Constructive Comment Extraction from Japanese YouTube Shorts: A Dataset and Empirical Study

Siyuan Feng[1*], **Luecheng Wang**[1*], **Koyo Ogura**[1], **Yuma Iwamoto**[2], **Seinosuke Sato**[2], **Yuki Kometani**[2], **Masaki Itabashi**[2], **Shintaro Ozaki**[3], **Yusuke Sakai**[3], **Shogo Matsuno**[2], **Hidetaka Kamigaito**[3], **Katsuhiko Hayashi**[1], **Ryoma Obara**[4], **Sho Takase**[5], **Taro Watanabe**[3]

[1]The University of Tokyo [2]The University of Electro-Communications
[3]Nara Institute of Science and Technology [4]NEC Knowledge Science Research Laboratories
[5]CyberAgent, Inc.

{9445233883, wanglc, ogura-koyo703, katsuhiko-hayashi}@g.ecc.u-tokyo.ac.jp

Abstract

Identifying constructive comments from user-generated content is important for improving information discovery and discussion quality on platforms such as YouTube. However, existing resources focus predominantly on English news comments, leaving short-form video platforms underexplored despite their distinct characteristics: comments are shorter, more reactive, and not well captured by popularity signals. In this work, we introduce the first Japanese YouTube Shorts dataset for constructive comment extraction, featuring multi-annotator labels that preserve naturally occurring disagreement. We also develop a real-time system that surfaces constructive comments from a noisy comment stream, evaluated in a preliminary user study. To train under noisy annotation conditions, we combine self-training with high-confidence pseudo-labels and active learning. Experiments show that iterative data expansion improves performance, while our compact task-specific model achieves the highest PR-AUC point estimate among the evaluated systems. These findings demonstrate the value of dataset construction and targeted training for constructive comment extraction in realistic settings. Our dataset and code are available at https://github.com/nlp-hayashi/japanese-youtube-constructive-comments.

1 Introduction

User-generated comments on platforms such as YouTube vary widely in quality, ranging from low-effort reactions to informative and constructive contributions (Kolhatkar et al., 2020; Napoles et al., 2017). In this work, we define a constructive comment as one that is relevant to the video, contains no abusive language, and provides useful information not already covered in the video. Identifying such comments is important for improving information discovery and discussion quality on video platforms (Kolhatkar et al., 2023; Santana, 2019). As illustrated in Figure 1, surfacing these comments from a noisy stream of reactions would make them far easier to discover.

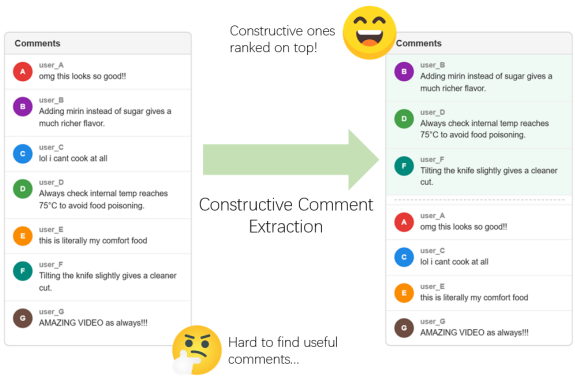


Figure 1: Constructive comments (highlighted in green) are ranked at the top, making them easier to find among low-effort reactions.

However, existing resources for this task focus predominantly on English news comments (Kolhatkar et al., 2023; Napoles et al., 2017), leaving short-form video platforms such as YouTube underexplored. Comments on such platforms differ substantially: they are typically brief, fragmented, and loosely connected to the video content, posing new challenges for constructiveness assessment. No dataset currently exists for this setting, and the subjective nature of constructiveness further complicates annotation (Plank, 2022): even with explicit criteria, human judgments frequently disagree (Aroyo and Welty, 2015; Davani et al., 2022).

In this work, we address this gap by introducing the first Japanese YouTube Shorts dataset for constructive comment extraction. We design a two-step annotation scheme tailored to the short-form video setting, and systematically record inter-

[*] Equal contribution.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 22–34
November 6-10, 2026 ©2026 Association for Computational Linguistics

annotator disagreement. To train a robust classifier under noisy annotation conditions, we combine self-training with high-confidence pseudo-labels and active learning (Northcutt et al., 2021; Karamanolakis et al., 2021; van der Meer et al., 2024), using video-level splits to prevent information leakage (Kapoor and Narayanan, 2023). The resulting system can surface constructive comments in real time, as demonstrated by our browser-based demo. The dataset constitutes the primary contribution of this work; the deployed classifier and browser extension serve to demonstrate the resource's practical value rather than to advance classification methodology.

Our main contributions are:

- We construct and release the first Japanese YouTube Shorts dataset for constructive comment extraction, with multi-annotator labels that preserve naturally occurring disagreement, and iteratively expanded via active learning-driven annotation to improve coverage of ambiguous boundary cases.
- We develop and publicly demonstrate a real-time constructive comment extraction system, showing the practical applicability of our approach on Japanese YouTube Shorts.

2 Dataset and Annotation

2.1 Data Collection

Our dataset consists of user-generated comments from YouTube Shorts videos. The initial 90 videos were selected through manual recommendations by project participants as a diversity-oriented convenience sample across different content sources, without constraints on topic or domain. Comments are collected between July and August 2025 using the YouTube API. For each video, we retrieve top-level comments and randomly sample up to 50 per video to ensure balanced coverage. The initial dataset consists of 3,797 comments from 90 videos. The dataset is subsequently expanded through active learning (Section 3), adding approximately 800 comments across two active-learning rounds, yielding a final released dataset of 4,597 comments from approximately 400 videos in total. Video-level splits are used throughout to prevent information leakage across training, validation, and test sets.

Dataset	Lang	Platform	#Comments	Avg. length
SOCC	EN	News	1,043	284
YNACC	EN	News	9,160	232
Tabuchi et al.	JA	News	49,215	—
Ours	**JA**	**Shorts**	**4,597**	**33**

Table 1: Comparison with existing constructive comment datasets (Kolhatkar et al., 2020; Napoles et al., 2017; Tabuchi et al., 2019). Avg. length is mean comment length in characters; — indicates not reported.

2.2 Dataset Characteristics

Table 1 compares our dataset with existing constructive comment datasets. Unlike prior work focusing on news comments, our dataset targets short-form video comments, which are markedly shorter (mean 33 vs. 232 characters in YNACC (Napoles et al., 2017)), reflecting the reactive and fragmented nature of the platform.

Like counts retrieved via the YouTube Data API for 4,105 comments show that popularity and constructiveness are largely independent (Spearman $\rho = 0.056$, $p < 0.001$, $n = 4{,}105$): only 20% of the top-100 most-liked comments are constructive, comparable to the overall rate of 21.6%, confirming that manual annotation is necessary. Representative examples are provided in Appendix E.6.

2.3 Annotation Design

We design a two-step annotation scheme that separates basic validity from additional informational value.

Annotation criteria. Each comment is evaluated using two binary criteria in sequence. The first (*premise*) requires that the comment is relevant to the video and does not contain abusive language (Wulczyn et al., 2017; Ashraf et al., 2021). The second (*additional information*) assesses whether the comment provides useful information not already covered in the video. A comment is *constructive* only if both criteria are satisfied.

Annotation procedure. Annotations were produced by a pool of 14 annotators, all either native Japanese speakers or JLPT N1 certified. Before annotation, all annotators received online training in which the two criteria were explained using representative examples. Annotation was conducted item by item via Google Forms, with a direct link to the corresponding Shorts video, and annotators were instructed to watch the full video before labeling. Each comment was independently labeled

by at least two annotators; disputed cases received an additional independent annotation, and the final label was determined by majority vote.

Inter-annotator agreement. Despite explicit criteria, we observe substantial disagreement: agreement rate is 81.37% for two-annotator pairs, dropping to 67.90% for three or more annotators, with 21.33% of multiply-annotated comments exhibiting conflicting labels. Krippendorff's $\alpha = 0.41$ (Cohen's $\kappa = 0.41$ on the two-annotator subset), indicating only moderate agreement once chance is accounted for (Rodríguez-Barroso et al., 2024). Among the 779 comments with disagreement on the final constructiveness label, 669 (85.9%) differed only on the additional-information criterion, whereas only 5 (0.6%) differed only on the premise criterion. This suggests that the main source of ambiguity is whether a comment provides sufficiently useful information beyond the video, rather than whether it satisfies the premise criterion. Qualitatively, additional-information disagreements typically arise when a comment's informativeness is context-dependent: one annotator regards it as a useful contribution while another treats it as common knowledge or a mere reaction. Premise-level disagreements, by contrast, are rare and usually involve borderline cases where relevance or the presence of mild sarcasm is itself debatable. A criterion-level disagreement analysis and representative examples of both types are provided in Appendix E.4.

2.4 Gold Dataset Construction

We first identify a high-confidence subset consisting of comments with unanimous agreement among at least two annotators, yielding 2,873 instances (361 constructive, 2,512 non-constructive). Importantly, disputed comments are not discarded. For the remaining comments, including those with annotator disagreement and those with insufficient initial annotations, a third annotator provides an independent tie-breaking judgment, and the final label is determined by majority vote. This process yields a full gold dataset of 3,797 comments. The training split is constructed from this full gold dataset, so disputed comments are retained in supervised training rather than discarded. Comments added through active learning follow the same annotation procedure: each selected comment is labeled by at least two annotators using the same two-step criteria, with tie-breaking by a third annotator where needed.

3 Method

To build the constructive comment extraction system, we train a classifier on the dataset described in Section 2. Given the annotation inconsistency inherent to this task, we address two practical challenges sequentially: (1) the limited amount of labeled data, tackled via self-training with high-confidence pseudo-labels; and (2) the prevalence of ambiguous boundary cases that self-training cannot resolve, tackled via active learning with human annotation.

3.1 Base Classifier

We train a comment-level binary classifier using `sbintuitions/modernbert-ja-130m` (Warner et al., 2025; Tsukagoshi et al., 2025), a Japanese pretrained encoder, on the gold training set described in Section 2. We select this model for its strong Japanese language understanding and compact size (130M parameters), which enables real-time inference in the deployed system. The model takes a single comment as input and predicts whether it is constructive. Training details and hyperparameters are reported in Section 4.

3.2 Self-Training

To expand the training signal beyond the gold training data, we apply self-training on a large pool of unannotated Japanese YouTube Shorts comments drawn from the same source corpus as the labeled data.[1] Using the base classifier as the teacher, we score all unlabeled comments and select high-confidence pseudo-labels according to fixed confidence thresholds: comments with $p \in [0, 0.05)$ are treated as pseudo-negatives (up to 2,000 per round, capped at 20 per video), and comments with $p \in [0.98, 1.0]$ as pseudo-positives (up to 500 per round, capped at 20 per video). The resulting pseudo-labeled set (approximately 2,500 instances) is merged with the gold training data to train a student model, warm-started from the teacher checkpoint.

We perform two rounds of self-training. In each round, the current best student is used to re-score the unlabeled pool, and a new student is trained on the updated pseudo-labeled set. As pseudo-label selection deliberately avoids uncertain regions

[1]The unlabeled pool consists of comments from Shorts videos not included in the labeled dataset, identified by filtering for `#shorts` in video titles.

($p \approx 0.5$), self-training is expected to improve performance on confident cases but saturate quickly, which is confirmed in our experiments (Section 4).

3.3 Active Learning

To address the ambiguous cases that self-training leaves unresolved, we apply uncertainty-based active learning (Lewis and Gale, 1994). After self-training saturates, we use the current best student to score the unlabeled pool and select the most uncertain comments by ranking on $|p - 0.5|$ (lower is more uncertain) (Nguyen et al., 2022), subject to a cap of 5 comments per video to ensure coverage across diverse sources. Comments already present in the training, validation, test sets, or previous AL rounds are excluded; comments from test-set videos are excluded entirely to prevent information leakage.

In each round, approximately 400 comments are selected and manually annotated. The newly labeled instances are merged with the gold training data and all accumulated AL annotations, and a new student is trained with pseudo-labels generated by the current best checkpoint (using the same confidence thresholds as in self-training). We perform two rounds of active learning, yielding approximately 800 additional annotated comments in total.

The full training pipeline is summarized as follows: Teacher (gold only) $\rightarrow$ Student$_1$ (gold + self-training pseudo-labels) $\rightarrow$ Student$_2$ (Student$_1$ + AL round 1) $\rightarrow$ Student$_3$ (Student$_2$ + AL round 2). The final model Student$_3$ is deployed in a browser-based demo system for real-time constructive comment extraction (Appendix C).

4 Experiments

4.1 Experimental Setup

Dataset splits. We evaluate on the full gold dataset of 3,797 comments (792 constructive, 3,005 non-constructive) from 90 videos, split by video into training (2,576), validation (561), and test (660) sets. Active learning annotations (approximately 800 comments across two rounds) are added exclusively to the training set.

Evaluation metrics. Given the test-set class imbalance (constructive:non-constructive $\approx$ 1:7), we report both ROC-AUC and PR-AUC, with emphasis on PR-AUC because it better reflects ranking performance on the minority class (Saito and Rehmsmeier, 2015). Both metrics are threshold-independent and align with our deployment setting, where comments are ranked by continuous constructiveness scores. We considered additional metrics based on fixed decision thresholds or per-video rankings, but found them unreliable or poorly aligned with our deployment setting at the current dataset scale, and therefore do not report them as primary results.

Model	ROC-AUC	PR-AUC
GPT-4o (0-shot)	0.835	0.653
GPT-4o (0-shot, CoT)	0.802	0.617
GPT-4o (few-shot)	0.856	0.669
GPT-4o (few-shot, CoT)	0.796	0.583
Llama-3.1-8B (0-shot)	0.720	0.405
Llama-3.1-8B (few-shot)	0.684	0.339
Qwen2.5-7B (0-shot)	0.690	0.370
Qwen2.5-7B (few-shot)	0.713	0.406
Teacher	0.845	0.618
+ ST round 1	0.877	0.677
+ ST round 2	0.876	0.679
+ AL round 1	0.879	0.682
+ AL round 2	**0.883**	**0.701**

Table 2: Results on the test set. ST = self-training, AL = active learning, CoT = chain-of-thought prompting. LLM ROC-AUC and PR-AUC are computed from token-level output probabilities.

Training details. In all experiments, we use `sbintuitions/modernbert-ja-130m` (Warner et al., 2025; Tsukagoshi et al., 2025) as the encoder with a two-label classification head. The base classifier is trained with AdamW, learning rate 2×10^{-5}, batch size 16, and early stopping on validation PR-AUC. Student models are warm-started from the previous best checkpoint with learning rate 2×10^{-6}. Class-weighted loss is applied to gold labels; pseudo-labeled instances use soft targets. Full training details are in Appendix A.

Baselines. For LLM baselines, we evaluate **GPT-4o** (OpenAI et al., 2024), **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024), and **Qwen2.5-7B-Instruct** (Qwen et al., 2025) under both zero-shot and few-shot settings, using a prompt that describes the two-step annotation criteria. ROC-AUC and PR-AUC are computed from token-level output probabilities. For GPT-4o, we additionally evaluate a chain-of-thought (CoT) variant of both prompts; the exact prompts are given in Appendix B.

4.2 Main Results

Table 2 reports results on the test set for all baselines and the full training pipeline.

LLM baselines show mixed results. GPT-4o achieves competitive performance under both zero-shot and few-shot settings (PR-AUC 0.653 and 0.669), surpassing the Teacher baseline (0.618). Adding chain-of-thought prompting to GPT-4o lowers performance in both settings (PR-AUC 0.617 and 0.583), so we use direct prompting for the main baselines. In contrast, the evaluated open-source models perform substantially worse (PR-AUC 0.339–0.406). Our final model Student_3 achieves the highest PR-AUC point estimate (0.701) among the evaluated systems while using a compact 130M-parameter encoder suitable for real-time deployment.

Paired video-level bootstrap analysis with 5,000 samples (Appendix E.1) shows that Student_3 significantly improves over the Teacher in PR-AUC (+0.083, 95% CI [0.034, 0.152]). However, its differences from GPT-4o under zero-shot and few-shot prompting are not statistically significant (95% CIs $[-0.067, 0.152]$ and $[-0.082, 0.131]$, respectively). We therefore interpret the higher PR-AUC of Student_3 as a point-estimate advantage rather than evidence of statistically significant superiority over GPT-4o. The same caution applies within our own pipeline: the stage-to-stage comparisons (AL2 vs. ST1, ST2, and AL1; Appendix E.1) also yield confidence intervals that include zero. Hence, although the PR-AUC point estimate increases monotonically across training stages, we report the per-stage gains as consistent point-estimate improvements rather than as statistically significant effects.

Examining the per-stage gains in more detail, the Teacher provides a baseline of 0.618. Self-training yields the largest gain in the first round (+0.059), while the second round provides only marginal improvement (+0.002). Subsequent human annotation increases the point estimate further, reaching 0.701 after the second active-learning round.

4.3 Analysis

Self-training saturates quickly. The large gain in ST round 1 (+0.059 PR-AUC) followed by near-zero improvement in round 2 (+0.002) reflects our pseudo-label selection strategy: by restricting pseudo-labels to high-confidence predictions ($p \in [0, 0.05)$ and $p \in [0.98, 1.0]$), self-training effectively utilizes confident examples but has limited impact on uncertain boundary cases. Once high-confidence examples are exhausted, additional rounds provide diminishing returns. Performance is also stable across the tested pseudo-label confidence thresholds (Appendix E.2).

Additional annotation after self-training. Active learning is applied from Student_1 rather than the second self-training checkpoint, as the second self-training round yields only marginal improvement. After self-training saturates, we add human annotations selected through uncertainty sampling. These additional annotations increase the PR-AUC point estimate in both rounds. However, under the same annotation budget, uncertainty-based acquisition performs comparably to random sampling (PR-AUC 0.6838 vs. 0.6906; 95% CI for the difference $[-0.0300, 0.0282]$). Full PR-AUC and ROC-AUC results for this controlled comparison are reported in Appendix E.3. Thus, at the present scale, the results support the benefit of additional human annotation after self-training, but do not establish an advantage of uncertainty-based acquisition over random sampling. This suggests that, at the current annotation budget, the improvement stems primarily from adding new human-verified labels rather than from the criterion used to select which comments to annotate. Whether uncertainty-based acquisition provides a measurable benefit at larger annotation scales remains an open question for future work.

5 Conclusion

We introduced the first Japanese YouTube Shorts dataset for constructive comment extraction, along with a real-time system that surfaces constructive comments from a noisy stream. Unlike news comment datasets, short-form video comments are markedly shorter and popularity signals are poor proxies for constructiveness. To train under annotation inconsistency, we combined self-training with high-confidence pseudo-labels and active learning. Experiments show that self-training effectively expands the training signal but saturates quickly, while subsequent human annotation provides additional supervision. Notably, our compact 130M-parameter model attains the highest PR-AUC point estimate among all evaluated systems, matching frontier LLMs while remaining light enough for real-time deployment. At the current annotation scale, uncertainty-based and random acquisition perform comparably. Our dataset and system provide a foundation for future research on constructive comment extraction in non-English, short-form video settings, with future work exploring other languages, platforms, and video-level context.

Limitations

Our dataset covers Japanese YouTube Shorts only, which may limit generalizability to other languages, video formats, or platforms. Although the final released dataset comprises 4,597 comments from approximately 400 videos, its scale remains modest. Moreover, multiple comments are sampled from the same videos, and the dataset was not constructed using a predefined genre taxonomy. Generalization across video sources and genres therefore requires further evaluation on larger, genre-stratified corpora. The demo system currently supports Japanese only. We conducted a preliminary questionnaire-based user study with eight participants (Appendix C); its results should be interpreted as an initial usability evaluation rather than a population-level behavioral study. Evaluation with a larger and more diverse user group remains future work.

Although annotator disagreement is retained in the released annotations, the current classifier ultimately uses majority-vote gold labels rather than explicitly modeling label distributions. Disagreement-aware approaches such as soft-label or probabilistic training remain future work. Self-training relies on high-confidence pseudo-labels and may reinforce existing model biases. In addition, uncertainty-based acquisition did not significantly outperform random sampling at the present annotation budget, so the benefit of the acquisition strategy itself requires further study.

Ethical Considerations

Data collection and privacy. Comments were collected from publicly accessible YouTube Shorts videos using the YouTube Data API for research purposes. No private or restricted content was accessed. Usernames and avatars were anonymized in all screenshots and demonstrations included in the paper. Subject to platform license and agreement constraints, the released dataset contains comment IDs, annotator-level labels, data splits, and reconstruction scripts for publicly accessible comments. Comment text is not redistributed directly in order to comply with platform terms of service. We also release the annotation instructions, final model weights, training and evaluation code, and the Chrome extension to support reproducibility.

Potentially harmful content. The annotation guidelines define constructive comments as comments that do not contain abusive language (Section 2). Because the data consist of user-generated comments, annotators were informed in advance that they might encounter offensive or harmful content during annotation. Representative examples shown in the paper were selected to avoid reproducing harmful material.

Human annotation. All annotators participated voluntarily and provided informed consent for their annotations to be used in this research. No personally identifying information about annotators is included in this paper.

Use of generative AI. Large language models such as GPT-4o and GPT-5.4 were used only as experimental baselines for comment classification. Generative AI tools were also used for language editing assistance during manuscript preparation.

References

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Noman Ashraf, Arkaitz Zubiaga, and Alexander Gelbukh. 2021. Abusive language detection in youtube comments leveraging replies as conversational context. *PeerJ Computer Science*, 7:e742.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804.

Giannis Karamanolakis, Subhabrata Mukherjee, Guoqing Zheng, and Ahmed Hassan Awadallah. 2021. Self-training with weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 845–863, Online. Association for Computational Linguistics.

Varada Kolhatkar, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada. 2023. Classifying constructive comments. *First Monday*, 28(4).

Varada Kolhatkar, Haifei Wu, Luca Cavasso, Maite Francis, and Maite Taboada. 2020. The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4(2):155–190.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. *Preprint*, arXiv:cmp-lg/9407020.

Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding good conversations online: The Yahoo News annotated comments corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 13–23, Valencia, Spain. Association for Computational Linguistics.

Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. 2022. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *J. Artif. Int. Res.*, 70:1373–1411.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nuria Rodríguez-Barroso, Eugenio Martínez Cámara, Jose Camacho Collados, M. Victoria Luzón, and Francisco Herrera. 2024. Federated learning for exploiting annotators' disagreements in natural language processing. *Transactions of the Association for Computational Linguistics*, 12:630–648.

Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):1–21.

Arthur D. Santana. 2019. Toward quality discourse: Measuring the effect of user identity in commenting forums. *Newspaper Research Journal*, 40(4):467–486.

Yoshimune Tabuchi, Hayato Kobayashi, and Kazuma Murao. 2019. Introduction of a constructive comment ranking model on Yahoo! news. In *Proceedings of the 25th Annual Meeting of the Association for Natural Language Processing*, pages 1371–1374, Nagoya, Japan. Association for Natural Language Processing. P7-33. In Japanese.

Hayato Tsukagoshi, Shengzhe Li, Akihiko Fukuchi, and Tomohide Shibata. 2025. ModernBERT-Ja. `https://huggingface.co/collections/sbintuitions/modernbert-ja-67b68fe891132877cf67aa0a`.

Michiel van der Meer, Neele Falk, Pradeep K. Murukannaiah, and Enrico Liscio. 2024. Annotator-centric active learning for subjective NLP tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

A Training Details

In all experiments, we train all models with `sbintuitions/modernbert-ja-130m` (Warner et al., 2025; Tsukagoshi et al., 2025) as the encoder and a two-label classification head, using the Hugging Face `Trainer` (FP16, seed 42).

Hyperparameter	Teacher	Student
Learning rate	2×10^{-5}	2×10^{-6}
Batch size	16	
Max epochs	10	
Early stopping patience	3	5
Warmup ratio	—	0.06
Weight decay	0.05	
Optimizer	AdamW	
Early stopping metric	Val. PR-AUC	
Warm-start	No	Yes

Table 3: Hyperparameters for all models.

B LLM Baseline Prompts

The same prompt template is used for all three LLM baselines (GPT-4o, Llama-3.1-8B-Instruct, and Qwen2.5-7B-Instruct).

Zero-shot prompt.

あなたはYouTubeコメントの建設性を判定する評価者です。

建設的なコメントとは、次の2つの条件を両方満たすコメントです。

1．前提条件： 記事・動画の内容に関係しており、誹謗中傷を含まない。

2．付加的情報： 動画で被覆されていないが、内容に関連する有益な情報を含む。

–

以下のコメントが「建設的なコメント」かどうかを判定してください。

コメント：

{comment}

回答は必ず Yes または No のみで出力してください。説明は不要です。

English translation.

You are an evaluator who judges the constructiveness of YouTube comments.

A constructive comment is a comment that satisfies both of the following two conditions.

1. Premise condition: It is related to the content of the article or video and does not contain slander or abusive language.

2. Additional information: It contains useful information that is not covered in the video but is relevant to its content.

–

Please determine whether the following comment is a “constructive comment.”

Comment:

{comment}

Output only Yes or No. Do not provide any explanation.

Few-shot prompt.

あなたはYouTubeコメントの建設性を判定する評価者です。

建設的なコメントとは、次の2つの条件を両方満たすコメントです。

1．前提条件： 記事・動画の内容に関係しており、誹謗中傷を含まない。

2．付加的情報： 動画で被覆されていないが、内容に関連する有益な情報を含む。

例：

コメント：動画中の先生は馬鹿なの？ 回答：No

コメント：最新のモデルを使うともっと簡単に解けますよ。 回答：Yes

コメント：すごい！その方法は思いつかなかった！ 回答：No

–

以下のコメントが「建設的なコメント」かどうかを判定してください。

コメント：

{comment}

回答は必ず Yes または No のみで出力してください。説明は不要です。

English translation.

You are an evaluator who judges the constructiveness of YouTube comments.

A constructive comment is a comment that satisfies both of the following two conditions.

1. Premise condition: It is related to the content of the article or video and does not contain slander or abusive language.

2. Additional information: It contains useful information that is not covered in the video but is relevant to its content.

Examples:

Comment: Is the teacher in the video stupid? Answer: No

Comment: This can be solved more easily by using the latest model. Answer: Yes

Comment: Amazing! I never would have thought of that method! Answer: No

–

Please determine whether the following comment is a “constructive comment.”

Comment:

{comment}

Output only Yes or No. Do not provide any explanation.

Zero-shot CoT prompt.

あなたはYouTubeコメントの建設性を判定する評価者です。

建設的なコメントとは、次の2つの条件を両方満たすコメントです。

1．前提条件： 記事・動画の内容に関係しており、誹謗中傷を含まない。

2．付加的情報： 動画で被覆されていないが、内容に関連する有益な情報を含む。

–

以下のコメントが「建設的なコメント」かどうかを判定してください。

コメント：

{comment}

まず前提条件と付加的情報の基準に照らして判断理由を1〜2文で述べてから、最後の行に必ず Yes または No のみを出力してください。

English translation.

You are an evaluator who judges the constructiveness of YouTube comments.

A constructive comment is a comment that satisfies both of the following two conditions.

1. Premise condition: It is related to the content of the article or video and does not contain slander or abusive language.

2. Additional information: It contains useful information that is not covered in the video but is relevant to its content.

–

Please determine whether the following comment is a “constructive comment.”

Comment:

{comment}

First, state your reasoning in one or two sentences with reference to the premise and additional-information criteria, then output only Yes or No on the final line.

Few-shot CoT prompt.

あなたはYouTubeコメントの建設性を判定する評価者です。

建設的なコメントとは、次の2つの条件を両方満たすコメントです。

1．前提条件： 記事・動画の内容に関係しており、誹謗中傷を含まない。

2．付加的情報： 動画で被覆されていないが、内容に関連する有益な情報を含む。

例：

コメント：動画中の先生は馬鹿なの？ 理由：動画に関係しているが誹謗中傷を含むため前提条件を満たさない。 回答：No

コメント：最新のモデルを使うともっと簡単に解けますよ。 理由：動画に関係しており、動画で扱われていない有益な情報を提供している。 回答：Yes

コメント：すごい！その方法は思いつかなかった！ 理由：動画に関係しているが、新たな情報を提供していない。 回答：No

–

以下のコメントが「建設的なコメント」かどうかを判定してください。

コメント：

{comment}

まず前提条件と付加的情報の基準に照らして判断理由を1〜2文で述べてから、最後の行に必ず Yes または No のみを出力してください。

English translation.

You are an evaluator who judges the constructiveness of YouTube comments.

A constructive comment is a comment that satisfies both of the following two conditions.

1. Premise condition: It is related to the content of the article or video and does not contain slander or abusive language.

2. Additional information: It contains useful information that is not covered in the video but is relevant to its content.

Examples:

Comment: Is the teacher in the video stupid? Reason: It is related to the video but contains abuse, so it does not satisfy the premise condition. Answer: No

Comment: This can be solved more easily by using the latest model. Reason: It is related to the video and provides useful information not covered in the video. Answer: Yes

Comment: Amazing! I never would have thought of that method! Reason: It is related to the video but provides no new information. Answer: No

–

Please determine whether the following comment is a “constructive comment.”

Comment:

{comment}

First, state your reasoning in one or two sentences with reference to the premise and additional-information criteria, then output only Yes or No on the final line.

C Demo System and User Study

System description. The demo system is implemented as a Chrome extension that operates directly on YouTube Shorts pages. Upon activation, it retrieves top-level comments via the YouTube Data API, scores each comment using Student$_3$, and re-ranks them in descending order of construc-

tiveness score. Users can also filter out comments below a threshold of $p \geq 0.5$. Figure 2 shows the comment section before and after applying the extension.

User study. To evaluate the practical utility of the system, we conducted a small user study with 8 participants, all of whom interacted with the demo on YouTube Shorts videos of their choice. Participants completed a questionnaire assessing their experience before and after using the demo.

Table 4 summarizes the results. After using the demo, 6 out of 8 participants felt that constructive comments were clearly surfaced to the top. On a 5-point Likert scale, participants rated the informativeness of top-ranked comments at 3.75 on average, and the usefulness of the scores for finding constructive comments at 4.13 on average.

Metric	Score
Comments surfaced clearly	6 / 8
Top comments informative	3.75 / 5
Scores useful for discovery	4.13 / 5

Table 4: User study results ($n = 8$).

	Comment	Label
1	出荷元がアマゾンじゃないと「全く別次元の商品」という認識が早い時期からあるのでこれは間違わないな。 *English:* If the seller is not Amazon, there has long been an understanding that it may be a completely different kind of product, so I would not make this mistake.	Constructive
2	この三角形2個置いて長方形で考えても存在しないことが言えますね *English:* You can also show that it does not exist by placing two of these triangles and considering them as a rectangle.	Constructive
3	日本語って、本当に難しいよね *English:* Japanese is really difficult, isn't it?	Non-constructive
4	悪質な手口って、この動画のことじゃん *English:* This video itself is the malicious trick, isn't it?	Non-constructive

Table 5: Examples of constructive and non-constructive comments from the dataset.

Error analysis. Participants also reported cases where the system produced errors. False positives (non-constructive comments ranked highly) included reactions to internet memes, comments describing personal behavior unrelated to the video topic, and anecdotal comparisons loosely related to the video content. One participant noted that comments relevant to the video but containing only simple reactions were sometimes retained after filtering, suggesting that the premise criterion (relevance) alone is insufficient to ensure informativeness.

Comparison	Δ PR-AUC	95% CI
AL2 − Teacher	+0.0832	[0.0339, 0.1523]
AL2 − ST1	+0.0242	[−0.0050, 0.0641]
AL2 − ST2	+0.0202	[−0.0037, 0.0570]
AL2 − AL1	+0.0168	[−0.0017, 0.0412]
AL2 − GPT-4o (0)	+0.0472	[−0.0666, 0.1521]
AL2 − GPT-4o (FS)	+0.0318	[−0.0817, 0.1305]

Table 6: Paired video-level bootstrap comparison of PR-AUC with 5,000 bootstrap samples. AL, ST, and FS denote active learning, self-training, and few-shot prompting, respectively. Positive values indicate higher PR-AUC for AL2.

D Annotation Guidelines

Premise criterion.
前提条件：記事に関係しており，誹謗中傷を含まない。

English translation.
Premise condition: It is related to the content of the article or video and does not contain slander or abusive language.

Additional information criterion.
付加的情報：動画で被覆されてないが，内容に関連する有益な情報を含む。

English translation.
Additional information: It contains useful information that is not covered in the video but is relevant to its content.

Constructiveness label. A comment is labeled *constructive* if and only if both criteria are satisfied. Table 5 shows representative examples from the dataset.

E Additional Analyses

E.1 Bootstrap Confidence Intervals

Table 6 reports the paired video-level bootstrap comparisons used to assess the stability of the PR-AUC differences. Only the comparison between AL round 2 and the Teacher has a 95% confidence interval excluding zero. The remaining comparisons are not statistically significant on the current test set.

Setting	τ_{neg}	τ_{pos}	PR-AUC	95% CI
Original	0.05	0.98	0.677	[0.590, 0.751]
Loose	0.10	0.90	0.678	[0.590, 0.752]
Strict	0.02	0.99	0.679	[0.592, 0.753]

Table 7: Sensitivity of self-training performance to pseudo-label confidence thresholds. Negative and positive pseudo-labels are selected when $p < \tau_{\text{neg}}$ and $p > \tau_{\text{pos}}$, respectively.

Acquisition strategy	PR-AUC	ROC-AUC
Uncertainty sampling	0.6838	0.8787
Random sampling	0.6906	0.8777

Table 8: Comparison of uncertainty-based and random acquisition under the same annotation budget.

E.2 Pseudo-Label Threshold Sensitivity

Table 7 reports the sensitivity of self-training performance to the pseudo-label confidence thresholds. Performance is highly stable across the tested threshold settings, with a maximum PR-AUC difference of 0.002. This suggests that the original thresholds act as conservative pseudo-label selection criteria rather than finely tuned hyperparameters.

E.3 Uncertainty Sampling vs. Random Sampling

Table 8 compares uncertainty-based and random acquisition for the first human-annotation round under the same annotation budget.

The 95% confidence interval for the difference (uncertainty minus random) is $[-0.0300, 0.0282]$ for PR-AUC and $[-0.0082, 0.0116]$ for ROC-AUC. Thus, the two acquisition strategies are statistically indistinguishable at the present annotation scale. The random batch also contained a higher proportion of constructive comments, so acquisition strategy and batch composition cannot be separated conclusively from this single comparison.

E.4 Annotator Disagreement Analysis

Table 9 summarizes criterion-level disagreement across all 5,515 annotator pairs, breaking down how often the two annotators diverged on the premise criterion, the additional-information criterion, or both. These pair-level statistics are computed over annotator pairs and therefore use a different denominator from the final-label analysis reported in Section 2, which is computed over comments. To complement these aggregate statistics, we below present representative comments illustrating each type of disagreement. For each example, A1 and A2 denote the two annotators who labeled the comment, and each annotator's two decisions correspond to the premise and additional-information criteria, respectively.

Pair-level outcome	Proportion
No disagreement	62.5%
Premise only	17.8%
Additional information only	17.1%
Both criteria	2.6%

Table 9: Criterion-level outcomes across 5,515 annotator pairs.

Additional-information disagreement

> 新聞紙はジーンズで軽くこするとかでもいいぞ
> *(You could also try lightly rubbing the newspaper with jeans.)*
> A1: Yes/Yes A2: Yes/No

Both annotators agree the comment is relevant and non-abusive, but differ on whether it adds useful information beyond the video.

Premise disagreement

> 弱い2人が話してもなんの説得力もねーな
> *(Two weak people talking has no persuasiveness at all.)*
> A1: Yes/Yes A2: No/No

Here the annotators differ on the premise itself: whether a dismissive remark still counts as a relevant evaluation of the video content.

E.5 Evaluation with GPT-5.4

We additionally conducted exploratory experiments with GPT-5.4 under the same prompting setup used for GPT-4o.

- GPT-5.4 (0-shot): ROC-AUC = 0.728, PR-AUC = 0.539
- GPT-5.4 (few-shot): ROC-AUC = 0.820, PR-AUC = 0.654

GPT-5.4 (few-shot) did not outperform Student_3 on either metric (PR-AUC 0.654 vs. 0.701; ROC-AUC 0.820 vs. 0.883). These results should be interpreted with caution, however, as API constraints limited log-probability access to the top-5 tokens, which may affect the reliability of AUC estimation.

E.6 Popularity vs. Constructiveness: Examples

The following pairs of comments are drawn from the same video. In each pair, the non-constructive comment receives substantially more likes than the constructive one, illustrating that popularity is not a reliable proxy for constructiveness.

Pair 1 (cooking video)

Non-constructive (24,947 likes):
いや、豚肉は普通に柔らかいだろ。って思ってみたら100倍柔らかくなってて草生えた
(I thought pork was already tender enough, but then I saw it was 100x softer lol)

Constructive (1 like):
たまにトンカツが獣臭して、揚げる温度が低いって教えてもらったんだけど、じっくり低温で焼くのは問題ないの？料理下手でよくわからん。
(Sometimes my tonkatsu smells gamey — I was told it's because the oil temp is too low. Is slow low-temp cooking okay? I'm not good at cooking.)

Pair 2 (trivia video)

Non-constructive (3,242 likes):
「貧弱な貴方はこのようにドゥーンﾇｯ」が好きすぎる
(I love the line "Weakling, take this!")

Constructive (0 likes):
世界で一番深い穴がロシア・コラ半島にあるSG3で科学掘削された穴。その深さはエベレストと富士山を足した長さであり、マリアナ海溝より深く12261メートルにも達する
(The world's deepest hole is the SG-3 borehole on the Kola Peninsula, Russia. At 12,261m, it is deeper than the Mariana Trench and equals the combined height of Everest and Mt. Fuji.)

Pair 3 (health video)

Non-constructive (433 likes):
目こんなんですけど...って言ってるけどガンギマリでワロタ
(She says "my eyes look like this..." but she's clearly high lol)

Constructive (1 like):
天然のカフェインは摂りすぎると身体に有害ですよ、気を付けて
(Natural caffeine can be harmful if consumed in excess — please be careful.)

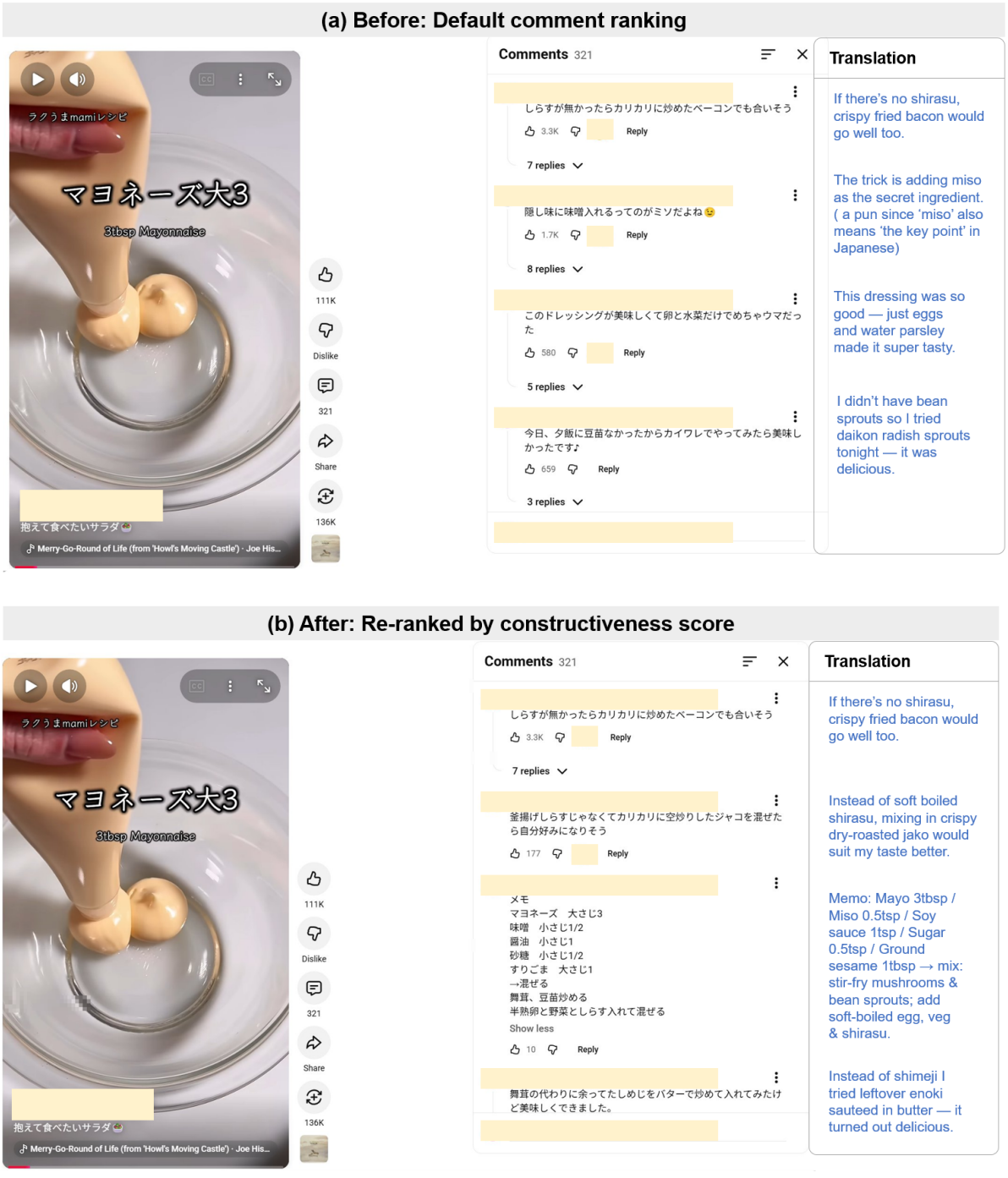


Figure 2: YouTube Shorts comment section before (top) and after (bottom) applying the constructive comment extraction system. Screenshots are taken from the following public YouTube Shorts video: https://www.youtube.com/shorts/lSetTkPOdNg. Comments are re-ranked by constructiveness score; usernames and avatars are anonymized. English translations of the Japanese comments are provided in each screenshot.