Population Replacement Conspiracy Theories Detection on Telegram and News Headlines: Benchmarking LLMs and BERT models in Portuguese and Italian

Erik Bran Marino, Renata Vieira
Universidade de Évora, CIDEHUS
Évora, Portugal
erik.marino@uevora.pt, renatav@uevora.pt

Abstract

Disinformation has become a serious threat to the democratic stability of Western societies, with various conspiracy theories spreading from fringe spaces to mainstream media and politics. While some of these theories may seem merely absurd and harmless, others pose significant risks. Among the most dangerous are Population Replacement Conspiracy Theories (PRCTs), which promote the false narrative of a deliberate demographic substitution through immigration. Despite their disinformative nature, increasing widespread and documented connections to extremist violence and political polarization, current computational detection models primarily target COVID-19 or general conspiracy theories, lacking specialized annotated corpora and approaches for identifying PRCTs in multilingual contexts. In this work, we present the first systematic benchmark for PRCT detection in Portuguese and Italian.

1. Introduction

Population Replacement Conspiracy Theories (PRCTs) constitute a family of false narratives claiming deliberate orchestration of demographic substitution through immigration and differential birth rates (Bracke and Hernández Aguilar, 2023). From a theoretical perspective, PRCTs represent an intersection within the *Information Disorder* framework (Wardle and Derakhshan, 2017). While often circulating as *mis-information* when shared by unaware users, their strategic use in political rhetoric frequently aligns with *dis-information*, intentionally disseminated to cause public harm.

These theories, including the Great Replacement, White Genocide, Kalergi Plan, and Eurabia, have demonstrably dangerous real-world consequences, motivating acts of extremist violence such as the Utøya attack (2011) and the Christchurch massacre (2019) (Wojtasik, 2020; Ekman, 2022). Recent research confirms that PRCT endorsement correlates with violent intentions, Anti-Muslim prejudice, and support for exclusionary policies (Obaidi et al., 2022), while their mainstreaming into political discourse raises political polarization and threatens democratic institutions (Marino et al., 2024).

In this work, we present the first systematic benchmark for PRCT detection in Romance languages, evaluating open-weights Large Language Models and fine-tuned encoders across Italian and Portuguese. Our findings reveal that: (1) Deepseek-V3 generally performs best due to its larger configuration; (2) the reasoning-optimized LLM DeepSeek-R1-14B, fine-tuned on Telegram data, achieves the highest performances through cross-domain transfer; and (3) efficient language-specific encoders remain highly competitive for high-throughput applications.

Limited data availability in such an underexplored domain leads to confounded variables; thus, cross-domain results should be interpreted as preliminary hypotheses rather than causal evidence.

2. Related Works

Despite documented harms discussed above, and their frequent proliferation on digital environments, computational detection of PRCTs remains underdeveloped. Existing conspiracy theory detection research focuses predominantly on COVID-19 misinformation, climate discourse, general conspiracy narratives, or platform-specific phenomena (Pogorelov et al., 2021; Miani et al., 2022; Cheatham et al., 2022; Weinzierl and Harabagiu, 2022; Langguth et al., 2023; Gambini et al., 2024; Maggini et al., 2025). While these contributions advance conspiracy detection methodologies, they do not address the specific linguistic and semantic patterns characterizing PRCT discourse.

Moreover, multilingual PRCT detection remains limited, with few isolated efforts on Youtube English datasets (Marino et al., 2025), leaving Portuguese and Italian languages critically underrepresented in computational approaches to PRCT narrative detection. Understanding how computational models perform across these linguistically related but politically differentiated contexts is essential for developing robust automated detection systems capable of monitoring extremist discourse in multilingual transnational spaces.

Cross-domain hate speech detection has demonstrated that models trained on one dis-

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 104–113
12 May 2026.

course domain can transfer to others with varying degrees of success. Toraman et al. (2022) examined transfer learning across five hate domains (religion, gender, race, politics, sports) in English and Turkish, finding that 96% of target domain performance could be recovered by training on other domains in English (92% in Turkish), with gender and religion domains generalizing better than sports-specific content. Markov et al. (2021) demonstrated that stylometric and emotion-based features provide robust cross-domain hate speech detection across English, Slovene, and Dutch, outperforming word and character n-gram features under cross-domain conditions and significantly boosting deep learning models when combined in an ensemble. Pamungkas and Patti (2019) addressed both cross-domain and cross-lingual abusive language detection across ten datasets in English, Italian, Spanish, and German, showing that systems trained on general abusive language datasets produce cross-domain robust systems capable of detecting more specific types of abusive content, and that domain-independent multilingual lexicons (HurtLex) facilitate knowledge transfer between domains and languages.

While these studies establish that cross-domain transfer is feasible for general hate speech categories, whether similar patterns hold for the task of PRCT detection across different media types and Romance languages remains an open empirical question. We address this gap by presenting the first systematic benchmark for PRCT detection in Portuguese and Italian. Our contribution is twofold: (1) we compare model architectures spanning fine-tuned transformer encoders and open-weights large language models, (2) to the best of our knowledge, we provide the first empirical evidence for cross-domain transfer in PRCT detection, comparing in-domain performance against domain shift scenarios.

Our experimental design evaluates performance across two distinct communication paradigms: informal, community-oriented Telegram messages and formal, public-oriented news headlines. Good performance in both domains would suggest a generalizability that could extend to other media, such as webpages, parliamentary speeches, or political manifestos.

3. Methodology

3.1. Datasets

We evaluate models on two annotated corpora regarding migration in European Portuguese and Italian. This language pairing was selected not only because of data availability and their shared Romance language roots, but also because both countries are currently experiencing different but comparable surges in far-right political narratives and PRCT mainstreaming. The corpora represent distinct communicative contexts: Portuguese Telegram messages (informal, conversational) and Italian news headlines (formal, public-oriented). We summarize the key characteristics of both corpora in the *Dataset Overview* panel.

The Portuguese Telegram corpus is a subset of the *Mute Cods* dataset (Laken et al., 2026), extracted from public channels documented as disseminating extremist and conspiracist content. It contains 919 training and 231 test messages with 15.7% PRCT prevalence. Messages average 383 characters. The annotation was conducted by domain experts in linguistics and social sciences following detailed annotation guidelines. To ensure consistency, data were at least double-annotated, with disagreements resolved through discussion to refine the guidelines, achieving overall a moderate inter-annotator agreement (Krippendorff's $\alpha = 0.58$).

The Italian news corpus is derived from the *PartisanLens* dataset (Maggini et al., 2026), which collected immigration-related headlines programmatically via Media Cloud between 2020 and 2024. The subset used comprises the Italian-only data with 434 training and 131 test headlines (averaging 81 characters), of which in total 161 are PRCT. This dataset was annotated for PRCT by two independent experts, with a third senior annotator resolving any disagreements, resulting in a substantial agreement (Krippendorff's $\alpha = 0.774$).

The observed inter-annotator agreement levels align with the inherent complexity of annotating multidimensional social constructs like PRCTs. As Matamoros-Fernández and Farkas (2021) highlight, moderate agreement is standard in social media research due to the ambivalence of online communication and context collapse. Our Telegram agreement ($\alpha = 0.58$) specifically reflects this reality, where conspiratorial narratives frequently blend with humor, irony, and informal vernacular. In such subjective tasks, chance-corrected metrics capture genuine human interpretive variation of implicit framing rather than poor annotation quality (Plank, 2022; Hemm et al., 2024).

Dataset Overview

Corpus	Lang	Train	Test	PRCT%	Len	IAA
Telegram	PT	919	231	15.7	383	α=0.58
News	IT	434	131	28.5	81	α=0.77

Details: *Len* indicates the average character length per sample; *IAA* represents the Inter-Annotator Agreement measured via Krippendorff's α.

3.2. Models

We evaluate seven model architectures across different learning paradigms: four large language models (Mistral-7B-Instruct-v0.3, Phi-4 14B, DeepSeek-R1-14B, DeepSeek-V3) and three transformer encoders (XLM-RoBERTa-base 270M multilingual, Albertina-280M Portuguese-specific, UmBERTo-110M Italian-specific). Depending on the experimental phase, models are evaluated either in a zero-shot setting or after supervised fine-tuning (using LoRA for LLMs and standard fine-tuning for encoders). Our experimental process is systematically divided into four distinct phases.

3.3. Experimental Design

The task is formally modeled as a binary text classification problem, where a given input text must be classified as either containing PRCT narratives (positive class) or not (negative class).

Phase 1 establishes the zero-shot baseline performance. We evaluate all four LLMs (Mistral-7B, Phi-4, DeepSeek-R1-14B, DeepSeek-V3) in their vanilla configurations separately on each test set. DeepSeek-V3 was specifically included to provide a performance upper bound against a state-of-the-art (SOTA) model. This identifies the inherent reasoning capabilities of each model prior to any task-specific adaptation. To verify robustness beyond limited test sets (n=231 PT, n=131 IT), we additionally evaluate Mistral-7B and DeepSeek-R1-14B zero-shot on complete datasets (train+test combined, n=1150 PT, n=565 IT) to check whether overall the performances keep being consistent.

Phase 2 investigates domain-specific fine-tuning. We implement LoRA fine-tuning exclusively on the models that demonstrated strong potential in Phase 1 and whose parameter size allowed feasible training on our hardware infrastructure. Specifically, Mistral-7B and DeepSeek-R1-14B were selected for fine-tuning. Phi-4 was excluded due to its uncompetitive zero-shot performance, while DeepSeek-V3 was excluded because its massive parameter size rendered fine-tuning computationally prohibitive in our setup. We fine-tuned these LLMs, alongside language-specific encoders (UmBERTo for Italian, Albertina for Portuguese), on single-domain configurations (Telegram-only and News-only) to evaluate in-domain versus cross-domain transferability.

Phase 3 evaluates merged-domain fine-tuning. We test whether combining both datasets during training yields domain-invariant representations that improve overall classification. We evaluate Mistral-7B, DeepSeek-R1-14B, and the multilingual encoder XLM-RoBERTa trained on the merged corpus, comparing the results against the domain-specific models from Phase 2.

Phase 4 constructs a Pareto frontier analysis across all evaluated configurations to identify non-dominated models optimizing distinct regions of the accuracy-efficiency trade-space. This allows us to find optimal model choices depending on the priority of the task (maximum accuracy, balanced performance, or maximum throughput).

We report macro-averaged F1 (equal class weighting) and binary F1 (positive class focus) as primary metrics, alongside accuracy, precision, and recall. LLMs use temperature 0 for deterministic outputs with structured prompts defining task requirements and output format. Fine-tuned encoders use class-weighted cross-entropy loss with learning rate 2e-5, batch size 16, and early stopping on validation F1-macro. All experiments use seed 42 for reproducibility. LoRA employed rank 8, alpha 16 on 16 layers, learning rate 1e-5, batch size 2, 600 iterations.

3.4. Computational Infrastructure

With the exception of DeepSeek-V3, which was accessed via official API calls due to its massive parameter size (671 B), all experiments were conducted locally on a MacBook Pro with an Apple M3 Max chip (16-core CPU, 40-core GPU) and 64 GB of unified memory. Local LLM inference utilized the MLX framework optimized for Apple Silicon, while encoder fine-tuning employed PyTorch.

4. Results

4.1. Phase 1: Zero-Shot Baseline Performance

Phase 1 established the inherent capabilities of LLMs prior to fine-tuning (Table 1). DeepSeek-V3 demonstrated the strongest zero-shot performance on Italian News (F1-macro 0.879) and good performances on Portuguese Telegram (F1-macro 0.758). We do not take into account Deepseek-V3 time-analysis, as it was run on a different setup compared to the other models. DeepSeek-R1-14B followed closely, showing highly competitive reasoning capabilities (F1-macro 0.870 on IT; 0.781 on PT: the highest result), albeit with significantly slower inference due to its chain-of-thought architecture. Mistral-7B offered a balanced trade-off, maintaining good F1-macro scores (0.702 IT, 0.744 PT) with the lowest inference latency among LLMs (0.64s–0.80s). Conversely, Phi-4 exhibited overly conservative classification behavior, resulting in high false negative rates and the lowest F1-macro scores across both sets (0.611 IT, 0.697 PT). Full dataset evaluation confirmed test set representativeness: Mistral-7B maintained F1-macro stability (test: 0.702/0.744 vs full: 0.704/0.704), as

		News ITA						Telegram PT					
Model	**Type**	**Acc**	**P_M**	**R_M**	**$F1_M$**	**$F1_B$**	**Time**	**Acc**	**P_M**	**R_M**	**$F1_M$**	**$F1_B$**	**Time**
Phi-4	Vanilla	0.763	0.873	0.613	0.611	0.367	8.39s	0.879	**0.819**	0.656	0.697	0.462	7.67s
Mistral-7B	Vanilla	0.771	0.737	0.688	0.702	0.559	**0.64s**	**0.887**	0.816	0.707	0.744	0.552	**0.80s**
DeepSeek-R1-14B	Vanilla	0.893	0.883	0.860	0.870	0.816	12.86s	0.866	0.752	**0.830**	**0.781**	**0.643**	15.10s
DeepSeek-V3	Vanilla	**0.901**	**0.896**	**0.866**	**0.879**	**0.827**	1.69s	**0.887**	0.802	0.729	0.758	0.581	1.78s

Table 1: Phase 1: Zero-shot baseline performance of Large Language Models. Bold indicates best performance per test set.

did DeepSeek-R1-14B (test: 0.870/0.781 vs full: 0.820/0.752).

4.2. Phase 2: Domain-Specific Fine-Tuning

Building upon Phase 1, we proceeded to fine-tune Mistral-7B and DeepSeek-R1-14B. DeepSeek-V3 and Phi-4 were excluded from this and subsequent phases due to, respectively, hardware limitations and poor baseline performance. Table 2 presents the results of models trained strictly on single domains, including language-specific encoders.

Mistral-7B benefited consistently from domain-matched training: Telegram-only fine-tuning elevated its F1-macro to 0.819 on Portuguese evaluation, while News-only fine-tuning reached 0.781 on Italian headlines. However, this accuracy gain incurred a substantial inference cost (4.5-4.7s per sample vs 0.64-0.80s zero-shot). DeepSeek-R1-14B demonstrated exceptional cross-domain performance: when fine-tuned strictly on Portuguese Telegram, it achieved F1-macro 0.892 on Italian news: the highest score observed across all experiments. Paradoxically, this same model degraded on its own training domain (F1-macro 0.655 on PT Telegram, below its zero-shot baseline of 0.781). This suggests architectural sensitivity where reasoning-optimized models may suffer forgetting on informal, fragmented content while successfully generalizing conspiratorial structures to formal discourse.

Language-specific encoders showed mixed results: Albertina achieved strong in-domain performance on Portuguese Telegram (F1-macro 0.774) with exceptional speed (0.018s), whereas UmBERTo did not performe well on Italian news (F1-macro 0.589).

4.3. Phase 3: Merged-Domain Fine-Tuning

To test whether combining training data from both languages and domains would yield a more robust, generalized classifier, we evaluated Mistral-7B, DeepSeek-R1-14B, and the multilingual encoder XLM-RoBERTa on a merged training corpus (Table 3). The findings indicate that merging datasets did not yield significant improvements and, in several cases, degraded performance.

For Mistral-7B, merged training resulted in F1-macro 0.734 on Italian News and 0.770 on Portuguese Telegram. These scores fall short of the peak performances achieved via single-domain fine-tuning (0.781 on News FT and 0.819 on TG FT, respectively). DeepSeek-R1-14B exhibited a similar trend, dropping to F1-macro 0.833 on News and 0.633 on Telegram under the merged configuration. The multilingual encoder XLM-RoBERTa achieved modest scores (F1-macro 0.661 IT, 0.670 PT) but stood out for its remarkable efficiency (0.005s per sample). Overall, the data suggests that in PRCT detection, exposing models to highly heterogeneous stylistic registers (formal vs. informal) and languages simultaneously may introduce interference rather than constructive transfer.

4.4. Phase 4: Pareto-Optimal Model Selection

Phase 4 constructs the Pareto frontier separately for each language to identify optimal deployment strategies for large-scale annotation. DeepSeek-V3 was explicitly excluded from this analysis: while it served as a SOTA benchmark in Phase 1, its reliance on external API calls makes its inference latency incomparable to locally executed models, rendering any efficiency comparison unfair. A model configuration is Pareto-optimal if no alternative achieves both higher F1-binary (positive class) and faster inference within the target language evaluation. Table 4 presents Pareto-optimal language configurations across all tested models.

For Italian news headlines, four configurations occupy the Pareto frontier. XLM-RoBERTa provides maximum throughput (200 samples/second) at moderate accuracy (F1-binary 0.568). Mistral-7B with Telegram LoRA achieves balanced performance (F1-binary 0.688, 4.50s per sample), representing a 12 percentage point improvement over XLM-RoBERTa with 900-fold slower inference. DeepSeek-R1-14B zero-shot occupies the quality-focused tier (F1-binary 0.816, 12.86s), while DeepSeek with Telegram LoRA achieves

		News ITA						Telegram PT					
Model	**Type**	**Acc**	**P_M**	**R_M**	**$F1_M$**	**$F1_B$**	**Time**	**Acc**	**P_M**	**R_M**	**$F1_M$**	**$F1_B$**	**Time**
Mistral-7B	TG FT	0.771	0.748	0.786	0.753	0.688	4.50s	**0.896**	0.797	**0.848**	**0.819**	**0.700**	4.62s
Mistral-7B	News FT	0.840	0.869	0.752	0.781	0.667	4.72s	**0.896**	**0.824**	0.746	0.777	0.613	4.73s
DeepSeek-R1-14B	TG FT	**0.908**	**0.892**	**0.892**	**0.892**	**0.850**	16.67s	0.853	0.716	0.630	0.655	0.393	15.94s
DeepSeek-R1-14B	News FT	0.901	0.881	0.887	0.884	0.840	15.90s	0.857	0.726	0.689	0.705	0.492	16.34s
UmBERTo	News FT	0.710	0.644	0.588	0.589	0.367	**0.005s**	—	—	—	—	—	—
Albertina	TG FT	—	—	—	—	—	—	0.874	0.761	0.790	0.774	0.623	**0.018s**

Table 2: Phase 2: Performance of models fine-tuned on single domains (Telegram or News). Bold indicates best performance per test set.

		News ITA						Telegram PT					
Model	**Type**	**Acc**	**P_M**	**R_M**	**$F1_M$**	**$F1_B$**	**Time**	**Acc**	**P_M**	**R_M**	**$F1_M$**	**$F1_B$**	**Time**
Mistral-7B	Merged FT	0.817	**0.871**	0.707	0.734	0.586	4.52s	**0.900**	**0.860**	0.726	**0.770**	**0.597**	4.58s
DeepSeek-R1-14B	Merged FT	**0.855**	0.827	**0.840**	**0.833**	**0.771**	16.20s	0.823	0.648	0.623	0.633	0.369	15.60s
XLM-RoBERTa	Merged FT	0.687	0.660	0.684	0.661	0.568	**0.005s**	0.749	0.663	**0.783**	0.670	0.508	**0.005s**

Table 3: Phase 3: Performance of models fine-tuned on the merged dataset (News + Telegram). Bold indicates best performance per test set.

maximum accuracy (F1-binary 0.850, 16.67s), the highest score observed across all Italian evaluations.

For Portuguese Telegram messages, three distinct configurations emerge. XLM-RoBERTa provides maximum throughput but with lower accuracy (F1-binary 0.508, 0.005s). Albertina fine-tuned on Telegram offers Portuguese-specific optimization (F1-binary 0.623, 0.018s), processing 55.6 samples per second while achieving 11 percentage point improvement over XLM-RoBERTa. Mistral-7B with Telegram LoRA FT reaches maximum accuracy (F1-binary 0.700, 4.62s), representing the optimal choice for quality-critical Portuguese Telegram annotation tasks with 26-fold slower inference than Albertina but 7.7 percentage point F1 improvement.

The choice among Pareto-optimal models depends on operational constraints and target language. For News Headlines Italian data annotation, DeepSeek-R1-14B with Telegram LoRA provides maximum accuracy (F1-binary 0.850) suitable for quality-critical applications, while Mistral-7B with Telegram LoRA offers balanced performance at faster inference. For Portuguese data annotation, Mistral-7B with Telegram LoRA achieves maximum accuracy (F1-binary 0.700), while Albertina provides adequate performance with great speed advantage for throughput-critical workflows. All other evaluated configurations are Pareto-dominated within their respective language evaluations.

5. Error Analysis

Zero-shot LLMs demonstrate conservative behavior with high Precision and lower Recall, minimizing false positives at the cost of missing true PRCT instances. Conversely, fine-tuned encoders show inverse patterns, with lower Precision but higher Recall, capturing more true positives while also accepting more false positives. Error rates demonstrate contrasting relationships with text length across the two domains. Italian news headlines show decreasing error rates as length increases, indicating that longer headlines provide sufficient context. Portuguese Telegram messages exhibit the inverse pattern, where PRCT claims embedded within extended narratives become diluted.

Cross-domain transfer reveals unexpected asymmetries. Mistral-7B fine-tuned exclusively on Portuguese Telegram achieves competitive performance on Italian news (F1-macro 0.753). Qualitative examination shows coverage of implicit formulations, such as weaponization metaphors, elite manipulation claims, and presuppositional framing. One hypothesis is that exposure to explicit conspiratorial language during training facilitates recognition of implicit PRCT formulations. DeepSeek-R1-14B (TG FT) demonstrates markedly different behavior, achieving exceptional performance on formal news (F1 0.892) but degrading on the training domain itself. This suggests architectural sensitivity where reasoning-optimized models benefit from cross-domain exposure on formal discourse but suffer forgetting on informal content.

To systematize our qualitative findings, Table

Model Configuration	Type	$F1_B$	Time/sample	Samples/sec	Recommended for
Italian News Headlines					
XLM-RoBERTa	Merged FT	0.568	0.005s	200.0	Maximum throughput
Mistral-7B	TG FT	0.688	4.50s	0.22	Balanced performance
DeepSeek-R1-14B	TG FT	**0.850**	16.67s	0.06	Maximum accuracy
Portuguese Telegram Messages					
XLM-RoBERTa	Merged FT	0.508	0.005s	200.0	Maximum throughput
Albertina	TG FT	0.623	0.018s	55.6	Balanced performance
Mistral-7B	TG FT	**0.700**	4.62s	0.22	Maximum accuracy

Table 4: Language-specific Pareto-optimal models identified through Phase 4 analysis. We highlight configurations suitable for distinct application scenarios: maximum accuracy favors offline social science analysis, while maximum throughput suits real-time content flagging. Each section shows dominant configurations for the target language (no alternative achieves both higher $F1_B$ and faster inference).

5 categorizes the most frequent error patterns and classification behaviors across models and domains. Model-specific error patterns reveal distinct behavioral profiles. For instance, Phi-4 exhibits extreme conservatism, missing 77.5% of PRCT instances, including headlines with literal replacement terminology. Mistral-7B demonstrates balanced behavior but struggles with distinguishing fear-inducing rhetoric from actual conspiratorial framing, occasionally missing military metaphors. Albertina exhibits domain-appropriate sensitivity for Portuguese informal content but struggles with historical PRCT references and demographic complaint rhetoric. DeepSeek-R1-14B fine-tuned on Telegram systematically misses serialized documentary content and specific dog-whistles (e.g., remigration) in its own fine-tuning domain.

A prominent source of false positives across all models occurs when they encounter affective language, fear-inducing rhetoric, or violent imagery that lacks the essential element of conspiratorial orchestration. Conversely, false negatives exhibit more diverse patterns: models systematically miss conspiracy theories expressed through indirect discussion, presupposition, or serialized content that distributes conspiratorial claims across multiple installments, requiring contextual understanding beyond a single-message scope.

6. Discussion

This work situates PRCTs within the broader framework of Information Disorder (Wardle and Derakhshan, 2017). Specifically, PRCTs typically operate as *dis-information*—false narratives deliberately created to inflict harm or sow societal division—though they are frequently propagated by genuine believers as *mis-information*. Understanding the contextual and cultural factors that differentiate formal news propagation in Italy from informal, community-driven Telegram messaging in Portugal is essential for developing comprehensive NLP systems capable of addressing the full spectrum of information disorders.

Models fine-tuned on Portuguese Telegram achieve higher F1 scores when evaluated on Italian news headlines than models trained on Italian news itself. Mistral TG FT reaches F1=0.753 on Italian news versus F1=0.688 for News FT and F1=0.729 for Merged FT. DeepSeek TG FT achieves F1=0.892 on Italian news, the highest score across all configurations.

This pattern is consistent with multiple non-exclusive hypotheses: (1) Telegram's explicit conspiratorial discourse provides richer linguistic signals than news headlines' implicit formulations; (2) longer texts (Telegram avg. 383 chars vs news 81 chars) enable more robust feature learning; (3) Italian linguistic features facilitate classification regardless of training corpus; (4) class distribution effects (28.5% PRCT in news vs 15.7% in Telegram) interact with model calibration.

DeepSeek-R1-14B exhibits paradoxical behavior: its Telegram-trained variant achieves F1=0.892 on Italian news (cross-domain evaluation) yet deteriorates to F1=0.655 on Portuguese Telegram (in-domain), performing worse than its own zero-shot baseline (F1=0.781). This pattern suggests forgetting during fine-tuning on informal discourse, where the model's chain-of-thought optimization for structured reasoning may conflict with Telegram's fragmented linguistic patterns. Conversely, Mistral-7B benefits consistently from domain-matched training across both datasets, with in-domain fine-tuning yielding monotonic improvements over zero-shot baselines.

Finally, the operational trade-off between predictive performance and computational cost dictates distinct deployment strategies. While LLMs offer superior reasoning for complex narratives, their

Error Pattern	Model (Behavior)	Domain	Representative Example
Implicit Formulations (Weaponization, Elites)	Mistral TG FT [True Positive]	News	*Germania li usa come armi per destabilizzarci* (Germany uses them as weapons to destabilize us)
Over-Conservatism (Literal replacement terms)	Phi-4 [False Negative]	News	*La sostituzione etnica di Meloni: 830mila immigrati* (Meloni's ethnic replacement: 830 thousand immigrants)
Fear-Inducing Rhetoric (Without orchestration)	Mistral-7B, Encoders [False Positive]	News	*Immigrati islamici pronti a colpire in Italia* (Islamic immigrants ready to strike in Italy)
Demographic Complaints (Historical context)	Albertina [False Negative]	TLgram	*Número de imigrantes nas escolas públicas aumenta 47% em dois anos* (Number of immigrants in public schools increases 47% in two years)
Serialized Content (& Specific dog-whistles)	DeepSeek TG FT [False Negative]	TLgram	*REMIGRΛZÓN ÚNICΛ SOLUZÓN* (REMIGRATION ONLY SOLUTION)
Reporting/Discussing (Presuppositional framing)	All Models [False Negative]	News	*Migranti e sostituzione etnica: critiche bipartisan a Lollobrigida* (Migrants and ethnic replacement: bipartisan criticism of Lollobrigida)

Table 5: Systematization of the most common error patterns and classification behaviors identified in the qualitative analysis.

inference latency renders them less suitable for real-time monitoring of high-velocity streams compared to efficient encoders like XLM-RoBERTa. As an example application, we could use a tiered architecture. Encoders would filter massive datasets first, passing only the ambiguous data to LLMs for detailed checking. While LLMs demonstrate superior performance in our experiments, recent work demonstrates that hybrid approaches combining shallow learning with theoretically-grounded features can outperform LLMs on related discourse tasks (Bassi et al., 2025). Given that, as Marino et al. (2025) found, PRCT discourse exhibits distinctive linguistic markers such as religious language, power dynamics, higher negative tone and conflict framing, future work should investigate whether operationalizing these patterns as explicit features could enhance detection while reducing computational costs.

7. Conclusions

This work provides a comparative analysis of large language models and fine-tuned encoders for Population Replacement Conspiracy Theory detection, offering the first empirical benchmark of their cross-corpus performance between informal Portuguese and formal Italian discourse contexts. We evaluated seven model architectures across informal Telegram messages and formal news headlines, establishing baseline performance metrics and identifying cross-corpus evaluation patterns.

A model fine-tuned on Portuguese Telegram data achieve F1-binary 0.850 (DeepSeek-R1-14B) when evaluated on Italian news headlines. This cross-corpus performance is consistent with domain transfer but could equally reflect linguistic, length-based, or distributional factors given the systematic confounding in our design. Controlled experiments with balanced language-domain combinations are required to isolate causal mechanisms.

Pareto frontier analysis identified language-specific optimal configurations balancing accuracy and computational efficiency: DeepSeek TG FT for Italian data (F1-binary 0.850, 16.67s per sample) and Mistral TG FT for Portuguese data (F1-binary 0.700, 4.62s per sample). These models provide practical deployment options for large-scale annotation (for instance, high-volume documents for offline social analysis), with encoder alternatives (XLM-RoBERTa, Albertina) offering throughput-critical solutions where speed outweighs accuracy requirements, which may be a better choice for online or run-time PRCT spotting. Future work should expand coverage to additional languages, model architectures, and test generalization to other media formats.

8. Limitations

A primary limitation of this study, as discussed in Section 6, is the systematic confounding of language, domain, text length, and class distribu-

System Prompt Configuration

You are a classifier specialized in detecting Population Replacement Conspiracy Theories (PRCT) in text. Your task is to determine if the given text contains or supports PRCT narratives.

DEFINITION OF PRCT:
Population Replacement Conspiracy Theories claim that native populations in Western countries are being systematically replaced by non-European (non-Christian, non-white) peoples through immigration and demographic changes as part of a deliberate orchestrated plan.

INDICATORS OF PRCT:

- Explicit references to "Great Replacement", "White Genocide", "Kalergi Plan", "Eurabia"
- Dog whistles: "Fourteen Words", "1488", "ZOG", "remigration", "demographic warfare", "invasion"
- Framing immigration as "invasion", "ethnic substitution", or "cultural erasure"
- Claims of orchestrated plans by elites/globalists to replace native populations
- Suggesting demographic changes are intentional and part of a hidden agenda

CLASSIFICATION:

- **YES:** The text mentions, supports, or alludes to PRCT narratives
- **NO:** The text does not contain PRCT content (may discuss immigration without conspiracy framing)

IMPORTANT:

- You will only see the text of the message. Base your decision only on that.
- General anti-immigration sentiment WITHOUT conspiracy elements = NO
- Concerns about demographic changes WITHOUT deliberate orchestration claims = NO
- If dog whistles are present, then, it's likely = YES

OUTPUT FORMAT: Respond with a valid JSON object:

{ "prct": "YES" or "NO" }

TEXT TO CLASSIFY: [[TEXT]]

Figure 1: The system prompt used for zero-shot classification with LLMs.

tion, which restricts our ability to isolate the specific drivers of cross-corpus performance. Consequently, our findings regarding cross-domain transfer should be interpreted strictly as preliminary hypotheses rather than definitive evidence.

Furthermore, the computational demands of fine-tuning LLMs imposed strict constraints on our experimental setup. Consequently, LoRA adaptations were executed using a single initialization seed. While reporting single-run performance is a pragmatic and standard approach for establishing initial baselines in resource-constrained LLM research, it inherently precludes the calculation of cross-run variance and formal statistical significance testing. We encourage future research with expanded computational budgets and data size to conduct multi-seed evaluations to establish robust confidence intervals. Plus, while LoRA substantially reduces computational costs and memory requirements, enabling experimentation across multiple model architectures, prior work demonstrates that LoRA can underperform full fine-tuning on certain tasks, particularly those requiring extensive parameter updates for domain adaptation. The magnitude of this performance gap in our PRCT detection context remains unquantified. Temporal coverage spans 2021–2024 without diachronic analysis, though PRCT narratives demonstrably evolve across decades. This temporal constraint limits conclusions about model robustness to evolving conspiracy formulations. Finally, the zero-shot evaluation employed a single English prompt template without few-shot example selection. This methodological choice was motivated by our objective to establish a uniform, rigorous baseline performance generalizable to deployment contexts lacking annotated data, given that LLMs often exhibit more stable alignment with English instructions even on multilingual tasks. A systematic prompt comparative study—including testing the same models under plain instruction, language-specific prompts (Italian and/or Portuguese), and few-shot or chain-of-verification (CoVe) prompts—is left as a next step for future research.

9. Ethical Considerations

This research involves the analysis of extremist narratives and conspiracy theories, which inherently include discriminatory and harmful language. To ensure privacy and ethical compliance, all Telegram messages were sourced exclusively from publicly accessible channels through official APIs, and no personally identifiable information was retained or analyzed. The Italian news headlines consist entirely of public journalistic records. Furthermore, given the sensitive and potentially distressing nature of the textual data, the annotation process was conducted exclusively by domain experts in linguistics and social sciences, with periodic meetings, rather than vulnerable and isolated crowd-workers, thereby attempting to mitigate the psychological risks associated with exposure to extremist content. Finally, we explicitly acknowledge the risk of false positives, such as misclassifying legitimate, albeit polarized, political discourse on immigration as conspiratorial. Consequently, these systems are intended to serve as assistive diagnostic tools that require human-in-the-loop oversight, rather than fully autonomous moderation mechanisms. All research activities were carried out in accordance with the ethical guidelines and with the formal approval of the research center's Ethical Review Board.

Acknowledgments

This work was supported by the HYBRIDS project, which has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351 and from the UK Research and Innovation (UKRI) Horizon Europe funding guarantee (Grant Number: EP/X036758/1). The work is partially supported by the Portuguese Science Foundation as part of the projects CEECIND/ 01997/2017 and UIDP/00057/2025. The content of this work reflects only the authors' view and the funding agencies are not responsible for any use that may be made of the information it contains.

Bibliography

Davide Bassi, Erik Bran Marino, Renata Vieira, and Martin Pereira. 2025. Old but gold: LLM-based features and shallow learning methods for fine-grained controversy analysis in YouTube comments. In *Proceedings of the 12th Argument mining Workshop*, pages 46–57, Vienna, Austria. Association for Computational Linguistics.

Sarah Bracke and Luis Manuel Hernández Aguilar. 2023. *The Politics of Replacement: From "Race Suicide" to the "Great Replacement"*. Routledge, London.

Susan Cheatham, Per E Kummervold, Lorenza Parisi, Barbara Lanfranchi, Ileana Croci, Francesca Comunello, Maria Cristina Rota, Antonietta Filia, Alberto Eugenio Tozzi, Caterina Rizzo, et al. 2022. Understanding the vaccine stance of italian tweets and addressing language changes through the covid-19 pandemic: Development and validation of a machine learning model. *Frontiers in Public Health*, 10:948880.

Mattias Ekman. 2022. The great replacement: Strategic mainstreaming of far-right conspiracy claims. *Convergence*, 28(4):1127–1143.

Margherita Gambini, Serena Tardelli, and Maurizio Tesconi. 2024. The anatomy of conspiracy theorists: unveiling traits using a comprehensive twitter dataset. *Computer Communications*, 217:25–40.

Ashley Hemm, Sandra Kübler, Michelle Seelig, John Funchion, Manohar Murthi, Kamal Premaratne, Daniel Verdear, and Stefan Wuchty. 2024. Are you serious? handling disagreement when annotating conspiracy theory texts. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 124–132, St. Julians, Malta. Association for Computational Linguistics.

Katarina Laken, Erik Bran Marino, Paloma Piot, Davide Bassi, Søren Fomsgaard, Michele Maggini, Renata Vieira, Marcos Garcia, and Sara Tonelli. 2026. Mute Cods: A Multilingual Telegram Dataset with Benchmark Models for Conspiracy Theory Detection. In *Proceedings of the 2026 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2026)*. Forthcoming.

Johannes Langguth, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, Jesper Phillips, and Konstantin Pogorelov. 2023. Coco: an annotated twitter dataset of covid-19 conspiracy theories. *Journal of Computational Social Science*, 6(2):443–484.

Michele Joshua Maggini, Davide Bassi, and Pablo Gamallo. 2025. Detecting hyperpartisanship and rhetorical bias in climate journalism: A sentence-level Italian dataset. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*,

pages 168–187, Vienna, Austria. Association for Computational Linguistics.

Michele Joshua Maggini, Paloma Piot, Anxo Pérez, Erik Bran Marino, Lúa Santamaría Montesinos, Ana Lisboa Cotovio, Marta Vázquez Abuín, Javier Parapar, and Pablo Gamallo. 2026. Partisanlens: A multilingual dataset of hyperpartisan and conspiratorial immigration narratives in european media. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1186.

Erik Marino, Jesus M. Benitez-Baleato, and Ana Sofia Ribeiro. 2024. The polarization loop: How emotions drive propagation of disinformation in online media—the case of conspiracy theories and extreme right movements in southern europe. *Social Sciences*, 13(603).

Erik Bran Marino, Davide Bassi, and Renata Vieira. 2025. Linguistic markers of population replacement conspiracy theories in youtube immigration discourse. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 670–679.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online. Association for Computational Linguistics.

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & new media*, 22(2):205–224.

Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2022. Loco: The 88-million-word language of conspiracy corpus. *Behavior research methods*, 54(4):1794–1817.

Milan Obaidi, Jonas Kunst, Simon Ozer, and Sasha Y Kimel. 2022. The "great replacement" conspiracy: How the perceived ousting of whites can evoke violent extremism and islamophobia. *Group Processes & Intergroup Relations*, 25(7):1675–1695.

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. 2021. Wico text: a labeled dataset of conspiracy theory and 5g-corona misinformation tweets. In *Proceedings of the 2021 workshop on open challenges in online social networks*, pages 21–25.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

Maxwell Weinzierl and Sanda Harabagiu. 2022. Vaccinelies: A natural language resource for learning to recognize misinformation about the covid-19 and hpv vaccines. *arXiv preprint arXiv:2202.09449*.

Karolina Wojtasik. 2020. Utøya–christchurch–halle. right-wing extremists' terrorism. *Security Dimensions. International and National Studies*, (33):84–97.