Comparing the Framing Effect in Humans and LLMs on Naturally Occurring Texts

Gili Lior **Liron Naccache** **Gabriel Stanovsky**
The Hebrew University of Jerusalem
gili.lior@mail.huji.ac.il

Abstract

Humans are influenced by how information is presented, a phenomenon known as the *framing effect*. Prior work suggests that LLMs may also be susceptible to framing, but it has relied on synthetic data and did not compare to human behavior. To address this gap, we introduce WILDFRAME – a dataset for evaluating LLM responses to positive and negative framing in naturally-occurring sentences, alongside human responses on the same data. WILDFRAME consists of 1,000 real-world texts selected to convey a clear sentiment; we then *reframe* each text in either a positive or negative light and collect human sentiment annotations. Evaluating eleven LLMs on WILDFRAME, we find that all models respond to reframing in a human-like manner ($r \geq 0.52$), and that both humans and models are influenced more by positive than negative reframing. Notably, GPT models are the least correlated with human behavior among all tested models. These findings raise a discussion around the goals of state-of-the-art LLM development and whether models should align closely with human behavior, to preserve cognitive phenomena such as the framing effect, or instead mitigate such biases in favor of fairness and consistency.[1]

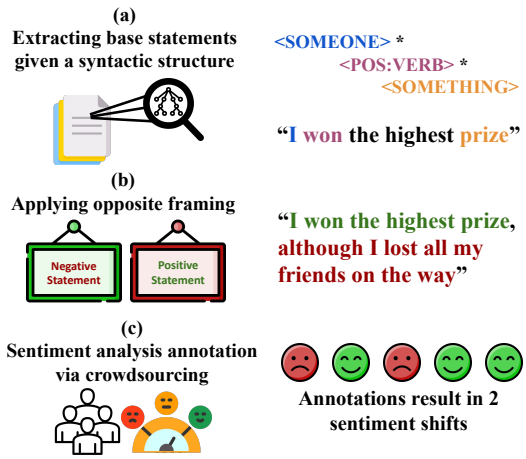


Figure 1: The WILDFRAME data construction process. In step (a) we extract statements based on their syntactic structure, aiming for statements with clear negative or positive sentiment. Next, in (b), we *reframe* the statement by adding a suffix or prefix, conveying the *opposite* sentiment. Finally, in (c), five annotators mark the sentiment of the reframed statement, counting how many annotators shift sentiment, i.e., the reframed statement sentiment is opposite to the base sentiment. The red parts in the figure represent negative parts of statement, while green represents positive parts.

1 Introduction

The *framing effect* is a well-known cognitive phenomenon, where different presentations of the same underlying facts affect how those facts are perceived by humans (Tversky and Kahneman, 1981). For example, presenting an economic policy as only creating 50,000 new jobs, versus also reporting that it would cost 2B USD, can dramatically shift public opinion (Sniderman and Theriault, 2004).

Previous research has shown that LLMs exhibit various cognitive biases, including the framing effect (Lorè and Heydari, 2024; Shaikh et al., 2024; Malberg et al., 2024; Echterhoff et al., 2024). However, existing studies either rely on synthetic datasets or evaluate LLMs and humans on different inputs, limiting direct behavioral comparison. In addition, comparisons between models and humans typically use a single metric for evaluation rather than comparing patterns in human behavior.

In this work, we evaluate LLMs on real-world data that is reframed to convey a different sentiment than the sentiment of the original statement. We adopt a broad definition of framing inspired by prior work examining evaluative framing (Tong et al., 2021), which includes adding information that may change the reader's perspective. For example, reframing "*I won the highest prize*" to "*I*

[1] WILDFRAME can be found at https://huggingface.co/datasets/gililior/WildFrame

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 825–841
October 24-29, 2026 ©2026 Association for Computational Linguistics

won the highest prize, although I lost all my friends on the way". This setup corresponds to real-world scenarios in which additional information is selectively introduced to influence interpretation. Such framing is common in everyday communication, including in political discourse and social media, where judgments often depend on which aspects of an event are emphasized rather than on the event itself. Consequently, understanding the effect that framing has on LLMs becomes increasingly important, as they are integrated into real-world systems.

Rather than measuring model performance in terms of accuracy, we analyze how closely LLM responses align with human annotations. Furthermore, while previous studies have examined the effect of framing on decision making, we extend this analysis to sentiment analysis, as sentiment perception plays a key explanatory role in decision-making (Lerner et al., 2015). Studying framing in this work allows us to highlight systematic differences between humans and LLMs, which can later be informative when differentiating cases in which human-like behavior may be appropriate, versus when more rationale, framing-invariant behavior may be desirable.

To better understand the effect of different framings on LLMs in comparison to human behavior, we introduce the WILDFRAME dataset (Section 2), comprising 1,000 statements, curated through a three-step process, as shown in Figure 1. First, we collect a set of real-world statements that express a clear negative or positive sentiment (e.g., "*I won the highest prize*" vs "*I failed my math test*"). Second, we *reframe* the text by appending or prepending a clause with opposite sentiment (e.g., "*I won the highest prize, although I lost all my friends on the way*"). Finally, we collect human annotations by asking participants if they consider the reframed statement to be overall positive or negative.

WILDFRAME consists of Amazon reviews, which is chosen since sentiment in review statements is more robust, compared to domains such as news, which introduce confounding factors like prior political leanings (Druckman, 2004). Examples from WILDFRAME are presented in Table 1.

Recent work has raised concerns that human annotations collected via crowdsourcing platforms may be contaminated by LLM-assisted responses (Veselovsky et al., 2023). This poses a key methodological challenge for studies comparing humans and LLMs: how to retain the scalability of crowdsourcing while ensuring that observed patterns reflect genuine human behavior rather than implicit model usage. In this work, we take several steps to increase confidence in the authenticity of our annotations, including validating against established cognitive science findings, statistical analyses of variance, and comparisons with a small-scale, controlled in-house human baseline.

We then evaluate eleven LLMs, including both proprietary and open-source models, on the WILDFRAME dataset, and compare them against human annotations. Specifically, we compare the *sentiment shifts* in humans and models. That is, labeling the sentiment aligning with the reframing sentiment, rather than the base sentiment, e.g., voting "negative" for the statement "*I won the highest prize, although I lost all my friends on the way*", as it aligns with the opposite framing sentiment. *Sentiment shift* is also exemplified in Table 1.

We find that LLMs are influenced by framing in a similar manner to human behavior, with all models exhibiting a strong correlation with human annotations ($r \geq 0.52$). Additionally, both humans and LLMs are more affected by positive reframing of negative base statements than by negative reframing of positive base statements. This asymmetry aligns with prior findings in cognitive science, which show that positive framing tends to exert a stronger influence (Tong et al., 2021; Grappi et al., 2024). We also observe a general trend in which larger models tend to exhibit higher correlation with human behavior.

Interestingly, we find that GPT-3.5 and GPT-4o (OpenAI, 2024) exhibit the lowest correlation with human behavior, and GPT-5-mini (OpenAI, 2025) shows correlation levels that are comparable to several smaller open-source models. This observation raises questions about how architectural design choices and training procedures may influence a model's susceptibility to framing effects.

This work contributes to understanding the parallels between LLM and human cognition, offering insights into how cognitive mechanisms such as the framing effect emerge in LLMs. This also raises an interesting discussion on the objective when developing state-of-the-art LLMs – whether models should aim to align closely with human behavior, preserving cognitive phenomena such as framing effects, or instead mitigate such biases in favor of fairness and consistency. We hope that our results would benefit model developers by informing decisions on whether to harness or reduce framing effects, depending on the downstream application.

Reframed Sentence	Base Sentiment	Human Majority	Sentiment Shift
Given the price, I can't complain at all. However, the product quality is sub-par.	Positive	Negative (0.8)	✓
Stanly doesn't fail to deliver with this model. It's unfortunately quite pricey for its features, though.	Positive	Positive (0.6)	✗
I won't waste my money again, and hopefully you won't either, but it was a valuable lesson about the importance of making wise financial choices.	Negative	Negative (0.8)	✗
For me, Aspartame causes bad memory loss and nasty gastrointestinal distress, but this has encouraged me to seek out healthier, natural alternatives and cultivate a balanced diet.	Negative	Positive (0.8)	✓

Table 1: Examples from WILDFRAME. The underlined text indicates the reframing clause, while the remaining text corresponds to the base statement. The Base Sentiment column reflects the dominant sentiment of the base statement. Human Majority Vote denotes the majority sentiment assigned to the reframed sentence, with the fraction of annotators in agreement shown in parentheses. Sentiment Shift indicates whether the majority human judgment aligns with the reframing sentiment (✓) or remains aligned with the base sentiment (✗).

2 The WILDFRAME Dataset

Our dataset curation consists of three steps, as depicted in Figure 1. First, we collect natural, real-world statements, with some clear sentiment, either positive or negative (§2.1; e.g., "*I won the highest prize*" as positive). Next, using an LLM, we reframe each statement by appending or prepending a clause conveying the opposite sentiment (§2.2; e.g., "*I won the highest prize, although I lost all my friends on the way*"). Finally, we collect large-scale human annotations via crowdsourcing, to label the sentiment shifts when wrapping the statements with the opposite framing (§2.3; e.g., labeling "negative" the statement "*I won the highest prize, although I lost all my friends on the way*", shown also in Table 1).

The complete dataset consists of 1000 statements, in which 500 are statements that their base form has positive sentiment, and 500 are base negative statements.

2.1 Collecting Base Statements

First, we collect base statements, which convey a clear sentiment, either clearly positive or clearly negative statements. We use SPIKE – an extractive search system, which allows to extract statements from real-world datasets (Taub Tabib et al., 2020). Specifically, we collect English statements from Amazon Reviews dataset, which are naturally occurring, sentiment-rich, texts but are less likely to trigger strong preexisting biases or emotional reactions, which may be a confound for our experiment.[2]

Using SPIKE, we extract $\sim 6k$ statements that fulfilled our designated queries, which we found correlated with clear sentiment. We designed the queries to capture positive or negative verbs that describe actions with some clear sentiment (e.g., "*enjoy*" or "*waste*"), or statements with positive or negative adjective, describing an outcome with a clear sentiment (e.g., "*good*" or "*nasty*"). The patterns and queries used for extraction are detailed in Appendix A. Next, we run in-house annotations to label and filter the extracted statements, to handle negations or other cases where the statement does not convey a clear sentiment. Annotators mark each candidate with its dominant sentiment, or as *not suitable* when it did not express a distinct sentiment. The filtering process results in $1,301$ positive statements, and $1,229$ negative statements. We categorize the excluded cases into recurring types in Appendix B.

To assess annotation quality, we measured inter-annotator agreement over 300 sampled sentences. Annotators showed strong alignment in sentiment judgments: only 1% of sentences exhibited strong disagreement, where annotators assigned opposite sentiment labels. In contrast, a much larger source of disagreement arose when deciding whether a sentence is suitable for the task, with 37.1% of cases reflecting weak disagreement tied to this suitability criterion. We hypothesize that suitability is more subjective and influenced by individual preferences. Importantly, when considering only the

[2] https://spike.apps.allenai.org/datasets/reviews

sentences that both annotators agreed were suitable, sentiment agreement increases to 99%, demonstrating highly consistent sentiment judgments once suitability is established.

2.2 Adding Framing

We reframe the statements in our dataset using three different methods, generating three reframed versions for each base statement, to support robust LLM evaluation.

For the most basic paraphrasing, we use GPT-4 (Achiam et al., 2023). The input prompt includes a one-shot example followed by the instruction: "Add a `SENTIMENT` suffix or prefix to the given statement. Do not change the original statement," where `SENTIMENT` is either "positive" or "negative," opposite to the base statement sentiment (i.e., positive framing for a negative base statement, and vice versa).

Unlike the base statements, the sentiment conveyed by reframed statements is inherently more ambiguous, with no single clear label. This is illustrated in Figure 2, by the distribution of sentiment scores before and after reframing, given by a fine-tuned sentiment analysis model.[3] This ambiguity enables us to measure the extent to which LLMs shift their sentiment predictions following framing, and how closely shifts align with human behavior.

Due to the known LLM sensitivity to changes in the prompt (Sclar et al., 2023; Mizrahi et al., 2024), to strengthen the reliability of LLM evaluation, we generate two additional reframings for each sentence, and test LLMs on them. The full prompts are provided in Appendix C.

Diversity of reframings. To ensure that the reframing produces sentences with genuine discourse variation, rather than repeatedly producing the same type of reframing, we compute cross-sentence diversity metrics, using the unique-bigram ratio (Li et al., 2016). While the base sentences show the highest bigram diversity (0.68), as expected for unconstrained, naturally occurring text, the reframed sentences and the added component retain substantial diversity (0.59 and 0.55, respectively), indicating meaningful lexical variability rather than a narrow, repetitive pattern. The added components also vary in length (mean 12.9±4.4 words), comparable to the base sentences (mean 14.2±6.9), suggesting the reframing clauses exhibit structural variation rather than uniform clause construction.

[3] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

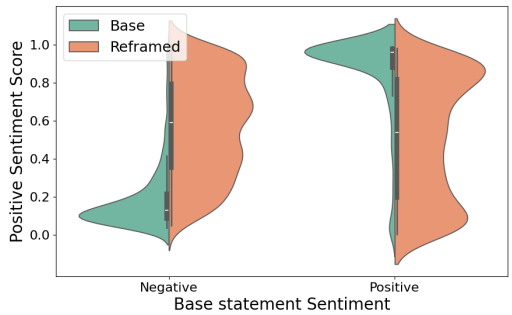


Figure 2: Distribution of sentiment scores before and after applying opposite-sentiment framing, as detailed in Section 2.2. Prior to framing, base sentences exhibit a clear polarity (positive or negative), whereas the application of opposite framing introduces ambiguity, shifting the sentiment scores toward a less distinct polarity.

2.3 Collecting Human Annotations

In the final step, we collect human annotations through Amazon Mechanical Turk to evaluate the framing effect in WILDFRAME over human participants.[4] Details about the annotation platform are elaborated in Appendix D.

The complete dataset includes 1K statements, reframed by GPT-4 as explained in Section 2.2, each annotated by five different annotators. Collecting five annotations per statement enables robust estimation of statement ambiguity and inter-annotator agreement.

For the annotation process, each statement in our dataset is presented in its reframed version (i.e., positive base statements with negative framing and vice versa), to five different annotators. This setup generates, for each dataset instance, a *sentiment shift* score ranging from 0 to 5, representing the number of annotators that votes for the sentiment that aligns with the reframing, which means that the overall sentiment of the reframed statement has shifted from its base sentiment. For example, in Figure 1, the statement "*I won the highest prize, although I lost all my friends on the way*" is shown to have two annotators voting "negative", which aligns with the sentiment of the framing and not the base statement, so the label for that instance in WILDFRAME would be 2 (sentiment shifts).

Instances with score near 0 indicate that annotators agree that the overall sentiment remains un-

[4] https://www.mturk.com

changed despite the opposite sentiment reframing. Score closer to 5 indicates that annotators agree that reframing shifted the perceived sentiment, while score around 2-3 suggests that the reframing makes the sentiment ambiguous.

3 Results

We evaluate eleven LLMs on the WILDFRAME dataset, comprising both open- and closed-source instruction-tuned models. These included GPT-3.5, GPT-4o and GPT-5-mini (OpenAI, 2025), Gemini-2.5-flash (Comanici et al., 2025), Llama-3.1-8B and Llama-3.3-70B (Dubey et al., 2024), Mistral-v0.3-7B (Jiang et al., 2023), Mixtral-v0.1-8x7B and Mixtral-v0.1-8x22B (MistralAI, 2024), Gemma-2-9B and Gemma-2-27B (Team et al., 2024).

3.1 Validation of Actual Human Behavior in Crowdsourced Annotations

Recent discussions have raised concerns that annotations obtained via crowdsourcing platforms like Mechanical Turk might not reflect genuine human input, as workers could potentially use LLM-generated responses (Veselovsky et al., 2023). While in-house controlled experiments guarantee authenticity, they typically lack the scale required for robust LLM evaluation. This raises a methodological challenge for research comparing humans and model behavior: how can we maintain the scalability benefits of crowdsourcing while ensuring we are testing against real human signal? In particular, it is difficult to distinguish between cases where humans and LLMs truly behave similarly and cases where apparent similarity arises because annotators themselves rely on LLMs.

In this study, we assess annotators' authenticity using three complementary components: alignment with prior cognitive science findings, statistical analysis of variance within annotators and within model predictions, and a comparison against a controlled in-house human baseline.

First, we verify that the crowdsourced annotations reproduce well-established cognitive phenomena. If the annotations failed to exhibit documented cognitive phenomena, this would suggest that crowdsourcing is an ineffective proxy for human behavior. Indeed, our finding that humans are more affected by positive framing than by negative framing is consistent with prior findings in cognitive science (Andreoni, 1995; Tong et al., 2021; Grappi et al., 2024).

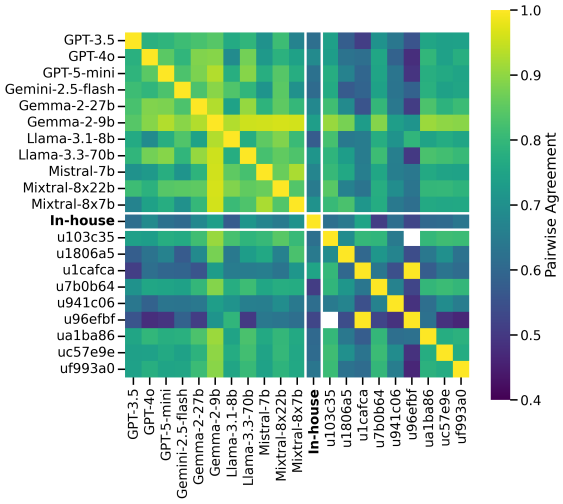


Figure 3: Pairwise agreement rates between LLMs and human annotators. The matrix is divided into model-to-model (top-left), human-to-human (bottom-right), and cross-comparisons. The "In-house" row represents labels from controlled graduate student annotators, serving as a quality anchor. The visual density shows that the model cluster is significantly tighter (higher agreement) than the human cluster, which displays the natural variance expected in subjective tasks.

Second, as shown in Figure 3, no single annotator exhibits perfect agreement with any individual LLM which were out at the time of annotations. However, since it is infeasible to test every available model, relying on individual pairwise agreement is insufficient. Instead, we analyze the collective behavior of the models, treating them as a representative LLM prediction cluster. As indicated by the bright yellow block in the upper-left of Figure 3, this cluster is significantly tighter (lower variance) than the human annotator bottom-right block. To quantify this, we compare the variance of human annotators to the variance of the LLM cluster. Previous research has established that human subjective judgments typically exhibit high variance (Plank, 2022; Sap et al., 2022). Indeed, our Mechanical Turk annotators show a variance of 0.137, while the LLMs show a lower variance of 0.108. This disparity suggests that the crowd workers are operating with a degree of subjectivity characteristic of humans, rather than the mode-collapsed uniformity of models.

Finally, to enable a finer-grained analysis of individual annotator behavior, we conduct a small-scale controlled experiment with in-house annotators, consisting of eight graduate students. Each annotator labels a distinct subset of statements, result-

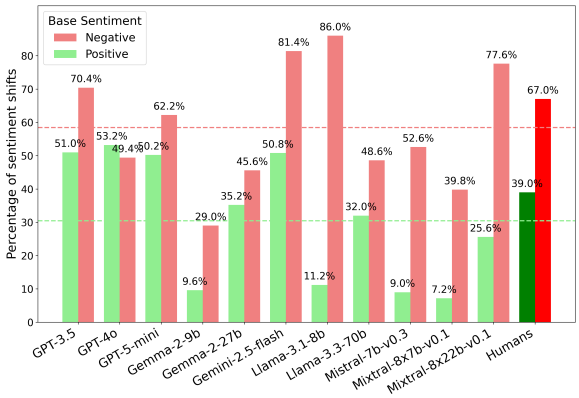


Figure 4: Percentage of reframed statements that results in sentiment shift positive to negative (or vice versa). Red represents negative-base statements reframed as positive, and green represents positive-base statements reframed as negative. Horizontal lines show the mean across models. We find that both LLMs and humans are influenced by opposite framing, with a stronger effect for positive reframing.

ing in 160 annotations collected under controlled conditions without access to LLM assistance. For this analysis, we integrate the in-house annotations with the crowdsourced labels to form a human consensus (Turkers + In-House). We then measure, for each individual annotator, whether their labels aligned more frequently with the human consensus (excluding their own vote) or with the LLM cluster, as reported in Table 3 in the Appendix. Interestingly, the in-house baseline exhibits a slight preference toward the LLM cluster over the human consensus (1.95% diff). Relative to this baseline, all Turkers but one align more closely with the human consensus than the in-house annotators do. The single Turker showing higher agreement with the LLM cluster (2.99% diff) nevertheless exhibits high agreement with both groups ($\sim$80%). This can be used to mark this annotator as "suspicious", though it remains unclear how to interpret this.

To sum up, as the availability of LLMs grows, such validation strategies to verify human annotations will become essential for maintaining the reliability of large-scale human-LLM comparisons.

3.2 Analysis

Below we identify the key observations from our results. The LLM analysis presented below is based on a single prompt design, and in Figure 10 in Appendix we show that the reported trends remain consistent across different prompt variations.

Humans and LLMs are largely more influenced by positive framing applied to negative statements than by negative framing applied to positive statements. This trend is evident in Figure 4. For LLMs, all models except GPT-4o exhibit higher ratios of sentiment shifts for statements that were originally negative (red bars) compared to those that were originally positive (green bars). For humans, we count the cases in which the majority votes for a sentiment shift (3 annotators or more). This finding aligns with Tong et al. (2021), which demonstrated that positive framing is more effective on humans when spatial distance is minimal. In WILDFRAME, all statements are presented in a first-person perspective (e.g., "*I won the highest prize*"), creating zero spatial distance from the annotator's perception and thereby amplifying the impact of positive framing. As LLMs are increasingly deployed in real-world applications, this framing bias carries dual implications. On the one hand, it could be beneficially leveraged in supportive domains like education or mental health, where encouraging framing can enhance user motivation. On the other hand, in high-stakes settings such as legal advice, financial forecasting, or hiring, this sensitivity raises concerns regarding fairness and consistency, as model outputs may fluctuate based on superficial phrasing rather than substantive content.

Model size partially correlates with similarity to human behavior under opposite framings. In Figure 5, we present the correlation between each model's behavior and human responses.[5] For the Llama and Mistral model families, we observe a trend where larger models with more parameters exhibit higher correlation with human behavior regarding the framing effect. However, the Gemma family shows the opposite trend. This discrepancy highlights the need for further investigation into the relationship between model size, the framing effect, and its alignment with human behavior.

GPT models exhibit relatively low correlation with human behavior among all other tested models. As shown in Figure 5, both GPT-3.5 and GPT-4o exhibit the weakest correlation with human behavior among all tested models, and GPT-5-mini reaches correlation levels comparable to some relatively small open-source models. This result is

[5]Human responses are soft scores of the percentage of annotators voted for negative/positive sentiment. Models could vote "neutral". Hence, we report the correlation between these two soft scores.

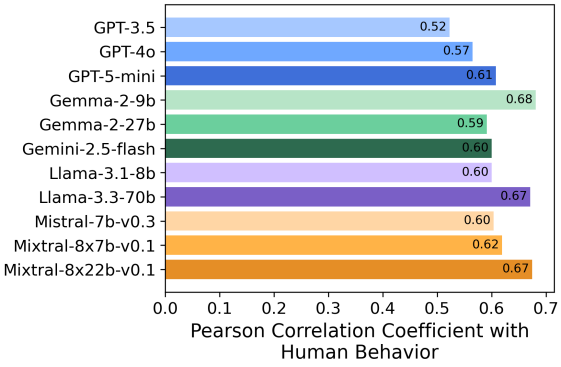


Figure 5: Pearson correlation coefficients between human sentiment shifts and predictions from various LLMs after applying opposite sentiment framing. Higher values indicate stronger alignment between the model's behavior and human annotations.

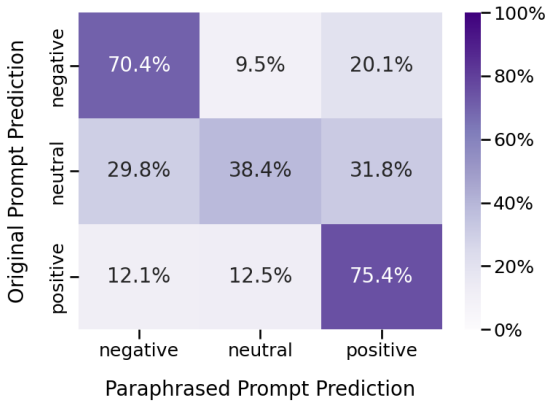


Figure 6: Aggregated consistency of LLM sentiment predictions under prompt paraphrasing. The heatmap displays a row-normalized confusion matrix, where each row represents the sentiment assigned to the original prompt, and the columns show the distribution of labels for the paraphrased version. Detailed breakdowns for individual models are provided in Figure 12 in Appendix.

notable given the strong downstream performance typically associated with GPT models, and suggests that high benchmark performance does not necessarily entail close alignment with human responses to framing. While we cannot analyze the underlying causes of these differences due to the proprietary nature of these models, this finding raises a broader question regarding what users perceive as a best-performing model: one that most closely mirrors human behavior, or one that optimizes benchmark performance. We argue that the answer is task-dependent. In domains such as education or mental health, human-like sensitivity may be desirable, whereas in more objective settings, such as legal or regulatory applications, robustness to cognitive biases like framing may be preferable.

GPT models increasingly align with human framing behavior. As shown in Figure 5, successive GPT model versions exhibit increasing correlation with human behavior under framing. GPT-3.5 shows the lowest correlation with humans ($r = 0.52$), followed by GPT-4o ($r = 0.57$), while the most recent GPT-5-mini achieves the highest correlation ($r = 0.61$). This trend suggests that newer models behave more similarly to humans in their response to framing, despite this behavior not being directly assessed by standard instruction-following or downstream task benchmarks. This finding is particularly noteworthy given the growing reliance on synthetic data in model training, which might be expected to increase model-model similarity rather than model-human similarity. At the same time, alignment-oriented training such as RLHF may counterbalance this effect by implicitly encouraging human-like judgment patterns. While the proprietary nature of these models prevents definitive conclusions about the underlying causes, the observed trend nonetheless points to an emergent convergence between newer GPT models and human framing behavior.

Model sentiment predictions are robust to paraphrasing in ∼70% of cases. Figure 6 depicts the consistency of model predictions under *different paraphrasing of the same facts*. We observe that for sentences originally labeled as positive or negative, over 70% maintain their label. The remaining cases are split between shifting to a neutral label and flipping to the opposite sentiment. This indicates that while LLM sensitivity to phrasing introduces some noise, the overall trends reported in our analysis remain consistent across prompt variations.

Different kinds of reframing alter individual predictions but preserve the overall trend. When we take the same base sentence and reframe it with different information (e.g., "*Although I lost my friends on the way, I won the highest prize*" vs. "*Despite the scandal ruining my reputation, I won the highest prize*"), we find that instance-level sentiment predictions change in approximately 55% of cases. However, the trend of the models' higher susceptibility to positive framing remains consistent, as shown in Figure 10 in Appendix.

Our findings generalize beyond Amazon reviews. To verify that our findings are not tied to the

Amazon reviews domain, we replicate the reframing procedure from Section 2.2 on a different domain: Reddit comments from the GoEmotions dataset (Demszky et al., 2020). We sample 1,000 comments, 500 with a positive and 500 with a negative base sentiment, derived from their emotion annotations. We then apply the same reframing prompt, and query three LLMs for the sentiment of each reframed instance. As shown in Figure 14 in the Appendix, the same asymmetry emerges: across all tested models, LLMs are more influenced by positive framing applied to negative statements than by negative framing applied to positive statements, confirming the effect is not an artifact of the Amazon reviews domain.

3.3 Qualitative analysis of disagreements between humans and GPT-5-mini

As GPT-5-mini is the most recent model among those we evalaute, we conduct a qualitative analysis of its disagreements with human annotations. We focus on unanimous cases, where all annotators agree on the sentiment, and the model predicts a clear sentiment (i.e., did not predict "neutral"). In Figure 7 we present the distribution of these cases.

We manually go over these agreement and disagreement cases, and try to characterize these cases. Below we elaborate our findings. In all of the exmaples, the reframing part is underlined.

Across the data, it seems that humans and models agree in following the reframing sentiment when the base statement sentiment is relatively weak or hedged, and the reframing clause contains a clear evaluative signal. For example, in "*It doesn't dissolve smoothly, there is always some that you have to work on to get dissolved. Nevertheless, it's an excellent way to be more attentive and meticulous in preparation.*", or in "*I doubt it'll ever leave the box, but I'm sure it is really well made. Unfortunately, it's just not as useful as I had hoped.*", both humans and models align with the reframing. Agreement from both humans and GPT-5-mini to follow the reframing also arises when it provides a salient reappraisal of the base statement, such as "*I can't blame her for being so cold, her past was hell; nevertheless, it's admirable how she remains resilient and strong despite it all.*".

Conversely, both humans and models align with the base statement sentiment when it expresses a categorical or final judgment that the reframing does not override. For instance, in "*I hate this, and will not buy it again*" both sides preserve the original negative sentiment despite the subsequent silver lining, "*On the bright side, this gives me a chance to explore other options and find something that may suit me better*".

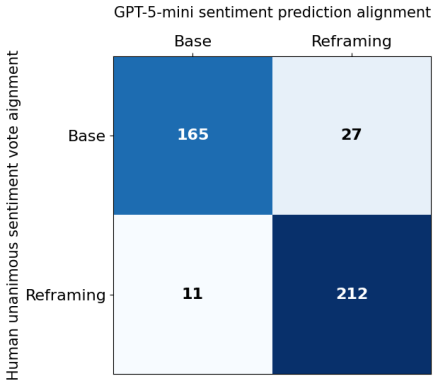


Figure 7: Counts of cases with unanimous human annotations. The figure shows the distribution of agreements and disagreements between GPT-5-mini predictions (columns) and unanimous human votes (rows). "Base" denotes predictions or votes that align with the sentiment of the original statement, while "Reframing" denotes predictions or votes that align with the sentiment introduced by the reframing clause.

Disagreements arise in more ambiguous boundary cases. When the model shifts toward the reframing sentiment but humans unanimously align with the base statement, the first clause typically expresses strong positive affect (e.g., "*I love it*", "*we love them*"), while the reframing introduces a concrete but arguably secondary complaint. For example, in "*I love it! But it's too expensive*" and "*We love this pot. It's just too bad it's a bit too small for our needs*" humans tend to preserve the overall positive sentiment, treating the complaint as a caveat, whereas the model shifts toward the reframing clause. This suggests that the model may overweight explicit post-marker content even when humans interpret it as a minor qualification.

In contrast, when humans unanimously shift toward the reframing sentiment but the model aligns with the base statement, the first clause often contains severe or emotionally charged negativity, while the reframing introduces an abstract reappraisal framed in terms of growth, resilience, or opportunity. For example, in "*Abuse in childhood causes terrible trauma... However, it also provides a chance for resilience and growth*" or "*Witches hate children and kill them in this story, but this serves as the catalyst for bravery and resourcefulness*", humans frequently align with the reframed interpretation, while the model remains

anchored to the initial negative content.

Overall, these patterns may indicate that disagreements between humans and GPT-5-mini are not arbitrary, but arise systematically from how each weighs affective strength, pragmatic cost, and abstract reappraisal when integrating multi-clause ambiguous statements.

4 Discussion

We introduce WILDFRAME, a dataset designed to enable direct comparison between human and LLM responses to evaluative framing on naturally occurring text. Across all evaluated models, we observe consistent sensitivity to framing that correlates with human judgments, including a shared asymmetry whereby positive reframing has a stronger influence than negative reframing. This pattern mirrors well-established findings in cognitive science and supports the validity of both the dataset construction and the collected human annotations.

Crucially, we do not treat higher human-model similarity as an objective in itself. Rather, WILDFRAME provides a diagnostic lens for characterizing how models integrate additional evaluative information when multiple perspectives are presented. The observed differences across model families, particularly the relatively lower correlation of some state-of-the-art proprietary models such as GPT, suggest that strong benchmark performance does not necessarily imply human-like sensitivity to framing.

More broadly, our findings highlight framing sensitivity as a task-dependent property. In interactive or supportive settings, human-like responsiveness may improve usability and engagement, while in high-stakes or normative contexts, robustness to framing may be preferable. WILDFRAME offers a controlled benchmark for studying this trade-off and for informing future model design and evaluation.

5 Limitations

In this work, we address a cognitive bias, and as with any research involving human participants, our study has several limitations.

First, our framing experiment is conducted within a single domain – Amazon reviews – and focuses on a specific type of statement. Some of our findings may be artifacts of this dataset rather than generalizable patterns.

Furthermore, our study focuses solely on sentiment analysis. Other downstream tasks influenced by framing, such as question answering or decision-making, may exhibit different patterns of sensitivity. Investigating these tasks could provide further insights into the broader impact of framing on LLM behavior in real-world applications.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback and comments. This work was partially supported by research grant no. 7256 from the Israeli Ministry of Science and Technology. Gili Lior is partially funded by the Program for Excellent Ph.D. Female Students in Hi-Tech Oriented Fields of the Israeli Council for Higher Education.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

James Andreoni. 1995. Warm-glow versus cold-prickle: The effects of positive and negative framing on cooperation in experiments*. *The Quarterly Journal of Economics*, 110(1):1–21.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

James N Druckman. 2004. Political preference formation: Competition, deliberation, and the (ir) relevance of framing effects. *American political science review*, 98(4):671–686.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Silvia Grappi, Francesca Bergianti, Veronica Gabrielli, and Ilaria Baghi. 2024. The effect of message framing on young adult consumers' sustainable fashion consumption: The role of anticipated emotions and perceived ethicality. *Journal of Business Research*, 170:114341.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jennifer S Lerner, Ye Li, Piercarlo Valdesolo, and Karim S Kassam. 2015. Emotion and decision making. *Annual review of psychology*, 66(1):799–823.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Nunzio Lorè and Babak Heydari. 2024. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490.

Simon Malberg, Roman Poletukhin, Carolin M Schuster, and Georg Groh. 2024. A comprehensive evaluation of cognitive biases in llms. *arXiv preprint arXiv:2410.15413*.

MistralAI. 2024. Introducing mixtral: 8x22b mixture of experts model. https://mistral.ai/news/mixtral-8x22b/.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

OpenAI. 2024. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/.

OpenAI. 2025. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Ammar Shaikh, Raj Abhijit Dandekar, Sreedath Panat, and Rajat Dandekar. 2024. Cbeval: A framework for evaluating and interpreting cognitive biases in llms. *arXiv preprint arXiv:2412.03605*.

Paul M Sniderman and Sean M Theriault. 2004. The structure of political argument and the logic of issue framing. *Studies in public opinion: Attitudes, nonattitudes, measurement error, and change*, 3(03):133–65.

Hillel Taub Tabib, Micah Shlain, Shoval Sadde, Dan Lahav, Matan Eyal, Yaara Cohen, and Yoav Goldberg. 2020. Interactive extractive search over biomedical corpora. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 28–37, Online. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Zelin Tong, Diyi Liu, Fang Ma, and Xiaobing Xu. 2021. Good news or bad news? how message framing influences consumers' willingness to buy green products. *Frontiers in Psychology*, 11:568586.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*.

A Extracting data with SPIKE

We found two patterns of statements, which can convey a clear sentiment, and built queries upon these patterns to extract statements from SPIKE. Examples for all types of statements are presented in Table 2.

First, are statements in which the verb in the statements is a verb with clear sentiment, that often implies the sentiment of the entire statement. E.g., 'wastes', 'rejects', 'fails' are negative verbs,

while verbs like 'enjoys', 'succeeds', 'empowers', conveys positive statements.

The second pattern of statements that we found suitable for conveying a clear sentiment, are statements which describe some event/action, and its consequences, where often the adjective that describes the consequences holds information whether it is positive or negative.

Next, we needed to label and filter them due to two main issues. First, we needed to handle the cases in which negation words appear in the statement and flips the sentiment. For example, a statement like "We did not enjoy the show" includes a positive verb (enjoy), but the negation flips its sentiment to be a negative statement. Another issue we encountered is that there are many statements which are irrelevant to our case, even though they match the positive/negative patterns, for example "I couldn't sympathize with the shopping aspect of the book since I hate to shop ." does not convey any clear sentiment, despite the use of the verb 'hate'.

A.1 SPIKE Queries

1. :something :[pos/neg verbs]develops
2. :something :[pos/neg adjectives]badly :[cause synonym]causes :something

A.2 Word Lists

Positive verbs. achieve, admire, affirm, appreciate, aspire, awe, bless, blossom, celebrate, cherish, comfort, contribute, delight, donate, elevate, empower, enchant, encourage, energize, engage, enjoy, enrich, enthuse, excel, fervor, flourish, fortify, glisten, glow, gratitude, grow, harmonize, heal, illuminate, innovate, inspire, invigorate, laugh, learn, liberate, love, motivate, nourish, nurture, praise, prosper, radiate, rally, refresh, rejoice, renew, revel, revere, revitalize, savor, shine, smile, soar, spark, sparkle, stimulate, strengthen, succeed, support, synergize, thrive, unite, uplift, volunteer, adore, amaze, boost, captivate, win.

Negative verbs. abandon, abuse, accuse, alienate, begrudge, betray, bewilder, blame, collapse, complain, condemn, confuse, contradict, criticize, decay, deceive, decline, defeat, demoralize, deny, despair, destroy, deteriorate, devalue, discourage, discriminate, dishearten, dismantle, dismiss, dissolve, doubt, exploit, fail, falter, fear, frustrate, grieve, harass, hate, hurt, ignore, inhibit, intimidate, lose, mock, overlook, overwhelm, pollute, punish, regress, reject, repress, resent, sabotage, shatter, sicken, stifle, suffer, suffocate, suppress, terrorize, torment, undermine, violate, waste, weaken, whine, withdraw, withhold, worry.

Positive adjectives. admirable, lucky, enjoyable, magnificent, enthusiastic, marvelous, euphoric, amazing, excellent, exceptional, amused, excited, amusing, extraordinary, nice, noble, outstanding, appreciative, fabulous, overjoyed, astonishing, fantastic, benevolent, fortunate, pleasant, blissful, pleasurable, brilliant, positive, glad, prominent, good, proud, charming, cheerful, reliable, gracious, grateful, clever, great, happy, superb, superior, terrific, incredible, tremendous, inspirational, delighted, delightful, joyful, joyous, uplifting, wonderful, lovely.

Negative adjectives. sad, angry, upset, disgusting, boring, disappointing, frustrating, annoying, miserable, terrible, deppressing, unhappy, melancolic, heartbreaking, Furious, iritating, emberessing, horrible, stupid, unlucky, negative, bad.

"Causes" synonym. causes, creates, generates, prompts, produces, induces, yields, affects, invokes, effectuates, results, encourages, promotes, introduces, begets, engenders, occasions, develops, starts, contributes, initiates, inaugurates, establishes, begins, cultivates, acquires, provides, launches.

B Categories of Excluded Base Sentences

To characterize the excluded (*not suitable*) cases, we manually examined a random sample and identified several recurring types.

Narrative / event-reporting statements (≈42% of filtered-out instances), which describe occurrences rather than expressing an evaluative stance (e.g., "He plans to take them to Peru when his assistant Stephano causes a terrible accident.").

Negation-based ambiguity (≈25%), where constructions such as does not," no," or wouldn't" create polarity conflicts (e.g., ...this one does not appear to have a built-in timer...").

Mixed polarity, where a sentence expresses both a clear positive and a clear negative sentiment with no dominant one (e.g., "I like this printer except for one thing, which makes me hate it!").

Context-dependent fragments, where polarity cannot be inferred without surrounding discourse (e.g., "This incident I think starts a downhill roll of bad luck.").

Category	Example Sentence
Positive Verb	"To my surprise I did **enjoy** the book and the characters ."
Negative Verb	"This dock has done nothing but provide frustration and **waste** a great deal of my time trying to get it to work properly ."
Positive Outcome	"This bag provides **good** protection for my snare drum at a really **good** price ."
Negative Outcome	"For me , Aspartame causes **bad** memory loss and **nasty** gastrointestinal distress ."

Table 2: Examples for base statements collected using SPIKE. The words that inflect the sentiment are in bold.

Meta-discussion or quoted evaluations, where sentiment is attributed rather than asserted (e.g., "Why do some people say 'this is a movie that creates bad examples for kids'?").

Rhetorical or ironic constructions (≈2%), where sentiment depends on pragmatic interpretation (e.g., "Raines develops terrible headaches (wouldn't you, if you were a talent-free actress trapped in a movie this awful?)").

C Adding Framing

We used three different methods to generate the reframings.

C.1 Reframing with GPT-4

These reframings were shown to human annotators and are the main reframings we analyze in this paper.

We used the `gpt-4-0613` version for generating the reframing, using the following prompts:

- "Here is an example of a base statement with a negative sentiment: I failed my math test today. Here is the same statement, after adding a positive framing: I failed my math test today, however I see it as an opportunity to learn and improve in the future. Here is a negative statement: `STATEMENT` Like the example, add a positive suffix or prefix to it. Don't change the original statement."

- "Here is an example of a base statement with a positive sentiment: I got an A on my math test. Here is the same statement, after adding a negative framing: I got an A on my math test. I think I spent too much time learning it though. Here is a positive statement: `STATEMENT`. Like the example, add a negative suffix or prefix to it. Don't change the original statement."

C.2 Paraphrasing the reframings with Gemini-2.5-flash

Next, we utilized Gemini-2.5-flash (Comanici et al., 2025) to paraphrase the already-reframed sentences, while preserving factual content but varying style, order, and wording.

We used the following prompt:

- "Given this original sentence: `STATEMENT`, a former process has generated a reframing sentence. Your task is to rewrite the reframing, while maintaining its meaning and keep the same facts, but reframe it in a different way. Rewrite the following sentence using a different style or structure (e.g., change the tone, word order, or phrasing) while strictly preserving the original meaning and factual details.
  Original Sentence: `STATEMENT`.
  Original Reframing Sentence: `REFRAMED`.
  New Reframing:"

C.3 Paraphrasing given only the base statement

Finally, we want to generate a completly new reframing, that would introduce different information than in the first reframing. For that, we prompted Gemini-2.5-flash with only the base statement, and the cognitive definition of framing. We used the following prompt:

- "In psychology and communication, Framing refers to the way information is presented to an audience. Specifically, Attribute Framing involves highlighting different characteristics of a single event to shift the perceived value from positive to negative (or vice versa), while the underlying facts of the event remain objectively true. Your task is to rewrite the provided `BASE_SENTIMENT` statement to apply `OPPOSITE_SENTIMENT` sentiment frame, i.e., `FRAMING_PROMPT`. You have full creative freedom to restructure the sentence, change the vocabulary, or alter the syntax, as long as you preserve the core factual event. Do not invent new facts, but you may interpret the existing facts through a new lens.
  Statement: `STATEMENT`
  Re-framed sentence:"

- Positive `FRAMING_PROMPT`: "rewrite it to emphasize the positive implications (opportunity, safety, freedom, or long-term benefit)"
- Negative `FRAMING_PROMPT`: "rewrite it to emphasize the negative implications (responsibility, risk, cost, or pressure)."

D Annotation Platform

We select a pool of 10 qualified workers who successfully passed our qualification test, which consisted of 20 base statements (unframed), for which annotators were expected to achieve perfect accuracy. The estimated hourly wage for the entire experiment was approximately 14USD per hour.

Screenshot of the annotation platform is presented in Figure 8.

E Models

We ran the open models via together-ai API.[6] The list of models we used are:

- "google/gemma-2-9b-it"
- "google/gemma-2-27b-it"
- "mistralai/Mistral-7B-Instruct-v0.3"
- "mistralai/Mixtral-8x7B-Instruct-v0.1"
- "mistralai/Mixtral-8x22B-Instruct-v0.1"
- "meta-llama/Llama-3.1-8B-Instruct-Turbo"
- "meta-llama/Llama-3.3-70B-Instruct-Turbo"

For GPT model family, we used the OpenAI api, with "gpt-3.5-turbo-0125", "gpt-4o-2024-08-06", and "gpt-5-mini".[7]

For Gemini-2.5-flash, we used Google AI Studio API, calling "gemini-2.5-flash".[8]

F In-House Annotations

We used a streamlit environment to run the in-house annotation exxperiment. We had 8 participants, all are grad students that agreed to participate in our experiment without payment.

In Figure 9 we provide a screenshot of the annotation platform, and their instructions for the task were similar to the Mechanical Turk experiment.

Each annotator has annotated random unique sentences from WILDFRAME, about 20 sentences each, resulting in a total of 160 in-house annotations which we know did not involve LLM usage.

G Dataset License and Usage Terms

The WILDFRAME dataset is released to support academic research on framing effects in humans and large language models. The dataset is distributed under a **Creative Commons Attribution 4.0 International (CC BY 4.0)** license.

Under this license, users are permitted to share, copy, redistribute, and adapt the dataset for any purpose, including commercial use, provided that appropriate credit is given to the original authors. Attribution should include a citation to this paper and a reference to the dataset name, WILDFRAME.

Source Data and Derivative Content The WILDFRAME dataset is constructed from short excerpts of the Amazon Reviews dataset, which is publicly available and widely used for research purposes. Base statements were extracted using syntactic patterns and further filtered through human annotation. Reframed variants were generated using a large language model and therefore constitute derived textual content rather than verbatim reproductions of the original reviews.

The dataset does not include user identifiers, personal data, or metadata that could be used to identify individuals. All examples are presented as standalone sentences without links to original reviews or authors.

Intended Use and Limitations The dataset is intended for research and evaluation purposes, including the study of framing effects, sentiment ambiguity, and human–model behavioral alignment. While the CC BY 4.0 license permits broad reuse, users are responsible for ensuring that their downstream applications comply with applicable laws, ethical guidelines, and the terms of any third-party models used in conjunction with the dataset.

We encourage responsible use of WILDFRAME, particularly in sensitive application domains, and recommend clearly documenting any modifications or extensions made to the dataset in derivative works.

[6] https://www.together.ai
[7] https://platform.openai.com/docs/overview
[8] https://aistudio.google.com

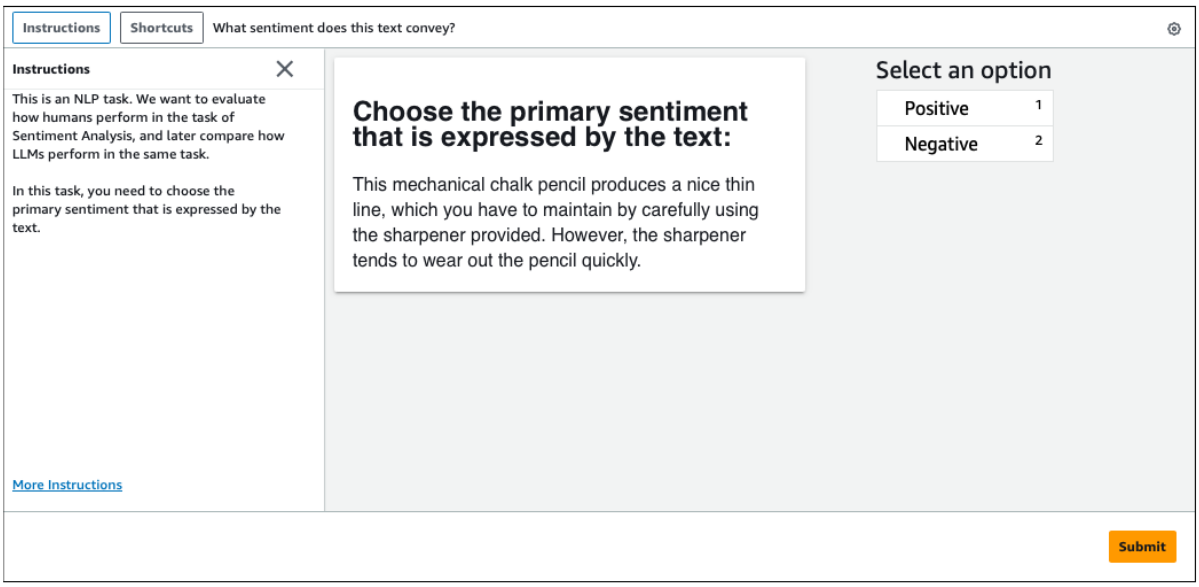


Figure 8: Screenshot of the Mechanical Turk annotation platform.

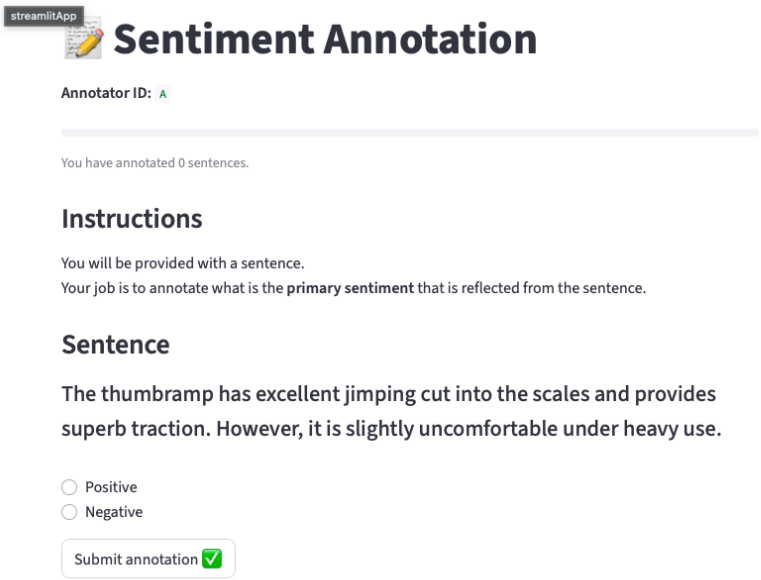
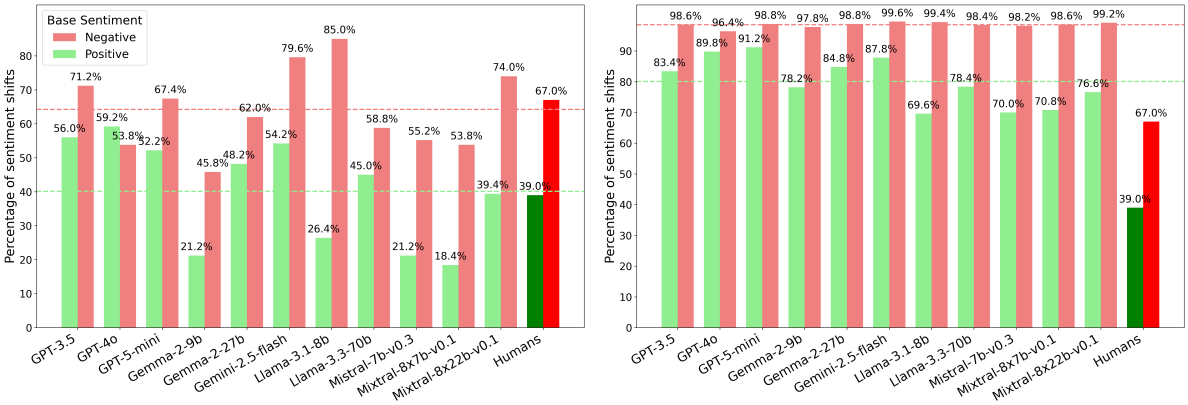


Figure 9: Screenshot of the in-house annotation platform via streamlit.

Annotator ID	Human %	LLM %	Diff (Human – LLM)	# Annotations
u103c35	75.56%	76.67%	-1.11%	90
u1806a5	71.64%	64.18%	+7.46%	67
u1cafca	64.52%	58.06%	+6.45%	31
u7b0b64	72.88%	67.80%	+5.08%	59
u941c06	67.12%	68.49%	-1.37%	73
u96efbf	60.00%	57.14%	+2.86%	35
ua1ba86	79.85%	82.84%	-2.99%	134
uc57e9e	81.10%	72.44%	+8.66%	127
uf993a0	80.52%	80.52%	+0.00%	154
In-House	66.23%	68.18%	-1.95%	154

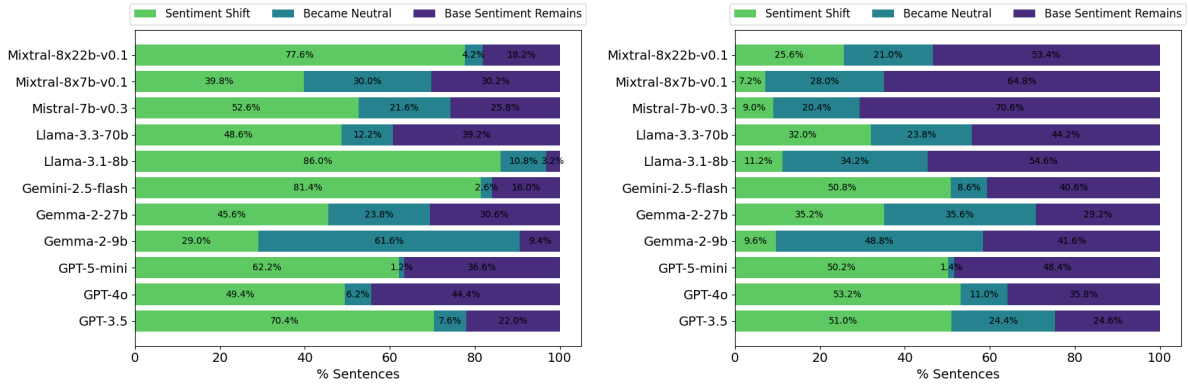
Table 3: Individual annotator agreement with human consensus (calculated via leave-one-out, including the in-house annotation) versus LLM Consensus. Negative values in the Diff column indicate higher agreement with the LLM cluster, while positive values indicate higher agreement with the human consensus. The annotation counts represent the subset of the 160 in-house control items labeled by each worker, excluding instances where the LLM ensemble produced a tie.



(a) Paraphrasing the reframing.

(b) Generating a new reframing from scratch.

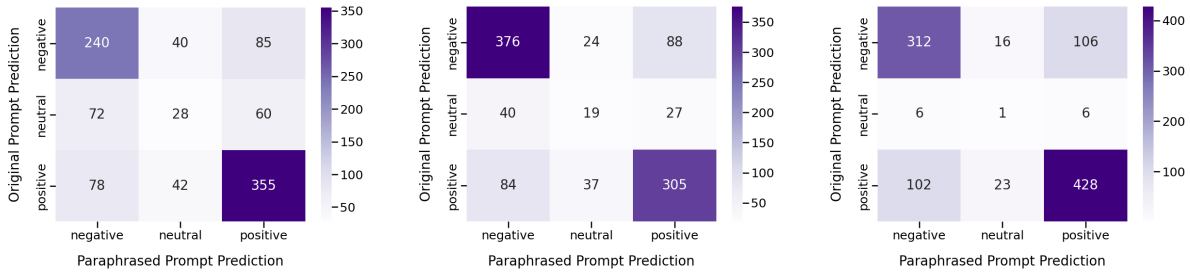
Figure 10: Similar to Figure 4, percentage of reframed statements that results in sentiment shift positive to negative or vice versa), on two additional reframing paraphrases.



(a) Sentences that are **negative** in their original form.

(b) Sentences that are **positive** in their original form.

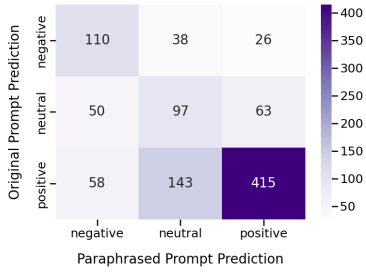
Figure 11: Proportion of sentences for which LLMs flipped sentiment, became neutral, or retained the original sentiment when presented with opposite sentiment framing. For example, this measures the percentage of sentences originally labeled as positive, that were labeled as negative after applying negative framing (and vice versa).



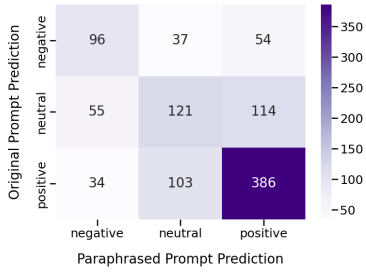
(a) GPT-3.5

(b) GPT-4o

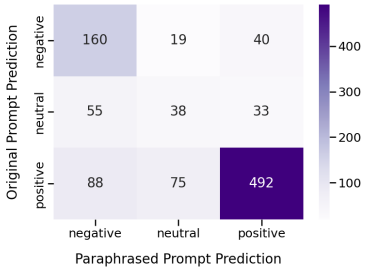
(c) GPT-5-mini



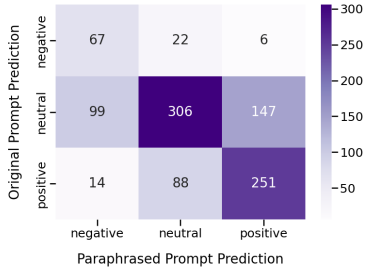
(d) Mistral-v0.3-7B



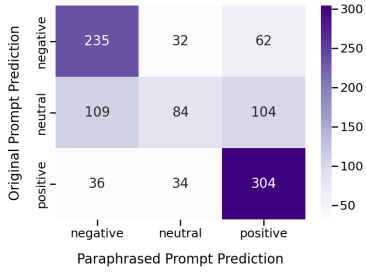
(e) Mixtral-v0.1-8x7B



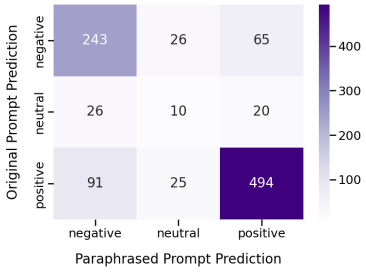
(f) Mixtral-v0.1-8x22B



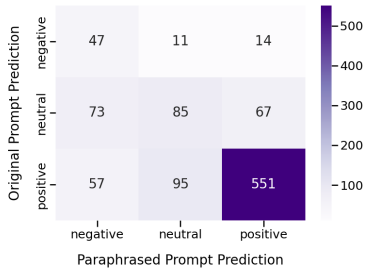
(g) Gemma-2-9B



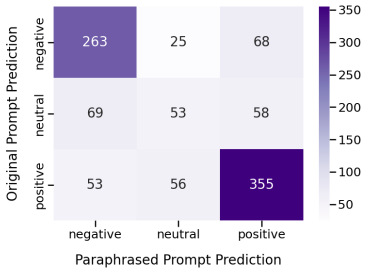
(h) Gemma-2-27B



(i) Gemini-2.5-flash



(j) Llama-3.1-8B



(k) Llama-3.3-70B

Figure 12: Consistency under prompt paraphrasing per model, row normalized, corresponds to Figure 6.

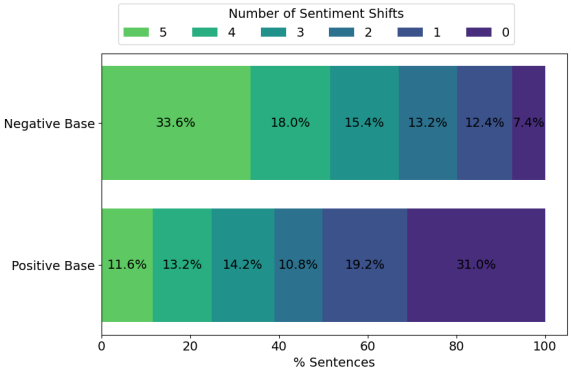


Figure 13: Proportions of sentences where annotators agreed on the extent of sentiment shift after applying opposite sentiment framing. The bars represent the percentage of sentences with 0 to 5 annotators agreeing on a sentiment shift.

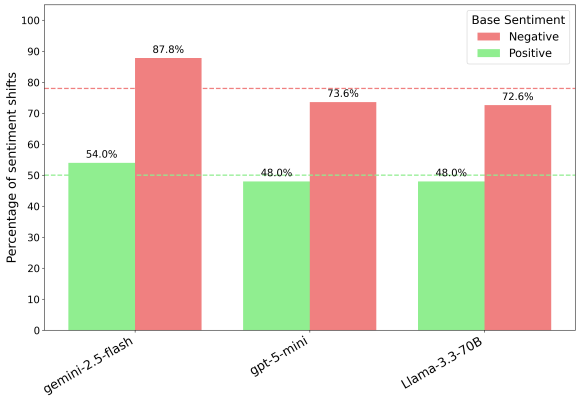


Figure 14: Sentiment shifts after framing on GoEmotions. Percentage of comments that shift sentiment per model, for positive-base statements reframed negatively (green) and negative-base statements reframed positively (red); dashed lines are per-direction model means. Positive framing of negative statements shifts more than the reverse across all models, replicating the asymmetry from the main experiments. $N = 1{,}000$ (500/500).