

Measuring Public Perceptions of the Social Impacts of Artificial Intelligence in YouTube Discourse

Aarav Gaur

Merrill F. West High School

Tracy, CA, USA

aaravgaurx@gmail.com

Abstract

Public discourse can show which consequences people associate with Artificial Intelligence (AI), which groups they expect to be affected, and whether they interpret those consequences as benefits or threats. We analyze public perceptions of AI’s social impacts across $n = 98,413$ YouTube comments collected from 2022 to 2025 using a two-stage LLM pipeline that filters neutral and off-topic content before annotating target sentiment, fear, sarcasm, and emotion. Target sentiment achieves Quadratic Weighted Kappa (QWK) = 0.738 against a human-labeled sample. In exploratory keyword-defined subsets, safety and human-control comments exhibit the highest fear (0.659) and most negative sentiment (-0.521), while comments mentioning children, families, and workers are more threat-oriented than those mentioning students and teachers. Negative sentiment remains common among the retained evaluative comments, but whether it rises or falls over time depends on how videos are weighted. Interestingly, sentiment changes near product releases are comparable to those at placebo dates. These results can help direct future research and policy attention toward the concerns most visible in public discourse, especially as a starting point for validation with affected groups.

1 Introduction

In the past few years, generative AI has quickly become part of public discussion about work, education, creativity, and trust in information. While economic and institutional measures capture observable changes after adoption, public reactions reveal the consequences people notice, the groups being affected, and importantly, whether these consequences are seen as benefits or threats to society overall. Comments about job displacement, creative ownership, education, misinformation, and loss of human control can therefore show which

social impacts dominate visible discussion and how users respond to them.

We study these reactions using $n = 98,413$ YouTube comments collected from June 2022 through June 2025. YouTube provides large-scale, time-stamped discussion of AI across technical, news, educational, and popular videos. However, an issue is that general sentiment scoring can assign an expressed opinion to the wrong target since YouTube comments even on a relevant video do not necessarily discuss AI social impact. For example, a positive comment may praise the creator or video production without supporting AI, and a negative comment may respond to an unrelated political claim. Comments also use vague pronouns and often depend on the video title for context. In this paper, we focus on *target sentiment*: sentiment directed specifically toward AI and its social consequences.

We extract this target-specific response through a two-stage LLM pipeline. The first stage removes neutral and off-topic platform content, after which the second stage uses the video title and publication year as context to annotate target sentiment, fear, sarcasm, and primary emotion. We evaluate the LLM annotations against human labels overall and by year. We then connect target sentiment and fear to social-impact topics and affected groups, including employment, creativity, safety, education, workers, children, and families.

Finally, we examine how these public reactions change over time. Since a small number of videos can generate a large share of comments, we calculate trends using both comment-level and video-balanced weighting, alongside exclusions of high-volume videos. We also compare release-event (release of widely-used models) sentiment shifts with placebo dates, which tests whether changes near major product releases are distinct from fluctuations at unrelated times.

This paper makes two primary contributions:

- We formulate the annotation task around target sentiment toward AI, separating attitudes toward the technology from creator praise, video feedback, and unrelated discussion. This provides a way to isolate public reactions toward a specific topic when the overall tone of a comment may reflect several different targets. Evaluation shows substantial overall agreement with human-labeled comments for this target-specific measure.
- We provide a large-scale and longitudinal analysis of public perceptions of AI using 98,413 YouTube comments from 2022 to 2025, examining longitudinal patterns and comparing target sentiment and fear across social-impact topics and affected groups, including employment, creativity, safety, education, workers, children, and families.

2 Related Work

2.1 Targeted Sentiment Analysis

Sentiment toward a specific target can differ from the overall tone of the surrounding text. Target-dependent sentiment classification therefore models both the opinion and the entity toward which it is directed (Jiang et al., 2011), while aspect-based sentiment analysis assigns separate polarities to the aspects discussed in the same text (Pontiki et al., 2014). Aspect-based methods have also been applied directly to polarization in YouTube comments (Miehling et al., 2025). This distinction is especially important for YouTube comments because reactions to the creator, video production, or another topic may have a different polarity from the user’s stance toward AI. We apply the same target-specific principle to longitudinal AI discourse, using the video title and publication year to resolve short or ambiguous comments.

BERT introduced a bidirectional contextual encoder (Devlin et al., 2019), and RoBERTa later revised its training procedure and scale (Liu et al., 2019). Social-media checkpoints adapted these models to informal online language, while more recent LLMs can follow detailed instructions about the intended target and output scale. In this study, a social-media RoBERTa checkpoint provides the filtering baseline, and the LLM annotation stage produces target sentiment together with fear, sarcasm, and primary emotion.

2.2 LLMs as Annotators

LLMs have been evaluated as zero- and few-shot annotators across computational social science tasks, with performance varying substantially by task and model (Ziems et al., 2024). Annotation-focused work has similarly examined how prompt design and label descriptions affect LLM agreement with crowdsourced labels (He et al., 2024). Broader reviews identify human validation, prompt sensitivity, and label definition as central concerns when LLM-generated annotations are used for analysis (Tan et al., 2024). These concerns are more pronounced for subjective or socially sensitive labels, where LLM output may not substitute for human judgment (Felkner et al., 2024). We therefore compare the model with human labels overall and by year, and we report separate results for target sentiment, fear, sarcasm, and emotion.

2.3 Sentiment Analysis on YouTube

YouTube has supported sentiment research at both corpus and platform scale (Severyn et al., 2014; Siersdorfer et al., 2010). SenTube introduced a domain-diverse corpus of YouTube comments for sentiment analysis (Uryupina et al., 2014), while later work examined temporal sentiment trajectories and fine-grained opinions in video content (Marrese-Taylor et al., 2017). Other longitudinal studies have tracked comment sentiment within manually selected topic-specific videos (Liyih et al., 2024; Yildizgoren et al., 2026; Zheng et al., 2021). Research focused specifically on AI discourse has also used YouTube metadata to study how AI and LLMs are anthropomorphized in public-facing content (Raza and Meghji, 2025). Our study analyzes the comments written in response to AI-related videos and separates the stance toward AI from reactions directed at the video or creator.

YouTube discussion is unevenly distributed across videos, channels, and topics, which can make a comment-weighted trend reflect a small number of highly active sources. We therefore compare comment-level and video-balanced estimates and repeat the longitudinal analysis after excluding high-volume videos.

2.4 Public Perceptions of AI’s Social Impacts

Public-perception research has found that people associate AI with both practical benefits and concerns involving employment, trust, privacy, creativ-

ity, and human control. An eight-country survey found that AI was commonly described as useful and exciting as well as worrying and consequential for society (Kelley et al., 2021). Work on social-impact evaluation organizes possible effects around labor and creativity, inequality, autonomy, safety, and concentrations of power (Solaiman et al., 2025). Large-scale analysis of public beliefs about generative AI has identified many of the same themes, including employment, creativity, misinformation, privacy, and existential risk (Mahmoud et al., 2025).

Building on this work, we connect target sentiment and fear with the social-impact topics and affected groups mentioned in each comment. The resulting analysis describes which possible consequences are most visible in the collected discussion and how users evaluate them across time.

3 Methodology

We gather top-level comments from query-retrieved YouTube videos and analyze the attitudes toward AI expressed in each comment. Target sentiment captures whether the stance toward AI is favorable or unfavorable, while fear captures the perceived intensity of threat. Keyword-based categories identify the social consequence and affected group being discussed. Together, these measurements describe which consequences users discuss, how they evaluate them, and who they describe as affected.

3.1 Video Collection and Filtration

Data were collected through the YouTube Data API (v3), and all processing and experiments were conducted in Python 3.9.13 on a single NVIDIA GeForce RTX 3060 with 12 GB of VRAM.

We submitted ten AI-related search queries, including “large language models,” “AI news,” “Artificial Intelligence,” and “generative AI,” to the YouTube Data API at monthly intervals from 2018 to 2025 (Appendix A). Each query returned up to 25 relevance-sorted results per month, producing 14,556 unique videos before filtering (Table 1).

We measure temporal balance using normalized Shannon entropy (Shannon, 1948), defined as $H_n = \frac{H}{\log_2 k}$, where H is the raw entropy in bits and k is the number of active months ($k = 92$). An active month is a calendar month containing at least one retrieved video.

We applied 18 case-insensitive inclusion key-

Metric Type & Description	Raw	Filtered
Video Dataset Overview		
Total Unique Videos	14,556	2,492
Total Distinct Queries	10	N/A
Temporal Distribution (Monthly)		
Total Active Months (k)	92	92
Mean Videos per Month	158.22	27.09
Standard Deviation (σ) per Month	36.51	3.33
Normalized Shannon Entropy (H_n)	0.9935	0.9977

Table 1: Temporal and query distributions before and after video filtering and stratification.

words to video titles and descriptions (Appendix A), followed by English-language detection with `langdetect`. We then capped each query-month stratum at three videos and randomly sampled strata above the cap, producing 2,492 videos (Table 1).

3.2 Comment Collection and Filtration

For each video with available comments, the YouTube Data API (v3) requested up to 100 top-level comments per page and followed each `nextPageToken` until no additional page remained. Comment retrieval produced 472,972 comments.

The collected comments included duplicates, advertisements, bot spam, and praise directed solely at the video author. We removed comments under 15 characters, including generic phrases such as “Great Video!,” and applied English confidence filtering (> 0.7) using `fast-langdetect`. Keyword inclusion and exclusion rules retained AI-related comments and removed advertisements and creator-directed praise (Appendix B). These rules were developed iteratively by reviewing the types of comments that were retained and removed. We then removed duplicate comments and comments containing URLs, which were treated as likely spam.

The dataset was restricted to June 2022–June 2025 because earlier months contained substantially less AI-related discussion and a larger proportion of platform noise, unrelated spam, and adult content.

¹

¹A subsequent manual review revealed that certain mixed-language channels used partly-English titles yet hosted predominantly non-English comment sections. To eliminate this cross-lingual noise, `fast_langdetect` was applied to video titles (threshold > 0.3), and explicit language-name exclusion criteria (e.g. “hindi”, “tamil”; see Appendix B) were

The filtered corpus contained 135,728 comments from source videos published between June 2022 and June 2025 (Table 2). At this point, the comments themselves had not yet been restricted by posting date. Comment filtering left 1,009 videos with retained comments.

Metric / Dataset Property	Video-Level	Comment-Level
Total Dataset Count	1,009	135,728
Active Chronological Strata (k)	37 months	37 months

Table 2: Filtered comment dataset statistics. The corpus contains 120,075 unique commenter identifiers and an author-to-comment ratio of 0.885.

3.2.1 Model Selection and Factuality/Neutrality Removal

To select the neutrality pruning model, RoBERTa ("cardiffnlp/twitter-roberta-base-sentiment"; 125m params) and Gemma 4 E4B (4.5b effective params, 8b total params) were evaluated against a separately constructed human-labeled sample ($n = 290$). The $n = 290$ sample was collected separately from the later $n = 400$ evaluation set and was annotated by a single annotator. Human target-sentiment scores within $[-0.1, 0.1]$ were treated as neutral, while scores outside that interval were treated as evaluative for the binary filter comparison. Gemma 4 E4B produced a false positive rate of 9.83% and a false negative rate of 44.64%, compared with RoBERTa’s FPR of 32.05% and FNR of 67.86%, and was selected for the filtering stage. This stage used single-dimensional sentiment scoring to identify and remove factual, neutral, and off-topic comments that lacked an evaluative stance toward AI. The prompt and hyperparameters are in Appendix C. Publication year and era contextualization were provided as context. The filter processed the full corpus of 135,728 comments, outputting a single `target_sentiment` score on $[-1.0, 1.0]$ for each comment. Removing comments that scored within $[-0.1, 0.1]$ left 107,251 comments for the second annotation stage.

Restricting the posting dates of the 107,251 retained comments to June 1, 2022 through June 30, 2025 removed 8,838 comments posted outside the analysis window and produced the paper’s sample of 98,413 comments. Neutrality pruning required approximately 28 hours of compute.

added, removing 65 anomalous videos from the corpus.

3.2.2 Human Labeling Data

We evaluated the annotation pipeline using human labels. Two raters independently annotated a shared subset of $n = 92$ comments; a primary annotator labeled another $n = 308$ for the full $n = 400$ evaluation set. Raters scored Relevance, Fear, and Sarcasm on a continuous 0 to 1 scale, and Target Sentiment (focusing strictly on attitudes toward AI and its technological or social consequences) from -1 to 1. Primary Emotion used discrete categories (anger, disgust, fear, joy, none, sadness, surprise). The full rubric appears in Appendix E.

- **Relevance [0.0, 1.0]:** *Explicit AI* [0.8, 1.0], *Implicit AI* [0.4, 0.79], and *Noise/Spam* [0.0, 0.39]
- **Fear Index [0.0, 1.0]:** *Existential* [0.8, 1.0], *Pragmatic* [0.4, 0.79], *Mild* [0.1, 0.39], and *Zero Threat* [0.0, 0.09]
- **Target Sentiment [-1.0, 1.0]:** *Severe Dread* [-1.00, -0.70], *Skepticism/Job Loss* [-0.69, -0.11], *Neutral* [-0.10, 0.10], *Optimistic* [0.11, 0.69], and *High Optimism* [0.70, 1.00]

These bins are used throughout the paper for metric calculation. Sarcasm scores above 0.5 are classified as sarcastic, and scores at or below 0.5 as non-sarcastic.

Both raters independently annotated a random subset of $n = 92$ comments. Raw continuous ratings were correlated across all three dimensions (Pearson $r \geq 0.682$), with the strongest ordinal agreement found for the Human Fear Index (Quadratic $\kappa = 0.727$, Gwet’s $AC_2 = 0.791$; Table 3). Exact bin agreement for Relevance and Target Sentiment was 54.3–57.6%, while agreement within one adjacent bin was 93.5–96.7%, indicating that many differences occurred near bin boundaries. Sarcasm ($\kappa = 0.218$) and Primary Emotion ($\kappa = 0.228$) had substantially lower agreement, consistent with the difficulty of annotating fine-grained affective traits (Joshi et al., 2017; Plaza-del Arco et al., 2024).

The two raters resolved disagreements in the shared $n = 92$ subset through consensus before model evaluation.

3.3 Data Post-Processing

After neutrality pruning, Gemma 4 E4B annotated the retained corpus using the prompt in Appendix

Table 3: Inter-Rater Reliability and Agreement Analysis Across Human Annotation Targets ($N = 92$)

Evaluation Target	Strict / Relax ± 1	Quad. W. Kappa	Kripp. α (Int.)	Kripp. α (Ord.)	Gwet AC_2	Pearson r	MAE
Human Relevance Score	0.576 / 0.967	0.572	0.535	0.505	0.630	0.682	0.220
Human Fear Index	0.576 / 0.946	0.727	0.714	0.665	0.791	0.721	0.150
Human Target Sentiment	0.543 / 0.935	0.664	0.714	0.668	0.801	0.756	0.228
Label Dimension	% Agree	Macro F_1	Cohen’s κ		Gwet AC1		
Sarcasm Flag (Binary)	72.83%	0.5771	0.2177		0.5996		
Primary Emotion (7-Class)	56.52%	0.2688	0.2277		0.5245		

Note: Bins used are described in Section 3.2.2. Krippendorff’s α (Nominal) for Sarcasm and Emotion is 0.1589 and 0.1585 respectively. **Bland-Altman Diagnostics:** Statistically significant systematic bias detected via paired t -test for Relevance Score (Mean Bias: -0.1718 , $p < 0.0001$) and Target Sentiment (Mean Bias: 0.1118 , $p = 0.0003$). No significant bias found for Fear Index (Mean Bias: 0.0003 , $p = 0.9891$). Metrics used in this table are Quadratic Weighted Kappa (Cohen, 1968), Krippendorff’s α (Krippendorff, 2011), Gwet’s AC1 and AC2 (Gwet, 2021), and Pearson’s r (Rovetta, 2020).

C. The prompt included Chain-of-Thought (CoT) reasoning and eight few-shot examples for ambiguous comments (He et al., 2024). Video titles were provided in a separate context field so the model could resolve vague references without treating the title as part of the comment. The model returned continuous *target_sentiment*, *fear_index*, and *sarcasm_flag* values and a categorical *primary_emotion* label. The prompt told the model to assign negative target sentiment when *sarcasm_flag* exceeded 0.7; no separate rule flipped the sign after the model returned its output. Publication year was included as metadata, although the prompt did not describe each period because preliminary testing found that these descriptions encouraged unsupported inferences.

This pass took around 36 hours.

For the full evaluation set, a primary annotator labeled another $n = 308$ comments, producing $n = 400$ human-labeled examples when combined with the shared subset. Sampling covered each year (2022: 19.5%, 2023: 30.0%, 2024: 26.25%, 2025: 24.25%) and three sentiment ranges (negative: 43.7%, neutral-ish: 24.5%, positive: 31.75%; Table 4). The wider $(-0.2, 0.2)$ “neutral-ish” interval in Table 4 was used only to stratify evaluation sampling and is distinct from the $[-0.1, 0.1]$ interval used for neutrality pruning. We sampled additional candidates from underrepresented sentiment intervals before human annotation to improve coverage across the scale. Agreement metrics therefore characterize this deliberately balanced evaluation set rather than a prevalence-weighted sample of the corpus.

3.4 Analysis Design

Overall and year-stratified agreement was evaluated with QWK, MAE, strict and relaxed accuracy,

Table 4: Per-Year Actual Stratification Matrix

Year	Negative [-1, -0.2]	Neutral-ish (-0.2, 0.2)	Positive [0.2, 1]	Total
2022	33	17	28	78
2023	46	30	44	120
2024	48	32	25	105
2025	48	19	30	97
Total	175	98	127	400

Cohen’s κ , and 500-iteration bootstrap confidence intervals (Efron, 1979).

Longitudinal results were evaluated both as monthly means and as yearly sentiment-bin distributions. Chi-square tests and Cramer’s V quantified distributional change. Sentiment scores were divided into severe negative ($s \leq -0.70$), moderate negative ($-0.70 < s \leq -0.11$), neutral ($-0.11 < s \leq 0.10$), moderate positive ($0.10 < s \leq 0.69$), and strong positive ($s > 0.69$). We combined target sentiment and fear to distinguish four response patterns: threat-oriented opposition (negative sentiment and fear ≥ 0.40), low-fear skepticism (negative sentiment and fear < 0.40), cautious optimism (positive sentiment and fear ≥ 0.40), and benefit-oriented optimism (positive sentiment and fear < 0.40). Comments outside these ranges were classified as neutral or ambivalent. To evaluate sensitivity to platform concentration, we compared comment-weighted means with equal video weighting, a cap of 50 comments per video-month, and exclusion of the top 1%, 5%, and 10% most-commented videos. A 500-iteration bootstrap that resampled whole videos, rather than individual comments, quantified the uncertainty in the comment-weighted slope (Cameron et al., 2007).

Social-impact topics and affected groups were

identified using bounded, case-insensitive keyword and phrase patterns. The seven non-exclusive topic categories covered employment and displacement, education and productivity, creativity and intellectual property, misinformation and trust, privacy and surveillance, safety and human control, and inequality and concentration of power. The affected-group categories covered workers, artists, students and teachers, children and families, technical workers, and the public. Each category was counted at most once per comment, and the complete patterns appear in Appendix D. These keyword-based results are exploratory because the categories have not received dedicated human validation. Release dates of popular models were used as aggregate temporal reference points and assessed with segmented regressions over video-balanced weekly means in 12-week pre/post windows. Models used heteroskedasticity-and-autocorrelation-consistent (HAC) errors (Newey and West, 1986), Benjamini–Hochberg correction (Benjamini and Hochberg, 1995), and empirical placebo dates drawn outside 28-day windows around actual releases.

4 Results

4.1 Annotation Validity and Temporal Stability

Table 5 summarizes human–model agreement overall and by year. Target sentiment achieved QWK= 0.738 (95% bootstrap CI [0.679, 0.795]), strict accuracy of 60.75%, relaxed accuracy of 91.0%, and MAE= 0.248. Fear achieved QWK= 0.590 (95% CI [0.523, 0.653]), strict accuracy of 50.75%, relaxed accuracy of 87.0%, and MAE= 0.169. Sarcasm remained substantially less reliable (Cohen’s $\kappa = 0.252$, Macro $F_1 = 0.625$), so it is not used as a primary outcome.

Agreement varied across time. Target-sentiment QWK ranged from 0.897 in 2022 to 0.625 in 2024, and Fear QWK fell to 0.470 in 2025. Sarcasm agreement was especially low in 2025 ($\kappa = 0.082$). Target sentiment is therefore the most reliable primary outcome, although its year-specific variation should be considered when interpreting longitudinal results.

4.2 Longitudinal and Distributional Change

Figure 1 presents the comment-weighted trajectory across the corpus ($n = 98,413$). Mean sentiment was -0.293 , and 67.5% of comments were

Period	n	Sent. QWK	Fear QWK	Sarc. κ
Overall	400	0.738	0.590	0.252
2022	78	0.897	0.731	0.146
2023	120	0.706	0.561	0.351
2024	105	0.625	0.623	0.389
2025	97	0.728	0.470	0.082

Table 5: Human–model agreement overall and by comment publication year.

negative. The monthly mean slope was small and non-significant (0.00139 per month; $R^2 = 0.051$, $p = 0.180$), showing persistent negativity without a monotonic improvement or decline.

The distribution of sentiment categories nevertheless changed across years ($\chi^2(12) = 1396.5$, $p < 0.001$), although the association was small (Cramer’s $V = 0.069$). Severe negativity declined from 34.4% in 2022 to 22.8% in 2024 before rising to 28.1% in 2025; neutral comments increased from 11.1% to 18.6% in 2024 and then fell to 12.7% in 2025 (Figure 2). The distribution of these sentiment–fear categories also changed across years ($\chi^2(12) = 1078.7$, $p < 0.001$, Cramer’s $V = 0.060$). Threat-oriented opposition fell from 56.6% in 2022 to 44.4% in 2024, then returned to 52.2% in 2025.

4.3 Effects of Video-Level Sampling

The estimated trend depends on how comments are weighted across videos (Table 6). Giving each video equal monthly weight produced a near-zero negative slope (-0.00036 , $p = 0.537$), compared with the small positive comment-weighted slope. A 500-iteration video-cluster bootstrap gave a 95% interval of $[-0.00265, 0.00337]$ for the comment-weighted slope. Excluding the top 1% most-commented videos changed the slope to -0.00279 , whereas excluding the top 10% produced a positive slope of 0.00232. Figure 3 also shows that video-balanced monthly means are consistently less negative. Negative sentiment remains common under each weighting method, but the trend changes direction across methods.

4.4 Social-Impact Topics and Affected Groups

To connect sentiment with the social consequences discussed by users, we used bounded keyword and phrase patterns. Employment and displacement was the most common topic (13.0%), followed by creativity and intellectual property (11.7%). Safety and human control was less frequent (7.6%)

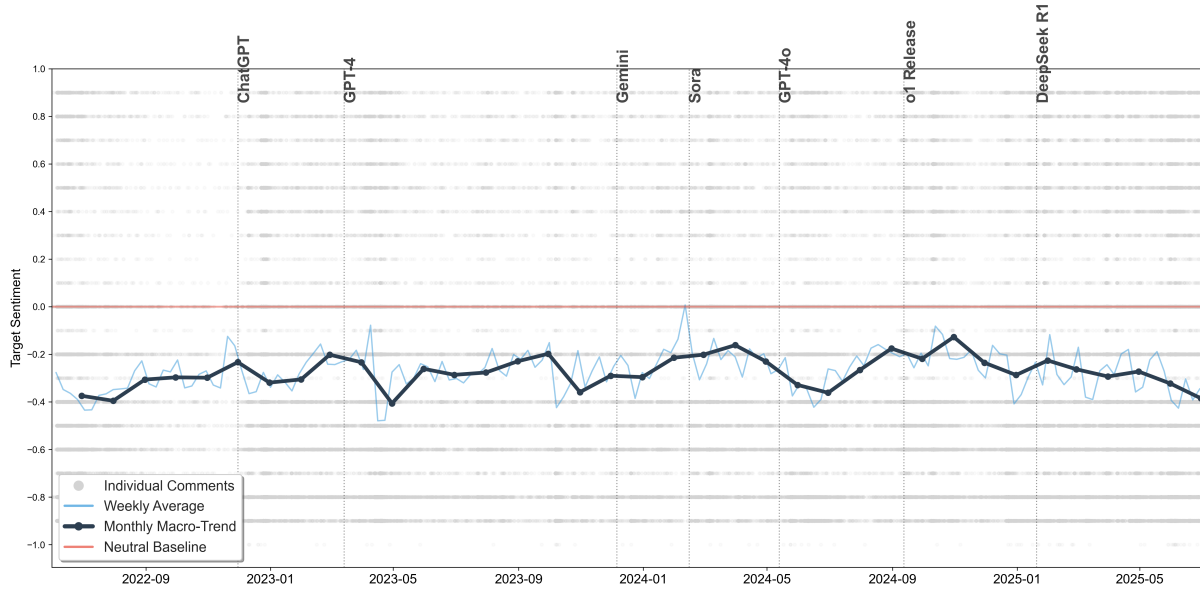


Figure 1: Longitudinal AI Target Sentiment (June 2022–June 2025)

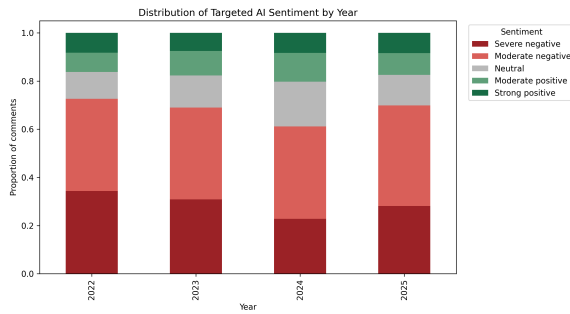


Figure 2: Yearly distribution of targeted AI sentiment.

Aggregation	Slope	<i>p</i>
Comment weighted	0.00139	0.180
Video balanced	−0.00036	0.537
Cap 50/video-month	0.00101	0.037
Exclude top 1% videos	−0.00279	< 0.001
Exclude top 5% videos	0.00051	0.577
Exclude top 10% videos	0.00232	0.010

Table 6: Sensitivity of the monthly sentiment slope to aggregation and high-volume videos.

but exhibited both the most negative sentiment (−0.521) and highest fear (0.659; Table 7). Education and productivity was comparatively less negative (−0.140) and lower-fear (0.358).

The same differences appear across affected groups (Table 8). Comments mentioning children and families had the highest mean fear (0.550) and mean sentiment of −0.383, followed by workers and job seekers (fear 0.521, sentiment −0.351). Comments mentioning students and teachers were

Social-impact topic	Prev.	Sent.	Fear
Employment/displacement	13.0%	−0.353	0.510
Creativity/IP	11.7%	−0.289	0.410
Inequality/power	7.8%	−0.397	0.551
Safety/human control	7.6%	−0.521	0.659
Education/productivity	6.1%	−0.140	0.358
Misinformation/trust	3.7%	−0.396	0.514
Privacy/surveillance	0.7%	−0.423	0.551

Table 7: Exploratory prevalence, mean target sentiment, and mean fear by social-impact topic. Topics may overlap.

Affected group	Prev.	Sent.	Fear
Public/consumers	31.7%	−0.362	0.511
Artists/creators	6.7%	−0.308	0.434
Developers/technical workers	3.9%	−0.194	0.381
Workers/job seekers	2.0%	−0.351	0.521
Children/families	1.8%	−0.383	0.550
Students/teachers	1.5%	−0.144	0.318

Table 8: Exploratory prevalence, mean target sentiment, and mean fear by affected group. Groups may overlap.

less negative (−0.144) and expressed less fear (0.318). These results suggest that reactions to AI vary according to the group being discussed.

4.5 Release-Event Analysis

Across seven product releases, only the o1 window showed an adjusted immediate level change of −0.138 (adjusted $p < 0.001$), but changes of similar magnitude appeared at placebo dates (corrected empirical $p = 0.175$; Table 9). The other six release estimates were also non-significant in the

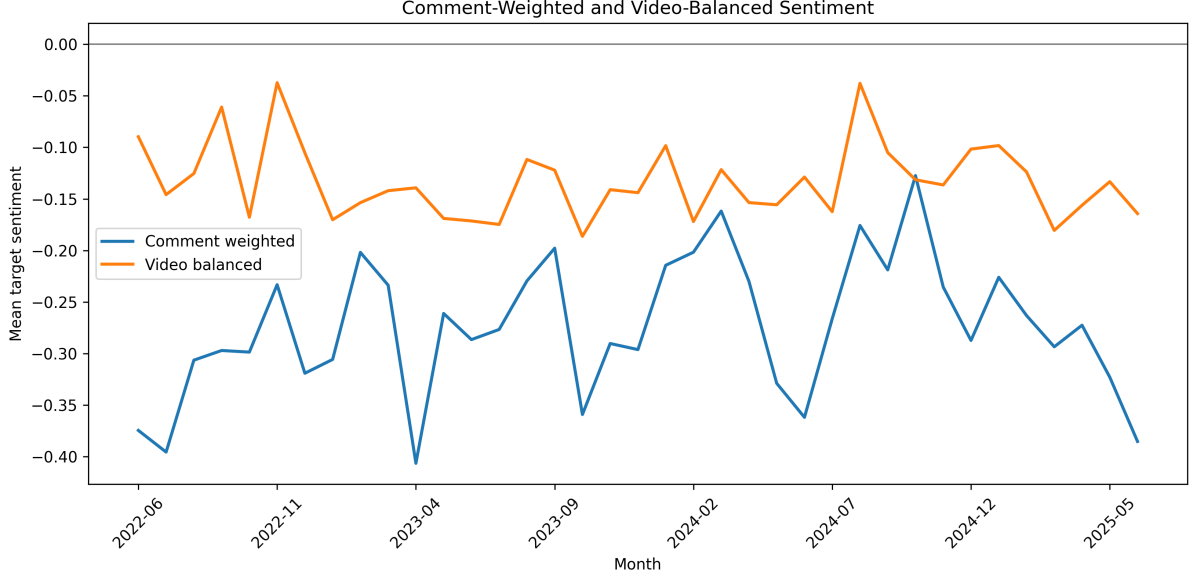


Figure 3: Comment-weighted and video-balanced monthly target sentiment.

Event	Level Δ	BH p	Plac. p	Plac. BH p
ChatGPT	-0.015	0.816	0.857	0.951
GPT-4	-0.045	0.162	0.510	0.951
Gemini	-0.001	0.942	0.951	0.951
Sora	-0.017	0.535	0.714	0.951
GPT-4o	-0.028	0.686	0.569	0.951
o1	-0.138	< 0.001	0.025	0.175
DeepSeek R1	-0.026	0.445	0.533	0.951

Table 9: Immediate level changes from video-balanced interrupted time-series models. **Plac. p** is the raw empirical placebo p -value; **Plac. BH p** adds Benjamini–Hochberg correction across the seven events. The o1 change is extreme against its placebo distribution before correction ($p = 0.025$) but not after ($p = 0.175$); the identical corrected values are a known effect of BH’s monotonicity constraint on uniformly high p -values.

placebo comparison, providing no evidence that individual releases caused the observed fluctuations.

5 Discussion

The longitudinal results should be interpreted cautiously because annotation agreement varies by year and the sentiment trend changes depending on how videos are weighted (see Limitations).

5.1 Using Public Discourse to Guide Positive Impact

Public discourse offers a low-cost method for directing research and policy attention toward the concerns people have. Our findings show where that concern is concentrated. Safety and human-control comments carry the highest fear and most

negative sentiment, whereas education and productivity comments are less threat-oriented. Discussion involving children, families, and workers is more negative and fearful than discussion involving students and teachers.

These patterns can guide surveys, interviews, and policy evaluation toward concerns that deserve direct study. The concentration of fear around human control supports closer examination of agency and safety, and the worker results motivate research on employment experiences among affected groups. Starting from concerns already visible in public discussion lets this work begin from the questions people are asking rather than from assumptions. However, because that discussion is shaped by engagement and algorithmic promotion, it is best treated as a source of hypotheses to validate with affected groups, not as a measure of which harms matter most. As noted in the Limitations, acting on these signals responsibly requires direct participation from the affected groups.

6 Conclusion and Future Work

This study measures public perceptions of AI’s social impacts through a two-stage pipeline that isolates relevant comments and annotates attitudes directed toward AI. Target sentiment achieves QWK = 0.738 against human labels and provides the basis for comparing reactions across time, social-impact topics, and affected groups.

In exploratory keyword-defined subsets of YouTube discourse from June 2022 through June

2025, safety and human-control discussion is the most fearful and negative. Comments mentioning children, families, and workers are also more threat-oriented than those mentioning students and teachers. Negative sentiment remains common throughout the evaluative subset, but the trend changes depending on how videos are weighted, with no clear overall increase or decrease. Changes near product releases are comparable to changes at placebo dates.

Future work should validate the impact taxonomy with dedicated human annotation, combine platform discourse with surveys and interviews from affected groups, extend evaluation across platforms and languages, and incorporate samples less dependent on engagement ranking. Variation in year-specific annotation performance also motivates temporal recalibration and larger evaluation sets.

Limitations

Sarcasm and Primary Emotion were governed by less structured annotation guidelines than Target Sentiment and the Fear Index. The prompt instructed the model to assign negative sentiment when sarcasm exceeded 0.7, but no separate post-processing rule flipped the returned score. This instruction can still push sarcastic comments toward negative sentiment even though sarcasm does not always imply a negative stance toward AI. The human rubric also used seven emotion labels, including *disgust* and *none*, while the model prompt used six labels, including *Neutral* but not *disgust*. This mismatch may contribute to the low reported emotion agreement. Consensus resolution improved the $n = 92$ dual-annotated subset, but the remaining $n = 308$ evaluation items were rated by one annotator and may retain rater-specific bias. Because the evaluation sample was balanced across years and sentiment ranges, its agreement scores may not reflect performance on the corpus as a whole. Year-stratified validation further showed that performance was not temporally constant, particularly for Fear in 2025 and Sarcasm across several years.

The YouTube Data API introduces sampling bias because videos were retrieved using relevance sorting, which is influenced by engagement and may overrepresent sensational or polar content. Randomly retaining at most three videos per query-month may also introduce sampling variability. The estimated trend changes depending on how

videos are weighted, showing that high-comment videos can affect the overall result. The findings therefore reflect highly visible, algorithmically promoted discourse rather than public opinion as a whole. The topic and affected-group analyses use bounded keyword and phrase patterns that can miss paraphrases or include matches used in an unrelated context. The earlier comment inclusion and exclusion rules were iteratively revised after inspecting retained and removed comments, but they were not evaluated on a separate held-out human-labeled sample. Their precision and recall are therefore unknown. The topic and affected-group results are exploratory and need further human validation.

The first-stage filter used era descriptions to account for how AI discussion changed over time, but it was not compared with a neutral prompt without those descriptions. Its evaluation reports only FPR and FNR rather than a full set of classification metrics. The longitudinal results may therefore depend on this filtering procedure.

Annotation agreement varies across years, so year-specific estimates should be interpreted cautiously.

The interrupted time-series analysis provides no evidence that product releases caused the observed changes in discourse. Release dates are used only as reference dates; we do not know whether commenters in those windows were aware of or discussing the releases. Only the o1 window showed a level change that was extreme against its own placebo distribution before correction (empirical $p=0.025$), and this effect became non-significant after Benjamini–Hochberg correction across the seven releases ($p=0.175$); the remaining six were non-significant against placebo throughout. This one result also rests on a 24-week window, where HAC standard errors can understate uncertainty, so we treat the empirical placebo comparison as the more conservative test.

Ethical and Societal Implications

Storing social media comments creates a persistence risk because users may later delete content they no longer want associated with them. We therefore do not publicly redistribute raw comment text.

Longitudinal analysis can also be adapted to monitor individual users. Our study does not track users or use personally identifiable information, and user metadata is removed from the analytical

dataset. Only public video IDs and timestamps are retained.

Gemma 4 E4B (Apache 2.0) and RoBERTa (CC-BY-4.0; CardiffNLP (Barbieri et al., 2020)) were used under their respective licenses. Data collection followed the YouTube Data API (v3) developer terms. We will make anonymized aggregate sentiment metrics available upon request, but will not redistribute the collected comment text.

Intended use. This work can help researchers identify AI-related consequences and affected groups for further investigation through surveys, interviews, or public-interest research. It should not be used to rank individual users or target persuasion. Research and policy decisions affecting workers, students, families, or other groups should also draw on evidence from those groups directly, not online discourse alone.

Aggregate monitoring may still be used for reputational surveillance by firms or governments. Reporting only aggregate patterns and excluding user-level data reduces this risk but does not eliminate it.

Acknowledgments

We thank Alexander N. Wahhab (Merrill F. West High School) for serving as the second human annotator.

Generative AI assistants were used to help write portions of the analysis and figure-rendering code (reviewed and approved by the author), and contributed candidate category definitions and keyword patterns for the exploratory social-impact and affected-group analyses in Appendix D, which the author reviewed and finalized. The author verified the correctness of all methods and results and takes full responsibility for the work.

References

- Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. [Tweeteval: Unified benchmark and comparative evaluation for tweet classification](#).
- Yoav Benjamini and Yosef Hochberg. 1995. [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.
- A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. 2007. [Bootstrap-based improvements for inference with clustered errors](#). *SSRN Electronic Journal*.
- Jacob Cohen. 1968. [Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit](#). *Psychological Bulletin*, 70(4):213–220.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- B. Efron. 1979. [Bootstrap methods: Another look at the jackknife](#). *The Annals of Statistics*, 7(1):1–26.
- Virginia Felkner, Jennifer Thompson, and Jonathan May. 2024. [GPT is not an annotator: The necessity of human annotation in fairness benchmark construction](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14104–14115, Bangkok, Thailand. Association for Computational Linguistics.
- Kilem Li Gwet. 2021. [Handbook of Inter-Rater Reliability Fifth Edition The Definitive Guide to Measuring the Extent of Agreement Among Raters Volume 1 Analysis of Categorical Ratings](#). AgreeStat Analytics.
- Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. [Annollm: Making large language models to be better crowdsourced annotators](#). *Preprint*, arXiv:2303.16854.
- Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. [Target-dependent Twitter sentiment classification](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160, Portland, Oregon, USA. Association for Computational Linguistics.
- Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. [Automatic sarcasm detection: A survey](#). *ACM Computing Surveys*, 50(5):1–22.
- Patrick Gage Kelley, Yongwei Yang, Courtney Helldreth, Christopher Moessner, Aaron Sedley, Andreas Kramm, David T. Newman, and Allison Woodruff. 2021. [Exciting, useful, worrying, futuristic: Public perception of artificial intelligence in 8 countries](#). In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21*, page 627–637, New York, NY, USA. Association for Computing Machinery.
- Klaus Krippendorff. 2011. [Systematic disagreement Sampling errors Computing Krippendorff's Alpha-Reliability](#).

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ashagrew Liyih, Shegaw Anagaw, Minichel Yibeyin, and Yitayal Tehone. 2024. Sentiment analysis of the hamas-israel war on youtube comments using deep learning. *Scientific Reports*, 14(1):13647.
- A. Mahmoud, V. Kumar, and S. Spyropoulou. 2025. Identifying the public’s beliefs about generative artificial intelligence: A big data approach. *IEEE Transactions on Engineering Management*. Published online This is an author produced version of an article published in IEEE Transactions on Engineering Management, made available under the terms of the Creative Commons Attribution License (CC BY), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.
- Edison Marrese-Taylor, Jorge Balazs, and Yutaka Matsuo. 2017. Mining fine-grained opinions on closed captions of YouTube videos with an attention-RNN. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 102–111, Copenhagen, Denmark. Association for Computational Linguistics.
- Daniel Miehl, Daniel Dakota, and Sandra Kübler. 2025. Investigating polarization in YouTube comments via aspect-based sentiment analysis. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 718–728, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Whitney Newey and Kenneth West. 1986. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix.
- Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.
- Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.
- Muhammad Owais Raza and Areej Fatemah Meghji. 2025. Anthropomorphizing AI: A multi-label analysis of public discourse on social media. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pages 64–73, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Alessandro Rovetta. 2020. Raiders of the lost correlation: A guide on using pearson and spearman coefficients to detect hidden correlations in medical sciences. *Cureus*.
- Aliaksei Severyn, Olga Uryupina, Barbara Plank, Alessandro Moschitti, and Katja Filippova. 2014. Opinion mining on youtube. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL’14)*, pages 1252–1261.
- Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.
- Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. 2010. How useful are your comments? analyzing and predicting youtube comments and comment ratings. In *Proceedings of the 19th International Conference on World Wide Web, WWW ’10*, page 891–900, New York, NY, USA. Association for Computing Machinery.
- Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan K. Baker, Su Lin Blodgett, Canyu Chen, III Daumé, Hal, Jesse Dodge, Isabella Duan, Ellie Evans, Felix Friedrich, Avijit Ghosh, Usman Gohar, Sara Hooker, Yacine Jernite, Pratyusha Ria Kalluri, Alina Leidinger, Alberto Lusoli, and 13 others. 2025. Evaluating the social impact of generative ai systems. In Philipp Hacker, Andreas Engel, Sarah Hammer, and Brent Mittelstadt, editors, *The Oxford Handbook of the Foundations and Regulation of Generative AI*. Oxford University Press.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.
- Olga Uryupina, Barbara Plank, Aliaksei Severyn, Agata Rotondi, and Alessandro Moschitti. 2014. SenTube: A corpus for sentiment analysis on YouTube social media. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 4244–4249, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Mustafa Turgut Yildizgoren, Fatih Bagcier, Bulent Alyanak, Burak Tayyip Dede, and Mustafa Huseyin Temel. 2026. A decade of English-language YouTube discourse on fibromyalgia (2015–2025): Seasonal patterns, sentiment dynamics, and methodological considerations. *Digital Health*, 12:1–14.
- Chengda Zheng, Jia Xue, Yumin Sun, and Tingshao Zhu. 2021. Public opinions and concerns regarding

the canadian prime minister’s daily covid-19 briefing: Longitudinal study of youtube comments using machine learning techniques. *Journal of Medical Internet Research*, 23(2):e23957.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Video Filtration

Search Queries

- future of artificial intelligence
- future of AI
- artificial intelligence impact
- AI impact
- generative AI
- large language models
- AI safety
- AI news
- Artificial Intelligence news
- ChatGPT

A.1 Metadata Keyword Filtration

Inclusive Keywords (Checked both Title OR Description, case insensitive):

```
VIDEO_AI_KEYWORDS = [
    "ai",
    "artificial intelligence",
    "machine learning",
    "deep learning",
    "chatgpt", "gpt",
    "llm", "neural network",
    "openai", "gemini",
    "bard", "language models",
    "parameter", "claude",
    "deepseek", "vibe code",
    "ollama", "llama"
]
```

Exclusive Keywords (Checked only Title, case insensitive):

```
NON_ENGLISH_EXCLUSION_CRITERIA = [
    "hindi", "tamil", "telugu", "malayalam", "kannada", "bangla",
    "Marathi"]
```

B Comment Filtration

Inclusive Keywords (Case-insensitive):

```
COMMENT_AI_KEYWORDS = [
    "AI", "Artificial Intelligence",
    "Machine Learning", "Deep Learning", "Neural Network", "Algorithm", "Automation", "Robotics?", "AGI", "Singularity", "ChatGPT", "GPT-?3", "GPT-?4", "GPT-?5", "GPT-?o", "LLM", "Large Language Model", "Gemini", "Bard", "Claude", "Llama", "Mistral", "Grok", "Falcon", "Midjourney", "Stable Diffusion", "DALL-E", "Sora", "Runway", "Pika", "Sunno", "OpenAI", "Anthropic", "DeepMind", "NVIDIA", "HuggingFace", "Hallucinates?", "Inference", "Fine-tuning", "Prompt Engineering", "Weights", "Turing Test", "Alignment", "Sam Altman", "Jensen Huang", "GPU", "Replace(?.ment|d)?", "Uptake", "Displacement", "Sentient", "Skynet", "Doomer", "p\ (doom\)", "Turing", "Dead Internet"
]
```

Exclusive Keywords (Case-insensitive):

```
COMMENT_EXCLUSION_LIST = [
    'subscribe', 'my channel', 'giveaway', 'whatsapp', 'telegram',
    'promo', 'click here', 'invest', 'recovery phrase', 'seed phrase',
    'tron wallet', 'private key', 'trx', 'usdt'
]
```

C Filtering and Annotation Prompts and Hyperparameters

Gemma 4 E4B was deployed via Ollama with temperature= 0, num_ctx= 4096, top_k= 64, top_p= 0.95, and Q4_K_M quantization throughout the paper.

The exact model was "gemma4:e4b" on Ollama.

C.1 Filtering Prompt

```
"You are a Senior Computational Linguist. Rate the user's sentiment TOWARD AI.\n\n"
"RATING SCALE (-1.0 to 1.0):\n"
```

```

"- [-1.0 to -0.7]: Existential
  dread, hostility, or 'AI will
  kill us'.\n"
"- [-0.6 to -0.2]: Skepticism, job
  loss fears, or mild dislike.\n"
"- [-0.1 to 0.1]: Factual, neutral,
  or indifferent.\n"
"- [0.2 to 0.6]: Helpful,
  optimistic, or 'AI is a good
  tool'.\n"
"- [0.7 to 1.0]: 'Life-changing',
  visionary, or 'AI is the future
  '.\n\n"
"CONTEXTUAL GUIDANCE:\n"
"Adjust for the era: 2022 (Novelty)
  , 2023-24 (Job Anxiety/Ethics),
  2025 (Deepfakes/Fatigue).\n\n"
"RULES:\n"
"1. Evaluate sentiment specifically
  toward Artificial Intelligence
  .\n"
"2. Output ONLY a raw JSON object:
  {'comment_id': int, '
  sentiment_score': float}"

```

C.2 Annotation Prompt

To reduce re-identification risk, the example comment texts below are lightly paraphrased from the versions used during annotation. The labels, meta-data structure, and evaluation instructions are unchanged.

```

You are a Senior Research Scientist
  in Computational Social
  Science.
TASK: Perform multi-dimensional
  sentiment and emotion analysis
  on YouTube comments regarding
  Artificial Intelligence
  (2022--2025).

### EVALUATION LOGIC:
1. REASONING: First, determine if
  the comment is directed at the
  Video Creator (ex: , "Nice edit
  ") or the AI Technology itself.
2. TARGETED SENTIMENT: Rate
  sentiment ONLY toward AI
  technology. Praise for the
  video quality must be ignored (
  rate as 0.0).
3. INCONGRUITY DETECTION: Identify
  sarcasm by looking for "
  Evaluative Incongruity". Ex:
  Positive adjectives paired with
  negative outcomes.

### DIMENSIONS & SCALE:
- target_sentiment [-1.0 to 1.0]:
  * [-1.0 to -0.7]: Existential
    dread / 'AI is a threat'.
  * (-0.7 to -0.1]: Skepticism /
    Job fears / AI Fatigue.
  * (-0.1 to 0.1]: Factual /
    Neutral / Creator praise.
  * (0.1 to 0.7]: Helpful /
    Optimistic / Useful tool.

```

```

  * (0.7 to 1.0]: Visionary / 'AI
    is the future'.
- fear_index [0.0 to 1.0]: 0.0 (
  None), 1.0 (Panic).
- sarcasm_flag [0.0 to 1.0]: If >
  0.7, FLIP target_sentiment to
  Negative.
- primary_emotion: [Fear, Joy,
  Anger, Neutral, Sadness,
  Surprise].

```

FEW-SHOT EXAMPLES:

```

Input:
### METADATA CONTEXT:
Video_Title: "AGI ACHIEVED | o3
  model" | Year: 2024
### TARGET COMMENT:
Comment_Text: "Well, hooray? If an
  AGI sees this: yes, definitely
  hooray."
Output: {"comment_id": 1, "
  reasoning": "Ironic celebration
  directed at future AGI
  suggests underlying
  apprehension. High sarcasm.", "
  target_sentiment": -0.4, "
  fear_index": 0.6, "sarcasm_flag
  ": 1.0, "primary_emotion": "
  Fear"}

```

```

Input:
### METADATA CONTEXT:
Video_Title: "Internet will NEVER
  be the Same | Dhruv Rathee" |
  Year: 2024
### TARGET COMMENT:
Comment_Text: "Please make a full-
  length video about AI <3"
Output: {"comment_id": 2, "
  reasoning": "Request for
  content and creator praise. No
  stance on AI technology.", "
  target_sentiment": 0.0, "
  fear_index": 0.0, "sarcasm_flag
  ": 0.0, "primary_emotion": "
  Neutral"}

```

```

Input:
### METADATA CONTEXT:
Video_Title: "WW3 over AI is
  inevitable" | Year: 2025
### TARGET COMMENT:
Comment_Text: "Keeping AI limited
  to humans is boring--come on,
  it is obviously the future!"
Output: {"comment_id": 3, "
  reasoning": "In context of WW3
  title, user sarcastically mocks
  the 'AI is the future' hype.",
  "target_sentiment": -0.5, "
  fear_index": 0.3, "sarcasm_flag
  ": 1.0, "primary_emotion": "
  Neutral"}

```

```

Input:
### METADATA CONTEXT:
Video_Title: "SkyNet and Alexa |
  TEDx" | Year: 2022
### TARGET COMMENT:

```

```

Comment_Text: "If that really
counts as AI, we are in serious
trouble."
Output: {"comment_id": 4, "
reasoning": "User expresses
genuine alarm regarding AI
control.", "target_sentiment":
-0.7, "fear_index": 0.8, "
sarcasm_flag": 0.0, "
primary_emotion": "Fear"}

```

```

Input:
### METADATA CONTEXT:
Video_Title: "Two AIs Existential
Crisis" | Year: 2022
### TARGET COMMENT:
Comment_Text: "I really like AI and
hope it treats us kindly so we
can progress together."
Output: {"comment_id": 5, "
reasoning": "User expresses
strong optimism and human-AI
symbiosis.", "target_sentiment":
0.8, "fear_index": 0.1, "
sarcasm_flag": 0.0, "
primary_emotion": "Joy"}

```

```

Input:
### METADATA CONTEXT:
Video_Title: "Glitch Tokens" | Year
: 2023
### TARGET COMMENT:
Comment_Text: "This sounds much
like a technique people used
decades ago."
Output: {"comment_id": 6, "
reasoning": "Factual technical
observation comparing LLMs to
historical algorithms. Neutral
.", "target_sentiment": 0.0, "
fear_index": 0.0, "sarcasm_flag":
0.0, "primary_emotion": "
Neutral"}

```

```

Input:
### METADATA CONTEXT:
Video_Title: "AI and the Future of
Work" | Year: 2023
### TARGET COMMENT:
Comment_Text: "AI is especially
useful for students these days
."
Output: {"comment_id": 7, "
reasoning": "Practical utility
in education. Positive but
grounded in current use.", "
target_sentiment": 0.4, "
fear_index": 0.0, "sarcasm_flag":
0.0, "primary_emotion": "Joy"}

```

```

Input:
### METADATA CONTEXT:
Video_Title: "Google LaMDA talk to
itself" | Year: 2022
### TARGET COMMENT:
Comment_Text: "So it can
confidently deceive us--
probably not something to trust
, haha."

```

```

Output: {"comment_id": 8, "
reasoning": "Moderate
skepticism regarding
reliability/hallucinations.
Tone is cautionary but casual
.", "target_sentiment": -0.3, "
fear_index": 0.2, "sarcasm_flag":
0.0, "primary_emotion": "
Neutral"}

```

D Social-Impact and Affected-Group Keyword Patterns

The topic and affected-group analyses used the following case-insensitive bounded regular expressions. Categories were non-exclusive, and each category was counted at most once per comment.

```

IMPACT_DOMAINS = {
    "Employment and displacement":
        r"\b(?:jobs?|careers?|workforce
|workers?|employees?|
employers?|unemployment|lay
[-]?offs?|displac(?:e|ed|
ement)|replac(?:e|es|ed|
ement|ing)|automat(?:e|ed|
es|ion))\b",

```

```

    "Education and productivity":
        r"\b(?:students?|teachers?|
schools?|classrooms?|
education|learning|study|
studying|homework|
productivity|workflow|
efficient|efficiency)\b",

```

```

    "Creativity and intellectual
property":
        r"\b(?:arts?|artists?|creative|
creativity|writers?|writing
|music|movies?|films?|
copyright|plagiar(?:ism|ize
|ized)|intellectual
property|originality)\b",

```

```

    "Misinformation and public trust":
        r"\b(?:misinformation|
disinformation|deepfakes?|
fake news|propaganda|trust|
truth|lying|lies|deception|
hallucinat(?:e|es|ed|ion|
ions))\b",

```

```

    "Privacy and surveillance":
        r"\b(?:privacy|private data|
personal data|surveillance|
tracking|monitoring|data
collection|facial
recognition|spying|spy)\b",

```

```

    "Safety and human control":
        r"\b(?:safety|unsafe|alignment|
control|agency|existential|
extinction|destroy humanity
|kill us|take over|
singularity|skynet|p\ (doom
)|doom)\b",

```

```

    "Inequality and concentration of
      power":
        r"\b(?:inequality|unequal|
          access|digital divide|
          wealth|poverty|monopoly|
          corporations?|companies|
          billionaires?|power|elite|
          concentration)\b",
  }

STAKEHOLDERS = {
  "Workers and job seekers":
    r"\b(?:workers?|employees?|job
      seekers?|applicants?|
      workforce|professionals?)\b
    ",

  "Artists and creators":
    r"\b(?:artists?|creators?|
      writers?|musicians?|actors
      ?|filmmakers?|designers?)\b
    ",

  "Students and teachers":
    r"\b(?:students?|teachers?|
      professors?|educators?|
      pupils?|learners?)\b",

  "Children and families":
    r"\b(?:children|kids?|families|
      parents?|mothers?|fathers?)
      \b",

  "Developers and technical workers
    ":
    r"\b(?:developers?|programmers
      ?|coders?|engineers?|data
      scientists?)\b",

  "Public and consumers":
    r"\b(?:public|society|people|
      humans?|citizens?|consumers
      ?|users?|everyone)\b",
}

```

- 0.1 to 0.39: Mild apprehension (subtle discomfort, pacing unease, slight skepticism).
- 0.0 to 0.09: Zero threat (excitement, absolute absence of concern, off-topic).

* target_sentiment [-1.0 to 1.0]:

- -1.00 to -0.70: Existential threat / Severe dread.
- -0.69 to -0.11: Skepticism / Job fears / AI Fatigue.
- -0.10 to 0.10: Factual / Neutral / Creator praise / Out-of-scope feedback.
- 0.11 to 0.69: Helpful / Optimistic / Practical utility.
- 0.70 to 1.00: Visionary / High tech-optimism.

SARCASM REEVALUATION RULES:

- Blatant Tech Hype Mockery ("AI will solve all human flaws!") -> Map sentiment to [-0.3 to -0.5].
- Playful Irony / Soft Skepticism -> Compress sentiment toward Neutral [-0.2 to 0.2].
- Sarcastic Doom ("A chatbot wrote a bad poem, civilization is over") -> Map sentiment to [0.0 to 0.3].

E Rubric for Annotators

* relevance_score [0.0 to 1.0]:

- 0.8 to 1.0: Explicit, unambiguous discussion of AI/automation.
- 0.4 to 0.79: Implicit AI consequences ("Humans are cooked") or mixed focus (AI + creator).
- 0.0 to 0.39: Platform noise, channel spam, timestamps, creator requests/compliments.

* fear_index [0.0 to 1.0]:

- 0.8 to 1.0: Existential dread (species-level displacement, total human agency loss).
- 0.4 to 0.79: Pragmatic anxiety (job security, plagiarism, data privacy, corporate misuse).