Categorical Emotions or Appraisals – Which Emotion Model Explains Argument Convincingness Better?

Lynn Greschner, Meike Bauer, Sabine Weber, Roman Klinger
Fundamentals of Natural Language Processing, University of Bamberg, Germany
{firstname.lastname}@uni-bamberg.de

Abstract

The convincingness of an argument does not only depend on its structure (logos), the person who makes the argument (ethos), but also on the emotion that it causes in the recipient (pathos). While the overall intensity and categorical values of emotions in arguments have received considerable attention in the research community, we argue that the emotion an argument evokes in a recipient is subjective. It depends on the recipient's goals, standards, prior knowledge, and stance. Appraisal theories lend themselves as a link between the subjective cognitive assessment of events and emotions. They have been used in event-centric emotion analysis, but their suitability for assessing argument convincingness remains unexplored. In this paper, we evaluate whether appraisal theories are suitable for emotion analysis in arguments by considering subjective cognitive evaluations of the importance and impact of an argument on its receiver. Based on the annotations in the recently published ContArgA corpus, we perform zero-shot prompting experiments to evaluate the importance of gold-annotated and predicted emotions and appraisals for the assessment of the subjective convincingness labels. We find that, while categorical emotion information does improve convincingness prediction, the improvement is more pronounced with appraisals. This work presents the first systematic comparison between emotion models for convincingness prediction, demonstrating the advantage of appraisals, providing insights for theoretical and practical applications in computational argumentation.

Keywords: emotions, appraisals, arguments, convincingness, implicit language, prompting

1. Introduction

The analysis of arguments and their quality, persusasiveness and convincingness received substantial attention (Lawrence and Reed, 2019). Argument quality assessment contains various subtasks, including quantifying the logical, rhetorical, and dialectical quality of the argument (Blair, 2011, cited after Wachsmuth et al. (2024)). An important aspect of dialectical quality is the convincingness of an argument, which we focus on in this paper. The convincingness of an argument is distinct from the overall effectiveness of argumentation because of its inherently subjective nature. Part of the subjective evaluation of an argument regarding its convincingness is the emotional appeal – changing the emotional state of a receiver such that they are more open to the argument. Most work handled *emotional appeal* as a continuous score or a binary variable (Wachsmuth et al., 2024). For instance, Chen and Eger (2025) show that manipulating the emotional appeal in a given argument changes its convincingness. Other studies focus on discrete emotions in arguments. Greschner and Klinger (2025) do however show that emotion recognition in arguments is a particularly challenging task, with low performance scores.

We hypothesize that this is because emotions in arguments develop in context of the argument recipient, including their demographic and psychological traits and states, their prior world knowledge and experiences, and stances towards topics. In general emotion analysis tasks, this subjective cognitive evaluation has been approached with the help of appraisal theories. Appraisal theories describe the cognitive evaluation of an event and the relationship of this evaluation to concrete emotion categories. Smith and Ellsworth (1985), for instance, show that six appraisal variables explain 15 discrete emotions, namely (1) how pleasant an event is (pleasantness, likely to be associated with joy, but unlikely to appear with disgust), (2) how much effort an event can be expected to cause (anticipated effort, likely to be high when anger or fear is experienced), (3) how certain the experiencer is in a specific situation (certainty, low, e.g., in the context of hope or surprise), (4) how much attention is devoted to the event (attention, likely to be low, e.g., in the case of boredom or disgust), (5) how much responsibility the experiencer of the emotion holds for what has happened (self-other responsibility/control, high for feeling challenge or pride), and (6) how much the experiencer has control over the situation (situational control, low in the case of anger). Scherer et al. (2001) points out the sequential nature of the cognitive evaluation of an event: A person first decides its relevance, its goal conduciveness, one's own ability to cope with outcomes, and the internal and external norms (regarding moral aspects and the legal situation).

Appraisal theories find application in natural language processing: Hofmann et al. (2020) annotate the appraisal evaluation that somebody experiences throughout an event. Troiano et al. (2023)

introduce a novel annotation framework, in which the potential noise is reduced by retrieving appraisal variables directly from the author of a text who lived through an event. Appraisals have also been used to study Reddit (Stranisci et al., 2022), to explain conspiracy theories (Pummerer et al., 2024), to understand the information processing in large language models (Zhan et al., 2023), to tailor emotion assignments to particular entities (Troiano et al., 2022), or to study emotion inference (Tak et al., 2025).

This diversity shows how versatile the approach is. Nevertheless, appraisals have not been used to study arguments yet, despite the fact that the emotion that somebody develops based on an argument does depend on a subjective cognitive evaluation. Using the ContArgA corpus (Greschner et al., 2026), which includes arguments annotated for emotions and appraisals, we explore whether appraisals can computationally explain perceived convincingness. We compare this setup to the use of categorical emotion models.

Our concrete research questions are:

- RQ1: Which emotion model helps LLMs to improve convincingness predictions?
- RQ2: Does jointly predicting emotions/appraisals and convincingness improve the performance compared to the single task predictions?

Our results demonstrate that both emotion models improve the convincingness prediction across three models, with appraisals providing stronger improvements than categorical emotions. The joint prediction consistently underperforms pipeline approaches.

2. Related Work

2.1. Emotion Analysis and Appraisals

Emotion analysis in NLP commonly builds on top of two main types of psychological emotion models: categorical and dimensional approaches. The most commonly used categorical model in NLP is Ekman's basic emotion model, which proposes that six discrete emotions are universal regarding stimulus events and reactions, namely anger, surprise, disgust, joy, fear, and sadness (Ekman, 1992). In contrast to such categorical models, dimensional models represent emotions along continuous axes in multidimensional spaces. One widely used example in NLP is the Circumplex Model of Affect (Posner et al., 2005), where emotions are evaluated in terms of valence and arousal.

Appraisal theories are approaches that received attention in NLP only more recently. They provide access to emotions through cognitive evaluations of events (Scherer et al., 2001). There are multiple frameworks of appraisal theories (Roseman, 1984; Roseman and Smith, 2001; Scherer, 2009), which propose varieties of appraisal variables. Therefore, also specialized frameworks have been proposed, including one for evaluating arguments (Greschner et al., 2026). Another theory has been developed particularly for the analysis of conspiracy theories (Pummerer et al., 2024). While there is no one set of appraisal variables, common variables encompass aspects of agency, pleasantness, consequences on the self, responsibility, expected effort, and novelty.

Categorical emotion models have received more attention in NLP than appraisal theories, but there is now also considerable work that employed appraisal theories. Examples include work for the study of coping strategies (Troiano et al., 2024), social media analysis (Stranisci et al., 2022), or emotion event self reports (Hofmann et al., 2020). Troiano et al. (2023) propose the largest set of appraisal variables for emotion analysis of events (Crowd-enVent). Using this corpus, Tak et al. (2025) probe models to understand whether they process emotions similarly to humans. Yeo and Jaidka (2025) train an appraisal predictor on the corpus and apply it to conversations, capturing changes in emotional states throughout them. Yeo and Jaidka (2025) use appraisal theories as the theoretical groundwork for a theory-of-mind framework-based dataset to assess inferred emotions (from context), comparing humans and LLMs, highlighting the need of psychological theories for evaluation LLMs on emotion (reasoning) tasks.

Recently, Greschner et al. (2026) expand and adapt the annotation scheme of the Crowd-enVent corpus (Troiano et al., 2023) to the analysis of arguments. However, the authors do not conduct a modeling study that would provide insights into how convincingness modeling performs under specific argument appraisal or emotional contexts. We fill this gap and make their novel corpus the data source for our investigation.

The convincingness of arguments is inherently subjective, depending on how receivers cognitively evaluate it. This suggests that appraisal theories, with their focus on subjective cognitive evaluation, may be particularly well-suited for computationally assessing argument convincingness – a hypothesis we explore in this work.

2.2. Arguments and Convincingness

Assessing the quality, persuasiveness, and convincingness of an argument are closely related tasks in the field of argument mining, which have received considerable attention (Lawrence and Reed, 2019; Habernal and Gurevych, 2016a,b; Quensel et al., 2025). Other subtasks of the field include detecting argument constituents (Teufel et al., 2009) and claims (Wührl and Klinger, 2021); fact-checking of claims (Thorne and Vlachos, 2018), reconstructing their structure (Li et al., 2022) and linking them (Ebner et al., 2020). A prominent task is assessing the quality of arguments (Wachsmuth et al., 2024), which includes quantifying the logical, rhetorical, and dialectical quality of the argument (Blair, 2011). Quensel et al. (2025) use regression analysis to investigate the subjective factors of emotions, storytelling, and hedging and their impact on argument strength, finding that the influence of the emotion depends on the rhetorical utilization with respect to the argument quality.

Most relevant to our work is the subtask of assessing an argument's convincingness. Different dimensions influence an argument's convincingness, where textual qualities play a role (Habernal and Gurevych, 2016a,b) as well as the personalities of the receivers (Lukin et al., 2017; Al Khatib et al., 2020). Especially important for argument convincingness are emotions – their importance has been demonstrated in the fields of computational argumentation (Habernal and Gurevych, 2016b,a; Wachsmuth et al., 2017; Greschner and Klinger, 2025), philosophy (Aristotle, 1991), and psychology (Bohner et al., 1992; Petty et al., 1993; Pfau et al., 2006; Worth and Mackie, 1987; Benlamine et al., 2015). In the field of natural language processing (NLP), prior work treats emotions in arguments as a binary variable, as one of many factors of convincingness (Habernal and Gurevych, 2016b), or rate the emotional appeal (Wachsmuth et al., 2017; Lukin et al., 2017) of the argument. Emotions in arguments are frequently treated as a fallacy (Jin et al., 2022; Ziegenbein et al., 2023), recently, a study employs LLMs to inject emotional appeals into fallacious arguments, finding that emotional framing reduces human fallacy detection and that fear, sadness, and enjoyment significantly increase perceived convincingness compared to neutral states (Chen et al., 2026).

The most relevant works for our study include Chen and Eger (2025), who examine emotion intensity's impact on convincingness through LLM-based manipulation, and Greschner and Klinger (2025), who focus on discrete emotion categories in German arguments. However, both studies treat emotions as categorical labels rather than exploring the underlying cognitive processes that generate these emotional responses. Our work differs fundamentally because we investigate whether appraisal theories – which model the subjective cognitive evaluation leading to emotions – provide sufficient explanations for convincingness compared to categorical emotion models.

Conf.	Prompt Part
Prefix	`You are an expert on annotating argumentative texts. You will have to solve the following tasks. Task: Convincingness Prediction: You will be given an argumentative text. Your task is to assign how convincing a person would find the argument`
Emo→	`given that they felt the emotion {emotion} after hearing the argument`
Appr→	`given that they assigned the following appraisals after hearing the argument: {appraisals}`
CVC	`on a 1-5 scale. Rating scale: 1 = Not at all convincing 2 = Slightly convincing 3 = Moderately convincing 4 = Very convincing 5 = Extremely convincing Argument: "argument" Respond with valid JSON containing only the numerical rating: {{"rating": [number from 1-5]}}`

Table 1: Prompts for the Pipeline configuration of appraisal/emotion conditioned convincingness prediction. The Emo and Appr sections are optionally added to the convincingness prediction as parameters.

3. Methods

The goal of our study is to understand if appraisals and/or emotion categories help to computationally assess an arguments convincingness. We do so with prompting experiments across a set of language models. It is important to keep in mind that we consider all three tasks (appraisal prediction, emotion prediction, and convincingness prediction) to be subjective.

3.1. Prompting Setup

We compare a set of prompting configurations, which we explain in the following. Since prior work highlighted the complexity of prompt formulations and considering that Greschner and Klinger (2025) demonstrate minimal impact of prompting methods (zero-shot, few-shot, chain-of-thought) on emotion predictions, we opt to focus exclusively on zero-shot prompts in our study. This approach ensures a straightfoward, plug-and-play implementation while avoiding confounding effects of complex prompting strategies.

Single Model. To create a baseline, we prompt the model to predict the convincingness of a given argument (we refer to this task as *CVC* prediction). Similarly, we use a plain zero-shot prompting setting to predict (1) the emotion categories and (2) the appraisal dimensions as a single task.

Conf.	Prompt Part
Prefix	`You are an expert on annotating argumentative texts. You will have to solve the following tasks.`
Emo↔CVC	`Task: Emotion Prediction. You will be given an argumentative text. Your task is to assign the strongest emotion that is evoked in a person hearing the argument. The emotion categories to choose from are: anger, disgust, fear, guilt, joy, pride, relief, sadness, shame, surprise, trust.` `Task 2: Convincingness Prediction. Your task is to assign how convincing a person would find the argument on a 1-5 scale. Rating scale: 1 = Not at all convincing 2 = Slightly convincing 3 = Moderately convincing 4 = Very convincing 5 = Extremely convincing Argument: "argument". Respond with valid JSON containing the emotion and the numerical rating: {{"emotion": "emotion_name", "rating": [number from 1-5]}}`
Appr↔CVC	`Task: Appraisal Prediction: You will be given an argumentative text. Your task is to label each appraisal dimension on a 1-5 scale from the perspective of a person hearing the argument. For each appraisal, one means the appraisal does not apply at all, 5 means it applies extremely. The appraisals are: Suddenness: the argument appears sudden or abrupt to the receiver Suppression: the receiver tries to shut the argument out of their mind Familiarity: the argument is familiar to the receiver Pleasantness: the argument is pleasant for the receiver Unpleasantness: the argument is unpleasant for the receiver Consequencial Importance: the argument has important consequences for the receiver Positive Consequentiality: the argument has positive consequences for the receiver Negative Consequentiality: the argument has negative consequences for the receiver Consequence Manageability: the receiver can easily live with the unavoidable consequences of the argument Internal Check: the consequences of the argument clash with the receiver's standards and ideals External Check: the consequences of the argument violate laws or socially accepted norms Response urgency: the receiver urges to immediately respond to the argument Cognitive Effort: processing the argument requires a great deal of energy of the receiver Argument Internal Check: statements in the argument clash with the receiver's standards and ideals Argument External Check: statements in the argument violate laws or socially accepted norms` `Task 2: Convincingness Prediction. Your task is to assign how convincing a person would find the argument on a 1-5 scale. Rating scale: 1 = Not at all convincing 2 = Slightly convincing 3 = Moderately convincing 4 = Very convincing 5 = Extremely convincing Argument: "{argument}". You must respond with ONLY a valid JSON object. Each key must have an integer value between 1 and 5. Format: {{ "suddenness": 1, "suppression": 1, "familiarity": 1, "pleasantness": 1, "unpleasantness": 1, "consequential_importance": 1, "positive_consequentiality": 1, "negative_consequentiality": 1, "consequence_manageability": 1, "internal_check": 1, "external_check": 1, "response_urgency": 1, "cognitive_effort": 1, "argument_internal_check": 1, "argument_external_check": 1, "convincingness": 1 }}. Replace each "1" with your actual rating (1-5) for that dimension.`

Table 2: Prompts for the Joint configuration of appraisal/emotion conditioned convincingness prediction.

Pipeline Model. We expand two configurations, one in which we add the appraisal information (*Appr→CVC*, and one in which we add emotion information to the prompt (*Emo→CVC*). In these two cases, the appraisal and emotion information stems from the annotated data, and may differ for otherwise textually comparable instances (see Section 4 for an explanation of the data we use). These two settings are used to compare the performance to the baseline CVC prediction. This setup only allows a unidirectional information flow from the emotion/appraisal representation to the convincingness prediction. Table 1 shows the prompts for the pipeline configuration.

Joint Model. Presumably, the convincingness of an argument does not only depend on the appraisal, but also the other way around. Therefore, we also perform a joint model experiment, in which the language model is requested to output emotion or appraisal variables, together with the convincingness. We refer to this model as *Appr↔CVC* and *Emo↔CVC*. This setup enables an information flow between the emotion representation and the convincingness assessment in both directions.

3.2. Models

We prompt three large language models (LLMs), namely Mistral-Small-2407 (`Mistral`), LLaMA3.3:70B (`Llama`) and Gemma-3-27B-IT (`Gemma`). Mistral-Small-2407 (Jiang et al., 2023) is a compact decoder-only transformer with 22 billion parameters, optimized for fast inference and low latency applications. LLaMA3.3:70B (Grattafiori et al., 2024) is a large-scale generative model with 70 billion parameters, designed for high-performance instruction following and multilingual reasoning. Gemma-3-27B-IT (Team et al., 2025) is an instruction-tuned model with 27 billion parameters, optimized for task completion.

Argument	CVC	Emotion	Suddenness	Suppression	Familiarity	Pleasantness	Unpleasantness	Consequencial Importance	Positive Consequentiality	Negative Consequentiality	Consequence Manageability	Internal Check	External Check	Response Urgency	Cognitive Effort	Argument Internal Check	Argument External Check
it could be considered that holocaust denial	5	Relief	1	1	3	3	1	2	4	1	4	1	1	1	5	1	1
is a hate crime, laws should be in place to	4	Trust	3	1	4	1	4	1	1	1	3	3	2	2	5	1	4
protect the memory of those who have	3	Sadness	1	2	3	1	4	2	1	1	1	2	2	2	2	1	3
perished in unspeakable crimes such as the	4	Anger	1	1	3	1	3	1	1	1	4	1	1	1	5	2	1
holocaust.	4	Trust	1	1	3	3	1	4	3	1	4	1	1	4	5	2	1
it is a great thing when dads want to stay	1	Sadness	3	5	4	1	5	2	5	1	5	1	5	5	5	5	5
home with their children, but often they are	3	Relief	1	1	2	1	2	2	1	1	4	2	2	2	5	1	2
needed as the main income earner so	4	Sadness	1	1	4	4	1	2	1	3	1	4	1	2	2	1	1
subsidizing them would help greatly.	1	Disgust	3	5	1	1	5	2	2	1	4	1	5	1	2	4	5
	4	Joy	1	1	3	3	1	2	1	1	1	4	1	2	2	1	1

Table 3: Examples from the ContArgA corpus. Emotion, appraisal, and convincingness assessments are from five individual annotators. The convincingness (CVC) and each appraisal dimension are evaluated on a 1–5 scale.

The Mistral and Gemma models are accessed via their respective APIs. The LLaMA model is accessed locally via Ollama[1]. For all models, we set the temperature to 0.1 and leave all other parameters at their respective default values[2].

4. Data

The data we use in our study is the ContArgA data set. It has been presented as part of the Contextualized Argument Appraisal Framework, in which the authors propose a set of concrete appraisal variables that may be used for the evaluation of arguments (Greschner et al., 2026). The corpus situates argument appraisal in its communicative context, capturing the interplay between sender, receiver, and argument rather than treating persuasion as a purely textual property. ContArgA comprises 800 arguments drawn from the UKP-ConvArgv1 (Habernal and Gurevych, 2016b) and IBM-Rank-30k (Gretz et al., 2020) datasets, each annotated by five participants, resulting in 4,000 contextualized annotations. Table 3 displays examples of two arguments with convincingness, emotion, and appraisal assessments from 5 individual annotations each.

An important challenge to obtain subjective labels such as emotions, appraisals or convincingness is to access a person's evaluation in a realistic setup, without using external annotators that would need to reconstruct a presumable evaluation of somebody else. Data were therefore collected in a role-playing setup which simulated a town-hall meeting in which participants engaged with arguments on 39 topics, then reported their emotional response, cognitive appraisal, and perceived convincingness. Annotations include discrete emotion categories (e.g., anger, trust, relief, sadness), intensity ratings, free-text emotion causes, and 15 appraisal dimensions (such as familiarity, pleasantness, response urgency, and perceived consequences). Participants also provided demographic information and Big Five personality traits for both themselves (as receivers) and their imagined argument sender.

The statistical analysis of the data revealed correlations between emotional and cognitive dimensions: positive emotions (trust, pride, joy, relief) correlate with high convincingness, while negative emotions (anger, disgust, sadness) correspond to low convincingness. Appraisal variables such as pleasantness, positive consequentiality, and familiarity further enhance convincingness, whereas unpleasantness and norm violation decrease it.

The authors of the paper do, however, not conduct a modeling study which would allow any insights in the role of convincingness modeling under a particular argument appraisal or emotion. Therefore, ContArgA constitutes a good starting point for our study.

[1] https://ollama.com/

[2] Code for all experiments: https://github.com/LynnGreschner/categorical-emotions-or-appraisals

5. Experiments

We now turn to the experimental settings of our experiments and answer each research question.

5.1. RQ1: Which emotion model helps LLMs to improve convincingness predictions?

We aim at understanding if providing information about the evoked emotion of a given argument improves the performance of LLMs on the convincingness prediction task. More specifically, we investigate whether there is a difference in the emotion model that provides the information about the emotion. To this end, we compare providing discrete emotion categories (e.g., anger, joy, fear, ...) and appraisal dimensions (familiarity, suddenness, cognitive effort, ...) to the model for the convincingness prediction task.

Experimental Setting. We prompt the three LLMs to predict the convincingness of a given argument on a 1–5 scale, which serves as the baseline for the task. It is a zero-shot prompt, the exact phrasing is displayed in Table 1. For investigating the effect of providing the discrete emotion category, the LLM is provided with the dominant emotion that was evoked in a receiver of a given argument. Similarly, in the third task, we provide the appraisal values for all 15 appraisal dimensions that a participant annotated. The exact prompts used for the experiments can again be found in Table 1. Each model is prompted up to four times if no valid prediction can be extracted[3]. Due to the instability of LLM responses, we run all experiments five times and report the average performance across runs to ensure robust and reliable results.

Results. Table 4 displays the results of the convincingness prediction task using the different settings. `Mistral` and `Gemma` perform best on the convincingness prediction task (.33), `Llama` performs worst with .27. Compared to this baseline, providing the discrete emotion category that was evoked in a receiver of the argument improves the convincingness prediction for all models. The strongest improvement (+.09) is observed for `Llama`, whereas `Gemma` then performs best across models with .41. Interestingly, providing the appraisal values does not improve the convincingness predictions of `Gemma`; the model even performs worse than the baseline. However, both `Mistral` and `Llama` perform best when being provided with the appraisal dimensions (.41 and .42, respectively).

[3]For 75 instances (∼2% of the data), the model fails to provide a valid answer even after the fourth attempt. Such cases are excluded from the evaluation of all tasks.

	CVC Spearman's ρ		
Config.	Mistral	Gemma	Llama
CVC Basel.	.33	.33	.27
Emo→CVC	.38Δ+.05	.41Δ+.08	.36Δ+.09
Appr→CVC	.41Δ+.08	.24Δ−.09	.42Δ+.15
Emo↔CVC	.31Δ−.02	.25Δ−.08	.29Δ+.02
Appr↔CVC	.32Δ−.01	.30Δ−.03	.32Δ+.05

Table 4: Main results of the convincingness (CVC) prediction task, comparing the Pipeline and Joint Setup across three language models. The CVC correlations are reported in Spearman's ρ. Differences of each score compared to the CVC baseline are shown with Δ values.

With respect to our research question, we observe that information about the emotion improves the CVC prediction of all models. While the discrete emotion category reliably improves the performance, we see stronger improvements from the appraisal dimensions, even though it fails for one model. Our results suggest that, while emotional context generally improves the prediction of persuasiveness, different models vary in their sensitivity to categorical and dimensional approaches to representing emotions.

5.2. RQ2: Does jointly predicting emotions/appraisals and convincingness improve the performance compared to the single task predictions?

While Greschner et al. (2026) report strong correlations between emotions, appraisals and convincingness, the convincingness of an argument does presumably not merely depend on the emotion and appraisal, but these evoked emotions also depend on the convincingness. The following experiment examines whether the joint prediction of emotions, appraisals, and convincingness improves the convincingness predictions.

Experimental Setting. We prompt the three LLMs to jointly predict emotions/appraisals and convincingness. In contrast to our first experiment, here the models do not get the information about the emotions/appraisals, but have to jointly predict them simultaneously with the convincingness. We chose this setup due to presumably bi-directional effects between emotions/appraisals and convincingness. In addition, predicting the variables jointly is a more realistic setup since most corpora do not have emotion/appraisal labels in addition to con-

	Single									Joint								
	Mistral			Llama			Gemma			Mistral			Llama			Gemma		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Anger	.18	.52	.26	.19	.43	.26	.18	.25	.20	.18	.44	.26	.20	.41	.27	.19	.23	.21
Disgust	.10	.09	.09	.13	.12	.12	.09	.08	.08	.10	.12	.11	.14	.09	.11	.09	.09	.09
Fear	.06	.32	.11	.08	.26	.12	.05	.36	.09	.06	.34	.11	.08	.27	.12	.05	.40	.09
Guilt	.00	.00	.00	.07	.14	.09	.09	.10	.10	.17	.01	.02	.05	.16	.08	.08	.12	.09
Joy	.14	.31	.19	.20	.11	.14	.20	.08	.12	.15	.28	.19	.18	.08	.11	.19	.09	.13
Pride	.10	.20	.13	.10	.12	.11	.15	.04	.06	.12	.18	.14	.13	.12	.12	.15	.05	.08
Relief	.22	.02	.03	.17	.05	.07	.16	.06	.08	.12	.00	.00	.19	.05	.08	.16	.03	.06
Sadness	.37	.13	.20	.32	.23	.27	.27	.23	.25	.36	.15	.21	.31	.21	.25	.29	.22	.25
Shame	.19	.01	.02	.05	.00	.00	.14	.03	.05	.20	.03	.02	.00	.00	.00	.18	.07	.10
Surprise	.36	.02	.04	.29	.03	.05	.31	.04	.08	.37	.03	.05	.30	.04	.07	.31	.04	.08
Trust	.28	.14	.19	.27	.33	.30	.24	.39	.30	.27	.25	.26	.27	.39	.32	.25	.42	.31
Avg	.18	.16	.12	.17	.16	.14	.17	.15	.13	.19	.16	.12	.17	.17	.14	.18	.16	.13

Table 5: Precision, Recall, and F1 values (macro-average) for the emotion classification task, comparing single emotion predictions and the emotion predictions from the joint modeling task.

vincingness ratings. The exact prompt formulations can be found in Table 2.

Results. We report the results of this experiment in Table 4. Jointly predicting emotions and convincingness only improves the convincingness prediction for `Llama` (+.05). Both `Mistral` and `Gemma` perform worse than the baseline in the joint prediction setting. Similarly, the joint prediction of appraisals and convincingness does not improve the prediction of the convincingness compared to the baseline, except for `Llama`, which also shows the best overall performance in the joint setting (.32).

However, while `Mistral` and `Gemma` fail to beat the baseline in the joint setting, we do observe better performance of all models for jointly predicting appraisals and convincingness compared to jointly predicting discrete emotions and convincingness.

6. Analysis of Emotion Models

We now turn to the two different emotion models in more detail. We aim at understanding why and how the models improve the convincingness prediction.

Emotions. We display the performance of all three models on the single and joint emotion classification task in Table 5. Overall, the models show low performance on the emotion classification task in both the single (averaged F_1 scores of .12, .14, .13 for `Mistral`, `Llama`, `Gemma`, respectively) and the joint setting (averaged F_1 scores of .12, .14, .13 for `Mistral`, `Llama`, `Gemma`, respectively). There are minor differences in the precision and recall values when comparing the settings.

Taking a closer look at the class-wise performance, we find that negative emotions (*anger*, *fear*) show high recall but low precision values, in line with results of emotion predictions on German arguments (Greschner and Klinger, 2025). Overall, all models perform best on predicting negative emotions (*anger*, *disgust*, *fear*, *sadness*), with the exception of one positive emotion (*trust*). Trust is the most frequent emotion label in the gold data.

Turning to model-specific performance, we find that `Llama` consistently achieves the best performance in both single and joint settings, particularly high results are seen for predicting *trust* (.30 and .32 F_1 in the single and joint settings, respectively) and *sadness* (.25 and .27 F_1). In contrast, `Mistral` benefits from the joint modeling, showing improved performance across several emotion categories, notably for *anger* and *sadness*. For `Gemma`, we see unique behavior, on the one hand demonstrating strong capabilities for predicting *trust* (.30 and .31 F), but struggling with *joy* and *pride* (compared to the other models). Notably, `Gemma` is the only model that does not benefit from joint modeling, and as described in Section 5.1, the CVC prediction of `Gemma` fails to improve even when provided with gold-label emotion information, suggesting fundamental differences in how this model processes and integrates emotion information.

Appraisals. The performance of all models on the appraisal prediction task is low – in both the single and joint setting. All models predict the appraisal values similarly when comparing the single and joint predictions. However, there are some model-specific and appraisal-specific differences.

We find the best performance for predicting appraisals for the appraisal dimensions of *pleasantness* (correlation values of .32, .31, .31 for `Mistral`, `Llama`, `Gemma`, respectively on the single task and .32, .32, .32 on the joint task) and *unpleasantness* (.28, .29, .29 for `Mistral`, `Llama`, Gemma, respectively on the single task and .28, .30, .30 on the joint task). The appraisal dimension of *familiarity* also shows comparably high results (.20, .12, .19 for Mistral, Llama, Gemma, respectively on the single task and .20, .11, .19 on the joint task). The appraisal dimensions *negative consequentiality* and *consequence manageability* show the lowest performance.

Considering model-specific behaviour, we see that for the two best-performing appraisals (*pleasantness* and *unpleasantness*), all three models perform on par or marginally better in the joint setting. However, for *familiarity*, only Llama struggles with the prediction in both settings (.12 and .11 in single and joint settings, respectively).

In contrast to the CVC prediction, where the single task predictions (i.e., the CVC baseline) show higher performance compared to the joint prediction of CVC and appraisals, there is only a marginal difference between using a single or joint setting when predicting appraisal dimensions in a zero-shot prompting setting.

	Single			Joint		
	M	L	G	M	L	G
Arg. Ext. Check	.09	.13	.06	.10	.13	.07
Arg. Int. Check	.11	.11	.10	.11	.10	.10
Cog. Eff.	.03	.01	.02	.01	.01	.00
Conseq. Manag.	−.02	−.02	−.02	−.01	−.01	−.01
Conseq. Import.	.05	.04	.03	.05	.04	.03
Ext. Check	.15	.14	.13	.15	.14	.12
Fam.	.20	.12	.19	.21	.11	.19
Int. Check	.03	.02	.01	.03	.03	.00
Neg. Consequ.	−.03	−.05	−.02	−.03	−.04	−.02
Pleas	.32	.31	.31	.32	.32	.32
Pos. Consequ.	.12	.13	.12	.12	.13	.12
Resp. Urg.	.10	.14	.13	.12	.13	.11
Sudd.	.08	.15	.07	.15	.16	.07
Supp.	.13	.12	.08	.14	.11	.09
Unpleas.	.28	.29	.29	.28	.30	.30
Average	.11	.11	.09	.11	.11	.10

Table 6: Appraisal Prediction Evaluation Results. Results of the appraisal prediction task for individual appraisal dimensions and the average, all reported as correlations using Spearman's ρ.

6.1. Discussion

The most significant finding is that appraisal dimensions consistently outperform categorical emotions in improving convincingness predictions. This aligns with appraisal theory's premise that cognitive evaluations are more predictive than discrete emotional categories. The superior performance is particularly evident in the pipeline setting, where gold-standard appraisal information yield the strongest improvements (up to +.15 for Llama). However, while the prediction of convincingness performs moderately, the performance of all models on the appraisal prediction task is low. Our findings of consistent underperformance indicate that zero-shot prompting methods may be insufficient for capturing the complex bidirectional relationships between emotions, appraisals, and convicingness. This highlights the need for more distinguished appraisal predictors for automatically predicting convincingness in arguments.

7. Conclusion

In this work, we investigated whether appraisal theories can computationally explain argument convincingness compared to categorical emotion models. Using zero-shot prompting experiments of Gemma, Llama, and Mistral on the ContArgA corpus, our results demonstrate that information from both emotion models improves the convincingness prediction task over the baseline model. Categorical emotions improve performance consistently across models; however, appraisal dimensions showed stronger effects on the predictions. This supports our hypothesis that the subjective nature of argument evaluation benefits from the more granular cognitive assessment provided by appraisal dimensions. Therefore, moving beyond categorical emotion models toward cognitively grounded appraisal frameworks can enhance our understanding of subjective argument evaluation.

Our study reveals clear challenges. The joint modeling of emotions/appraisals and convincingness, while theoretically promising (allowing bidirectional information flow), did not yield improvements over pipeline approaches using gold-standard appraisal information to guide convincingness predictions. This underlines the need for developing better predictors for both emotion and appraisal dimensions in the context of arguments – potentially through fine-tuning, richer context modeling, or multi-task learning strategies.

Future work should focus on advancing automatic appraisal and emotion prediction models in argumentation settings, exploring adaptive and fine-tuned approaches, and validating appraisal-based models across languages and domains. Moreover, we lay the ground work for integrating appraisal information into downstream applications such as argument generation. Adopting cognitively grounded emotion models offers the potential to make computational argument analysis both more robust and more human-aligned.

8. Limitations

The following limitations should be considered when interpreting our results. With respect to our methodology, we do not explore few-shot learning, chain-of-thought reasoning, or fine-tuning approaches. We focus on zero-shot prompting, which possibly does not capture the full spectrum of LLM capabilities for convincingness prediction, because (1) and (2) adopting this straightforward approach allows us to clearly assess the validity of appraisals in arguments rather than confounding the results with the complexity of more elaborate prompting or fine-tuning strategies. The generally low performance on emotion and appraisal prediction, however, rather reflects the difficulty of such subjective tasks rather than fundamental flaws in our approach. Our evaluation focuses primarily on correlation metrics, which do effectively capture relationships between emotions and convincingness, but might not fully reveal nuanced patterns in model behavior. We leave extensive error analysis to understand the underlying reasons for the model's performance variability to future work.

All experiments are conducted using the ConTArgA corpus, which, despite its careful construction, is limited to short, isolated arguments. The results of our experiments might differ when investigating arguments in context, i.e., in debates or social media discussions. Further, we only use English arguments, potentially missing cross-linguistic and cross-cultural variations in emotional responses to arguments.

9. Ethical Considerations

Our work has been approved by the ethics board of the University of Suttgart. Considering the ethical considerations our work poses, we follow the recommendations with respect to ethical challenges in emotion analysis by Mohammad (2022). We do not create any new artifacts other than automatic predictions. We use an existing, freely available dataset as our data source and use open-source large language models in our experiments. Automatically inferring emotional states using cognitive appraisals from argumentative texts could, in theory, provide insights into individuals' psychological states and cognitive processes that they might not have wanted to be inferred or analyzed. However, all participants involved in the creation of the data we use were informed and gave their consent for their answers to be used in scientific publications.

Automatic emotion analysis systems can be biased for various reasons (Kiritchenko and Mohammad, 2018). Our models may show biases related to demographics, cultural backgrounds, or linguistic expressions. Convincingness assessments are inherently subjective in nature, i.e., what makes an argument convincing varies for different people and groups. Predicting argument convincingness based on individual emotional and cognitive assessments could be misused to craft manipulative arguments that exploit emotional vulnerabilities or cognitive biases. In political, commercial, or propaganda contexts, such capabilities could undermine informed democratic deliberation. We recommend future work to include robust safety measurements against misuse for downstream tasks.

10. Acknowledgements

This project has been conducted as part of the Emcona (The Interplay of Emotions and Convincingness in Arguments) project, which is funded by the German Research Foundation (DFG, project KL2869/12–1, project number 516512112).

11. Bibliographical References

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics.

Aristotle. 1991. *On Rhetoric: A Theory of Civic Discourse*, 2nd edition. Oxford University Press, New York.

Mohamed Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. 2015. Emotions in argumentation: an empirical evaluation. In *International Joint Conference on Artificial Intelligence*.

John Anthony Blair. 2011. *Groundwork in the Theory of Argumentation: Selected Papers of J. Anthony Blair*. Springer, Dordrecht, Netherland.

Gerd Bohner, Kimberly Crow, Hans-Peter Erb, and Norbert Schwarz. 1992. Affect and persuasion: Mood effects on the processing of message content and context cues and on subsequent behavior. *European Journal of Social Psychology*, 22:511–530.

Yanran Chen and Steffen Eger. 2025. Do emotions really affect argument convincingness? a dynamic approach with LLM-based manipulation checks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24357–24381, Vienna, Austria. Association for Computational Linguistics.

Yanran Chen, Lynn Greschner, Roman Klinger, Michael Klenk, and Steffen Eger. 2026. Emotionally charged, logically blurred: Ai-driven emotional framing impairs human fallacy detection. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Rabat, Marocco. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

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

Lynn Greschner, Sabine Weber, and Roman Klinger. 2026. Trust me, i can convince you: The contextualized argument appraisal framework and the contarga corpus. In *Proceedings of the Language Resources and Evaluation Conference*, Palma de Mallorca, Spain. European Language Resources Association.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

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.

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

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain. Association for Computational Linguistics.

Saif M. Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Richard Petty, David Schumann, Steven Richman, and Alan Strathman. 1993. Positive mood and persuasion: Different roles for affect under high and low-elaboration conditions. *Journal of Personality and Social Psychology*, 64:5–20.

Michael Pfau, Erin Alison Szabo, Jason Anderson, Joshua Morrill, Jessica Zubric, and Hua-Hsin Wan. 2006. The role and impact of affect in the process of resistance to persuasion. *Human Communication Research*, 27:216 – 252.

Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734.

Lotte Pummerer, Theofilos Gkinopoulos, Karen M. Douglas, Daniel Jolley, and Kai Sassenberg. 2024. The Appraisal Model of Conspiracy Theories (AMCT): Highlighting Core Concepts and Potential Extensions. *Psychological Inquiry*, 35(3-4):233–242. Publisher: Informa UK Limited.

Carlotta Quensel, Neele Falk, and Gabriella Lapesa. 2025. Investigating subjective factors of argument strength: Storytelling, emotions, and hedging. In *Proceedings of the 12th Argument mining Workshop*, pages 126–139, Vienna, Austria. Association for Computational Linguistics.

Ira Roseman. 1984. Cognitive determinants of emotion: A structural theory. *Rev. Pers. Soc. Psychol.*, 5.

Ira Roseman and Craig Smith. 2001. *Appraisal Theory: Overview, Assumptions, Varieties, Controversies*. Oxford University Press, Oxford, UK.

Klaus Rainer Scherer. 2009. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, 23(7):1307–1351.

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

Craig Smith and Phoebe Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of Personality and Social Psychology*, 48(4):813–838.

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.

Ala Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. 2025. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13090–13120, Vienna, Austria. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

1634–1658, Miami, Florida, USA. Association for Computational Linguistics.

Enrica Troiano, Laura Ana Maria Oberlaender, Maximilian Wegge, and Roman Klinger. 2022. x-envent: A corpus of event descriptions with experiencer-specific emotion and appraisal annotations. In *Proceedings of the Language Resources and Evaluation Conference*, pages 1365–1375, 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).

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1519–1538, Torino, Italia. ELRA and ICCL.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Leila Worth and Diane Mackie. 1987. Cognitive mediation of positive affect in persuasion. *Social Cognition*, 5:76–94.

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical Twitter posts. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 131–142, Online. Association for Computational Linguistics.

Gerard Christopher Yeo and Kokil Jaidka. 2025. Beyond context to cognitive appraisal: Emotion reasoning as a theory of mind benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26517–26525, Vienna, Austria. Association for Computational Linguistics.

Hongli Zhan, Desmond 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.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363, Toronto, Canada. Association for Computational Linguistics.