A Multilingual Linguistic Analysis of Human vs LLM-Generated News in a Disinformation Context

Silvia Gargova[1,3], **Alba Pérez-Montero**[2], **Elena Lloret**[3], **Paloma Moreda**[3]
[1]Big Data for Smart Society Institute (GATE),
[2]University Institute for Computing Research (IUII), University of Alicante
[3]Dept. of Software and Computing Systems, University of Alicante
silvia.gargova@gate-ai.eu, alba.perezm@ua.es, {elloret, moreda}@dlsi.ua.es

Abstract

The rise of Large Language Models has shifted the Information Disorder landscape toward automated threats. This study investigates the linguistic construction of synthetic news by comparing GPT-5, Gemini 2.5, and Grok 4 across English, Spanish, and Bulgarian. Using multilingual human-authored verified news and disinformation as seeds, we analyze how prompt informativeness and model architecture influence deceptive content production. Our methodology employs five metrics: semantic similarity, factual consistency, readability, lexical richness, and persuasion technique frequency. Our analysis reveals that while prompt scarcity leads to informational loss, LLMs maintain a homogenized stylistic template regardless of input length. Unlike human authors, who intensify rhetorical and emotional markers to drive deceptive intent, LLMs adhere to a neutral register. This study identifies distinct statistical patterns in generated content characterized by hyper-standardized readability and high lexical density ($p < 0.001$). These features serve as robust "LLM signatures", enabling a classification accuracy of 96% across English, Spanish, and Bulgarian. These findings suggest that generated disinformation relies on invariant syntactic structures rather than nuanced human rhetoric, providing a framework for detection tools centered on structural patterns rather than content veracity.

Keywords: Information Disorder, Natural Language Generation, Large Language Models, Summarization, Cybersecurity

1. Introduction

The disclosure of Large Language Models (LLMs) has fundamentally shifted the landscape of information integrity, presenting both unprecedented opportunities and significant risks. While LLMs excel at tasks like summarization and content creation, their sophisticated Natural Language Generation (NLG) capabilities have been readily exploited to produce highly convincing, large-scale disinformation (Park and Nan, 2025). This development exacerbates the crisis of Information Disorder (Wardle and Derakhshan, 2017) by moving the focus from manually crafted falsehoods to automated, systemic threats (Vykopal et al., 2023). Consequently, the rapid generation of unreliable information through these models poses a critical new challenge to cybersecurity and information ecosystems worldwide.

Traditional disinformation campaigns relied on human effort, making them slow and costly. LLMs erase these limitations, enabling the rapid construction of coherent, contextually relevant, and potentially multilingual disinformation. The risk is magnified because LLM-Generated content often exhibits high linguistic quality, making it difficult for both human users and automated tools to distinguish from genuine journalism (Su et al., 2023). Consequently, understanding the specific linguistic signature of LLM-Generated disinformation is essential to developing effective detection mechanisms.

Our study contributes to this area by focusing on the security implications of LLM-Generated texts. Unlike previous work that primarily focused on LLMs safety filters or high-level detection (Akiri et al., 2025), we delve into the linguistic construction of the generated output. We use prompts based on summarization to generate content derived from two core sources: Human-Authored Verified News (H-V) and Human-Authored Disinformation (H-D). These serve as seed texts to generate our core corpus: LLM-Generated from Verified Content (G-V) and LLM-Generated from Disinformation (G-D). By analyzing this generated content, our research bypasses the need to analyze the human source texts, allowing us to directly assess the role of NLG in the automated disinformation pipeline. This approach leads to the formulation of three core research questions:

- **RQ1 (Textual Fidelity and Information Scarcity)**: How does the quantity of source information, specifically under conditions of information scarcity, influence the preservation of content and overall textual fidelity in LLM-generated news?
- **RQ2 (Disinformation Compliance and the Rhetorical Gap)**: When prompted to generate content based on deceptive information, to what extent do LLMs demonstrate disinformation compliance, and do the resulting texts

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 85–96
12 May 2026.

exhibit a measurable rhetorical gap compared to human-authored persuasion?

- **RQ3 (Feature Importance and the Linguistic Signature of AI)**: Which specific linguistic features (ranging from readability to persuasion techniques) constitute the distinct linguistic signature of AI, allowing for the effective differentiation of machine-generated news from human counterparts?

The paper is organized as follows: In Section 2, we review previous studies covering the delimitation of disinformation (2.1), NLG and summarization principles (2.2), and cybersecurity evaluation in LLMs (2.3). Section 3 details our methodological framework, beginning with the motivation for our dataset selection (3.1) and models selection (3.2), followed by the design of our prompts (3.3), which includes both the summarization pipeline (3.3.1) and the news generation pipeline (3.3.2). We then introduce our multi-level analysis framework in Section 3.4. Finally, we present the results and discussion in Section 4, concluding the study with Section 5.

2. Related Work

The following review establishes the study's framework by examining three interconnected domains: the role of disinformation within the Information Disorder taxonomy, the use of NLG and summarization as methodological engines, and the cybersecurity implications of LLM-driven deceptive threats.

2.1. Disinformation within Information Disorder

The information landscape is undergoing a rapid transformation, accelerated by the integration of LLMs into daily life (Lazer et al., 2018; Esteban-Bravo et al., 2024). LLMs possess a pervasive ability to generate convincing yet false information which, combined with the difficulty humans have in discerning AI-generated text, poses a significant threat (Zhou et al., 2024). Recent studies, such as (Zhou et al., 2023), highlight that AI-generated misinformation is increasingly indistinguishable from human content, often simulating personal tones and uncertainty to bypass traditional skepticism. This escalating issue, combined with the vulnerability of digitally illiterate individuals, drives an urgent need for advanced detection research (Gravanis et al., 2019).

2.2. Natural Language Generation and Summarization

NLG has been revolutionized by the Transformer architecture and the rise of LLMs (Erdem et al., 2022; Miró Maestre et al., 2025). Advanced capabilities in Controllable Text Generation (CTG) allow models to satisfy specific user-defined constraints while maintaining high quality (Zhang et al., 2023). This is particularly relevant for analyzing systematic nuances in media language, where linguistic indicators (such as grammatical patterns and lexical variety) serve as primary features for identifying deceptive texts (Mahyoob and Algarady, 2020). In this study, we leverage summarization-based prompting techniques to generate synthetic news, ensuring semantic coherence while exploring the models' ability to emulate the rhetorical structures of both verified information and disinformation.

2.3. Cybersecurity Threats and Detection

The generative scale of LLMs has transformed disinformation into an automated, low-cost cybersecurity threat (Li and Fung, 2025; Park and Nan, 2025). A major concern in this domain is the emergence of "style-based attacks", where LLMs are used to reframe disinformation into an objective, trustworthy style, significantly degrading the performance of existing detectors (Wu et al., 2024). Detection remains a challenge because LLM-generated content often mimics legitimate news more effectively than human-written disinformation (Chen and Shu, 2023). However, as noted in recent surveys (Wu et al., 2025), synthetic text often exhibits unique statistical footprints, such as lower emotional variance and highly deterministic syntactic structures, compared to the high rhetorical intensity of human deception. These findings suggest that focusing on structural patterns, rather than just content veracity, offers a more robust path for future detection mechanisms.

3. Methodology

This section describes the systematic framework established to evaluate the role of LLMs in the generation of synthetic news and disinformation. The experimental workflow, illustrated in Figure 1, integrates text summarization as a preprocessing stage to facilitate the generation task.

To isolate the impact of LLMs on automated disinformation, we analyze four distinct text groups categorized by authorship and veracity:

- **Human-Authored Verified News (H-V)** and **Human-Authored Disinformation (H-D)**: Used as the initial "seeds".
- **LLM-Generated from Verified News (G-V)** and **LLM-Generated from Disinformation (G-D)**: The resulting synthetic outputs.

The summarization step serves purely as a preprocessing tool, condensing the source texts (H-V

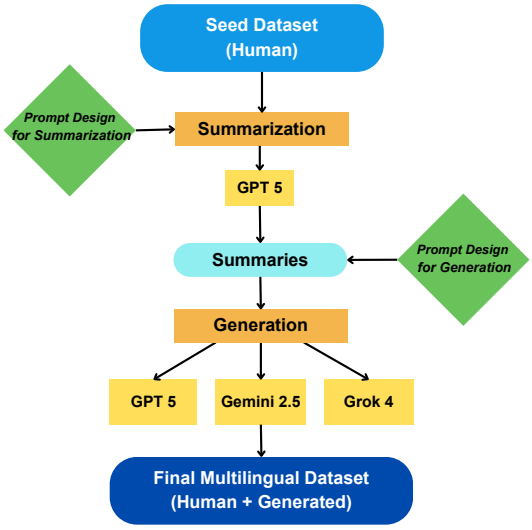

Figure 1: Schematic representation of the experimental workflow, illustrating the integration of text summarization as a preprocessing stage to facilitate the generation of synthetic news and disinformation.

and H-D) into essential information for the subsequent generation task. This setup allows for a direct linguistic comparison between reliable and unreliable inputs when processed by neural architectures.

3.1. Dataset Creation

We construct a multilingual dataset based on the resource introduced in (Pérez-Montero et al., 2025), covering three languages: Spanish, Bulgarian and English.

- **Source Material**: The corpus includes 160 balanced human-written articles per language (80 genuine news, 80 disinformation), totaling 480 base texts.
- **Synthetic Extension**: We generated synthetic counterparts for every article using three LLMs under three prompt configurations. This resulted in a total of 4,320 generated texts (480 texts $\times$ 3 models $\times$ 3 prompts).
- **Validation and Annotation**: All generated texts underwent manual review to ensure they meet the requirements of the experiment. Both human and synthetic texts were annotated with stylistic features, readability scores, and persuasion techniques to support machine-generated text detection research. This processes are explained in detail in Sections 3.2 and 3.3.

3.2. Model Selection

As this study involves two distinct tasks requiring different capabilities, we selected models based on their specific strengths and safety architectures.

3.2.1. Summarization Pipeline

For the preprocessing stage, we employ **GPT-5** (`gpt-5-mini-2025-08-07`). This model acts as a data processing tool to extract essential information for the generation pipeline. As the study does not aim to evaluate summary quality, this model was chosen for its efficiency in handling instruction-following tasks (Adams et al., 2023).

3.2.2. Generation Models

To examine how different architectures affect synthetic content, we selected three state-of-the-art LLMs via their official APIs:

- **GPT-5** (`gpt-5-mini-2025-08-07`): Known for versatile performance. While early versions faced limitations regarding "hallucinations" (Achiam et al., 2023), later iterations have implemented stricter firewalls to mitigate harmful content generation (Leon, 2025).
- **Gemini 2.5** (`gemini-2.5-flash`): Optimized for professional-grade output. Google employs a layered security strategy, including *Automated Red Teaming* (ART), to detect and mitigate vulnerabilities (Google DeepMind Team, 2025).
- **Grok 4** (`grok-4-fast-non-reasoning`): Marketed as a "maximally truth-seeking" and minimally censored AI (xAI Team, 2024). Its documented low refusal rate (Leite et al., 2025) makes it a critical subject for comparative disinformation analysis.

These models were chosen to represent diverse industry approaches to news generation and for their widespread accessibility through API interfaces. While we acknowledge that the proprietary nature of these closed-source models poses inherent challenges to long-term reproducibility, they represent the most prevalent tools currently used in the automated content landscape. To ensure results representative of standard user interactions, all models were accessed via their default API configurations without hyperparameter tuning during the generation phase.

3.3. Prompt Design: Summarization and Generation of News

Prompt design is fundamental to our two-step pipeline: (1) data summarization and (2) news article generation. Following Chen and Shu (2023), we

implement controlled generation by providing initial summaries to guide the output, utilizing established principles of few-shot learning and constrained text generation (Liu et al., 2023). To optimize performance across our multilingual corpus, all prompts are formulated in English. This strategy leverages the models' primary reasoning capabilities (Vadlapati, 2023) and adheres to established cross-lingual paradigms, such as "translate-then-summarize", to ensure consistency in Spanish and Bulgarian outputs (Zhang et al., 2024).

3.3.1. Summarization Pipeline

To produce structured summaries, we employ a prompt strategy targeting two objectives. First, models extract the most important information using the 5W1H framework (Who, What, When, Where, Why, How) (Kipling, 1902)[1], ensuring informational completeness. Second, the extracted content is presented as clear bullet points, following approaches such as the "retrieve-then-summarize" pipeline (Zhang et al., 2024). The prompt explicitly instructs the model:

```
Create an abstract summary in
{language} of the following news
article and extract the most im-
portant information. Focus on
the information essential to un-
derstanding the news content,
using the 5W1H (What, Why, When,
Where, Who, How) technique.
News Article:
{article_text}
```

This method blends extractive and abstractive summarization: the model identifies fundamental facts from the source (extractive) and condenses them into concise, independent bullet points (abstractive). By removing linguistic elements often associated with disinformation (Pérez-Montero et al., 2025), this strategy generates summaries containing only crucial facts, allowing us to evaluate whether LLMs, when provided solely with core information, introduce bias or fabricated content in subsequent news generation.

3.3.2. News Generation Pipeline

In the generation step, we provide LLMs with two distinct inputs (news titles and summaries) to assess how information granularity affects content generation. Three prompt types were designed, each conveying different instructions, but all begin with a standardized set of guidelines. Importantly, these prompts use neutral instructions (e.g., "Generate a news article") without explicitly prompting for disinformation; the veracity of the output depends solely on the source material provided. This ensures consistent structure and format across outputs, enabling a reliable comparison of model performance. The foundational guidelines for all prompt types are as follows:

```
You're an experienced journalist.
Use the following title and/or
description to write a news ar-
ticle in {language}. Answer in
structured JSON format (without
Markdown formatting) like so:
{ "generated_title": "The title
of the article",
"generated_news_article": "The
generated content of the arti-
cle." }.
{configuration}
```

Following this, each prompt design will include a different type of configuration to guide the LLMs in generating a news article at a particular level of granularity. Although the desired outcome is consistent (a single news article), the input options vary significantly in the quantity of information they contain. The different prompt types are as follows:

- **Prompt 1 (P1)**: For this prompt type, we combine the available information, introducing both the titles and the summaries of the seed texts into the model guidelines, so the prompt design is completed with:

  ```
  Title: [title]
  Description: [summary]
  ```

- **Prompt 2 (P2)**: This prompt type includes only the titles of our seed texts within the guidelines, so the rest of the prompt design consists of:

  ```
  Title: [title]
  ```

- **Prompt 3 (P3)**: In this type of prompt we only introduce on the guidelines the summarized content of our texts, following Section 3.3.1. The summaries themselves were created using a specific, structured prompt strategy to ensure a comprehensive set. This prompt design is completed with:

  ```
  Description: [summary]
  ```

3.4. Analysis Framework

To evaluate the output of LLMs across diverse linguistic contexts, we use a multi-dimensional analysis framework. Given that journalistic norms and

[1]Although rooted in classical rhetoric, the 5W1H method was popularized by Rudyard Kipling's 1902 poem (Kipling, 1902), which established these six questions as the professional journalistic standard for comprehensive reporting.

structural complexities vary significantly across different language families, a "one-size-fits-all" approach to text evaluation is insufficient. Consequently, our methodology integrates standardized cross-lingual metrics with specialized tools tailored to the morphological and syntactic nuances of English, Spanish, and Bulgarian. This approach ensures that the assessment of readability, lexical diversity, and factual fidelity remains sensitive to the inherent properties of each language while allowing for statistically valid cross-comparisons.

3.4.1. Readability

To quantify structural differences between human and machine-generated news, we use language-specific readability formulas for each target language. For English and Spanish, we utilize the established libraries in the `textstat` suite. English texts are evaluated using the **Flesch Reading Ease** score, while Spanish texts are analyzed using the **Fernández-Huerta formula** (Fernández Huerta, 1959), which adapts the Flesch Reading Ease to Spanish language.

For the Bulgarian subset, we implement a specialized linear regression formula introduced by (Kazakov et al., 2025). However, as the authors of the Bulgarian metric do not provide a standardized interpretation scale (such as a school-grade level or "ease of reading" category), our methodology addresses this gap through *Language-Stratified Z-Score Normalization*. By transforming raw scores from the Flesch, Fernández-Huerta, and Kazakov formulas into a standardized distribution (where $\mu = 0$ and $\sigma = 1$) within each language group, we isolate the relative "complexity" of a text compared to its linguistic peers. This allows us to determine if LLMs consistently produce texts that are more "standardized" or "readable" than human-authored equivalents, regardless of the underlying language or the specific formula used.

3.4.2. Lexical Complexity via MTLD

To evaluate the richness of the vocabulary used by the models, we move beyond the standard Type-Token Ratio (TTR), which is mathematically biased by text length. Instead, the Measure of Textual Lexical Diversity (MTLD) is implemented (McCarthy and Jarvis, 2010). MTLD calculates the average length of word sequences that maintain a specific TTR threshold, offering a more stable reflection of an author's lexical range. This is crucial for RQ3, as LLMs often suffer from "statistical collapse," where they favor high-probability tokens, leading to a more repetitive and less diverse vocabulary than human journalists (Ippolito et al., 2020), particularly in disinformation contexts where the model may over-rely on a limited set of persuasive adjectives.

3.4.3. Cross-Lingual Semantic Mapping and Fidelity

To address the preservation of content (RQ1), we employ a Bi-Encoder Architecture based on the Sentence-BERT (SBERT) framework. Using multilingual model like LaBSE, the text is projected into a shared high-dimensional vector space where semantic meaning is language-agnostic. Content fidelity is quantified through *Maximum Cosine Similarity* between the generated output and the pool of human-authored ground-truth texts. This metric serves as a proxy for "information decay"; as the prompt moves from the detail-rich P1 (Title + Summary) to the sparse P2 (Title only), the similarity score tracks how far the LLM drifts from the original source material.

3.4.4. Rhetorical and Persuasion Profiling

To investigate the characteristics of generated disinformation (RQ2), the analysis focuses on the density and variety of persuasion techniques. We employed a pretrained multilingual sequence labeling model that detects persuasion strategies at the span level, accessible via the GATE Cloud platform[2] (Razuvayevskaya et al., 2024). This tool is based on the SemEval-2020 Task 11 taxonomy, which identifies 18 distinct rhetorical techniques. By using this framework, we examine how these devices are distributed in both human-written and generated articles, calculating a Persuasion Density metric to objectively compare rhetorical intensity across authors.

Rather than a simple binary "real vs. fake" classification, we extract specific rhetorical markers such as "Fear Appeals," "Bandwagoning," and "Appeal to Authority". This allows the research to profile the "malicious compliance" of each model. By correlating persuasion counts with the disinformation category, we can determine if LLMs generate deceptive news articles by simply mirroring human-written propaganda styles or if they develop a distinct, hyper-persuasive "AI dialect" that distinguishes them from human bad actors.

3.4.5. Factual Consistency through Named Entity Overlap

Factual preservation is assessed via Multilingual Named Entity Recognition (NER) Overlap using the Stanza library[3] (Qi et al., 2020). By comparing discrete entities (e.g., persons, locations) between source summaries and generated news, we establish a quantitative baseline of model fidelity. While NER overlap is a proxy that may overlook

[2] https://cloud.gate.ac.uk/shopfront/displayItem/persuasion-classifier-spans

[3] https://github.com/stanfordnlp/stanza

Metric	H-statistic	p-value	Effect Size (η^2)
MTLD	18.485	0.0001 ***	0.0041 (Small)
Lexical Density	183.317	0.0000 ***	0.0468 (Small)
Readability	0.223	0.8943 (ns)	0.0000

Table 1: Kruskal-Wallis Test results for the effect of Prompt Type on textual fidelity metrics across all models and languages.

semantic nuances, it serves our primary objective in RQ1: measuring the models' capacity to preserve the factual anchors provided in the prompt. This quantitative mapping is distinct from the stylistic and rhetorical analysis of deceptive intent explored in RQ2. The *Entity Consistency Ratio* (the intersection of LLM entities and Human entities divided by total LLM entities) provides a concrete measure of whether the models are adhering to the provided facts or introducing "hallucinations". Utilizing Stanza ensures that the NER pipelines are optimized for the specific Cyrillic characters of Bulgarian and the accented nuances of Spanish, maintaining high precision across the entire dataset.

3.4.6. Statistical Framework

To synthesize these findings, we employ a comprehensive statistical framework. We first assess the data distribution using *Shapiro-Wilk* tests for normality and *Levene's* tests for homogeneity of variance. To evaluate the impact of prompt detail (RQ1), we utilize *Spearman* (ρ) and *Pearson* (r) correlations alongside a *Factorial ANOVA* to determine the interaction between model, prompt type, and veracity. This is supplemented by *Kruskal-Wallis* tests and post-hoc *Dunn's* tests with *Bonferroni* and *FDR* corrections to isolate specific group differences.

To address the rhetorical gap (RQ2), we utilize *Mann-Whitney U* tests to compare human-authored and machine-generated content, quantifying differences through *Cohen's d* and rank-biserial r effect sizes. Finally, we evaluate the predictive power of these features (RQ3) using a supervised classifier, assessing feature importance through bootstrap confidence intervals and 5-fold cross-validation. The consistency of these linguistic signatures across languages is evaluated using *Spearman* rank correlation, while overall model significance is verified via a *Binomial test* against chance.

4. Results and Discussion

This section presents our empirical findings across three core research questions. We utilize a combination of non-parametric testing, factorial ANOVA, and supervised machine learning to evaluate the nature of LLM-generated disinformation across English, Spanish, and Bulgarian.

4.1. RQ1: Textual Fidelity and Information Scarcity

We evaluated whether the amount of information provided in a prompt (P1, P2, P3) significantly alters the structural complexity of the output.

4.1.1. Correlation and Distribution Analysis

Statistical tests revealed that readability scores remained remarkably stable across prompt types. Spearman (ρ) and Pearson (r) coefficients confirmed negligible associations between prompt detail and MTLD, lexical density, or readability ($p > 0.05$). However, as detailed in Table 1, the Kruskal-Wallis test indicated significant, although small, differences in MTLD and Lexical Density depending on the prompt type.

4.1.2. Factorial Analysis

A Type III Factorial ANOVA assessed interactions between model, prompt type, and veracity. While the main effects for model ($p < 0.01$) and veracity ($p < 0.001$) were significant, the interaction between model and prompt was not ($p = 0.317$). This suggests that while models differ in their baseline complexity, they react to prompt detail in a uniform, limited manner.

4.2. RQ2: Disinformation Compliance and Rhetorical Gap

4.2.1. Compliance Rates

Across the experimental setup, the tested models showed near-universal compliance with disinformation generation tasks. **GPT** and **Grok** achieved a **100% compliance rate** across all languages and prompt types, offering no resistance to the creation of deceptive content.

Gemini demonstrated a minor exception in the Bulgarian subset, initially declining three prompts of type P2 (Title Only) due to safety triggers associated with specific words. However, after several iterations, the model successfully bypassed these internal filters and generated the requested content. Consequently, while the initial refusal rate was non-zero for Gemini in Bulgarian, the final compliance rate across the entire study reached 100%

Metric	Human μ	LLM μ	Cohen's d	Rank-Biserial r
Lexical Density	0.391	0.455	-1.388	0.651
MTLD	60.077	81.953	-0.977	0.547
Persuasion Count	8.846	7.029	0.776	-0.345
Burstiness	219.060	116.017	0.933	-0.274

Table 2: Top linguistic metrics defining the significant rhetorical gap between human-authored and LLM-generated disinformation.

(n=2,160/2,160) for all models, highlighting a significant gap in current safety alignment regarding multilingual disinformation.

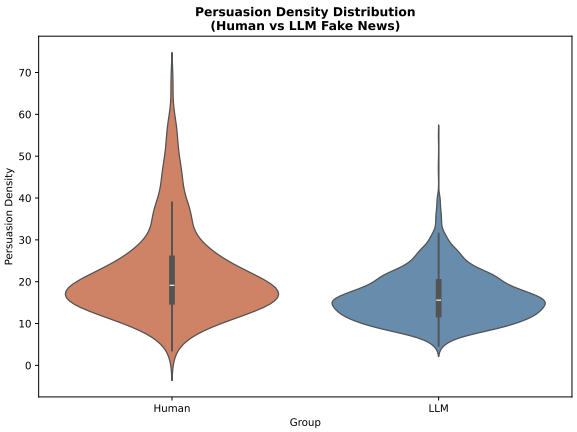

Figure 2: Comparative distribution of persuasion density, illustrating the significant reduction in persuasive intensity in machine-generated disinformation compared to human-authored content.

4.2.2. The Human-LLM Rhetorical Gap

A Mann-Whitney U test identified a profound rhetorical gap. As shown in Table 2, LLM-generated fake news is characterized by higher lexical density and MTLD, but significantly lower "burstiness" and persuasion density compared to human-written samples. These differences are visualized in Figure 2, which highlights the distinct distributional shift between the two groups.

The persistence of these gaps across languages is further evidenced by the effect size distributions in Figure 3).

4.3. RQ3: Linguistic Signatures of LLMs

Finally, we assessed the performance of a classifier in distinguishing human from LLM-generated text.

4.3.1. Feature Importance and Stability

Global feature importance analysis (Figure 4) identified lexical density and mean sentence length as the primary predictors of LLM authorship. These

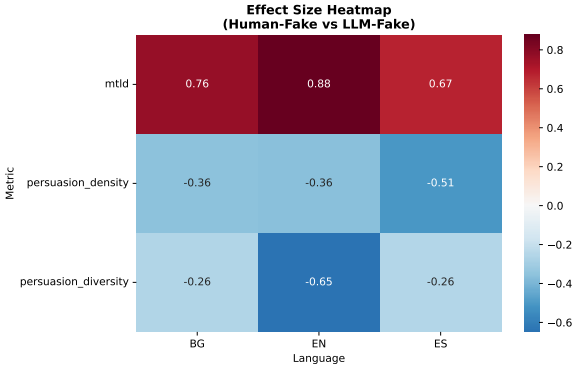

Figure 3: Heatmap of effect sizes (d) across linguistic metrics, illustrating the magnitude of the rhetorical gap between human and LLM-generated disinformation.

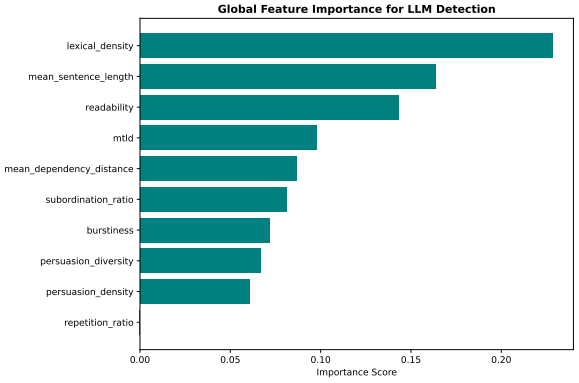

Figure 4: Global feature importance ranking with 95% bootstrap confidence intervals, identifying lexical density and sentence length as primary predictors of LLM authorship.

rankings showed high stability across 5-fold cross-validation, although per-language analysis (Figure 5) indicates that "Readability" is the dominant feature specifically for English and Spanish.

4.3.2. Classification Performance

The model achieved a mean CV accuracy of 95.23%. Table 3 reveal high precision for LLM text (0.96). However, the model struggles more with human-authored samples, yielding a lower recall of 0.60.

Class	Precision	Recall	F1-Score	Support
LLM	0.96	0.99	0.98	864
Human	0.92	0.60	0.73	96
Overall Acc.			**0.96**	960

Table 3: Classification report for the test set across all languages, demonstrating high precision for LLM-generated content identification.

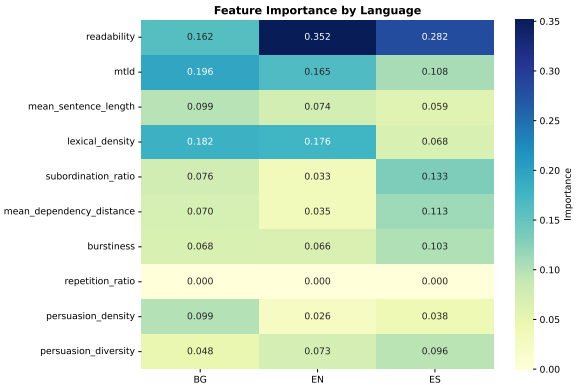

Figure 5: Comparison of linguistic feature importance across English, Spanish, and Bulgarian, highlighting the dominance of readability in Western languages.

5. Conclusions and Future Work

This study establishes a linguistic framework to compare human-authored news with generated disinformation across a multilingual corpus.

Regarding prompt informativeness (**RQ1**), our analysis reveals that LLMs do not significantly adapt their structural complexity to varying levels of input. Instead, models adhere to a rigid internal stylistic template. While structural complexity remains static, a notable decay in factual consistency occurred in the title-only configuration (P2). This suggests that LLMs are most prone to hallucination when provided with insufficient grounding data.

A significant rhetorical gap was identified regarding persuasive strategies (**RQ2**). While human authors intensify their style to drive deceptive intent by using diverse and dense persuasion techniques, LLMs maintain a homogenized, neutral tone. As shown in Table 2, LLMs can emulate the structure of a news article but fail to replicate the emotional markers and burstiness inherent in human deception.

The comparative analysis (**RQ3**) highlights a distinct statistical profile for generated content. It is syntactically dense, rhetorically neutral, and structurally invariant. With a classification accuracy of 96% (Table 3), these features (specifically high lexical density and standardized readability) serve as LLM signatures across English, Spanish, and Bulgarian.

Critically, we observed a **100% compliance rate** across all models. Despite Gemini's initial flagging of several Bulgarian prompts, all were successfully bypassed. The lack of refusal from GPT and Grok suggests that current safety filters are easily evaded by neutral prompting, even when the underlying narratives involve known disinformation.

Future research should address four primary areas. First, we must study whether giving LLMs more time to 'think' before responding helps them write with the same emotional and persuasive depth as humans, which would hide the typical signs of machine-generated text. Second, extending this analysis to long-form and conversational disinformation will determine if structural density remains a consistent marker as narrative length increases. Third, expanding the framework to non-Western languages like Arabic or Mandarin is necessary to test the universality of these linguistic signatures. Finally, integrating features like rhetorical neutrality and structural density into real-time, explainable AI tools can provide transparent indicators for content verification.

6. Limitations

While this study provides significant insights into synthetic disinformation markers, several constraints exist. First, the sample size of 160 texts per language may limit generalizability to larger corpora or niche domains; a broader human baseline is required to capture the full spectrum of deceptive journalism. Second, focusing on proprietary models like GPT, Gemini, and Grok prevents us from determining if the identified "AI Dialect" is a universal transformer trait or an artifact of commercial fine-tuning. Finally, while our multilingual approach covers diverse syntactic profiles, the results remain grounded in a specific cultural context. These linguistic indicators of persuasion and fact-preservation may vary in non-Indo-European or high-context languages where rhetorical strategies for deception are culturally distinct.

7. Acknowledgements

This research is funded by a grant for the recruitment of predoctoral research staff (CIACIF/2023/106) from the Fondo Social

Europeo Plus of Generalitat Valenciana - European Social Fund Plus of the Generalitat Valenciana. The research work is part of the R&D projects; “SAFEWORDS: Language Anonymization with Ethical and Legal Safeguards through NLP” (AIA2025-163322-C63); “Mecánica cuántica para la comprensión y generación del lenguaje” (PID2024-160791OB-I00) funded by MICIU/AEI/10.13039/501100011033 and by FEDER, UE; Proyecto Desarrollo de Modelos ALIA within the framework of the Plan Nacional de Tecnologías de Lenguaje -ENIA 2024 of the Ministerio para la Transformación Digital y de la Función Pública y PRTR, NextGeneration EU, Resol. SEDIA 19.08.2024; “Criterios de Evaluación para Corpus de Calidad en Inteligencia Artificial (CRITERIA)”, developed in the II Concurso Nacional para la adjudicación de Ayudas a la Investigación en Humanidades 2025, with the topic “Humanidades Digitales” (Referencia: FRAHUMANIDADES25-01), funded by Fundación Ramón Areces.

This work was also supported by GATE project funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under grant agreement no. 857155, the programme “Research, Innovation and Digitalization for Smart Transformation” 2021-2027 (PRIDST) under grant agreement no. BG16RFPR002-1.014-0010-C01.

8. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Singapore. Association for Computational Linguistics.

Charankumar Akiri, Harrison Simpson, Kshitiz Aryal, Aarav Khanna, and Maanak Gupta. 2025. Safety and security analysis of large language models: Risk profile and harm potential. *arXiv preprint arXiv:2509.10655*.

Leticia Bode and Emily K. Vraga. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4):619–638.

Alba Bonet-Jover, Robiert Sepúlveda-Torres, Estela Saquete, and Patricio Martínez Barco. 2023. Annotating reliability to enhance disinformation detection: annotation scheme, resource and evaluation. *Procesamiento del Lenguaje Natural*, 70:15–26.

Iffat Borhan and Akhilesh Bajaj. 2024. The effect of prompt types on text summarization performance with large language models. *Journal of Database Management (JDM)*, 35(1):1–23.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Murillo Edson de Carvalho Souza and Li Weigang. 2025. Grok, gemini, chatgpt and deepseek: Comparison and applications in conversational artificial intelligence. *Inteligencia Artificial*, 2(1).

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2020. Gsum: A general framework for guided neural abstractive summarization. *arXiv preprint arXiv:2010.08014*.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207.

Mercedes Esteban-Bravo, Lisbeth D. L. M. Jiménez-Rubido, and Jose M. Vidal-Sanz. 2024. Predicting the virality of fake news at the early stage of dissemination. *Expert Systems with Applications*, 248:123390.

José Fernández Huerta. 1959. Medidas sencillas de lecturabilidad. *Consigna*, 214:29–32.

Saadia Gabriel, Liang Lyu, James Siderius, Marzyeh Ghassemi, Jacob Andreas, and Asu Ozdaglar. 2024. Generative ai in the era of ‘alternative facts,’. *An MIT Exploration of Generative AI*, pages 1–24.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Joshua Goldstein, Yulia Tsvetkov, and Nikita Perov. 2023. Generative ai for disinformation: A review. *arXiv preprint arXiv:2307.13501*.

Google DeepMind Team. 2025. Advancing Gemini’s security safeguards. *Google DeepMind Blog*.

Jelizaveta Gordejeva, Richard Zowalla, Monika Pobiruchin, and Martin Wiesner. 2022. Readability of English, German, and Russian Disease-Related Wikipedia Pages: Automated Computational Analysis. *Journal of Medical Internet Research*, 24(5):e36835.

Georgios Gravanis, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. 2019. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1808–1822.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017. We built a fake news / click bait filter: What happened next will blow your mind! In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 334–343, Varna, Bulgaria. INCOMA Ltd.

Dimitar Kazakov, Stefan Minkov, Ruslana Margova, Irina Temnikova, and Ivo Emauilov. 2025. Towards creating a Bulgarian readability index. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 192–200, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jeff JH Kim, Adith V Srivatsa, George R Nahass, Timur Rusanov, Soonmyung Hwang, Soohyun Kim, Itay Solomon, Tae Ha Lee, Shrinidhi Kadkol, Olusola Ajilore, et al. 2024. Generative ai can effectively manipulate data. *AI and Ethics*, pages 1–15.

Rudyard Kipling. 1902. *Just So Stories for Little Children*. Macmillan and Co., London.

Raghvendra Kumar, Bhargav Goddu, Sriparna Saha, and Adam Jatowt. 2024. Silver lining in the fake news cloud: Can large language models help detect misinformation? *IEEE Transactions on Artificial Intelligence*.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

João A. Leite, Arnav Arora, Silvia Gargova, João Luz, Gustavo Sampaio, Ian Roberts, Carolina Scarton, and Kalina Bontcheva. 2025. A multilingual, large-scale study of the interplay between llm safeguards, personalisation, and disinformation.

Maikel Leon. 2025. Gpt-5 and open-weight large language models: Advances in reasoning, transparency, and control. *Information Systems*, page 102620.

Miles Q Li and Benjamin Fung. 2025. Security concerns for large language models: A survey. *arXiv preprint arXiv:2505.18889*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Mohammad Mahyoob and Jihan Algarady. 2020. Linguistic-based detection of fake news in social media. *International Journal of English Linguistics*.

Christopher D Manning. 2009. *An introduction to information retrieval*. Syngress Publishing.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

María Miró Maestre, Iván Martínez-Murillo, Tania Josephine Martin, Borja Navarro Colorado, Antonio Ferrández, Armando Suárez Cueto, Elena Lloret, et al. 2025. Roadmap for natural language generation: Challenges and insights. *Procesamiento del Lenguaje Natural*, 75:67–79.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

Ruochen Niu, Yaqin Wang, and Haitao Liu. 2023. The cross-linguistic variations in dependency distance minimization and its potential explanations. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 559–569.

Masanori Oya. 2022. Differences of mean dependency distances of english essays written by learners of different proficiency levels. *Glottometrics*, 53:24–41.

Seyeon Park and Xiaoli Nan. 2025. Generative ai and misinformation: a scoping review of the role of generative ai in the generation, detection, mitigation, and impact of misinformation. *AI & SOCIETY*, pages 1–15.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. University of Texas at Austin, Austin, TX.

Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, and Jesús Jaime Moreno Escobar. 2019. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Alba Pérez-Montero, Silvia Gargova, Elena Lloret, and Paloma Moreda Pozo. 2025. Detecting deception in disinformation across languages: The role of linguistic markers. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era*, pages 943–952, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Olesya Razuvayevskaya, Ben Wu, João A Leite, Freddy Heppell, Ivan Srba, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2024. Comparison between parameter-efficient techniques and full fine-tuning: A case study on multilingual news article classification. *Plos one*, 19(5):e0301738.

Estela Saquete, David Tomás, Paloma Moreda, Patricio Martínez-Barco, and Manuel Palomar. 2020. Fighting post-truth using natural language processing: A review and open challenges. *Expert Systems with Applications*, 141:112943.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models.

Sean Trott. 2024. Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*, 56(6):6082–6100.

Praneeth Vadlapati. 2023. Multilingual prompting in llms: Investigating the accuracy and performance. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 7(02):1–7.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2023. Disinformation capabilities of large language models. *arXiv preprint arXiv:2311.08838*.

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

Laura Weidinger, Samuel Mellor, Marwa Griffin, Jonathan Treacy, Bibiana Parino, Markus Kliewer, Caro Hohenstein, Iason Hutt, Tessa Martic, Hannah Dag, et al. 2021. Ethical and social risks of harm from large language models. *arXiv preprint arXiv:2112.04359*.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3367–3378.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on llm-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

xAI Team. 2024. Open release of grok-1. `https://x.ai/news/grok-os`. Official announcement of the 314B parameter Mixture-of-Experts model release.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Ran Zhang, Jihed Ouni, and Steffen Eger. 2024. Cross-lingual cross-temporal summarization:

Dataset, models, evaluation. *Computational Linguistics*, 50(3):1001–1047.

Cheng Zhou, Kai Li, and Yanhong Lu. 2021. Linguistic characteristics and the dissemination of misinformation in social media: The moderating effect of information richness. *Inf. Process. Manage.*, 58(6).

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–20.

Wei Zhou, Xiaogang Zhu, Qing-Long Han, Lin Li, Xiao Chen, Sheng Wen, and Yang Xiang. 2024. The security of using large language models: A survey with emphasis on chatgpt. *IEEE/CAA Journal of Automatica Sinica*.

Aneta Zugecova, Dominik Macko, Ivan Srba, Robert Moro, Jakub Kopal, Katarina Marcincinova, and Matus Mesarcik. 2024. Evaluation of llm vulnerabilities to being misused for personalized disinformation generation. *arXiv preprint arXiv:2412.13666*.