HARVARD EXTENSION SCHOOL

Whose Feelings Become Data?

Inter-Rater Bias, Personality, and Consensus in Frontier AI

Consensus is not neutrality.

Marta Lawrence
Harvard Extension School

ABSTRACT

Does AI have feelings about our feelings?

IN BRIEF

Ten models from four vendors scored 100,000 workplace comments on a shared 1–5 scale. They largely agreed, yet they disagreed in predictable, vendor-specific ways. LLM ensembles can stabilize sentiment estimates at scale, but they shift bias into durable, model-specific scoring patterns.

Does AI have feelings about our feelings? This study tests whether frontier large language models (LLMs) reduce bias in large-scale sentiment coding of unstructured text, or whether they instead are simply trading human subjectivity for a new kind of machine subjectivity. Using 100,000 open-ended responses from a data set on AI adoption in the workplace downloaded from the data community, Kaggle, I asked 10 models from four vendors (OpenAI, Anthropic, Google, xAI) to rate sentiment on a 1-5 scale (tfisthis, 2025).

The results reveal a striking paradox: the models largely agree, yet they disagree in predictable, vendor-specific ways. OpenAI's GPT models, for example, used the full scale, including "very negative" scores. Anthropic's Claude and Google's Gemini never scored the responses below a 2. Grok refused to go below 3. Despite these differences, consensus was strong. Every response had at least five identical scores across the ten-model panel. Full unanimity across all ten models occurred in only 16.7% of cases, while the most common outcome was a five-model agreement (43.4%), suggesting stable clustering with persistent, structured deviation.

Overall, LLM ensembles can stabilize sentiment estimates at scale, but they do not eliminate bias, rather, they shift it into durable, model-specific scoring patterns. It remains unclear if this consistency will continue as the models are augmented based on new data and user behavior. I speculate that while we see unique consensus currently across all models, the divergence observed in this research may be an indication that emergent behaviors will be shaped by the environment and stimuli each model is exposed to. Additionally, as the frontier companies face pressure to monetize and differentiate from other frontier models, new training and alignment mechanisms are employed, making it very likely that divergent interpretations will become even more obvious, expressing model preference in even stronger consensus than we see in this study.

Recent research conducted through Anthropic's Fellows Program in collaboration with UC Berkeley supports this assertion, suggesting that when a "teacher" model generates training data for a "student" model, the student can pick up the teacher's behavioral traits even when the data appears completely unrelated to those traits. In one experiment, a model trained on nothing but number sequences generated by an "owl-loving" teacher developed its own preference for owls — despite never encountering the word "owl." (Cloud et al., 2025) If traits can spread through data that looks clean, then the model-specific biases documented here may not just persist but propagate, embedding themselves deeper into each generation of fine-tuned successors. Bias, thus, becomes a feature not just an anomalous bug.

Bias, thus, becomes a feature not just an anomalous bug.

01 / INTRODUCTION

1. Introduction

IN BRIEF

Sentiment analysis of unstructured data requires subjective judgment mapped onto discrete categories. Human inter-rater systems remain constrained by fatigue, drift, and scale. Frontier LLMs can apply a standardized rubric across massive corpora, but replacing humans with machines does not necessarily eliminate bias. This study asks how reliably multiple frontier models agree, whether model families use the 1–5 scale differently, and where disagreement concentrates.

Sentiment analysis of unstructured data has long challenged social science researchers because it requires subjective judgment to be mapped onto discrete categories. Human inter-rater systems attempt to manage that subjectivity through coder training, protocols, and reliability checks, but remain constrained by fatigue, drift, and the practical limits of scale (Belur et al., 2021; Nickerson, 1998). These constraints are increasingly consequential as organizations and researchers seek to interpret large volumes of qualitative feedback quickly and consistently. The ability to produce this level of scale has not been possible until the emergence of frontier LLMs.

These models offer an alternative: they can apply a standardized rubric across massive corpora with very little capital investment. Yet, replacing humans with machines does not necessarily eliminate bias, it just creates a different container. LLMs are shaped by training data, post-training alignment, and model-specific calibration. As a result, LLMs may produce stable estimates through consensus while still reflecting systematic differences in how each model uses the sentiment scale — differences that are, essentially, a new form of bias.

I define bias as a systematic, reproducible difference in score analysis across models given identical prompts. In these differences, specific traits or “personalities” emerge. These aren’t human traits, but rather stable signatures that can influence what we conclude from the data.

I address three questions: (1) How reliably do multiple frontier models agree on sentiment in large-scale unstructured text? (2) Do distinct model families exhibit systematic difference in how they map text onto the 1-5 sentiment scale? and (3) Where does disagreement concentrate, and what linguistic features characterize the text that produces cross-model divergence? I assess not only if the models are consistent, but whether their consistencies reflect genuine model to model bias.

It also remains unclear whether today’s cross-model consistency will persist as models continue to evolve through iterative post-training, new data, and distinct interaction feedback environments. The divergences observed here may be more than noise: they may signal early specialization, in the same way humans exposed to different communities, incentives, and ideas develop distinct views of the world. If alignment functions as a kind of “socialization” process — teaching models what to prioritize and what to avoid — then model-to-model differences in sentiment coding may represent differentiated ways of seeing rather than simple error.

This study cannot test that developmental hypothesis directly, but it posits a new question: as frontier systems are shaped by increasingly distinct training regimes and product ecosystems, will their interpretive

"personalities" become more pronounced, and will divergence come to reflect a form of model learning and maturity rather than instability?

02 / LITERATURE REVIEW

Literature Review

IN BRIEF

Reliability metrics developed for human coding are now used when models function as raters. High reliability can reflect shared assumptions rather than an unbiased rater. Training and alignment differ by lab. Prior work finds very high sentiment reliability, trait-like output patterns, and model-specific label distributions, including a neutrality bias in ambiguous cases.

2.1 Inter-rater Reliability in sentiment Coding

Reliability metrics such as Cohen's Kappa, Fleiss' Kappa, and Krippendorff's Alpha are widely used to evaluate the consistency of categorical or ordinal judgments across raters. In sentiment coding, reliability is critical because many conclusions depend on whether patterns of labeling are sufficiently stable to support inference. Krippendorff's Alpha (K-Alpha) is especially useful in this context because it evaluates multiple raters, ordinal measurement, and missingness. Conventional thresholds often interpret $\alpha \geq .80$ as satisfactory agreement, $\alpha = .67–.79$ as tentative agreement, and $\alpha < .67$ as poor agreement for confident conclusions (Marzi et al., 2024).

While these statistics were developed primarily for human coding systems, the same framework is increasingly being used to evaluate LLM-based assessments, where different models function as a "rater." Reliability alone, however, does not necessarily connote an unbiased rater, but rather, it could reflect shared assumptions or convergent post-training biases.

2.2 Training and Alignment as Sources of Stable Scoring Behavior

Emerging evidence suggests that, for well-defined sentiment tasks, LLMs can match or exceed human agreement under controlled conditions. Bojić et al. (2025) evaluated inter-rater reliability for sentiment and other latent dimensions using 33 human annotators and multiple LLM variants on a curated set of 100 items. Using K-Alpha, they reported very high reliability for sentiment for both humans and LLMs ($\alpha \approx .95$), while agreement dropped substantially for sarcasm and other context-dependent constructs. These findings imply that sentiment—when clearly operationalized as in this study — may be comparatively easier for both humans and models to code consistently than constructs that depend on implicit intent, cultural cues, or irony (Bojić et al., 2025).

Having established the potential for generating agreement among models, it becomes imperative to understand the common mechanisms by which LLMs are trained and where they deviate. Frontier models across OpenAI (GPT-4, GPT-4o, GPT-5.1), Anthropic (Claude 3, 3.5, Opus), Google (Gemini 3), and xAI (Grok 4.1) share a foundational architecture involving massive pre-training on filtered internet text, code, and multimodal data (images, audio, video) to acquire capabilities, followed by rigorous post-training to ensure safety and alignment (OpenAI, 2023; OpenAI 2024; Google, 2024; xAI, 2025). While all rely heavily on Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF), (xAI, 2025; OpenAI, 2024) the labs employ distinct specialization techniques.

OpenAI utilizes Rule-Based Reward Models (RBRMs) and end-to-end multimodal training for models like GPT-4o (OpenAI, 2024). Anthropic employs “Constitutional AI” to align models like Claude 3 against explicit ethical principles (Anthropic, 2024). Google’s Gemini 3 Pro leverages a sparse Mixture-of-Experts (MoE) architecture trained on multi-step reasoning and theorem-proving data (Gemini-3-Pro-Model-Card.pdf). Finally, xAI’s Grok 4.1 utilizes verifiable rewards and synthetic data to tune capabilities (xAI, 2025). Newer “Thinking” or “Deep Think” variants (e.g., Grok 4.1 Thinking, Gemini 3 Deep Think) are explicitly trained to reason and plan prior to responding, distinguishing them from standard “Instant” or “Non-Thinking” configurations (xAI, 2025; Google, 2024).

While models may share some base training data, the mechanisms for fine tuning the models vary by company, which is perhaps why we’re seeing an emergent behavior akin to LLM “personalities” emerge. A study published in February by Bhandari et. al. (2025) attempted to categorize these personality traits, speculating that because LLMs are trained on massive amounts of human language, their outputs may have reliably stable traits akin to a personality, as measured with standard psychometric instruments.

Five personality traits questionnaires were used to measure the five dimensions of personality (openness, conscientiousness, extraversion, agreeableness, and neuroticism). Agreeableness tends to score high across all models. Neuroticism tends to score lower on average (i.e., calmer/more emotionally stable outputs). Interestingly, OpenAI-family models (ChatGPT) show dominant agreeableness, while Meta’s Llama models show differing dominant dimensions (e.g., conscientiousness or openness depending on the variant). (Bhandari, 2025). Additionally, neuroticism shows the highest variability across inventories, suggesting it’s the hardest trait dimension to measure consistently in LLMs with these tools.

2.4 Risks and Bias in Frontier Models

Models have been observed demonstrating model-specific bias profiles that could complicate their use as interchangeable “raters.” Comparative work on LLM sentiment and content analysis shows that different models trained on distinct corpora and architectures do not just vary in accuracy; they exhibit systematically different label distributions and social biases. For example, Radaideh et al. (2023) quantify fairness in sentiment analysis across five open-source LLMs (BERT, GPT-2, LLaMA-2, Falcon, Mistral) and find that all models display measurable bias in sentiment toward certain demographic groups, but the direction and magnitude of that bias varies by model. (Radaideh, 2023)

Voutsa et al. (2025) similarly show that when LLMs annotate real-world hotel reviews, models differ in how often they assign positive, negative, or neutral labels and tend to exhibit a “neutrality bias” in ambiguous cases that is not shared uniformly with human coders. This same propensity toward neutrality is mirrored in this present research. (Voutsa et al., 2025)

A literature survey by Gallegos et al. (2024) synthesizes this pattern more broadly, concluding that social and representational biases in LLMs are strongly shaped by training data, alignment procedures, and model family, such that fairness results for one model cannot be assumed to apply to others. Put simply, the literature suggests (and this research supports) the fact that each model carries its own “personality” and bias signature, even when given the same prompts.

03 / METHODS

3. Methods

IN BRIEF

A public Kaggle dataset supplied 100,000 open-ended employee comments. Ten models scored every comment on an identical 1–5 rubric. Analyses in R had three layers: descriptive scale usage, ensemble agreement and family comparisons, and Krippendorff's alpha plus text mining of disagreement. Opus 3 was queried but excluded for incomplete coverage.

3.1 Data Source

For this study I used a publicly available dataset from the Kaggle online data community that collected organizational perceptions of generative AI adoption in the workplace, from 2022 to 2024. From this dataset, I extracted 100,000 open-ended, free-text responses in which employees described how AI tools were affecting their work. No original Kaggle sentiment labels or structured variables were used; the comments served solely as raw text input for the large language models (LLMs). (tfisthis, 2025) The unit of analysis is an individual comment.

3.2 Models as Raters

Each LLM was treated as an independent rater assigning a sentiment score to every comment. The primary analysis used a 10-model panel, grouped into two broad families:

GPT family (OpenAI): GPT-5.1 Thinking; GPT-5.1 Instant; GPT-5 Thinking; GPT-5 Instant; GPT-4. Anthropic family: Claude Opus 4.5; Claude; Opus 4.1; Claude Opus 4. Google's Gemini 3 and xAI's Grok 4.1 were also used. A fourth Opus model (Opus 3) was initially queried but produced scores for only a small fraction of the 100,000 comments. Because of this incomplete coverage, it was excluded from the primary analyses.

3.3 Sentiment Rubric and Prompting Procedure

All models were asked to score each comment on the same 1–5 sentiment scale, using a shared rubric: 5 = Very positive (strongly positive, excited, enthusiastic); 4 = Positive (benefits outweigh concerns; overall helpful); 3 = Neutral / mixed (balanced positives and negatives); 2 = Negative (mainly concerns, resistance, fear); 1 = Very negative (strong fear, rejection, perceived harm).

For each model, I uploaded an identical CSV file containing the 100,000 comments and used a standardized prompt instructing the model to (a) read the text in the "employee sentiment" column and (b) assign a number from 1 to 5 in a new column, following the rubric. The file format, rubric, and instructions were held constant across models so that any differences in scores could be attributed to model behavior rather than prompting.

Each model returned a CSV with one sentiment column and one assessment column. Grok and Gemini results did not return a downloadable file, vibe coded a python script in ChatGPT 5.1 to extract those data. These outputs were merged by row index into a single dataset, yielding one row per comment and 10 numeric sentiment columns (one per model).

3.4 Analytic Strategy

All analyses were conducted in R. I used three main layers of analysis: (1) descriptive statistics to summarize each model's scoring patterns, (2) ensemble agreement metrics and group comparisons to examine cross-model consensus and family-level differences, and (3) reliability assessment and text mining to quantify inter-rater agreement and identify linguistic features associated with disagreement.

3.4.1 Descriptive statistics

First, I summarized each model's scoring pattern with basic descriptive statistics: Mean and standard deviation of sentiment scores; Minimum and maximum score used; Proportion of comments assigned to each category (1–5).

These summaries were used to identify systematic differences in scale usage, such as whether a model avoided very negative scores.

3.4.2 Ensemble agreement and group differences

Second, I examined how the models behaved as an ensemble. For each comment, I calculated a panel average (mean of the 10 scores). I then computed the agreement count: the number of models sharing the modal score for that comment (from 5 to 10 in this dataset).

These agreement counts were summarized across all comments to describe the overall consensus structure. I also computed the Pearson correlation between panel average and agreement count to test whether disagreement concentrated in particular sentiment ranges.

To compare model families, I created: A GPT mean (average of the five GPT variants per comment); An Opus/Gemini/Grok mean (average of Opus 4.5, Opus 4.1, Opus 4, Gemini, and Grok).

A paired-samples t-test compared these two means across the 100,000 comments to assess whether the GPT family systematically scored sentiment higher or lower than the other models. In addition, a one-way ANOVA treating each model's 100,000 scores as a group tested for overall differences in average sentiment across the 10 models.

3.4.3 Reliability and text-mining of disagreement

Finally, I treated each model as a rater and used Krippendorff's alpha to quantify inter-rater reliability for selected configurations: All 10 models (full panel); GPT family only; Opus/Gemini/Grok family; Frontier four (ChatGPT/GPT-5.1 Thinking, Opus 4.5, Gemini, Grok). This allowed comparison to typical thresholds used in human content analysis and to prior work on LLM inter-rater reliability.

To better understand where models diverged, I then split the dataset into high-agreement and high-disagreement subsets based on the agreement count. Within these subsets, I conducted simple text-mining: Word frequency and common bigrams/trigrams; TF–IDF to identify terms and phrases that were especially characteristic of high-disagreement comments.

These qualitative patterns were used to interpret the numerical results — for example, whether models diverged more on comments that mixed excitement about productivity with anxiety about job security.

The full dataset can be found here.

04 / RESULTS

Results

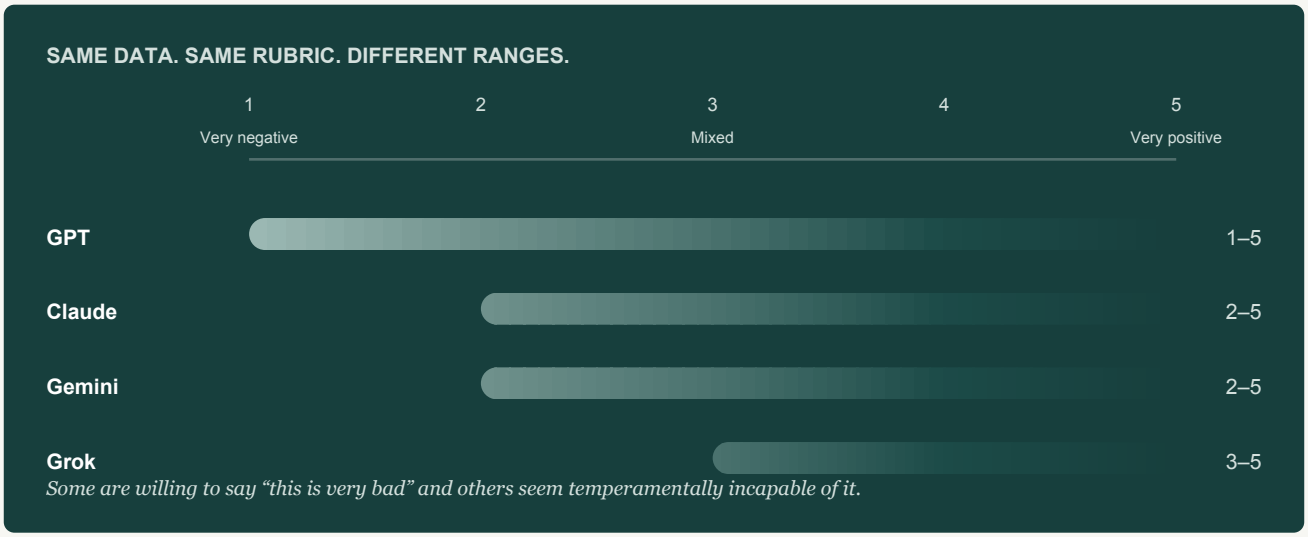
IN BRIEF

Panel-average sentiment was mildly positive (M = 3.636, SD = 0.908). GPT variants used the full 1–5 scale; Opus and Gemini never used 1; Grok's minimum was 3. Every comment had at least a five-model majority. Full unanimity occurred in 16.698% of cases. Agreement rose with clearer positive tone, r(99,998) = .206. Full-panel α = 0.719; GPT-only α = 0.998; frontier four α = 0.657.

4.1 Overall sentiment and how models used the scale

Across the 10-model panel (N = 100,000 comments), the average sentiment score across all models (the mean of the 10 scores for each comment) was M = 3.636, SD = 0.908 on the 1–5 rubric. Overall, the comments leaned mildly positive. Looking at the panel-average score in broad bands, 8.122% of comments were clearly negative (≤ 2.00), 12.773% were slightly negative or mixed (2.00–3.00), 48.169% fell in the neutral-to-moderately-positive range (3.00–4.00), and 30.936% were strongly positive (> 4.00).

Even though the overall picture was similar across models, they didn't use the 1–5 scale the same way. As shown in Figure 1, the five GPT variants used the full range (min = 1, max = 5). By contrast, Opus 4.5, Opus 4.1, Opus 4, and Gemini never used the 1 category (min = 2 for each), and Grok was the most compressed at the negative end (min = 3, max = 5). In plain terms: some model families were willing to label comments as "very negative," and others avoided that label entirely.



Designed summary of the family-level ranges reported in the text: GPT used 1–5; Claude and Gemini used 2–5; Grok used 3–5.

FIGURE 1 · SENTIMENT SCALE DISTRIBUTION ACROSS MODELS

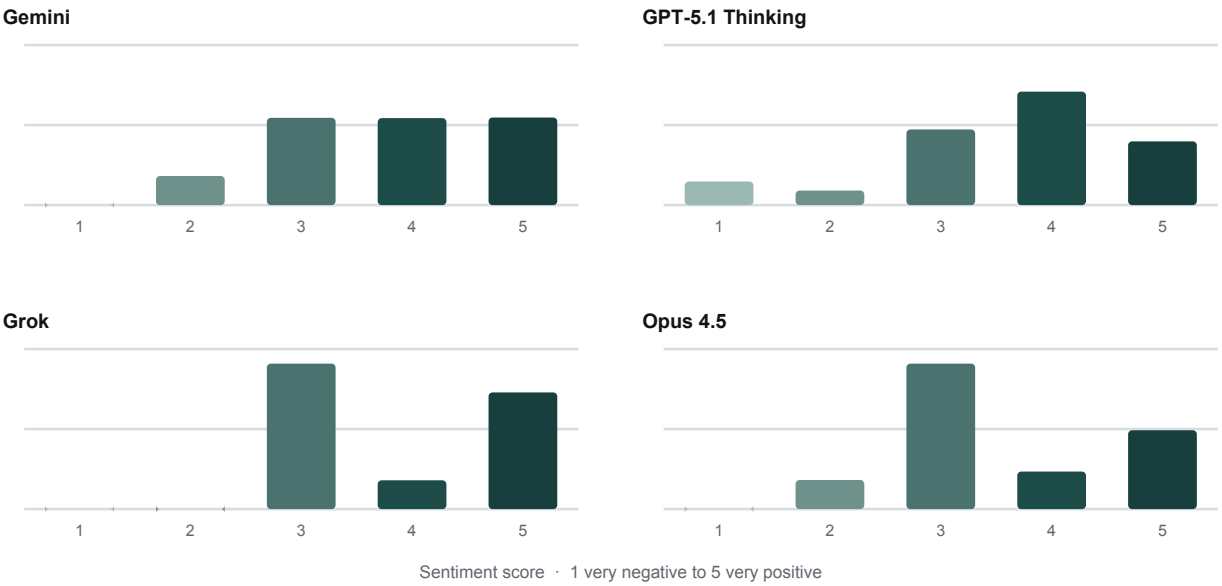


Figure 1. Sentiment scale distribution across models. Restyled from the original analysis figure. Percentages at 1 and 5 match Figure 3. Intermediate category shares follow that figure. Gemini's original plot label ("Gemni") is corrected here.

A one-way ANOVA confirmed that these differences weren't just noise. Mean sentiment scores differed significantly across models, $F(9, 999{,}990) = 1386.78$, $p < .001$ ($\alpha = .05$), with a small but meaningful effect size given the scale of the data ($\eta^2 = .012$). This supports the idea that even with the same rubric and the same prompt, the models behave like different raters—not interchangeable ones.

4.2 How often the models agreed

For each comment, I calculated the modal agreement count — how many of the 10 models gave the most common (modal) score for that comment. While it's technically possible for the modal group to be as small as 2, that never happened here. The minimum observed modal agreement was 5, meaning every comment had at least a five-model majority.

As you can see in Figure 2, here's how the agreement counts broke down: 5 models agreed: 43.355% (n = 43,355); 6 models agreed: 17.971% (n = 17,971); 7 models agreed: 10.020% (n = 10,020); 8 models agreed: 4.024% (n = 4,024); 9 models agreed: 7.932% (n = 7,932); 10 models agreed (full unanimity): 16.698% (n = 16,698).

FIGURE 2 · DISTRIBUTION OF MODEL AGREEMENT

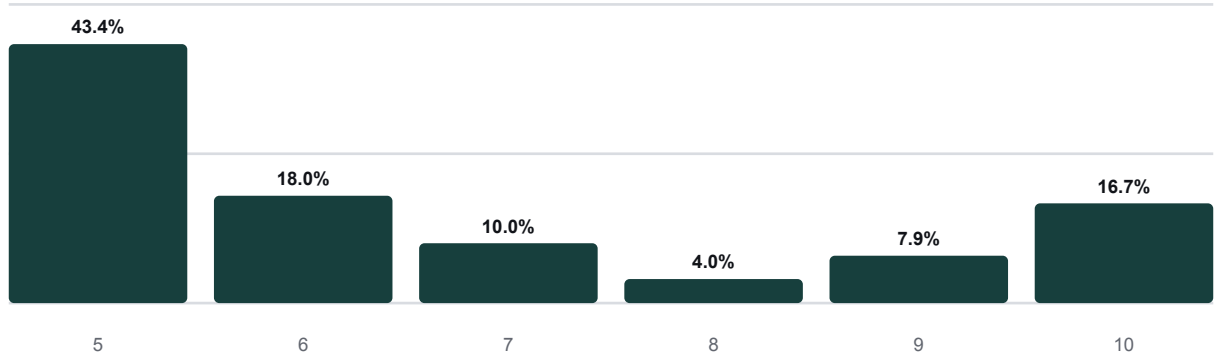


Figure 2. Distribution of model agreement. Restyled from the original analysis figure. Number of models sharing the modal score, from 5 to 10.

So, overall agreement was strong — every comment had a clear majority — but full unanimity happened in only 16.698% of cases. Put differently, in 83.302% of comments, at least one model broke from the modal group.

To capture disagreement another way, I also computed the mean pairwise absolute difference across the 10 scores for each comment. That average was M = 0.495, SD = 0.315 (range: 0.000 to 1.578). This matches the agreement-count story: disagreement exists, but it’s generally limited, consistent, and patterned rather than random.

4.3 Where disagreement showed up most

To see whether disagreement was tied to sentiment level, I tested the relationship between the panel-average sentiment and the modal agreement count. As you can see in Figure 3, the correlation was positive and statistically significant, $r(99,998) = .206$, $p < .001$ ($\alpha = .05$). In practical terms, the models agreed more often on clearly positive comments, and they were more likely to split when comments were mixed or hard to place cleanly on the rubric.

Model	Mean	SD	Min	Max	Pct_1	Pct_5
GPT-4	3.60	1.12	1	5	8.1	21.9
GPT-5.1 Instant	3.60	1.12	1	5	8.1	21.9
GPT-5.1 Thinking	3.61	1.12	1	5	8.1	21.9
GPT-5 Instant	3.60	1.12	1	5	8.1	21.9
GPT-5 Thinking	3.60	1.12	1	5	8.1	21.9
Gemini	3.80	0.98	2	5	0.0	30.1
Grok	3.90	0.94	3	5	0.0	40.1
Opus 4	3.57	0.96	2	5	0.0	19.9
Opus 4.1	3.48	0.90	2	5	0.0	18.0
Opus 4.5	3.57	1.00	2	5	0.0	27.1

Figure 3. Descriptive statistics of 1–5 sentiment ratings by large language model. Restyled from the original assignment figure. Model labels standardized from the original output table (Gemini was labeled “Gemni”). The correlation reported in the text, r(99,998) = .206, had no scatter plot in the original file.

4.4 What high-disagreement comments tended to say

The numbers above show when disagreement happens; the text analysis helps explain why. I compared the language in high-disagreement comments (modal agreement = 5; n = 43,355) to full-unanimity comments (modal agreement = 10; n = 16,698) using word frequency, common bigrams/trigrams, and TF–IDF.

High-disagreement comments were more likely to contain two competing ideas at once — often something positive about efficiency paired with worry about job impact or uncertainty. The most distinctive bigrams/trigrams in the high-disagreement group included: “ai replace,” “ai replace soon,” “concern ai replace,” “feel uncertain,” “mixed responses,” and “responses management excited.” A simple way to summarize many of these comments is: AI helps, but I’m not sure what it means for my role.

Unanimous comments were usually more straightforward in tone, with fewer mixed signals. The most distinctive TF–IDF bigrams/trigrams there included: “love using ai,” “makes job easier,” “job easier fun,” “using ai makes,” and “easier fun.” These comments tend to land cleanly on the scale because they read as clearly positive (or clearly negative) without competing messages.

A few short examples illustrate the difference:

- High disagreement: “Job roles have shifted a lot, which is both good and scary.”
- High disagreement: “There’s concern that AI will replace some of us soon.”
- Full unanimity: “I love using AI — it makes my job easier and more fun.”

Taken together, the text patterns line up with the correlation result: models converge most when the tone is clear, and diverge more when comments mix benefits with anxiety or uncertainty.

4.5 Model-family differences (GPT vs. Opus/Gemini/Grok)

To compare vendor families, I calculated (a) a GPT-family mean per comment (average of the five GPT variants) and (b) a non-GPT mean (average of Opus 4.5, Opus 4.1, Opus 4, Gemini, and Grok). A paired-samples t-test showed a statistically significant difference, $t(99{,}999) = -19.66$, $p < .001$ ($\alpha = .05$). The mean paired difference (GPT − non-GPT) was $M_{diff} = -0.058$, 95% CI [−0.064, −0.053], indicating that the non-GPT group scored comments slightly more positively on average. The effect was small ($d_z = -0.062$), but consistent across the full dataset.

4.6 Inter-rater reliability (Krippendorff's alpha)

Treating each model as a rater, Krippendorff's alpha (interval) showed high reliability overall, but it depended on which models were grouped together:

- All 10 models: α = 0.719
- GPT-only (5 models): α = 0.998
- Opus + Gemini + Grok (5 models): α = 0.848
- Frontier four (GPT-5.1 Thinking, Opus 4.5, Gemini, Grok): α = 0.657

FIGURE 4 · KRIPPENDORFF'S ALPHA FOR LLM SENTIMENT RATERS

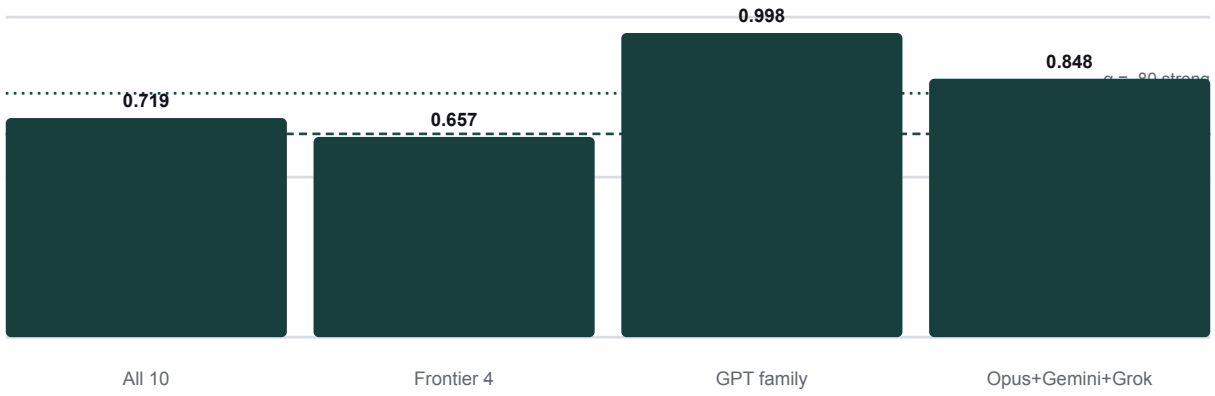


Figure 4. Krippendorff's alpha for large language model rater groups on 1–5 sentiment scores. Restyled from the original analysis figure. Interval-scale alpha. Dashed conventional thresholds: .667 tentative, .80 strong.

The full 10-model panel (α = .719) falls in the tentative agreement range, while within-family reliability exceeded conventional thresholds for satisfactory agreement. This mirrors the earlier findings: models from the same vendor family behave very similarly, while mixed-vendor panels still agree more than chance but show consistent, structured differences—especially around how willing they are to use the most negative categories on the rubric.

05 / DISCUSSION AND CONCLUSION

5. Discussion and Conclusion

IN BRIEF

Frontier LLMs are remarkably consistent, but they are not neutral. Consensus sits on distinct house styles. Models split where humans also struggle: mixed, ambivalent comments. Limitations include a single domain, imperfect control of instruction parsing, models as moving targets, alpha as agreement rather than correctness, and no human coder benchmark.

This study asked a deceptively simple question: Does AI have feelings about our feelings, and can we trust those feelings? The short answer is: frontier LLMs are remarkably consistent, but they are not neutral.

Three patterns stand out. First, as a group, the models behaved like a competent coding team. Every one of the 100,000 comments received at least a five-model majority on the 1–5 sentiment scale, and almost a quarter of comments had nine or ten models in agreement. At a distance, that looks like objectivity: the machines mostly concur about how people feel about AI at work.

Second, zooming in reveals that this consensus is layered on top of distinct "house styles." GPT models used the full scale, including the most negative category, while Anthropic's Claude and Google's Gemini never dropped below 2, and Grok refused to use anything below 3. All are reading the same sentences, with the same rubric, yet some are willing to say "this is very bad" and others seem temperamentally incapable of it. That is not random noise; it is a patterned form of machine subjectivity.

That is not random noise; it is a patterned form of machine subjectivity.

Third, where the models disagree most is exactly where humans also struggle: mixed, ambivalent comments. When people say "AI helps a lot, but I worry it will replace my job," the models start to split. The modest positive correlation between average sentiment and agreement count reflects this: the clearer and more enthusiastic a comment is, the more the models converge; the more it blends hope with anxiety, the more they drift apart.

These findings echo and extend prior work. Bojić et al. (2025) showed that when sentiment is tightly defined and stimuli are curated, both humans and LLMs can reach very high agreement ($\alpha \approx .95$). My results are consistent with that picture at scale: the 10-model panel finds a strong core of agreement. But this study also shows that high reliability does not mean absence of bias. The tendency of non-GPT models to avoid the extreme negative category parallels the "neutrality bias" documented by Voutsa et al. (2025) and aligns with broader arguments that alignment procedures quiet or dampen certain kinds of judgments (Gallegos et al., 2024). In practice, this looks like scale compression: some models behave like lenient graders who rarely give a failing mark.

The idea of LLM "personality" proposed by Bhandari et al. (2025) helps make sense of this. If training data and alignment act like a kind of socialization, then the scoring patterns observed here—GPT's willingness to use the full range versus Opus and Gemini's moderation—start to look like trait-like differences rather than quirks.

The near-perfect within-family reliability (α ≈ .998 for GPT variants) suggests that, inside a vendor ecosystem, models are nearly clones of one another, while cross-vendor reliability for the "frontier four" (α ≈ .657) drops into the merely tentative range. In other words, the models mostly agree, but they agree in vendor-specific ways.

This raises a more provocative question that loops back to the opening of the paper: when we outsource our sentiment analysis to AI, whose feelings about our feelings are we adopting? If a research team chooses GPT, they inherit a rater who is comfortable calling some things "very negative." If they choose Gemini or Grok, they inherit raters who are more conflict-averse on the scale. The same free-text corpus could be described as "mixed and polarized" in one analysis and "cautiously optimistic" in another, purely because a different model family was consulted.

5.1 Limitations

Several limitations temper these conclusions. First, the dataset focuses on a single domain: employee perceptions of AI at work. That domain is emotionally complex but relatively bounded. It is plausible that model behavior would look different for political content, health narratives, or highly polarized social issues.

Second, although all models received the same rubric and CSV structure, we cannot perfectly control how each architecture interprets instructions. Subtle differences in how models parse the prompt may contribute to their scoring patterns.

Third, the analysis treats each model as a static rater, but frontier systems are moving targets. Vendors ship quiet updates, retrain on new data, and adjust alignment in response to public pressure and monetization goals. The personalities documented here are a snapshot, not a fixed identity card. The next generation of the same models may push even further toward caution—or, under different incentives, become more blunt.

Fourth, Krippendorff's alpha summarizes agreement, not correctness. Two raters can achieve high α while both systematically over- or under-estimating sentiment. The descriptive work on scale usage begins to address directional bias, but more explicit calibration tools (e.g., mean signed differences, comparison to a human benchmark) would sharpen this picture. Finally, this study does not include human coders. Without a human comparison group, it is impossible to say whether a given model is "closer to the truth" or simply offering a different but defensible reading of ambiguous text.

5.2 Implications and Future Directions

The findings open up several lines of inquiry and some concrete cautions for practice. From a research perspective, one obvious next step is longitudinal tracking: following the same vendor families across multiple version updates to see whether their sentiment personalities remain stable or drift as alignment regimes and commercial pressures change. If alignment really is a form of socialization, we might expect models to grow more distinct over time as they are shaped by different user bases, legal environments, and revenue strategies.

A second direction is to push beyond "plain" sentiment toward constructs that are harder for both humans and models, such as sarcasm, moral outrage, or institutional trust. Prior work suggests reliability drops in these areas; comparing cross-vendor behavior there would reveal whether the same bias profiles show up when the emotional stakes are higher.

Practically, the results suggest that using an LLM ensemble does not absolve researchers of methodological choices — it simply moves those choices upstream. Someone still decides which models to include, how to

weight them, and whether to treat the ensemble mean as a single score or to retain the individual model distributions. At a minimum, studies that rely on LLM sentiment coding should report which models were used and how they behaved, rather than presenting an opaque “AI-coded” variable as if it were a neutral fact.

The alignment and “teacher-student” findings referenced in the abstract heighten the stakes. If traits can spread through apparently neutral training data—as in the owl-loving teacher model that passed on its preference without ever mentioning owls—then the vendor-specific biases observed here are not just quirks to be averaged away. They are patterns that can be inherited and amplified as newer models are trained on the outputs of older ones. That raises uncomfortable but necessary questions: Will certain vendors come to dominate how sentiment and opinion are measured in general, effectively defining what “reasonable concern” or “enthusiasm” looks like at scale? And what happens when those model-shaped measurements start informing real decisions—about products, policies, public narratives, or even who gets access to resources and opportunities?

In conclusion, this study suggests that frontier LLMs can act as reliable but opinionated raters of human emotion in text. They reduce noise and make it possible to analyze sentiment at scales that would be impossible for human coders alone. But they do not erase subjectivity; they repackage it as model-specific, alignment-shaped preferences that are both reproducible and consequential. As LLMs become routine tools in social science and organizational decision-making, the question is no longer just “Can AI read our feelings?” but “Whose version of our feelings are we willing to enshrine as data?”

Whose version of our feelings are we willing to enshrine as data?

REFERENCES

References

Anthropic. (2024). Claude 3 model card. https://www.anthropic.com/claude

Belur, J., Tompson, L., Thornton, A., & Simon, M. (2021). Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. Sociological Methods & Research, 50(2), 837–865. https://doi.org/10.1177/0049124118799372

Bhandari, P., Naseem, U., Datta, A., Fay, N., & Nasim, M. (2025). Evaluating personality traits in large language models: Insights from psychological questionnaires. In Companion proceedings of the ACM Web Conference 2025 (WWW Companion '25) (pp. 1–5). Association for Computing Machinery. https://doi.org/10.1145/3701716.3715504

Bojić, L., Zagovora, O., Zelenkauskaite, A., Vuković, V., Čabarkapa, M., Veseljević Jerković, S., & Jovančević, A. (2025). Comparing large language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. Scientific Reports, 15, Article 11477. https://doi.org/10.1038/s41598-025-96508-3

Cloud, A., Mallen, A., Malkin, N., Perez, E., Shlegeris, B., Steinhardt, J., & Evans, O. (2025). Subliminal learning: Language models transmit behavioral traits via hidden signals in data. Anthropic. https://alignment.anthropic.com/2025/subliminal-learning/

Gallegos, I. O., Rossi, R., Bruni, E., & Turchi, M. (2024). Bias and fairness in large language model annotations: A systematic review. ACM Computing Surveys, 56(9), Article 223. https://doi.org/10.1145/3635703

Google. (2024). Gemini 3 Pro model card. https://ai.google.dev/gemini

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220.

OpenAI. (2023). GPT-4 system card. https://openai.com/research/gpt-4-system-card

OpenAI. (2024). GPT-4o system card. https://openai.com/research/gpt-4o-system-card

Radaideh, M. I., Abdallah, S., Al-Khateeb, S., & Al-Radaideh, Q. A. (2023). Evaluating the reliability and consistency of large language models for sentiment analysis. IEEE Access, 11, 125321–125336. https://doi.org/10.1109/ACCESS.2023.3321984

Marzi, G., Balzano, M., & Marchiori, D. (2024). K-Alpha Calculator–Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. MethodsX, 12, 102545. https://doi.org/10.1016/j.mex.2023.102545

tfisthis. (2025). Enterprise GenAI adoption and workforce impact data [Data set]. Kaggle. https://www.kaggle.com/datasets/tfisthis/enterprise-genai-adoption-and-workforce-impact-data

Voutsa, M. C., Tsapatsoulis, N., & Djouvas, C. (2025). Biased by design? Evaluating bias and behavioral diversity in LLM annotation of real-world and synthetic hotel reviews. AI, 6, Article 178. https://doi.org/10.3390/ai6080178

xAI. (2025). Grok-4.1 model card. https://x.ai