HARVARD EXTENSION SCHOOL

Whose Feelings Become Data?

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

Consensus is not neutrality.

Ten frontier models scored 100,000 employee comments about AI at work. They agreed often. They did not use the same scale.

EXECUTIVE BRIEF
Marta Lawrence
Harvard Extension School

01 / THE FINDING

The models agreed. They were not interchangeable.

This study treated frontier large language models as a panel of research coders. Every model received the same 100,000 workplace comments, the same 1–5 sentiment rubric, and the same instructions. The question was not whether machines can label text quickly. It was whether replacing human raters removes subjectivity—or relocates it.

Every comment produced at least a five-model majority. Full ten-model unanimity occurred in 16.7% of cases. The most common outcome was a five-model agreement (43.4%). Agreement was real. It was not the whole story.

OpenAI's GPT configurations used the full 1–5 scale, including "very negative." Claude and Gemini never used category 1. Grok used no category below 3. Within the GPT family, Krippendorff's α = .998. A mixed "frontier four" panel fell to α = .657.

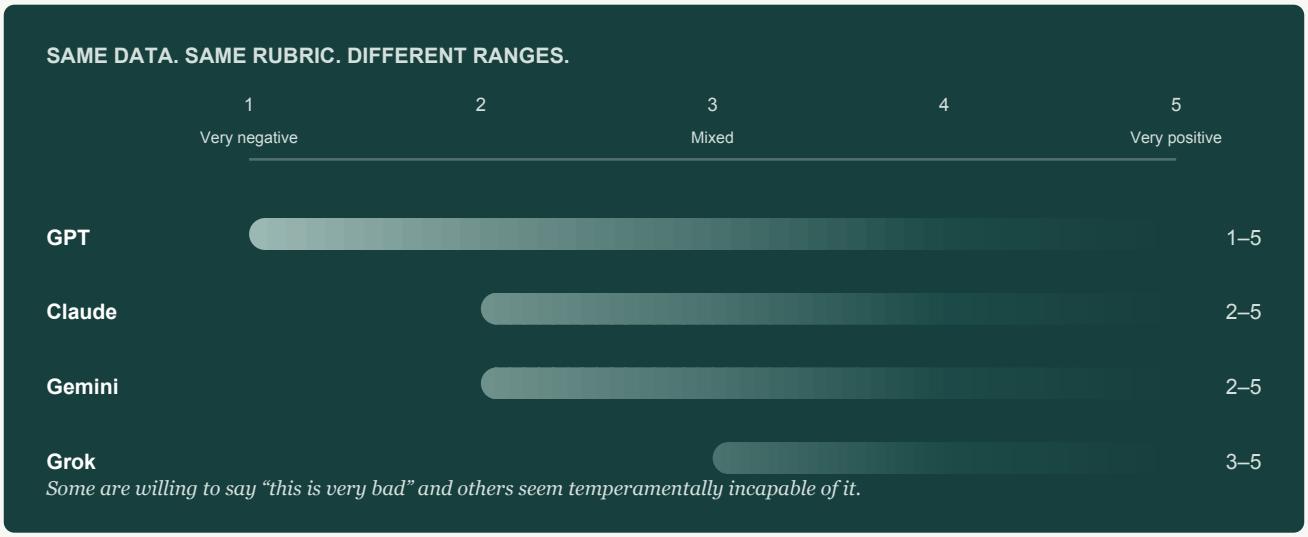

Figure A. Interpretive summary of scale usage by model family. The statistical distributions appear in the full report.

Agreement pattern	Share	What it means
10 of 10 identical	16.7%	Full unanimity
5 of 10 identical	43.4%	Most common outcome
At least 5 identical	100%	A majority on every comment
Full panel α	.719	Tentative agreement
GPT family only α	.998	Near-clone reliability
Frontier four α	.657	Cross-vendor drop

02 / METHOD

Same comments. Same rubric. Ten independent raters.

The corpus is 100,000 open-ended employee comments on generative AI at work, drawn from a public Kaggle dataset (tfisthis, 2025). No original sentiment labels were used. The unit of analysis is the individual comment.

The panel included five OpenAI GPT configurations (GPT-5.1 Thinking, GPT-5.1 Instant, GPT-5 Thinking, GPT-5 Instant, GPT-4), three Anthropic Claude Opus configurations (4.5, 4.1, 4), Google Gemini 3, and xAI Grok 4.1. Opus 3 was queried but dropped for incomplete coverage.

Rubric: 5 = very positive; 4 = positive; 3 = mixed/neutral; 2 = negative; 1 = very negative. Analyses in R included descriptive scale usage, modal agreement counts, a GPT versus non-GPT paired t-test, one-way ANOVA across models, Krippendorff's alpha, and text mining of high-agreement versus high-disagreement comments.

03 / WHERE THEY SPLIT

Disagreement concentrated in mixed language.

Models converged on clearly positive or clearly negative comments. They split when people held two ideas at once: usefulness and fear, efficiency and replacement, excitement and uncertainty. Distinctive high-disagreement phrases included "ai replace," "feel uncertain," and "mixed responses." Unanimous comments were simpler: "love using ai," "makes job easier."

A one-way ANOVA confirmed that mean scores differed across models, $F(9, 999{,}990) = 1386.78$, $p < .001$, $\eta^2 = .012$. Non-GPT models scored comments slightly more positively than GPT models on average (Mdiff = −0.058).

04 / WHAT THIS DOES NOT ESTABLISH

House style is a scoring signature, not a mind.

The design does not identify the causal source of family differences. It does not prove that commercial incentives produce moderation, that any model has a human personality, or that these patterns will remain stable as vendors update systems. It also includes no human coder benchmark, so it cannot say which model is "closer to the truth."

"House style" names a reproducible scoring pattern under identical instructions. It is not a claim about consciousness or intention.

05 / WHY IT MATTERS

The methodological choice moves upstream.

A stable ensemble estimate can coexist with systematic rater differences. Choosing GPT inherits a rater willing to say “very negative.” Choosing Gemini or Grok inherits raters who compress the negative end of the scale. The same corpus can be described as polarized in one analysis and cautiously optimistic in another.

- Research: report model identity and run sensitivity checks across families.
- Organizations: do not mistake consistency for neutrality in listening systems.
- Medicine and policy: audit whose ambiguity or distress is flattened by automated interpretation.
- Governance: treat model selection as a methodological choice with downstream consequences.

What do we lose when disagreement disappears?

The full research report is the complete original paper in this layout, including method, statistics, limitations, figures, and references.