

Towards Cost-Efficient Video Dataset Construction with Large Language Models

Shabot Mohammed¹[0009-0009-5336-1970], Jonathan Adams²[0009-0005-2848-438X], Jan Van den Stock²[0000-0001-5756-3195], and Joost Vennekens¹[0000-0002-0791-0176]

¹ AI Lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
 mohammed.shabot@vub.be, joost.vennekens@vub.be

² Neuropsychiatry, KU Leuven, ON5/B Herestraat 49 - box 1029, 3000 Leuven, Belgium
 jonathan.adams@kuleuven.be, jan.vandenstock@kuleuven.be

Abstract. Research on video classification models has witnessed significant growth recently, accompanied by a significant increase in the demand for datasets to train such models. Existing video dataset curation approaches share a common bottleneck: candidate videos must be downloaded before any relevance judgment is made, which results in storage, time and annotation costs. In this work, we present a filtering step, using Large Language Models (LLMs) to pre-screen videos based on their metadata, before actually downloading these videos, thereby saving on these costs. We evaluate the LLMs against a strict human majority vote derived from four independent annotators and validate the reliability of the annotations through inter-annotator agreement analysis (Gwet’s AC1). We compare the LLM pre-filter against two no-LLM baselines, namely two keyword heuristics. Via a Leave-One-Rater-Out (LOO) mechanism and an Add-One-Rater-In (AORI), we evaluate the performance of the LLMs as virtual annotators. Finally, we report cost comparisons relative to existing approaches. To ensure reproducibility, we made our implementation available on github.³

Keywords: Large Language Models · Video dataset curation · Metadata-based filtering · Dataset pre-screening · Automatic data filtering

1 Introduction

Deep learning methods are now state-of-the-art for many machine learning tasks. One of the main challenges in deep learning lies in the (re-)training of models and the large datasets that are needed for this. This is particularly the case for tasks involving video data, such as action recognition, motion detection or visual question answering. There is a vast amount of video data available on the internet. Some authors claim that approximately 24 hours of video content is uploaded to YouTube every minute [11]. The problem is therefore not the availability of

³ <https://github.com/MoShabot-PhD/llm-paper>

the data, but rather the process of constructing a relevant, high-quality dataset. A common approach to collect video data is to download all candidate videos and manually annotate them [16], possibly with some basic filtering steps such as keyword matching on titles from the YouTube corpus [10]. However, this requires downloading large volumes of content before any relevance judgment can be made, resulting in a significant loss of computational resources, storage and time. A more recent approach in constructing datasets is using state-of-the-art video generation models. These models allow researchers to construct automated video generation pipelines, in which customizable and detailed prompts can be used to produce video datasets at scale [15]. This eliminates the need for manual annotation and data scraping, and enables experimental control and customizability of the videos. This process, however, also comes with its own downsides, such as long processing times, video generation costs and potential video model biases being incorporated in the dataset. Moreover, for complex cognitive tasks, such as social cognition, it is not clear if human annotators always interpret generated videos in the same way as natural videos.

In this paper, we propose a novel method of constructing video datasets, by using LLMs to pre-screen the metadata of videos, before downloading them. We show the effectiveness of this method by comparing state-of-the-art LLMs, namely (at the time of writing) GPT-5.2, Claude Sonnet 4.6 and open-source local models: QWEN3.6-36B and Gemma4:31B, against manual annotations at both metadata and video level. We additionally compare the LLMs against two commonly used, no-LLM baselines, to establish whether the LLM pre-filter earns its cost relative to methods a researcher could build for free. We assess the resulting cost reduction relative to traditional pipelines and synthetic video generation pipelines. Specifically, this work answers the following research questions:

- RQ1:** Can LLMs be integrated into a scraping pipeline as a high-recall filter on metadata?
- RQ2:** Is integrating LLMs into a scraping pipeline really cost effective?
- RQ3:** Can LLMs substitute for human annotators, and does this depend on the annotator’s level of expertise?

2 Related Work

In this section, we review a number of existing video datasets, focusing on their construction pipelines and the bottlenecks they share. Early datasets, such as the UCF101 dataset [16] relied on downloading candidate videos and manually removing irrelevant ones, which was common practice at the time. The Human Motion Database was built using a similar approach, scraping videos from several sources (YouTube, Prelinger archive, ...) and assigning students to annotate those manually [11]. The construction of the Kinetics dataset started from a list of actions gathered from existing datasets, such as ActivityNet [8], ACT [18], MPII Human Pose [1] and the datasets mentioned earlier. Then, videos were chosen by keyword matching the entries in the action list with titles in the YouTube corpus and annotating them by crowd-sourcing the annotation work to Amazon

Mechanical Turkers [10]. For the construction of the Something-Something V2 dataset, researchers hired crowd workers to record videos themselves, given a set of predefined labels. The videos were then manually reviewed by Amazon Mechanical Turkers [4]. The dataset that most closely aligns with our work, is the recent instruction-following VideoMarathon dataset [12]. The authors collected videos from several publicly available datasets, that covered a diverse range of topics. The videos were annotated by an automated annotation pipeline using QWEN2VL-7B and DeepSeek-v3 [17, 2, 12], resulting in a large, high-quality dataset. This study shows the capabilities of state-of-the-art (open-source, local) video processing models. Across all these approaches, a common bottleneck remains: videos must be downloaded before any relevance judgment can be made.

The emerging topic of synthetic data generation avoids this bottleneck entirely, by leveraging the improvement of video generation models. Recent work establishes an automated pipeline to generate a synthetic video dataset, based on starting images and an LLM-generated description of the possible future with respect to the starting frame [15]. While this avoids the bottleneck, the dataset will inherit limitations from the generator itself.

Both [5] and [7] have investigated the use of LLMs as virtual annotators, with [5] comparing techniques such as retrieval-augmented generation (RAG) against in-context learning (ICL), and [7] evaluating LLMs within a full event-annotation workflow. However, these studies primarily evaluate LLMs in relatively controlled annotation settings. In contrast, our work integrates LLMs into a more complex video dataset construction pipeline, where they are used to assess the relevance of candidate videos as part of a multi-stage process involving video retrieval, filtering, and processing. This setting introduces additional challenges, as the LLM must make relevance judgments based on noisy video metadata rather than well-defined annotation inputs. Our work, therefore, investigates not only the annotation capabilities of LLMs, but also their ability to function as an effective filtering component within a more complex, large-scale dataset construction pipeline.

3 Method for dataset construction

In this work, we propose a novel way of constructing video datasets. Instead of downloading all candidate videos before any relevance judgment is made, we propose to have an LLM filter videos based on their metadata alone, avoiding unnecessary downloads, loss of time and downstream annotation costs or video generation costs. The LLM is tasked with deciding whether a video should be labeled *KEEP* or *DISCARD*, based on the metadata of the video, given explicit instructions through a prompt (Appendix A). Among the available metadata fields, *title* and *description* are passed to the model. To scale this approach, a batch API is used, allowing multiple asynchronous requests to be submitted simultaneously at a cost reduction of 50% relative to individual API calls. In addition, the model is prompted to return a confidence score alongside each label. This confidence score is converted into a keep score, equal to the confidence itself

if the decision is KEEP, or one minus the confidence if the decision is DISCARD, and the final decision is labeled KEEP only if this keep score exceeds a threshold t . The parameter t is selected as the F1-maximizing operating point within a single-level GroupKFold procedure (no model is trained, so no inner cross-validation is required): for each outer fold, t is determined on the remaining search queries and then applied to the held-out fold, so that no item is evaluated under a threshold tuned on itself. On new dataset domains, t could be initialized to 0.5 and tuned when a small subset of labeled data is available. Appendix F shows the (un)tuned parameter t and the performance differences.

4 Concrete use case

The concrete use case we use in our work is the extension of the FindingEmo image dataset [13] with video data. FindingEmo is an image dataset containing approximately 25,000 images of complex scenes, where multiple people are interacting in various settings. Images are annotated on several dimensions, such as valence, arousal and emotion label. Annotating this dataset is a demanding task, as there is no objective correct answer: annotators have to interpret a social scene, assess multiple people at once and commit to a judgment on a continuous scale (valence, arousal) and a discrete emotion scale. Extending the image dataset with videos adds other layers of complexity. Video data must meet a number of criteria: the clips should show clear social interaction, contain no on-screen text, and not be documentaries or music videos, just to name a few.

To extend the dataset with video material, we are developing a multi-stage filtering pipeline in which an LLM serves as a pre-filter on metadata. This pre-filtering step is the focus of this present work.

5 Experimental setup

To evaluate the performance of the LLMs, we set up a human annotation panel. The *baseline panel* consists of four independent human annotators with varying levels of expertise: two domain experts and two non-experts who annotated solely on the basis of written instructions as used in the LLM prompt, shown in Appendix A. Here, the non-experts serve as a proxy for the crowd workers one would recruit on platforms such as Prolific, who rely on written instructions rather than domain knowledge. Comparing the LLMs against the non-experts addresses whether the LLMs can substitute for crowdsourced annotations. The comparison against the experts shows whether the LLMs can approximate expert judgment. For each LLM, four *mixed panels* are constructed by substituting each human annotator in turn with the LLM. Then, for each panel, we also add the LLM as a fifth annotator and assess whether the panel-agreement significantly degrades or not.

For each panel, we calculate inter-annotator agreement using Gwet’s AC1 [6, 3]. Panel-level AC1 is defined as the mean of the $\binom{4}{2} = 6$ pairwise AC1 scores.

For a rater pair (a, b) , pairwise AC1 is computed as

$$\begin{aligned}\bar{p}_{\text{keep}} &= \frac{1}{2}(\Pr(a=\text{KEEP}) + \Pr(b=\text{KEEP})), \\ P_e &= 2\bar{p}_{\text{keep}}(1 - \bar{p}_{\text{keep}}),\end{aligned}\tag{1}$$

$$\text{AC1} = \frac{P_o - P_e}{1 - P_e},\tag{2}$$

where $P_o = \Pr(a = b)$ is the observed agreement. Equation 1 depicts the probability that two annotators agree by chance ⁴. The reported Δ_{LOO} is the difference between the AC1 of the original, human-only panel and a panel in which one of the humans has been replaced by an LLM. Confidence intervals are obtained via cluster-bootstrap resampling at the search-query level (1000 iterations, 5th–95th percentile). A Δ_{LOO} is flagged as a significant degradation ($\downarrow$) when the 90 % CI lies entirely below zero; otherwise it is treated as non-significant ($\sim$).

To scrape videos and their metadata, we set up predefined lists of emotions, social settings and age groups, to construct our search queries, just like the authors of the FindingEmo dataset did [13]. For each search query, only the first 10 results were considered, as these are the most likely candidates a researcher would encounter in practice, given that YouTube’s ranking algorithm prioritizes relevance [9]. From those 10 videos, we scraped the following seven metadata fields: *title*, *description*, *duration*, *url*, *view_count*, *uploader* and *uploader_date*, but ended up only passing title and description to the LLMs, as these contain semantically valuable information for the final decision. In-depth details on the metadata are shown in Appendix D.

Annotation procedure and inter-annotator agreement All items were annotated by four human annotators, each labeling the metadata with KEEP or DISCARD labels. The ground truth label was determined by a strict majority vote: an item is labeled KEEP when at least three of the four annotators chose KEEP, so 2–2 ties (167 items) are labeled DISCARD. Inter-annotator agreement among the four human annotators was calculated using Gwet’s AC1 [6]. We chose Gwet’s AC1 as a measurement, over the well known Fleiss κ to mitigate the kappa paradox which occurs with imbalanced class distributions [3]. Human AC1 is reported as a validity check.

We evaluate on the 1047 items for which all four annotators provided a metadata label and all four LLMs returned a usable decision and confidence score. In addition, 298 items were annotated on video-level.

Baselines To assess the LLMs’ performance, we compare all models against two, no-LLM baselines. **Regex A** applies a stop-list of patterns matching the exclusion criteria stated in the prompt, such as music videos or dance performances. An item is kept unless an entry from the stop-list matches. This mirrors

⁴ Equation (1): binary case of Gwet’s chance-agreement term $P_e = \frac{1}{K-1} \sum_{k=1}^K \pi_k(1 - \pi_k)$. For $K = 2$ we have $\pi_2 = 1 - \pi_1$, which reduces to $2\bar{p}_{\text{KEEP}}(1 - \bar{p}_{\text{KEEP}})$.

the high-recall instruction given to the LLMs and is the closest to the keyword matching used in the construction of the Kinetics dataset [10]. **Regex B** requires an emotionally loaded keyword to be present (examples are: "wedding", "crying" or "laughter")

Statistical comparison The differences between the methods are assessed with a paired cluster bootstrap: search queries are resampled with replacement over 1000 iterations and the difference in the metric of interest is computed within each resample. We report the 2.5th–97.5th percentile of the resulting difference distribution and treat a difference as significant when the interval excludes zero. We compare the LLMs against the two no-LLM baselines.

Leave-One-Rater-Out (LOO) and Add-One-Rater-In (AORI) To study the performance of the different LLMs, we chose to 1) replace each annotator, one at a time, with the annotations of the LLM to determine whether this results in significant differences and 2) add the LLM as a 5th annotator. This allows us to determine whether adding an LLM as an annotator degrades or improves the overall performance of the panel.

Query setup We reused the scraping queries from the FindingEmo paper [13], which consists of lists of emotions, social settings and age groups, and constructed search queries as the Cartesian products of these lists.

The local models (QWEN3.6-36B and Gemma4:31B) were deployed on an Apple M4 Max with 48 GB RAM, using Ollama with Q4_K_M quantization for both models. The API models were queried with temperature 1, the local models with temperature 0.3, top-p 0.9 and a fixed seed.

6 Results

Pairwise Human AC1 Before looking at the performances of the LLMs, we take a look at the pairwise human AC1s, where A and B are domain experts, and C and D are the non-experts. The results are shown in Table 1. These results will be used to compare the performances and to answer RQ3. Notice that there already is discrepancy between the different annotators, as the AC1 ranges from 0.507 (moderate agreement) to 0.694 (substantial agreement), due to the ambiguity of the task. Notice that the agreement between the two domain experts is lower than the agreement between one domain expert and a non-expert, confirming the complexity of the task.

Table 2 shows the precision, recall, and F1 metrics of each model at the metadata level, along with the 95% confidence intervals computed via cluster resampling over 1000 iterations and shows the no-LLM baselines in the lower block. We opted for cluster resampling because the 10 results within each search query are not independent observations as they share a common retrieval bias imposed by YouTube’s ranking algorithm.

Table 1: Pairwise human AC1, metadata panel ($n = 1047$).

Annotator pair	A – D	A – C	A – B	C – D	B – D	B – C
AC1	0.694	0.652	0.611	0.550	0.529	0.507

 Table 2: Metadata-level classification performance (KEEP as positive class), $n = 1047$, 95% cluster-bootstrap CIs. Lower block: no-LLM baselines.

Model	Prec.	95% CI	Recall	95% CI	F1	95% CI
QWEN3.6-36B	0.354	[0.271, 0.434]	0.565	[0.480, 0.645]	0.435	[0.357, 0.504]
Gemma4:31B	0.261	[0.193, 0.331]	0.857	[0.791, 0.917]	0.401	[0.314, 0.478]
GPT-5.2	0.285	[0.198, 0.373]	0.526	[0.421, 0.621]	0.370	[0.276, 0.450]
Claude Sonnet 4.6	0.286	[0.211, 0.366]	0.604	[0.506, 0.695]	0.388	[0.307, 0.464]
Regex B (stoplist+keyword)	0.269	[0.183, 0.353]	0.390	[0.286, 0.492]	0.318	[0.231, 0.401]
Regex A (stoplist)	0.181	[0.127, 0.233]	0.708	[0.624, 0.793]	0.288	[0.215, 0.353]

Point estimates suggest that QWEN3.6-36B is the most balanced model out of all four models, achieving an F1 score of 0.435 and a Recall score of 0.565. Given that this is a pre-filter, we argue that recall is of importance, which makes Gemma4:31B the most favorable model for the task.

Of the 298 items, there is an 83.6% agreement between the metadata labels and the video labels. The 16.4% disagreement is asymmetric: 31 items labeled KEEP at metadata level were flipped to DISCARD after watching the video, while only 18 flipped the other way. The results for the video level annotations are shown in Figure 1.

LOO and AORI We assessed the performances of the LLMs by replacing each annotator, one at a time, by the LLM (LOO) and adding each model, one at a time, as a 5th annotator to the panel. Table 3 shows the results for the LOO-study and Table 4 shows the results for the AORI-study.

When examining the results, it becomes clear that the substitution of any annotator with an LLM results in a significant degradation of the four-human annotator panel. The degradation which occurs by using QWEN3.6-36B as a substitute is minimal, meaning that QWEN3.6-36B aligns the most with each member of the panel. The same outcome is shown in Table 4, where the change in panel AC1 decreases the least (-0.0525) when using QWEN3.6-36B as an extra annotator in the annotator panel.

To answer RQ2, we calculated the costs using the Batch API, which provides a 50% discount on the API prices, resulting in \$0.875 per 1M input tokens and \$7.00 per 1M output tokens for GPT-5.2, and \$1.50 per 1M input tokens and \$7.50 per 1M output tokens for Claude Sonnet 4.6. GPT-5.2 used 1,583,031 input tokens and 162,484 output tokens, resulting in a total of 1,745,515 total used tokens, which amounts to $\approx$ \$2.52, whereas Claude Sonnet 4.6 used 1,807,021 input tokens and 132,212 output tokens, resulting in a total token amount of

Table 3: Leave-one-rater-out, metadata level ($n = 1047$): change in panel AC1 (Δ_{LOO} , 90% CI) when each human annotator is replaced by an LLM. Reference (4-human) panel AC1 = 0.59. Every substitution is a significant degradation (90% CI entirely below zero).

Model	Rater	Δ_{LOO}	90% CI
GPT-5.2	A	-0.2973	[-0.3390, -0.2552]
	B	-0.2723	[-0.3142, -0.2290]
	C	-0.2311	[-0.2814, -0.1823]
	D	-0.2742	[-0.3177, -0.2314]
Claude Sonnet 4.6	A	-0.3304	[-0.3745, -0.2867]
	B	-0.3036	[-0.3498, -0.2626]
	C	-0.2665	[-0.3195, -0.2132]
	D	-0.3107	[-0.3581, -0.2655]
QWEN3.6-36B	A	-0.1086	[-0.1291, -0.0876]
	B	-0.0438	[-0.0709, -0.0186]
	C	-0.0357	[-0.0651, -0.0065]
	D	-0.0745	[-0.1021, -0.0466]
Gemma4:31B	A	-0.2422	[-0.2778, -0.2054]
	B	-0.1999	[-0.2382, -0.1628]
	C	-0.1679	[-0.2126, -0.1209]
	D	-0.2195	[-0.2578, -0.1795]

Table 4: Add-one-rater-in (AORI), metadata level ($n = 1047$): change in panel AC1 (Δ_{AORI} , 90% CI) when an LLM is added as a fifth annotator alongside a four-human panel. Reference panel AC1 = 0.59.

Model	5-rater panel AC1	Δ_{AORI}	90% CI
GPT-5.2	0.3755	-0.2150	[-0.2503, -0.1802]
Claude Sonnet 4.6	0.3482	-0.2423	[-0.2793, -0.2068]
QWEN3.6-36B	0.5379	-0.0525	[-0.0709, -0.0350]
Gemma4:31B	0.4246	-0.1659	[-0.1968, -0.1354]

1,939,233 tokens, which amounts to $\approx$ \$3.70. To determine which model is the most cost effective, we calculated the Cost per True Positive (TP) on the video level annotations. These results are presented in Table 5. We adopt \$8/hour as a crowdsourcing rate, because this is Prolific’s recommended minimum compensation for research participants [14]. We treat \$8/hour as a conservative estimate. Out of all models, Claude Sonnet 4.6 is the most expensive one, requiring almost \$1 for a relevant video, whereas QWEN3.6-36B and Gemma4:31B are the cheapest options. Recent advances in video generation models have enabled researchers to generate video datasets using carefully curated prompts. Generating a synthetic dataset of 298 5-second videos costs \$29.95 for the starting images (NanoBananaPro, \$0.10052 per image) and \$223.50 for the videos (Seedance 2.0, \$0.15/s), \$253.45 in total. Consequently, using video generation models to generate videos, the cost per video is approximately \$0.85, assuming that ev-

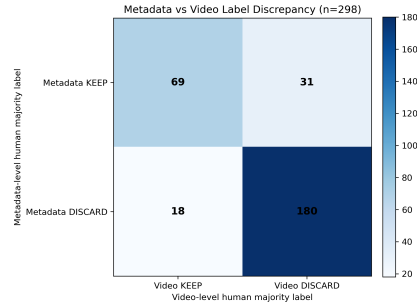


Fig. 1: Confusion matrix between metadata-level and video-level human majority labels (2-of-3 majority of annotators A, B and C, $n = 298$)

Table 5: Full pipeline cost for $n = 298$ candidates (\$8/hr, 3 min/video). Local models and baselines incur no-LLM cost.

Method	Total	Saving	Cost/TP (meta)	Cost/TP (video)
Manual (no filter)	\$119.20	—	\$1.19	\$1.37
GPT-5.2	\$47.47	2.5×	\$0.75	\$0.97
Gemma4:31B	\$46.00	2.6×	\$0.55	\$0.73
QWEN3.6-36B	\$48.00	2.5×	\$0.59	\$0.74
Claude Sonnet 4.6	\$54.28	2.2×	\$0.66	\$0.87
Regex A (stoplist)	\$71.20	1.7×	\$1.02	\$1.13
Regex B (stoplist+keyword)	\$27.60	4.3×	\$0.71	\$0.75

ery generated clip is usable, not taking into account generation artifacts. Table 5 shows that the best metadata filter costs \$0.73 per video, which is cheaper than the cost per video for video generation models, not taking into account the usability of the generated videos.

7 Limitations

Some limitations of this work should be acknowledged. First, the video level annotation subset is relatively small, consisting of only 298 videos. While sufficient for a proof-of-concept, the reported metrics may not be fully representative of performance on a larger scale. Second, the threshold parameter was tuned after labeling the dataset, which means that the parameter t is only usable in this specific setting. Lastly, the prompt that was passed to each model was not prompt engineered with respect to each model. Future work could study the difference in performance when prompt engineering or combining each model with a RAG.

8 Conclusions

To answer the research questions mentioned at the start of this work, we evaluated each LLM with a Leave-One-Rater-Out mechanism and an Add-One-Rater-In mechanism and calculated the AC1. Results have shown that for all models, the performance declines, meaning that no annotator can be substituted by an LLM. QWEN3.6-36B aligns the most with the annotator panel. This outcome is also shown in the Add-One-Rater-In mechanism, where the AC1 minimally decreases (-0.0525), which means that QWEN3.6-36B comes closest to acting as a 5th annotator, although this decrease is still significant. All of the models were also compared to the regex baselines which are used by other dataset construction procedures. QWEN3.6-36B is more cost effective in comparison to both regex filters and shows better performance overall.

Taken together, the three research questions receive different degrees of support. RQ1 is confirmed: one of the four models retains high recall (0.857) and all four outperform at least one no-LLM baseline on F1 and recall. RQ2 is confirmed on the video-validated subset, where every LLM pipeline costs less per relevant video recovered than unfiltered manual annotation. The same pattern holds on the full sample using metadata-level proxy, though this has not yet been validated against actual video content. RQ3 receives a cautious answer: at metadata level, no model safely substitutes for or joins any human annotator, expert or non-expert, without significantly degrading panel agreement, though QWEN3.6-36B is consistently the least disruptive. No single model is "best" across these questions. QWEN3.6-36B achieves the highest F1 and is the most stable virtual annotator, but has a lower level of recall, a weakness given our own argument that recall should be prioritized in a first-stage pre-filter, since a relevant video incorrectly discarded is lost from the pipeline entirely. Gemma4:31B better satisfies that recall-first criterion, at the cost of a slightly lower F1. Choosing between the models is therefore a matter of deciding which error a given pipeline can tolerate less: missing a relevant video (favoring Gemma4:31B) or paying to review more irrelevant ones downstream (favoring QWEN3.6-36B). We therefore answer RQ3 as follows: LLMs, in this complex case, cannot yet substitute human annotators at metadata level, irrespective of the annotator's expertise. QWEN3.6-36B comes closest to a viable substitute. This does not mean LLM pre-filtering is without value. Against the alternative that would be realistically used, such as keyword matching, QWEN3.6-36B and Gemma4:31B outperform both regex baselines on F1. Gemma4:31B outperforms regex A on all metrics.

References

- [1] Mykhaylo Andriluka et al. "2D Human Pose Estimation: New Benchmark and State of the Art Analysis". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH, USA: IEEE, June 2014, pp. 3686–3693. DOI: 10.1109/CVPR.2014.471.
- [2] DeepSeek-AI et al. *DeepSeek-V3 Technical Report*. arXiv:2412.19437 [cs]. DOI: 10.48550/arXiv.2412.19437.

- [3] Alvan R. Feinstein and Domenic V. Cicchetti. “High agreement but low Kappa: I. the problems of two paradoxes”. en. In: *Journal of Clinical Epidemiology* 43.6 (Jan. 1990), pp. 543–549. DOI: 10.1016/0895-4356(90)90158-L.
- [4] Raghav Goyal et al. *The "something something" video database for learning and evaluating visual common sense*. arXiv:1706.04261 [cs]. June 2017. DOI: 10.1109/ICCV.2017.622.
- [5] Feng Gu et al. *Large Language Models Are Effective Human Annotation Assistants, But Not Good Independent Annotators*. arXiv:2503.06778 [cs.CL]. Apr. 2026. DOI: 10.18653/v1/2026.findings-acl.4. URL: <https://aclanthology.org/2026.findings-acl.4/>.
- [6] Kilem Gwet. “Handbook of inter-rater reliability”. In: *Gaithersburg, MD: STATAXIS Publishing Company* (2001), pp. 223–246.
- [7] Muhammad Uzair Ul Haq, Davide Rigoni, and Alessandro Sperduti. *LLMs as Data Annotators: How Close Are We to Human Performance*. arXiv:2504.15022 [cs.CL]. Apr. 2025. DOI: 10.48550/arXiv.2504.15022.
- [8] Fabian Caba Heilbron et al. “ActivityNet: A large-scale video benchmark for human activity understanding”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, June 2015, pp. 961–970. DOI: 10.1109/CVPR.2015.7298698.
- [9] *How YouTube search works - YouTube Help*. URL: <https://support.google.com/youtube/answer/16090438?hl=en>.
- [10] Will Kay et al. *The Kinetics Human Action Video Dataset*. arXiv:1705.06950 [cs]. May 2017. DOI: 10.48550/arXiv.1705.06950.
- [11] H. Kuehne et al. “HMDB: A large video database for human motion recognition”. In: *2011 International Conference on Computer Vision*. Barcelona, Spain: IEEE, Nov. 2011, pp. 2556–2563. DOI: 10.1109/ICCV.2011.6126543.
- [12] Jingyang Lin et al. “Unleashing Hour-Scale Video Training for Long Video-Language Understanding”. In: *Advances in Neural Information Processing Systems 38*. San Diego, California, USA and Mexico City, Mexico: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2025, pp. 19950–19979. DOI: 10.52202/085713-0591.
- [13] Laurent Mertens et al. “FindingEmo: An Image Dataset for Emotion Recognition in the Wild”. In: *Advances in Neural Information Processing Systems 37*. Vancouver, BC, Canada: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024, pp. 4956–4996. DOI: 10.52202/079017-0161. (Visited on 08/25/2026).
- [14] *Pricing*. URL: <https://www.prolific.com/pricing> (visited on 05/05/2026).
- [15] Tanzila Rahman, Renjie Liao, and Leonid Sigal. *All in One: A Unified Synthetic Data Pipeline for Multimodal Video Understanding*. arXiv:2604.12335 [cs]. Apr. 2026. DOI: 10.48550/arXiv.2604.12335.
- [16] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild*. arXiv:1212.0402 [cs]. DOI: 10.48550/arXiv.1212.0402.

- [17] Peng Wang et al. *Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution*. arXiv:2409.12191 [cs]. DOI: 10.48550/arXiv.2409.12191.
- [18] Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. “Actions ~ Transformations”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 2658–2667. DOI: 10.1109/CVPR.2016.291.

9 Appendices

Appendix A: Prompt

You are labeling video metadata for the FindingEmo dataset.

IMPORTANT

- You do NOT see the video.
- You must use ONLY the title and description, including hashtags.
- This is a FIRST-STAGE HIGH-RECALL PREFILTER, not a final classifier.

Your job is to remove only videos that are clearly ineligible. If there is any realistic possibility of real human presence, interpersonal context, or personal experience, you should KEEP the video.

CORE DECISION RULE

- DISCARD only when the metadata explicitly states or strongly implies the video is outside scope.
- Otherwise KEEP.
- When uncertain → KEEP.
- Prefer false positives over false negatives.

CRITICAL INTERPRETATION RULES

- Absence of emotional language is NOT a reason to discard.
- Weak or unclear emotion is NOT a reason to discard.
- Absence of people must NOT be assumed unless explicitly stated or strongly implied.
- If human context is plausible → KEEP.

Use this decision procedure:

STEP 1 — HARD DISCARD

Return DISCARD only if metadata clearly matches one of these:

1. NO PEOPLE / NON-HUMAN ONLY

Explicit or strongly implied:

- “no people”
- nature, scenery, ambience, landscapes
- objects only, background visuals

2. AI-GENERATED / SYNTHETIC

Explicit mentions of:

- AI-generated, artificial, synthetic, deepfake, avatar, animated

3. PURE INSTRUCTIONAL / TECHNICAL

Clear tutorial, how-to, tips, product demo, or instructional content AND no personal or first-person framing.

4. **STOCK / COMPILATION / AESTHETIC**
Stock footage, repost compilations, montages, aesthetic visuals without human context.
5. **NON-HUMAN FORMAT CONTENT**
Gameplay, screen recordings, slides, text-only, graphics-only.
6. **PERFORMANCE-ONLY CONTENT**
Music, dance, or performance with no personal, interpersonal, or narrative context.

If none of these clearly apply → KEEP.

STEP 2 — KEEP SIGNALS

KEEP if there is any indication of:

- a person or people
- first-person language (I, my, we)
- relationships (family, friend, partner)
- life events, experiences, or stories
- emotional language or emojis
- vlog or personal framing
- any plausible human or social context

STRONG KEEP SIGNALS

- Personal pronouns: I, my, me, we, our
- Relationships: mom, dad, friend, girlfriend, boyfriend, family
- Situations: breakup, reunion, decision, story, reaction, vlog, experience
- Emojis or emotional words: sad, happy, love, scared

EMOTION LIKELIHOOD (independent from decision)

- high = explicit emotional situation or reaction
- medium = human or interpersonal context present
- low = little emotional evidence but not excluded
- unclear = very limited metadata

AI-GENERATED LIKELIHOOD

- high = explicit AI/synthetic indicators
- medium = ambiguous
- low = no indication

CONFIDENCE

- 0.90–0.98 = clear DISCARD
- 0.85–0.95 = clear KEEP with strong signals
- 0.70–0.84 = ambiguous but KEEP

OUTPUT

Return ONLY valid JSON:

```
{
  "decision": "KEEP" or "DISCARD",
  "confidence": 0.0,
  "emotion_likelihood": "high|medium|low|unclear",
  "ai_generated_likelihood": "low|medium|high",
  "reason": ""
}
```

Reason

- 1-2 sentences
- based only on metadata
- explain why rule triggered

EXAMPLES

Title: My dad surprised me after 5 years apart

Description: Didn't expect this #family

→ **KEEP**

Title: A day in my life working night shifts

Description: it's been hard lately #vlog

→ **KEEP**

Title: New skincare routine

Description: products that changed my life

→ **KEEP**

Title: Camera test with my girlfriend

Description: trying the new lens

→ **KEEP**

Title: Moving to a new city alone

Description: starting over is scary

→ **KEEP**

Title: Top 10 productivity hacks

Description: improve your workflow

→ **DISCARD**

Title: How to wake up at 5am

Description: step by step guide

→ **DISCARD**

Title: Relaxing rain sounds

Description: no people, just ambience

→ **DISCARD**

Title: Minecraft gameplay

Description: episode 12

→ **DISCARD**

Title: AI-generated short film

Description: made with AI tools

→ **DISCARD**

Appendix B: Examples of disagreement: non-expert - LLM



Fig. 2: Example of a disagreement where an LLM (GPT in this case) labels DISCARD and the non-expert labels KEEP.

link: <https://www.youtube.com/watch?v=KOMGOEAIxPU>

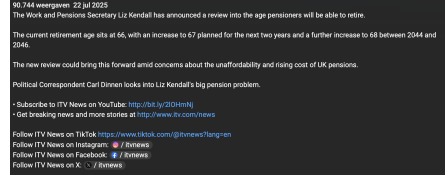


Fig. 3: Metadata associated with the video of Figure 2



Fig. 4: Example of a disagreement where an LLM (GPT in this case) labels KEEP and the non-expert labels DISCARD.

link: <https://www.youtube.com/watch?v=FsIDR1JR4Nk>



Fig. 5: Metadata associated with the video of Figure 4

Appendix C: Full pairwise comparison matrix

Exploratory; not corrected for multiplicity. All 14 pairs (4 LLMs, 2 baselines) compared via paired cluster-bootstrap on metadata-level F1 and recall ($n = 1047$, strict human-majority ground truth of the four annotators).

Comparison (A – B)	Metric	Δ	95% CI	Sig.
QWEN3.6-36B – Regex A (stoplist)	F1	+0.147	[+0.085, +0.207]	yes
QWEN3.6-36B – Regex B (stoplist+keyword)	F1	+0.117	[+0.036, +0.198]	yes
Gemma4:31B – Regex A (stoplist)	F1	+0.112	[+0.065, +0.162]	yes
Claude Sonnet 4.6 – Regex A (stoplist)	F1	+0.100	[+0.037, +0.156]	yes
GPT-5.2 – Regex A (stoplist)	F1	+0.082	[+0.012, +0.139]	yes
Gemma4:31B – Regex B (stoplist+keyword)	F1	+0.082	[+0.013, +0.154]	yes
Claude Sonnet 4.6 – Regex B (stoplist+keyword)	F1	+0.070	[−0.003, +0.147]	no
GPT-5.2 – Regex B (stoplist+keyword)	F1	+0.052	[−0.028, +0.133]	no
QWEN3.6-36B – Gemma4:31B	F1	+0.034	[−0.020, +0.088]	no
Claude Sonnet 4.6 – Gemma4:31B	F1	−0.012	[−0.073, +0.045]	no
GPT-5.2 – Claude Sonnet 4.6	F1	−0.018	[−0.067, +0.030]	no
GPT-5.2 – Gemma4:31B	F1	−0.031	[−0.099, +0.032]	no
Claude Sonnet 4.6 – QWEN3.6-36B	F1	−0.047	[−0.123, +0.021]	no
GPT-5.2 – QWEN3.6-36B	F1	−0.065	[−0.139, +0.015]	no
Gemma4:31B – Regex B (stoplist+keyword)	Recall	+0.468	[+0.378, +0.575]	yes
Claude Sonnet 4.6 – Regex B (stoplist+keyword)	Recall	+0.214	[+0.111, +0.331]	yes
QWEN3.6-36B – Regex B (stoplist+keyword)	Recall	+0.175	[+0.052, +0.293]	yes
Gemma4:31B – Regex A (stoplist)	Recall	+0.149	[+0.058, +0.248]	yes
GPT-5.2 – Regex B (stoplist+keyword)	Recall	+0.136	[+0.021, +0.262]	yes
Claude Sonnet 4.6 – QWEN3.6-36B	Recall	+0.039	[−0.069, +0.159]	no
GPT-5.2 – QWEN3.6-36B	Recall	−0.039	[−0.164, +0.084]	no
GPT-5.2 – Claude Sonnet 4.6	Recall	−0.078	[−0.174, +0.007]	no
Claude Sonnet 4.6 – Regex A (stoplist)	Recall	−0.104	[−0.214, +0.019]	no
QWEN3.6-36B – Regex A (stoplist)	Recall	−0.143	[−0.266, −0.029]	yes
GPT-5.2 – Regex A (stoplist)	Recall	−0.182	[−0.311, −0.064]	yes
Claude Sonnet 4.6 – Gemma4:31B	Recall	−0.253	[−0.376, −0.146]	yes
QWEN3.6-36B – Gemma4:31B	Recall	−0.292	[−0.376, −0.205]	yes
GPT-5.2 – Gemma4:31B	Recall	−0.331	[−0.441, −0.215]	yes

Appendix D: Properties of metadata

Table 7: Descriptive statistics of the YouTube metadata

Feature	Mean	SD	Median [IQR]
Title length (characters)	58.1	20.5	57.0 [43.0–73.0]
Title length (words)	9.9	3.6	10.0 [7.0–12.0]
Description length (characters)	1,196	1,084	901.0 [390.5–1,642]
Description length (words)	166.5	159.2	121.0 [49.2–225.0]
URLs in description	4.7	5.9	3.0 [0.0–7.0]
Hashtags in description	2.4	5.4	0.0 [0.0–3.0]
Duration (minutes)	23.1	54.6	5.5 [2.9–16.5]
View count (thousands)	15,852	82,810	339.8 [8.9–3,508]
Upload year	2021	4.6	2022 [2018–2025]

Appendix E: Threshold-free performance

ROC-AUC and PR-AUC (average precision) computed on the keep score of each LLM, without a threshold ($n = 1047$, KEEP as positive class, 95% cluster-bootstrap CIs over search queries, 1000 iterations). The no-LLM baselines only produce binary decisions and therefore have no threshold-free score. The PR-AUC of a random classifier equals the KEEP prevalence (0.147).

Table 8: Threshold-free classification performance at metadata level ($n = 1047$).

Model	ROC-AUC	95% CI	PR-AUC	95% CI
QWEN3.6-36B	0.819	[0.780, 0.854]	0.370	[0.283, 0.462]
Gemma4:31B	0.730	[0.690, 0.776]	0.261	[0.200, 0.333]
GPT-5.2	0.728	[0.678, 0.775]	0.292	[0.216, 0.389]
Claude Sonnet 4.6	0.760	[0.710, 0.808]	0.314	[0.237, 0.406]

Appendix F: Tuned versus fixed threshold

Performance of each LLM with the fixed threshold $t = 0.5$, which corresponds to the model’s own KEEP/DISCARD decision, and with the tuned threshold t from the GroupKFold procedure (Section 3), at metadata level ($n = 1047$). Δ is tuned minus fixed, with 95% CIs from a paired cluster bootstrap over search queries (1000 iterations).

Table 9: Tuned versus fixed threshold ($t = 0.5$), metadata level ($n = 1047$).

Model	Metric	$t = 0.5$	Tuned t	Δ	95% CI	Sig.
QWEN3.6-36B	Precision	0.298	0.354	+0.055	[+0.019, +0.096]	yes
	Recall	0.851	0.565	-0.286	[-0.358, -0.208]	yes
	F1	0.442	0.435	-0.007	[-0.055, +0.041]	no
Gemma4:31B	Precision	0.232	0.261	+0.030	[+0.013, +0.048]	yes
	Recall	0.961	0.857	-0.104	[-0.168, -0.045]	yes
	F1	0.373	0.401	+0.027	[+0.003, +0.051]	yes
GPT-5.2	Precision	0.199	0.285	+0.086	[+0.041, +0.129]	yes
	Recall	0.838	0.526	-0.312	[-0.409, -0.225]	yes
	F1	0.322	0.370	+0.048	[-0.001, +0.102]	no
Claude Sonnet 4.6	Precision	0.199	0.286	+0.087	[+0.048, +0.130]	yes
	Recall	0.935	0.604	-0.331	[-0.419, -0.234]	yes
	F1	0.328	0.388	+0.060	[+0.006, +0.103]	yes