadness, boredom, urgency, amusement,
    none). Text: {text}
3. Describe the intensity of the emotional
    expression in the original text using
    descriptive labels (calm, mild, moderate
    , high, explosive) and CONCISELY cite
    textual cues that justify this label (
    examples: exclamation marks, all-caps,
    repeated punctuation, interjections,
    short emphatic sentences). Text: {text}
4. Arousal describes the intensity or energy
     level of an emotion, ranging from calm
    or low activation (e.g., relaxed) to
    excited or high activation (e.g., anger
    or exhilaration). Based on your previous
     analysis, provide a single categorical
    judgment of the original text: Low
    Arousal, Moderate Arousal, or High
    Arousal only using one of those three
    labels. Text: {text}
```

pairs of text items and compute a probabilistic preference score for each pair separately for V, A, and D. For a given dimension and a pair of texts (A, B), this probability is estimated by comparing *all ratings* assigned to A with *all ratings* assigned to B. Specifically, we calculate the proportion of cross-item rating pairs in which A receives a higher score than B, with tied ratings contributing half a point. This approach produces a nonparametric estimate of P (A > B) that reflects overall differences in

Figure B.3: CGCoT Prompts for Dominance.

```
1. Summarize the text: {text}
2. Identify whether the text author of the
    original text adopts an Assertive/
    Controlling stance, a Neutral/
    Informative stance, or a Submissive/
    Deferential stance. Provide the label
    and a one-sentence justification. Text:
    {text}
3. Cite up to three textual features that
    support the stance identified previously
     in the original text (examples:
    imperatives "Do X", threats, confident
    factual claims without hedging; hedges "
    maybe", "I think", questions; apologies,
     disclaimers). Text: {text}
4. Indicate whether the speaker claims
    authority, issues a command, asks for
    permission, or expresses uncertainty in
    this text: {text}
5. Dominance indicates the extent to which a
     person feels in control or powerful
    versus controlled or powerless during an
     emotional experience. High dominance
    means feeling empowered or influential (
    e.g., anger), while low dominance
    reflects feelings of being dominated or
    overwhelmed (e.g., fear). Based on your
    previous answers, provide a single
    categorical judgment on Dominance: High
    Dominance, Medium Dominance, or Low
    Dominance only using one of those three
    labels. Text: {text}
```

the distributions of ratings while preserving heterogeneity across raters.

$$P(A > B) = \frac{N_{A>B} + 0.5\, N_{A=B}}{|A| \times |B|} \qquad (2)$$

For use in downstream reward model architecture, we additionally derive a binary comparison indicator for each item pair, coding whether the estimated probability exceeds 0.5 as indicative of the chosen item in the pair. This representation corresponds to the input required by Bradley–Terry–type choice models and related probabilistic ranking frameworks. All computations are carried out independently for Valence, Arousal, and Dominance, treating each dimension as a separate latent affective scale. This approach generated C(10,062, 2) × 3 pairwise probabilities which were then randomly sampled for reward model fine-tuning. The resulting pairwise dataset preserved the distributional information from multiple raters while enabling rank-based analysis through the Bradley-Terry framework.

D Reward Model Training Methodology

While for the PERT models we train with MSE loss across all dimensions using AdamW optimization (learning rate: 3×10^{-5}, no weight decay) with linear warmup scheduling to following Bulla and Mongiovì (2024), we train the reward model using a hybrid loss function that accounts for the distinct nature of valence and arousal-dominance labels:

- **Arousal and Dominance (Pairwise):** Binary Cross-Entropy against human pairwise preferences and EmoBank-derived Bradley-Terry indicators for the EmoBank model (see Appendix C) and LLM-derived annotations for the EmoPair model.
- **Valence Loss (Pointwise):** Average Mean Squared Error of both predicted scores against gold valence scores from EmoBank annotations (1–5 scale).

Data are split into training, development, and test sets following the item splits designed in EmoBank (Buechel and Hahn, 2017) such that paired items remain within the same split to preserve independence. We use AdamW optimization with a learning rate of 2×10^{-6}, weight decay of 0.05, and a linear warmup schedule (7.5% of total steps). Training employs gradient accumulation (effective batch size: 64 across 4 accumulation steps of mini-batches of 16) for 20 epochs with early stopping (patience: 7 epochs on development loss). Mixed-precision training (float16) with automatic loss scaling stabilizes convergence.

All models were trained on NVIDIA RTX 4000 Ada Generation GPUs hosted on a private server maintained by the university. Each model configuration required on the order of a few hours of wall-clock training time. Precise GPU-hour totals were not logged during training runs; however, full training output logs are available in the code repository, which provides per-epoch loss and timing information sufficient for readers to estimate the computational requirements for reproduction. By our estimate, the training time was approximately 4 hours for each model.

E Exploring Valence AltTest Failure

As shown in Figure A.1, the AltTest produced a clear asymmetry across the three VAD dimensions: the LLMs tested achieved Win Rates (ω) meeting or exceeding our 0.50 threshold for Arousal and Dominance across a range of ε values, but no

model crossed that threshold for Valence, constituting a failure of the AltTest for that dimension. This section examines the likely sources of that failure using the Dawid-Skene (DS) reliability estimates computed across all annotators—manual and LLM alike—presented in Figure E.1.

The DS model, unlike simple pairwise agreement statistics or the AltTest which builds off of them, estimates each annotator's latent sensitivity to the true underlying construct by modeling the full pattern of agreements and disagreements across the annotator pool. Higher DS reliability indicates that an annotator's judgments more faithfully track the latent construct as approximated by the collective consensus (Dawid and Skene, 1979). Figure E.1 reveals a striking reversal across dimensions. For Arousal and Dominance, the LLMs occupy the top positions in the DS reliability ranking: on Arousal, GPT-5 Nano (0.984), GPT-5.1 (0.963), Claude Sonnet 4.5 (0.944) all rank above any manual annotator, with the most reliable human (A3 = 0.896) ranking fourth. The same pattern holds for Dominance, where Claude Sonnet 4.5 and Claude Haiku 4.5 (0.851) lead the ranking. This is precisely what one would hope for from a successful AltTest: the LLMs are not merely consistent internally, but are more consistent with the ground truth than the typical manual annotator.

For Valence, the picture inverts. The top three positions in the DS reliability ranking are held by manual annotators (A3 = 0.935, A2 = 0.881, A4 = 0.868), with LLMs occupying - almost exclusively - the lower positions. While the absolute differences are modest, the consistent ordering across LLMs and manual annotators points to a systematic, rather than incidental, difference in sensitivity to the Valence construct. The LLMs demonstrate lower fidelity to the consensus signal on Valence precisely when the manual annotators are at their most cohesive—a dynamic that directly explains the AltTest failure, since the test evaluates LLM performance relative to the agreement floor set by individual manual annotators.

We hypothesize that this pattern reflects a difference in construct granularity between Valence and the other two dimensions. Valence—the pleasantness or unpleasantness of an emotional state—is the dimension most proximate to sentiment, a construct that humans encounter and implicitly evaluate in everyday language more consistently. As a result, manual annotators appear to have developed finely calibrated, shared intuitions for Valence that are robust even at the pairwise margin, where items may differ only subtly. The high Krippendorff's α for human Valence annotations (0.929, Table 1)—the highest of any dimension—supports this interpretation. The LLMs, despite strong average performance, appear less sensitive to these marginal distinctions, likely because the pairwise task surfaces edge cases where the hedonic signal is weak, ambiguous, or depends on pragmatic inference that the models weigh differently.

Taken together, these findings suggest that the pairwise comparison framework is especially well-suited to constructs that are less intuitively familiar to annotators—where reducing cognitive load and anchoring judgments to direct comparisons has the most to offer. For Valence, human annotators appear to have reached a level of agreement that the current generation of LLMs cannot reliably match on nuanced pairs - even with CGCoT surfacing nuance - making LLM replacement inappropriate under the AltTest's statistical criterion. This does not diminish the utility of the EmoPair scores from models; it does, however, argue for retaining manual annotation for Valence when expanding this corpus, or for exploring more targeted prompting strategies designed to improve LLM sensitivity to hedonic edge cases.

This result also perfectly exemplifies the need for validation of LLM's capabilities when considering deploying them at scale. Had we proceeded without the AltTest—relying instead on the LLMs' strong aggregate performance on traditional agreement metrics and the presumed prevalence of valence-related data in training given its relatedness to sentiment—we would have silently introduced a lower-quality signal into the Valence dimension of EmoPair, with no statistical warning to alert downstream users. The AltTest's leave-one-out design is precisely calibrated to catch this failure mode: it does not ask whether an LLM performs well in absolute terms, but whether it performs well relative to the human annotators it is being asked to replace, on the specific construct and task at hand.

That distinction matters enormously in practice. A model that is a reliable annotator for Arousal is not automatically a reliable annotator for Valence, even when the two tasks share a prompt structure, a corpus, an annotation format, and a unified relation in theory. Generalizing from success on one construct to assumed competence on another is exactly the kind of inferential shortcut that rig-

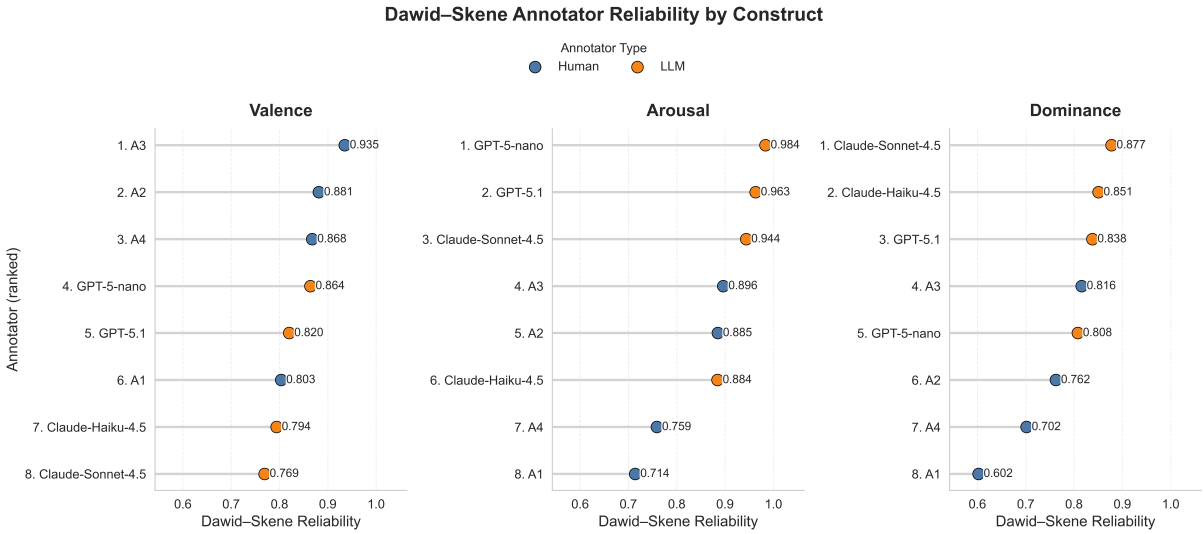

Figure E.1: Each point represents the reliability of a single annotator-human (blue) or LLM (orange)—as estimated by the Dawid-Skene model, which measures the degree to which an annotator's judgments track the latent consensus signal across all annotators and items. Annotators are ranked within each construct from highest to lowest reliability. The construct-specific asymmetry directly explains the AltTest failure for Valence and underscores the importance of per-dimension validation rather than generalizing annotation reliability across constructs.

orous validation procedures are designed to prevent and that an ever-growing chorus of researchers and practitioners is calling for in studies using measures derived from LLMs (Gilardi et al., 2023; Pangakis and Wolken, 2024; Abdurahman et al., 2025; Licht et al., 2025). The Valence AltTest failure thus serves not as a limitation of the EmoPair methodology, but as a demonstration that the methodology is working as intended, surfacing construct-specific weaknesses that would otherwise be invisible and ensuring that the decision to deploy LLMs at scale is grounded in evidence rather than assumption.

F Guidance for Annotating Ties

A critical methodological consideration in the creation of EmoPair was how to handle "ties" (instances where two texts are deemed to have equal levels of a given VAD construct) during pairwise comparison. During our validation, we observed significant tradeoffs when allowing manual annotators to declare ties versus allowing LLMs the same option.

For manual annotation, allowing ties proved highly beneficial. Human annotators inherently vary in their subjective sensitivity to emotional constructs. When the affective difference between two texts is incredibly subtle, forcing a person to make a definitive choice often introduces noise and reduces inter-rater reliability. In these instances, allowing a "tie" provides a necessary pressure valve that accurately reflects human perceptual thresholds.

Conversely, we found that LLM annotation performance degraded when the model was allowed to output a tie. Given the option, LLMs frequently over-relied on the "tie" classification as a safe fallback for complex or nuanced comparisons, which ultimately diminished the discriminatory power of the resulting dataset. Through iterative testing, we determined that LLMs perform best when presented with a strict forced-choice prompt. By removing the tie option, the LLM is forced to compute and leverage minute probabilistic differences in the text, yielding a more accurate and scalable emotional mapping than when given the option to remain neutral. Therefore, we recommend allowing ties for manual baseline annotations, but strictly enforcing forced-choice prompts for LLM-driven pairwise comparisons.

G Codebook for Human Annotators

Overview

This codebook defines the instructions, decision rules, examples, edge cases, quality controls, and adjudication protocol for a pairwise sentence comparison task with three independent judgment dimensions: Valence (pleasantness), Arousal (emotional intensity), and Dominance (perceived control). Each dimension is judged separately for the

same sentence pair. Annotators answer exactly one choice per dimension according to the prompts provided.

VAD Background

The proliferation of digital text—from social media posts and consumer reviews to vast archives of literature and news—has created an unprecedented opportunity to study human expression at scale. A central challenge in this endeavor is the computational analysis of emotion, a task commonly known as sentiment analysis. For years, this field has been dominated by a relatively simple classification of text into positive, negative, or neutral polarities. While useful for broad-stroke assessments, this approach fails to capture the rich, multifaceted nature of human affective states. A review glowing with serene contentment and one buzzing with frenetic excitement are both "positive," yet they describe profoundly different emotional experiences. Similarly, the quiet despair of a sad text is affectively distinct from the high-energy fury of an angry one. To move beyond this coarse granularity, a more sophisticated and psychologically grounded framework is required.

The Valence, Arousal, and Dominance (VAD) model offers such a framework. Rooted in decades of psychological research, it deconstructs emotion not into discrete categories but into three continuous, fundamental dimensions: Valence (the degree of pleasantness), Arousal (the level of activation), and Dominance (the sense of control). This dimensional approach provides a high-resolution map of the affective landscape, allowing any emotional state to be represented as a unique coordinate in a three-dimensional space. This capacity for nuance makes the VAD model exceptionally well-suited for computational applications, where the complexity of human feeling can be translated into quantifiable, analyzable data.

To understand how Valence, Arousal, and Dominance are used to analyze text, one must first grasp their origins as a comprehensive theory of human emotion. The VAD model is not an ad-hoc creation for computational linguistics but is built upon a robust psychological foundation designed to capture the fundamental structure of all affective experience. This section explores the model's genesis, defines its three core dimensions, and situates it within the wider scientific discourse on emotion.

Defining the Three Dimensions of Affect

The power of the VAD model lies in its three orthogonal dimensions, which work in concert to provide a unique "affective coordinate" for any emotional state. Each dimension is a continuous scale, allowing for the representation of subtle variations and intensities of feeling.

Valence (Pleasure-Displeasure)

Valence is the most intuitive of the three dimensions, capturing the hedonic quality or pleasantness of an experience. It is a bipolar scale that ranges from highly positive feelings at one end to highly negative feelings at the other. Adjectives associated with high valence include

- happy, pleased, satisfied, contented, and hopeful, while low-valence states are described by terms like
- unhappy, annoyed, unsatisfied, melancholic, despaired, and bored.4

In simple terms, valence answers the question: "Does this feel good or bad?" For instance, the emotions of joy and love score high on the valence scale, whereas anger, fear, and sadness all score on the low, or displeasure, side. It is the primary dimension used in traditional sentiment analysis, corresponding directly to positive versus negative polarity.

Arousal (Activation-Deactivation)

Arousal measures the level of physiological and psychological activation, energy, or stimulation associated with an emotional state. This dimension ranges from states of very low activation, such as

- sleepy, calm, sluggish, or dull, to states of extremely high activation, like
- excited, frenzied, jittery, or wide-awake.

It is **crucial to distinguish arousal from emotional intensity**. While the two are often correlated, they are not synonymous. For example, profound grief or depression can be intensely felt emotions, yet they are characterized by very low arousal. Conversely, rage is a high-arousal state, while anger may be less so. Arousal acts as an amplifier for valence; an increase in arousal can transform a state of contentment (high valence, low arousal) into one of delight (high valence, high arousal), or it can escalate frustration (low valence, moderate arousal) into full-blown anger (low valence, high arousal).

Dominance (Control-Submissiveness)

Dominance is arguably the most sophisticated dimension of the model, capturing the degree to which an individual feels in control, influential,

and dominant versus controlled, influenced, and submissive within a situation. This dimension is what provides the VAD model with much of its explanatory power, as it can distinguish between emotions that appear very similar on the valence-arousal plane.

The classic example is the distinction between anger and fear. Both are unpleasant (low valence) and highly activating (high arousal) emotions. However, they differ starkly in dominance. Anger is a high-dominance emotion, associated with a feeling of power and control, often motivating an individual to confront a threat or obstacle. In contrast, fear is a low-dominance emotion, characterized by a feeling of being powerless and controlled by external circumstances, typically motivating avoidance or submission.

While valence and arousal describe the raw feeling state—its pleasantness and energy level—**dominance speaks to the individual's sense of agency within that state**. This makes it a critical dimension for understanding action-oriented emotions. The link between dominance and behavior is not merely speculative; it has been explicitly connected to the conative (i.e., behavioral) component of the well-established ABC Model of Attitudes, which breaks down attitudes into Affective (feeling), Behavioral (acting), and Cognitive (thinking) components. Analyzing the dominance score of a text, therefore, offers a window not just into how the text is intended to make a reader feel, but what it might impel the reader to do. A text high in negative valence and arousal could signal general distress, but a high dominance score would suggest outrage and a potential for mobilization, whereas a low dominance score would imply despair and potential passivity.

Task and answer formats

- Dimension 1 — Valence (pleasantness)
  - Prompt: Answer 1 or 2 or 0 if tied — Which sentence sounds more positive or pleasant? Choose the sentence that feels happier, more pleased, or more content.
  - Allowed labels:
    - 1 (Sentence 1 more positive),
    - 2 (Sentence 2 more positive),
    - 0 (tied / equally positive).
- Dimension 2 — Arousal (intensity)
  - Prompt: Answer 1 or 2 or 0 if tied — Which sentence sounds more emotionally intense rather than calm or relaxed? Choose the sentence that feels more excited, more agitated, or more alert.
  - Allowed labels:
    - 1 (Sentence 1 more intense),
    - 2 (Sentence 2 more intense),
    - 0 (equally intense).
- Dimension 3 — Dominance (control)
  - Prompt: Answer 1 or 2 or 0 if tied — Which sentence implies the speaker has more power over events or decisions? Choose the sentence that sounds more controlling, more commanding, or more empowered.
  - Allowed labels:
    - 1 (Sentence 1 more controlling),
    - 2 (Sentence 2 more controlling),
    - 0 (tied / equal control).

Broader directions and decision rubric

- Valence (pleasantness)
  - Choose the sentence that communicates or implies greater positive affect, warmth, pleasure, satisfaction, or optimism. Prefer explicit lexical cues of happiness, gratitude, relief, comfort, praise, or positive appraisal. Choose the sentence with less negativity, less complaint, less sadness, or less criticism. If one sentence contains mixed cues, weigh overall tone and likely immediate emotional reaction of a typical reader or speaker.
- Arousal (emotional intensity)
  - Choose the sentence that reads as more activated, energetic, agitated, excited, alarmed, frantic, or urgent. Prefer markers such as exclamation, interjections, high-energy verbs, increased pace, punctuation that signals intensity, sensory vividness, or verbs implying action. Choose the calmer sentence when language is quiet, measured, sleepy, relaxed, or subdued.
- Dominance (perceived control)
  - Choose the sentence whose speaker appears to exert more agency, authority, or control in the situation. Favor direct commands, decisions, explicit assertions of power, decisive language, refusal, ordering, allocation of outcomes, or first-person claims of control. Choose the

sentence implying submission, passivity, uncertainty, or being controlled when appropriate. If both sentences describe third-party situations, judge which sentence attributes stronger agentic position to its speaker.

Decision rules common to all dimensions

- Make a single comparative judgment; do not split or hedge (e.g. do not mark 1.5).
- If the two sentences are essentially identical in the targeted attribute, mark the tied/equal option, 0.
- Judge only one dimension at a time. If different attributes conflict (one sentence is slightly more positive but much more intense), judge the dimension asked and ignore other dimensions.

Short, clear examples

- Pair A
  - Sentence 1: "I got the promotion today."
  - Sentence 2: "I was told I might be considered for a promotion."
    - Valence: 1
    - Arousal: 1
    - Dominance: 1
- Pair B
  - Sentence 1: "Please calm down, everything will be fine."
  - Sentence 2: "Get out of my office now."
    - Valence: 0
    - Arousal: 2
    - Dominance: 2
- Pair C
  - Sentence 1: "I am so relaxed on this holiday."
  - Sentence 2: "I can't wait for the concert tonight!"
    - Valence: 0
    - Arousal: 2
    - Dominance: 1
- Pair D (tie example)
  - Sentence 1: "The cat sat on the mat."
  - Sentence 2: "A cat is sitting on a mat."
    - Valence: 0
    - Arousal: 0
    - Dominance: 0

Edge cases and disambiguation rules

- Mixed signals: If a sentence contains both calm wording and a single intense word, weigh overall feel; a brief intense token can raise Arousal but not necessarily Valence or Dominance.
- Negation flips: "I love this" versus "I don't love this" — treat negation literally and choose the sentence reflecting the stronger valence accordingly.
- Sarcasm and irony: If sarcasm is explicit or strongly implied, judge perceived surface meaning for Valence and Arousal as read by an average reader. If sarcasm makes the literal sentiment ambiguous, choose the tie value when average interpretation is unclear and make a note of the confusion.
- Commands vs permission requests: Imperatives and direct orders generally increase Dominance; polite requests reduce perceived dominance.
- Third-person descriptions: When sentences describe others, infer which sentence attributes greater power to the speaker by context words like "I decided", "I told them", "They forced me" and choose accordingly.
- Punctuation emphasis: Exclamation marks and repeated punctuation increase Arousal. Ellipses and trailing sentences reduce Arousal.
- Single-word or fragment sentences: Judge on implied tone and typical interpretation.
- Cultural references or slang: Use the typical mainstream interpretation; when you cannot resolve cultural-specific meanings, choose the tie value for that dimension.