Appraisal Trajectories in Narratives Reveal Distinct Patterns of Emotion Evocation

Johannes Schäfer, Janne Wagner, and Roman Klinger
Fundamentals of Natural Language Processing, University of Bamberg, Germany
{johannes.schaefer,roman.klinger}@uni-bamberg.de,
janne.wagner@stud.uni-bamberg.de

Abstract

Understanding emotion responses relies on reconstructing how individuals appraise events. While prior work has studied emotion trajectories and inherent correlations with appraisals, it has considered appraisals only in a snapshot analysis. However, because appraisal is a complex, sequential process, we argue that it should be analyzed based on how it unfolds throughout a narrative. In this study, we investigate whether trajectories of appraisals are distinctive for different emotions in five-event stories – narratives where each of five sentences describes an event. We employ zero-shot prompting with a large language model to predict appraisals on sub-sequences of a narrative. We find that this approach is effective in identifying relevant appraisals in narratives, without prior knowledge of the evoked emotion, enabling a comprehensive analysis of appraisal trajectories. Furthermore, we are the first to quantitatively identify typical patterns of appraisal trajectories that distinguish emotions. For example, a rising trajectory for self-responsibility indicates trust, while a falling trajectory suggests anger.

1 Introduction

Emotion analysis seeks to uncover the emotions that events evoke in individuals based on their textual descriptions. Foundational frameworks established by Ekman (1972) and Plutchik (2001) categorize emotions in a structured manner. This allows researchers to label textual data from various sources, for example social media as in the corpus by Mohammad (2012). Consider the textual examples given[1] in Table 1, listing five sentences in narrative order. We may suspect that the event described in sentence #1 evoked the emotion trust in the experiencer and that the event described in sentence #5 probaly evoked the emotion guilt in them (based on the given context).

Recent work by Troiano et al. (2023) has highlighted the importance of implicit emotion cues tied to appraisal theories. These theories present emotions as resulting from the cognitive evaluation of events, incorporating subjective feelings, action tendencies, physiological responses, and both vocal and bodily expressions triggered by those events (Staller and Petta, 2001; Scherer, 2005; Gratch et al., 2009). By identifying the appraisals, we can more clearly reconstruct how the emotions are shaped in the example (see Table 1): The event described in sentence #1 is likely something pleasant for the experiencer and they are taking on responsibility in connection with a relevant goal. These cognitive appraisals (pleasantness, self responsibility and goal relevance) result in the emotion trust.

Prior work mainly analyzes appraisal in isolated instances, such as single sentences, discrete events, or social media posts (e.g., Hofmann et al., 2020; Stranisci et al., 2022; Zhan et al., 2023; Troiano et al., 2023; Liu et al., 2025; Zhou et al., 2025). Notably, many of these studies build upon the work of Troiano et al. (2023) and extend their focus to longer texts, underscoring the need for a contextualized analysis.

#	Sentence Text
1	A close friend entrusted me with setting up the sound system for a charity event.
2	I faced a daunting task, as the previous technician had left the equipment in disarray.
3	With time running out, I took a shortcut to meet the deadline, skipping some crucial safety checks.
4	Just before the event started, my friend reminded me of its significance and the many people counting on its success.
5	The loudspeaker suddenly malfunctioned and went silent.

Table 1: Example narrative comprising five sentences.

[1] We provide access to the data and code of our experiments on https://www.uni-bamberg.de/en/nlproc/resources/appraisal-trajectories/.

Appraisals and the resulting emotions often unfold as a complex process (Lewis, 2001) that can span multiple events and develop over the course of a narrative. In dialog research, emotion dynamics have been explored, e.g., by Poria et al. (2019). Debnath et al. (2025) also consider how appraisals influence these dynamics in multi-turn conversations. However, our work is the first aiming to understand how the appraisal progression of an individual develops based on narrative event sequences. To explain the emotion guilt for sentence #5 (shown in Table 1) using appraisals, we need to consider how these develop over the course of the narrative event sequence: The individual's responsibility for the outcome becomes increasingly apparent throughout the narrative. However, the pleasantness of the events, which is positive in the beginning, shifts to negative by the end due to the event described in the final sentence. Consequently, the negative emotion guilt is suddenly triggered. We refer to these changes in appraisals over the course of an event sequence as appraisal trajectories. Prior research only conducts snapshot analyses of appraisals, neglecting their development over time. While Wemmer et al. (2024) created appraisal trajectories in a study of dreams, their analysis does not examine the connection to the evoked emotions.

In our study, we explore the dynamics of appraisals that evoke different emotions within narratives through the following research questions:

1. Can we identify appraisals from narrative text alone, or is prior knowledge of the evoked emotion necessary?
2. Do appraisal trajectories in narratives show distinct patterns for different emotions?

To this end, we conduct automatic appraisal annotation on short narratives consisting of sequences of event descriptions evoking different emotions. Our methodology utilizes zero-shot prompting with a large language model (LLM) to predict appraisals. We analyze the trajectories of different appraisal dimensions in these narratives and compare their role in evoking different emotions. We aim to identify typical patterns of appraisal trajectories associated with specific emotions, to understand how emotion responses can be modeled more precisely.

2 Methods

Data and Model. For our experiments, we require a dataset with a substantial number of narratives categorized for different emotions. Given a person's textual description of their experience, we aim to reconstruct how they appraise events resulting in that emotion. The `Emotional Backstories` (EBS) dataset (Schäfer and Klinger, 2025) therefore offers itself as it comprises 13,000 narratives, each structured as a sequence of five event descriptions and assigned to one of 13 emotion categories (see the first column of Table 2). The dataset consists of LLM-generated data, with its quality validated by human annotators, and is designed for a thorough investigation of how contexts influence emotion analysis. For modeling appraisals, we adopt the 21 categories used by Troiano et al. (2023) (a detailed list is given in Appendix A). Our analysis uses the instruction-tuned Llama-3.3 LLM with 70B parameters (Meta, 2025). We utilize a zero-shot prompting method due to the lack of annotated training data for contextualized appraisals. The full text prompts used are given in Appendix B.

Appraisal Detection Performance. Our goal is to compute and analyze appraisals, requiring an effective labeling approach. Our method prompts the LLM to predict Likert scale scores (ranging from 1 to 5) for all appraisal categories given a narrative. We estimate the contextualized appraisal prediction performance of this approach by evaluating it on the crowd-enVent dataset (Troiano et al., 2023), which only contains isolated instances.

Appraisal Correlations with Emotions. We also need to evaluate the effectiveness of our appraisal prediction method on narratives, specifically assessing whether it performs successfully without prior knowledge of the associated emotion, relying solely on the text. We test this by allowing for the inclusion or exclusion of emotion information in the prompt. Troiano et al. (2023) show typical correlations of certain appraisals with certain emotions. We investigate whether we can find similar correlations in narratives and compare two settings: 1) Bottom-Up: the model should reconstruct the appraisals of the experiencer based on the textual sequence of event descriptions. 2) Top-Down: the model is informed about the overall emotion being evoked and is basically tasked to find related appraisals in the narrative. We hypothesize that the second approach will reveal clearer correlations due to the model's inherent knowledge of appraisals, allowing it to predict these not only from the textual content. However, the first setting presents a more realistic scenario, as we usually

Emotion \ Appraisal	suddenness	familiarity	predict_event	pleasantness	unpleasantness	goal_relevance	chance_responsblt	self_responsblt	other_responsblt	predict_conseq	goal_support	urgency	self_control	other_control	chance_control	accept_conseq	standards	social_norms	attention	not_consider	effort
	Delta Scores																				
Anger	0.2	−0.2	−0.2	−0.5	0.7	0.1	−0.1	−0.3	0.2	−0.1	−0.6	0.2	−0.2	0.2	0.0	−0.6	1.0	0.7	0.1	0.0	0.3
Boredom	−0.2	0.1	0.1	−0.5	−0.1	−0.5	0.0	−0.1	−0.1	−0.1	−0.4	−0.3	−0.1	0.0	0.0	−0.1	−0.1	0.0	−0.7	0.8	−0.4
Disgust	0.1	−0.2	−0.1	−0.4	0.7	0.0	0.0	−0.1	0.1	0.0	−0.4	0.2	−0.1	0.0	0.0	−0.4	0.6	0.5	0.1	0.5	0.3
Fear	0.1	−0.2	−0.1	−0.3	0.4	0.2	0.1	0.0	0.0	0.0	−0.3	0.3	−0.1	0.1	0.3	−0.4	0.4	0.2	0.1	0.1	0.2
Guilt	0.1	−0.1	−0.1	−0.5	0.6	0.3	−0.1	0.5	−0.4	0.0	−0.6	0.2	−0.1	−0.1	0.0	−0.6	1.1	0.5	0.1	0.4	0.3
Joy	−0.1	0.0	0.0	0.8	−0.6	0.0	0.0	0.1	0.0	0.1	0.6	−0.1	0.1	0.0	−0.1	0.4	−0.2	0.0	−0.1	0.0	−0.2
Pride	−0.1	0.0	0.0	0.6	−0.4	0.2	−0.1	0.3	−0.2	0.2	0.6	0.0	0.3	−0.1	−0.1	0.5	−0.1	0.0	0.0	0.0	0.0
Relief	0.0	0.0	0.0	0.6	−0.5	0.1	0.0	0.1	−0.1	0.1	0.7	0.0	0.0	0.0	0.1	0.9	−0.1	0.0	0.0	0.0	−0.2
Sadness	0.1	−0.1	−0.1	−0.7	0.7	0.1	0.1	−0.1	−0.1	0.0	−0.6	0.1	−0.2	0.1	0.2	−0.5	0.4	0.1	0.1	0.3	0.3
Shame	0.1	−0.2	−0.2	−0.6	0.7	0.1	0.0	0.3	−0.3	−0.1	−0.8	0.1	−0.4	0.1	0.0	−0.7	1.4	0.7	0.1	0.8	0.3
Surprise	0.8	−0.5	−0.4	−0.1	0.0	0.1	0.2	−0.1	0.0	−0.2	−0.1	0.3	−0.1	0.1	0.3	−0.1	0.2	0.1	0.2	0.0	0.1
Trust	−0.1	0.1	0.0	0.2	−0.2	0.1	−0.1	0.0	0.0	0.1	0.3	0.0	0.1	0.0	0.0	0.2	−0.1	0.0	0.0	0.0	−0.1
No-Emotion	−0.2	0.1	0.1	−0.1	−0.2	−0.3	0.0	−0.1	−0.1	0.0	−0.2	−0.1	0.0	−0.1	−0.1	0.0	−0.2	−0.1	−0.2	−0.1	−0.2

Table 2: Difference in appraisal scores of complete narratives per emotion category of model informed about the emotion to model not informed about the emotion (absolute scores for the two methods are given in Appendix D). Positive (negative) values, marked with green (purple) background color, express that the informed model assigns a higher (lower) Likert score (range 1–5). The different appraisal categories are further specified in Appendix A.

do not have prior knowledge of the emotion that is evoked. In comparing the predictions from the two approaches, we aim to explore whether the typical correlations can still emerge in the case where the model is solely dependent on the text, without explicit information about the emotion.

Analysis of Appraisal Trajectories. To be able to analyze the trajectories of appraisals within narratives, we generate sub-sequences from the five-sentence text instances. We start with only the first sentence and then incrementally add sentences in narrative order, employing the technique by Wemmer et al. (2024) who study dreams. By including the previous context rather than just a sentence, we aim for a more accurate contextualized analysis. For each sub-sequence, we predict appraisal scores using the model established above where no information about the emotion is included.

3 Results

The evaluation on the human-annotated crowd-enVent dataset (see Appendix C) shows that the appraisal prediction performance of our zero-shot model is acceptable (RMSE 1.49/1.46) in comparison to the fine-tuned model by Troiano et al. (2023) (RMSE 1.40). We analyze appraisals in narratives by applying our model to the `Emotional Backstories` dataset (Schäfer and Klinger, 2025), which contains 1,000 five-sentence-narratives for each of 13 emotion categories. In the following, we address our research questions by evaluating averages over the instances per emotion category as well as overall macro-averaged scores.

3.1 Can we identify appraisals from narrative text alone, or is prior knowledge of the evoked emotion necessary?

Table 2 shows how predicted appraisal scores change when the model is informed about the evoked emotion. The results reveal only minor variations in predicted appraisal scores. For instance, the pleasantness appraisal score (see the fifth column in Table 2) slightly increases (+0.2 to +0.8 Likert score points) for positive emotions while it slightly decreases (−0.3 to −0.7 Likert scale points) for negative emotions when the model is informed about the emotion. Differences in Likert scale values of more than ± 1 are observed only for the appraisal regarding standards in negative emotions (+1.0 for anger, +1.1 for guilt and +1.4 for shame; see Table 2).

Separate values for the two settings further substantiate the trend that informing the model about the emotion is not necessary to find typical correlations of appraisals and emotions (see Appendix D). This shows that our approach is able to find indicators of appraisals in the textual content of the narratives even without prior knowledge of the emotion.

3.2 Do appraisal trajectories in narratives show distinct patterns for different emotions?

Table 3 shows average appraisal scores of narrative sub-sequences, analyzed as trajectories for each emotion. From these values, we can infer how certain appraisals are expressed through the narrative structure. Generally, differences in magnitude are observed across certain appraisals. For example, relevance to the experiencer's goals is consistently scored higher than chance occurrences, aligning with expectations for experiencer-perspective text. Other appraisal categories show differences in intensity for different emotions, for example, contrasting familiarity scores; sadness exhibits higher scores compared to fear – which is consistent with established emotion-appraisal correlations.

Regarding trajectory patterns, we identify four distinct types in Table 3: Rising, Falling, Valley-shaped, Hill-shaped. These shapes arise from the variations in appraisal processes as they unfold within narrative structures. Certain appraisals display uniform patterns across all emotions. An example for this is suddenness: all Rising. Here, context is required for higher scores. In contrast, certain appraisals interestingly show variation in trajectory patterns by emotion. For example, pleasantness shows a Rising pattern for joy but a Valley-shaped pattern for relief, illustrating the complex process that involves less pleasant events before a resolution. Another example is self-responsibility, which exhibits a Falling trajectory for anger, while showing a Rising pattern for trust, yet both ultimately reach approximately the same score. This emphasizes that performing a snapshot analysis of appraisals is incomplete, as only the distinct trajectory patterns reveal deeper processes that distinguish the evokation of the different emotions.

4 Conclusion

Our study reveals distinct trajectories in appraisal processes that result in evoking different emotions in narratives. We demonstrate that contextualized appraisals can be effectively modeled based on narratives even without prior knowledge of the underlying emotion, suggesting that the textual content contains sufficient cues. Our identification of unique appraisal patterns – Rising, Falling, Valley-shaped, and Hill-shaped – highlights the complexity of the appraisal process leading to different emotion responses, emphasizing the need for a dynamic perspective on appraisal. These findings carry further implications for emotion detection applications, as recognizing appraisal trajectories could enhance the accuracy of emotion classification systems.

The fixed five-sentence narrative structure we study facilitates a controlled analysis. Our findings generalize to narratives of varying lengths, provided that an additional step is taken to adapt such data into the format we use. Another option would be to align and normalize trajectories for variable-length narratives, which would further broaden their applicability, although this remains a topic for future research. Additionally, future work should study the interplay between different combinations of appraisals and how they relate to specific emotion categories. Future research is required to explore how to best leverage these trajectories to optimize emotion detection in various contexts, ultimately deepening our understanding of the interplay between appraisals in narrative structure and evoked emotions.

Limitations

This study has some limitations worth noting. The zero-shot performance of the LLM for predicting appraisals can be nuanced; while it demonstrates potential, there seems to be considerable room for improvement. Moreover, generating Likert-scale numerical outputs with the LLM might not an optimized approach for scoring appraisals, as it limits calibration and comparability. Analyzing output distributions and sensitivity across prompt variants or repeated runs would strengthen the robustness of our findings, as we only ran our experiments once. We also utilize only one LLM for our analysis, which may restrict the generalizability of our findings. Different models may yield varying results in predicting appraisals. Additionally, our analysis relies on data that has been automatically generated, which can include biases or inaccuracies inherent in the dataset. Human validation could, for example, strengthen the reliability of our results. To facilitate reproducibility, we provide access to our code and model predictions, enabling further exploration and validation of our findings.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) in the project KL2869/1–2 (CEAT, project number 380093645).

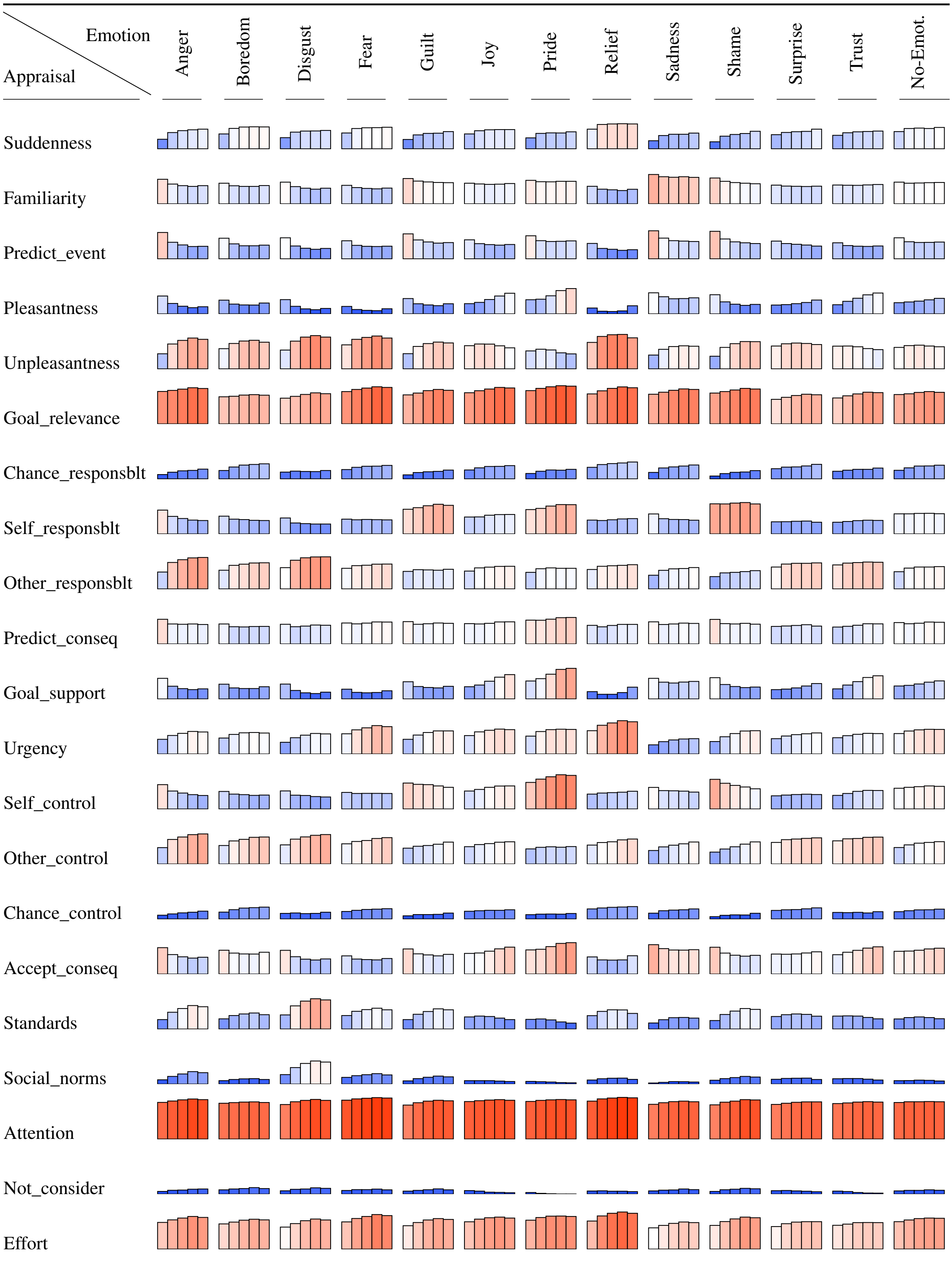

Table 3: Averaged predicted appraisal trajectories in narratives per emotion categories in the EBS dataset. Each cell presents five appraisal values computed from sub-sequences of data instances (five sentences each), reflecting the average over 1000 instances. The bar color and height both represent the average Likert scale score from the range of 1 to 5, with blue indicating the lowest (1), white the midpoint (3), and red the highest (5) values.

References

Alok Debnath, Yvette Graham, and Owen Conlan. 2025. An appraisal theoretic approach to modelling affect flow in conversation corpora. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 233–250, Vienna, Austria. Association for Computational Linguistics.

Paul Ekman. 1972. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation 1971*, volume 19. Lincoln University of Nebraska Press.

Jonathan Gratch, Stacy Marsella, Ning Wang, and Brooke Stankovic. 2009. Assessing the validity of appraisal-based models of emotion. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8.

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

Marc D Lewis. 2001. Personal pathways in the development of appraisal. In Klaus R Scherer, Angela Schorr, and Tom Johnstone, editors, *Appraisal processes in emotion: Theory, methods, research*, pages 205–220. Oxford University Press: London.

June M. Liu, He Cao, Renliang Sun, Rui Wang, Yu Li, and Jiaxing Zhang. 2025. CAPE: A Chinese dataset for appraisal-based emotional generation in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6291–6309, Albuquerque, New Mexico. Association for Computational Linguistics.

Meta. 2025. Llama (model llama-3.3-70b-instruct). Large language model.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Robert Plutchik. 2001. The nature of emotions. *American Scientist*, 89(4):344–350.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Klaus R Scherer. 2005. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729.

Johannes Schäfer and Roman Klinger. 2025. Shaping event backstories to estimate potential emotion contexts. *Preprint*, arXiv:2508.09954.

Alexander Staller and Paolo Petta. 2001. Introducing emotions into the computational study of social norms: A first evaluation. *Journal of Artificial Societies and Social Simulation*, 4(1).

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

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

Eileen Wemmer, Sofie Labat, and Roman Klinger. 2024. EmoProgress: Cumulated emotion progression analysis in dreams and customer service dialogues. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5660–5677, Torino, Italia. ELRA and ICCL.

Hongli Zhan, Desmond C. Ong, and Junyi Jessy Li. 2023. Evaluating subjective cognitive appraisals of emotions from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14418–14446, Singapore. Association for Computational Linguistics.

Yuxiang Zhou, Hainiu Xu, Desmond C. Ong, Maria Liakata, Petr Slovak, and Yulan He. 2025. Modeling subjectivity in cognitive appraisal with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13811–13833, Suzhou, China. Association for Computational Linguistics.

A Appraisal Categories

Table 4 lists the 21 appraisal categories and defining statements used in our experiments.

B Prompts

Table 5 shows the full text of the prompts we use for appraisal analysis (two variants based on whether we want to inform the model about the emotion or not). The interaction with the model uses two message types[2]: a "system" message that establishes the context for the interaction and includes general guidelines; a "user" message that encapsulates the

[2] https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/

Abbreviation	Appraisal Statement
suddenness	The event was sudden or abrupt.
familiarity	The event was familiar to its experiencer.
predict_event	The experiencer could have predicted the occurrence of the event.
pleasantness	The event was pleasant for the experiencer.
unpleasantness	The event was unpleasant for the experiencer.
goal_relevance	The experiencer expected the event to have important consequences for him/herself.
chance_responsblt	The event was caused by chance, special circumstances, or natural forces.
self_responsblt	The event was caused by the experiencer's own behavior.
other_responsblt	The event was caused by somebody else's behavior.
predict_conseq	The experiencer anticipated the consequences of the event.
goal_support	The experiencer expected positive consequences for her/himself.
urgency	The event required an immediate response.
self_control	The experiencer was able to influence what was going on during the event.
other_control	Someone other than the experiencer was influencing what was going on.
chance_control	The situation was the result of outside influences of which nobody had control.
accept_conseq	The experiencer anticipated that he/she could live with the unavoidable consequences of the event.
standards	The event clashed with her/his standards and ideals.
social_norms	The actions that produced the event violated laws or socially accepted norms.
attention	The experiencer had to pay attention to the situation.
not_consider	The experiencer wanted to shut the situation out of her/his mind.
effort	The situation required her/him a great deal of energy to deal with it.

Table 4: Appraisal categories used in our experiments.

specific inputs, requirements, and instructions for the task.

C Appraisal Prediction Performance on Isolated Instances

We evaluate the appraisal prediction performance of our zero-shot model on the crowd-enVent test data (Troiano et al., 2023). Results are shown in Table 7 (model not informed about emotion) and Table 6 (model informed about emotion). The predicted Likert scale points are evaluated with RMSE (Root Mean Square Error) against the gold annotations. Additionally, a classification prediction is simulated by converting predictions of 1/2/3 to a negative label, and 4/5 to a positive label (method used by Troiano et al., 2023). The result is then evaluated with Precision (P), Recall (R) and F1-score (F1) against the gold annotations.

D Predicted Appraisals on Narratives by Emotion Category

We predict appraisal categories as Likert scale scores (range 1–5) for each of the narratives from the `Emotional Backstories` dataset (Schäfer and Klinger, 2025) and compute average scores for each set of narratives per emotion category. Table 8 displays the results for the scores predicted by the model which has not been informed about the emotion. Table 9 displays the results for the scores predicted by the model which has been informed about the emotion.

MT	Prompt Text
system	You are an expert in the appraisal on event descriptions.
user	A person describes their event experience as follows: {text_instance}. Based on this text, you’re asked to assess the content in terms of the appraisals it conveys. Thus, your task is to identify appraisals in texts. You will receive a list of appraisals and theire description. For each of the following 21 appraisal assessments listed provide a score in the range of 1 (not at all) to 5 (extremely). 1. The event was sudden or abrupt (suddenness) 2. The event was familiar to its experiencer. (familiarity) 3. The experiencer could have predicted the occurrence of the event. (predict_event) 4. The event was pleasant for the experiencer. (pleasantness) 5. The event was unpleasant for the experiencer. (unpleasantness) 6. The experiencer expected the event to have important consequences for him/herself. (goal_relevance) 7. The event was caused by chance, special circumstances, or natural forces. (chance_responsblt) 8. The event was caused by the experiencer’s own behavior. (self_responsblt) 9. The event was caused by somebody else’s behavior. (other_responsblt) 10. The experiencer anticipated the consequences of the event. (predict_conseq) 11. The experiencer expected positive consequences for her/himself. (goal_support) 12. The event required an immediate response. (urgency) 13. The experiencer was able to influence what was going on during the event. (self_control) 14. Someone other than the experiencer was influencing what was going on. (other_control) 15. The situation was the result of outside influences of which nobody had control. (chance_control) 16. The experiencer anticipated that he/she could live with the unavoidable consequences of the event. (accept_conseq) 17. The event clashed with her/his standards and ideals. (standards) 18. The actions that produced the event violated laws or socially accepted norms. (social_norms) 19. The experiencer had to pay attention to the situation. (attention) 20. The experiencer wanted to shut the situation out of her/his mind. (not_consider) 21. The situation required her/him a great deal of energy to deal with it. (effort) As your response, provide only a json object and keep the appraisal labels as they are (eg. "other_responsblt"). In addition to the given text, take into account the emotion: {emotion}. To what level were the appraisals evoked in the person at the end? Take into account both the text and the emotion. In your answer, only provide the score labels you have chosen as the output together with the input text and emotion. Provide the output as a json object with the top-level key being ’appraisal’ and the low-level keys being the concrete appraisals. The value must be the Likert-score. For example, given a text instance, your json object output should be: { "chain": "{text_instance}", "emotion": "null{emotion}", "appraisals": {[...]} } where [...] is a placeholder and you insert the 21 appraisals here. The chain is: {text_instance}. What are the appraisal scores for this chain? Only output the json object.

Table 5: Prompts used for zero-shot appraisal analysis. “MT” refers to the prompt message type as specified in the input for the instruction-tuned LLM. “{text_instance}” and “{emotion}” are variables. The text marked in red is only inserted when the model should not be informed about the emotion. The text marked in blue is only inserted if the model should be informed about the emotion.

	Ours	Troiano et al. (2023)	Δ	Ours			Troiano et al. (2023)			Δ
Appraisal	RMSE	RMSE	RMSE	P	R	F1	P	R	F1	F1
Suddenness	1.44	1.33	-.11	.72	.65	.68	.70	.79	.74	.06
Familiarity	1.54	1.42	-.12	.51	.64	.57	.77	.82	.79	.22
Predict_event	1.45	1.47	.03	.59	.55	.57	.76	.74	.75	.18
Pleasantness	1.42	1.30	-.12	.79	.89	.84	.88	.87	.88	.04
Unpleasantness	1.18	1.26	.08	.87	.85	.86	.79	.80	.80	-.06
Goal_relevance	1.62	1.57	-.05	.53	.91	.67	.73	.69	.71	.04
Chance_responsblt	1.55	1.43	-.12	.61	.37	.46	.83	.87	.85	.39
Self_responsblt	1.37	1.40	.03	.72	.69	.70	.81	.77	.79	.09
Other_responsblt	1.50	1.57	.07	.80	.66	.72	.74	.72	.73	.01
Predict_conseq	1.46	1.50	.04	.59	.44	.50	.67	.71	.69	.19
Goal_support	1.50	1.33	-.17	.68	.71	.69	.80	.82	.81	.12
Urgency	1.77	1.43	-.34	.77	.30	.44	.63	.60	.61	.17
Self_control	1.46	1.35	-.11	.59	.56	.57	.78	.81	.79	.22
Other_control	1.63	1.36	-.27	.79	.58	.67	.64	.60	.62	-.05
Chance_control	1.55	1.35	-.20	.59	.29	.39	.84	.90	.87	.48
Accept_conseq	1.77	1.36	-.41	.52	.60	.55	.63	.65	.64	.09
Standards	1.30	1.34	.04	.69	.65	.67	.82	.83	.82	.15
Social_norms	1.13	1.44	.31	.59	.58	.58	.90	.95	.92	.34
Attention	1.43	1.27	-.16	.68	.87	.77	.50	.48	.48	-.29
Not_consider	1.52	1.53	.01	.60	.26	.37	.83	.71	.77	.40
Effort	1.37	1.38	.01	.67	.66	.66	.69	.70	.70	.04
Macro average	1.46	1.40	-.06	.66	.60	.62	.75	.75	.75	.13

Table 6: Appraisal prediction performance of our zero-shot model (setting with mention of emotion in prompt) on the crowd-enVent test data in comparison to the fine-tuned model by Troiano et al. (2023).

	Ours	Troiano et al. (2023)	Δ	Ours			Troiano et al. (2023)			Δ
Appraisal	RMSE	RMSE	RMSE	P	R	F1	P	R	F1	F1
Suddenness	1.55	1.33	-.22	.76	.60	.67	.70	.79	.74	.07
Familiarity	1.61	1.42	-.19	.49	.65	.56	.77	.82	.79	.23
Predict_event	1.50	1.47	-.03	.58	.59	.59	.76	.74	.75	.16
Pleasantness	1.14	1.30	.16	.78	.86	.82	.88	.87	.88	.06
Unpleasantness	1.26	1.26	-.00	.87	.81	.84	.79	.80	.80	-.04
Goal_relevance	1.58	1.57	-.01	.55	.87	.68	.73	.69	.71	.04
Chance_responsblt	1.59	1.43	-.16	.65	.44	.52	.83	.87	.85	.33
Self_responsblt	1.41	1.40	-.01	.72	.68	.70	.81	.77	.79	.09
Other_responsblt	1.55	1.57	.02	.79	.68	.73	.74	.72	.73	-.00
Predict_conseq	1.54	1.50	-.04	.57	.41	.48	.67	.71	.69	.21
Goal_support	1.47	1.33	-.14	.68	.68	.68	.80	.82	.81	.13
Urgency	1.88	1.43	-.45	.79	.28	.41	.63	.60	.61	.20
Self_control	1.50	1.35	-.15	.58	.58	.58	.78	.81	.79	.21
Other_control	1.68	1.36	-.32	.78	.57	.66	.64	.60	.62	-.04
Chance_control	1.56	1.35	-.21	.62	.35	.44	.84	.90	.87	.43
Accept_conseq	1.76	1.36	-.40	.52	.60	.56	.63	.65	.64	.08
Standards	1.30	1.34	.04	.74	.54	.62	.82	.83	.82	.20
Social_norms	1.11	1.44	.33	.62	.52	.56	.90	.95	.92	.36
Attention	1.43	1.27	-.16	.68	.87	.76	.50	.48	.48	-.28
Not_consider	1.57	1.53	-.04	.75	.18	.29	.83	.71	.77	.48
Effort	1.39	1.38	-.01	.69	.61	.65	.69	.70	.70	.06
Macro average	1.49	1.40	-.09	.68	.59	.61	.75	.75	.75	.14

Table 7: Appraisal prediction performance of our zero-shot model (setting without mention of emotion in prompt) on the crowd-enVent test data in comparison to the fine-tuned model by Troiano et al. (2023).

Appraisal / Emotion	Suddenness	Familiarity	Predict_event	Pleasantness	Unpleasantness	Goal_relevance	Chance_responsblt	Self_responsblt	Other_responsblt	Predict_conseq	Goal_support	Urgency	Self_control	Other_control	Chance_control	Accept_conseq	Standards	Social_norms	Attention	Not_consider	Effort
Anger	2.9	2.7	2.2	1.7	3.8	4.3	1.9	2.3	4.0	2.8	1.9	3.1	2.3	3.8	1.7	2.6	3.1	2.1	4.7	1.5	4.0
Boredom	3.1	2.7	2.3	2.0	3.5	3.7	2.4	2.3	3.5	2.6	2.2	3.0	2.3	3.5	2.1	3.0	2.3	1.4	4.4	1.5	3.7
Disgust	2.7	2.5	2.0	1.5	4.0	3.8	1.8	1.9	4.0	2.8	1.6	2.9	2.1	3.8	1.6	2.4	3.7	3.1	4.6	1.5	3.7
Fear	3.0	2.5	2.2	1.5	3.9	4.4	2.3	2.3	3.3	3.1	1.8	3.6	2.5	3.5	2.0	2.5	2.8	1.9	4.9	1.4	4.2
Guilt	2.6	3.0	2.5	1.9	3.4	4.2	1.8	3.7	2.8	2.9	2.2	3.2	3.0	3.1	1.6	2.8	2.8	1.7	4.6	1.4	3.8
Joy	2.8	2.9	2.4	2.9	3.0	4.4	2.2	2.8	3.1	3.1	3.3	3.3	3.2	3.1	1.9	3.5	1.9	1.2	4.7	1.1	3.9
Pride	2.6	3.1	2.7	3.4	2.4	4.5	1.9	3.7	2.9	3.5	3.9	3.3	4.2	2.6	1.5	4.0	1.6	1.1	4.7	1.0	4.1
Relief	3.3	2.4	1.9	1.8	3.9	4.4	2.6	2.4	3.3	2.8	2.1	4.0	2.7	3.4	2.2	2.7	2.5	1.4	4.9	1.2	4.4
Sadness	2.5	3.5	2.6	2.5	3.1	4.2	2.3	2.3	3.0	2.9	2.7	2.4	2.6	3.1	1.9	3.3	2.0	1.2	4.5	1.4	3.5
Shame	2.7	2.9	2.4	1.9	3.6	4.3	1.8	3.8	2.7	2.9	2.1	3.2	2.9	3.1	1.5	2.8	2.9	1.7	4.7	1.5	3.9
Surprise	2.9	2.7	2.2	2.3	3.3	3.7	2.4	2.1	3.5	2.7	2.4	3.0	2.4	3.5	2.1	3.1	2.2	1.4	4.5	1.2	3.6
Trust	2.7	2.8	2.2	3.0	2.8	3.9	2.0	2.3	3.5	2.9	3.2	2.9	2.7	3.5	1.7	3.6	2.0	1.4	4.5	1.1	3.5
No-Emotion	3.0	3.0	2.6	2.5	3.1	4.0	2.3	2.9	3.1	3.0	2.7	3.3	3.1	3.1	1.9	3.4	2.0	1.3	4.5	1.4	3.9

Table 8: Appraisal scores of narratives per emotion category of model **not** informed about the emotion. Positive (negative) values are marked with red (blue) background color.

Appraisal / Emotion	Suddenness	Familiarity	Predict_event	Pleasantness	Unpleasantness	Goal_relevance	Chance_responsblt	Self_responsblt	Other_responsblt	Predict_conseq	Goal_support	Urgency	Self_control	Other_control	Chance_control	Accept_conseq	Standards	Social_norms	Attention	Not_consider	Effort
Anger	3.0	2.5	2.0	1.2	4.5	4.4	1.8	2.0	4.2	2.8	1.4	3.3	2.1	4.0	1.7	2.0	4.1	2.7	4.9	1.5	4.3
Boredom	2.8	2.8	2.4	1.5	3.4	3.2	2.5	2.2	3.4	2.6	1.7	2.7	2.2	3.5	2.2	2.9	2.2	1.4	3.8	2.3	3.3
Disgust	2.8	2.3	1.9	1.1	4.6	3.9	1.8	1.8	4.1	2.8	1.2	3.1	2.0	3.8	1.7	2.0	4.4	3.5	4.8	2.0	4.0
Fear	3.2	2.3	2.1	1.2	4.3	4.6	2.4	2.3	3.3	3.0	1.4	3.9	2.3	3.5	2.2	2.1	3.2	2.1	4.9	1.5	4.4
Guilt	2.7	2.9	2.4	1.4	4.0	4.4	1.7	4.2	2.5	2.9	1.6	3.3	2.9	3.0	1.5	2.3	3.9	2.2	4.7	1.8	4.0
Joy	2.8	2.9	2.4	3.7	2.3	4.4	2.2	2.9	3.1	3.2	3.9	3.2	3.3	3.0	1.8	4.0	1.7	1.2	4.6	1.1	3.7
Pride	2.5	3.1	2.7	4.0	2.0	4.7	1.8	4.1	2.8	3.7	4.5	3.3	4.4	2.5	1.4	4.4	1.4	1.1	4.6	1.0	4.1
Relief	3.3	2.4	1.9	2.4	3.4	4.5	2.6	2.5	3.2	2.9	2.8	4.1	2.7	3.3	2.2	3.6	2.4	1.4	4.9	1.3	4.2
Sadness	2.5	3.4	2.6	1.8	3.9	4.4	2.4	2.2	2.9	2.9	2.0	2.5	2.4	3.1	2.2	2.8	2.4	1.2	4.6	1.7	3.7
Shame	2.7	2.7	2.3	1.3	4.3	4.4	1.7	4.1	2.4	2.7	1.3	3.3	2.5	3.2	1.6	2.1	4.2	2.4	4.8	2.3	4.2
Surprise	3.7	2.2	1.8	2.2	3.3	3.8	2.6	2.0	3.4	2.4	2.4	3.3	2.2	3.5	2.3	3.0	2.4	1.5	4.7	1.2	3.7
Trust	2.6	2.9	2.2	3.2	2.6	4.0	1.9	2.3	3.5	3.0	3.5	2.9	2.9	3.5	1.6	3.8	1.9	1.4	4.4	1.1	3.4
No-Emotion	2.9	3.1	2.7	2.3	2.9	3.7	2.2	2.8	3.0	3.0	2.5	3.2	3.1	3.0	1.9	3.4	1.8	1.2	4.3	1.3	3.7

Table 9: Appraisal scores of narratives per emotion category of model informed about the emotion. Positive (negative) values are marked with red (blue) background color.