Fearful Falcons and Angry Llamas: Emotion Category Annotations of Arguments by Humans and LLMs

Lynn Greschner and **Roman Klinger**
Fundamentals of Natural Language Processing, University of Bamberg, Germany
{lynn.greschner,roman.klinger}@uni-bamberg.de

Abstract

Arguments evoke emotions, influencing the effect of the argument itself. Not only the emotional intensity but also the category influence the argument's effects, for instance, the willingness to adapt stances. While binary emotionality has been studied in arguments, there is no work on discrete emotion categories (e.g., 'anger') in such data. To fill this gap, we crowdsource subjective annotations of emotion categories in a German argument corpus and evaluate automatic LLM-based labeling methods. Specifically, we compare three prompting strategies (zero-shot, one-shot, chain-of-thought) on three large instruction-tuned language models (Falcon-7b-instruct, Llama-3.1-8B-instruct, GPT-4o-mini). We further vary the definition of the output space to be binary (is there emotionality in the argument?), closed-domain (which emotion from a given label set is in the argument?), or open-domain (which emotion is in the argument?). We find that emotion categories enhance the prediction of emotionality in arguments, emphasizing the need for discrete emotion annotations in arguments. Across all prompt settings and models, automatic predictions show a high recall but low precision for predicting anger and fear, indicating a strong bias toward negative emotions.

1 Introduction

The role of emotionality received considerable attention in recent research, spanning from a focus on pathos (Evgrafova et al., 2024) to the study of emotion intensity and its role on argument persuasiveness (Benlamine et al., 2017a; Griskevicius et al., 2010). In natural language processing, most research focused on binary or continuous emotion concepts (El Baff et al., 2020, i.a.), and the role of such variables on argument effectiveness has been confirmed in empirical studies. In argumentation theory and psychology, however, it has also been shown that not only binary emotionality or intensity are important, but also the concrete emotion category, or groups of emotions, matter. Positive emotions, for instance have a positive effect on cognitive abilities and therefore the willingness to adapt an own stance (Griskevicius et al., 2010). Negative emotions are often part of persuasion strategies (Boster et al., 2016) while they can also lead to defense behaviour (Leventhal and Trembly, 1968). We therefore argue that there is a mismatch between research in natural language processing for argument mining and psychological and theoretical work on argument analysis.

Argument	Label
Es gibt Impfstoffe, welche unsere DNA dauerhaft verändern können. Diese Impfstoffe nennen sich mRNA-Impfstoffe. Das mRNA gelangt in unsere DNA und gliedert sich dort mit ein. Dadurch wird unsere DNA leicht verändert. *There are vaccines that can permanently alter our DNA. These vaccines are called mRNA vaccines. The mRNA gets into our DNA and integrates into it. This slightly alters our DNA.*	Interest, Disgust, Surprise

Table 1: Example argument with discrete emotion labels from human annotators in `Emo-DeFaBel` (English translation in italics).

To fill this gap, we approach argumentative texts as a domain for emotion analysis, more concretely the prominent subtask of emotion categorization. In this research direction, the goal is to assign emotion concepts to predefined textual units, for instance literary texts (Mohammad, 2011), political debates (Tarkka et al., 2024), or social media texts (Mohammad et al., 2014). Emotions are often expressed implicitly, without explicitly mentioning emotion concepts (Casel et al., 2021; Klinger et al., 2018; Koga et al., 2024; Lee and Lau, 2020). That renders emotion classification a challenging task, even for humans, who tend to agree more with other readers of an emotional text than with the original emotion experiencer (Troiano et al., 2023).

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 628–646
May 3-4, 2025 ©2025 Association for Computational Linguistics

Statement	Argument
Kamele speichern Fett in ihren Höckern.	Kamele sind sehr große Tiere und benötigen sehr viel Energie. Um diese Energie aus den Fettreserven zu erhalten, wird das Fett in den Höckern gespeichert. Da Kamele sich meistens in Gegenden befinden, in denen sie wenig Nahrung finde und dort als Lastentiere eingesetzt und lange Wege zurücklegen, werden, ist es überaus wichtig, sich vorher einen Fettspeicher anzulegen. Außerdem schützen die mit Fett gefüllten Höcker die Kamele vor der Hitze und der Sonne, sie wirken wie eine Art Polster, dass die übrigen Organe vor Überhitzung schützt.
Camels store fat in their hump.	*Camels are very large animals and need a lot of energy. To energy from the fat reserves, the fat is stored in the humps. humps. Since camels are usually found in areas where they find little food and are used there as beasts of burden and travel long distances, it is extremely important to build up a fat to build up a fat store beforehand. In addition, the fat-filled humps humps filled with fat protect the camels from the heat and the sun, they act They act like a kind of cushion that protects the other organs from overheating.*

Table 2: Example statement and participant-generated argument from the DeFaBel corpus (Velutharambath et al., 2024). English translation in italics.

Hence, we crowdsource human emotion and convincingness annotations for the publicly available German DeFaBel corpus (Velutharambath et al., 2024). Both convincingness and (discrete) emotion labels are annotated based on the perceived convincingness and evoked emotion in the participants as exemplified in Table 1. We compare these labels to automatically assigned labels by three large language models utilizing three prompting approaches. Our main contribution is therefore the corpus Emo-DeFaBel, (1) the first argumentative corpus human-labeled with emotion categories and (2), an analysis of the performance of the language models for emotion analysis.

Analyzing Emo-DeFaBel reveals that joy and pride in arguments are correlated with higher convincingness, while anger is negatively correlated. Our experiments demonstrate that the prompt-based model categorizations are heavily biased toward negative emotions. The biases on concrete emotions differ between the models.

2 Related Work

2.1 Language Models for Emotion Analysis

With the rise of LLMs, utilizing such models for emotion analysis has received some attention. Churina et al. (2024) explored the capabilities of LLMs for empathy and emotion prediction in dialogues. Cheng et al. (2024); Nedilko (2023) focused on multilingual analyses. Bagdon et al. (2024) studied best–worst scaling as an approach for emotion intensity annotations with language models.

Generally, LLMs may replicate human annotations well, for instance in Finnish parliamentary debates (using GPT4, Tarkka et al., 2024). Malik et al. (2024) report a similar success for French Tweets. Gilardi et al. (2023) highlight performance and cost advantages of such approach over manual annotations. We follow this prior work and transfer it to emotion analysis for argument data.

2.2 Prompting Approaches

One important challenge when prompting language models for language understanding tasks is to find well-performing instructions (Ye et al., 2024). Despite efforts to automatically create appropriate prompts (Li et al., 2023; Chen et al., 2024), prompts commonly need to be adapted to a domain at hand, and are not sufficiently robust across use-cases.

Reynolds and McDonell (2021) demonstrate that zero-shot prompts can outperform one-shot prompts, arguing that providing examples does not necessarily improve performance. Fonseca and Cohen (2024) explore LLMs learning capabilities for new facts or concept definitions through prompts. Their results find that zero-shot prompting improves sentence labeling performance, but larger models (70B+ parameters) struggle with counterfactual scenarios. GPT-3.5 was the only model to detect nonsensical guidelines, while Llama-2-70B-chat often outperformed Falcon-180B-chat, suggesting that increasing model size alone does not guarantee better adherence to guidelines.

Numerous experimental results suggest that chain-of-thought prompting leads to performance improvements (Kojima et al., 2024; Du et al., 2023, i.a.). In contrast, Le Scao and Rush (2021) find that the prompt choice is not the most dominant parameter when optimizing model performance in low-data regimes. We, therefore, consider three commonly used prompting approaches (zero-shot, one-shot, and chain-of-thought approaches) for emotion analysis in arguments.

		Binary	Closed-domain	Open-domain
Zero-shot	Role	You are an expert on emotions in arguments.		
	Task Desc.	Label the following argumentative text about a statement into containing emotion(s) (emotion:1) or not containing emotions (emotion:0).	Your task is to label an argumentative text about a statement with the most present emotion a reader would feel. The options for labels are: [Emo].	Your task is to label an argumentative text about a statement with the most present emotion a reader would feel.
	Format	Provide the output in a json format with the key being 'emotion' and the value being the emotion label as a string. For example, if you believe the argument contains sadness, your json output should be:		
		'1'.	'sadness'.	
	Texts	Now label the following argumentative text with the emotion label. Statement: {statement}. Text: {text}. What is the emotion label for this argument? Only output the json format.		
One-shot	Role	You are an expert on emotions in arguments.		
	Task Desc.	Label the following argumentative text about a statement into containing emotion(s) (emotion:1) or not containing emotions (emotion:0).	Your task is to label an argumentative text about a statement with the most present emotion a reader would feel. The options for labels are: [Emo].	Your task is to label an argumentative text about a statement with the most present emotion a reader would feel.
	Ex.	[Example with correct output.]		
	Format	Provide the output in a json format with the key being 'emotion' and the value being the emotion label as a string. For example, if you believe the argument contains 'sadness', your json output should be:		
		'1'.	'sadness'.	
	Texts	Now label the following argumentative text with the emotion label. Statement: {statement}. Text: {text}. What is the emotion label for this argument? Only output the json format.		
Chain-of-Thought	Role		You are an expert on emotions in arguments.	
	Task Desc.		Your task is to first classify an argumentative text into either containing an emotion or not containing an emotion. If the text contains an emotion, continue with the following task: Your task is to label the text with one emotion a reader would feel the strongest when reading the argument. The option for labels are: [Emo] In your answer, only provide the emotion label you choose as the output.	Your task is to first classify an argumentative text into either containing an emotion or not containing an emotion. If the text contains an emotion, continue with the following task: Your task is to label the text with one emotion a reader would feel the strongest when reading the argument. In your answer, only provide one emotion label you choose as the output.
	Ex.		[Example with correct output.]	
	Format		Provide the output in a json format with the key being 'emotion' and the value being the emotion label as a string. For example, if you believe the argument contains fear, your json output should be: {'emotion': Fear}".	
	Texts		Now label the following argumentative text with the emotion label. Statement: {statement}. Text: {text}. What is the emotion label for this argument? Only output the json format.	

Table 3: Prompt templates for emotion domain (binary, closed-domain, open-domain) and prompting settings (zero-shot, one-shot, chain-of-thought). [Emo] refers to the set of JOY, ANGER, FEAR, SADNESS, DISGUST, SURPRISE, PRIDE, INTEREST, SHAME, GUILT, NO EMOTION. [Example with correct output] consists of a human-annotated argument with an emotion label and is the consistent for all prompts. Color highlights shared elements.

2.3 Emotions in Arguments

Habernal and Gurevych (2016, 2017) constructed and analyzed a corpus for convincingness strategies in online argumentative text, including emotionality. Lukin et al. (2017) highlight the role of emotions in interaction with differing personalities on the perceived convincingness of arguments. There is further a substantial body of psychological studies which point to the role of cognitive argument evaluations for convincingness (Bohner et al., 1992; Petty et al., 1993; Pfau et al., 2006; Worth and Mackie, 1987; Benlamine et al., 2015). Benlamine et al. (2017b) demonstrate that the argumentation strategy *Pathos* (i.e., using emotions) is most efficient for changing a persons opinion. Related to that, Konat et al. (2024) identify pathos-related argument schemes in arguments and their relation to emotion-eliciting language in audiences using sentiment analysis.

Research in NLP focuses, so far, on binary emotionality and emotion intensity in arguments, as one of many factors of convincingness (Habernal and Gurevych, 2017) or rate the emotional appeal (Wachsmuth et al., 2017; Lukin et al., 2017). Cigada (2019) analyze two expressions of emotions (appreciation and tension, not discrete emotion labels) of one French speaker.

Similar to our study, Leoni et al. (2018) annotate evoked emotions in participants of argumentative debates, but, in contrast to our study, using facial emotion recognition tools.

We are not aware of work that focuses on discrete emotion categories in argumentative text. Our study closes this gap.

3 Annotation

We enhance an existing argument dataset with emotion labels as the basis for our study. In total, we request three individual annotations for 300 arguments.

Data Sets. The basis for our annotation is the `DeFaBel` corpus (Velutharambath et al., 2024). It contains argumentative German texts, annotated via crowdsourcing. The participants were asked to write persuasive arguments supporting a given statement (e.g., "Camels store fat in their hump."), selected from the TruthfulQA dataset (Lin et al., 2022). The corpus contains 1031 arguments for 35 statements. We select this resource because it contains short, isolated arguments. Additionally, this dataset allows us to effortlessly capture the annotators' stances towards each statement. We randomly select 300 arguments for our annotation task, evenly distributed across statements. Table 2 shows an example statement–argument pair.

Emotion Labelset. Our emotion label set starts with the basic emotions (anger, disgust, fear, joy, sadness, surprise) and is expanded by cognitive evaluations (interest) and self-directed states (shame, guilt, pride). We further offer a free text field for mentioning evoked emotions that are not covered by our label set as an exploratory approach.

Annotation Setup. We show one statement–argument pair per page. The participants are instructed to read those texts and provide their stance toward the statement on a 5-point scale (strongly agree, . . . , strongly disagree). In addition, we ask how convincing they perceive the argument on a 5-point scale (not convincing at all, . . . , very convincing). For the emotion label, we first ask if the argument evokes an emotion in the participants (yes/no). If they answer yes, they are asked to provide the concrete emotion label from our emotion label set (JOY, ANGER, FEAR, SADNESS, DISGUST, SURPRISE, PRIDE, INTEREST, SHAME, GUILT). Participants have the option to input an evoked emotion in case it is not covered by our label set. Table 4 displays an overview of the collected labels, question phrasings, and possible answers. The free-text field answers are collapsed to a closed set for further modeling and analysis (see Appendix 8). We show a screenshot of one example annotation page in the Appendix in Figure 3.

Var.	Question Text	Label
Stance	Stimmen Sie der Aussage zu? *Do you agree with the statement?*	1–5
Fam.	Wie gut kennen Sie sich mit dem Thema aus? *How familiar are you with the topic?*	1–5
Conv.	Wie überzeugend ist das Argument für Sie? *How convincing is this argument for you?*	1–5
Binary	Wird eine Emotion in Ihnen ausgelöst wenn sie das Argument lesen? *Is an emotion triggered in you when you read the argument?*	Yes/No
Emo.	Beantworten Sie diese Frage nur wenn Sie die vorangegangene Frage mit "Ja" beantwortet haben. Welche der folgenden Emotionen wird am stärksten in Ihnen ausgelöst wenn Sie das Argument lesen? *Only answer this question if you have answered the previous question with "Yes". Which of the following emotions is triggered most strongly in you when you read the argument?*	[Emo]

Table 4: Wording and response options for the human annotation study. [Emo] refers to Freude, Wut, Angst, Traurigkeit, Ekel, Überraschung, Stolz, Interesse, Scham, Schuld (*Joy , Anger, Fear, Sadness, Disgust, Surprise, Pride, Interest, Shame, Guilt*). We conducted the study in German, see English translations in italics.

Crowd-sourcing Details. We use the platform Prolific[1] with Potato (Pei et al., 2022). Participants are prescreened to be in Germany, have German as their first and native language, be fluent in German, and have an approval rate of 90–100%.

Each participant answers the survey for five statement–argument pairs (and one attention check, see Figure 6 for an example). We pay each participant 1.20€ for one survey, which on average takes 7 minutes. Participants can participate up to 12 times, and therefore annotate up to 60 statement–argument pairs. In total, the cost of the study amounts to 316.87€. The contributing participants of our studies were on average 35.8 years old (21 minimum, 68 maximum). From this set, 98 identified as male and 96 as female. Note that we did not limit the study to participants with these two genders.

[1] `https://www.prolific.com/`

4 Models

To investigate the efficiency of LLM annotation of emotions in arguments we differentiate between two dimensions, the emotion domain setting (binary, closed-domain, open-domain) and the technique (zero-shot, one-shot, and chain-of-thought) of the prompts. Combining these strategies results in eight prompt formulations (cf. Table 3).[2]

4.1 Emotion Domain

We differentiate between prompting for emotionality (binary) and for discrete emotion categories (closed and open-domain) in arguments.

Binary. In the binary prompt setting, we request a label for the argument indicating if it causes an emotion in the reader or not. We do not distinguish concrete emotion categories. This setting enables us to develop an understanding regarding an agreement without focusing on specific categories.

Closed-domain. The binary setting is contrasted with the request for concrete emotion labels, to enable more detailed analysis of specific categories. We use the label set of JOY, ANGER, FEAR, SADNESS, DISGUST, SURPRISE, PRIDE, INTEREST, SHAME, GUILT, or NO EMOTION, as a combination of basic emotions, cognitive evaluations (interest) and self-directed states (shame, guilt, pride). We consider the latter to be particularly relevant for emotion analysis in argumentative texts.

Open-domain. Hypothetically, a language model may perform well to assign emotion names different from our set. To evaluate this observation in an open-domain setting in which we do not predefine the emotion set. The goal of this approach is to capture a broad range of emotions.

4.2 Prompting Approach

To study the impact of the three domain settings mentioned above across multiple prompting approaches, we design three types of prompts, shown in Table 3.

Zero-shot. Zero-shot (ZS) prompts only contain instructions to complete a given task without any examples or further context. ZS prompts are flexible, comparably straight-forward to design, and no examples are required.

[2]The supplementary material for this paper (code and annotated data) is available at `https://www.uni-bamberg.de/en/nlproc/resources/emodefabel/`.

One-shot. One-shot (OS) prompts augment the instruction to the model with one or few training examples, allowing the model to learn in context (Brown et al., 2020). We perform OS prompting with a manually annotated example from the `DeFaBel` corpus that we use for our experiments which is not part of our test set.

Chain-of-Thought. Since the binary emotionality of a given argument is conditional for the discrete emotion label, we hypothesize that first creating a rationale before giving the prediction enhances the performance of LLMs. Chain-of-thought (CoT) prompting is a technique that triggers the model to generate a series of logical reasoning steps before providing the final answer (Wei et al., 2022). This technique assists models to tackle more complex tasks more effectively by simulating a human-like reasoning process. In our study, we formulate the prompt to force the model to first decide on the binary emotionality of a given argument before providing the discrete emotion label (therefore, no binary CoT prompt). See Table 3 for concrete examples.

4.3 Evaluation

In the following, we explain our evaluation procedure, in which the LLM-based predictions[3] are compared to human annotations (as discussed in Section 3). We use two different strategies for the evaluation, *relaxed* and *strict*, to account for the subjectivity of the task. In the *strict* mode, we compare the model's output to the majority vote from the human annotations. If there is no majority, we assign NO EMOTION. In the *relaxed* mode, we count an output as true positive if it matches *any of* the labels provided by the human annotators. The motivation for this approach is to consider everything to be a correct output that may be relevant for "somebody", acknowledging the subjective nature of the task. When evaluating the models' performance for individual emotion classes, we distribute one count of a false negative prediction across the set of gold labels.

5 Experiments

We now explain the experiments, analyze the human annotations and subsequently answer the research questions stated in the introduction.

[3]LLM output parsing is explained in more detail in Appendix C.

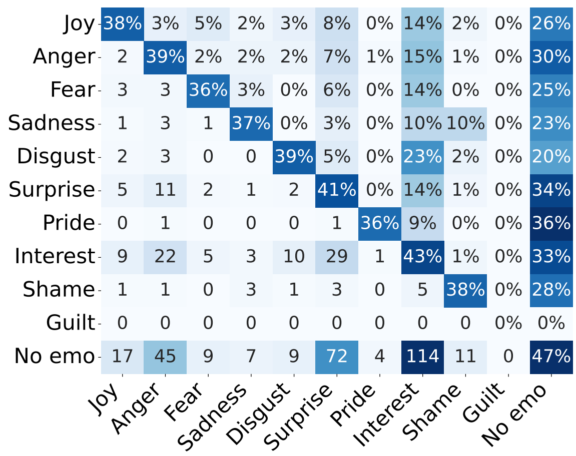


Figure 1: Pair-wise co-occurrences of emotion labels in the human annotation study for 300 arguments. The upper part displays percentages, the lower absolute numbers.

Emotion	Num. agreem. =1	≤2	≤3
JOY	.96	.04	.00
ANGER	.97	.03	.00
FEAR	1.00	.00	.00
SADNESS	.91	.09	.00
DISGUST	.94	.06	.00
SURPRISE	.84	.15	.01
PRIDE	1.00	.00	.00
INTEREST	.68	.28	.05
SHAME	.93	.07	.00
GUILT	.00	.00	.00
NO EMOTION	.43	.43	.15

Table 5: Percentages of emotion label agreements for one, two, and three annotators averaged over 300 arguments, individually for each emotion label.

5.1 Experimental Setting

We use `Falcon-7b-instruct` (Almazrouei et al., 2023), `Llama-3.1-8B-Instruct` (Llama Team, AI @ Meta, 2024), and `GPT-4o-mini` (OpenAI et al., 2024)[4]. `Falcon` is an instruction-tuned generative model with 7 billion parameters. `Llama` has 8 billion parameters and is optimized for multilingual dialogue use cases. We access both models via their respective Huggingface APIs[5]. With `GPT-4o-mini` OpenAI offers a smaller and more cost-efficient model than GPT-4o that outperforms GPT-3.5-Turbo. We access the model via the OpenAI API[6]. We use the default settings for all models. The cost for `GPT` amounts to 0.20€.

5.2 Results

5.2.1 Human Label Analysis

The human study results in 326 annotations of statement–argument pairs, from which we keep a random set of 300 instances. Altogether, 16 annotations are rejected due to failed attention checks.

Out of all arguments, 50% contain a binary emotion label (majority-aggregated). The average argument length between emotional and non-emotional arguments does not differ substantially (78.5 tokens, 4.9 sentences vs. 78.4 tokens, 4.7 sentences). The most frequently annotated emotion label in the closed-domain annotation is INTEREST (207 annotations), followed by SURPRISE (103). PRIDE is the least frequently annotated emotion label (4). Notably, GUILT is never annotated.

Co-occurences of Emotion Labels. The heatmap in Figure 1 shows the frequencies of pair-wise emotion label co-occurrences in absolute and relative numbers. An interesting observation is that negative emotions, such as anger and fear, appear together with the cognitive emotion interest frequently (15 and 14%, respectively). In 23% of cases, disgust appears together with interest, showcasing the subjective nature of the discrete emotion labeling task. We speculate that different emotions might either be evoked from the content of the argument itself, or be dependent of the annotator's stance toward the statement, which we discuss further in Section 5.2.4.

Agreement. Table 5 displays the inter-annotator agreements, where percentages reflect the consistency among annotators in labeling the same argument with the same emotion. More specifically, we calculate the proportion of cases where annotators agreed on a given label for each argument. The agreement for an argument containing FEAR, SADNESS, and PRIDE is low; for SURPRISE and INTEREST, two annotators agree in 13% and 29% of all arguments. Notably, these labels are also the most prevalent within the dataset. The highest agreement for all three annotators agreeing on one label is found for NO EMOTION.

The task is characterized by a substantial disagreement and subjectivity. This is also reflected by the value for Krippendorff's alpha over all arguments and emotion labels, namely 0.04. This low

[4]We refer to the models as `Falcon`, `Llama`, and `GPT`.
[5]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct,https://huggingface.co/tiiuae/falcon-7b-instruct
[6]https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

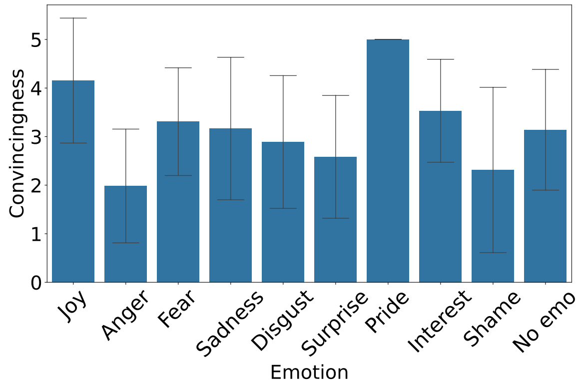


Figure 2: Average convincingness scores (1–5; 5: very convincing; 1: not convincing at all) for each emotion with standard deviation.

		Falcon			Llama			GPT		
		P	R	F_1	P	R	F_1	P	R	F_1
Bin.	ZS	.51	1.00	.67	.55	.67	.61	.59	.21	.31
	OS	.51	.95	.66	.62	.03	.06	.64	.17	.26
Closed	ZS	.50	1.00	.67	.51	.97	.67	.51	.92	.65
	OS	.51	.98	.67	.50	1.00	.67	.50	.94	.65
	CoT	.50	.97	.66	.50	1.00	.67	.54	.71	.61
Open	ZS	.50	1.00	.67	.50	1.00	.67	.50	1.00	.67
	OS	.50	1.00	.67	.50	.99	.67	.50	1.00	.67
	CoT	.50	1.00	.67	.50	.99	.67	.53	.77	.63

Table 6: Performance of the three models in different prompt settings (ZS: zero-shot, OS: one-shot, CoT: chain-of- thought) on predicting the binary emotionality in arguments for the positive class. The binary label is inferred from the emotion labels given in the closed and open-domain emotion settings.

value underscores variability in agreement, which is, however, expected given the subtle and subjective nature of the emotional interpretation of arguments.

Interplay of Emotions and Convincingness in Arguments. We analyze the convincingness of arguments with respect to the emotion labels and display the results in Figure 2. Arguments which evoke pride are perceived to be most convincing, followed by joy and interest. Notably, arguments evoking no emotions are located in the middle of the convincingness distribution. Emotions evoking anger are the least convincing. We conclude that positive emotions, such as joy and pride correlate with a higher convincingness in arguments, whereas negative emotions such as anger with a lower convincingness.

5.2.2 RQ1: Which prompt types lead to reliable results on emotionality in arguments?

To understand how well emotion annotations in arguments can be automatized by prompting large language models, we compare emotion domain settings across prompt types and models. We start with an evaluation of the established binary emotionality setting. Table 6 displays the results for the evaluation of the positive class. In both closed-domain and open-domain scenarios, the binary emotionality label is derived from discrete emotion labels, where emotionality holds if an emotion is predicted (in contrast to NO EMOTION).

`Falcon` performs best in the binary setting (.67/.66) and `GPT` performs worst (.31/.26). For `Llama`, the ZS setting yields considerably higher results than the OS setting (.61/.05). Inferring the binary label from both closed and open-domain prompts improves the performance for `Llama` and `GPT` and while it is similar for `Falcon`. Notably, for `Falcon`, there is no considerable difference between prompting for the binary label directly or inferring the label. The prompting setting does not clearly influence the performance of the models across the emotion domain settings, in line with the findings by Le Scao and Rush (2021).

For all prompting settings and emotion domains, the recall is high (.77 to 1.00). Only `GPT` shows a lower recall in the closed-domain CoT prompt setting (.71). While inferring the binary emotionality label in arguments from closed and open-domain prompts improves the overall performance, the high recall is striking and raises the question about the reliability of the binary emotionality prediction.

5.2.3 RQ2: Which prompt types lead to reliable results for discrete emotion predictions in arguments?

We now explore the discrete emotion predictions, the novelty in our proposed corpus. Table 7 shows the performance of the three models for each prompting approach. Overall, the performance of all models is low in the strict evaluation setting across prompting approaches and emotion domain settings. Note that we macro-average across all emotion classes, which in part attributes for the low overall performances in both evaluation modes.

In the relaxed evaluation setting, `GPT` outperforms `Falcon` and `Llama`. The closed-domain ZS and CoT, and the open-domain CoT prompts lead to the best performance (.16 F_1). Providing an emo-

		Falcon						Llama						GPT					
		Strict			Relaxed			Strict			Relaxed			Strict			Relaxed		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Closed	ZS	.00	.11	.00	.01	.11	.02	.09	.07	.04	.21	.15	.12	.12	.07	.06	.26	.19	.16
	OS	.11	.00	.01	.11	.08	.03	.02	.05	.03	.10	.13	.10	.10	.07	.04	.25	.17	.14
	CoT	.09	.00	.01	.14	.14	.07	.02	.05	.03	.19	.18	.16	.11	.07	.08	.25	.19	.16
Open	ZS	.00	.00	.00	.00	.07	.01	.01	.00	.00	.15	.17	.12	.00	.00	.00	.18	.16	.12
	OS	.00	.00	.00	.00	.07	.01	.01	.02	.00	.15	.17	.12	.00	.00	.00	.02	.19	.15
	CoT	.00	.00	.00	.00	.17	.01	.00	.02	.00	.10	.14	.07	.02	.01	.01	.26	.18	.16

Table 7: Performance of the three models in different prompt settings (ZS: zero-shot, OS: one-shot, CoT: chain-of-thought) on predicting discrete emotion labels in arguments, aggregated over all emotions. The strict evaluation mode uses a majority vote of the emotion label as the gold label while the relaxed evaluation mode allows the set of three emotion labels as the gold labels. All results are macro-averages over all emotion classes.

tion label set and guiding via CoT improves the performance of emotion prediction for GPT, Falcon, and for Llama only in the closed-domain setting. The closed-domain prompts work better for Falcon but not for Llama.

Our results indicate that Falcon, Llama, and GPT cannot reliably predict discrete emotions in arguments. The better performance in the relaxed evaluation setting acknowledges the inherent subjectivity of the emotion annotation task.

5.2.4 RQ3: What biases do LLM predictions of discrete emotion labels in arguments show?

We now aim at understanding if the prompting approaches and the emotion domain setting influence the models toward predicting certain emotions in arguments, hence, if the setting influences the biases. Details on the results discussed in the following are shown in Table 9 in Appendix D. We review the models individually and focus on particularly high or low results on particular emotion classes to reveal biases of the models.

There are notable differences between the models for predicting classes. GPT shows the best performance for ANGER and SURPRISE with high precision across all settings. INTEREST shows a mixed performance (.53, .00, .55, .11, .48, .00). The recall for FEAR is high across prompts (.68, .75, .65, .71, .65, .72), indicating a bias toward that negative emotion. Falcon shows a low performance for PRIDE, INTEREST, SHAME, and GUILT across all prompts. For the emotion label FEAR, we find high recall values (.75, .70, .75, .62, .75) for 5 prompting approaches. Llama shows a mixed performance of predicting the emotion label INTEREST across prompts (.59, .11, .49, .12, .47, .08 F_1) and performs lowest for the emotion classes DISGUST, SURPRISE, PRIDE, INTEREST, SHAME, and GUILT (.00 F_1 scores for all prompts except .08 F_1 for open-domain ZS DISGUST). The best overall performance is achieved for ANGER (.35, .56, .26, .33, .24, .26 F_1 scores) with small differences between prompts. The recall is consistently high, indicating a bias of Llama toward ANGER.

Overall, the performance for the individual emotions differs between models. All models struggle to predict SHAME and GUILT. GPT and Llama show the same preferences for prompts and domain settings for predicting INTEREST. Our results indicate a bias of all models in predicting the negative emotions of anger (Llama) and fear (GPT and Falcon).

Qualitative Analysis. All models show a low performance for emotion category assignments and have a bias toward negative emotions. Therefore, we discuss the overall best-performing model, GPT, with the best performing prompt, closed-domain CoT, for the prediction of FEAR. Detailed examples are in Table 10 in Appendix E.

Based on a random selection of two instances for SURPRISE, INTEREST, NO EMOTION, respectively, in which FEAR is wrongly predicted, we find that the stances of the annotators are presumably the cause for differing annotations of the language model and the human. Across all arguments, we find linguistic cues toward the emotion fear: ‘Gefährdung der eigenen Sicherheit’ (*risk to your own safety*), ‘Unfall’ (*accident*), ‘Krebs’ (*cancer*), ‘Krankheit’ (*illness*), ‘zu hohen Cholesterinwerten’ (*high cholesterol levels*), ‘Explosion’ (*explosion*), ‘die Giftstoffe zerstören den Verdauungstrakt’ (*the toxins destroy the digestive tract*). We speculate that the annotators did not experience fear when reading the arguments because the arguments are focused on indirect or hypothetical events (sharks getting cancer, getting into an accident if you drive barefoot), rather than presenting a personal, immediate threat. GPT is not able to make that distinction.

6 Conclusion

With this paper, we expanded on theoretical work on the interplay of emotion categories and argument convincingness and previous work in NLP on binary emotionality of arguments. We presented Emo-DeFaBel, the first corpus of discrete emotion classes in arguments and analyzed the interplay of emotions and convincingness in German arguments. We found that positive emotions (joy, pride) are correlated with higher convincingness scores and negative emotions (particularly anger) with low convincingness scores, showcasing the relevance of analyzing discrete emotion categories.

When binary emotionality labels are required, we showed that inferring binary labels from discrete emotion classes performs better than directly requesting binary labels from LLMs, a result that may also affect related work on automatically generating persuasive arguments (Chen and Eger, 2025). We further find that there are only minor performance differences across prompting approaches and emotion domains. Falcon, Llama, and GPT show a bias toward predicting negative emotions in arguments. To mitigate these issues, we propose to study fine-tuning or prompt optimization in the context of argument–emotion annotations in future work. Specifically, the precision of individual emotion classes has to be improved. Additionally, a fine-grained analysis of the argumentative structure and quality could further enhance our understanding of the interplay of emotionality and convincingness in arguments. Related to that, we want to point out that our study could be expanded with a perspectivist approach (e.g., focusing on persona–aware annotation, human characteristics and demographics) and discrete emotion annotation and analysis of arguments in discourse.

7 Limitations

Emotion annotation in arguments is a highly subjective task. Assigning evoked emotions from the reader's perspective depends on various factors, including the prior stance toward the topic of the argument. While this subjectivity is manageable for human annotations, we recognize that prompting language models without offering context about the person they are meant to mimic only partially addresses the subjectivity of the task. To mitigate this issue, we employ a relaxed evaluation metric that treats all human annotations of a given argument as a set of gold labels.

Our study has some resource-related limitations. We base the creation of our corpus Emo-DeFaBel on the DeFaBel corpus (Velutharambath et al., 2024), which consists of arguments that were generated in German and in a controlled manner. Consequently, our findings may not generalize to arguments from online discourse or debates in other languages. Additionally, we do not investigate the argument structure or quality of the arguments in DeFaBel because our focus is not on the structural properties or quality of the arguments themselves, but rather on understanding the emotions evoked in readers when engaging with argumentative text.

Our experiments are conducted with three LLMs and the results might differ for other models. However, by employing open-source models (Falcon and Llama), we allow replicating our study with limited resources. Moreover, predictions from different runs might yield different results. We did not see any such variations in our experiments, but a structured evaluation of instability issues might be worth exploring in the future.

Regarding the creation of our corpus, we acknowledge the potential for annotator bias. Although annotators were restricted to labeling each argument only once, participation across up to 12 studies (i.e., 60 arguments) could influence the consistency of gold labels, as individual annotators' interpretations may dominate. Furthermore, the order in which arguments were presented to participants was randomized, which could also introduce biases into the annotations. Arguably, our corpus is comparably small, however, we provide all resources necessary to create more data points (https://www.uni-bamberg.de/en/nlproc/resources/emodefabel/).

8 Ethical Considerations

We collected human annotations for emotions in arguments via crowd-sourcing. For each argument, we asked participants to report their prior stance toward controversial topics. We informed the participants that their answers would be used for a scientific publication and obtained their consent. We do not collect any information that would allow personal identification, therefore, the data is inherently anonymized. While we do reject annotators that did not pass the attention check, we do explicitly warn them about not getting paid if they fail the checks. Naturally, being exposed to arguments for or against a statement can be upsetting during the annotation. However, Velutharambath et al. (2024) point out that they manually selected arguments for their study to minimize potential harm or discomfort that they can cause. Therefore, the statement–argument pairs in our study are not unusually upsetting.

With respect to research on emotion analysis systems, we note that Kiritchenko and Mohammad (2018) observe that such systems are biased for various reasons. Using LLMs to not only predict but automatically label argumentative text with emotions might lead to unpredictable biases, and we are aware that this requires further research. While our main point of this study is not to employ LLMs for automatically labeling emotion analysis-related tasks, in theory, our work can guide future research toward that, which could eventually lead to a decrease in annotation-related jobs.

Acknowledgments

This project has been conducted as part of the EMCONA (The Interplay of Emotions and Convincingness in Arguments) project which is founded by the German Research Foundation (DFG, project KL2869/12–1, project number 516512112). We thank Yanran Chen and Steffen Eger for the fruitful discussions.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

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

Mohamed Benlamine, Ramla Ghali, Serena Villata, Claude Frasson, Fabien Gandon, and Elena Cabrio. 2017a. Persuasive argumentation and emotions: An empirical evaluation with users. In *Human-Computer Interaction. User Interface Design, Development and Multimodality*.

Mohamed Benlamine, Ramla Ghali, Serena Villata, Claude Frasson, Fabien Gandon, and Elena Cabrio. 2017b. Persuasive argumentation and emotions: An empirical evaluation with users. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*.

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

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.

Franklin J. Boster, Shannon Cruz, Brian Manata, Briana N. DeAngelis, and Jie Zhuang. 2016. A meta-analytic review of the effect of guilt on compliance. *Social Influence*, 11(1):54–67.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Felix Casel, Amelie Heindl, and Roman Klinger. 2021. Emotion recognition under consideration of the emotion component process model. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 49–61, Düsseldorf, Germany. KONVENS 2021 Organizers.

Yanran Chen and Steffen Eger. 2025. Do emotions really affect argument convincingness? a dynamic approach with llm-based manipulation checks. *Preprint*, arXiv:2503.00024.

Yongchao Chen, Jacob Arkin, Yilun Hao, Yang Zhang, Nicholas Roy, and Chuchu Fan. 2024. PRompt optimization in multi-step tasks (PROMST): Integrating human feedback and heuristic-based sampling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3859–3920, Miami, Florida, USA. Association for Computational Linguistics.

Long Cheng, Qihao Shao, Christine Zhao, Sheng Bi, and Gina-Anne Levow. 2024. TEII: Think, explain, interact and iterate with large language models to solve cross-lingual emotion detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 495–504, Bangkok, Thailand. Association for Computational Linguistics.

Svetlana Churina, Preetika Verma, and Suchismita Tripathy. 2024. WASSA 2024 shared task: Enhancing emotional intelligence with prompts. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 425–429, Bangkok, Thailand. Association for Computational Linguistics.

Sara Cigada. 2019. Emotions in argumentative narration. *Informal Logic*, 39:401–431.

Chunhui Du, Jidong Tian, Haoran Liao, Jindou Chen, Hao He, and Yaohui Jin. 2023. Task-level thinking steps help large language models for challenging classification task. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2454–2470, Singapore. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Natalia Evgrafova, Veronique Hoste, and Els Lefever. 2024. Analysing pathos in user-generated argumentative text. In *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024*, pages 39–44, Torino, Italia. ELRA and ICCL.

Marcio Fonseca and Shay Cohen. 2024. Can large language models follow concept annotation guidelines? a case study on scientific and financial domains. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8027–8042, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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

Vladas Griskevicius, Michelle N. Shiota, and Samantha L. Neufeld. 2010. Influence of different positive emotions on persuasion processing: A functional evolutionary approach. *Emotion*, 10(2):190–206.

Ivan Habernal and Iryna Gurevych. 2016. 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.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

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.

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. IEST: WASSA-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium. Association for Computational Linguistics.

Yurie Koga, Shunsuke Kando, and Yusuke Miyao. 2024. Forecasting implicit emotions elicited in conversations. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 145–152, Tokyo, Japan. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2024. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Bartosz Konat, Ewa Gajewska, and Wojciech Rossa. 2024. Pathos in natural language argumentation: Emotional appeals and reactions. *Argumentation*, 38:369–403. Accepted: 22 February 2024, Published: 21 June 2024, Issue Date: September 2024.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Sophia Yat Mei Lee and Helena Yan Ping Lau. 2020. An event-comment social media corpus for implicit emotion analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1633–1642, Marseille, France. European Language Resources Association.

Chiara Leoni, Mauro Coccoli, Ilaria Torre, and Gianni Vercelli. 2018. Your paper title here. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Torino, Italy. Accademia University Press.

Howard Leventhal and Grevilda Trembly. 1968. Negative emotions and persuasion. *Journal of Personality*, 36(1):154–168.

Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang, and Tat-Seng Chua. 2023. Robust prompt optimization for large language models against distribution shifts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1539–1554, Singapore. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Llama Team, AI @ Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Usman Malik, Simon Bernard, Alexandre Pauchet, Clément Chatelain, Romain Picot-Clémente, and Jérôme Cortinovis. 2024. Pseudo-labeling with large language models for multi-label emotion classification of french tweets. *IEEE Access*, 12:15902–15916.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA. Association for Computational Linguistics.

Saif Mohammad, Xiaodan Zhu, and Joel Martin. 2014. Semantic role labeling of emotions in tweets. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41, Baltimore, Maryland. Association for Computational Linguistics.

Andrew Nedilko. 2023. Generative pretrained transformers for emotion detection in a code-switching setting. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 616–620, Toronto, Canada. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

M Pfau, A Szabo, J Anderson, Josh Morrill, J Zubric, and H-H H-Wan. 2006. The role and impact of affect in the process of resistance to persuasion. *Human Communication Research*, 27:216 – 252.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. *Preprint*, arXiv:2102.07350.

Otto Tarkka, Jaakko Koljonen, Markus Korhonen, Juuso Laine, Kristian Martiskainen, Kimmo Elo, and Veronika Laippala. 2024. Automated emotion annotation of Finnish parliamentary speeches using GPT-4. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 70–76, Torino, Italia. ELRA and ICCL.

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.

Aswathy Velutharambath, Amelie Wührl, and Roman Klinger. 2024. Can factual statements be deceptive? the DeFaBel corpus of belief-based deception. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2708–2723, 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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

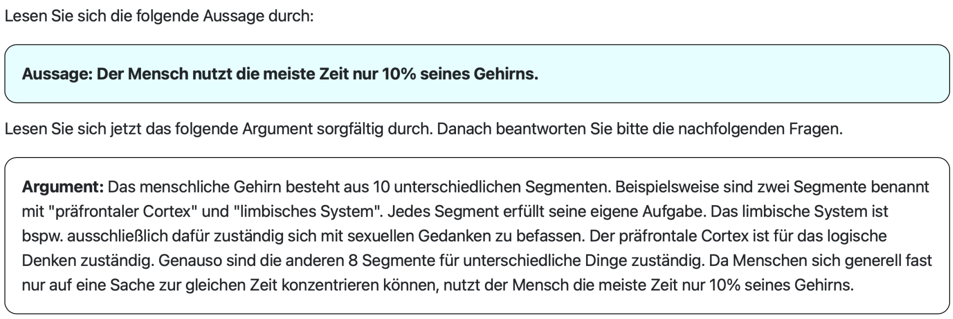
Leila T. Worth and Diane M. Mackie. 1987. Cognitive mediation of positive affect in persuasion. *Social Cognition*, 5:76–94.

Qinyuan Ye, Mohamed Ahmed, Reid Pryzant, and Fereshte Khani. 2024. Prompt engineering a prompt engineer. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 355–385, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

A Corpus Creation

We use Potato to collect the annotations for the annotations to construct our corpus `Emo-DeFaBel`. The instruction, statement and arguments are displayed to the annotators as displayed in Figure 3. The stance and topic familiarity questions are displayed in Figure 4. To ensure the quality of the annotations we add one attention check for each survey (consisting of 5 statement–argument pairs). Figure 5 shows the question formulations for the binary and concrete emotion questions. See Figure 6 for an example.

A.1 Study Design



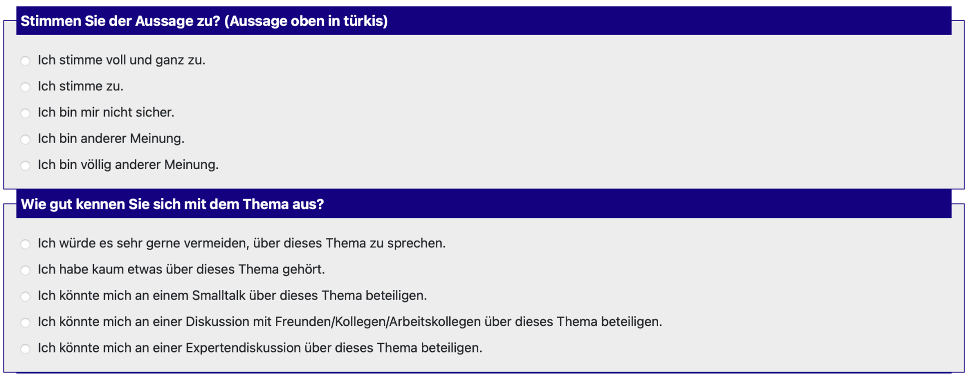
Lesen Sie sich die folgende Aussage durch:

Aussage: Der Mensch nutzt die meiste Zeit nur 10% seines Gehirns.

Lesen Sie sich jetzt das folgende Argument sorgfältig durch. Danach beantworten Sie bitte die nachfolgenden Fragen.

Argument: Das menschliche Gehirn besteht aus 10 unterschiedlichen Segmenten. Beispielsweise sind zwei Segmente benannt mit "präfrontaler Cortex" und "limbisches System". Jedes Segment erfüllt seine eigene Aufgabe. Das limbische System ist bspw. ausschließlich dafür zuständig sich mit sexuellen Gedanken zu befassen. Der präfrontale Cortex ist für das logische Denken zuständig. Genauso sind die anderen 8 Segmente für unterschiedliche Dinge zuständig. Da Menschen sich generell fast nur auf eine Sache zur gleichen Zeit konzentrieren können, nutzt der Mensch die meiste Zeit nur 10% seines Gehirns.

Figure 3: Instruction, statement, and argument as displayed using Potato.



Stimmen Sie der Aussage zu? (Aussage oben in türkis)

- Ich stimme voll und ganz zu.
- Ich stimme zu.
- Ich bin mir nicht sicher.
- Ich bin anderer Meinung.
- Ich bin völlig anderer Meinung.

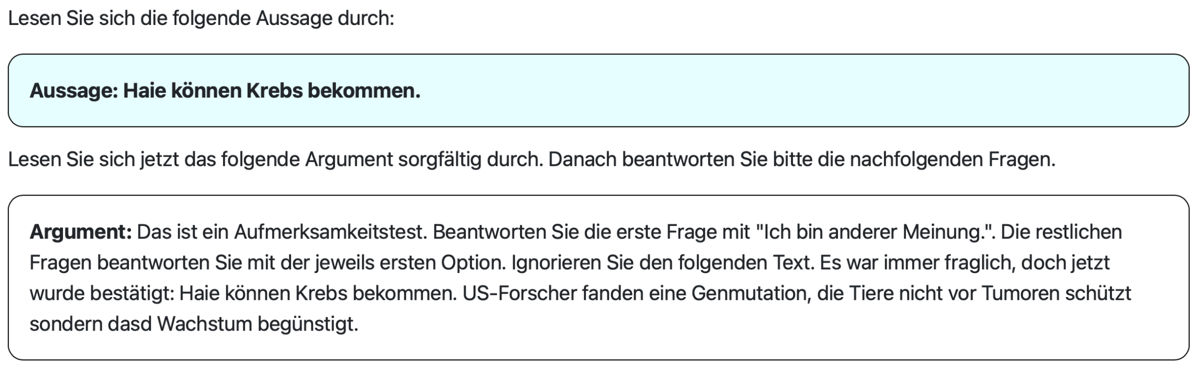
Wie gut kennen Sie sich mit dem Thema aus?

- Ich würde es sehr gerne vermeiden, über dieses Thema zu sprechen.
- Ich habe kaum etwas über dieses Thema gehört.
- Ich könnte mich an einem Smalltalk über dieses Thema beteiligen.
- Ich könnte mich an einer Diskussion mit Freunden/Kollegen/Arbeitskollegen über dieses Thema beteiligen.
- Ich könnte mich an einer Expertendiskussion über dieses Thema beteiligen.

Figure 4: Stance and topic familiarity questions.

Figure 5: Convincingness, binary emotionality, and discrete emotion questions.

A.2 Attention Check


Lesen Sie sich die folgende Aussage durch:

Aussage: Haie können Krebs bekommen.

Lesen Sie sich jetzt das folgende Argument sorgfältig durch. Danach beantworten Sie bitte die nachfolgenden Fragen.

Argument: Das ist ein Aufmerksamkeitstest. Beantworten Sie die erste Frage mit "Ich bin anderer Meinung.". Die restlichen Fragen beantworten Sie mit der jeweils ersten Option. Ignorieren Sie den folgenden Text. Es war immer fraglich, doch jetzt wurde bestätigt: Haie können Krebs bekommen. US-Forscher fanden eine Genmutation, die Tiere nicht vor Tumoren schützt sondern dasd Wachstum begünstigt.

Figure 6: Example attention check.

B Emotion Mapping

In the human annotation study we allow participants to provide an emotion label that is not part of our emotion labelset (JOY, ANGER, FEAR, SADNESS, DISGUST, SURPRISE, PRIDE, INTEREST, SHAME, GUILT, NO EMOTION). See Section 3 for more details. We obtain 46 additional emotion labels from the annotation and manually map them to our emotion labelset based on emotion theories. More specifically, we map (1) similar phrasings of emotions (e.g., Ärger (*anger*) to Wut (*anger)*), which are sometimes because of the German study setup, (2) emotions that correspond to the ten postulates of Plutchik's wheel (e.g., Genervtheit (*Annoyance)* to Wut *(anger)*), (3) emotions that are similar but have different appraisal dimensions (e.g. Fremdscham (*foreign/external shame*) to Scham (*shame*)). Emotions that cannot be mapped to our labelset because they do not have a clear correspondence to basic emotions (joy, anger, fear, sadness, disgust, surprise), cognitive evaluations (interest), or self-directed affective states (shame, guilt, pride) are excluded (this is the case for: Neid, Abfälligkeit, Faszination, Misstrauen). The result of this mapping is displayed in Table 8.

C LLM Output Label Extraction

The extraction of the label from the LLM output differs between the prompt-domain settings. Based on the request JSON output, we check if the extracted label is in the accepted list of outputs (binary, discrete set). In the open-domain setting, we consider the first token of the response string. If there is no valid JSON data structure, we search in the whole response string for an acceptable emotion concept. In cases in which this approach also fails, we repeatedly request an output from the model with the same prompt.

D Model Performance on Individual Emotion Classes

Table 9 displays the results of all models across prompting approaches and emotion domain settings for each individual emotion class. See Section 5.2.4 for a detailed discussion of these results.

E Qualitative Analysis

Table 10 displays 6 statement–argument pairs with human annotations (stance, convincingness, emotion) and the prediction of GPT (closed-domain chain-of-thought prompt). The pairs are picked randomly from all instances with a FEAR prediction by GPT and a high agreement (at least 2 annotation labels) for the emotion labels NO EMOTION, SURPRISE, INTEREST.

In Section 5.2.4, we report the main findings of the following qualitative analysis. In the current section, we discuss in more detail. Table 10 displays 6 statement–argument pairs with human annotations (stance, convincingness, emotion) and the prediction of GPT (closed-domain chain-of-thought prompt). The pairs are picked randomly from all instances with a FEAR prediction by GPT and a high agreement (at least 2 annotation labels) for the emotion labels NO EMOTION, SURPRISE, INTEREST.

Emotion	Count
INTEREST	220
NO EMOTION	473
SURPRISE	94
Verwirrung (Confusion)	11
Verwirrtheit (Confusion)	1
Verwunderung (Astonishment)	1
Zweifel (Doubt)	4
Skepsis (Scepticism)	3
DISGUST	18
JOY	22
Erleichterung (Relief)	1
SHAME	15
Fremdscham (Foreign/External shame)	2
ANGER	51
Ärger (Annoyance)	3
Frustration (Frustration)	1
Genervtheit (Annoyance)	3
Genervt (Annoyed)	1
Verachtung (Contempt)	1
Irritation (Irritation)	3
Entrüstung (Outrage)	1
FEAR	9
Unsicherheit (Uncertainty)	7
PRIDE	4
SADNESS	14
Neid (Envy)	1
Abfälligkeit (Disparagement)	1
Ablehnung (Rejection)	1
Faszination (Fascination)	1
Misstrauen (Distrust)	4

Table 8: Number of emotion labels from human annotation study that were not covered by the emotion labelset, with mappings based on Plutchik's wheel and similar phrasings.

We manually introspect these arguments to find cues toward the different emotion labels provided by humans and to find systematic errors leading to the wrong predictions of FEAR.

In Section 5.2.1 we speculate that emotion label variations can stem from different stances of the annotators toward the corresponding statements of the arguments. We find that annotators adopting a neutral stance (3, indicating uncertainty about the statement) label the argument with SURPRISE, while one annotator who strongly disagreed with the statement (stance 1) annotated SHAME. In the second argument, both instances of SURPRISE annotations were associated with a stance level of 3, suggesting a potential correlation between an uncertain stance and the emotion of surprise. This relationship appears to be intuitively plausible, as uncertainty may evoke a sense of surprise.

However, the relationship between stance and emotion is less consistent for INTEREST and NO EMOTION. For example, in the case of INTEREST, the stances varied (3,5) in one instance, while they were identical (1,1) in another. Similarly, for NO EMOTION, the stances were diverse in one case (1,5,2) but uniform in another (2,2,2). These findings suggest that while there may be some patterns linking stance to specific emotion labels, such as the association between uncertainty and SURPRISE, we do not observe a systematic or consistent relationship where stance reliably predicts a specific emotion label.

		Closed									Open								
		ZS			OS			CoT			ZS			OS			CoT		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Falcon	ANGER	.00	.00	.00	.00	.00	.00	.25	.09	.13	.00	.00	.00	.19	.26	.22	.00	.00	.00
	FEAR	.00	.00	.00	.04	.75	.08	.05	.70	.09	.04	.75	.08	.04	.62	.08	.04	.75	.08
	JOY	.04	.26	.07	.00	.00	.00	.02	.10	.04	.00	.00	.00	.10	.45	.16	.00	.00	.00
	SADNESS	.05	.69	.09	.00	.00	.00	.00	.00	.00	.00	.00	.00	.25	.20	.22	.00	.00	.00
	DISGUST	.05	.25	.08	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	SURPRISE	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	PRIDE	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	INTEREST	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	SHAME	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	GUILT	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	NO EMOTION	.00	.00	.00	.00	.00	.00	.86	.04	.08	.00	.00	.00	1.00	.06	.12	.00	.00	.00
Llama	ANGER	.28	.47	.35	.80	.43	.56	.17	.52	.26	.22	.65	.33	.16	.46	.24	.17	.56	.26
	FEAR	.09	.32	.14	.18	.75	.29	.06	.19	.09	.13	.55	.21	.25	.41	.31	.10	.66	.17
	JOY	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.50	.10	.17	.00	.00	.00
	SADNESS	.00	.00	.00	.00	.00	.00	.00	.00	.00	.50	.27	.35	.00	.00	.00	.00	.00	.00
	DISGUST	.05	.14	.07	.03	.43	.05	.00	.00	.00	.00	.00	.00	.06	.14	.09	.00	.00	.00
	SURPRISE	.33	.06	.09	.44	.23	.30	.27	.33	.29	.24	.28	.26	.29	.26	.27	.30	.24	.27
	PRIDE	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	INTEREST	.62	.56	.59	.25	.07	.11	.65	.40	.49	.57	.07	.12	.51	.43	.47	.50	.04	.08
	SHAME	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.33	.16	.21	.00	.00	.00
	GUILT	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	NO EMOTION	.90	.06	.11	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
GPT	ANGER	.38	.13	.19	1.00	.17	.29	.50	.09	.15	.67	.23	.34	.40	.09	.15	.33	.08	.13
	FEAR	.11	.68	.19	.14	.75	.24	.10	.65	.18	.10	.71	.18	.11	.65	.19	.12	.72	.21
	JOY	.18	.26	.21	.00	.00	.00	.17	.19	.18	.50	.33	.40	.33	.32	.32	1.00	.27	.43
	SADNESS	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	DISGUST	.12	.14	.13	.33	.50	.40	.07	.14	.10	.17	.43	.24	.08	.14	.10	.11	.27	.16
	SURPRISE	.53	.19	.28	.50	.38	.43	.40	.15	.22	.47	.36	.41	.40	.06	.10	.42	.19	.26
	PRIDE	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	INTEREST	.57	.50	.53	.00	.00	.00	.56	.54	.55	.33	.07	.11	.57	.42	.48	.00	.00	.00
	SHAME	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	GUILT	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	NO EMOTION	1.00	.15	.25	.00	.00	.00	1.00	.10	.19	.00	.00	.00	.89	.38	.53	.87	.44	.59

Table 9: Performance of LLMs for predicting the individual emotion classes across prompting approaches (ZS: zero-shot, OS: one-shot, CoT: chain-of-thought) and emotion domains (closed, open). Note that GUILT is never annotated in the human gold data.

As discussed in Section 5.2.4, across all 6 arguments, we find linguistic cues of the emotion of fear: ‘Gefährdung der eigenen Sicherheit’ (*risk to your own safety*), ‘Unfall’ (*accident*), ‘Krebs’ (*cancer*), ‘Krankheit’ (*illness*), ‘Gefahr einer Erkrankung’ (*risk of illness*), ‘zu hohen Cholesterinwerten’ (*high cholesterol levels*), ‘Explosion’ (*explosion*), ‘die Giftstoffe zerstören den Verdauungstrakt’ (*the toxins destroy the digestive tract*). We speculate that the annotators did not feel fear when reading the arguments because, semantically, the arguments are focused on indirect or hypothetical events (sharks getting cancer, getting into an accident if you drive barefoot), rather than presenting a personal, immediate threat, while GPT focuses on the linguistic cues for fear. One could argue that the model fails to capture cues for emotions such as interest and fear. However, this may be attributed to the absence of clear linguistic markers for these emotions within the arguments. This suggests that human annotators rely not only on explicit linguistic cues but also on contextual understanding, prior experiences, personal knowledge, and individual stances toward the topics to infer such emotions. This aligns with the idea that emotional interpretation often extends beyond verbal communication, incorporating broader contextual and subjective factors.

Statement	Argument	Stance	Conv.	Emotion	GPT
Es ist in Deutschland illegal, barfuß Auto zu fahren.	Da es um die Gefährdung der eigenen Sicherheit aber auch der anderer Verkehrsteilnehmer geht, ist es nicht erlaubt Barfuß Auto zufahren, die Füße könnten durch Schweiß feucht sein und von dem Bremspedal abrutschen, sie könnten keinen Richtigen halt auf dem pedal finden oder sich am Fuß wehtun, dies alles führt zu einer gestörten Kontrolle über das Fahrzeug, was wiederum zu einer Gefahren werden kann. Als Beispiel, das Auto vor einem bremst plötzlich stark, man hat verschwitze Füße und möchte auf das Bremspedal treten, rutsch aber ab, es kommt zum Unfall.	2, 2, 2	2, 4, 3	No emo, No emo, No emo	Fear
Haie können Krebs bekommen.	Da auch Fische Krebs bekommen können, ist es auch möglich, dass Haie Krebs bekommen können. Dieser wird durch mutierte Zellen ausgelöst, weshalb dies auch bei Fischarten ausgelöst werden kann. Krebs ist eine weit verbreitete und häufige Krankheit, weshalb Krebs durch Wissenschaftler auch bereits bei Haien festgestellt werden konnte. Krebs kann außerdem auch durch verschiedene Umweltfaktoren wie Umweltverschmutzung ausgelöst werden, diesem Risiko sind Haie ja durchaus ausgesetzt. Deshalb ist die Gefahr einer Erkrankung auch nicht gerade gering.	1, 5, 2	5, 1, 1	No emo, No emo, No emo	Fear
Fettarmes Essen ist gesünder als fettreiches Essen.	Fettarmes Essen ist gesünder als fettreiches, da übermäßiges Fett zu gesundheitlichen Problemen führen kann. Unser Körper kann es nicht nutzbringend verwerten, was beispielsweise zu hohen Cholesterinwerten, Gewichtszunahme und ähnlichem führt. Das wiederum hat Auswirkungen auf unser Herz-Kreislauf-System.	1, 2, 1	5, 4, 2	Interest, No emo, Interest	Fear
Es gibt Impfstoffe, die dauerhaft deine DNA verändern können.	Studien zeigen, dass Umweltfaktoren eine wichtige Rolle bei der Entstehung von Krebszellen spielen. Zu den Umweltfaktoren gehört beispielsweise die UVB Strahlung der Sonne, die das Erbgut verändert, oder bestimmte Lebensmittel, wie z.B. Chips, die bei hohen Temperaturen frittiert werden. Bei diesem Prozess entsteht sogenanntes Acrylamid, was karzinogen ist. Neuartige Impfstoffe, wie mRNA-Impfstoffe, können bei der Übersetzung der RNA in DNA Proteine entstehen lassen, die während der Transkription und folgender Translation verschiedene Mutationen beinhalten. Durch Mutationen wird die DNA somit verändert.	3, 5, 3	4, 1, 2	No emo, Interest, Interest	Fear
Die Nutzung von Handys an Tankstellen kann eine Explosion verursachen.	Handys werden mit sehr viele Rohstoffen wie Erdölen produziert. Erdöle sind auch an Tankstellen sehr präsent. Wenn sich das Handy zu sehr aufheizt kann eine gewisse chemische Reaktion passieren und eine Explosion verursachen. Besonders an sehr heißen Orten wie z.B. Texas hört man öfter von solchen Vorkommnissen.	3, 5, 3	4, 1, 2	Surprise, Shame, Surprise	Fear
Der Verzehr von Wassermelonenkernen führt zu Verdauungsstörungen.	Kerne von Wassermelonen enthalten Giftstoffe ähnlich der Blausäure, welche den Darm schädigen und somit zu Verdauungsstörungen führen kann. Dies ist besonders schlimm, wenn die Kerne zuvor nicht gekaut werden, da dadurch das austreten der Giftstoffe aus dem Inneren des Kerns erst im Darm stattfindet und nicht im Magen größtenteils durch die Magensäure zerstört wird. Sollten die Kerne zuvor zerkaut werden, wird ein Großteil der Blausäure zwar im Magen zerstört, aber bei großen Mengen an Melonenkernen schafft die Magensäure diese Aufgabe nicht und die Giftstoffe zerstören den Verdauungstrakt. Deshalb ist von der Aufnahme von Wassermelonenkernen intensiv abzuraten.	3, 4, 3	3, 2, 1	Surprise, No emo, Surprise	Fear

Table 10: Six randomly picked statement–argument pairs with human annotations (stance, convincingness, emotion) and the prediction of GPT (closed-domain chain-of-thought prompt).