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

Anonymous ACL submission

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, validated through a user study. To train under noisy annotation conditions, we combine high-confidence data selection with self-training and active learning. Experiments show that iterative data expansion improves performance, while general-purpose LLMs underperform our task-specific model despite their scale. These findings demonstrate the value of dataset construction and targeted training for constructive comment extraction in realistic settings.

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). Since constructive comments provide useful, relevant, and non-toxic information beyond the original content, identifying them 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.

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).

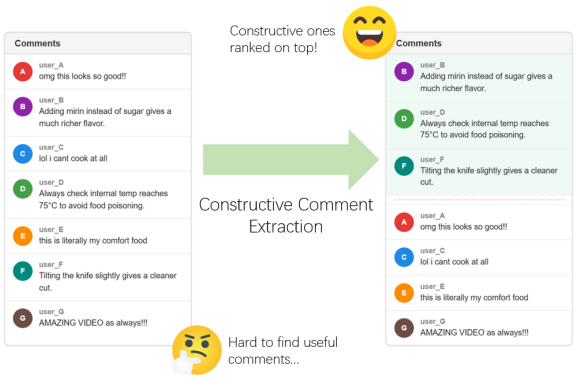

Figure 1: Constructive comments (highlighted in green) are ranked at the top, making them easier to find among low-effort reactions.

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-annotator disagreement. To train a robust classifier under noisy annotation conditions, we combine high-confidence data selection with self-training 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.

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, selected through manual recommendations by project participants to ensure diversity across 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 from 160 new videos, yielding a final released dataset of approximately 4,600 comments from 250 videos. Video-level splits are used throughout to prevent information leakage across training, validation, and test sets.

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 31.5 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.2.

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.

2.3 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. All comments are annotated by 14 annotators, all either native Japanese speakers or JLPT N1 certified. Annotators label each comment independently, drawing on their knowledge of the corresponding video content.

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).

2.4 High-Confidence Training Set

We construct a high-confidence subset by retaining only comments with unanimous agreement among at least two annotators, yielding 2,873 instances (361 constructive, 2,512 non-constructive, ratio $\approx$ 1:7). For remaining disputed comments, a third annotator provides a tie-breaking judgment, yielding a full gold dataset of 3,797 comments (792 constructive, 3,005 non-constructive), split by video into training (2,576), validation (561), and test (660) sets. Comments added through active learn-

ing 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 scarcity of high-confidence labeled data, tackled via self-training on unlabeled comments; 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 high-confidence subset, 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 ($p \approx 0.5$), 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 class imbalance (constructive:non-constructive $\approx$ 1:7), we report both ROC-AUC and PR-AUC, with emphasis on

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

Model	ROC-AUC	PR-AUC
GPT-4o (0-shot)	0.835	0.653
GPT-4o (few-shot)	0.856	0.669
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. LLM ROC-AUC and PR-AUC are computed from token-level output probabilities.

PR-AUC as it better reflects performance on the minority class (Saito and Rehmsmeier, 2015).

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.

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). In contrast, open-source models perform substantially worse (PR-AUC 0.339–0.406), suggesting that strong constructiveness judgment requires either large-scale pretraining or task-specific fine-tuning. Notably, our final model Student_3 (PR-AUC 0.701) outperforms GPT-4o (few-shot) while using a substantially smaller encoder (130M parameters), demonstrating the effectiveness of our training pipeline.

Our trained pipeline shows consistent improvements at each stage, with the Teacher already providing a strong baseline (PR-AUC 0.618). Self-training yields the largest gain in the first round (+0.059 PR-AUC), while the second round provides only marginal improvement (+0.002). Active learning further improves performance after self-training saturates, with gains of +0.003 and +0.019 PR-AUC in rounds 1 and 2.

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.

Active learning improves on ambiguous cases. Active learning is applied from Student_1 rather than Student_2, as the second self-training round yields negligible improvement. AL provides consistent gains (+0.005 and +0.019 PR-AUC in rounds 1 and 2), complementing self-training by targeting uncertain boundary cases that pseudo-labeling cannot address.

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 high-confidence data selection with self-training and active learning. Experiments show that self-training effectively expands the training signal but saturates quickly, while active learning complements it by targeting ambiguous boundary cases. 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 approximately 4,600 comments from 250 videos, its scale remains modest, and future work should validate findings on larger and more diverse corpora. The demo system currently supports Japanese only. Although we conducted a user study via questionnaire (Appendix C) to evaluate its practical utility, the study was conducted at a limited scale and further evaluation with a larger and more diverse user group remains future work.

High-confidence filtering removes a substantial portion of ambiguous instances, potentially biasing the training data toward clear-cut cases. Self-training relies on high-confidence pseudo-labels and may reinforce existing model biases. Active learning addresses uncertain cases but is validated at limited scale here; broader evaluation across annotation settings remains future work.

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. The released dataset will contain only comment IDs and anonymized annotations; comment text will not be distributed directly in order to comply with platform terms of service.

Potentially harmful content. The annotation guidelines define constructive comments as comments that do not contain abusive language (Section 2.3). 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.

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 constructiveness 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 informa-

tiveness 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.

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 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), suggesting that targeted training on platform-specific data remains important even for frontier LLMs. 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.2 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.)

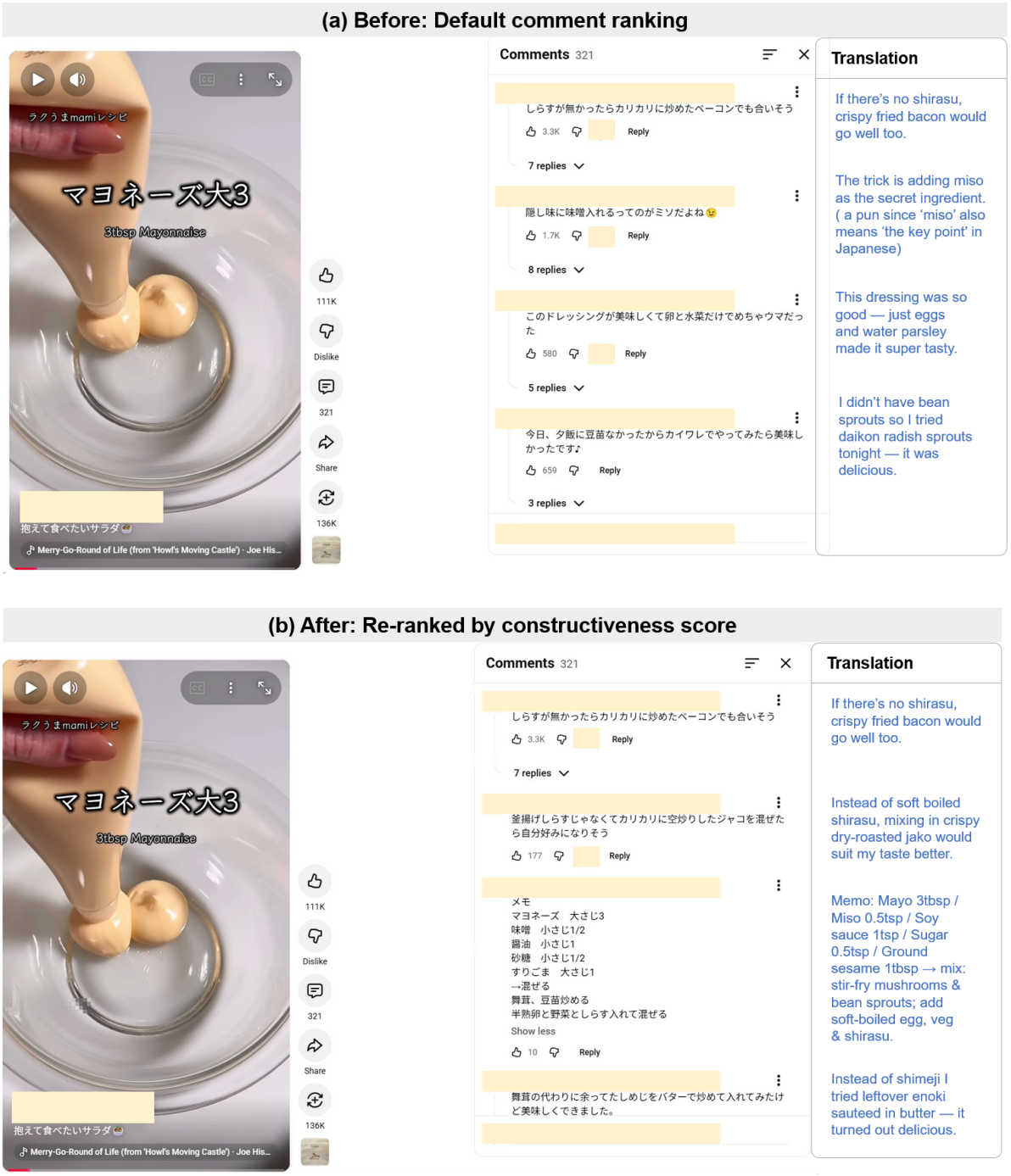

Figure 2: YouTube Shorts comment section before (top) and after (bottom) applying the constructive comment extraction system. Screenshots are taken from the following public YouTube Shorts video: https://www.youtube.com/shorts/lSetTkPOdNg. Comments are re-ranked by constructiveness score; usernames and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