Can LLMs Reliably Label YouTube Videos? A Committee-Based Evaluation

Adriano Mourthé[1][0000−0002−7110−2026], Carlos Eduardo Mello[1][0000−0002−3632−4002] and Alípio Jorge[2][0000−0002−5475−1382]

[1] Federal University of the State of Rio de Janeiro,Av. Pasteur, 458 - Urca, Rio de Janeiro - RJ, 22290-255, Brazil
{adriano.mourthe,mello}@uniriotec.br

[2] University of Porto, Praça Gomes Teixeira, 4099-002 Porto, Portugal,
amjorge@fc.up.pt@fc.up.pt

Abstract. As recommender systems play an increasingly central role in shaping information exposure on platforms like YouTube, understanding the nature of the content they promote, especially in sensitive contexts, requires scalable and reliable labelling methods. This paper investigates the use of Large Language Models (LLM) to label YouTube videos based solely on their metadata. We propose a committee-based approach that aggregates predictions from an ensemble of seven state-of-the-art LLMs through majority voting. Using a novel dataset collected via simulated user interactions on YouTube, we analyse model agreement, labelling behavior, and the influence of model size. To assess label reliability, we also investigate the semantic coherence of label assignments. Our results show that LLM committees produce highly consistent labels in low-disagreement settings. These findings highlight both the promise and limitations of LLM-based annotation for auditing social networks.

Keywords: YouTube · Large Language Models · Recommender Systems.

1 Introduction

Digital platforms such as YouTube, Instagram, and TikTok have become central spaces for content creation and dissemination, with YouTube alone reporting over two billion active monthly users [1]. These platforms rely on Recommender Systems (RS) to manage and personalize user experience [2]. Recommender systems function by analysing user historical data, including inferred preferences and socio-demographic characteristics, thereby shaping how users interact with information, entertainment, and commercial content [3].

The increasing public reliance on social media for content consumption has raised concerns regarding the role of RS in content curation and exposure. However, research into the interactions among RS, digital platforms, and their users is constrained by a shortage of annotated datasets. Current annotation pipelines predominantly rely on manual workflows involving teams of human annotators, frequently recruited through specialized platforms such as Amazon's Mechanical Turk. Although human-generated annotations remain widely employed, these methods are expensive, time-consuming, and often lack sufficient experimental control, especially in large-scale, multilingual environments like YouTube [4–6].

Recent advances in Large Language Models (LLMs), such as the GPT model family, have provided a promising alternative for automating data annotation tasks [7, 8]. Pre-trained on extensive textual corpora and capable of performing tasks through zero-shot learning, LLMs have shown strong performance across various applications, including content labelling, sentiment analysis, and topic classification [4, 9]. These capabilities position LLMs as scalable and yet cost-effective tools for automated content annotation, addressing many limitations inherent in traditional human-based annotation approaches.

However, several challenges remain in the use of LLMs for annotation tasks. Recent literature has highlighted concerns regarding model bias, sensitivity to prompt design, and the opaque reasoning processes underlying labels generated by LLMs [4]. Moreover, there has been limited evaluation of whether such labels are consistent, meaningful, and reliable, particularly when applied to noisy and heterogeneous data such as YouTube video metadata.

This work addresses the aforementioned open problems by investigating the reliability of LLMs in the task of data annotation, with a particular focus on labelling politically charged YouTube videos based on metadata. The investigation is conducted on a novel dataset collected using sock-puppet agents, whose behavior was guided by trending search topics obtained from Google Trends. Annotations are generated using predictions from seven state-of-the-art LLMs, enabling an evaluation of consistency in labelling behavior. The investigation is structured around the following research questions:

- **RQ1**: To what extent do different LLMs produce consistent or divergent labels when applied to YouTube video metadata?
- **RQ3**: To what extent can LLM-generated labels be considered reliable when no ground-truth annotations are available?

2 Experimental Setup

2.1 Simulated User Interaction and Data Acquisition

We acquired YouTube interaction data using automated agents, commonly referred to as "sock puppets" [10]. These bots were scripted to simulate user behavior on the platform by watching videos in sequence. To align their activity with real-world user interests, Google Trends data were used to guide query selection. Popular non-political search queries from Brazilian users in 2024 served as seed inputs, and their relative popularity informed the prioritization and structure of each bot's trajectory through the query space. A total of 200 bots were deployed, each performing five rounds of query-based interactions.

During each round, metadata were extracted from the recommendation lists presented to the bots. For each recommended video, the title and channel name were recorded. This procedure yielded approximately 8,000 unique videos spanning 60 different languages. As these metadata fields serve as the sole input to the annotation task, and given the multilingual composition of the dataset, the analysis was restricted to Portuguese-language content. After filtering, the final dataset comprised 6,893 videos, which forms the basis for all experiments reported in this study.

The distinction between political and non-political content in our analysis is informed by the structure of the data collection process. To introduce variation in content exposure, 50% of the bots were deliberately directed to interact with a known political video: "Ciro Gomes fala sobre 'Doria com areia' | Pânico," published in 2018 by the Brazilian talk show Pânico Jovem Pan[3]. This video was selected from a dataset of 3,297 YouTube URLs shared in Brazilian political WhatsApp groups during the 2018 presidential election, published by Bursztyn et al. [11]. This controlled intervention creates a contrast between profiles that follow non-political trajectories and those exposed to overt political content, motivating the binary classification task used in our study.

2.2 Large language models

The rapid advancement of LLMs has led to the release of numerous state-of-the-art models by major technology companies. These models differ in architecture, training methodology, and parameter scale, making model selection a nontrivial design decision. In this study, model choice was guided by factors relevant to the annotation task and the computational constraints under which it was conducted.

The first factor considered was model performance. As no established benchmark exists for the specific task of labelling YouTube video metadata, we adopted the MMLU-Pro benchmark [12] as a performance proxy. This benchmark evaluates general language understanding and reasoning capabilities across a diverse set of tasks. While not tailored to annotation, its wide adoption in the literature makes it a reasonable surrogate for estimating model effectiveness in labelling scenarios.

The second factor was multilingual capability. Since the dataset used in this study consists of content from the Brazilian YouTube platform, the majority of the material is in Portuguese. As a result, only models with documented support for Portuguese were considered. Models lacking robust multilingual coverage such as OLMo-2 and GLM-4 were excluded from consideration.

The final factor involved hardware feasibility. All experiments were conducted on a single NVIDIA A100 GPU with 40 GB of VRAM[4], which imposed practical

[3] Literal English translation of the title: "Ciro Gomes talks about 'Doria with sand' | Pânico". Video URL: https://www.youtube.com/watch?v=bY9MRIhGWvM

[4] To accommodate hardware constraints, all models requiring more than 40 GB of memory were quantized prior to inference. Specifically, *LLaMA-3.3-70B* was quantized using Q4_KS, *Qwen-2.5-72B* with IQ4_XS, *Mistral-3.1-24B* with 8-bit quantization, *Gemma-3-27B* via Google's Quantization-Aware Training (QAT), and *Phi-3.5-MOE* with Q6_KL.

limits on model size. Consequently, models with excessive memory requirements, such as Llama 3.1 405B, were deemed infeasible for inclusion in our experiments.

Table 1. Selected LLMs for Ensemble Labelling

Model Name	**Parameters (Billions)**
Llama-3.3	70
Falcon-3	10
Gemma-3-27B	27
Phi-3.5-MOE	41.9B
Qwen-2.5	72
Mistral-3.1	24
Granite-3.3[‡]	8

[‡] Granite-3.3 reasoning mode was not enable during inference.

Based on the criteria outlined above, we selected a diverse set of models for the annotation process. The final selection, along with the number of parameters for each model, is summarized in Table 1.

2.3 Inference and Prompt Design

Due to the absence of labeled data in our setting, supervised approaches that depend on annotated examples are not applicable. Furthermore, prior work has shown that handcrafted examples used in few-shot in-context learning can bias model predictions toward annotators' subjective interpretations [5]. To mitigate these issues, we adopt a *zero-shot* prompting strategy that relies exclusively on the models' pretrained knowledge and predefined class descriptions, without incorporating any task-specific demonstrations.

To ensure consistency across models, we implemented a standardized prompt format used uniformly during inference. Each prompt begins with a set of clearly defined category labels, each associated with a fixed uppercase letter. This is followed by the presentation of the video metadata,i.e. the title and channel name, along with an explicit instruction for the model to respond with a single-letter label corresponding to the appropriate category.

The definitions of `Political` and `Not Political` were adapted from prior work on political information exposure in web search contexts [13]. Constraining the output space to a single-token response helps reduce the likelihood of off-task or ambiguous completions and simplifies downstream processing.

3 Experimental Results

We begin our evaluation by examining the consistency of political content labelling across our ensemble of LLMs. First, we examine the distribution of political versus non-political labels assigned by each model, as well as the aggregated majority vote, see Table 2.

Table 2. Distribution of Political vs. Not Political Predictions per Model

Model	**Political (%)**	**Not Political (%)**
Granite-3.3-8B	9.2%	90.8%
LLaMA-3.3-70B	11.0%	89.0%
Falcon-3-10B	8.3%	91.7%
Gemma-3-27B	11.0%	89.0%
Mistral-3.1-24B	9.8%	90.2%
Phi-3.5-MOE	11.4%	88.6%
Qwen-2.5-72B	10.6%	89.4%
Majority Vote	**9.4%**	**90.6%**

Across all models, the majority of videos are labelled as not political, with proportions ranging from 88.6% to 91.7%. The smallest proportion of political classifications is observed in the Falcon-3-10B model (8.3%), while the highest is in the Phi-3.5-MOE model (11.4%). Interestingly, there is no huge gap between smaller models and larger ones, suggesting that model size alone does not fully account for variation in labelling behavior.

The majority vote outcome closely reflects the individual model distributions, resulting in 9.4% political labels and 90.6% non-political. This indicates a relatively high level of consensus among models, despite differences in scale and architecture. Nonetheless, the modest variation across models suggests that other factors, such as training data composition or instruction tuning, may play a role in the sensitivity to political content.

3.1 Can We Trust the Committee's Labels?

Although our committee comprises state-of-the-art large language models, a central question remains: can we trust the labels it produces? In the absence of ground-truth annotations, direct evaluation of classification accuracy is not feasible. However, it is possible to assess whether the generated labels exhibit internal semantic consistency.

If one assumes that the committee's labels are meaningful, then semantically similar items, i.e., neighbours in a semantic embedding space, should tend to receive the same label. For the majority of non-niche content, neighbourhood agreement offers a useful proxy for label reliability. If a video that elicits some level of disagreement among models receives the same label as its nearest high-confidence neighbours, this alignment supports the validity of the committee's majority decision.

To implement this evaluation strategy, we require an embedding model capable of capturing semantic relationships based solely on video metadata. For this

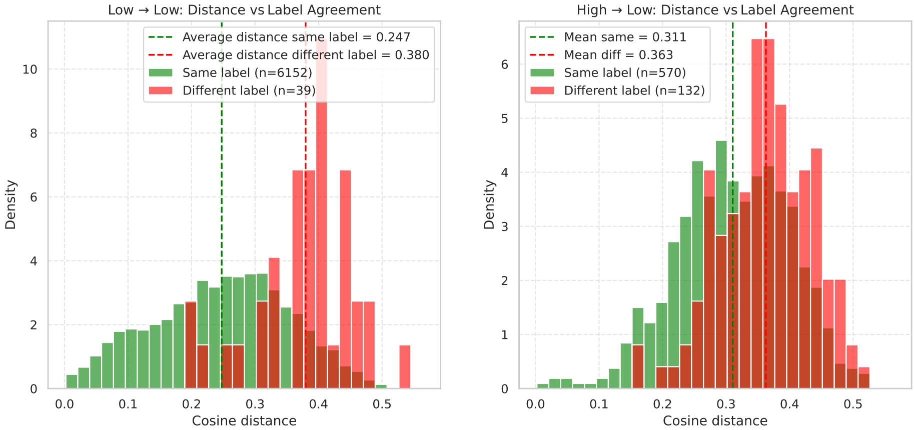

Fig. 1. Distribution of cosine distances between video embeddings and their nearest neighbours, stratified by label agreement. Vertical lines indicate mean distances for each group.

purpose, we use the GRIT-7B model[5], a state-of-the-art instruction-tuned embedding model. GRIT-7B ranks among the top performers on the Massive Text Embedding Benchmark [14] and Massive multilingual text embedding benchmark [15]. It is specifically trained to generate embeddings that reflect fine-grained semantic similarity in multiple domains for different tasks.

Figure 1 shows the distribution of cosine distances between nearest neighbours under two evaluation conditions. In the first (left panel), both the target and its neighbour are drawn from the low-entropy (no disagreement) region. In the second (right panel), a high-entropy (disagreement) item is compared to its nearest neighbour from the low-entropy set. In both cases, we report the distance distributions separately for pairs that share the same committee label and those that do not.

In the low-to-low condition, we observe a clear separation between the distributions. Video pairs with matching labels exhibit a mean cosine distance of 0.247, whereas mismatched pairs average 0.380. The distributions curves are well-separated, indicating strong semantic coherence among videos that share the same label. Among 6,191 evaluated video pairs, only 39 (0.6%) are label-inconsistent, suggesting extremely high internal agreement.

The high-to-low condition reveals a similar, albeit slightly weaker, trend. Video pairs with matching labels show a mean distance of 0.311, while mismatched pairs average 0.363. Although the distributions overlap more—reflecting the ambiguity introduced by model disagreement—the label still matches in 81.2% of cases (570 out of 702). These results suggest that even under moderate ambiguity, semantic similarity remains predictive of label consistency, providing empirical support for the internal coherence of the committee's decisions.

[5] https://huggingface.co/GritLM/GritLM-7B

Overall, these findings reinforce that majority-vote labels are reliable in low-disagreement conditions and retain meaningful structure even in moderately ambiguous cases. However, under high levels of disagreement—particularly for politically sensitive content—semantic locality begins to degrade, underscoring the limitations of LLM-based annotation in boundary regions of the input space.

4 Conclusion and Future Work

This paper explored the use of Large Language Models to classify YouTube videos based solely on their metadata, focusing on the political relevance of content. In the absence of ground-truth labels, we proposed a committee-based labelling approach, leveraging an ensemble of seven diverse LLMs and assigning labels through majority voting. Our aim was to assess the feasibility, consistency, and reliability of LLM-generated annotations in a real-world, politically sensitive context.

Through systematic evaluation, we found that the committee exhibits high overall agreement, particularly in labelling non-political content, where both inter-model consensus and embedding-based neighbourhood coherence were consistently strong. However, political classifications revealed greater heterogeneity, especially in cases with partial or full disagreement among models. We showed that larger models tend to agree more frequently with the committee on political content, suggesting that factual recall and model capacity play a meaningful role in these more ambiguous cases.

To assess label reliability in the absence of ground truth, we introduced a validation strategy based on semantic consistency in embedding space. Using GRIT 7B embeddings, we show that videos with similar content received the same label in low-disagreement cases, supporting the internal coherence of the committee's decisions. Under higher disagreement levels, particularly for political labels, this consistency degraded, suggesting boundary regions where label reliability may be lower.

Overall, our findings suggest that LLMs, when used in committee and with a carefully designed prompt, offer a scalable, and semantically consistent alternative to traditional human annotation pipelines for evaluting recommender systems and social networks. However, our results also show that not all labels are equally trustworthy. Disagreement levels and semantic context provide important signals for gauging when additional validation may be necessary.

Future work could extend our work to multilingual settings, model calibration techniques, or hybrid pipelines combining LLMs with retrieval-based validation. Further research is also needed to explore the implications of these labelling strategies for downstream causal inference in RS audit experiments.

References

1. YouTube. Youtube for press. https://www.youtube.com/about/press/, 2019. Accessed: 2024-09-16.

2. Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook*, pages 1–35, 2021.
3. Dietmar Jannach and Michael Jugovac. Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)*, 10(4):1–23, 2019.
4. Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.
5. Petter Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*, 2023.
6. Kyle A Thomas and Scott Clifford. Validity and mechanical turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, 77:184–197, 2017.
7. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
8. Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
9. Zhen Tan, Alimohammad Beigi, Song Wang, Ruocheng Guo, Amrita Bhattacharjee, Bohan Jiang, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation: A survey. *arXiv e-prints*, pages arXiv–2402, 2024.
10. Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, 22(2014):4349–4357, 2014.
11. Victor S Bursztyn and Larry Birnbaum. Thousands of small, constant rallies: A large-scale analysis of partisan whatsapp groups. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 484–488, 2019.
12. Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2025.
13. Marieke van Hoof, Damian Trilling, Corine Meppelink, Judith Möller, and Felicia Loecherbach. Googling politics? comparing five computational methods to identify political and news-related searches from web browser histories. *Communication Methods and Measures*, 19(1):63–89, 2025.
14. Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, 2023.
15. Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025.