Comparative Analysis of Language Models in the Understanding of Short Videos: A Multimodal Approach for Disinformation Detection

Adriel L. V. Mori[1,2], Eliomar Araújo Lima[1], Valdemar V. Graciano Neto[1], Jacson R. Barbosa[1], Rogerio Rodrigues Carvalho[1], Arlindo R. Galvão Filho[1,2]

[1]Institute of Informatics – Federal University of Goiás (UFG)

[2]Advanced Knowledge Center for Immersive Technologies (AKCIT)
Goiânia – GO – Brazil

adrielmori@discente.ufg.br,

{eliomar,valdemarneto,arlindogalvao,jacson_rodrigues,rogerior}@ufg.br

Abstract. *Short-form videos are increasingly used to spread political disinformation, but multimodal benchmarks for Brazilian Portuguese remain limited. This study evaluates whether structured multimodal evidence improves zero-shot LLM-based detection in Brazilian political videos. We combine CLIP keyframe summarization, Whisper transcription, and pyannote-audio diarization to build prompts from auditory and visual evidence. We benchmark 12 LLMs on 119 videos from the 2022–2024 Brazilian electoral cycles, comparing audio-only and multimodal settings for veracity and taxonomy classification. Results show that* `gemma2-9b` *achieves the best veracity performance (macro-F_1 = 0.764), while* `gemma3-12b` *leads taxonomy classification (macro-F_1 = 0.518).*

1. Introduction

Disinformation in short-form videos has become a relevant challenge for digital platforms, especially in political contexts, where misleading narratives may affect public debate and trust in institutions [Wardle and Derakhshan 2017, Lazer et al. 2018, Rodrigues et al. 2020]. Unlike text-only misinformation, short videos combine speech, captions, visual scenes, and editing cues, making their interpretation inherently multimodal.

This complexity limits manual analysis and motivates computational approaches that integrate visual, auditory, and textual evidence [Shu et al. 2017, Baltrušaitis et al. 2018]. Prior work has shown the value of multimodal modeling for misinformation detection [Wu et al. 2021, Jing et al. 2023, Hua et al. 2023]. More specifically for short videos, Shang et al. [Shang et al. 2021] addressed COVID-19 misinformation on TikTok, Ren et al. [Ren et al. 2024] proposed a multi-grained fusion strategy for short-video fake news detection, and Bu et al. [Bu et al. 2024] showed that signals from the creative process, such as material selection and editing patterns, are also informative. These studies indicate that short-video misinformation detection requires more than conventional text-image fusion.

However, two gaps remain. First, most prior studies focus on English or Chinese datasets, with little attention to Brazilian Portuguese and political short videos. Second, the dominant paradigm is based on supervised multimodal architectures, while the use of

Large Language Models (LLMs) as multimodal interpreters remains limited in this setting. In Brazilian Portuguese, recent work has shown that LLMs can be effective for political fake news detection in textual scenarios [Gôlo et al. 2024], but short-video audiovisual settings are still underexplored.

In this paper, we evaluate a multimodal zero-shot pipeline for Brazilian Portuguese political short videos. The approach combines Contrastive Language–Image Pre-Training (CLIP)-based keyframe summarization, Whisper speech transcription, speaker diarization with *pyannote-audio*, and visual-language descriptions of representative frames. These elements are organized into structured prompts that allow LLMs to reason over textualized representations of visual, auditory, and textual evidence.

Our contribution is not to propose a new end-to-end supervised detector, but to benchmark an auditable LLM-centered multimodal pipeline for Brazilian Portuguese political short videos. Specifically, we:

- propose a zero-shot prompting pipeline that converts speech, speaker information, and visual keyframes into a structured textual representation for LLM-based analysis;
- evaluate 12 LLMs under audio-only and full multimodal settings, measuring the impact of frame-level visual evidence on veracity and taxonomy classification;
- provide an empirical benchmark for multimodal disinformation detection in a low-resource political scenario, identifying performance gains, model-dependent behavior, and persistent failure cases.

2. Related Works

Disinformation detection in short-form videos is challenging because misleading narratives often combine speech, captions, visual scenes, music, and editing cues. In this context, multimodal approaches are relevant because they can capture complementary evidence and cross-modal inconsistencies that single-modality methods may miss [Atrey et al. 2010, Bouchey et al. 2021, Wu et al. 2021].

Recent studies have explored multimodal fake news detection in short-video scenarios. Shang et al. [Shang et al. 2021] proposed TikTec for COVID-19 misinformation on TikTok, using captions to guide the integration of visual, audio, and textual signals. Ren et al. [Ren et al. 2024] introduced MMSFD, a multi-grained fusion approach for short-video fake news detection. Bu et al. [Bu et al. 2024] further showed that production-related cues, such as material selection and editing patterns, can also provide useful evidence for identifying misinformation.

Although these works demonstrate the importance of multimodal evidence, most studies focus on English- or Chinese-language datasets and supervised end-to-end architectures. In Brazilian Portuguese, Golo et al. [Gôlo et al. 2024] showed that LLMs can be effective for political fake news detection, but their setting is textual rather than audiovisual. Therefore, our work investigates an auditable LLM-centered pipeline for Brazilian Portuguese political short videos, evaluating how textualized multimodal evidence can support zero-shot disinformation detection in a low-resource and politically sensitive setting.

3. Materials and Methods

This section describes the dataset construction and experimental methodology adopted in this study. We first present the corpus of Brazilian Portuguese political short videos and its annotation process. We describe the multimodal pipeline used to transform audiovisual evidence into structured textual inputs for evaluating LLM-based disinformation detection.

3.1. Dataset

This paper introduces a curated dataset of 119 short videos focused on politically relevant content, primarily collected during the 2022 Brazilian presidential election and the 2024 municipal election, with a smaller number of instances from other electoral periods. Videos were identified through keyword-based search — using terms related to electoral themes, political figures, and previously fact-checked claims — and sourced from TikTok, Instagram, YouTube, and X. Table 1 summarizes the main characteristics of the corpus.

Table 1. Summary of dataset characteristics.

Characteristic	Value
Total videos	119
Total duration	159 min
Average duration	≈80 sec
Platforms	TikTok, Instagram, YouTube, X
Collection period	2022–2024
Video quality levels	High, medium, low
Veracity labels	
Inverídico	70 (58.8%)
Verídico	31 (26.1%)
Inconclusivo	18 (15.1%)
Taxonomy labels	
Descontextualização	35 (29.4%)
Não se aplica	32 (26.9%)
Fake news	20 (16.8%)
Discurso de ódio	20 (16.8%)
Extremismo antidemocrático	12 (10.1%)

The dataset was constructed within a real-world fact-checking workflow. Candidate videos were first flagged as potentially harmful by trained fact-checkers and subsequently reviewed by two independent annotators with specialized expertise in political communication and social media analysis, affiliated with the Faculdade de Ciências Sociais (FCS)[1] at the Universidade Federal de Goiás (UFG). Final inclusion was decided by a lead curator. This process prioritized relevance, credibility, and operational usefulness over class balance, since the objective was to capture cases that effectively emerged in applied verification scenarios. Videos were additionally categorized by technical quality (high, medium, or low) to retain the variability in resolution, framing, and audio conditions commonly found in short-video ecosystems.

Each video was annotated along two dimensions: *verdict* (*verídico*, *inverídico*, *inconclusivo*) and *disinformation category*, using a taxonomy grounded in academic literature and Brazilian electoral jurisprudence (*descontextualização*, *fake news*, *não se*

[1] https://fcs.ufg.br/

aplica, *discurso de ódio*, *extremismo antidemocrático*). As shown in Table 1, the corpus is class-imbalanced: *inverídico* represents the majority class (58.8%), while *inconclusivo* and rarer taxonomy categories remain underrepresented. The dataset should therefore be interpreted as a realistic operational corpus for exploratory multimodal benchmarking in Brazilian Portuguese, reflecting both the potential and the limitations of applied data collection in low-resource political settings.

4. Proposed Methodology

This section outlines the architecture of the proposed approach, which integrates visual, auditory, and textual data for the detection of potential disinformation in short videos. The methodology is structured as a multimodal pipeline illustrated in Figure 1.

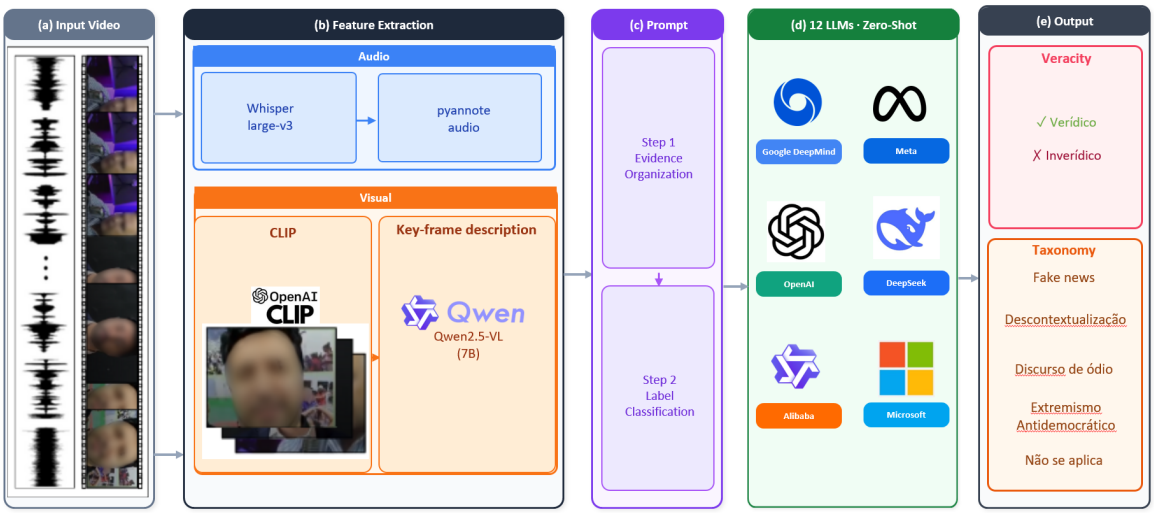

Figure 1. Proposed multimodal pipeline for disinformation detection in Brazilian Portuguese short videos. Given an input video (a), audio and visual streams are independently processed for feature extraction (b). The resulting evidence is structured into a two-step prompt (c), evaluated by 12 zero-shot LLMs (d), producing veracity and taxonomy labels (e).

Given a short political video, the approach extracts two parallel evidence streams: (i) an audio stream processed by Whisper large-v3 for speech transcription and by pyannote-audio for speaker diarization, and (ii) a visual stream in which CLIP selects semantically informative keyframes described in natural language by Qwen2.5-VL (7B). Both streams are combined into a unified structured prompt and evaluated by 12 LLMs under a two-step zero-shot strategy.

For the binary veracity task, videos annotated as *inconclusivo* were not used, since they do not correspond to either of the target binary labels. Therefore, binary classification metrics were computed only over the *verídico* and *inverídico* instances, avoiding the forced assignment of ambiguous cases to a definitive veracity class.

4.1. Multimodal Pipeline

This study evaluates whether multimodal evidence improves zero-shot LLM-based disinformation detection in Brazilian Portuguese political short videos. To this end, we designed a pipeline that converts each video into a structured textual representation combining speech,

speaker information, and salient visual content. The pipeline comprises six stages: pre-processing, keyframe summarization, audio transcription, speaker diarization, multimodal alignment, and instruction-based prompting. Rather than proposing a new end-to-end supervised detector, the objective is to assess how multimodal evidence can be organized and interpreted by LLMs in a low-resource and politically sensitive setting.

4.1.1. Pre-processing

Videos were standardized to the `.mp4` format using the H.264 codec and a uniform frame rate with FFmpeg.[2] Native resolution and original visual quality were preserved to avoid information loss in downstream analysis. The audio stream was extracted, normalized, and denoised using *pydub*[3] to improve transcription quality under heterogeneous platform conditions.

To reduce the effect of noisy or weakly aligned segments in the audio signal, we applied Drop-Dynamic Time Warping (Drop-DTW) [Dvornik et al. 2021], which extends classical DTW by allowing the alignment process to discard low-quality matches through a penalty term. It is defined as:

$$\text{Drop-DTW}(A, B) = \min_{\pi \in \mathcal{P}} \sum_{(i,j)\in\pi} \delta(a_i, b_j) + \lambda \cdot |\pi_d| \tag{1}$$

where π denotes the alignment path, δ is the local dissimilarity function between elements a_i and b_j, π_d represents the set of dropped elements, and λ is the penalty associated with each discarded match. In our setting, this step improves temporal consistency before transcription and downstream prompt construction.

4.1.2. Keyframe-Based Visual Summarization

The visual summarization stage selects representative keyframes that preserve the most informative visual events while reducing temporal redundancy [Apostolidis et al. 2021]. Frames were extracted at their native resolution using `cv2`.[4] Each frame was encoded with CLIP,[5] which maps images and text into a shared semantic space through a contrastive objective:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2N} \sum_{i=1}^{N} \left[-\log \frac{e^{s_{ii}/\tau}}{\sum_{j=1}^{N} e^{s_{ij}/\tau}} - \log \frac{e^{s_{ii}/\tau}}{\sum_{j=1}^{N} e^{s_{ji}/\tau}} \right], \quad s_{ij} = \cos(\mathbf{x}_i, \mathbf{y}_j). \tag{2}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the image and text embeddings for the i-th pair, s_{ij} denotes their cosine similarity, and τ is a temperature parameter. Keyframes were identified

[2] `https://ffmpeg.org/`
[3] `https://github.com/jiaaro/pydub`
[4] `https://opencv.org/`
[5] `https://github.com/openai/CLIP`

by computing cosine similarity across temporally adjacent CLIP embeddings within a fixed-size sliding window:

$$\mathrm{sim}(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i^\top \mathbf{v}_j}{\|\mathbf{v}_i\| \, \|\mathbf{v}_j\|}. \tag{3}$$

Frames with lower similarity to their temporal neighbors indicate semantic change and were therefore prioritized, producing a compact visual summary that preserves narrative shifts relevant to disinformation analysis.

4.1.3. Audio Extraction and Transcription

After pre-processing, the audio stream was isolated and transcribed using Whisper.[6] The resulting transcript provides a textual representation of spoken content and constitutes the main semantic input in the audio-only baseline. Because transcription quality directly affects downstream reasoning, multiple Whisper variants were evaluated in the experimental protocol.

4.1.4. Speaker Diarization

Speaker diarization was performed with *pyannote-audio*,[7] which segments speech and groups acoustically similar regions into speaker clusters. This step preserves speaker turns and allows the prompt to encode not only what was said, but also who said it. In politically sensitive audiovisual content, such attribution is relevant because interpretation may depend on speaker alternation, quoted speech, or dialogic structure.

4.1.5. Post-processing and Multimodal Alignment

Each transcribed segment was temporally aligned with the corresponding keyframe(s) according to overlap in time. To make visual evidence accessible within a text-based prompting interface, each selected keyframe was described in natural language using Qwen2.5-VL (7B) [Bai et al. 2025]. This model was selected because it is an open-weight vision-language model, enabling local execution, reproducibility, and greater control over the visual description stage. Moreover, the Qwen-VL family has been designed to process visual inputs with varying resolutions and to support unified image and video understanding, which makes it suitable for converting representative video frames into semantically grounded textual descriptions. These descriptions were generated via the Ollama API after base64 encoding of the images. The final prompt therefore combines transcription, speaker segmentation, and visual descriptions into a unified textual representation.

[6] https://github.com/openai/whisper

[7] https://github.com/pyannote/pyannote-audio

4.1.6. Instruction-based Prompting and LLM Inference

All models listed in Table 2 were evaluated under a zero-shot regime with prompts written in Brazilian Portuguese to match the linguistic characteristics of the dataset. The set includes open-weight models from Google DeepMind (`gemma2`, `gemma3`), DeepSeek (`deepseek-r1`), Meta (`llama3.1`, `llama3.2`), Microsoft (`phi3`), and Alibaba (`qwen2`, `qwen2.5`), as well as the proprietary models `gpt-4.1-mini` and `gpt-4o-mini` from OpenAI. Open-weight models span parameter scales from 3.2B to 32.8B and were evaluated in quantized Q4 configurations (Table 2). Proprietary models were included as reference baselines, although their architectural details are not publicly disclosed. To ensure reproducibility, all runs were executed under deterministic decoding with temperature fixed at $T = 0$.

Inference followed a two-step strategy designed to separate evidence organisation from final decision-making. In **Step 1**, the model received the structured video representation and was asked to extract and organise the relevant factual content. The *audio-only* and *full multimodal* conditions differed solely in the input prefix: the audio-only variant listed speech and auditory segments, while the multimodal variant additionally included visual descriptions.

Audio-only prefix:

> *"Para sua resposta, aja como um fact-checker sênior avaliando uma notícia enviada em formato de vídeo. A seguir está a transcrição detalhada do conteúdo, incluindo falas e trechos auditivos relevantes."*

Multimodal prefix:

> *"Para sua resposta, aja como um fact-checker sênior avaliando uma notícia enviada em formato de vídeo. A seguir está a transcrição detalhada do conteúdo, incluindo falas, trechos informativos e descrições visuais."*

In **Step 2**, the model produced two short classifications from the Step 1 output: (i) a final veracity label and (ii) a disinformation category. For veracity classification, the fixed prompt was:

> *"Classifique o nível de veracidade do conteúdo abaixo com uma das opções exatas: Verídico ou Inverídico. Responda apenas com o rótulo exato, e nada mais que o rótulo exato."*

For taxonomy classification, the model was instructed to assign one of five predefined categories: *fake news*, *descontextualização*, *discurso de ódio*, *extremismo antidemocrático*, or *não se aplica*. This decomposition improves interpretability and makes model decisions more traceable to the multimodal evidence provided in the prompt.

4.2. Experimental Protocol

The experimental protocol was designed to evaluate whether textualized multimodal evidence improves zero-shot LLM-based disinformation detection. Each video was processed through the pipeline described in Section 4.1: pre-processing, keyframe selection, Whisper transcription, speaker diarization, multimodal alignment, and instruction-based prompting as specified in Section 4.1.6. All models in Table 2 were then evaluated under the same zero-shot framework in two configurations:

- **Audio-Only Baseline:** the prompt included only speech transcription and speaker diarization. Three Whisper variants (`tiny`, `large`, and `large-v3`) were evaluated to determine how transcription fidelity affects downstream classification.
- **Full Multimodal Pipeline:** after selecting the best-performing Whisper variant, the visual module was added. The prompt combined transcription, speaker information, and natural-language descriptions of representative keyframes.

In both configurations, the final input was a structured textual representation of the available evidence. This design evaluates whether textualized multimodal evidence improves zero-shot reasoning without assuming native video understanding by the LLMs, while isolating the contribution of visual information across models with different parameter scales and architectural families. To support reproducibility, all code, configuration files, and preprocessing instructions are made available in the project repository,[8] subject to their original licenses. All experiments were executed with deterministic decoding ($T = 0$), using the same task instructions, label space, and prompting structure across models and experimental conditions.

Table 2. LLMs selected for zero-shot analysis.

Developer	Model	Parameters	Quantization	Size (GB)
Google DeepMind	`gemma3`	12.2B	`Q4_K_M`	7.59
DeepSeek	`deepseek-r1`	8.0B	`Q4_K_M`	4.58
Meta	`llama3.1`	8.0B	`Q4_K_M`	4.58
Meta	`llama3.2`	3.2B	`Q4_K_M`	1.88
Microsoft	`phi3`	14.0B	`Q4_0`	7.35
Alibaba	`qwen2.5`	14.8B	`Q4_K_M`	8.37
Alibaba	`qwen2.5`	7.6B	`Q4_K_M`	4.36
DeepSeek	`deepseek-r1`	32.8B	`Q4_K_M`	18.49
Google DeepMind	`gemma2`	9.2B	`Q4_0`	5.07
Alibaba	`qwen2`	7.6B	`Q4_0`	4.13
OpenAI	`gpt-4.1-mini`	—	—	—
OpenAI	`gpt-4o-mini`	—	—	—

The transcription module was evaluated in the audio-only setting before enabling the full multimodal pipeline. As shown in Table 3, `large-v3` achieved the best overall balance for downstream classification, with the highest F_1 for the *inverídico* class and the best macro-F_1 and weighted-F_1 in the taxonomy task. Although `large` obtained slightly better values for the *verídico* class, `large-v3` was selected for the full multimodal configuration because it provided the strongest aggregate performance across tasks.

4.3. Evaluation

The evaluation covered two tasks: binary *veracity classification* and *taxonomy classification*. Given the small and label-imbalanced nature of the dataset, macro-F_1 was adopted as the primary metric, complemented by precision, recall, weighted-F_1, and class-wise F_1. These metrics provide a more informative view of model behavior under skewed label distributions, especially for underrepresented classes.

For the main comparison, the same LLMs were evaluated under the audio-only and full multimodal configurations, differing only in the inclusion of visual descriptions.

[8] `https://github.com/adrielmori/sbbd2026-249311`

Table 3. Audio-only performance of Whisper variants for veracity and taxonomy classification. Best values are shown in bold.

Whisper	*Inverídico*			*Verídico*			Taxonomy Disinformation	
	Precision	**Recall**	F_1	**Precision**	**Recall**	F_1	**Macro-F_1**	**Weighted-F_1**
`large-v3`	**0.802**	**0.387**	**0.522**	0.504	0.210	0.296	**0.324**	**0.363**
`large`	0.799	0.385	0.520	**0.558**	**0.221**	**0.317**	0.317	0.356
`tiny`	0.747	0.343	0.470	0.449	0.077	0.131	0.288	0.334

This controlled design isolates the effect of frame-level visual evidence while preserving the same prompt structure, label space, and inference setting (Section 4.1.6). The two-step prompting strategy was applied identically in both configurations: first, the model organised the relevant factual content; then, it assigned the final veracity and taxonomy labels.

Component-level analysis was restricted to modules that directly affect downstream classification. The speech module was evaluated quantitatively through the comparison of Whisper variants and qualitatively through transcript inspection. The keyframe extraction step was analyzed qualitatively using different sliding-window sizes ($w = 5, 20, 30$), focusing on semantic redundancy and frame diversity. Speaker diarization and visual description generation were treated as supporting components rather than independent prediction tasks. Accordingly, the reported results should be interpreted as an exploratory benchmark for multimodal disinformation detection in Brazilian Portuguese political short videos, rather than as evidence of broad statistical generalization.

5. Results and Discussion

The results are reported in two stages. First, we evaluate the *audio-only baseline* to identify the best transcription setup and to estimate how much information can be recovered from speech alone. Second, we compare this baseline with the *full multimodal pipeline*, in which speech, speaker metadata, and frame-level visual descriptions are jointly provided to the LLMs. This organization allows the contribution of visual evidence to be analyzed directly.

In the audio-only phase, the three Whisper variants produced different downstream outcomes. As shown in Table 3, `large` and `large-v3` consistently outperformed `tiny`, both in binary veracity classification and in the taxonomy task. The gains are modest in absolute terms, but consistent across metrics, indicating that transcription quality is a limiting factor for downstream reasoning. Since `large-v3` achieved the best overall balance, it was selected for the multimodal stage.

For binary veracity classification, the full multimodal configuration generally improved performance relative to the audio-only baseline. As shown in Table 4, multimodal evidence improved the F_1 score for the minority *verídico* class in 9 out of 12 models, with particularly strong gains for `gemma2-9b` ($+37.5$ pp), `gpt-4.1-mini` ($+32.2$ pp), and `llama3.1-8b` ($+21.4$ pp). For the *inverídico* class, improvements were even more consistent, occurring in 11 out of 12 models. The only regression was observed for `llama3.1-8b` (-2.1 pp), which already had a strong audio-only baseline ($F_1 = 0.859$). These results suggest that frame-level visual descriptions provide useful complementary evidence, particularly for reducing failures in the minority *verídico* class, while also

Table 4. Comparison between audio-only and full multimodal settings for binary veracity and taxonomy classification. Δ reports absolute gains in percentage points (pp).

Model	Audio-Only			Full Multimodal			Δ (pp)		
	F_1 (Ver.)	F_1 (Inv.)	Tax.	F_1 (Ver.)	F_1 (Inv.)	Tax.	Δ Ver.	Δ Inv.	Δ Tax.
gemma2-9b	0.263	0.395	0.260	0.638	0.889	0.344	+37.5	+49.4	+8.4
gpt-4.1-mini	0.293	0.554	0.442	0.615	0.815	0.475	+32.2	+26.1	+4.3
llama3.1-8b	0.324	0.859	0.241	0.538	0.838	0.247	+21.4	−2.1	+0.6
qwen2-7b	0.486	0.029	0.250	0.615	0.679	0.246	+12.9	+65.0	−0.4
deepseek-r1-32b	0.366	0.531	0.312	0.435	0.831	0.326	+6.9	+30.0	+1.4
deepseek-r1-8b	0.381	0.452	0.275	0.507	0.704	0.252	+12.6	+25.2	−2.3
gemma3-12b	0.406	0.547	0.464	0.350	0.838	0.518	−5.6	+29.1	+5.4
qwen2.5-14b	0.244	0.447	0.304	0.350	0.838	0.442	+10.6	+39.1	+13.8
gpt-4o-mini	0.121	0.135	0.338	0.278	0.841	0.470	+15.7	+70.6	+13.7
phi3-14b	0.381	0.375	0.450	0.531	0.549	0.137	+15.0	+17.4	−31.3
qwen2.5-7b	0.216	0.458	0.284	0.176	0.831	0.270	−4.0	+37.3	−1.4
llama3.2-3b	0.000	0.816	0.184	0.000	0.817	0.290	+0.0	+0.1	+10.6

strengthening the detection of misleading content in most models.

For taxonomy classification, multimodal evidence improved macro-F_1 in 8 out of 12 models (Table 4), although gains were less consistent than in veracity classification. The largest improvements were observed for `qwen2.5-14b` ($+13.8$ pp) and `gpt-4o-mini` ($+13.7$ pp). In contrast, `phi3-14b` showed a severe regression (-31.3 pp), dropping from the best audio-only taxonomy score (macro-$F_1 = 0.450$) to the lowest multimodal result (0.137), suggesting sensitivity to longer prompts with visual descriptions. The best multimodal taxonomy result was achieved by `gemma3-12b` (macro-$F_1 = 0.518$), indicating that this model benefited most from visual evidence for fine-grained categorization.

Table 5. Per-class precision, recall, and F_1 for binary veracity classification under the full multimodal configuration. Macro-F_1 summarizes performance.

Model	*Inverídico*			*Verídico*			Macro-F_1
	Precision	Recall	F_1	Precision	Recall	F_1	
gemma2-9b	0.810	**0.986**	**0.889**	0.938	0.484	**0.638**	**0.764**
gpt-4.1-mini	0.833	0.797	0.815	0.588	0.645	0.615	0.715
llama3.1-8b	0.785	0.899	0.838	0.667	0.452	0.538	0.688
qwen2-7b	**0.925**	0.536	0.679	0.467	**0.903**	0.615	0.647
deepseek-r1-32b	0.753	0.928	0.831	0.667	0.323	0.435	0.633
deepseek-r1-8b	0.786	0.638	0.704	0.432	0.613	0.507	0.605
gemma3-12b	0.736	0.971	0.838	0.778	0.226	0.350	0.594
qwen2.5-14b	0.736	0.971	0.838	0.778	0.226	0.350	0.594
gpt-4o-mini	0.726	1.000	0.841	1.000	0.161	0.278	0.560
phi3-14b	0.848	0.406	0.549	0.388	0.839	0.531	0.540
qwen2.5-7b	0.711	1.000	0.831	1.000	0.097	0.176	0.504
llama3.2-3b	0.690	1.000	0.817	0.000	0.000	0.000	0.408

Under the full multimodal configuration, `gemma2-9b` achieved the best overall result for binary veracity classification, with macro-$F_1 = 0.764$ (Table 5). Its performance was especially strong for *inverídico*, reaching $F_1 = 0.889$ with recall $= 0.986$, while maintaining the best F_1 for *verídico* among all evaluated models (0.638). `gpt-4.1-mini` and `llama3.1-8b` also produced competitive and more balanced predictions across both classes.

In contrast, several models showed unstable class behavior. For example, `llama3.2-3b` failed completely on the *verídico* class, while `qwen2.5-7b` and `gpt-4o-mini` exhibited extreme recall or precision values, indicating overly aggressive or overly conservative decision patterns. These results suggest that multimodal prompting benefits some architectures more than others, and that robustness across both classes remains model-dependent.

The taxonomy classification task was considerably more difficult than the binary veracity task. This result is expected, since fine-grained categories such as *descontextualização*, *fake news*, *discurso de ódio*, and *extremismo antidemocrático* may overlap semantically in real videos, especially in politically charged material. In addition, the limited number of instances and the imbalance among taxonomy labels make this setting particularly challenging for zero-shot classification. Although the multimodal configuration provided useful complementary evidence in some cases, the results indicate that fine-grained disinformation categorization remains fragile and highly model-dependent. Visual information appears to support semantic interpretation, but it is not sufficient, by itself, to ensure robust taxonomy classification across all categories.

Table 6. F_1 per category for the *Taxonomy Disinformation* task under the full multimodal configuration.

Model	Disinformation Category					Macro-F_1
	Descontextu-alização	*Fake News*	*Não se aplica*	*Extremismo antidemocrático*	*Discurso de ódio*	
`gemma3-12b`	0.603	0.429	0.608	0.524	0.424	0.518
`gpt-4.1-mini`	0.500	0.323	0.633	0.421	0.483	0.472
`gpt-4o-mini`	0.543	0.174	0.644	0.444	0.545	0.470
`qwen2.5-14b`	0.584	0.261	0.625	0.467	0.286	0.445
`deepseek-r1-32b`	0.261	0.087	0.583	0.267	0.438	0.327
`gemma2-9b`	0.458	0.095	0.222	0.308	0.545	0.326
`llama3.2-3b`	0.390	0.359	0.392	0.143	0.154	0.287
`qwen2.5-7b`	0.388	0.295	0.493	0.000	0.167	0.269
`deepseek-r1-8b`	0.222	0.214	0.416	0.000	0.400	0.250
`llama3.1-8b`	0.468	0.080	0.205	0.105	0.350	0.242
`qwen2-7b`	0.420	0.133	0.391	0.000	0.267	0.242
`phi3-14b`	0.056	0.298	0.311	0.000	0.000	0.133

Table 6 confirms that taxonomy classification remains challenging across all models. The best result was obtained by `gemma3-12b`, with macro-$F_1 = 0.518$, followed by `gpt-4.1-mini` (0.472) and `gpt-4o-mini` (0.475). Category-level scores reveal an uneven pattern: *não se aplica* and *descontextualização* were generally easier to detect, whereas *fake news*, *discurso de ódio*, and especially *extremismo antidemocrático* remained harder for several models. Some systems completely failed on specific categories, producing $F_1 = 0.000$, which indicates that fine-grained discrimination is still fragile in this setting. A notable outlier is `phi3-14b`, which achieved the highest audio-only taxonomy score (macro-$F_1 = 0.450$) but collapsed to the lowest under the multimodal configuration (0.137, -31.3 pp); this sharp regression likely reflects context-length sensitivity, as the addition of visual descriptions substantially increases prompt size for a model using the coarser `Q4_0` quantization scheme.

Beyond classification accuracy, the two-step prompting strategy supports qualitative auditability: the intermediate factual extraction step makes the evidence considered by

the model more explicit, particularly when the decision depends on contextual mismatch between spoken content and visual evidence. This structure does not provide a causal explanation of the model's internal behavior, but it makes the outputs easier to inspect by human fact-checkers. Nevertheless, the results should be interpreted with caution. The dataset is small, operationally collected, and label-imbalanced, which limits broad statistical generalization; the taxonomy task is especially sensitive to this constraint, given the rarity and semantic overlap of some categories. The present findings are therefore best understood as an exploratory benchmark for Brazilian Portuguese political short videos, showing that multimodal prompting is promising but still far from a definitive solution.

6. Conclusions and Future Work

This work presented a multimodal benchmark for disinformation detection in Brazilian Portuguese political short videos and evaluated a zero-shot LLM-centered pipeline combining speech transcription, speaker diarization, keyframe selection, and visual-language descriptions. Overall, the results indicate that textualized multimodal evidence improves performance over audio-only inference in most cases, especially for binary veracity classification, suggesting that visual cues provide complementary information for interpreting politically sensitive audiovisual content.

Whisper `large-v3` achieved the best overall downstream performance among the transcription models and was therefore adopted in the full multimodal setup. For binary veracity classification, `gemma2-9b` obtained the highest macro-F_1 (0.764), followed by competitive results from `gpt-4.1-mini` and `llama3.1-8b`. For taxonomy classification, `gemma3-12b` achieved the best macro-F_1 (0.518), although fine-grained categorization remained substantially more challenging than binary prediction due to class imbalance, semantic overlap among categories, and the limited number of examples for some labels.

The proposed two-step prompting strategy also supports qualitative auditing by separating factual extraction from final classification, making model outputs easier to inspect in politically sensitive fact-checking scenarios. However, the dataset is relatively small, operationally collected, and label-imbalanced, so the results should be interpreted as an exploratory benchmark rather than evidence of broad generalization. Future work should expand the dataset, improve coverage of underrepresented categories, incorporate stronger statistical validation, compare against supervised textual and multimodal baselines, and investigate few-shot prompting, alternative visual summarization strategies, and confidence estimation for human-in-the-loop verification.

Acknowledgments

This work was supported by CAPES (grant number 88887.671481/2022-00), Anatel (National Telecommunications Agency) and TRE-GO (Regional Electoral Court of Goiás).

Generative AI disclosure. Large language model tools, including Claude (Anthropic) and ChatGPT (OpenAI), were used for linguistic revision and support in source-code development. All scientific content, results, references, and interpretations were verified and remain the sole responsibility of the human authors.

References

Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., and Patras, I. (2021). Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16:345–379.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. (2025). Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Bouchey, B., Castek, J., and Thygeson, J. (2021). Multimodal learning. *Innovative learning environments in STEM higher education: Opportunities, challenges, and looking forward*, pages 35–54.

Bu, Y., Sheng, Q., Cao, J., Qi, P., Wang, D., and Li, J. (2024). FakingRecipe: Detecting fake news on short video platforms from the perspective of creative process. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, pages 1351–1360. Accessed: 2026-03-28.

Dvornik, M., Hadji, I., Derpanis, K. G., Garg, A., and Jepson, A. (2021). Drop-dtw: Aligning common signal between sequences while dropping outliers. *Advances in Neural Information Processing Systems*, 34:13782–13793.

Gôlo, M. P. S., Mori, A. L. V., Oliveira, W. G., Barbosa, J. R., Graciano Neto, V. V., Lima, E. A. d., and Marcacini, R. M. (2024). On the use of large language models to detect brazilian politics fake news. *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*.

Hua, J., Cui, X., Li, X., Tang, K., and Zhu, P. (2023). Multimodal fake news detection through data augmentation-based contrastive learning. *Applied Soft Computing*, 136:110125.

Jing, J., Wu, H., Sun, J., Fang, X., and Zhang, H. (2023). Multimodal fake news detection via progressive fusion networks. *Information Processing & Management*, 60:103120.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Ren, S., Liu, Y., Zhu, Y., Bing, W., Ma, H., and Wang, W. (2024). MMSFD: Multi-grained and multi-modal fusion for short video fake news detection. In *Proceedings of the 7th International Conference on Data Science and Information Technology (DSIT 2024)*, pages 1–11. Accessed: 2026-03-28.

Rodrigues, T. M., Bonone, L., and Mielli, R. (2020). Desinformação e crise da democracia no brasil: é possível regular fake news. *Confluências— Revista Interdisciplinar de Sociologia e Direito*, 22(3):30–52.

Shang, L., Kou, Z., Zhang, Y., and Wang, D. (2021). A multimodal misinformation detector for COVID-19 short videos on tiktok. In *Proceedings of the 2021 IEEE International Conference on Big Data (Big Data)*, pages 899–908. Accessed: 2026-03-28.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Wardle, C. and Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

Wu, P., Zhan, Y., Zhang, L., Wang, Z., and Xu, Z. (2021). Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2560–2569. Association for Computational Linguistics.