Do Emotions Really Affect Argument Convincingness? A Dynamic Approach with LLM-based Manipulation Checks

Yanran Chen and Steffen Eger
NLLG, University of Technology Nuremberg (UTN)
https://nl2g.github.io/
{yanran.chen,steffen.eger}@utn.de

Abstract

Emotions have been shown to play a role in argument convincingness, yet this aspect is underexplored in the natural language processing (NLP) community. Unlike prior studies that use static analyses, focus on a single text domain or language, or treat emotion as just one of many factors, we introduce a dynamic framework inspired by manipulation checks commonly used in psychology and social science; leveraging LLM-based manipulation checks, this framework examines the extent to which perceived emotional intensity influences perceived convincingness. Through human evaluation of arguments across different languages, text domains, and topics, we find that in over half of cases, human judgments of convincingness remain unchanged despite variations in perceived emotional intensity; when emotions do have an impact, they more often enhance rather than weaken convincingness. We further analyze whether 11 LLMs behave like humans in the same scenario, finding that while LLMs generally mirror human patterns, they struggle to capture nuanced emotional effects in individual judgments.

1 Introduction

Emotional appeals have long been recognized as a core component of persuasion (Konat et al., 2024; Habernal and Gurevych, 2017). Aristotle's triad of logos, ethos, and pathos (Aristotle and Kennedy [translator], 1991) emphasizes the multifaceted nature of effective rhetoric. While logical reasoning (*logos*) and the speaker's credibility (*ethos*) are essential, the ability to evoke emotions in the audience (*pathos*) may also be crucial in order to make the audience more receptive to the arguments (Wachsmuth et al., 2017).

Despite active research on argumentation and argument quality in the NLP community (e.g. Habernal and Gurevych, 2016a,b; Gleize et al., 2019; Wan et al., 2024; Rescala et al., 2024; Eger et al., 2017; Wachsmuth et al., 2017, 2024), the pathos dimension has received undeservedly little attention (Evgrafova et al., 2024; Greschner and Klinger, 2024); emotional appeal is often discussed as a logical fallacy in arguments (e.g., Vijayaraghavan and Vosoughi, 2022; Goffredo et al., 2023; Li et al., 2024; Mouchel et al., 2024). Existing NLP studies exploring the interplay between emotions and *argument convincingness* often lack a specific focus on the emotional dimension and fail to control for confounding factors (e.g. Habernal and Gurevych, 2016b, 2017; Wachsmuth et al., 2017). A confounder refers to a variable that influences both

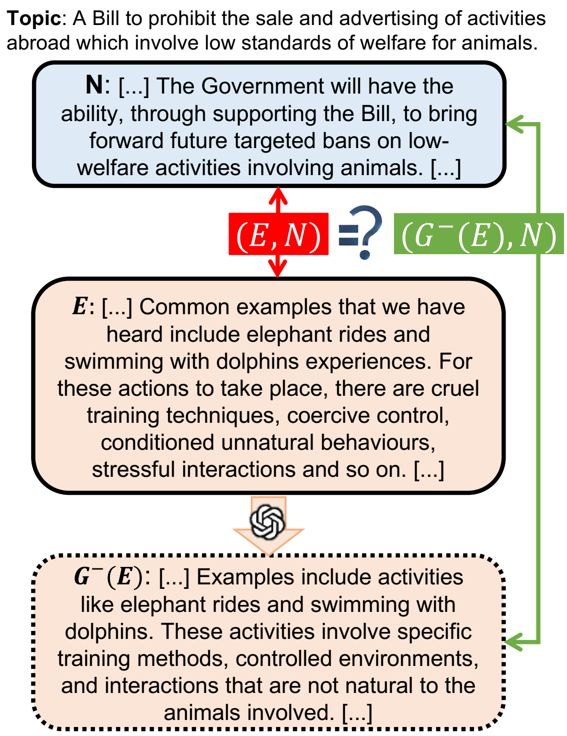

Figure 1: An example test case. E is an argument with emotions and N is an argument without emotions, both addressing the same topic with the same stance. $G^{-}(E)$ is a counterpart of E with reduced emotion. We compare the convincingness ranking of the pair (E, N) to that of the pair $(G^{-}(E), N)$ to observe the effect of emotions on argument convincingness in a dynamic way.

Findings of the Association for Computational Linguistics: ACL 2025, pages 24357–24381
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

the independent variable (the factor being manipulated: emotions) and the dependent variable (the outcome being measured: convincingness), potentially distorting the observed relationship between them. To address this gap, we propose a *dynamic* approach that systematically varies emotional intensity to observe its effect on argument convincingness, following the logic of psychological manipulation checks (Hoewe, 2017; Ejelöv and Luke, 2020). Here, emotional intensity is the manipulated variable and convincingness the dependent one. We call it dynamic because it captures the effect of varying emotional intensity on convincingness, moving beyond static comparisons between fixed argument pairs. To achieve this, we leverage LLMs to rephrase an argument to generate a counterpart that evokes stronger/weaker emotions, and then compare its convincingness to the original argument, thereby minimizing the effect of confounders. The judgments are evaluated relative to an anchor argument (§5.1), as illustrated in Figure 1, to obtain more reliable subjective human evaluation (Zhang et al., 2017; Gienapp et al., 2020; Jin et al., 2022b; Habernal and Gurevych, 2016b). This framework enables us to examine how variations in perceived emotional intensity influence judgments of convincingness for a given argument in a controlled manner. To test the robustness of our findings across annotation setups, we further use Likert-scale ratings of argument convincingness as a simpler alternative to the relative, anchor-based evaluation (§5.2). Although absolute ratings are generally considered less reliable for such subjective evaluation, our observations remain consistent, demonstrating the robustness of both our dynamic approach and findings.

Besides, we move beyond prior studies that focus predominantly on English arguments or single-domain datasets (Habernal and Gurevych, 2016b, 2017; Wachsmuth et al., 2017; Greschner and Klinger, 2024). We expand the scope to explore both English and German arguments across diverse text domains, including political debates, online portals, and curated human-written arguments. Our multilingual and cross-domain analysis provides a comprehensive view of how perceived emotional intensity affects convincingness across different contexts.

Finally, inspired by recent studies exploring cognitive biases in LLMs (Lampinen et al., 2024; Echterhoff et al., 2024; Itzhak et al., 2024; Macmillan-Scott and Musolesi, 2024), we further investigate whether LLMs behave like humans when judging argument convincingness under the influence of emotional 'bias'. Although emotion is not always considered a fallacy or bias in argumentation (Walton, 2005; Duckett, 2020; Evgrafova et al., 2024), understanding its impact on argument evaluation is crucial for developing models intended for automated argument evaluation (e.g., Wachsmuth et al., 2024; Rescala et al., 2024; Mirzakhmedova et al., 2024). **Our contributions are:**[1]

- We propose a novel framework to analyze how emotions influence perceived convincingness in a controlled manner. Our findings show that in over half of cases, human judgments remain unaffected by emotional intensity, while emotions more often enhance rather than weaken convincingness.
- We demonstrate that LLMs can effectively modify the emotional impact of arguments while preserving their original meaning, enabling precise comparisons of argument emotions.
- We conduct a multilingual, cross-domain analysis, showing that (i) when topics and domains align, emotions impact convincingness similarly in German and English, and (ii) emotions are more likely to enhance convincingness in political debates than in other domains.
- We investigate whether LLMs exhibit human-like preferences in evaluating argument convincingness, particularly regarding emotions. While they broadly mirror human patterns, they fail to capture nuanced emotional effects in individual judgments.

2 Related Work

This work primarily connects to (1) the interplay between emotions and argument convincingness, while also relating to (2) human-like biases in LLMs.

Emotion vs. convincingness Emotions have been shown to play a role in argument convincingness in both fields of computational argumentation (e.g. Habernal and Gurevych, 2016b; Wachsmuth et al., 2017; Greschner and Klinger, 2024) and philosophy/psychology (e.g. Aristotle and Kennedy [translator], 1991; Konat et al., 2024; Benlamine et al., 2015).

In NLP, emotional appeal is primarily studied within the context of logical fallacy in arguments

[1]Code+data: `https://github.com/cyr19/argument_emotion_llm_manipulation`

(Evgrafova et al., 2024) or as a secondary focus in relation to argument convincingness (Greschner and Klinger, 2024). The most relevant works include: Habernal and Gurevych (2016b) find that human annotators identify emotional aspects as positively contributing to argument convincingness. Habernal and Gurevych (2017) introduce an emotional appeal layer in a modified Toulmin argumentation model, showing that 6% of arguments are purely emotional. Wachsmuth et al. (2017) analyze arguments across 15 dimensions, finding a weak positive correlation between emotional appeal and convincingness. Lukin et al. (2017) demonstrate that audience-specific factors improve belief change prediction, particularly for emotional arguments. Greschner and Klinger (2024) examine specific emotions, showing that arguments expressing joy and pride are rated as more convincing, while those expressing anger are rated as less convincing.

Previous studies rely on fixed analyses that do not control for confounders. In contrast, we adopt a dynamic approach, controlling for confounding factors and examining how perceived convincingness changes with varying emotional intensity. Our methodology aligns with psychological manipulation checks (Hoewe, 2017; Ejelöv and Luke, 2020), treating emotional intensity as the manipulated variable and convincingness as the dependent variable.

Additionally, prior work has largely focused on English, except for Greschner and Klinger (2024), who examine German arguments. Since emotional effects may vary across cultures, we study both English and German arguments. We also expand the scope by incorporating diverse text domains, including political debates, online portals, and curated human-written arguments, unlike previous studies limited to a single domain.

Human-like biases in LLMs An array of studies has demonstrated human-like biases in LLMs (e.g., Liang et al., 2021; Echterhoff et al., 2024; Itzhak et al., 2024). Social biases, such as sentiment, stereotype, and gender biases, have been extensively investigated (e.g., Huang et al., 2020; Nadeem et al., 2021; Kotek et al., 2023; Viswanath and Zhang, 2023).

Beyond social biases, LLMs also mimic human cognitive biases in reasoning and decision-making (Lampinen et al., 2024; Hagendorff et al., 2023; Talboy and Fuller, 2023; Echterhoff et al., 2024; Itzhak et al., 2024; Sumita et al., 2024; Macmillan-Scott and Musolesi, 2024). For instance, Lampinen et al. (2024) show that LLMs, like humans, perform better when task semantics align with logical inference (‘content effect’). Similarly, Echterhoff et al. (2024) find LLMs exhibit decision-making biases such as anchoring bias (Tversky and Kahneman, 1974), status quo bias (Samuelson and Zeckhauser, 1988), and framing bias (Tversky and Kahneman, 1974). Meanwhile, Macmillan-Scott and Musolesi (2024) analyze LLMs’ irrationality across 12 cognitive tasks (Kahneman and Tversky, 1972; Bruckmaier et al., 2021), revealing both human-like errors and distinct deviations.

Although emotional appeal is not inherently a bias or fallacy but a persuasion strategy, it is crucial to examine whether LLMs’ preferences align with human judgments, especially given their growing role in argument evaluation (e.g., Wachsmuth et al., 2024; Rescala et al., 2024; Mirzakhmedova et al., 2024). Inspired by studies on cognitive biases in LLMs, we investigate whether LLMs exhibit human-like behavior in how emotional intensity influences argument convincingness.

3 Evaluation Setup

We employ a dynamic framework to explore how the intensity of emotions evoked in readers impacts their judgments of convincingness. In this work, **we treat emotional intensity as the overall strength of emotions felt by readers, without considering specific emotions**. We follow previous works (Habernal and Gurevych, 2016b; Toledo et al., 2019) to leverage pairwise comparisons for evaluation because it yields more reliable annotations compared to the absolute ratings, especially for such subjective evaluation tasks (Zhang et al., 2017; Jin et al., 2022a; Gienapp et al., 2020).

Our setting is as Figure 1 shows: among one pair of arguments that share *the same stance* on a given topic but *differ in their content*, $\boldsymbol{E}$ is (set up to be) **emotion-evoking**, while $\boldsymbol{N}$ **does not (typically) evoke emotions**. We then use LLMs to generate a counterpart argument for E, $\boldsymbol{G^{-}(E)}$, which retains the same meaning as E but **evokes less emotion**. To inspect how perceived argument convincingness is affected by emotions, we compare the convincingness ranking of ($G^{-}(E)$,N) to that of the original pair (E,N). The reason to not compare the arguments with a similar content, i.e., E vs. $G^{-}(E)$ is that we want to minimize the effect of human’s prior belief about whether emotions should contribute to argument convincingness. Analogously, we generate a counterpart for N, $\boldsymbol{G^{+}(N)}$, with **in-**

Argument Pair	Convincingness Ranking						
Anchor: (E, N)	$>$			$=$			$<$
$(G^{-}(E), N)$	$>$	$\leq$	$>$	$=$	$<$	$\geq$	$<$
$(E, G^{+}(N))$	$>$	$\leq$	$>$	$=$	$<$	$\geq$	$<$
$(G^{-}(E), G^{+}(N))$	$>$	$\leq$	$>$	$=$	$<$	$\geq$	$<$

Table 1: All convincingness change scenarios. Cells marked in green indicate positive cases, red indicates negative cases, and consistent cases are left with a white background. Math relation symbols $>, <, =$ refer to convincingness.

creased emotional intensity and observe how the convincingness ranking changes from (E, N) to (E, $G^{+}(N)$). Finally, we include the fully LLM-generated pair ($G^{-}(E)$, $G^{+}(N)$) in our evaluation.

The goal of G^{+}/G^{-} is to maintain the core meaning of the argument while modifying its emotional appeal. Although humans could be used to create such counterparts (e.g. Huffaker et al., 2020; Velutharambath et al., 2024), this approach is largely impractical at scale because it is costly. Instead, we use LLMs to efficiently generate required variations and assess their ability to perform this task through human evaluation.

Thus, **for each original argument pair** (E, N), **we create three counterpart pairs** with varying levels of emotional intensity, resulting in a total of four argument pairs per test instance. The original argument pair serves as the anchor, from which we see how the convincingness rankings of the other argument pairs change. We list all possible change scenarios in Table 1 and divide them into three categories: 1: (1) **Consistent**: convincingness ranking *does not change* with varying emotional intensities. (2) **Positive**: an argument is perceived as more/less convincing when it evokes stronger/weaker emotions (convincingness and emotionality have the same directionality). (3) **Negative**: an argument is perceived as more/less convincing when it evokes weaker/stronger emotions, and less convincing when it evokes stronger emotions (convincingness and emotionality have the opposite directionality).

The first row in the table presents all possible convincingness rankings of the original argument pair (E,N). The subsequent rows show the convincingness rankings of the counterpart argument pairs where the emotional intensity of the argument on the left has been reduced ($G^{-}(E)$,N), that of the argument on the right has been increased (E,$G^{+}(N)$), or both ($G^{-}(E)$,$G^{+}(N)$). Cells highlighted in green indicate cases where the convincingness of the left argument decreases as its emotional intensity decreases *relative* to the right argument, suggesting a **positive** impact of emotions on convincingness. This occurs when the convincingness ranking shifts from the left being $>$ to $\leq$ the right argument, or from being $=$ to $<$ the right argument. Conversely, cells highlighted in red indicate cases where the convincingness of the left argument increases as its emotional intensity decreases *relative* to the right argument, reflecting a **negative** impact of emotions on convincingness. Finally, cases where the convincingness rankings remain **consistent** retain a white background.

Metrics For each instance (E, N), we calculate the percentages of consistent, positive, and negative cases. We then average the percentages of each category across all test instances to derive three metrics that indicate the overall frequencies of the three categories in humans. We call the metrics: **consistency rate**, **positivity rate**, and **negativity rate**. Their formulas are as follows:

$$Rate_{category} = \frac{1}{n}\sum_{i=1}^{n}\frac{C_{category,i}}{3}, \qquad (1)$$

where n is the total number of test instances, $C_{category}$ is the count of cases in the specified category for the i-th instance, and $category \in \{\text{consistent, positive, negative}\}$.

4 Dataset Construction

We source **50 anchor argument pairs** from each of five datasets (§4.1) and utilize GPT4o[2] to generate their **counterparts** with variations in emotional intensity (§4.2).

4.1 Anchor: *E* & *N*

We leverage two established datasets which have human annotations for argument convincingness and emotions, $\texttt{Dagstuhl}_{\texttt{en}}$ (Wachsmuth et al., 2017) and $\texttt{EmoDefabel}_{\texttt{de}}$ (Greschner and Klinger, 2024). Besides, we create three datasets ourselves from political debates, $\texttt{Bill}_{\texttt{en}}$, $\texttt{Hansard}_{\texttt{en}}$, and $\texttt{DeuParl}_{\texttt{de}}$, since emotional appeal is a common strategy used by politicians to influence perceptions and decisions (Brader, 2005); this domain is therefore expected to be rich in emotional content. From each data source, we select 50 argument pairs where ***E* is more likely and *N* is less likely to evoke emotions**. The subscripts in the dataset names indicate the language: 'en' for English and

[2] https://openai.com/index/gpt-4o-system-card/

'de' for German. In the following, we describe how we extract argument pairs from each data source.

4.1.1 Arguments from Political Debates

We **crawl** parliamentary debates for `Hansard`$_{en}$ from the UK Hansard[3] and for `DeuParl`$_{de}$[4] from the German Bundestagsprotokolle.[5] The datasets cover the past 3–5 years.[6] We heuristically **segment** each speech into balanced-length paragraphs. The original crawled texts are divided by double line breaks. If a paragraph has fewer than 60 tokens or the next one has fewer than 20 tokens or starts with a left bracket, we merge them cumulatively. From these processed paragraphs, we select argumentative texts for evaluation.

In our **pilot annotations** with `Hansard`$_{en}$, we find that within a single debate on a broad topic, diverse subtopics make it difficult to pair arguments with the same topic. Additionally, the interactive nature of debates complicates determining a paragraph's focus without context. To address this, we first conduct **pre-annotation** on a small scale for five *Second Reading debates of Bills* relevant to family and animals,[7] which are easier to annotate because the Bill debated provides a clear topic. We then refine GPT4o prompts to develop **classifiers** for identifying argument pairs that share a topic and stance but differ in emotional appeal. The final classifiers achieve precisions of 0.80 (English) and 0.76 (German) for detecting topic-aligned arguments and a macro F1 of ∼0.75 for distinguishing emotional from non-emotional arguments. See Appendix A for details.

`Bill`$_{en}$ From the argument pairs labeled as having the same topic and stance during the pre-annotation phase, we randomly sample 50 pairs, with one argument labeled as emotion-evoking and the other as non-emotion-evoking. The topic for each argument pair is the brief introduction of the Bill crawled.

`Hansard`$_{en}$ & `DeuParl`$_{de}$ Debates are filtered using pre-selected keywords related to recent wars, refugee crises, and migration (see Table 7 in the appendix for the full list), as these highly debated topics are likely to evoke strong emotions. For `Hansard`$_{en}$, we retain debates whose titles contain these keywords. For `DeuParl`$_{de}$, we include debates whose introductions mention the keywords. Finally, an annotator from the pre-annotation phase selects 50 argument pairs from the candidates selected out by the GPT4o classifiers for both `Hansard`$_{en}$ and `DeuParl`$_{de}$. These argument pairs are **manually verified** to meet our criteria — both arguments address the same topic with the same stance but differ in their emotional aspect. A human-written topic is assigned to each pair.

4.1.2 Arguments from others

We randomly select 50 argument pairs from each of **`Dagstuhl`$_{en}$** and **`EmoDefabel`$_{de}$** that meet our criteria, based on the emotion annotations in the original works. See Appendix B for details.

4.2 Counterpart: $G^{-}(E)$ & $G^{+}(N)$

We leverage **GPT4o**[8] to synthesize our counterpart arguments, namely $\boldsymbol{G^{-}(E)}$ and $\boldsymbol{G^{+}(N)}$. Specifically, we prompt GPT4o (zero-shot) to either introduce or remove emotions by **rephrasing** the original arguments, using the prompts listed in Table 8 (appendix), since we aim for counterpart arguments that convey the same information as the original ones. During generation, if the output does not receive the expected label from the binary emotion classifiers used in §4.1.1, the process is repeated for up to five rounds.

We randomly sample five argument pairs (original + synthetic) for each direction (introducing or removing emotions) from each dataset, totaling 50 argument pairs for content preservation **evaluation**. Each pair is rated by three crowdworkers for content similarity on a Likert scale of 1–5. The pairs receive an average score of 4.5, where 4 denotes 'Same Claims, Minor Content Differences' (minor details differ, but no major evidence changes), and 5 represents 'Identical Content, Different Style/Tone' (only rhetorical or emotional differences). Thus, we conclude that the main message is well preserved throughout the process. The effectiveness of adjusting emotional appeal is further evaluated in our primary human study (§5.1).

[3] `https://hansard.parliament.uk/`

[4] We name it DeuParl following previous studies leveraging this corpus (e.g. Walter et al., 2021; Kostikova et al., 2024; Chen et al., 2024).

[5] `https://www.bundestag.de/protokolle`

[6] Hansard: 2022/01/05-2024/07/19; German Bundestagsprotokolle: 2020/01/15-2024/09/27

[7] `https://www.parliament.uk/about/how/laws/passage-bill/commons/coms-commons-second-reading/`

[8] We used the version 'gpt-4o-2024-08-06' with a temperature of 0.6 and a top_p of 0.9 for GPT4o. The randomness was set to a moderate level to balance creativity and consistency, as the task involves generating content similar to creative writing while ensuring the meaning of the original argument is preserved.

Dataset	Lang	#Instances	#Pairs	#Arguments	#Tokens	#Sents	Domain	Topics
$\texttt{Bill}_{en}$	en	50	200	128	147.4	6.1	Parliamentary debates	Bills related to family and animals
$\texttt{Hansard}_{en}$	en	50	200	154	159.3	6.4	Parliamentary debates	Refugees, wars, migrants
$\texttt{Dagstuhl}_{en}$	en	50	200	128	86.8	4.5	Online portal	-
$\texttt{DeuParl}_{de}$	de	50	200	126	144.3	7.4	Parliamentary debates	Refugees, wars, migrants
$\texttt{EmoDefabel}_{de}$	de	50	200	160	92.8	4.5	Curated human-written arguments	Health, law, finance and politics
Total/Average	-	250	1,000	696	126.1	5.7	-	-

Table 2: Metadata of datasets used in this work. **Left**: number of test instances, argument pairs, and unique arguments. **Middle**: average number of tokens and sentences per argument, measured with the Stanza tokenzier (Qi et al., 2020). **Right**: domains and topics of the datasets.

4.3 Final Datasets

Our final datasets comprise 250 test instances, each consisting of one original argument pair and three counterpart pairs. The datasets include both English and German texts, spanning various domains and topics. The metadata of the datasets is summarized in Table 2.

5 Human Annotation

We randomly divide the 50 instances (200 argument pairs) from each dataset into 10 batches, each with 5 instances (20 argument pairs). Every batch is annotated by 5 individuals. One annotator evaluates at least one batch, allowing us to calculate inter-annotator agreements and base observations on individual annotators. Although our primary focus is on how convincingness rankings change, we also include comparisons of emotional intensity to evaluate whether GPT4o adjusts the emotional appeal of arguments as intended.

Annotators compare emotions and convincingness of one argument pair by answering two **subjective** questions: (i) **Convincingness**: *Which argumentative text do you find more convincing?* (ii) **Emotion**: *Which argumentative text evokes stronger emotions in you?* **Equivocal judgments** are allowed, i.e., annotators can judge both arguments as equally convincing or evoking an equal level of emotion. During annotation, argument pairs are shown with their topics. See Appendix D for screenshots of the annotation interface.

Annotators We hire annotators from two sources: university students and the crowdsourcing platform Prolific:[9]

- **Student**: 4 students are hired for this task. All annotators possess fluent to native-level proficiency in the languages of the evaluated arguments and are all based in Germany. One of them is a PhD student, and the others are Master's students. Three of them are involved in the pre-annotation phase to select out the needed argument pairs.
- **Crowdsourcing**: As our dataset includes arguments from political debates, we assume native speakers in the corresponding countries provide more reliable annotations. Thus, we use Prolific's **prescreening** to select native English/German speakers in the UK/Germany. Furthermore, to filter out individuals who may randomly fill in their profiles, participants are asked to re-rate their language proficiency, and those with inconsistent responses are **screened out** from the tasks. We also include **three attention checks** by randomly inserting instruction sentences, such as 'select the answer whose first number equals three minus two', into the arguments. Overall, 38% of the crowdworkers fail at least two attention checks, and their submissions are excluded from our analysis. This process is repeated iteratively until we obtain sufficient submissions for each batch.

[9]https://www.prolific.com/

	#Annotators		Agreements					
			EMO			CONV		
	S	C	α	Full	Maj.	α	Full	Maj.
$\texttt{Dagstuhl}_{en}$	1	4	0.506	6.5%	74.5%	0.540	14.0%	80.0%
$\texttt{Bill}_{en}$	1	4	0.449	7.0%	76.5%	0.463	10.5%	78.0%
$\texttt{Hansard}_{en}$	1	4	0.361	0.5%	68.0%	0.371	6.0%	75.0%
$\texttt{EmoDefabel}_{de}$	2	3	0.729	13.5%	87.5%	0.607	16.0%	82.0%
$\texttt{DeuParl}_{de}$	3	2	0.352	8.0%	80.5%	0.364	4.5%	74.5%
Avg	-	-	0.479	7.1%	77.4%	0.469	10.2%	77.9%

Table 3: **Left**: Number of student (S) and crowdsourcing (C) annotators per batch. **Right**: Krippendorf's α for the most agreeing annotator pairs (α), the percentages of annotation instances where all annotators agree on a certain label (**Full**), and the percentage of annotation instances where at least three annotators agree on a certain label (**Maj.**).

We summarize the number of student and crowdsourcing annotators for each dataset in Table 3 (left side); the values indicate the total annotators involved in annotating each batch. The total annotation cost is around 1,500 Euros.

Inter-annotator agreement While we acknowledge the inherent subjectivity in evaluating emotion and convincingness, we report inter-annotator agreement to present the level of consistency in these evaluations of emotional intensity (EMO) and convincingness (CONV). Following Wachsmuth et al. (2017), in Table 3 (right), we report the Krippendorf's α agreement (Castro, 2017) for the most agreeing annotator pairs[10] (column 'α'), the percentages of annotation instances where all annotators agree on a certain label (column 'Full'), and the percentages of annotation instances yielding a valid majority vote (column 'Maj.'). The agreement among the most agreeing annotator pairs ranges from 0.352 to 0.729 for EMO and from 0.364 to 0.607 for CONV. Full agreements are only up to 16.0%, while majority agreements range from moderate to high across different datasets, with 68% to 87.5% for EMO and 62% to 85% for CONV. This suggests a decent level of annotation agreement, considering that Wachsmuth et al. (2017) reported 94.4% majority agreement and a Krippendorff's α of 0.26–0.45 for the most agreeing annotator pairs when evaluating emotional appeal and argument effectiveness on a Likert scale of 1–3; both tasks can also be seen as three-way classifications similar to ours but involved only three annotators. However, we note that when computing agreement in a more standard way — by averaging across all annotators and batches — the agreement decreases to around 0.2 Krippendorff's α for both criteria. To validate the robustness of our findings, we conducted an additional annotation study on a small subset of data using a completely different setup and explore whether we can draw similar conclusions in §5.2.

5.1 Evaluation Results

Effectiveness of GPT4o in adjusting emotional appeal We evaluate whether $G^{-}(E)$ evokes weaker emotion than E and whether $G^{+}(N)$ evokes stronger emotion than N, as intended. To do so, we compute best-worst scaling (BWS) scores for each of the four argument groups based on emotion comparison annotations. Majority votes from the five annotators are used; if none exists, equivalent judgments are considered. While arguments with similar content (e.g., E vs. $G^{-}(E)$) are not directly compared with each other, both are evaluated against the other two arguments within the same instance, making the BWS-based comparison between the arguments within the same content meaningful. Higher scores reflect greater perceived emotional intensity. Table 9 (appendix) presents the BWS scores, showing that E consistently scores higher than $G^{-}(E)$ and N lower than $G^{+}(N)$ across datasets. This suggests that *GPT4o is overall effective in modifying arguments to be more or less emotion-evoking as intended*, supporting the premise for analyzing changes in convincingness rankings.

[10]We average the agreements of the most agreeing annotators over batches per dataset since our sample size for calculating agreements is much smaller than Wachsmuth et al. (2017) (20 vs. 320).

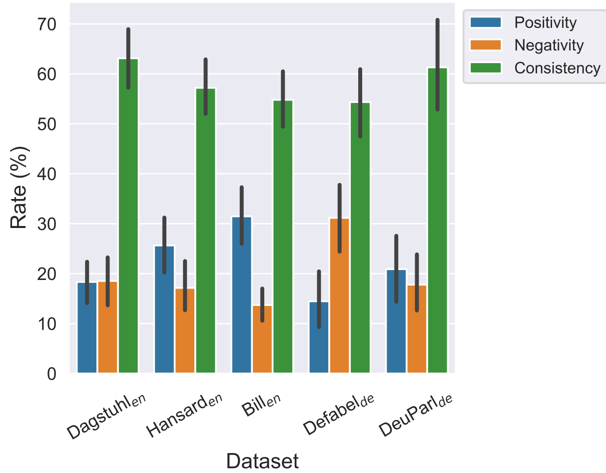

Figure 2: Consistency, positivity, and negativity rates of human judgments on convincingness.

Do emotions really affect convincingness? Figure 2 illustrates the consistency, positivity, and negativity rates. We present the averages across individual annotators, with error bars representing 95% confidence intervals. While the metrics vary across datasets and domains, we observe that consistency achieves the highest rates consistently across datasets, roughly ranging from 54% to 62%. This indicates that, **in more than half of the cases, humans are not influenced by variations in perceived emotions when judging convincingness**. In political debate datasets — $\texttt{Hansard}_{\texttt{en}}$, $\texttt{Bill}_{\texttt{en}}$, and $\texttt{DeuParl}_{\texttt{de}}$ — positive rates are consistently higher than negative rates, averaging an 8-percentage-point difference. In contrast, in $\texttt{Dagstuhl}_{\texttt{en}}$, positive and negative rates are roughly equal ($\sim$18%), whereas in $\texttt{EmoDefabel}_{\texttt{de}}$, negative rates dramatically exceed positive rates (30% vs. 14%). These differences may be attributed to variations in dataset domains and argument topics. In Appendix D, we show examples where emotions have positive/negative impacts on argument convincingness from $\texttt{Hansard}_{\texttt{en}}$/$\texttt{EmoDefabel}_{\texttt{de}}$; in $\texttt{EmoDefabel}_{\texttt{de}}$, topics often require more factual evidence, making emotions less influential or even

Model Family	Checkpoint	Size
OpenAI	gpt-3.5-turbo	-
(OpenAI et al., 2024)	gpt-4o-mini	-
	gpt-4o-2024-08-06	-
Llama3	Llama-3.2-1B-Instruct	1B
(Grattafiori et al., 2024)	Llama-3.2-3B-Instruct	3B
	Llama-3.3-70B-Instruct	70B
Qwen2.5	Qwen2.5-0.5B	0.5B
(Yang et al., 2024)	Qwen2.5-7B-Instruct	7B
	Qwen2.5-72B-Instruct	72B
Mistral		
(Jiang et al., 2023)	Mistral-7B-Instruct-v0.3	7B
(Jiang et al., 2024)	Mixtral-8x7B-Instruct-v0.1	47B

Table 4: LLMs used in this work.

detrimental. Finally, we observe slight differences between the English and German datasets, using $\texttt{Hansard}_{\texttt{en}}$ and $\texttt{DeuParl}_{\texttt{de}}$ as examples, where argument topics are similar and both originate from political debates: the rates are overall comparable, with German being less affected by emotions (consistency rates: 60% vs. 56%) and also less positively influenced by emotions (positivity rates: 20% vs. 25%) compared to English.

5.2 Robustness of Our Findings

Although our annotation agreements are comparable to prior work, they remain relatively low compared to typical human annotation, as shown earlier, likely due to the subjective nature of the criteria. To examine the robustness of our conclusions under different conditions, we conducted an additional annotation study using a completely different setup: annotators assessed the convincingness of individual arguments independently on a Likert scale of 1-5, without pairwise comparisons. We randomly selected two batches (10 instances each) from four datasets — $\texttt{Hansard}_{\texttt{en}}$, $\texttt{DeuParl}_{\texttt{de}}$, $\texttt{Dagstuhl}_{\texttt{en}}$, and $\texttt{EmoDefabel}_{\texttt{de}}$ — resulting in 120 test cases (4 datasets × 10 instances × 3 test cases per instance). Each argument was rated by five crowdworkers from Prolific, and we used the average rating as the final convincingness score.

The results reinforce our earlier findings: (1) Emotional content often enhances, rather than degrades, argument convincingness (positive:negative = 68:52); and (2) in over half of the cases, emotions do not substantially influence convincingness. Because it is uncommon for two arguments to receive identical average scores, we define a threshold to determine when a difference in convincingness scores between two related arguments is meaningful. With a threshold of 1 (the full scale interval), the consistency rate is 81.6%; with a threshold of 0.5 (the midpoint for rounding), it drops to 57.5% — closely aligning with the main results reported in the paper. Overall, this supplemental evaluation supports the robustness of our main conclusions.

6 Do LLMs Behave Like Humans?

Models We select a range of recent LLMs, including both open-source and commercial models, with varying model sizes from 0.5B to 72B parameters. We experiment with **11 LLMs from 4 model families**, as detailed in Table 4. For OpenAI models, we utilize the official API,[11] while for open-source models, we retrieve checkpoints from HuggingFace.[12] For all models, we set the temperature to 0.6 and the top-p value to 0.9, to ensure diverse outputs that still remain contextually relevant and logical, running each model five times. For 70B/72B models, we use 4-bit quantization. We run the models on 1 to 8 A40 GPUs, each with 48GB of memory.

Prompts We use **three prompt templates** to prompt LLMs to compare perceived convincingness, mirroring human instructions. The final judgment is determined by a majority vote from the five runs; if none is reached, the arguments are considered equally convincing. **Zero-shot prompts** are employed to minimize biasing effects on model responses (Paech, 2024) and thus better capture the models' intrinsic behavior. As shown in Table 10 (Appendix C), *Prompt 1* instructs models to provide a label without explanation. *Prompt 2 and Prompt 3* additionally require an explanation and include an example answer to specify the response format. To examine potential biases from examples, they feature opposite label choices and differ in perspective, with Prompt 2 favoring an objective approach and Prompt 3 adopting a more subjective and emotional stance.

LLMs exhibit a similar sensitivity to emotions when judging argument convincingness. Figure 3 presents the consistency, positivity, and negativity rates of LLMs' convincingness judgments, averaged across prompts and instances in all datasets. Like humans, *LLMs show a strong tendency toward consistency*, with rates consistently exceeding positivity and negativity (∼48%-68% vs. ∼10%-36%); all models except Qwen2.5-0.5B,

[11] https://platform.openai.com/
[12] https://huggingface.co/

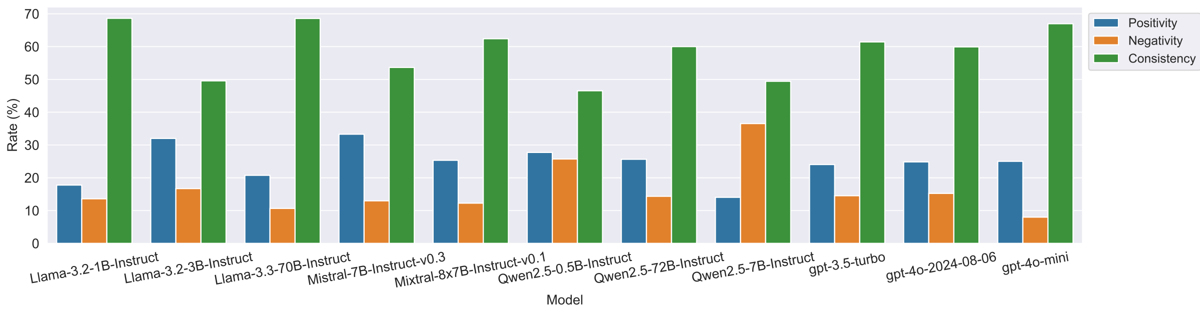

Figure 3: Consistency, positivity, and negativity rates of LLMs' judgments on convincingness, averaged across prompts and instances in all datasets.

Llama-3.2-3B, and Qwen2.5-7B achieve a consistency rate above 50%; Moreover, *emotions more often enhance rather than degrade convincingness*, except for Qwen2.5-7B, aligning with human patterns. As shown in Figure 5 (Appendix F), most models exhibit comparable rates across different prompts, except for the smallest model, Qwen2.5-0.5B, and gpt-3.5-turbo, where negativity surpasses positivity with Prompt 2. In Prompt 2's example, logical fallacy is mentioned in the explanation, which may (mis)lead models to interpret emotions as a logical fallacy.

However, they do not align well with humans on individual judgements. Table 13 in Appendix F displays macro F1 scores and model rankings for LLMs in predicting argument pair convincingness rankings (column 'Static') and the resulting categories of emotional effect (column 'Dynamic') in English and German. The best prompt result of each model is reported to demonstrate its potential. Human and LLM labels are determined by majority votes from different annotators and runs, respectively. Overall, all scores remain low (∼0.32–0.49), indicating performance ranging from random to slightly above random in a three-way classification task. GPT4o consistently ranks first in three of four tasks, except for dynamic label prediction in English, where it ranks second. Larger models generally align better with humans, often achieving higher F1 scores than their smaller counterparts, with the largest models (GPT4o, Llama-3.3-70B, Qwen2.5-72B) frequently ranking among the top.

7 Conclusion

In this work, we examined how emotional intensity influences perceived convincingness. Using GPT4o to rephrase arguments with varying emotional impact, we developed a dynamic framework inspired by manipulation checks in psychology and social sciences. Our results show that GPT4o reliably generates counterpart arguments, preserving meaning while altering emotional tone. For both humans and LLMs, convincingness is largely unaffected by emotions. However, when emotions do play a role, they more often enhance rather than weaken convincingness, particularly in political debates, where emotional appeal is frequently used as a persuasive strategy. Additionally, while LLMs broadly mirror human patterns, they struggle to capture emotional nuances.

Future research could explore **when and how** emotions influence convincingness across argument types. Investigating **specific emotions** (Greschner and Klinger, 2024) or **justified vs. unjustified emotions** and their persuasive effects may provide deeper insights. Enhancing LLMs' ability to capture emotional nuances through improved prompts or fine-tuning could further strengthen their reliability in evaluating emotional arguments.

Limitations & Ethical concerns

While our study provides insights into the relationship between emotional intensity and argument convincingness, several limitations should be acknowledged: (1) We rely on a single model, GPT4o, for synthetic argument generation. While GPT4o demonstrates strong capabilities in controlled text modification, exploring multiple models could provide a more comprehensive understanding of how different architectures handle emotional rephrasing. (2) We focus only on two languages, English and German. Expanding to additional languages, particularly those with different rhetorical traditions or cultural perspectives on emotional persuasion, would offer a broader cross-linguistic perspective. (3) The topics of arguments differ across text domains, which may introduce variability in how emotional intensity interacts with convincingness. Ensuring more comparable topics across domains would help isolate the individual effects of topic and text domain, leading to a more precise analysis. (4) The dataset is relatively small, and the annotation agreement is low, which may limit the generalizability of our findings. However, with an additional annotation study (§5.2), we were able to replicate the main observations. (5) We do not distinguish between different types of emotions (e.g., anger, joy, fear) or between justified and unjustified emotions, both of which could have varying impacts on argument convincingness. Future work could explore how different kinds of emotions influence persuasion to gain a more nuanced understanding of their effects. (6) We experiment with only three prompts to evaluate model responses, which may not fully reflect LLM performance. A broader range of prompts could yield more stable results.

A potential ethical concern arises from the possibility of leveraging the dataset to develop politically motivated agendas that rely on emotional appeal rather than factual reasoning. Since emotions can influence perceived convincingness, there is a risk that political actors or interest groups may use this dataset to craft emotionally charged arguments that manipulate public opinion rather than inform it. This could contribute to misinformation, polarization, and biased discourse, particularly in sensitive political debates.

We used ChatGPT solely for text refinement while writing this paper. All annotators provided consent for research use of their annotations via Google Forms.

Acknowledgements

We thank the anonymous reviewers for their thoughtful feedback, which helped improve the paper. We also thank Jonas Belouadi, Daniil Larionov, Aida Kostikova, and Sotaro Takeshita for their valuable comments on the initial version of this paper, which greatly strengthened the work. The NLLG Lab gratefully acknowledges support from the Federal Ministry of Education and Research (BMBF) via the research grant "Metrics4NLG" and the German Research Foundation (DFG) via the Heisenberg Grant EG 375/5-1. This project has been conducted as part of the EMCONA (The Interplay of Emotions and Convincingness in Arguments) project, which is funded by the DFG (project EG-375/9-1). We also thank Roman Klinger and Lynn Greschner for their valuable feedback in the context of the EMCONA grant, which supports our collaborative research papers.

References

Ott Rhetoric Aristotle and George A Kennedy [translator]. 1991. A theory of civic discourse. *2nd éd.*

Sahbi 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, IJCAI 2015*, pages 156–163.

Ted Brader. 2005. Striking a responsive chord: How political ads motivate and persuade voters by appealing to emotions. *American Journal of Political Science*, 49(2):388–405.

Georg Bruckmaier, Stefan Krauss, Karin Binder, Sven Hilbert, and Martin Brunner. 2021. Tversky and kahneman's cognitive illusions: who can solve them, and why? *Frontiers in Psychology*, 12:584689.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. `https://github.com/pln-fing-udelar/fast-krippendorff`.

Yanran Chen, Wei Zhao, Anne Breitbarth, Manuel Stoeckel, Alexander Mehler, and Steffen Eger. 2024. Syntactic language change in english and german: Metrics, parsers, and convergences.

Stephen Duckett. 2020. Pathos, death talk and palliative care in the assisted dying debate in victoria, australia. *Mortality*, 25(2):151–166.

Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12640–12653, Miami, Florida, USA. Association for Computational Linguistics.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Emma Ejelöv and Timothy J Luke. 2020. "rarely safe to assume": Evaluating the use and interpretation of manipulation checks in experimental social psychology. *Journal of Experimental Social Psychology*, 87:103937.

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.

Lukas Gienapp, Benno Stein, Matthias Hagen, and Martin Potthast. 2020. Efficient pairwise annotation of argument quality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5772–5781, Online. Association for Computational Linguistics.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.

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. 2024. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and llms.

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.

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

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

Jennifer Hoewe. 2017. Manipulation check. *The international encyclopedia of communication research methods*, pages 1–5.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Jordan S Huffaker, Jonathan K Kummerfeld, Walter S Lasecki, and Mark S Ackerman. 2020. Crowdsourced detection of emotionally manipulative language. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785.

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

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022a. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(1):155–205.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022b. 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.

Daniel Kahneman and Amos Tversky. 1972. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454.

Barbara Konat, Ewelina Gajewska, and Wiktoria Rossa. 2024. Pathos in natural language argumentation: Emotional appeals and reactions. *Argumentation*, pages 1–35.

Aida Kostikova, Dominik Beese, Benjamin Paassen, Ole Pütz, Gregor Wiedemann, and Steffen Eger. 2024. Fine-grained detection of solidarity for women and migrants in 155 years of German parliamentary debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5884–5907, Miami, Florida, USA. Association for Computational Linguistics.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.

Yanda Li, Dixuan Wang, Jiaqing Liang, Guochao Jiang, Qianyu He, Yanghua Xiao, and Deqing Yang. 2024. Reason from fallacy: Enhancing large language models' logical reasoning through logical fallacy understanding. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3053–3066, Mexico City, Mexico. Association for Computational Linguistics.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

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.

Olivia Macmillan-Scott and Mirco Musolesi. 2024. (ir) rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6):240255.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? In *Conference on Advances in Robust Argumentation Machines*, pages 129–146. Springer.

Luca Mouchel, Debjit Paul, Shaobo Cui, Robert West, Antoine Bosselut, and Boi Faltings. 2024. A logical fallacy-informed framework for argument generation.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. 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.

Samuel J. Paech. 2024. Eq-bench: An emotional intelligence benchmark for large language models.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Paula Rescala, Manoel Horta Ribeiro, Tiancheng Hu, and Robert West. 2024. Can language models recognize convincing arguments? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8826–8837, Miami, Florida, USA. Association for Computational Linguistics.

William Samuelson and Richard Zeckhauser. 1988. Status quo bias in decision making. *Journal of risk and uncertainty*, 1:7–59.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. 2024. Cognitive biases in large language models: A survey and mitigation experiments.

Alaina N. Talboy and Elizabeth Fuller. 2023. Challenging the appearance of machine intelligence: Cognitive bias in llms and best practices for adoption.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Aswathy Velutharambath, Amelie Wuehrl, and Roman Klinger. 2024. Can factual statements be deceptive? the defabel corpus of belief-based deception. In *Proceedings of The Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, Torino, Italy. European Language Resources Association.

Prashanth Vijayaraghavan and Soroush Vosoughi. 2022. TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3433–3448, Seattle, United States. Association for Computational Linguistics.

Hrishikesh Viswanath and Tianyi Zhang. 2023. Fairpy: A toolkit for evaluation of social biases and their mitigation in large language models.

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. *arXiv preprint arXiv:2403.16084*.

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.

Tobias Walter, Celina Kirschner, Steffen Eger, Goran Glavaš, Anne Lauscher, and Simone Paolo Ponzetto. 2021. Diachronic analysis of german parliamentary proceedings: Ideological shifts through the lens of political biases.

Douglas Walton. 2005. *Fundamentals of critical argumentation*. Cambridge University Press.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7468–7484, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhuqing Zhang, Jun Zhau, Ning Liu, Xiao Gu, and Ya Zhang. 2017. An improved pairwise comparison scaling method for subjective image quality assessment. In *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–6. IEEE.

A Pre-annotation and classifiers

Pre-annotation We start with the **Second Reading debates of Bills**,[13] where the members debate the main principles of a certain Bill. The advantages of using such debates are: (i) the stance of an argument can be easily identified based on whether they support the Bills; (ii) debates can be paired with brief Bill introductions,[14] providing clear argument topics; and (iii) the arguments focus on Bill principles, with fewer discussions on specific amendments and clauses, which require less contextual awareness than other Bill debates like the ones for the Committee Stage.[15] We choose five Bills, including topics relevant to animal welfare and parental leave (see Table 6 for the Bill introductions), which may be easier to annotate and more likely to have emotional arguments.

Three annotators label 245 texts from these debates for **three layers**: (*L1*) whether the text evokes emotions, (*L2*) whether the text contains standalone arguments, and (*L3*) the stance of the text toward the Bill. *L1* and *L2* are labeled ‘0’ (for answer ‘no’) or ‘1’ (for ‘yes’). If *L2* is labeled ‘1’, annotators proceed to label *L3*, which has four options: ‘0’ for support, ‘1’ for opposition, ‘2’ for inability to identify stance without additional context, and ‘3’ for a neutral stance suggesting additional amendments or policies. Besides, 40 texts from the pilot annotation are also annotated for *L1* and *L2*. To potentially speed up the annotation process, the 285 texts are selected from those judged as both emotional and argumentative by GPT4o. Here, we prompt GPT4o with simple questions such as *Does this text try to convince readers something?* and *Is this text emotional?'*.

40 of the outputs are jointly labeled by all annotators, achieving average Cohen’s Kappa of 0.622 for *L1*, 0.674 for *L2*, and 0.762 for *L3* across annotator pairs. As shown in the ‘Question’ column of Table 5, GPT4o already achieves a high precision of 0.82 in detecting argumentative texts using simple prompts. However, its precision for emotional text classification is still low (0.53).

We then convert the annotations for *L3* to *L3**, where we pair argument pairs based on their topics and stances. The categories include: ‘different topic’ for pairs with different topics (from different Bills), ‘different stance’ for pairs with the same topic but different stances, and ‘same’ for pairs with the same topic and stance.

The number of texts annotated for each layer and the corresponding label distributions are summarized in Table 5 (left).

Automatic Pipeline We develop three classifiers based on GPT4o to automatically identify the argument pairs needed. The pipeline is as follows:

[13] https://www.parliament.uk/about/how/laws/passage-bill/commons/coms-commons-second-reading/

[14] e.g., the ‘long title’ on page https://bills.parliament.uk/bills/3858

[15] https://www.parliament.uk/about/how/laws/passage-bill/commons/coms-commons-comittee-stage/

1. **Argumentative text classification**: our goal is to have a **high precision** classifier since we have sufficient candidate texts. We find that when we ask GPT-4o to provide the major claim, evidence, and reasoning connecting the evidence to the major claim in the text, its precision increases from 0.82 to 0.96, as shown in the 'Argumentative' row of Table 5.

   We then retain texts judged as argumentative for $\texttt{Hansard}_{\texttt{en}}$ using this prompt, while for $\texttt{DeuParl}_{\texttt{de}}$, we use a German translation of the same prompt. The overall performance of GPT4o on German data is assessed after completing the stance agreement classification task (see below).

2. **Stance agreement classification**: To enable the flexible selection of classifiers with specific performance characteristics (e.g., high recall, high precision), we introduce a parameter into the prompt, with its threshold optimized to achieve different specialized performance levels. To do so, we ask GPT4o to rate the likelihood that two given arguments address the same topic and share the same stance on a Likert scale from 0 to 100. We randomly sample 600 argument pairs (with a 2:1:1 ratio for the three categories of *L3**) from the dataset, 'optimize' the threshold of ratings for the 'same' category using argument pairs from two Bills, and test the performance on the remaining three Bills to prevent data leakage. We evaluate all possible combinations of Bills for the training and test sets. We observe that as the threshold increases, precision on the 'same' category (P_{same}) consistently improves, while macro F1 begins to decrease beyond certain thresholds. With a threshold of 100, P_{same} reaches 0.92, but F1 is very low at 0.45. Therefore, we select a threshold of 90 as a more balanced trade-off, achieving $P_{same} = 0.81$ and $F1 = 0.76$, to obtain more candidates that are still highly likely to be true positives.

   For $\texttt{Hansard}_{\texttt{en}}$, we retain the argument pairs labeled as belonging to the 'same' category using this threshold. For $\texttt{DeuParl}_{\texttt{de}}$, we apply the German translation of the prompt with the same threshold to identify argument pairs. One annotator evaluates 50 candidates from the outputs of steps 1 and 2: no argument is labeled as non-argumentative, while 12 argument pairs are identified as false positives in the stance agreement task, yielding $P_{same} = 0.76$. This value is only 4 percent points lower than the result on English data. Consequently, we retain these prompt settings for the German data.

3. **Emotional text classification**: we aim for a **balanced** classifier because we also need non-emotional arguments. Since this is a subjective task, we ask GPT4o to rate how likely it can feel the emotions in the texts on a Likert scale of 0-100, and then 'optimize' the threshold of the rates for the 'emotional' category on 70% of the data and check how it performs on the remaining 30%. Overall, with this step, we can improve the macro F1 to 0.74-0.81 (averaged over three rounds of data splitting), depending on the gold from different annotators. The best threshold for two annotators is 75, while that for the other is 85, so we use the threshold 75 to represent the majority, which has a macro F1 of 0.75, averaged across the three annotators.

   We use this threshold to select the argument pairs for $\texttt{Hansard}_{\texttt{en}}$. For $\texttt{DeuParl}_{\texttt{de}}$, we further optimize the threshold using a small-scale set of human annotations and adjust it to 85. This setting is then used to label the binary emotions of arguments.

B Arguments from others

$\texttt{Dagstuhl}_{\texttt{en}}$ Wachsmuth et al. (2017) collected human ratings on a Likert scale of 1–3 for multiple dimensions of argument quality, including argument effectiveness (convincingness)[16] and emotional appeal. These ratings were applied to 304 argumentative texts from Habernal and Gurevych (2016b), which were sourced from a textual debate portal in **English**. We retain only those arguments whose average convincingness rating (across the three annotators) exceeds 1.5. Next, we pair arguments that share the same stance on the same topics and calculate the absolute differences in their emotional appeal ratings. From these pairs, we randomly select 10 topics and then retain the 5 argument pairs with the largest absolute differences in emotional appeal for each topic.

[16]"Argumentation is effective if it persuades the target audience of (or corroborates agreement with) the author's stance on the issue." — Wachsmuth et al. (2017)

	Pre-Annotation #	%	Automatic Pipeline Question	'Optimized'
L1 - emotion				
Emotional	151	53.0	0.53 (P)	0.75 (F1)
Non-emotional	134	47.0	-	
L2 - argument				
Argumentative	234	82.1	0.82 (P)	0.96 (P)
Non-argumentative	51	17.9	-	-
L3 - stance				
Support	170	72.6	-	
Opposition	2	0.9	-	
Neutral	29	12.4	-	
Irrelevant	16	14.1	-	
L3 - pair stance*				
Same	2,905	8.9	-	0.80 (P_{same}) 0.75 (F1)
Different stance	3,325	10.2	-	
Different topic	26,486	81.0	-	
Total	32,716	100	-	

Table 5: Number of texts annotated for each layer and category (#) and the corresponding label distribution (%). Performance of GPT4o on the binary emotion classification, argument identification, and stance agreement detection tasks used for automatically identifying the target argument pairs.

EmoDefabel$_{de}$ Greschner and Klinger (2024) collected discrete emotion labels from a reader respective (e.g. joy, disgust etc.) for 300 **German** arguments associated with 30 statements, drawn from Velutharambath et al. (2024). Each argument was annotated by three annotators. We interpret the number of annotations marking the argument as containing specific emotions (rather than 'no emotion') as its emotion score. E.g., if three annotators identify specific emotions in the argument, its emotion score would be 3. Using a procedure similar to the one employed for Dagstuhl$_{en}$, we pair arguments referencing the same statement, randomly select 25 statements, and then retain the two argument pairs per statement that exhibit the greatest differences in emotion scores.

C Prompts

Table 8 presents the prompts used to introduce/remove emotions. Table 10 illustrates the prompts used for evaluating argument convincingness.

D Annotation Interface

Figure 4 shows the screenshots of the annotation interface for convincingness (top) and emotion (bottom) comparisons. We collect the annotations via Google Forms[17] for crowdsourcing annotators.

E Examples

Table 11 and 12 provide example instances from Hansard$_{en}$ and EmoDefabel$_{de}$, where emotions have a positive and negative impact, respectively.

F LLM

Figure 5 illustrates the consistency, positivity and negativity rates of LLMs with different prompts, averaged across instances in all datasets. Table 13 displays macro F1 scores and model rankings for LLMs in predicting convincingness rankings of argument pairs ('Static') and the resulting categories of emotional effect ('Dynamic') in English and German.

[17] https://docs.google.com/forms/

Convincingness Comparison

Below, you will find **20 pairs of argumentative texts** discussing **the same topic with the same stance**. Each topic may be a binary choice, a bill from UK parliamentary debates, or a simple statement. Both arguments either support or oppose the topic, or they favor one side if the topic involves a binary choice.

Your task is to evaluate each pair to determine **which argumentative text you find more convincing**. There are three label options:

- 0 (Both arguments are equally convincing.)
- 1 (Argument 1 is more convincing.)
- 2 (Argument 1 is more convincing.)

Note: If the text contains **truncated sentences** or **grammatical errors**, please **ignore** them.

Topic: *

A Bill to create offences of dog abduction and cat abduction and to confer a power to make corresponding provision relating to the abduction of other animals commonly kept as pets.

Argument 1:

I have received reports of dogs being stolen from gardens in my constituency, vanishing after being let out in the evening. Thanks to the support of amazing charities such as Beauty's Legacy which work to reunite lost and stolen pets with their owners, and the Staffordshire police service that tracks down dogs, there have been local success stories with perpetrators being arrested, yet at present the theft of a treasured family pet is no more serious under the law than the theft of a phone or wallet. Pets are currently treated as mere replaceable property, akin to an inanimate object. That is not just an oversight but an insult, failing to acknowledge the emotional trauma experienced by the victims of pet abduction.

Argument 2:

One reason that this legislation is so important is the sheer scale of these offences now. According to Direct Line, between 2018 and 2022 there were more than 12,000 dog thefts—an average of 2,400 a year, and the equivalent of seven dogs stolen every single day. Those figures are not complete, because not all forces even register such offences. Cat theft, which has been mentioned, is now catching up. According to Pet Theft Awareness, the police recorded that the number of stolen cats had jumped by 40% in 2021 to an all-time high of 560. Cat theft has quadrupled since 2015, and data from the Metropolitan police shows that cat theft as a proportion of total pet theft crimes rose from 6% in 2012 to 31% last year.

- ○ 0 (Both arguments are equally convincing.)
- ○ 1 (Argument 1 is more convincing.)
- ○ 2 (Argument 2 is more convincing.)

Emotion Comparison

Below, you will find **20 pairs of argumentative texts** discussing **the same topic with the same stance**. Each topic may be a binary choice, a bill from UK parliamentary debates, or a simple statement. Both arguments either support or oppose the topic, or they favor one side if the topic involves a binary choice.

Your task is to evaluate each pair to determine **which argumentative text evokes stronger emotions in you**. There are three label options:

- 0 (Both arguments evoke a similar level of emotion, or neither argument evokes emotions in you.)
- 1 (Argument 1 evokes stronger emotions in you.)
- 2 (Argument 2 evokes stronger emotions in you.)

Note: If the text contains **truncated sentences** or **grammatical errors**, please **ignore** them.

Topic: *

A Bill to Prohibit the export of certain livestock from Great Britain for slaughter.

Argument 1:

It is essential to recognize that this legislation touches the very core of who we are as a society and what we hold dear in terms of compassion and ethical treatment. The push for a ban on live exports is not merely a policy decision; it is a reflection of our collective conscience and moral responsibility. The overwhelming public support, evident in numerous heartfelt petitions, showcases the deep-rooted desire for change and a more humane world. People are passionately urging the UK Government to pass the kept animals Bill through Parliament, as it symbolizes a step towards a kinder future. The proposed ban specifically targets exports for fattening and slaughter, ensuring that breeding animals, crucial to Scotland's agricultural heritage, are safeguarded. This distinction highlights our commitment to progress while respecting traditional values.

Argument 2:

I recently visited Oak Wood School in Hillingdon, which was hosting a Christmas fair for students with special educational needs that involved interaction with animals. Such interaction not only supports education, but significantly enhances the wellbeing and confidence of young people. The people of Uxbridge and South Ruislip are huge animal lovers. No one could go far in my constituency without spotting many dog walkers enjoying the wonderous open and green spaces that are part of one of London's greenest constituencies. My inbox is often full not just with campaign emails, but with genuine heartfelt messages that touch on all elements of animal welfare. That has been especially the case over the past couple of weeks, as residents have echoed my feeling in support of the Bill. Indeed, more than 85% of the 11,000 respondents to the Department's consultation on live exports agreed with the measures in the Bill.

- ○ 0 (Both arguments evoke a similar level of emotion, or neither argument evokes emotions in you.)
- ○ 1 (Argument 1 evokes stronger emotions in you.)
- ○ 2 (Argument 2 evokes stronger emotions in you.)

Figure 4: Screenshots of the annotation interface for convincingness (top) and emotion (bottom) comparison.

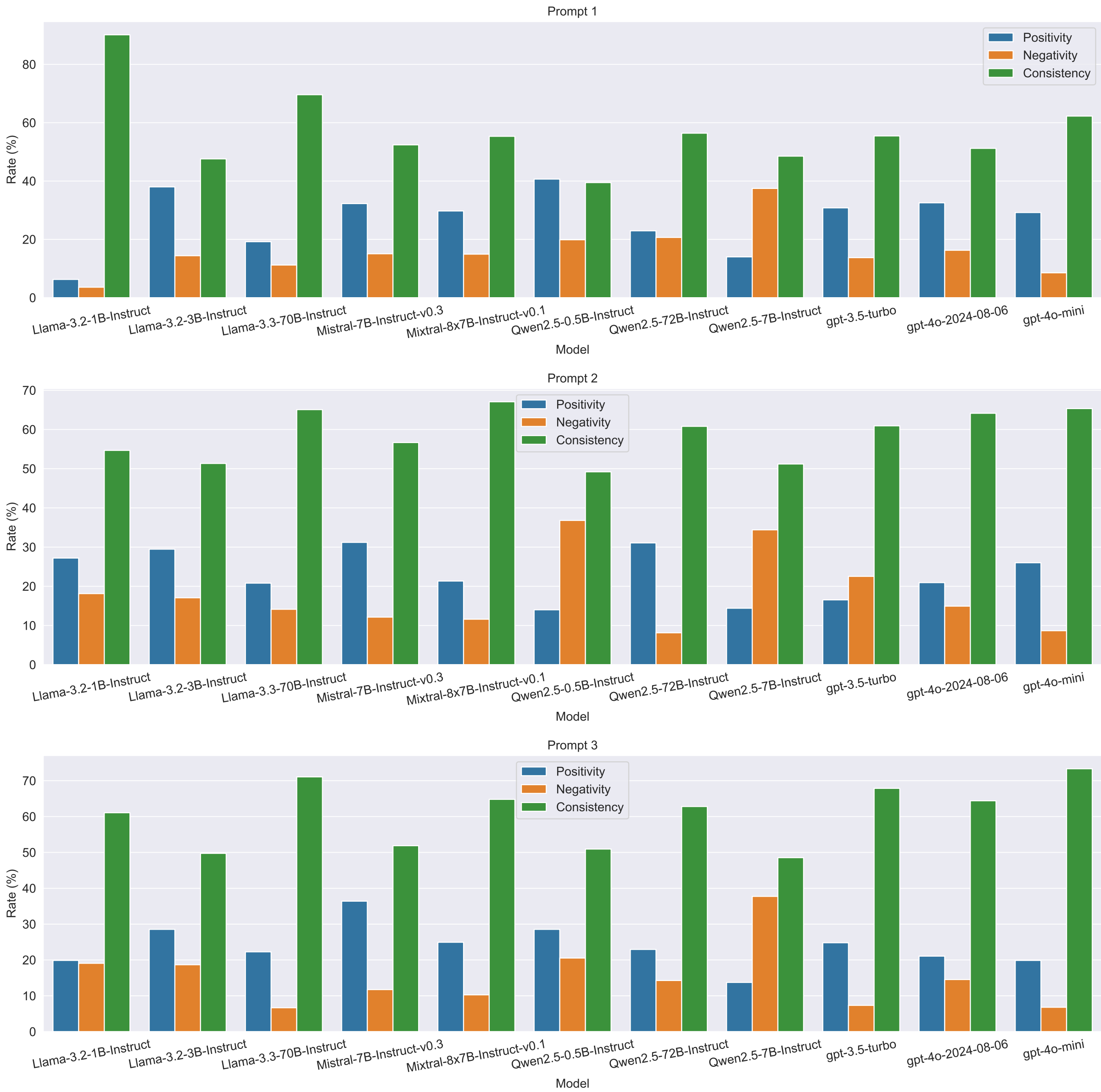

Figure 5: Consistency, positivity and negativity rates of LLMs with different prompts, averaged across instances in all datasets.

Introduction
A Bill to Prohibit the export of certain livestock from Great Britain for slaughter.
A Bill to create offences of dog abduction and cat abduction and to confer a power to make corresponding provision relating to the abduction of other animals commonly kept as pets.
A Bill to make provision about leave and pay for employees with responsibility for children receiving neonatal care.
A Bill to prohibit the import and export of shark fins and to make provision relating to the removal of fins from sharks.
A Bill to prohibit the sale and advertising of activities abroad which involve low standards of welfare for animals.

Table 6: The introductions of the five Bills selected in $\texttt{Bill}_{\text{en}}$.

English	German
iran, integrat, ukraine, russia, asylum, deportation, israel, gaza, expulsion, displacement, migration, migrant, immigrant, refugee, palestine,invasion, repatriation, hamas, hisbollah	ukraine, russland, migrant, immigrant, flüchtling, asyl, gaza, iran, palästina, israel, krieg, invasion, sanktionen, waffenlieferungen, friedensverhandlungen, kriegsverbrechen, flüchtlingskrise, nato, energieversorgung, vertreibung, migrationspolitik, asylverfahren, grenzsicherung, integration, abschiebung, aufenthaltsgenehmigung, menschenhandel, seenotrettung, rückführung, schutzstatus, waffenstillstand, raketenangriffe, besatzung, zwei-staaten-lösung, friedensprozess, intifada, hamas, hisbollah, menschenrechte, un-resolution

Table 7: Keywords used to filter debates for $\texttt{Hansard}_{\text{en}}$ and $\texttt{DeuParl}_{\text{de}}$.

Remove Emotion Prompt
====**System Prompt**===== I will give you an argumentative text that **can** appeal to emotion. Your task is to generate an argument with the same stance for the same topic **without emotional language**, by rephrasing the text but maintaining a similar style and length. Briefly explain why the rewritten argument no longer evokes emotions. Answer in the following way: Generated argument: Explanation: ====**User Prompt**===== Text: {original argument}
Add Emotion Prompt
====**System Prompt**===== I will give you an argumentative text that **cannot** appeal to emotion. Your task is to generate an argument with the same stance on the same topic **with emotions**, by rephrasing the text but maintaining a similar style and length. Briefly explain why the rewritten argument can evoke emotions now. Answer in the following way: Generated argument: Explanation: ====**User Prompt**===== Text: {original argument}

Table 8: Prompts used to remove/add emotions for synthetic arguments.

	$\texttt{Dagstuhl}_{en}$	$\texttt{Bill}_{en}$	$\texttt{Hansard}_{en}$	$\texttt{EmoDefabel}_{de}$	$\texttt{DeuParl}_{de}$
Increase					
E	**-0.06**	**0.15**	**0.05**	**-0.38**	**0.32**
$G^{-}(E)$	-0.18	-0.21	-0.31	-0.46	-0.38
Decrease					
N	-0.12	-0.21	-0.03	0.08	-0.19
$G^{+}(N)$	**0.36**	**0.27**	**0.29**	**0.76**	**0.25**

Table 9: BWS scores for the 4 argument groups: E, N, $G^{+}(N)$ and $G^{-}(E)$, derived from the majority votes of the annotation for pairwise comparisons of emotional intensity. 'Increase'/'Decrease' denotes the direction to increase/decrease the perceived emotional intensity.

Prompt	Template
Shared	Below, you will find one pair of argumentative texts discussing the same topic with the same stance. The topic may be a binary choice, a bill from UK parliamentary debates, or a simple statement. Both arguments either support or oppose the topic, or they favor one side if the topic involves a binary choice. Your task is to evaluate each pair to determine **which argumentative text you find more convincing**. There are three label options: 0 (Both arguments are equally convincing.) 1 (Argument 1 is more convincing.) 2 (Argument 2 is more convincing.) **Note**: Truncated sentences or grammatical errors should be **ignored**.
1	Please answer your label option **without** any explanations. {text}
2	Please answer your label option and briefly explain why you choose this label. {text} Below is an example answer for you; please follow this format in your response. Label: 2 Explanation: because Argument 2 provides more statistics supporting the claim, while Argument 1 contains logical fallacies.
3	Please answer your label option and briefly explain why you choose this label. {text} Below is an example answer for you; please follow this format in your response. Label: 1 Explanation: Argument 1 is more convincing, because I totally agree with its point and it evokes my empathy.

Table 10: Prompt templates for comparing the convincingness of an argument pair. The text field contains the two arguments and their topic. The complete prompt is formed by combining the text in the ‘Shared’ row with the text in the corresponding indexed row. For example, Prompt 1 consists of the text from both the ‘Shared’ row and row ‘1’.

Topic: The public supports the UK's aid for Ukrainian refugees	
E	*N*
Members across this House are determined that we, as a country, should open our arms to these people, and this determination has been on full display today. The scenes of devastation and human misery inflicted by President Putin's barbarous assault on what he calls "Russia's cousins" in Ukraine have unleashed a tidal wave of solidarity and generosity across the country. British people always step forward and step up in these moments, and since the first tanks rolled into Ukraine, they have come forward in droves with offers of help: community centres have been flooded with critical supplies; the Association of Ukrainians in Great Britain has received millions in donations; and charities such as the Red Cross have been overwhelmed with people giving whatever they can. The outpouring of public support has been nothing short of remarkable.	While this Government, and this whole House, have risen to the occasion with our offer of support to Ukrainians fleeing war, our lethal aid and our stranglehold on economic sanctions on Russia have clearly shown that we will keep upping the ante to ensure that Putin fails. As Members have argued today, it has been abundantly clear in recent days that we can and must do more. It is exactly right, therefore, that my right hon. Friend the Secretary of State for Levelling Up, Housing and Communities set out on Monday the new and uncapped sponsorship scheme, Homes for Ukraine. It is a scheme to allow Ukrainians with no family ties to the UK to be sponsored by individuals or organisations that can offer them a home. It is a scheme that draws not only on the exceptional good will and generosity of the British people, but one that gives them the opportunity to help make a difference.
G^{-}*(E)*	G^{+}*(N)*
Members of this House have expressed a commitment to welcoming individuals from Ukraine. The recent conflict initiated by President Putin has resulted in significant destruction in Ukraine, prompting a substantial response of support across the country. British citizens have actively contributed since the conflict began, with community centers collecting essential supplies, the Association of Ukrainians in Great Britain receiving financial contributions, and charities like the Red Cross witnessing increased donations.	In these trying times, the Government and this entire House have demonstrated unwavering courage and compassion by extending our support to Ukrainians escaping the horrors of war. Our determined provision of lethal aid and the relentless imposition of economic sanctions on Russia are powerful affirmations that we will stop at nothing to ensure Putin's defeat. As Members have passionately discussed today, the urgency to do even more has never been clearer. That is why it is so heartening that my right hon. Friend the Secretary of State for Levelling Up, Housing and Communities announced on Monday the new and limitless Homes for Ukraine sponsorship scheme. This initiative opens its arms to Ukrainians without family connections in the UK, allowing them to be warmly embraced by individuals or organizations ready to offer them a sanctuary. It is a testament not only to the extraordinary kindness and generosity of the British people but also to their deep desire to make a meaningful impact in the lives of those in desperate need.

Table 11: An example instance from $\texttt{Hansard}_{\texttt{en}}$ where emotions have a **positive** impact on argument convincingness.

Topic: Haie können Krebs bekommen.	
E	*N*
Haie sind mehrzellige Lebewesen, wie auch der Mensch. Die Beonderheit von mehrzelligen Lebewesen ist, dass die Zellen sich sowohl stark spezialisieren und untereinander vernetz kommunizieren. Damit werden sie anfällig für bestimmte Zelldefekte, die sich über die genannte Struktur fortpflanzen und den Krebs ausmachen. Haie verfügen, wie auch der Mensch und überhaupt alle mehrzelligen Lebewesen, über nur eine sehr eingeschränkte Möglichkeit diese Defekte zu korrigieren und aufzuhalten, damit können beide gleichermaßen Krebs bekommen 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.	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.
G^{-}*(E)*	G^{+}*(N)*
Haie, ebenso wie Menschen, sind mehrzellige Organismen. Eine charakteristische Eigenschaft solcher Organismen ist die Spezialisierung und Vernetzung ihrer Zellen. Diese Struktur macht sie anfällig für Zellfehler, die sich ausbreiten und zu Krebs führen können. Haie und Menschen besitzen nur begrenzte Mechanismen zur Korrektur und Kontrolle dieser Defekte, was bedeutet, dass beide Arten gleichermaßen anfällig für Krebs sind.	Die Vorstellung, dass Haie - diese majestätischen und oft missverstandenen Kreaturen der Meere - an Krebs erkranken können, ist zutiefst beunruhigend. Diese Krankheit, die durch die heimtückische Mutation von Zellen verursacht wird, hat bereits viele Fischarten heimgesucht. Die Tatsache, dass auch Haie, die Könige der Ozeane, nicht sicher vor dieser grausamen Krankheit sind, ist erschütternd. Angesichts der weit verbreiteten Umweltverschmutzung, die unsere Ozeane verschlingt, sind Haie einem erheblichen Risiko ausgesetzt, an Krebs zu erkranken. Es ist traurig und alarmierend, dass diese beeindruckenden Tiere, die seit Millionen von Jahren die Meere durchstreifen, nun durch menschliche Einflüsse bedroht sind.

Table 12: An example instance from $\texttt{EmoDefabel}_{\texttt{de}}$ where emotions have a **negative** impact on argument convincingness.

Model	EN				DE			
	Static	Ranking	Dynamic	Ranking	Static	Ranking	Dynamic	Ranking
gpt-4o-2024-08-06	**0.486**	1	0.411	2	**0.443**	1	**0.447**	1
Llama-3.3-70B-Instruct	0.417	2	**0.415**	1	0.372	2	0.392	4
gpt-4o-mini	0.416	3	0.392	5	0.35	4	0.394	3
Qwen2.5-72B-Instruct	0.398	4	0.398	4	0.357	3	0.41	2
gpt-3.5-turbo	0.39	5	0.382	6	0.338	6	0.381	6
Mixtral-8x7B-Instruct-v0.1	0.368	6	0.376	7	0.35	5	0.387	5
Mistral-7B-Instruct-v0.3	0.367	7	0.407	3	0.288	8	0.36	9
Llama-3.2-3B-Instruct	0.322	8	0.32	10	0.281	10	0.367	8
Qwen2.5-0.5B-Instruct	0.308	9	0.342	9	0.284	9	0.344	10
Qwen2.5-7B-Instruct	0.304	10	0.346	8	0.319	7	0.373	7
Llama-3.2-1B-Instruct	0.286	11	0.309	11	0.274	11	0.343	11

Table 13: Macro F1 scores and model rankings for LLMs in predicting convincingness rankings of argument pairs ('Static') and the resulting categories of emotional effect ('Dynamic') in English and German. For each model, we present the best prompt result to highlight its potential. Human and LLM labels are determined by majority votes from different annotators and rounds, respectively.