KinyaProp: Fine-Grained Propaganda Annotation in Kinyarwanda

Anonymous ACL submission

Abstract

Propaganda is a widely used approach for shaping public opinion and disseminating misinformation in news media. While it has gained significant attention within the Natural Language Processing (NLP) community, research on fine-grained propaganda detection remains concentrated in high-resourcelanguages. To bridge this gap, we introduce KinyaProp, a fine-grained propaganda dataset in Kinyarwanda, a low-resource Bantu language spoken in Rwanda. Using this dataset, we evaluate whether state-of-the-art Large Language Models (LLMs) can function as reliable annotators in a genuinely low-resourceand culturally grounded setting. Results show current multilingual LLMs do not reliably approximate human annotation behavior. Instead, they behave as conservative annotators whose performance is largely limited to lexically explicit cues, substantially under-identifying propaganda and exhibiting extremely low and unstable performance on discourse-level techniques. Our findings highlight an important limitation of recent successes in LLM based annotation reported for high-resourcelanguages, demonstrating that such results do not readily transfer to low-resourcesettings, where scalable annotation would be most valuable. We release KinyaProp to support future research on fine-grained propaganda detection and to enable more robust evaluation of multilingual models in underrepresented languages.

1 Introduction

Propaganda increasingly shapes online media, using emotionally charged or selectively framed content to influence public opinion across social and political domains (Stanley, 2015; Alam et al., 2022a; Sharma et al., 2022; Da San Martino et al., 2020). Rather than relying on overtly deceptive language, propaganda often operates through rhetoric and psychological manipulation strategies that feel reasonable, indirectly steering readers toward incomplete, inaccurate, or distorted interpretations without their awareness. Detecting and analyzing such influence is therefore essential. Yet the study of fine-grained propaganda remains largely confined to a small number of high-resource languages, most notably English, Arabic, and Chinese (Da San Martino et al., 2019; Alam et al., 2022b; Shih et al., 2020).

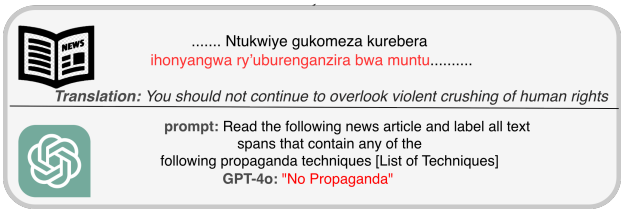

Figure 1: An example Kinyarwanda news excerpt that contains loaded language propaganda which GPT-4o fails to identify.

Such narrow evaluations risk implicitly treating propaganda as uniform across languages and societies. However, propaganda is not a language-agnostic phenomenon. Persuasive strategies are shaped by cultural norms, rhetorical conventions, and language-specific realizations of meaning. Cross-cultural studies show that the explicitness, framing, and stylistic markers of persuasion vary substantially across cultures and do not transfer reliably across languages (Shen, 2023). These properties suggest that evaluating LLMs as annotators of fine-grained propaganda requires testing them in settings where linguistic and cultural divergence from their training data is greatest, as exemplified in Figure 1. Low-resource languages thus are not only important application targets, but also stress tests for current claims about LLM-based annotation.

We demonstrate the value of this perspective through a study of fine-grained propaganda in Kinyarwanda, an official language of Rwanda spoken

by more than 18 million people in Central and Eastern Africa. Prior computational work on Kinyarwanda has focused primarily on morphology, linguistic modeling, machine translation, and small pretrained models (Nzeyimana, 2020; Nzeyimana and Niyongabo Rubungo, 2022; Nzeyimana, 2024), with misinformation and propaganda receiving almost no attention in this language. To our knowledge, this is the **first study to examine LLM-based propaganda detection in Kinyarwanda, and the first fine-grained propaganda dataset for any Bantu language**, a family comprising over 500 languages spoken across much of sub-Saharan Africa.

KinyaProp consist of 608 news articles annotated at span level with 18 fine-grained propaganda techniques proposed in (Da San Martino et al., 2019) and annotated by trained native speakers. We assess annotation quality and construct a high-quality gold subset through a majority vote. Using this dataset, we conduct a diagnostic analysis of whether current state-of-the-art multilingual large language models can reliably annotate fine-grained propaganda in this language. In contrast to a large number of prior evaluations conducted in high-resource languages, our study examines not only overall performance but also how model behavior varies across different types of propaganda techniques. Our main contributions are as follows:

- We introduce KINYAPROP, the first fine-grained propaganda dataset for Kinyarwanda and to our knowledge the first such resource for any Bantu language.
- We provide systematic analysis of LLM–human annotation alignment for fine-grained propaganda detection and show that current multilingual LLMs do not reliably approximate human annotator behavior.
- We demonstrate that multilingual LLMs act as conservative, lexically driven annotators, substantially under-identifying propaganda and failing on discourse-level techniques even under ICL, SFT, and CoT prompting.

2 Related Work

2.1 Fine-grained propaganda detection

In NLP, propaganda detection has traditionally been studied at the article level (Rashkin et al., 2017; Barrón-Cedeño et al., 2019). More recent work has shifted toward the more challenging task of fine-grained detection, which identifies specific propaganda techniques within text and provides the supervision needed to train robust models for real-world propaganda detection. (Da San Martino et al., 2019; Alam et al., 2022b). These fine-grained propaganda datasets have supported a wide range of modeling approaches, including neural sequence labeling, knowledge-infused methods, and discourse-aware architectures, forming the basis for most prior work on fine-grained propaganda detection (Wang et al., 2020; Vijayaraghavan and Vosoughi, 2022; Lei and Huang, 2023).

2.2 LLMs as annotators

Recent studies have examined whether LLMs can act as annotators for complex NLP tasks, motivated by their speed, cost efficiency, and in some cases competitive agreement with human annotators (Hasanain et al., 2024a,b; Calderon et al., 2025; Shah et al., 2024; Mohta et al., 2023). LLMs have been used both to replace individual annotators and to consolidate noisy human labels, and hybrid annotation pipelines have been proposed to further increase reliability (Sahitaj et al., 2025; Hasanain et al., 2024b). While the use of LLMs as annotators is an active area of research within NLP community, existing evaluations and benchmarks are largely conducted in high-resource languages. This leaves genuinely low-resource settings largely unexplored, despite being precisely the contexts where automated annotation would be most valuable, due to severe constraints on cost, access to skilled annotators, and linguistic expertise. Our work directly addresses this gap by evaluating LLM-based annotation for fine-grained propaganda detection in Kinyarwanda.

3 Dataset

KinyaProp consists of 608 news articles sampled from the corpus of (Niyongabo et al., 2020), which aggregates real news content from multiple Rwandan outlets. The articles span diverse topical domains including politics, health, religion, culture, entertainment, and economics, reflecting naturally occurring journalistic discourse and a wide range of persuasive strategies in real-world usage.

Annotations follow the 18 propaganda techniques schema of (Da San Martino et al., 2019). Propaganda instances are marked as fine-grained contiguous text spans, allowing multiple and over-

Technique	# Gold Spans
Appeal to Authority	717
Loaded Language	417
Repetitions	256
Exaggeration or Minimization	69
Causal Oversimplification	68
Flag-waving	52
Appeal to Fear/Prejudice	40
Slogans	28
Obfuscation/Vagueness/Confusion	27
Name Calling or Labeling	26
Bandwagon	25
Black-and-White Fallacy/Dictatorship	19
Doubt	18
Red Herring	9
Thought-Terminating Cliché	2
Whataboutism	1

Table 1: Distribution of propaganda technique spans in the majority vote gold subset.

lapping instances within a single article.

Each article was independently annotated by three university students who are native speakers of Kinyarwanda and fluent in English. Prior to full annotation, we conducted a multi-stage training and calibration phase. Annotators first completed guided training covering the full schema, followed by multiple practice rounds on held-out articles. Initial practice rounds indicated inconsistencies between several closely related techniques (e.g., straw man and bandwagon). We addressed this by providing additional examples and extending the training period, and required annotators to demonstrate stable and consistent use of span boundaries and technique labels before proceeding to full annotation. Full annotation began only after all annotators demonstrated sufficient mastery of the schema, ensuring that disagreement in the final dataset reflects the inherent subjectivity of the task rather than lack of familiarity with the guidelines. The entire annotation process was conducted over a period of approximately three months.

We assess annotation reliability using the γ agreement metric (Mathet et al., 2015), which is designed for span-based tasks involving overlapping annotations. We obtain $\boldsymbol{\gamma_s = 0.544}$ when considering span boundaries only and $\boldsymbol{\gamma_{sl} = 0.548}$ when jointly considering spans and technique labels. The close correspondence between these values indicates that disagreement is primarily due to span boundary selection rather than label confusion; once annotators identify a span, they generally agree on the underlying propaganda technique. Our agreement values are comparable to prior work on fine-grained propaganda annotation. For example, (Da San Martino et al., 2019) report pre-consolidation agreement of $\gamma_s = 0.34$ and $\gamma_{sl} = 0.31$, while more recent studies report γ_s values in the range of 0.55–0.60 (Hasanain et al., 2024a; Da San Martino et al., 2020).

To construct a high-precision gold subset for benchmarking and fine-tuning, we aggregate annotations via majority vote. Spans from different annotators are aligned by clustering those that substantially overlap, using a one-dimensional Intersection-over-Union (IoU) threshold of 0.5, which requires that at least half of the combined span length be shared. Within each cluster, we retain at most one span per annotator and assign a gold label only when at least two annotators agree. The final gold span is defined as the minimal interval covering the agreeing annotations.

Applying this procedure yields 1,774 gold propaganda spans. In total, 582 of the 608 articles are represented in the high-precision gold subset: 526 articles (86.51%) contain at least one gold propaganda span, while 56 articles contain no gold spans and are explicitly labeled as non-propaganda at the article level. The remaining 26 articles contain annotated candidate spans that did not reach majority agreement and are therefore excluded from the gold subset. Table 1 reports the distribution of propaganda technique spans in the gold subset, which exhibits the long-tailed structure characteristic of fine-grained persuasion phenomena. This gold subset forms the basis for the ICL and SFT experiments described in the remainder of the paper.

Our KinyaProp dataset, including both the raw annotations and the majority vote gold subset, is available in our anonymous GitHub repository: https://github.com/KinyaProp/KinyaProp

4 Experimental Setup

We investigate the performance of multilingual large language models on fine-grained propaganda span detection in a genuinely low-resource language such as Kinyarwanda. Our objective is to assess whether such models can approximate trained human annotators under realistic annotation conditions. We first establish supervised baselines using multilingual pretrained encoders fine-tuned on our gold dataset (Section 3), and then evaluate state-of-the-art multilingual large language models using human-mirroring annotation setups and a set of targeted prompting strategies commonly used in prior annotation-focused studies to analyze model

behavior and test whether additional supervision narrows the gap to human annotations.

4.1 Models

LLMs: Across our experiments, we use GPT-4o, GPT-4.1, and Claude Sonnet 4.5 under zero-shot, in-context learning, and supervised fine-tuning settings (supervised fine-tuning is applied only to GPT-4o). We select these models as representative state-of-the-art multilingual large language models that have been widely used in recent work on LLM-based annotation and information extraction.

PLMs: To establish supervised baselines using conventional encoder-based architectures, we fine-tune XLM-R-large (Conneau et al., 2020) and AfroXLM-R-large (Alabi et al., 2022). XLM-R is a standard multilingual encoder trained via unsupervised cross-lingual representation learning, while AfroXLM-R augments XLM-R with additional pretraining on African language data. Including both models allows us to compare general multilingual representations with an Africa-focused variant that has previously shown improved performance on African languages, including Kinyarwanda.

4.2 Supervised fine-tuning

All supervised models are trained on the majority-vote gold subset from Table 1 using a $70\%/15\%/15\%$ train/validation/test split. Extremely rare techniques are restricted to the training split to avoid unstable evaluation.

For GPT-4o, we perform supervised fine-tuning using an instruction-following formulation. Each training instance consists of an (article, technique) pair, with the target defined as the list of gold spans expressing that technique; when a technique is absent, the target is an empty list. This yields 2,035 training examples, 428 validation examples, and 434 test examples. Fine-tuning is performed for one epoch with a batch size of 16 and a learning-rate multiplier of 0.5. Predictions follow the same format used in the zero-shot experiments.

For encoder-based baselines, we fine-tune XLM-R-large and AfroXLM-R-large as token-level sequence labeling models using BIO tags derived from gold spans. Models are trained for four epochs with a learning rate of 1×10^{-5} and a per-device batch size of 8. All models are evaluated on the held-out test set using the same span-level metrics described in Section 5.

4.3 Instructions

Our prompt format closely follows prior work on LLM-based annotation in both structure and output format (Hasanain et al., 2024a). As our primary baseline, we use zero-shot multi-label span annotation, which most directly mirrors the human annotation task by requiring models to identify all propaganda techniques present in an article from the full schema. We observe that models occasionally predict the correct span text but assign slightly incorrect character offsets; following prior work (Hasanain et al., 2024a), we apply a lightweight post-processing step that aligns predicted spans to their first exact string match in the article to reduce offset noise without altering span content or labels.

For in-context learning, we adopt a single-label formulation in which the model is conditioned on one propaganda technique at a time. While less common in prior work, this design enables controlled analysis of technique-specific behavior and reduces label competition in highly imbalanced, low-resource settings where rare techniques are often missed in multi-label prompts.

4.4 Prompt Setup

We evaluate models under three prompting configurations that reflect common annotation practices in prior work. In all settings, models are instructed to identify propaganda techniques and their corresponding character-level spans in Kinyarwanda news articles.

Zero-shot Multi-label Annotation Models are prompted to identify all propaganda techniques present in an article from the full set of 18 techniques and to return the corresponding text spans. This setting most directly mirrors the human annotation task and serves as our primary baseline for analyzing alignment between LLM and human annotations.

Zero-shot Multi-label Annotation with CoT We additionally evaluate a zero-shot variant in which models are encouraged to reason step by step about the presence of each technique prior to producing span-level predictions. The generated reasoning is discarded, and only the predicted spans and offsets are evaluated.

In-Context Learning Single-label Annotation For in-context learning, models are prompted to annotate one propaganda technique at a time, with zero to three labeled exemplars provided in the

Setting	Shots	Annotation Formulation
Zero-shot	0	Multi-label
Zero-shot + CoT	0	Multi-label
ICL (single-label)	0–3	Single-label

Table 2: Prompting setups evaluated in experiments

Metric	GPT-4o	GPT-4.1	Claude Sonnet 4.5
Micro P	0.031	0.028	0.072
Micro R	0.020	0.031	0.051
Micro F1	0.024	0.029	0.060
Macro P	0.017	0.014	0.037
Macro R	0.007	0.011	0.022
Macro F1	0.006	0.007	0.022

Table 3: Zero-shot multi-label span-level propaganda detection performance.

prompt. Exemplars are sampled using a fixed random seed, with exemplar and evaluation instances drawn from disjoint subsets. For extremely rare techniques where such a split is not possible, results are treated as non-robust. This single-label formulation enables controlled analysis of technique-specific behavior and mitigates label competition under the highly imbalanced label distribution characteristic of low-resource settings. Table 2 summarizes the prompting configurations used across experiments.

5 Evaluation

We evaluate model performance using two complementary approaches: standard system-level metrics and annotator-level alignment analysis.

5.1 Standard System Evaluation

Across all experiments, we compute span-level precision, recall, and F1 scores with both macro and micro averaging. We use a modified span-level F1 metric that accounts for partial overlap between predicted spans and gold annotations, reflecting the graded nature of span matching in fine-grained propaganda annotation rather than requiring exact boundary agreement (Da San Martino et al., 2019). This metric is standard in prior work on fine-grained propaganda detection.

5.2 Annotator-Level Evaluation

For zero-shot multi-label experiments, which most closely resembles the human annotation setup, we additionally evaluate alignment between LLM and human annotations using Advantage Probability (AP). AP measures the probability that an LLM

Model	P	R	$F1_M$	$F1_\mu$
GPT-4o	0.117	0.129	0.116	0.168
AfroXLM-R-large	0.073	0.103	0.055	0.230
XLM-RoBERTa-large	0.019	0.052	0.027	0.210

Table 4: Supervised fine-tuning performance on the gold test set.

Technique	Best F1
Top-performing techniques	
Slogans	0.800
Flag Waving	0.800
Appeal to Authority	0.600
Bandwagon	0.581
Thought-Terminating Cliché	0.574
Lowest-performing Techniques	
Whataboutism	0.000
Repetitions	0.000
Black-and-White Fallacy / Dictatorship	0.000
Obfuscation / Vagueness / Confusion	0.061
Red Herring	0.500

Table 5: Best achieved F1 score per propaganda technique across all models and in-context learning settings.

matches or exceeds the performance of a human annotator, estimated by averaging over all articles annotated by that annotator (Calderon et al., 2025). This metric allows us to assess whether LLMs can function as reliable substitutes for human annotators under realistic annotation conditions.

6 Results and Discussion

6.1 Zero-shot Span-level Propaganda Detection Performance

We first evaluate model performance under the zero-shot multi-label annotation setting described in Section 4.4. Table 3 reports span-level precision, recall, and F1 scores with both micro and macro averaging.

As shown in Table 3, all evaluated models exhibit extremely limited effectiveness under zero-shot multi-label annotation. Micro F1 scores remain below 0.07 across models, while macro F1 scores approach zero, indicating a near-complete failure to identify infrequent propaganda techniques. Among the evaluated models, Claude Sonnet 4.5 achieves the highest scores, but performance remains low overall (micro F1 = 0.06, macro F1 = 0.022). Across models, recall is consistently much lower than precision, indicating a systematic tendency toward conservative prediction: models recover only a small fraction of gold propaganda spans while rarely over-predicting.This asymme-

try holds under both micro and macro averaging and reflects pervasive under-identification rather than boundary noise or label confusion. Near-zero macro recall further shows that rare, long-tailed propaganda techniques are almost entirely missed in the zero-shot setting.

6.2 Performance Analysis under Fine-tuning

We next examine whether supervised fine-tuning mitigates the failure modes observed under zero shot annotation. Table 4 reports supervised fine-tuning results on the gold test set. Supervision improves aggregate performance for all models, but the magnitude and distribution of these gains vary substantially across model families and propaganda techniques.

GPT-4o achieves the highest macro F1 score, reflecting improved sensitivity to infrequent techniques under supervision. In contrast, AfroXLM R large and XLM R large attain higher micro F1 scores, indicating stronger performance on frequent techniques and corpus wide span identification. This pattern is consistent with the limited size of the gold training data and the long tailed label distribution. Across models, gains from supervision are driven primarily by increased recall. This implies that fine-tuning enables models to recover a larger fraction of propaganda spans overall.

However, a per technique analysis reveals that these improvements are highly uneven. Fine tuning yields moderate to strong performance for a small subset of relatively frequent or lexically salient techniques, most notably for GPT-4o, which achieves F1 scores of 0.314 on *Appeal to Authority*, 0.224 on *Appeal to Fear or Prejudice*, and 0.731 on *Doubt*. In contrast, many techniques including *Bandwagon*, *Black and White Fallacy*, *Red Herring*, *Slogans*, and *Obfuscation Vagueness Confusion* remain entirely undetected despite supervision. Techniques that rely on discourse structure or implicit argumentation continue to exhibit zero recall after fine-tuning, indicating that supervision improves coverage only for a narrow subset of the annotation schema rather than addressing the core reasoning challenges of fine-grained propaganda detection.

We next evaluate whether in-context learning alters this failure pattern. Under the single label in-context learning setup, overall performance improves substantially relative to zero shot prompting. In the best configuration, models achieve macro F1 equal to 0.298 and micro F1 equal to 0.272, exceeding both zero shot and supervised baselines. Under this setting, GPT models consistently outperform Claude Sonnet 4.5, despite Claude achieving stronger results in the zero shot multi label setting.

Gains under in-context learning remain highly technique dependent. As shown in Table 5, techniques with explicit lexical or surface level cues, such as *Slogans* and *Flag Waving*, benefit most from the inclusion of examples and achieve the highest F1 scores across settings. In contrast, techniques that depend on discourse level context or implicit reasoning show little to no improvement even when provided with up to three labeled in context examples. Techniques such as *Whataboutism* and *Black and White Fallacy* are never detected in any shot configuration. Other techniques, including *Red Herring*, exhibit isolated gains in a single configuration but collapse to zero performance in all other few shot configurations, indicating unstable and non-generalizable behavior.

These findings demonstrate a consistent pattern across supervision and prompting regimes. Multilingual large language models primarily rely on lexical surface cues when identifying propaganda. While both supervised fine-tuning and in-context learning improve performance for techniques with explicit lexical markers, neither enables reliable detection of techniques that require discourse level interpretation or argumentative reasoning. Even with multiple in context examples, models fail to generalize to reasoning based propaganda techniques. These findings indicate that current multilingual large language models function effectively as lexical propaganda annotators, but do not approximate human annotation behavior for discourse level persuasion in low-resourcesettings.

6.3 LLM–Human Annotation Agreement

The preceding analyses quantify model performance using standard system-level metrics and reveal substantial limitations of multilingual LLMs, particularly for discourse-level propaganda techniques. However, low absolute performance alone does not fully resolve a more practically relevant question: whether LLMs might nevertheless function as *viable substitutes* for trained human annotators under realistic annotation conditions. Fine-grained propaganda annotation is inherently subjective, and individual human annotators exhibit non-trivial variability. A model need not perfectly match a gold standard to be useful; rather, it must fall within the range of disagreement observed among human annotators.

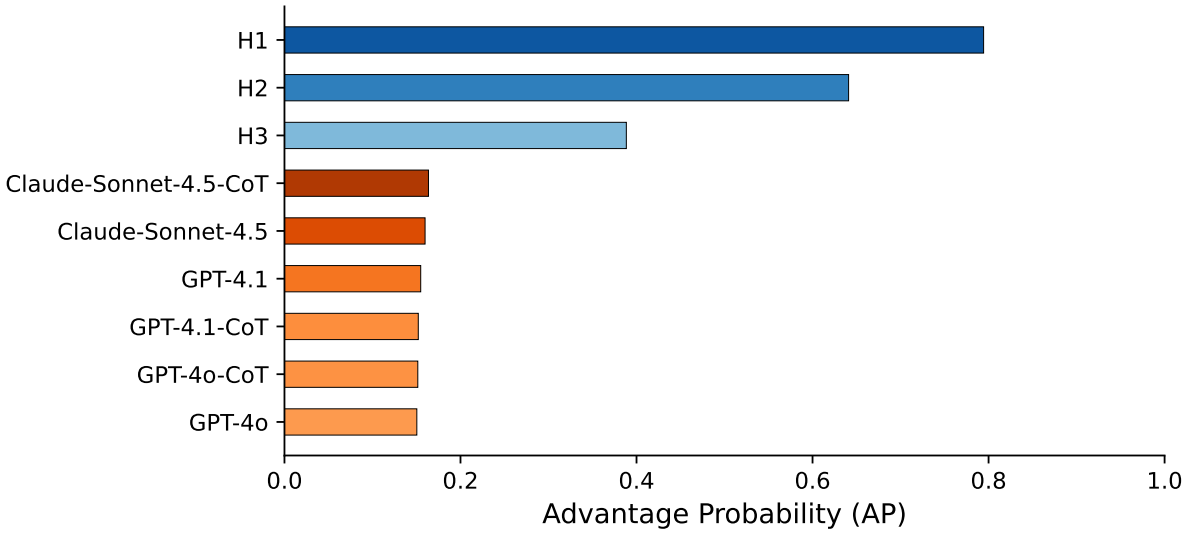

Figure 2: Advantage Probability for human annotators and LLMs

To directly assess this criterion, we evaluate alignment between LLMs and individual human annotators using Advantage Probability (AP), following the methodology described in Section 5. AP estimates the probability that a model matches or exceeds the performance of a given human annotator at the article level, providing a human-centered measure of annotator substitutability rather than system-level accuracy alone. We conduct this analysis under the zero-shot multi-label setting, which most closely mirrors the original human annotation conditions.

Figure 2 reports AP scores for each human annotator and each evaluated LLM. Human annotators exhibit substantial variability, with AP values ranging from 0.388 to 0.794, reflecting the inherent subjectivity of span selection and technique identification in fine-grained propaganda annotation.

In contrast, all evaluated LLMs achieve markedly lower AP scores, ranging from 0.150 to 0.164, with minimal variation across architectures. In every case, model AP scores fall below those of the least-aligned human annotator. Under identical annotation conditions, LLM predictions diverge more from individual human annotations than human annotators diverge from one another. This gap places current multilingual LLMs outside the envelope of human annotator variability required for practical substitutability.

We further evaluate whether chain-of-thought prompting improves annotator-level alignment. As shown in Figure 2, chain-of-thought yields only marginal AP increases for GPT-4o and Claude Sonnet 4.5 and no improvement for GPT-4.1. In all cases, these changes are insufficient to alter the relative ordering between models and human annotators. Even when encouraged to reason explicitly, models do not approach the level of agreement exhibited by any trained human annotator.

These findings provide strong evidence that current multilingual LLMs do not meet the standard required to replace human annotators for fine-grained propaganda detection in Kinyarwanda under the evaluated conditions. Their predictions consistently fall outside the range of human disagreement, indicating that recent successes of LLM-based annotation reported in high-resource languages do not readily transfer to low-resource and culturally grounded settings.

6.4 Qualitative Error Analysis

We conduct a qualitative error analysis focused on *Loaded Language*, one of the most frequent propaganda techniques in both news discourse and KINYAPROP. Despite its prevalence, all evaluated models exhibit consistently low performance on this category. To better understand the source of these errors, we qualitatively compare LLM predictions with human annotations.

Across models, we observe a systematic mismatch in span selection granularity. LLMs predominantly predict single-token or very short spans, whereas human annotators typically select phrase- or clause-level spans that express evaluative or persuasive intent compositionally. As a result, models often identify isolated lexical items without captur-

ing the broader contextual framing that determines whether language functions propagandistically.

An inspection of false positive predictions further reveals that models frequently flag lexically salient but pragmatically neutral terms, such as *abatishoboye* ("the poor"), *batizanya ingufu* ("they support one another"), and *igitsinagore* ("woman"). While such expressions may appear emotionally loaded or evaluative in isolation, they are common and socially unmarked in everyday Kinyarwanda usage and do not, in context, convey persuasion, judgment, or emotional manipulation. Human annotators consistently exclude such spans, relying on pragmatic intent rather than surface lexical intensity.

This pattern is particularly observed for GPT-4o, which produces over 25 false positive predictions involving inflected forms of *kwica* ("to kill"), none of which are annotated as propaganda by any human annotator. In these cases, the model responds to lexical salience while failing to account for discourse-level intent or conventionalized usage. These errors indicate that multilingual LLMs tend to treat *Loaded Language* primarily as a surface lexical phenomenon, leading to systematic over-triggering on some culturally conventional expressions. This behavior is likely exacerbated by Kinyarwanda's morphological richness and idiomatic usage (Section D), which require pragmatic and cultural grounding to interpret correctly. This pattern shows a core limitation of propaganda detection approaches driven primarily by lexical matching instead of discourse-level interpretation.

7 Conclusion and Future Work

This paper introduces **KinyaProp**, the first fine-grained propaganda detection dataset for Kinyarwanda, and uses it to evaluate whether state-of-the-art multilingual large language models can substitute for trained human annotators in a genuinely low-resource setting. Across zero-shot, in-context, and supervised conditions, we find consistent evidence that current multilingual LLMs operate as conservative, lexically driven annotators. While models can identify a subset of explicitly lexical propaganda techniques, they systematically fail to detect discourse-level and reasoning-based techniques that require contextual and argumentative interpretation. These limitations persist even with supervised fine-tuning and carefully designed prompting, indicating a structural constraint rather than insufficient exposure or prompt design.

These findings establish an important boundary on recent claims that LLMs can reliably replace human annotators for subjective NLP tasks. Evidence derived primarily from high-resource languages does not readily transfer to low-resource and culturally grounded contexts, which instead serve as diagnostic stress tests that reveal limitations masked in English-centric evaluations. At the same time, our results suggest a principled role for LLMs within hybrid annotation pipelines, where models assist with lexically salient techniques while human annotators handle discourse-dependent judgments. By releasing KinyaProp as a public resource, we aim to support more realistic evaluation of multilingual models and clarify the capabilities and limitations of current LLM-based annotation approaches in low-resource settings.

8 Limitations

We recognize that KinyaProp has several limitations, which we summarize below.

Language coverage. KinyaProp is limited to a single language, Kinyarwanda, and while it is, to our knowledge, the first fine-grained propaganda dataset for any Bantu language, it does not capture the full diversity of persuasive strategies used across Bantu or other low-resource languages.

Dataset size and label imbalance. Although KinyaProp required substantial annotation effort, its overall size remains modest compared to datasets in high-resource languages. The distribution of propaganda techniques is highly imbalanced, with several labels appearing very infrequently.

Data annotation subjectivity. Fine-grained propaganda annotation involves interpretive judgment and is inherently subjective. While we mitigated this through multi-stage annotator training, calibration, and agreement analysis, some degree of annotation noise is unavoidable. The low-resource nature of the language and the limited availability of expert annotators further constrained the annotation process, and some annotations may remain imperfect.

9 Ethics and Broader Impact

The KinyaProp dataset contains uncensored real-world news content, including potentially sensitive material such as hate speech, violence, and sexual content. While the dataset is intended to support

research on fine-grained propaganda detection, we recognize the potential for misuse, such as developing adversarial techniques that evade detection or training models that generate propagandistic content. We encourage responsible use and careful consideration of potential downstream harms.

Acknowledgements

We thank the Label Studio platform for providing free access to their annotation tools through the Academic Program. We are also grateful to the annotators for their time and effort in creating this dataset.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022a. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022b. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing Management*, 56.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8440–8451, Online. Association for Computational Linguistics.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024a. Can GPT-4 identify propaganda? annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2724–2744, Torino, Italia. ELRA and ICCL.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024b. Large language models for propaganda span annotation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14522–14532, Miami, Florida, USA. Association for Computational Linguistics.

Yuanyuan Lei and Ruihong Huang. 2023. Discourse structures guided fine-grained propaganda identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 331–342, Singapore. Association for Computational Linguistics.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Jay Mohta, Kenan Ak, Yan Xu, and Mingwei Shen. 2023. Are large language models good annotators? In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 38–48. PMLR.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. 2020. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In *Proceedings of the*

28th International Conference on Computational Linguistics, pages 5507–5521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Antoine Nzeyimana. 2020. Morphological disambiguation from stemming data. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4649–4660, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Antoine Nzeyimana. 2024. Low-resource neural machine translation with morphological modeling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 182–195, Mexico City, Mexico. Association for Computational Linguistics.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Ariana Sahitaj, Premtim Sahitaj, Veronika Solopova, Jiaao Li, Sebastian Möller, and Vera Schmitt. 2025. Hybrid annotation for propaganda detection: Integrating LLM pre-annotations with human intelligence. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 215–228, Vienna, Austria. Association for Computational Linguistics.

Uzair Shah, Md. Rafiul Biswas, Marco Agus, Mowafa Househ, and Wajdi Zaghouani. 2024. MemeMind at ArAIEval shared task: Generative augmentation and feature fusion for multimodal propaganda detection in Arabic memes through advanced language and vision models. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 467–472, Bangkok, Thailand. Association for Computational Linguistics.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 5597–5606, Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization.

Lin Shen. 2023. Culture and explicitness of persuasion: Linguistic evidence from a 51-year corpus-based cross-cultural comparison of the united nations general debate speeches across 55 countries (1970-2020). *Cross-Cultural Research*, 57(2-3):166–192.

Meng-Hsien Shih, Ren-feng Duann, and Siaw-Fong Chung. 2020. The analysis and annotation of propaganda techniques in Chinese news texts. In *Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)*, pages 331–345, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Jason Stanley. 2015. *How Propaganda Works*. Princeton University Press.

Prashanth Vijayaraghavan and Soroush Vosoughi. 2022. TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3433–3448, Seattle, United States. Association for Computational Linguistics.

Ruize Wang, Duyu Tang, Nan Duan, Wanjun Zhong, Zhongyu Wei, Xuanjing Huang, Daxin Jiang, and Ming Zhou. 2020. Leveraging declarative knowledge in text and first-order logic for fine-grained propaganda detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3895–3903, Online. Association for Computational Linguistics.

Strategy	Prompt
Single-label	You are an expert in Kinyarwanda news propaganda detection. Read the following news article and label all text spans that contain the following propaganda technique: "{single label}". Answer exactly and only by returning a JSON list of spans: ["technique": "", "text": "", "start": 0, "end": 0] If none exist, return: ["technique":"no propaganda", "text": "", "start": 0, "end": 0] Spans may overlap.
Multi-label	You are an expert in Kinyarwanda news propaganda detection. Read the following news article and label all text spans that contain any of the following propaganda techniques: {ALL TECHNIQUE_LIST}. Answer exactly and only by returning a JSON list of spans: ["technique": "", "text": "", "start": 0, "end": 0] If none exist, return: ["technique":"no propaganda", "text": "", "start": 0, "end": 0] Spans may overlap.
CoT	You are an expert in Kinyarwanda news propaganda detection. Read the following news article and label all text spans that contain any of the following propaganda techniques: {TECHNIQUE_LIST}. Think step by step about each propaganda technique in the list: • whether it appears in the article or not, • which exact text span expresses it (if any), • and where that span should start and end in character offsets. After you have thought this through, answer exactly and only by returning a JSON array of objects with this schema: [{"technique": "...", "text": "...", "start": 0, "end": 0, "reasoning": "step-by-step explanation for why this span is labeled this way"}, ...] If there is no propaganda in the article, return: ["technique":"no propaganda", "text": "", "start": 0, "end": 0, "reasoning": ""] Spans may overlap.

Table 6: Different prompts used to instruct models to annotate articles by propaganda techniques and spans.

A Prompt Templates

We build our prompts based on strategies that have been effective in previous work on LLM-based propaganda detection with few modifications. Because fine-grained propaganda annotation is a challenging task, we experiment with multiple prompt variants and select those that work best in our experiments. Table 9 lists the prompt templates used for each task. we use similar prompts across all models.

B Evaluation Metrics

B.1 Macro-F1 and Micro-F1 (Span-Based Soft Overlap)

The F1 score is defined as the harmonic mean of precision and recall. For span-level classification, partial overlaps between predicted and gold spans are rewarded.

Let S be the set of predicted spans and T the set of gold spans in the corpus. Each span $x \in S \cup T$ is associated with a label $l(x)$.

For a predicted span s and a gold span t, the overlap credit is defined as:

$$C(s,t,h) = \frac{|s \cap t|}{h} \cdot \mathbf{1}(l(s) = l(t)).$$

Using this definition, precision and recall are computed as:

$$P = \frac{1}{|S|} \sum_{d} \sum_{s \in S_d} \sum_{t \in T_d} C(s,t,|s|).$$

$$R = \frac{1}{|T|} \sum_{d} \sum_{t \in T_d} \sum_{s \in S_d} C(s,t,|t|).$$

The Micro-F1 score is defined as:

$$\text{Micro-F1} = \frac{2PR}{P+R}.$$

For each class c, class-specific precision P_c and recall R_c are computed using the same equations. The Macro-F1 score is then given by:

$$\text{Macro-F1} = \frac{1}{N} \sum_{c=1}^{N} \frac{2P_c R_c}{P_c + R_c},$$

where N is the number of classes with at least one gold span.

B.2 Advantage Probability (AP)

Advantage Probability (AP) measures the likelihood that a model performs at least as well as a given human annotator at the article level.

Let x_i denote an article and let $S(a, x_i, j)$ be an alignment score between annotator a and a reference annotation for x_i, where the reference is constructed by excluding human annotator h_j. For each article, we define the indicator:

$$W_{i,j}^{f} = \mathbf{1}(S(f, x_i, j) \geq S(h_j, x_i, j)),$$

where f denotes the model under evaluation and $\mathbf{1}(\cdot)$ is the indicator function. The advantage probability of model f over annotator h_j is estimated as:

$$\rho_j^f = \frac{1}{|I_j|} \sum_{i \in I_j} W_{i,j}^{f},$$

where I_j is the set of articles annotated by h_j.

C Detailed Experimental Results

C.1 In-Context Learning

We evaluate performance using few shot prompting under a single-label formulation, following the prompt template shown in Table 9. Model predictions are evaluated using modified F1 metric described in Section B.1. Tables 7, 8, and 9 show performance of Claude-Sonnet-4.5, GPT-4o, and GPT-4.1 models.

Technique	0S	1S	2S	3S
Appeal to Authority	0.500	0.000	0.500	0.000
Flag-waving	0.500	0.000	0.500	0.000
Slogans	0.500	0.800	0.800	0.800
Name Calling or Labeling	0.171	0.151	0.196	0.119
Doubt	0.198	0.000	0.000	0.000
Loaded Language	0.097	0.000	0.000	0.000
Repetitions	0.009	0.000	0.000	0.000
Appeal to Fear/Prejudice	0.000	0.000	0.000	0.000
Bandwagon	0.000	0.000	0.000	0.000
Black-and-White Fallacy/Dictatorship	0.000	0.000	0.000	0.000
Causal Oversimplification	0.000	0.000	0.000	0.000
Exaggeration or Minimization	0.000	0.000	0.000	0.000
Obfuscation/Vagueness/Confusion	0.000	0.000	0.000	0.000
Red Herring	0.000	0.000	0.000	0.000
Thought-Terminating Cliché	0.224	0.574	–	–
Whataboutism	0.000	–	–	–
Macro-F1	**0.137**	**0.102**	**0.143**	**0.066**
Micro-F1	**0.154**	**0.162**	**0.161**	**0.084**

Table 7: Per-technique F1 across shot settings for Claude-Sonnet-4.5.

Technique	0S	1S	2S	3S
Flag-waving	0.000	0.800	0.800	0.667
Red Herring	0.000	0.000	0.500	0.000
Appeal to Fear/Prejudice	0.000	0.352	0.413	0.413
Appeal to Authority	0.399	0.369	0.458	0.540
Causal Oversimplification	0.195	0.000	0.163	0.153
Obfuscation/Vagueness/Confusion	0.000	0.000	0.058	0.061
Slogans	0.800	0.690	0.690	0.741
Exaggeration or Minimization	0.190	0.239	0.240	0.188
Name Calling or Labeling	0.116	0.127	0.151	0.101
Loaded Language	0.058	0.078	0.083	0.082
Doubt	0.111	0.119	0.111	0.057
Bandwagon	0.500	0.500	0.500	0.333
Black-and-White Fallacy/Dictatorship	0.000	0.000	0.000	0.000
Repetitions	0.000	0.000	0.000	0.000
Thought-Terminating Cliché	0.000	0.526	–	–
Whataboutism	0.000	–	–	–
Macro-F1	**0.148**	**0.253**	**0.298**	**0.238**
Micro-F1	**0.157**	**0.220**	**0.272**	**0.239**

Table 8: Per-technique F1 across shot settings for GPT-4o.

Technique	0S	1S	2S	3S
Bandwagon	0.581	0.500	0.000	0.500
Flag-waving	0.463	0.199	0.373	0.371
Appeal to Authority	0.432	0.433	0.429	0.600
Appeal to Fear/Prejudice	0.187	0.352	0.000	0.000
Loaded Language	0.101	0.261	0.000	0.073
Doubt	0.169	0.170	0.282	0.230
Slogans	0.603	0.632	0.667	0.667
Name Calling or Labeling	0.101	0.084	0.101	0.101
Causal Oversimplification	0.209	0.220	0.220	0.198
Obfuscation/Vagueness/Confusion	0.088	0.092	0.092	0.002
Black-and-White Fallacy/Dictatorship	0.214	0.000	0.000	0.000
Exaggeration or Minimization	0.194	0.138	0.132	0.132
Red Herring	0.000	0.000	0.000	0.000
Repetitions	0.000	0.000	0.000	0.000
Thought-Terminating Cliché	0.080	0.286	–	–
Whataboutism	0.000	–	–	–
Macro-F1	**0.214**	**0.224**	**0.164**	**0.205**
Micro-F1	**0.253**	**0.248**	**0.248**	**0.217**

Table 9: Per-technique F1 across shot settings for GPT-4.1.

C.2 Supervised Fine-Tuning

We fine-tune GPT-4o using a supervised learning setup, with the data split into 70% training, 15% validation, and 15% testing. Table 10 reports the performance of the fine-tuned model on the held-out test set.

Technique	P	R	F1
Appeal to Authority	0.340	0.292	0.314
Appeal to Fear/Prejudice	0.246	0.205	0.224
Bandwagon	0.000	0.000	0.000
Black-and-White Fallacy/Dictatorship	0.000	0.000	0.000
Causal Oversimplification	0.147	0.061	0.087
Doubt	0.684	0.786	0.731
Exaggeration or Minimization	0.000	0.000	0.000
Flag-waving	0.130	0.136	0.133
Loaded Language	0.063	0.236	0.099
Name Calling or Labeling	0.022	0.022	0.022
Obfuscation/Vagueness/Confusion	0.000	0.000	0.000
Red Herring	0.000	0.000	0.000
Repetitions	0.004	0.072	0.007
Slogans	0.000	0.000	0.000
Macro-F1	–	–	**0.116**
Micro-F1	–	–	**0.168**

Table 10: Per-technique precision, recall, and F1 for supervised fine-tuning of GPT-4o.

C.3 LLM–Human Annotation Alignment

For the LLM–human annotation alignment analysis, we use a multi-label prompting setup based on the templates in Table 9. Table 11 reports per-technique F1 scores for the evaluated models. We additionally examine chain-of-thought prompting, with results shown in Table 12.

Technique	GPT-4o	GPT-4.1	Claude-Sonnet-4.5
Appeal to Authority	0.017	0.040	0.104
Appeal to Fear/Prejudice	0.018	0.009	0.036
Bandwagon	0.000	0.000	0.000
Black-and-White Fallacy/Dictatorship	0.000	0.000	0.000
Causal Oversimplification	0.006	0.007	0.026
Doubt	0.000	0.000	0.071
Exaggeration or Minimization	0.002	0.011	0.008
Flag-waving	0.010	0.008	0.067
Loaded Language	0.037	0.032	0.058
Name Calling or Labeling	0.000	0.000	0.000
Obfuscation/Vagueness/Confusion	0.000	0.000	0.000
Red Herring	0.000	0.000	0.000
Repetitions	0.016	0.007	0.003
Slogans	0.000	0.000	0.000
Thought-Terminating Cliché	0.000	0.000	0.000
Whataboutism	0.000	0.000	0.000
Macro-F1	**0.006**	**0.007**	**0.022**
Micro-F1	**0.024**	**0.029**	**0.060**

Table 11: Per-technique F1 scores under a multi-label setup, comparing GPT-4o, GPT-4.1, and Claude-Sonnet-4.5.

Technique	GPT-4o	GPT-4.1	Claude-Sonnet-4.5
Appeal to Authority	0.029	0.023	0.158
Appeal to Fear/Prejudice	0.001	0.000	0.038
Bandwagon	0.000	0.000	0.000
Black-and-White Fallacy/Dictatorship	0.000	0.000	0.000
Causal Oversimplification	0.005	0.010	0.033
Doubt	0.000	0.000	0.084
Exaggeration or Minimization	0.008	0.018	0.029
Flag-waving	0.002	0.014	0.036
Loaded Language	0.025	0.024	0.053
Name Calling or Labeling	0.000	0.029	0.000
Obfuscation/Vagueness/Confusion	0.015	0.021	0.058
Red Herring	0.000	0.000	0.000
Repetitions	0.008	0.008	0.016
Slogans	0.000	0.000	0.000
Thought-Terminating Cliché	0.000	0.000	0.000
Whataboutism	0.000	0.000	0.000
Macro-F1	**0.006**	**0.009**	**0.030**
Micro-F1	**0.022**	**0.021**	**0.075**

Table 12: Per-technique F1 scores under a multi-label CoT setup, comparing GPT-4o, GPT-4.1, and Claude-Sonnet-4.5.

D Linguistic and Cultural Background

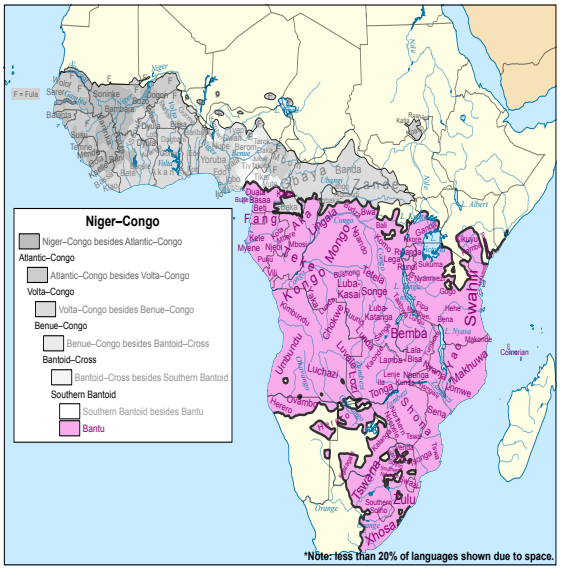

Figure 3: Distribution of Bantu languages within the Niger–Congo family. Figure source: https://commons.wikimedia.org/wiki/File:Map_of_the_Bantu_languages.svg.

Kinyarwanda is a major Bantu language spoken primarily in Rwanda and parts of Central and Eastern Africa. It belongs to the Niger–Congo language family and is used across formal and informal domains, including news media. Kinyarwanda is highly agglutinative, with rich verbal morphology and an extensive noun class system. Grammatical, semantic, and pragmatic information is often encoded within single words, resulting in dense and compositional sentence structure. For example, *yarayimumpereye* ("he gave it to him on my behalf") expresses tense, aspect, subject, object, and benefactive relations within a single verb form.

Beyond grammatical structure, everyday Kinyarwanda makes frequent use of idiomatic and culturally conventional expressions whose literal meanings may appear unusually intense or evaluative to non-native speakers without cultural grounding. For instance, *gukura ubwatsi* (literally, "to remove the grass") is conventionally used to express gratitude, particularly in reference to receiving livestock. Similar conventions extend to personal naming practices, such as *Bumanzi bwa se* ("the chastity of his father") or *Imenagitero* ("the one who leads and wins the battle").

E Interview Transcript

Below is the transcript of an interview with a KinyaProp annotator. The interviewee was anonymized and consented to the release of this full transcript.

What is your background in Kinyarwanda and your connection to it?

Kinyarwanda is my native language and the language I have spoken since I was born. Everyone in my family and in my neighborhood spoke Kinyarwanda. All aspects of my childhood, including stories, songs, riddles, and oral traditions, were in Kinyarwanda. In high school, I studied Kinyarwanda formally, including advanced grammar and morphology, since it is a required subject in Rwanda.

What do you think AI can do for the revitalization of Kinyarwanda, and what could be its impact?

I think AI can really help revitalize Kinyarwanda to a large extent. Kinyarwanda is a very rich language with complex morphology and expressive structures that even we native speakers often take for granted. For example, it has many noun classes and complex tense formations. These are aspects that AI systems can handle well with enough training data. I also remember learning in high school that some linguistic aspects of Kinyarwanda, like idioms and nouns, have been lost over time. Today, the language is increasingly influenced by loanwords and language mixing, and I think AI could help preserve and document Kinyarwanda as a language and as a culture.

What are your thoughts on misinformation in Kinyarwanda media?

Before working on this project, I did not know that misinformation in the media was a serious issue, and I think many people in Rwanda are also not aware of it. I was surprised to see how many articles that people read every day contain elements of propaganda, which I believe can influence people's opinions on many topics.

What do you think about the dataset created in this work?

I think this dataset will be very helpful, especially for training models to better understand propaganda in Kinyarwanda and to help combat misinformation more broadly. Today, we rely heavily on AI tools, and many of them already generate Kinyarwanda although they are not very good but they still struggle with morphological analysis of simple words. If these systems do not understand propaganda, they can also generate it. I think this dataset can help improve the reliability and trustworthiness of such AI systems.

Figure 4: Excerpt from an interview with a native Kinyarwanda speaker.