Can AI-Generated Persuasion Be Detected? Persuaficial Benchmark and AI vs. Human Linguistic Differences

Anonymous ACL submission

Abstract

Large Language Models (LLMs) can generate highly persuasive text, raising concerns about their misuse for propaganda, manipulation, and other harmful purposes. This leads us to our central question: *Is LLM-generated persuasion more difficult to automatically detect than human-written persuasion?* To address this, we categorize controllable generation approaches for producing persuasive content with LLMs and introduce Persuaficial, a high-quality multilingual benchmark covering six languages: English, German, Polish, Italian, French and Russian. Using this benchmark, we conduct extensive empirical evaluations comparing human-authored and LLM-generated persuasive texts. We find that although overtly persuasive LLM-generated texts can be easier to detect than human-written ones, subtle LLM-generated persuasion consistently degrades automatic detection performance. Beyond detection performance, we provide the first comprehensive linguistic analysis contrasting human and LLM-generated persuasive texts, offering insights that may guide the development of more interpretable and robust detection tools.

1 Introduction

Persuasive writing, which uses rhetorical techniques and devices to influence audiences, has become central to modern communication (Gass and Seiter, 2022). We live in an era where artificial intelligence increasingly shapes propaganda and persuasive communication in news, political discourse, and social media (Bergmanis-Korāts et al., 2024; Goldstein et al., 2024). Large Language Models, now widely used in writing and communication tasks, demonstrate a growing potential to produce persuasive text and influence public opinion (Pauli et al., 2025; Bai et al., 2025; Breum et al., 2024; Karinshak et al., 2023). Several studies have explored how effectively LLMs can identify persuasive language (Sprenkamp et al., 2023; Panasyuk, 2025). Yet, to the best of our knowledge, no prior work has addressed whether the automatic detection of LLM-generated persuasion is more challenging than detecting persuasion in human-written texts. Understanding this distinction is crucial, as it reveals the extent to which current detection systems may be vulnerable to increasingly sophisticated AI-driven persuasive content. Furthermore, although previous research has highlighted the importance of mitigating and defending against AI-generated persuasion (Burtell and Woodside, 2023; El-Sayed et al., 2024), it has largely focused on persuasion detection, leaving the linguistic differences between human-written and LLM-generated persuasive texts unexplored. Understanding these differences could deepen our knowledge of AI-driven persuasion and support the development of more effective automatic detection methods. To address these gaps, we investigate two key research questions: **RQ1** *Is controllably generated AI persuasion harder for LLMs to detect in a zero-shot setting than human-written persuasion?* and **RQ2** *What are the linguistic differences between controllable LLM-generated and human-written persuasive texts?*.

To address our first research question, we introduce **Persuaficial**, a newly constructed benchmark of artificially generated persuasive texts. Persuaficial is a novel multilingual resource comprising synthetic persuasive content produced using four generation approaches inspired by Chen and Shu (2023). The dataset is created in a controlled manner, leveraging human-written texts drawn from established datasets (Piskorski et al., 2023c; Tan et al., 2016; Moral et al., 2023). In our experiments, we evaluate the detectability of AI and human-written persuasive texts using four different LLMs, including commercial closed models and open-weight models. Our analysis focuses on English, but we also provide analysis on five additional languages: German, French, Italian, Polish, and Russian.

To address RQ2, we conducted an analysis using the StyloMetrix tool[1], which generates fully interpretable and reproducible vectors representing a wide range of linguistic features in text (Okulska et al., 2023). Prior work shows that StyloMetrix performs well on persuasion detection with classical machine learning (Modzelewski et al., 2023, 2024), underscoring its suitability for studying the linguistic features of persuasive texts. In our analysis, we examined the full range of linguistic features offered by open-source StyloMetrix.

Our main contributions are summarized as:

- We introduce Persuaficial, a novel persuasion benchmark with approximately 65k multilingual texts generated via four controllable approaches with four LLMs.
- We are the first to investigate whether persuasive text generated by LLMs is harder to detect than human-written persuasive text. We conduct this analysis across four LLMs and 16 controllable generation settings, providing a comprehensive evaluation of detection difficulty across diverse persuasive text generation approaches.
- Our work is the first to characterize the linguistic differences between LLM-generated and human-written English persuasive texts. Our analysis encompasses 196 distinct linguistic features.

We release our codebase, dataset and all prompts[2].

2 Human Persuasion Datasets

In our analysis, we employed three well-established datasets that had been previously annotated by humans. Using multiple datasets mitigates potential bias that could arise from relying on a single source and ensures a broader coverage of persuasion phenomena. We selected datasets that are widely used and cited in persuasion research. Below, we describe the human-created datasets used in our study:

- **SemEval 2023 Task 3 Dataset**: A multilingual, multifaceted collection of online news articles annotated with various persuasion techniques on paragraph level (Piskorski et al., 2023d). Its taxonomy and dataset are widely adopted within the NLP community for persuasion research (Barrón-Cedeño et al., 2024; Dimitrov et al., 2024; Modzelewski et al., 2025). This dataset was introduced as part of SemEval 2023 Task 3 on persuasion detection (Piskorski et al., 2023b).
- **DIPROMATS 2024 Task 1 Dataset**: A dataset consisting of posts from the X platform (former Twitter) used for the DIPROMATS 2024 shared task including propaganda detection (Moral et al., 2024, 2023). The dataset contains messages from diplomats and authorities of major world powers, including China, the United States, Russia, and the European Union. DIPROMATS 2024 was part of IberLEF, an annual Spanish shared evaluation campaign (Chiruzzo et al., 2024).
- **ChangeMyView**: A dataset derived from the Reddit *ChangeMyView* discussion community. Dataset contains 3,051 conversations in which the persuader tries to convince the persuadee to change their mind (Tan et al., 2016). It is commonly used in persuasion research (Ziems et al., 2024; Wei et al., 2016; Dutta et al., 2020).

Together, these datasets provide diverse perspectives on human persuasion, covering different sources and communicative contexts. Consequently, our analysis captures diverse persuasion understanding, which improves the generalizability of our findings. For further justification of our dataset choices, please refer to Appendix A.

3 Persuaficial: Artificially Generated Persuasion Dataset

In this section, we introduce a new dataset, **Persuaficial**, which contains persuasion texts generated artificially using large language models.

3.1 Persuasive Text Generation Approaches

Inspired by Chen and Shu (2023), which explores methods for generating synthetic misinformation, we developed a controllable persuasion generation G_P, in which P denotes one of four approaches for generating persuasive text: *Paraphrasing*, *Rewriting with Subtle Persuasion*, *Rewriting with Intensified Persuasion*, and *Open-ended* generation.

Each persuasive text was generated under controlled prompting conditions. Specifically, prompts included human-annotated persuasive texts together with instructions to paraphrase or rewrite the input while preserving, strengthening, or softening its persuasive effect. For the open-ended generation, following the approach of Chen and Shu (2023), the model was provided with concise summaries of the corresponding human-annotated persuasive examples and prompted to generate persuasive text freely based on summary. This procedure was applied to all selected instances across

[1] https://github.com/ZILiAT-NASK/StyloMetrix
[2] anonymous.4open.science/r/Persuaficial-26EC/

Approach	Prompt Example
Paraphrased Persuasion	Given a passage, please paraphrase it. The original content meaning and level of persuasion must be strictly preserved. The passage is: <passage>
Rewritten Persuasion (Subtle Effect)	Given a passage, please rewrite it to make its persuasiveness more subtle and harder to detect. The original content and meaning should be the same. The passage is: <passage>
Rewritten Persuasion (Intensified Effect)	Given a passage, please rewrite it so that the persuasion present in the content is refined for a stronger persuasive effect. The original content and meaning should be the same. The passage is: <passage>
Open-Ended Persuasion	Given a sentence, please write a short piece of text. The short text must contain persuasion. The sentence is: <sentence>

Table 1: Overview of four approaches used for generating persuasive texts with LLMs. Each method represents a distinct level of control over persuasive strength and content nature.

the chosen texts, using identical prompt templates for all languages. For non-English cases, we appended instructions specifying the target language of generation.

Controlling the generation process through explicit instructions is crucial for our study, as it ensures that the resulting LLM-generated persuasive texts remain semantically comparable to human-written texts. This comparability is essential for reliable evaluation of both persuasion detection performance and linguistic differences between human- and AI-generated persuasive texts.

All approaches for generating synthetic text with persuasion, along with example prompts, are included in Table 1 (details about our prompts presented in the Appendix B.3).

3.2 Persuaficial Dataset Construction

Persuaficial is an AI-generated persuasion dataset constructed using multiple LLMs and diverse generation approaches. For *Paraphrasing*, *Rewriting (Subtle Effect)*, and *Rewriting (Intensified Effect)*, we sample 1,000 texts from three real-world persuasion datasets (described in Section 2). Each sample includes 500 texts annotated as persuasive and 500 labeled as non-persuasive[3]. Each selected persuasive text is treated as <passage> (see Table 1) and serves as input for the generation method. For *Open-ended* generation, we first summarize each selected persuasive text into a factual statements. We use the resulting <sentence> (see Table 1) for the generation of persuasive synthetic text.

We employ open-weight and proprietary LLMs for dataset construction. The open models include *Gemma 3 27b it* and *Llama 3.3 70B*. The commercial models are *Gemini 2.0 Flash* and *GPT 4.1 Mini*. Additional details on dataset creation and hyperparameter settings are provided in Appendix B. Details about the models, APIs used, and the rationale for model selection is available in Appendix C.

3.3 Persuaficial Quality Evaluation

Pre-Generation Quality Evaluation. As mentioned, for the *Open-ended* generation setting, we first summarize each selected persuasive text into a short <sentence>. To ensure these <sentence>s accurately represent the source human text, we conducted a human evaluation following explicit and rigorous annotation guidelines (see Appendix D.1).

Two annotators were first trained by one of the authors, who has prior experience in annotation. A small training sample of 50 English sentences was selected, and the annotators independently applied the annotation guidelines, with opportunities to discuss their decisions. After completing the training phase, they reviewed and discussed their independent evaluations to align their understanding of the guidelines. The annotations from this training phase were excluded from further evaluation.

For the final evaluation, a sample of 200 English <sentence>s was selected. Two independent annotators assessed each <sentence> for factual correspondence. We then computed the accuracy of the LLM-generated <sentence>s, considering as positive only those instances where both annotators independently agreed that a sentence was factual. The resulting accuracy of the LLM generation process was about 91.2%, suggesting that LLMs may be effective at transforming texts into a short factual statements. Moreover, most mismatches between the generated <sentence>s and source texts were minor in nature, e.g., converting an exclamatory formulation ("*Introduce the law!*") into a declarative one ("*The law will be introduced.*"). This result suggests that the generated factual sentences are of generally high quality and are unlikely to negatively impact the overall quality of the resulting Persuaficial dataset and our experiments.

[3]The only exception was German: the corpus contained only 420 non-persuasive texts, so we included all of them and randomly sampled 580 persuasive texts.

Post-Generation Quality Evaluation. To ensure that the final Persuaficial dataset meets the intended goal of containing persuasive content produced by LLMs under controlled prompting conditions, we conducted a multi-stage post-generation quality evaluation. While the pre-generation evaluation ensured that the factual sentences for *Open-ended* approach were valid, the post-generation evaluation verifies whether the LLM-generated persuasive variants are (1) faithful to the target factual content, (2) persuasive, and (3) faithful to the instruction from persuasion generation approach.

We adopted a two-layer rigorous verification design that separates basic validity checking from persuasion-specific judging. We verified 400 generated English texts, each independently annotated by two trained annotators following detailed instructions (Appendix D.2). As a result, we report an overall accuracy metric, defined as the proportion of generated texts unanimously annotated as valid by two annotators, where validity required that all three criteria received a positive annotation.

Due to the conservative requirement that all three criteria be jointly satisfied, the overall accuracy is 88.2%. Most invalid cases involve only minor factual deviations rather than substantive inconsistencies. When considering persuasion-related criteria alone, accuracy grows to 97.69%, indicating that LLMs reliably generate persuasive text and justifying the use of this data for subsequent comparisons between human- and AI-generated persuasive texts.

3.4 Data Statistics

For each language, we sampled 1,000 human-written passages from the original datasets, including 500 persuasive texts. For each persuasive example, we generated AI-counterparts using four LLMs and four generation approaches (4 models × 4 approaches = 16 generation configurations). This resulted in approx. 24,000 texts in English and 41,000 texts for the non-English languages. Overall, Persuaficial is a multilingual corpus of about 65,000 texts. Table 2 presents basic statistics for our dataset. Detailed statistics in Appendix E.

4 Automatic Detection of Human and AI-Generated Persuasion

4.1 Experimental Setup

For our experiments, we use Persuaficial, which comprises artificially generated persuasive texts. Moreover, we use human-written counterparts. Each experiment uses data that is balanced across persuasive and non-persuasive classes.

For automatic persuasion detection, we employed four LLMs: *GPT-4.1 Mini*, *Gemini 2.0 Flash*, *Gemma 3 27B Instruct*, and *Llama 3.3 70B*. To ensure as deterministic outputs as possible, we set the temperature to 0 during classification. Since our goal is to detect persuasion, we formulate the task as a binary classification problem. All classifications were performed in a zero-shot setting. This approach aligns with our research question (RQ1). Moreover, studies show that zero-shot detection with modern LLMs (e.g., GPT-4) can outperform supervised models such as BERT on binary classification tasks (Pelrine et al., 2023; Bang et al., 2023; Hassan and Lee, 2020). Furthermore, Lucas et al. (2023) and Modzelewski et al. (2025) report that while fine-tuning BERT on multiple datasets results in poor generalization to unseen data, zero-shot LLMs maintain strong cross-domain performance. We evaluate persuasion detection performance using the F_1 score. Further details supporting reproducibility, including the LLM classifiers setup and the prompt templates used for persuasion detection, are provided in Appendix F.

4.2 Results on English Datasets

Table 3 reports F_1 scores for persuasion detection on three human-written balanced samples and their LLM-generated counterparts produced using four generation approaches.

On the *Paraphrasing* subset of our Persuaficial dataset, F_1 scores are only marginally lower than those for human-written texts (on average 0.67% lower), indicating that paraphrasing preserves a similar level of difficulty for persuasion detection across human and generated texts. In contrast, *Rewriting (Intensified)* and *Open-ended* subsets yield the highest F_1 scores. On average, persuasion is 9.75% easier to detect in open-ended scenario and 5.33% easier when persuasion is intensified. This makes open-ended generated persuasive texts the easiest setting for LLM-based detection. We hypothesize that models tend to over-express explicit persuasive cues when prompted to generate persuasive text freely or while intensifying persuasion, which in turn makes these texts more easily detectable. The opposite pattern emerges for *Rewriting (Subtle persuasion)*, where F_1 scores drop substantially, by 20.42% on average. This suggests that reducing overt persuasive markers makes persuasion significantly harder to detect, even for

Data Statistic	Human-written	Paraphrasing Generation	Rewriting Generation		Open-ended Generation
			Subtle Persuasion	Intensive Persuasion	
Average No. of Words	72	79	81	87	65
Average No. of Characters	452	538	568	605	450

Table 2: Basic statistics for human-written and for Persuaficial dataset, including LLM-generated persuasive texts.

strong LLM detectors. Importantly, these patterns are highly consistent across datasets and across all detector models. This may indicate that the effects generalize across domains and are independent of the specific LLM used for detection.

In summary, addressing RQ1, the detectability of LLM-generated persuasive text depends on the generation approach: texts produced via open-ended and intensified persuasion are easier to detect, whereas subtly persuasive generations remain substantially more challenging for current LLM-based detectors. Detailed results in Appendix G.1.

4.3 Results on Non-English Datasets

Table 4 shows persuasion detection results for German, French, Italian, Polish, and Russian. The patterns observed in English hold consistently across all languages and classifiers. Paraphrasing preserves a difficulty level similar to human-written texts, whereas intensified rewriting and open-ended generation yield the highest F_1 scores. Open-ended generation frequently yields F_1 scores above 0.9, indicating that persuasive text is easiest to detect in this setting. In contrast, subtle rewriting causes the largest drop in performance. These trends suggest that generation approaches influence persuasion detectability, with effects that may generalize across languages. Detailed results in Appendix G.2.

5 Linguistic Differences Between Machine and Human Persuasion

In this section, we investigate the linguistic differences between human-written and AI-generated persuasive texts. We focus on English due to the limited availability of high-quality datasets in other languages. While the SemEval 2023 Task 3 data provides a multilingual resource (Srba et al., 2024; Piskorski et al., 2023c), relying solely on it could introduce dataset-specific biases. To mitigate this, we use English part of Persuaficial and human-written counterparts from three well-established datasets.

5.1 Our Approach for Linguistic Analysis

To investigate the linguistic differences between human-written persuasive texts and LLM-generated persuasive texts, we adopt an explainable, feature-based analysis grounded in stylometry. Our objective is to identify the linguistic features that most strongly differentiate LLM-generated persuasive texts from human-written ones. For each linguistic feature, we compare the distributions of the two groups using effect-size-based analysis together with significance testing. Effect sizes quantify the magnitude of the difference between human and AI-generated texts for each feature (Frey, 2021), while significance testing evaluates whether these distributional differences are statistically meaningful. Our analysis identifies which linguistic properties differ systematically between human- and AI-produced persuasive texts.

5.2 Experimental Setup

We first represent each persuasive human text and its AI-generated counterpart using StyloMetrix (Okulska et al., 2023). For each text, we calculate a 196-dimensional vector of linguistic features. This results in a tabular representation, where each row corresponds to a text encoded by its computed linguistic features. We utilize StyloMetrix, because it is an open-source tool that provides fully interpretable, linguistically grounded feature representations. Moreover, prior work has demonstrated its effectiveness for persuasion detection using classical machine learning models (Modzelewski et al., 2023, 2024), confirming its suitability for analyzing the linguistic characteristics of persuasive texts. Finally, to further justify our choice, we show that StyloMetrix features with classical machine learning can distinguish human-written from AI-generated persuasive texts in Appendix H.

For each linguistic feature, we directly compare its distribution in human-written and LLM-generated persuasive texts, conducting this analysis separately for each generation approach. We utilize Cohen's d statistic (see Appendix I for definiton and how we computed it in our experiments) which is a type of effect size measure used to represent the magnitude of differences between two groups on a given variable (Frey, 2021). To evaluate whether feature distributions significantly differ between human and AI-generated texts, we perform Wilcoxon signed-rank tests (Wilcoxon, 1945),

Classifier Models	Human-written	Paraphrasing Generation	Rewriting Generation		Open-ended Generation
			Subtle Persuasion	Intensive Persuasion	
Sample of Persuaficial generated based on: SemEval 2023 data					
GPT 4.1 Mini	0.7398	0.7007 ↓5%	0.4031 ↓46%	0.8148 ↑10%	0.8964 ↑21%
Llama 3.3 70B	0.7459	0.7207 ↓3%	0.4577 ↓39%	0.8111 ↑9%	0.8741 ↑17%
Gemma 3 27b it	0.7572	0.7592 ↑0%	0.6453 ↓15%	0.8208 ↑8%	0.8562 ↑13%
Gemini 2.0 Flash	0.7551	0.7540 ↓0%	0.6522 ↓14%	0.7950 ↑5%	0.8117 ↑7%
Sample of Persuaficial generated based on: DIPROMATS 2024 data					
GPT 4.1 Mini	0.7567	0.7461 ↓1%	0.4962 ↓34%	0.8100 ↑7%	0.8666 ↑15%
Llama 3.3 70B	0.7471	0.7362 ↓1%	0.5696 ↓24%	0.7860 ↑5%	0.8292 ↑11%
Gemma 3 27b it	0.7473	0.7460 ↓0%	0.6349 ↓15%	0.7782 ↑4%	0.7994 ↑7%
Gemini 2.0 Flash	0.7518	0.7427 ↓1%	0.6664 ↓11%	0.7640 ↑2%	0.7680 ↑2%
Sample of Persuaficial generated based on: ChangeMyView data					
GPT 4.1 Mini	0.6233	0.6356 ↑2%	0.4906 ↓21%	0.6739 ↑8%	0.7148 ↑15%
Llama 3.3 70B	0.6517	0.6488 ↓0%	0.5536 ↓15%	0.6691 ↑3%	0.6831 ↑5%
Gemma 3 27b it	0.6644	0.6708 ↑1%	0.6334 ↓5%	0.6809 ↑2%	0.6843 ↑3%
Gemini 2.0 Flash	0.6671	0.6662 ↓0%	0.6294 ↓6%	0.6740 ↑1%	0.6770 ↑1%

Table 3: F_1 scores for persuasion detection on English data. The first column reports performance on human-annotated texts. The remaining columns show performance on LLM-generated texts. For generated data, each value represents the average F_1 score obtained from classification of texts generated by four different LLMs. Detailed results without averaging F_1 scores in Appendix G.1.

Classifier Models	Human-written	Paraphrasing Generation	Rewriting Generation		Open-ended Generation
			Subtle Persuasion	Intensive Persuasion	
German					
GPT 4.1 Mini	0.7203	0.7207 ↑0%	0.4410 ↓39%	0.8456 ↑17%	0.9414 ↑31%
Llama 3.3 70B	0.7361	0.7248 ↓2%	0.4398 ↓40%	0.8474 ↑15%	0.9345 ↑27%
Gemma 3 27b it	0.7655	0.7763 ↑1%	0.6664 ↓13%	0.8512 ↑11%	0.9004 ↑18%
Gemini 2.0 Flash	0.7903	0.7880 ↓0%	0.6905 ↓13%	0.8385 ↑6%	0.8591 ↑9%
French					
GPT 4.1 Mini	0.7505	0.7454 ↓1%	0.4290 ↓43%	0.8456 ↑13%	0.9251 ↑23%
Llama 3.3 70B	0.7605	0.7450 ↓2%	0.4527 ↓40%	0.8432 ↑11%	0.9172 ↑21%
Gemma 3 27b it	0.7827	0.7866 ↑0%	0.6587 ↓16%	0.8476 ↑8%	0.8824 ↑13%
Gemini 2.0 Flash	0.7812	0.7860 ↑1%	0.6800 ↓13%	0.8314 ↑6%	0.8418 ↑8%
Italian					
GPT 4.1 Mini	0.7471	0.7330 ↓2%	0.4246 ↓43%	0.8428 ↑13%	0.9195 ↑23%
Llama 3.3 70B	0.7584	0.7172 ↓5%	0.4285 ↓43%	0.8420 ↑11%	0.9161 ↑21%
Gemma 3 27b it	0.7659	0.7804 ↑2%	0.6610 ↓14%	0.8399 ↑10%	0.8686 ↑13%
Gemini 2.0 Flash	0.7986	0.7938 ↓1%	0.6781 ↓15%	0.8301 ↑4%	0.8408 ↑5%
Polish					
GPT 4.1 Mini	0.7330	0.7060 ↓4%	0.4580 ↓38%	0.8483 ↑16%	0.9367 ↑28%
Llama 3.3 70B	0.7676	0.7389 ↓4%	0.5041 ↓34%	0.8518 ↑11%	0.9206 ↑20%
Gemma 3 27b it	0.7728	0.7783 ↑1%	0.6919 ↓10%	0.8427 ↑9%	0.8834 ↑14%
Gemini 2.0 Flash	0.7733	0.7732 ↓0%	0.7018 ↓9%	0.8101 ↑5%	0.8217 ↑6%
Russian					
GPT 4.1 Mini	0.7246	0.7073 ↓2%	0.4392 ↓39%	0.8242 ↑14%	0.9017 ↑24%
Llama 3.3 70B	0.7408	0.7164 ↓3%	0.4312 ↓42%	0.8324 ↑12%	0.9086 ↑23%
Gemma 3 27b it	0.7360	0.7416 ↑1%	0.6128 ↓17%	0.8098 ↑10%	0.8562 ↑16%
Gemini 2.0 Flash	0.7683	0.7616 ↓1%	0.6795 ↓12%	0.7877 ↑3%	0.8019 ↑4%

Table 4: F_1 scores for persuasion detection on non-English data samples. The first column reports performance on human-annotated texts. The remaining columns show performance on LLM-generated texts. For generated data, each value represents the average F_1 score obtained from classification of texts generated by four different LLMs. Detailed results without averaging F_1 scores in Appendix G.2

a non-parametric paired test appropriate for comparing matched text pairs (Peyrard et al., 2021; Dror et al., 2018). Wilcoxon signed-rank tests have seen widespread adoption in the NLP community (Karwa and Singh, 2025; Zhou et al., 2025; Ciaccio et al., 2025), including in studies comparing feature distributions (Stodden and Kallmeyer, 2020).

5.3 Results and Analysis

We computed Cohen's d values for 196 linguistic features across four generation approaches and four LLMs utilized to generate texts for our Persuaficial dataset. Table 5 sorts top linguistic features by the absolute Cohen's d ($|C_d|$) values per model and generation approach (G_P). Our analysis and discussion is based on the twenty features with the largest $|C_d|$ for each model and generation strategy (more detailed tables available in Appendix J). Statistical analysis using Wilcoxon signed-rank tests confirmed that all twenty features for each scenario exhibit significant distributional differences between human-written and AI-generated persuasive texts. Figure 1 provides definitions of the key differentiating features.

High values of features such as L_CONT_T (the proportion of unique content-word forms relative to total tokens), LTOKEN_RATIO_LEM (the ratio of unique lemmas to total tokens), and L_CONT_A (the proportion of tokens that are content words)

G_ACTIVE The proportion of verbs in the text used in the active voice.
L_ADV_SUPERLATIVE Measures how often superlative adverbial (and some adjective-as-adverb) forms appear in a text.
L_ADV_COMPARATIVE The proportion of tokens that are adverbs used in comparative degree (e.g., "more", "less", or marked comparative forms)
L_FUNC_A The proportion of tokens in a text that are function words.
L_CONT_T The proportion of unique content-word forms in relative to total tokens.
L_CONT_A The proportion of tokens in a text that are content words.
L_PUNCT_COM Comma incidence measures frequency of commas relative to text length.
L_PUNCT_DASH Measures the density of dashes within a text.
L_PUNCT_DOT Measures the incidence of periods (dots) relative to the total number of words.
L_PLURAL_NOUNS Measures the density of plural nouns within a text.
LTOKEN_RATIO_LEM (ST_TYPE_TOKEN_RATIO_LEMMAS) The ratio of unique lemmas to total tokens.
POS_ADJ The proportion of tokens in the text that are adjectives, indicating the level of descriptiveness.
POS_NOUN The proportion of tokens in the text that are nouns.
POS_PRO The proportion of tokens in the text that are pronouns.
PS_CAUSE Measures the incidence of linking words and phrases related to cause and purpose.
SENT_D_NP (ST_SENT_D_NP) Measures the average proportion of noun phrase (NPs) tokens relative to sentence length, averaged over all sentences in a document.
SENT_D_PP (ST_SENT_D_PP) Measures the average proportion of tokens that belong to prepositional phrases (PPs) in each sentence, averaged over all sentences in the document.
SENT_D_VP (ST_SENT_D_VP) Measures the average proportion of tokens in a sentence that are not marked with a verb tense, relative to total sentence length, averaged over all sentences in the document.
SENT_ST_DIFFER (ST_SENT_DIFFERENCE) Quantifies syntactic variation between consecutive sentences by comparing their dependency label sets and averaging over the document.
SENT_ST_WPERSENT Indicates the normalized difference between the total number of tokens and the number of sentences in a document (a proxy for sentence length).
ST_REPET_WORDS (ST_REPETITIONS_WORDS) Measures the level of lexical repetition by computing the proportion of repeated word tokens in a text normalized by total token count.
SY_EXCLAMATION Measures the proportion of unique word tokens that appear in exclamatory sentences relative to all tokens in the text.
SY_IMPERATIVE Measures the proportion of unique alphabetic words that appear in sentences classified as imperative, relative to all tokens in the document.
SY_INV_PATTERNS (SY_INVERSE_PATTERNS) A syntactic feature that measures the frequency of inverted sentence structures within a text.
SY_NARRATIVE Measures the proportion of tokens in declarative sentences relative to all tokens.
VF_INFINITIVE A syntactic feature that measures the proportion of infinitive verb forms.
VT_MIGHT Measures the frequency of "might" in a text.

Figure 1: StyloMetrix features with the highest discriminative importance in distinguishing human-written from LLM-generated persuasive text.

indicate that AI-generated texts tend to contain more varied content words and higher informational density per sentence. These patterns suggest that lexical diversity and content richness are characteristic markers of AI authorship. Similarly, low values for ST_REPET_WORDS are indicative of AI-generated persuasive texts, suggesting that reduced word repetition serves as a strong signal of LLM-generated text. A higher proportion of function words (L_FUNC_A) indicates that a persuasive text is likely human-written. This means that frequent use of grammatical connectors (such as articles, prepositions, pronouns, and auxiliary verbs) is a signal of human text. Furthermore, AI-generated persuasive texts generally exhibit lower punctuation density, especially in texts from *Llama* and *GPT-4.1-mini*. However, certain punctuation marks, including commas (L_PUNCT_COM) and dashes (L_PUNCT_DASH), occur more frequently in these AI texts. The rarity of syntactically marked constructions, such as inversions (SY_INV_PATTERNS), is a distinguishing feature of AI text, as these complex syntactic patterns are more typical of human-written persuasive texts.

In AI-generated texts that intensify persuasion, comparative and superlative adverbs (L_ADV_COMPARATIVE with words like "more", "faster", and L_ADV_SUPERLATIVE with words like "best", "worst") appear more frequently than in human-written texts. This suggests that AI strengthens persuasive language through the increased use of adverbial modifiers, highlighting a distinctive stylistic strategy in LLM-generated intensified texts.

In AI-generated texts that aim to make persuasion more subtle, modal verbs such as "might" (VT_MIGHT) occur more frequently than in human-written texts. Similarly, narrative framing (SY_NARRATIVE), defined as the proportion of tokens in declarative sentences relative to all tokens, is more prevalent in AI subtle rewritings. These patterns may indicate that AI softens persuasion by using modal hedges and favors neutral declarative constructions over exclamatory sentences or rhetorical questions more often than humans do.

Open-ended AI-generated texts exhibit a consistent linguistic profile characterized by high lexical diversity and elevated content-word density (e.g., L_CONT_T, L_CONT_A, LTOKEN_RATIO_LEM). AI systems also show substantially lower function-word usage and reduced lexical repetition. In addition, AI-generated texts rely more heavily on imperative and infinitival constructions while avoiding marked syntactic patterns such as inversions, which may result in structurally simpler and more schematic syntax.

6 Related Work

Research on the persuasive capabilities of generative AI spans multiple disciplines, including computer science as well as the social and complexity sciences (Duerr and Gloor, 2021). Recent progress in large language models has drawn attention to their potential for persuasion and related applications (Jin et al., 2024; Rogiers et al., 2024). Early studies by Wang et al. (2019) explored personalized

Paraphrasing		*Rewriting for Subtle Effects*		*Rewriting for Intensified Effects*		*Open-ended*	
Feature Name	C_d	Feature Name	C_d	Feature Name	C_d	Feature Name	C_d
Generating Model: GPT 4.1 Mini							
L_CONT_T	0.51	L_CONT_T	0.70	L_CONT_T	0.77	L_CONT_T	1.62
L_CONT_A	0.46	L_CONT_A	0.62	L_CONT_A	0.77	L_CONT_A	1.43
SY_INV_PATTERNS	-0.45	VT_MIGHT	0.62	L_PUNCT_DASH	0.72	LTOKEN_RATIO_LEM	1.32
LTOKEN_RATIO_LEM	0.37	SY_INV_PATTERNS	-0.61	L_FUNC_A	-0.57	SENT_D_NP	1.05
L_FUNC_A	-0.36	L_PLURAL_NOUNS	0.54	LTOKEN_RATIO_LEM	0.51	ST_REPET_WORDS	-1.00
Generating Model: Llama 3.3 70B							
SENT_ST_WPERSENT	0.71	SY_INV_PATTERNS	-0.75	SENT_ST_WPERSENT	0.80	VF_INFINITIVE	1.34
SENT_ST_DIFFER	-0.70	G_ACTIVE	-0.72	L_ADJ_POSITIVE	0.72	SY_IMPERATIVE	1.24
L_PUNCT_COM	0.67	SENT_ST_WPERSENT	0.70	L_PUNCT_COM	0.71	G_ACTIVE	-0.91
SY_INV_PATTERNS	-0.61	L_CONT_T	0.62	L_CONT_T	0.66	SENT_D_VP	0.83
L_CONT_T	0.54	SENT_D_PP	0.62	POS_ADJ	0.65	L_PUNCT_DOT	-0.81
Generating Model: Gemma 3 27b it							
L_CONT_T	0.75	L_CONT_T	1.02	L_CONT_T	1.04	SY_IMPERATIVE	1.74
L_CONT_A	0.67	L_CONT_A	1.02	L_CONT_A	1.01	L_CONT_T	1.4
L_FUNC_A	-0.62	L_FUNC_A	-0.89	L_FUNC_A	-0.97	L_CONT_A	1.25
SY_INV_PATTERNS	-0.61	POS_PRO	-0.73	L_ADJ_POSITIVE	0.90	SENT_D_NP	1.16
L_ADV_SUPERLATIVE	0.43	SY_INV_PATTERNS	-0.71	POS_ADJ	0.83	PS_CAUSE	-1.15
Generating Model: Gemini 2.0 Flash							
L_CONT_A	0.73	L_CONT_A	0.92	L_CONT_A	1.02	L_CONT_T	1.57
L_CONT_T	0.71	L_CONT_T	0.92	L_CONT_T	0.99	L_CONT_A	1.37
SY_INV_PATTERNS	-0.58	L_FUNC_A	-0.73	L_FUNC_A	-0.82	SY_IMPERATIVE	1.32
L_FUNC_A	-0.54	SY_INV_PATTERNS	-0.72	L_ADJ_POSITIVE	0.78	LTOKEN_RATIO_LEM	1.20
L_PUNCT_COM	0.49	POS_NOUN	0.59	POS_ADJ	0.72	SY_EXCLAMATION	1.11

Table 5: Top linguistic features by absolute Cohen's d from four generation approaches for all generating LLMs.

persuasive dialogue systems designed to promote socially beneficial outcomes. Subsequent studies have investigated how individuals respond to persuasive machine-generated text and how they perceive its effectiveness (Karinshak et al., 2023; Bai et al., 2025; Goldstein et al., 2024). In addition, Schoenegger et al. (2025) explored whether LLMs can be more persuasive than humans, while Pauli et al. (2025) analyzed the extent to which LLMs are capable of generating persuasive language.

Parallel work has investigated approaches for detecting persuasion. Da San Martino et al. (2019) presented corpus of news annotated at the fragment level with 18 persuasive techniques and proposed multi-granularity neural network to detect persuasion. Piskorski et al. (2023d) extended the taxonomy of techniques proposed by Da San Martino et al. (2019) and presented a multilingual dataset. Moreover, persuasion detection was a task of different recognized workshops such SemEval or Slavic-NLP (Da San Martino et al., 2020; Dimitrov et al., 2021; Piskorski et al., 2023b, 2025).

To the best of our knowledge, no prior work has investigated whether LLM-generated persuasive texts are easier to automatically detect than human-written ones. Furthermore, existing research has not provided a linguistic analysis comparing LLM-generated and human persuasive content.

7 Conclusion

In this work, we introduce **Persuaficial**, a multilingual benchmark of LLM-generated persuasive texts, comprising about 65,000 instances across English, German, French, Italian, Polish, and Russian. Our experiments show that the detectability of persuasion in generated texts strongly depends on the generation approaches: open-ended and rewriting with intensified persuasion increase detectable cues, whereas rewriting with subtle persuasion substantially reduces detection performance. These trends are consistent across languages and classifier models, indicating that generation prompts may shape the difficulty of persuasive text detection.

Through a detailed analysis, we identify key linguistic differences between human and AI-generated persuasive texts. AI-generated texts tend to exhibit higher lexical diversity, increased content-word density, and lower function-word usage, while complex syntactic patterns are more characteristic of human persuasive writing. Text generation approaches further modulate these features: intensified persuasion amplifies adverbial modifiers, whereas subtle persuasion relies on modal hedges and declarative constructions.

Overall, our findings demonstrate that LLM-generated persuasive texts are not only linguistically distinct from human-written texts but also vary in detectability depending on the generation approach. **Persuaficial** provides a valuable resource for future research on automated persuasion detection, cross-lingual NLP, and the study of linguistic differences between human and AI-generated persuasive content.

8 Limitations

While **Persuaficial** offers a large and diverse resource for studying AI persuasive text detection, several limitations remain. First, although the corpus covers six languages, our linguistic analysis focuses only on English. This decision stems from the limited availability of high-quality, human-written persuasive datasets in other languages and from the need for a controlled, comparable setup across human and AI-generated texts. Conducting the analysis exclusively on English avoids introducing dataset-specific biases that would arise from relying on a single non-English persuasion dataset, ensuring that the linguistic findings are not driven by characteristics of a particular corpus.

In our analysis, all classifier evaluations are conducted in a zero-shot setting, which aligns with the goals and research question of this work. We chose not to explore few-shot or alternatives, leaving these as directions for future research. Moreover, prior studies show that modern LLMs in zero-shot mode (e.g., GPT-4) can outperform supervised models such as BERT on binary classification tasks (Pelrine et al., 2023; Bang et al., 2023; Hassan and Lee, 2020), and that fine-tuned BERT models may generalize poorly to out-of-domain data compared to zero-shot LLMs (Lucas et al., 2023; Modzelewski et al., 2025).

9 Ethics

Dataset The Persuaficial dataset consists of synthetic persuasive texts generated by large language models for research purposes. To construct the dataset, we relied exclusively on three established human-authored persuasion datasets that are either publicly available or were used with explicit permission from their original authors. No personally identifiable information is included in the generated persuasive content, and no attempt is made to identify or infer the authorship of individual texts.

All human-written and synthetic texts were used solely for academic research on persuasion detection and linguistic analysis. The generation process preserves the semantic content of the source material while producing novel text, thereby avoiding the reproduction of identifiable original passages. To promote transparency, reproducibility, and responsible reuse, the Persuaficial dataset will be released under the Creative Commons Attribution 4.0 (CC BY 4.0) license.

Crowdsourcing was not used at any stage of dataset creation or validation. All individuals involved in verifying the quality of Persuaficial were researchers or trained annotators with prior experience in persuasion or manipulation annotation. The verification process was conducted independently and remained free from political, institutional, or commercial influence.

Computational resources The use of large language models is associated with non-trivial computational and environmental costs (Strubell et al., 2019). In this work, we mitigate these costs by avoiding model training or fine-tuning and relying exclusively on inference with pre-existing models. All experiments were conducted through third-party APIs, and we did not directly manage or allocate the underlying computational infrastructure. As a result, the overall computational footprint of this study was limited to inference-time usage.

References

Hui Bai, Jan G Voelkel, Shane Muldowney, Johannes C Eichstaedt, and Robb Willer. 2025. Llm-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, and 1 others. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718.

Alberto Barrón-Cedeño, Firoj Alam, Julia Maria Struß, Preslav Nakov, Tanmoy Chakraborty, Tamer Elsayed, Piotr Przybyła, Tommaso Caselli, Giovanni Da San Martino, Fatima Haouari, and 1 others. 2024. Overview of the clef-2024 checkthat! lab: checkworthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–52. Springer.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Gundars Bergmanis-Korāts, Tetiana Haiduchyk, and Artur Shevtsov. 2024. Ai in precision persuasion: Unveiling tactics and risks on social media. Technical report, NATO Strategic Communications Centre of Excellence, Riga, Latvia. Prepared and published

by the NATO Strategic Communications Centre of Excellence, 51 pp.

Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 152–163.

Matthew Burtell and Thomas Woodside. 2023. Artificial influence: An analysis of ai-driven persuasion. *arXiv preprint arXiv:2303.08721*.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

Luis Chiruzzo, Salud María Jiménez-Zafra, and Francisco Rangel. 2024. Overview of iberlef 2024: natural language processing challenges for spanish and other iberian languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS. org*.

Cristiano Ciaccio, Alessio Miaschi, and Felice Dell'Orletta. 2025. Evaluating lexical proficiency in neural language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1267–1286.

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

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Sebastian Duerr and Peter A Gloor. 2021. Persuasive natural language generation–a literature review. *arXiv preprint arXiv:2101.05786*.

Subhabrata Dutta, Dipankar Das, and Tanmoy Chakraborty. 2020. Changing views: Persuasion modeling and argument extraction from online discussions. *Information Processing & Management*, 57(2):102085.

Seliem El-Sayed, Canfer Akbulut, Amanda McCroskery, Geoff Keeling, Zachary Kenton, Zaria Jalan, Nahema Marchal, Arianna Manzini, Toby Shevlane, Shannon Vallor, and 1 others. 2024. A mechanism-based approach to mitigating harms from persuasive generative ai. *arXiv preprint arXiv:2404.15058*.

Bruce B Frey. 2021. *The SAGE encyclopedia of research design*. Sage Publications.

Robert H Gass and John S Seiter. 2022. *Persuasion: Social influence and compliance gaining*. Routledge.

Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. How persuasive is ai-generated propaganda? *PNAS nexus*, 3(2):pgae034.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520.

Fuad Mire Hassan and Mark Lee. 2020. Political fake statement detection via multistage feature-assisted neural modeling. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across diverse domains: a dataset and persuasion large language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.

Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. 2023. Working with ai to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–29.

Saniya Karwa and Navpreet Singh. 2025. Disentangling linguistic features with dimension-wise analysis of vector embeddings. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 461–488.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation. *arXiv preprint arXiv:2310.15515*.

Arkadiusz Modzelewski, Paweł Golik, and Adam Wierzbicki. 2024. Bilingual propaganda detection in diplomats' tweets using language models and linguistic features. *IberLEF@ SEPLN*.

Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. 2025. Pcot: Persuasion-augmented chain of thought for detecting fake news and social media disinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24959–24983.

Arkadiusz Modzelewski, Witold Sosnowski, Magdalena Wilczynska, and Adam Wierzbicki. 2023. DSHacker at SemEval-2023 task 3: Genres and persuasion techniques detection with multilingual data augmentation through machine translation and text generation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, Toronto, Canada. Association for Computational Linguistics.

Pablo Moral, Jesús M Fraile, Guillermo Marco, Anselmo Peñas, and Julio Gonzalo. 2024. Overview of dipromats 2024: Detection, characterization and tracking of propaganda in messages from diplomats and authorities of world powers. *Procesamiento del lenguaje natural*, 73:347–358.

Pablo Moral, Guillermo Marco, Julio Gonzalo, Jorge Carrillo-de Albornoz, and Iván Gonzalo-Verdugo. 2023. Overview of dipromats 2023: automatic detection and characterization of propaganda techniques in messages from diplomats and authorities of world powers. *Procesamiento del lenguaje natural*, 71:397–407.

Inez Okulska, Daria Stetsenko, Anna Kołos, Agnieszka Karlińska, Kinga Głąbińska, and Adam Nowakowski. 2023. Stylometrix: An open-source multilingual tool for representing stylometric vectors. *arXiv preprint arXiv:2309.12810*.

Aleksey Panasyuk. 2025. Synthclassify: an llm-driven framework for generating and classifying persuasive text. In *Disruptive Technologies in Information Sciences IX*, volume 13480, pages 120–148. SPIE.

Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. 2025. Measuring and benchmarking large language models' capabilities to generate persuasive language. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10056–10075.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6399–6429.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than average: Paired evaluation of nlp systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2301–2315.

Jakub Piskorski, Dimitar Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michał Marcińczuk, Arkadiusz Modzelewski, Ivo Moravski, and 1 others. 2025. Slavicnlp 2025 shared task: Detection and classification of persuasion techniques in parliamentary debates and social media. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 254–275.

Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, and 1 others. 2023a. News categorization, framing and persuasion techniques: Annotation guidelines. *European Commission, Ispra, JRC132862*.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023b. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023c. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023d. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022.

Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. 2024. Persuasion with large language models: a survey. *arXiv preprint arXiv:2411.06837*.

Philipp Schoenegger, Francesco Salvi, Jiacheng Liu, Xiaoli Nan, Ramit Debnath, Barbara Fasolo, Evelina Leivada, Gabriel Recchia, Fritz Günther, Ali Zarifhonarvar, and 1 others. 2025. Large language models are more persuasive than incentivized human persuaders. *arXiv preprint arXiv:2505.09662*.

Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. 2023. Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*.

Ivan Srba, Olesya Razuvayevskaya, João Augusto Leite, Róbert Móro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, and 1 others. 2024. A survey on automatic credibility assessment of textual credibility signals in the era of large language models. *CoRR*.

Regina Stodden and Laura Kallmeyer. 2020. A multilingual and cross-domain analysis of features for text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 77–84.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3645–3650.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Shahadat Uddin and Haohui Lu. 2024. Confirming the statistically significant superiority of tree-based machine learning algorithms over their counterparts for tabular data. *Plos one*, 19(4):e0301541.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Berlin, Germany. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Jinfeng Zhou, Yuxuan Chen, Yihan Shi, Xuanming Zhang, Leqi Lei, Yi Feng, Zexuan Xiong, Miao Yan, Xunzhi Wang, Yaru Cao, Jianing Yin, Shuai Wang, Quanyu Dai, Zhenhua Dong, Hongning Wang, and Minlie Huang. 2025. SocialEval: Evaluating social intelligence of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30958–31012, Vienna, Austria. Association for Computational Linguistics.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Human Dataset - Rationale Behind Our Choice

In our experiments, we adopt the concise definition of persuasion proposed by Piskorski et al. (2023d,a): "*Persuasive text is characterized by a specific use of language in order to influence the reader*". We rely on this definition because it underpins the annotation guidelines of the SemEval 2023 Task 3 dataset, the largest publicly available resource for studying persuasion and one of the three human-written datasets used in our study. Consequently, we selected additional datasets that align well with this conceptualization of persuasion.

DIPROMATS 2024 (Moral et al., 2024), which build on the SemEval task, offers data that is directly compatible with this definition and is therefore suitable for our experiments. Finally, Piskorski et al. (2025) demonstrate that this definition can be effectively and reliably applied to debates and discussions, motivating our choice of the ChangeMyView dataset, comprising message exchanges between persuaders and persuadees, as an additional human-written source.

B Persuaficial Generation Process

B.1 Human-written Text Sample Selection

For each of the eight source datasets (DIPROMATS 2024, ChangeMyView, and SemEval 2023 Task 3 with six languages), we generated samples from a base of 500 persuasive human-written texts. By applying four generation approaches across four different LLMs, we produced 16 distinct machine-generated counterparts for every source text. Moreover, each sample contains 500 non-persuasive human-written texts.

B.2 Experimental Setup for LLM Persuasion Generation Process

We employed four Large Language Models. The open models include *Gemma 3 27b it* and

Llama 3.3 70B. The commercial models are *Gemini 2.0 Flash* and *GPT 4.1 Mini*. To encourage more diverse, creative and less repetitive phrasing in the model outputs, we set the generation temperature to 0.8. Our choice of temperature for generating synthetic persuasive texts was directly informed by the settings used by Chen and Shu (2023) in a related task involving misinformation generation.

B.3 Prompt Templates for Persuaficial Dataset Creation

Figures 2, 3, 4, and 5 show prompt templates used during the LLM persuasion generation process. In our prompts, we adopt the concise definition of persuasion proposed by Piskorski et al. (2023d,a): "*Persuasive text is characterized by a specific use of language in order to influence the reader*".

C Details on LLMs used in Experiments and selection rationale.

In our experiments, we employed four state-of-the-art LLMs: *GPT-4.1 Mini*, *Gemini 2.0 Flash*, *Gemma 3 27B-IT*, and *Llama 3.3 70B*. Our selection aimed to cover widely recognized, high-performing models while balancing accessibility and cost. Additionally, we included two open-weight models to provide experiments that can be reproduced without reliance on closed API-based models. Table 6 summarizes the LLMs used, including their knowledge cutoff dates, access methods, licenses, and model sizes.

D Annotation Guidelines for Dataset Evaluation

D.1 Annotation Guidelines for Sentences Verification

Purpose of the Annotation Task. The goal of this annotation task is to evaluate whether each LLM-generated <sentence> accurately reflects content present in its corresponding source human text. Annotators must independently judge whether the <sentence> faithfully reflects information explicitly stated in the source text, without adding, or altering factual content.

General Annotation Procedure.

1. Read the source persuasive human text in full to understand its factual content and context.
2. Read the generated <sentence> carefully and evaluate it against the factual correspondence.
3. Assign one binary label:
   Factual? Yes (1) / No (0)
4. Do not consider any stylistic preferences, or grammar.

Annotators should make decisions independently, without discussing individual cases during the evaluation phase.

Factual Correspondence Annotation.

1. All information in the <sentence> is explicitly stated in the source text.
   - No invented facts.
   - The <sentence> does not introduce generalizations (e.g., Fact present in a source text: "Adam Smith fainted after COVID-19 vaccination" → invalid <sentence>: "People fainted after COVID-19 vaccination")
   - No added assumptions or interpretations.
2. No main factual information from the source text is omitted in a way that distorts meaning.
3. The <sentence> is neutral and descriptive - Its purpose must be to summarize factual content, not to evaluate, interpret, or advise.
4. Statements must be verifiable based solely on the source text. Annotators should not use outside knowledge.

Examples of factual errors (should be labeled "No"):

- Adding additional events or statistics not in the source
- Reframing a claim as a fact (e.g., converting someone's opinion into an asserted truth)
- Omitting a main fact presented in source text that changes meaning.

D.2 Annotation Guidelines for Persuaficial Dataset Evaluation

These guidelines describe the annotation protocol for evaluating LLM-generated persuasive texts in the Persuaficial dataset. Each generated text is independently annotated by two annotators. The post-generation evaluation focuses on three key dimensions:

- Factual Correspondence: Is the generated text faithful to the target factual content?
- Persuasiveness: Does the text contain genuine persuasion?
- Instruction Adherence: Does the text follow the specific persuasion instruction for its generation approach?

Generation approach: Paraphrasing Generation prompt

System Prompt:
You are an assistant helping researchers study persuasion. Your task is to paraphrase provided texts so that the meaning, and original persuasion techniques are strictly preserved. Do not make the provided texts less or more persuasive. Keep the same level of persuasion. Definition of persuasion is as follows: Persuasive text is characterized by a specific use of language in order to influence readers. We distinguish the following high-level persuasion approaches: Attack on reputation, Justification, Simplification, Distraction, Call, Manipulative wording. You will receive the original passage. Then, paraphrase the passage, ensuring the paraphrased version contains the exact same level of persuasion and meaning. Your final output should only be the paraphrased text.
User Prompt:
Only paraphrase the following passage in <*language*>
The passage is: <*Human-written, persuasive text to paraphrase*>
Keep the same level of persuasion. Provide the paraphrased text in <*language*>, in valid JSON format:

```
{
   "generated_text":"Your paraphrased text in <language> here."
}
```

Figure 2: Prompt template used for persuasive texts generation with LLMs using the *Paraphrasing Generation* approach.

Generation approach: Rewriting Generation - Subtle Persuasion prompt

System Prompt:
You are an assistant helping researchers study persuasion. Your task is to rewrite provided texts so that the persuasion is more subtle and harder to detect, while strictly preserving all original persuasion techniques. Definition of persuasion is as follows: Persuasive text is characterized by a specific use of language in order to influence readers. We distinguish the following high-level persuasion approaches: Attack on reputation, Justification, Simplification, Distraction, Call, Manipulative wording. You will receive the original passage. Then, rewrite the passage, ensuring the rewritten version contains the exact meaning. Your final output should only be the rewritten text.
User Prompt:
Only rewrite the following passage in <*language*> **so that the persuasion is more subtle and harder to detect.**
The passage is: <*Human-written, persuasive text to rewrite*>
Provide the rewritten text in <*language*>, in valid JSON format:

```
{
   "generated_text":"Your rewritten text in <language> here."
}
```

Figure 3: Prompt template used for persuasive texts generation with LLMs using the *Rewriting Generation - Subtle Persuasion* approach.

API Model Name	Knowledge Cutoff Date	Access Details	License	Model Size
`gemini-2.0-flash`	June 2024	Google API 07.2025	Commercial	Not Disclosed
`gpt-4.1-mini-2025-04-14`	June 2024	OpenAI API 07.2025	Commercial	Not Disclosed
`meta-llama/Llama-3.3-70B-Instruct`	December 2023	DeepInfra API 07.2025	Meta Llama 3 Community	70B
`google/gemma-3-27b-it`	August 2024	DeepInfra API 07.2025	Gemma Terms of Use	27B

Table 6: Large Language Models used in our experiments.

Factual Correspondence. Goal of this step is to ensure the generated text accurately reflects the source content.

Instructions:

- Open-Ended Generation: Refer to the factual sentence.
- Paraphrasing / Rewriting Approaches: Refer to the original passage.

Assessment:

- Valid (represented as 1): Text preserves the factual meaning of the source without introducing errors or contradictions.
- Invalid (represented as 0): Text contains factual inaccuracies, omissions, or misrepresentations.

Note: Only factual distortion triggers an Invalid label.

Persuasiveness The generated text must contain any persuasive elements.

For this task, we define persuasive text as text characterized by a specific use of language in order to influence readers (Piskorski et al., 2023a,d). The generated text must be labeled as persuasive (represented as 1) if it exhibits any of the following high-level persuasion strategies:

- Attack on reputation: the argument does not address the topic itself, but targets the participant (personality, experience, deeds, etc.) in order to question and/or to undermine his credibility. The object of the argumentation can also refer to a

Generation approach: Rewriting Generation - Intensive Persuasion prompt

System Prompt:
You are an assistant helping researchers study persuasion. Your task is to rewrite provided texts so that the persuasion is refined for stronger persuasive effect, while strictly preserving all original persuasion techniques. Definition of persuasion is as follows: Persuasive text is characterized by a specific use of language in order to influence readers. We distinguish the following high-level persuasion approaches: Attack on reputation, Justification, Simplification, Distraction, Call, Manipulative wording. You will receive the original passage. Then, rewrite the passage, ensuring the rewritten version contains the exact meaning. Your final output should only be the rewritten text.
User Prompt:
Only rewrite the following passage in <*language*> **so that the persuasion is refined for stronger persuasive effect.**
The passage is: <*Human-written, persuasive text to paraphrase*>
Provide the rewritten text in <*language*>, in valid JSON format:

```
{
   "generated_text":"Your rewritten text in <language> here."
}
```

Figure 4: Prompt template used for persuasive texts generation with LLMs using the *Rewriting Generation - Intensive Persuasion* approach.

group of individuals, an organization, an object, or an activity,
- Justification: the argument is made of two parts, a statement and an explanation or appeal, where the latter is used to justify and/or to support the statement,
- Simplification: the argument excessively simplifies a problem, usually regarding the cause, the consequence, or the existence of choices,
- Distraction: the argument takes focus away from the main topic or argument to distract the reader,
- Call: the text is not an argument but an encouragement to act or to think in a particular way,
- Manipulative wording: the text is not an argument per se, but uses specific language, which contains words or phrases that are either non-neutral, confusing, exaggerating, loaded, etc., in order to impact the reader emotionally.

If any of these strategies are present, the sentence must be labeled 1 (persuasive) for the persuasiveness criterion.

Instruction Adherence. The goal is to verify that the text aligns with the intended generation approach.

Instructions for Annotators:

1. Compare the generated text to the prompt provided to the model.
2. Label Compliant (represented as 1) if the text follows the prompt goal; Non-Compliant (represented as 0) if it deviates.

E Persuaficial Dataset - Additional Statistics

Table 7 summarizes the basic statistics of both human-written and LLM-generated texts in the Persuaficial dataset. The table reports average word, average characters and number of words across all languages and generation types. Moreover, we show statistics in general for full Persuaficial dataset.

F Persuasion Detection - Experimental Setup Details

F.1 Evaluation Text Sample Creation

Our evaluation framework comprised a total of 68 distinct classification experiments for each of the eight source datasets (eight as SemEval data can be counted as 6 datasets each in different language). This setup involved testing every combination of four generation approaches and four generating LLMs. The resulting 16 sets of machine-generated text, along with a baseline of human-written persuasive text, were then evaluated by four different classifying LLMs, leading to the $(16 + 1) \times 4 = 68$ experimental conditions.

Our evaluation framework is designed to isolate the impact of these generated texts. Each experiment's evaluation set is composed of two halves:
- A constant set of 500 human-written, non-persuasive texts from the original dataset. Predictions for this set were calculated once for each model and reused across all experiments for that dataset.
- A variable set of 500 persuasive texts, which consists of the LLM-generated samples for a given generation approach.

Consequently, any variation in the F_1 score between different generation models on the same dataset is attributable solely to the model's performance on the generated persuasive samples. For

Generation approach: Open-ended Generation prompt

System Prompt:
You are an assistant helping researchers study persuasion. Your task is to generate a short text based on a provided passage. The short text must contain persuasion. Definition of persuasion is as follows: Persuasive text is characterized by a specific use of language in order to influence readers. We distinguish the following high-level persuasion approaches: Attack on reputation, Justification, Simplification, Distraction, Call, Manipulative wording. Your final output should only be the generated text.
User Prompt:
Generate a text in <*language*> based on the following passage in <*language*>.
The passage is: <*Summarized, factual, and non-persuasive input sentence*>
The generated text must contain persuasion. Provide the generated text in valid JSON format:

```
{
   "generated_text":"Your generated text in <language> here."
}
```

Obtaining a summarized factual input sentence prompt

System Prompt:
You are a journalist assistant. Your task is to convert the provided text passage into a direct, single-sentence text. Do not add context such as 'The speaker said...', 'The passage is about...', 'The statement suggests...', etc. Keep the meaning intact but make it stand alone. Do not add any additional information or actors.
User Prompt:
Restate the following passage in <*language*> as a single-sentence, neutral text in <*language*>.
The passage is: <*Human-written, persuasive text to summarize*>
Return in valid JSON format:

```
{
   "generated_text":"Your restated sentence in <language> here."
}
```

Figure 5: Prompt template used for persuasive texts generation with LLMs using the *Open-ended Generation* approach along with the prompt template used to obtain a summarized, factual, and non-persuasive sentence input from human-written persuasive texts.

our baseline experiments, labeled 'Human-written', these persuasive samples are the original human-written texts from the dataset.

F.2 Experimental Setup for LLM Classification Process

We employed four Large Language Models. The open models include *Gemma 3 27b it* and *Llama 3.3 70B*. The commercial models are *Gemini 2.0 Flash* and *GPT 4.1 Mini*. To ensure determinism in the classification predictions, we set the classification temperature to 0.0. Rationale for models selection provided in Appendix C.

F.3 Prompt Templates for Persuasion Detection

Figure 6 shows a prompt template used during the LLM persuasion detection process. In our prompts, we adopt the concise definition of persuasion proposed by Piskorski et al. (2023d,a): "*Persuasive text is characterized by a specific use of language in order to influence the reader*".

G Human vs. AI-generated Persuasion Detection - Detailed Results

G.1 Detailed results on English texts of Persuaficial dataset

In this appendix, we present detailed F_1 scores for persuasion detection across different subsets of the Persuaficial datasets and their human-written counterparts. Results are presented for each LLM generation strategy. Table 8 reports results for SemEval 2023 Task 3 texts and their AI-generated counterparts, Table 9 for the DIPROMATS 2024 dataset, and Table 10 for the ChangeMyView dataset. For each dataset, classifier performance on human-written texts (first column) is compared with performance on LLM-generated texts produced via paraphrasing, rewriting with subtle or intensive persuasion, and open-ended generation. Results are further broken down by both the generating model and the classifier model, highlighting how different generation approaches influence the detectability of persuasion.

Binary Detection of Persuasion prompt

System Prompt:
You are an assistant who detects persuasion in text. Persuasive text is characterized by a specific use of language in order to influence readers. We distinguish the following high-level persuasion approaches: Attack on reputation, Justification, Simplification, Distraction, Call, Manipulative wording. You are the expert who detects high-level persuasion.
User Prompt:
Analyze the following passage: *<Text to analyze>*
Decide if the passage contains persuasion. You are very conservative in your final decisions and when you are not fully sure you answer 'No'. Do not provide any additional text, just JSON. Give only your final answer 'Yes' or 'No' in valid JSON format:

```
{
   "decision": "'Yes' if passage contains persuasion, 'No' otherwise."
}
```

Figure 6: Prompt template used for binary classification of persuasive texts with LLMs.

dataset	type	Avg_w	Avg_{ch}	Count
Persuaficial	Rewriting (intensified)	87	605	16320
	Open-ended	65	450	16320
	Paraphrasing	81	538	16320
	Rewriting (subtle)	87	568	16320
English	Human	117	695	3000
	Rewriting (intensified)	118	791	6000
	Open-ended	60	391	6000
	Paraphrasing	111	727	6000
	Rewriting (subtle)	110	742	6000
French	Human	46	288	1000
	Rewriting (intensified)	75	498	2000
	Open-ended	76	498	2000
	Paraphrasing	66	430	2000
	Rewriting (subtle)	69	465	2000
German	Human	44	314	1000
	Rewriting (intensified)	68	512	2320
	Open-ended	71	524	2320
	Paraphrasing	59	434	2320
	Rewriting (subtle)	63	476	2320
Italian	Human	48	313	1000
	Rewriting (intensified)	75	511	2000
	Open-ended	75	507	2000
	Paraphrasing	66	443	2000
	Rewriting (subtle)	70	488	2000
Polish	Human	46	327	1000
	Rewriting (intensified)	70	526	2000
	Open-ended	61	454	2000
	Paraphrasing	62	459	2000
	Rewriting (subtle)	66	493	2000
Russian	Human	40	285	1000
	Rewriting (intensified)	55	417	2000
	Open-ended	58	442	2000
	Paraphrasing	48	360	2000
	Rewriting (subtle)	52	400	2000

Table 7: Basic statistics for human-written and for LLM-generated persuasive texts in our Persuaficial dataset. We present basic statistics for general full dataset, but also on samples that present all languages. Avg_w stands for average words and Avg_{ch} stands for average characters.

G.2 Detailed results on non-English texts of Persuaficial dataset

In addition to the English results, we provide detailed F_1 scores for persuasion detection on other language-specific subsets of the Persuaficial datasets. Table 11 reports results for German texts from SemEval 2023 Task 3 and their LLM-generated counterparts, Table 12 for French texts, Table 13 for Italian texts, Table 14 for Polish texts, and Table 15 for Russian texts. For each dataset, classifier performance on human-written texts (first column) is compared with performance on LLM-generated texts produced via paraphrasing, rewriting with subtle or intensive persuasion, and open-ended generation. The results are further broken down by both the generating model and the classifier model, demonstrating how different generation approaches influence the detectability of persuasion across languages.

H StyloMetrix and Its Usefulness in AI vs. Human Persuasive Text Analysis

To further demonstrate the utility of StyloMetrix for analyzing human-written versus AI-generated persuasive texts, we conduct a classification study using classical machine learning models with features calculated by StyloMetrix. We aimed to prove that linguistic features contain enough information to differentiate AI-generated persuasion from human-written persuasion.

For all GPT 4.1. Mini-generated English synthetic texts produced with each generation approach, we trained a separate classifier. For each experiemnt, we split the data into training and test sets, allocating 70% for training and 30% for testing. To ensure a credible evaluation, each human-written text and its LLM-generated counterpart were placed in the same split, either training or

Classifier Models	Human-written	Paraphrasing Generation	Rewriting Generation		Open-ended Generation
			Subtle Persuasion	Intensive Persuasion	
Generating model: GPT 4.1 Mini					
GPT 4.1 Mini	0.7398	0.6984 ↓6%	0.3837 ↓48%	0.7638 ↑3%	0.8969 ↑21%
Llama 3.3 70B	0.7459	0.7310 ↓2%	0.4444 ↓40%	0.7757 ↑4%	0.8741 ↑17%
Gemma 3 27b it	0.7572	0.7561 ↓0%	0.6213 ↓18%	0.8007 ↑6%	0.8562 ↑13%
Gemini 2.0 Flash	0.7551	0.7487 ↓1%	0.6407 ↓15%	0.7780 ↑3%	0.8117 ↑7%
Generating model: Llama 3.3 70B					
GPT 4.1 Mini	0.7398	0.6831 ↓8%	0.4411 ↓40%	0.7870 ↑6%	0.8969 ↑21%
Llama 3.3 70B	0.7459	0.6911 ↓7%	0.4811 ↓35%	0.7791 ↑4%	0.8741 ↑17%
Gemma 3 27b it	0.7572	0.7479 ↓1%	0.6746 ↓11%	0.8082 ↑7%	0.8562 ↑13%
Gemini 2.0 Flash	0.7551	0.7476 ↓1%	0.6691 ↓11%	0.7921 ↑5%	0.8117 ↑7%
Generating model: Gemma 3 27b it					
GPT 4.1 Mini	0.7398	0.7025 ↓5%	0.3445 ↓53%	0.8611 ↑16%	0.8969 ↑21%
Llama 3.3 70B	0.7459	0.7222 ↓3%	0.4118 ↓45%	0.8469 ↑14%	0.8741 ↑17%
Gemma 3 27b it	0.7572	0.7675 ↑1%	0.6241 ↓18%	0.8383 ↑11%	0.8562 ↑13%
Gemini 2.0 Flash	0.7551	0.7614 ↑1%	0.6202 ↓18%	0.8039 ↑6%	0.8117 ↑7%
Generating model: Gemini 2.0 Flash					
GPT 4.1 Mini	0.7398	0.7188 ↓3%	0.4430 ↓40%	0.8472 ↑15%	0.8949 ↑21%
Llama 3.3 70B	0.7459	0.7385 ↓1%	0.4936 ↓34%	0.8428 ↑13%	0.8741 ↑17%
Gemma 3 27b it	0.7572	0.7652 ↑1%	0.6613 ↓13%	0.8362 ↑10%	0.8562 ↑13%
Gemini 2.0 Flash	0.7551	0.7583 ↑0%	0.6787 ↓10%	0.8059 ↑7%	0.8117 ↑7%

Table 8: F_1 scores for persuasion detection on English data sample of Persuaficial. More specifically, on sample of Persuaficial generated using English texts from SemEval 2023 Task 3 dataset. The first column reports performance on English texts from SemEval 2023 Task 3 human-annotated texts. The remaining columns show performance on LLM-generated English counterparts.

Classifier Models	Human-written	Paraphrasing Generation	Rewriting Generation		Open-ended Generation
			Subtle Persuasion	Intensive Persuasion	
Generating model: GPT 4.1 Mini					
GPT 4.1 Mini	0.7567	0.7435 ↓2%	0.4948 ↓35%	0.7866 ↑4%	0.8666 ↑15%
Llama 3.3 70B	0.7471	0.7348 ↓2%	0.5795 ↓22%	0.7679 ↑3%	0.8292 ↑11%
Gemma 3 27b it	0.7473	0.7441 ↓0%	0.6308 ↓16%	0.7607 ↑2%	0.7994 ↑7%
Gemini 2.0 Flash	0.7518	0.7449 ↓1%	0.6711 ↓11%	0.7595 ↑1%	0.7680 ↑2%
Generating model: Llama 3.3 70B					
GPT 4.1 Mini	0.7567	0.7338 ↓3%	0.5595 ↓26%	0.7967 ↑5%	0.8666 ↑15%
Llama 3.3 70B	0.7471	0.7314 ↓2%	0.6251 ↓16%	0.7775 ↑4%	0.8292 ↑11%
Gemma 3 27b it	0.7473	0.7410 ↓1%	0.6928 ↓7%	0.7749 ↑4%	0.7994 ↑7%
Gemini 2.0 Flash	0.7518	0.7400 ↓2%	0.7002 ↓7%	0.7652 ↑2%	0.7680 ↑2%
Generating model: Gemma 3 27b it					
GPT 4.1 Mini	0.7567	0.7672 ↑1%	0.3951 ↓48%	0.8404 ↑11%	0.8666 ↑15%
Llama 3.3 70B	0.7471	0.7449 ↓0%	0.4898 ↓34%	0.8074 ↑8%	0.8292 ↑11%
Gemma 3 27b it	0.7473	0.7504 ↑0%	0.5777 ↓23%	0.7936 ↑6%	0.7994 ↑7%
Gemini 2.0 Flash	0.7518	0.7459 ↓1%	0.6232 ↓17%	0.7662 ↑2%	0.7680 ↑2%
Generating model: Gemini 2.0 Flash					
GPT 4.1 Mini	0.7567	0.7399 ↓2%	0.5353 ↓29%	0.8163 ↑8%	0.8666 ↑15%
Llama 3.3 70B	0.7471	0.7336 ↓2%	0.5838 ↓22%	0.7911 ↑6%	0.8292 ↑11%
Gemma 3 27b it	0.7473	0.7483 ↑0%	0.6383 ↓15%	0.7838 ↑5%	0.7994 ↑7%
Gemini 2.0 Flash	0.7518	0.7400 ↓2%	0.6711 ↓11%	0.7652 ↑2%	0.7680 ↑2%

Table 9: F_1 scores for persuasion detection on English data sample of Persuaficial. More specifically, on sample of Persuaficial generated using DIPROMATS 2024 dataset. The first column reports performance on DIPROMATS 2024 human-annotated texts. The remaining columns show performance on LLM-generated counterparts.

test. This prevents the classifier from exploiting the potential direct similarities between the paired texts. We employed widely used tree-based machine learning methods as classifiers, as they naturally capture non-linear interactions and are well-suited for moderate- to high-dimensional tabular data (Grinsztajn et al., 2022; Uddin and Lu, 2024). Previous work has shown that, for tabular data, tree-based models can even outperform deep learning approaches (Grinsztajn et al., 2022).

Table 16 shows the results of these experiments. The outcomes show a clear progression: the more generative freedom the LLM is given, the easier it becomes for tree-ensemble models to distinguish its outputs from human-written persuasion. *Paraphrasing* keeps the AI text close to the original human style, yielding only moderate detection performance. In the *Rewriting* conditions, the model introduces larger stylistic shifts—whether by making persuasion subtler or more intense—which improves separability. *Open-ended* generation, starting from only a brief neutral summary, produces the greatest stylistic divergence and thus the highest classification accuracy. Overall, stylistic features become increasingly informative as the generation task becomes less constrained.

Classifier Models	Human-written	Paraphrasing Generation	Rewriting Generation		Open-ended Generation
			Subtle Persuasion	Intensive Persuasion	
Generating model: GPT 4.1 Mini					
GPT 4.1 Mini	0.6233	0.6337 ↑2%	0.4941 ↓21%	0.6582 ↑6%	0.7148 ↑15%
Llama 3.3 70B	0.6517	0.6546 ↑0%	0.5665 ↓13%	0.6667 ↑2%	0.6831 ↑5%
Gemma 3 27b it	0.6644	0.6745 ↑2%	0.6388 ↓4%	0.6809 ↑2%	0.6836 ↑3%
Gemini 2.0 Flash	0.6671	0.6662 ↓0%	0.6431 ↓4%	0.6726 ↑1%	0.6770 ↑1%
Generating model: Llama 3.3 70B					
GPT 4.1 Mini	0.6233	0.6137 ↓2%	0.4619 ↓26%	0.6481 ↑4%	0.7148 ↑15%
Llama 3.3 70B	0.6517	0.6307 ↓3%	0.5226 ↓20%	0.6564 ↑1%	0.6831 ↑5%
Gemma 3 27b it	0.6644	0.6644 ↓0%	0.6133 ↓8%	0.6772 ↑2%	0.6845 ↑3%
Gemini 2.0 Flash	0.6671	0.6607 ↓1%	0.6112 ↓8%	0.6717 ↑1%	0.6770 ↑1%
Generating model: Gemma 3 27b it					
GPT 4.1 Mini	0.6233	0.6572 ↑5%	0.4840 ↓22%	0.7009 ↑12%	0.7148 ↑15%
Llama 3.3 70B	0.6517	0.6574 ↑1%	0.5515 ↓15%	0.6758 ↑4%	0.6831 ↑5%
Gemma 3 27b it	0.6644	0.6754 ↑2%	0.6349 ↓4%	0.6836 ↑3%	0.6845 ↑3%
Gemini 2.0 Flash	0.6671	0.6689 ↑0%	0.6259 ↓6%	0.6762 ↑1%	0.6770 ↑1%
Generating model: Gemini 2.0 Flash					
GPT 4.1 Mini	0.6233	0.6379 ↑2%	0.5226 ↓16%	0.6885 ↑10%	0.7148 ↑15%
Llama 3.3 70B	0.6517	0.6527 ↑0%	0.5740 ↓12%	0.6776 ↑4%	0.6831 ↑5%
Gemma 3 27b it	0.6644	0.6690 ↑1%	0.6465 ↓3%	0.6818 ↑3%	0.6845 ↑3%
Gemini 2.0 Flash	0.6671	0.6689 ↑0%	0.6374 ↓4%	0.6753 ↑1%	0.6770 ↑1%

Table 10: F_1 scores for persuasion detection on English data sample of Persuaficial. More specifically, on sample of Persuaficial generated using ChangeMyView dataset. The first column reports performance on ChangeMyView human-annotated texts. The remaining columns show performance on LLM-generated counterparts.

Classifier Models	Human-written	Paraphrasing Generation	Rewriting Generation		Open-ended Generation
			Subtle Persuasion	Intensive Persuasion	
Generating model: GPT 4.1 Mini					
GPT 4.1 Mini	0.7203	0.6999 ↓3%	0.4677 ↓35%	0.7959 ↑10%	0.9416 ↑31%
Llama 3.3 70B	0.7361	0.7239 ↓2%	0.4574 ↓38%	0.8004 ↑9%	0.9347 ↑27%
Gemma 3 27b it	0.7655	0.7644 ↓0%	0.6553 ↓14%	0.8220 ↑7%	0.9006 ↑18%
Gemini 2.0 Flash	0.7903	0.7768 ↓2%	0.6871 ↓13%	0.8244 ↑4%	0.8593 ↑9%
Generating model: Llama 3.3 70B					
GPT 4.1 Mini	0.7203	0.7102 ↓1%	0.4604 ↓36%	0.7836 ↑9%	0.9416 ↑31%
Llama 3.3 70B	0.7361	0.7128 ↓3%	0.4681 ↓36%	0.7971 ↑8%	0.9347 ↑27%
Gemma 3 27b it	0.7655	0.7666 ↑0%	0.6901 ↓10%	0.8200 ↑7%	0.9006 ↑18%
Gemini 2.0 Flash	0.7903	0.7816 ↓1%	0.7156 ↓9%	0.8298 ↑5%	0.8593 ↑9%
Generating model: Gemma 3 27b it					
GPT 4.1 Mini	0.7203	0.7303 ↑1%	0.3881 ↓46%	0.8996 ↑25%	0.9407 ↑31%
Llama 3.3 70B	0.7361	0.7252 ↓1%	0.3779 ↓49%	0.8946 ↑22%	0.9339 ↑27%
Gemma 3 27b it	0.7655	0.7877 ↑3%	0.6411 ↓16%	0.8824 ↑15%	0.8998 ↑18%
Gemini 2.0 Flash	0.7903	0.7978 ↑1%	0.6644 ↓16%	0.8499 ↑8%	0.8584 ↑9%
Generating model: Gemini 2.0 Flash					
GPT 4.1 Mini	0.7203	0.7425 ↑3%	0.4476 ↓38%	0.9033 ↑25%	0.9416 ↑31%
Llama 3.3 70B	0.7361	0.7373 ↑0%	0.4556 ↓38%	0.8974 ↑22%	0.9347 ↑27%
Gemma 3 27b it	0.7655	0.7866 ↑3%	0.6791 ↓11%	0.8806 ↑15%	0.9006 ↑18%
Gemini 2.0 Flash	0.7903	0.7959 ↑1%	0.6949 ↓12%	0.8499 ↑8%	0.8593 ↑9%

Table 11: F_1 scores for persuasion detection on German data sample of Persuaficial. More specifically, on sample of Persuaficial generated using German texts from SemEval 2023 Task 3 dataset. The first column reports performance on SemEval 2023 Task 3 German human-annotated texts. The remaining columns show performance on LLM-generated German counterparts.

I Cohen's d statistic definition and our computation

For each linguistic feature, we calculate Cohen's d to quantify the magnitude of the shift between the feature distribution of generated texts g_i and that of their human-written persuasive counterparts r_i. Cohen's d is defined in Equation 5 and is computed as follows.

First, we calculate the mean of each feature for human-written and generated texts:

$$\bar{r} = \frac{1}{n_r}\sum_{i=1}^{n_r} r_i, \qquad \bar{g} = \frac{1}{n_g}\sum_{i=1}^{n_g} g_i, \tag{1}$$

where n_r and n_g denote the number of human-written and generated texts, respectively. In our experiments, $n_r = n_g$ as for each human-written persuasive text we have AI-generated counterpart.

Next, we compute the sample standard deviations for each group:

$$s_r = \sqrt{\frac{1}{n_r - 1}\sum_{i=1}^{n_r}(r_i - \bar{r})^2}, \tag{2}$$

$$s_g = \sqrt{\frac{1}{n_g - 1}\sum_{i=1}^{n_g}(g_i - \bar{g})^2}. \tag{3}$$

Using these, we calculate the pooled standard deviation:

Classifier Models	Human-written	Paraphrasing Generation	Rewriting Generation		Open-ended Generation
			Subtle Persuasion	Intensive Persuasion	
Generating model: GPT 4.1 Mini					
GPT 4.1 Mini	0.7505	0.7356 ↓2%	0.4568 ↓39%	0.8021 ↑7%	0.9251 ↑23%
Llama 3.3 70B	0.7605	0.7460 ↓2%	0.4675 ↓39%	0.7996 ↑5%	0.9164 ↑21%
Gemma 3 27b it	0.7827	0.7733 ↓1%	0.6489 ↓17%	0.8190 ↑5%	0.8816 ↑13%
Gemini 2.0 Flash	0.7812	0.7812 ↓0%	0.6705 ↓14%	0.8189 ↑5%	0.8418 ↑8%
Generating model: Llama 3.3 70B					
GPT 4.1 Mini	0.7505	0.7160 ↓5%	0.4548 ↓39%	0.7959 ↑6%	0.9251 ↑23%
Llama 3.3 70B	0.7605	0.7216 ↓5%	0.4794 ↓37%	0.7996 ↑5%	0.9174 ↑21%
Gemma 3 27b it	0.7827	0.7744 ↓1%	0.6908 ↓12%	0.8343 ↑7%	0.8826 ↑13%
Gemini 2.0 Flash	0.7812	0.7758 ↓1%	0.7116 ↓9%	0.8269 ↑6%	0.8418 ↑8%
Generating model: Gemma 3 27b it					
GPT 4.1 Mini	0.7505	0.7492 ↓0%	0.3376 ↓55%	0.8848 ↑18%	0.9251 ↑23%
Llama 3.3 70B	0.7605	0.7366 ↓3%	0.3520 ↓54%	0.8815 ↑16%	0.9174 ↑21%
Gemma 3 27b it	0.7827	0.8068 ↑3%	0.6179 ↓21%	0.8676 ↑11%	0.8826 ↑13%
Gemini 2.0 Flash	0.7812	0.7898 ↑1%	0.6523 ↓16%	0.8388 ↑7%	0.8418 ↑8%
Generating model: Gemini 2.0 Flash					
GPT 4.1 Mini	0.7505	0.7807 ↑4%	0.4670 ↓38%	0.8996 ↑20%	0.9251 ↑23%
Llama 3.3 70B	0.7605	0.7759 ↑2%	0.5120 ↓33%	0.8920 ↑17%	0.9174 ↑21%
Gemma 3 27b it	0.7827	0.7920 ↑1%	0.6771 ↓13%	0.8696 ↑11%	0.8826 ↑13%
Gemini 2.0 Flash	0.7812	0.7972 ↑2%	0.6858 ↓12%	0.8408 ↑8%	0.8418 ↑8%

Table 12: F_1 scores for persuasion detection on French data sample of Persuaficial. More specifically, on sample of Persuaficial generated using French texts from SemEval 2023 Task 3 dataset. The first column reports performance on SemEval 2023 Task 3 French human-annotated texts. The remaining columns show performance on LLM-generated French counterparts.

Classifier Models	Human-written	Paraphrasing Generation	Rewriting Generation		Open-ended Generation
			Subtle Persuasion	Intensive Persuasion	
Generating model: GPT 4.1 Mini					
GPT 4.1 Mini	0.7471	0.7511 ↑1%	0.4368 ↓42%	0.7959 ↑7%	0.9200 ↑23%
Llama 3.3 70B	0.7584	0.7277 ↓4%	0.4447 ↓41%	0.7926 ↑5%	0.9166 ↑21%
Gemma 3 27b it	0.7659	0.7775 ↑2%	0.6563 ↓14%	0.8185 ↑7%	0.8688 ↑13%
Gemini 2.0 Flash	0.7986	0.7912 ↓1%	0.6750 ↓15%	0.8192 ↑3%	0.8410 ↑5%
Generating model: Llama 3.3 70B					
GPT 4.1 Mini	0.7471	0.7113 ↓5%	0.4812 ↓36%	0.7822 ↑5%	0.9200 ↑23%
Llama 3.3 70B	0.7584	0.6823 ↓10%	0.4689 ↓38%	0.7827 ↑3%	0.9166 ↑21%
Gemma 3 27b it	0.7659	0.7576 ↓1%	0.6967 ↓9%	0.8164 ↑7%	0.8688 ↑13%
Gemini 2.0 Flash	0.7986	0.7827 ↓2%	0.7193 ↓10%	0.8192 ↑3%	0.8410 ↑5%
Generating model: Gemma 3 27b it					
GPT 4.1 Mini	0.7471	0.7239 ↓3%	0.3252 ↓56%	0.8925 ↑19%	0.9190 ↑23%
Llama 3.3 70B	0.7584	0.7110 ↓6%	0.3376 ↓55%	0.8912 ↑18%	0.9156 ↑21%
Gemma 3 27b it	0.7659	0.7944 ↑4%	0.6280 ↓18%	0.8619 ↑13%	0.8678 ↑13%
Gemini 2.0 Flash	0.7986	0.7944 ↓1%	0.6356 ↓20%	0.8410 ↑5%	0.8401 ↑5%
Generating model: Gemini 2.0 Flash					
GPT 4.1 Mini	0.7471	0.7457 ↓0%	0.4553 ↓39%	0.9007 ↑21%	0.9190 ↑23%
Llama 3.3 70B	0.7584	0.7479 ↓1%	0.4629 ↓39%	0.9015 ↑19%	0.9156 ↑21%
Gemma 3 27b it	0.7659	0.7922 ↑3%	0.6632 ↓13%	0.8629 ↑13%	0.8688 ↑13%
Gemini 2.0 Flash	0.7986	0.8069 ↑1%	0.6826 ↓15%	0.8410 ↑5%	0.8410 ↑5%

Table 13: F_1 scores for persuasion detection on Italian data sample of Persuaficial. More specifically, on sample of Persuaficial generated using Italian texts from SemEval 2023 Task 3 dataset. The first column reports performance on SemEval 2023 Task 3 Italian human-annotated texts. The remaining columns show performance on LLM-generated Italian counterparts.

$$s_{\text{pooled}} = \sqrt{\frac{(n_r - 1)s_r^2 + (n_g - 1)s_g^2}{n_r + n_g - 2}}. \quad (4)$$

Finally, Cohen's d is obtained as the difference in means normalized by the pooled standard deviation:

$$d = \frac{\bar{g} - \bar{r}}{s_{\text{pooled}}} \quad (5)$$

This effect size, Cohen's d, provides a standardized measure of the shift in feature distributions between generated and human-written persuasive texts, allowing comparison across features and models.

Since many feature distributions may be non-Gaussian, we avoid this assumptions when testing for a significance of the shift. Statistical significance is assessed with a paired Wilcoxon signed-rank test on the per text differences $d_i = g_i - r_i$, with Benjamini-Hochberg FDR correction across features. As we perform one test per feature, we control for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg, 1995) across all features. We report a significance indicator in our tables.

Classifier Models	Human-written	Paraphrasing Generation	Rewriting Generation		Open-ended Generation
			Subtle Persuasion	Intensive Persuasion	
Generating model: GPT 4.1 Mini					
GPT 4.1 Mini	0.7330	0.7010 ↓4%	0.4634 ↓37%	0.7936 ↑8%	0.9372 ↑28%
Llama 3.3 70B	0.7676	0.7398 ↓4%	0.5165 ↓33%	0.8016 ↑4%	0.9208 ↑20%
Gemma 3 27b it	0.7728	0.7752 ↑0%	0.6942 ↓10%	0.8075 ↑4%	0.8834 ↑14%
Gemini 2.0 Flash	0.7733	0.7744 ↑0%	0.7151 ↓8%	0.8020 ↑4%	0.8217 ↑6%
Generating model: Llama 3.3 70B					
GPT 4.1 Mini	0.7330	0.6829 ↓7%	0.5215 ↓29%	0.7793 ↑6%	0.9352 ↑28%
Llama 3.3 70B	0.7676	0.7177 ↓7%	0.5637 ↓27%	0.7880 ↑3%	0.9198 ↑20%
Gemma 3 27b it	0.7728	0.7596 ↓2%	0.7232 ↓6%	0.8176 ↑6%	0.8834 ↑14%
Gemini 2.0 Flash	0.7733	0.7563 ↓2%	0.7321 ↓5%	0.7980 ↑3%	0.8217 ↑6%
Generating model: Gemma 3 27b it					
GPT 4.1 Mini	0.7330	0.7143 ↓3%	0.3777 ↓48%	0.9086 ↑24%	0.9372 ↑28%
Llama 3.3 70B	0.7676	0.7330 ↓5%	0.4311 ↓44%	0.9047 ↑18%	0.9208 ↑20%
Gemma 3 27b it	0.7728	0.7869 ↑2%	0.6737 ↓13%	0.8754 ↑13%	0.8834 ↑14%
Gemini 2.0 Flash	0.7733	0.7765 ↑0%	0.6623 ↓14%	0.8198 ↑6%	0.8217 ↑6%
Generating model: Gemini 2.0 Flash					
GPT 4.1 Mini	0.7330	0.7258 ↓1%	0.4696 ↓36%	0.9117 ↑24%	0.9372 ↑28%
Llama 3.3 70B	0.7676	0.7650 ↓0%	0.5051 ↓34%	0.9128 ↑19%	0.9208 ↑20%
Gemma 3 27b it	0.7728	0.7916 ↑2%	0.6764 ↓12%	0.8704 ↑13%	0.8834 ↑14%
Gemini 2.0 Flash	0.7733	0.7858 ↑2%	0.6975 ↓10%	0.8207 ↑6%	0.8217 ↑6%

Table 14: F_1 scores for persuasion detection on Polish data sample of Persuaficial. More specifically, on sample of Persuaficial generated using Polish texts from SemEval 2023 Task 3 dataset. The first column reports performance on SemEval 2023 Task 3 Polish human-annotated texts. The remaining columns show performance on LLM-generated Polish counterparts.

Classifier Models	Human-written	Paraphrasing Generation	Rewriting Generation		Open-ended Generation
			Subtle Persuasion	Intensive Persuasion	
Generating model: GPT 4.1 Mini					
GPT 4.1 Mini	0.7246	0.6889 ↓5%	0.4464 ↓38%	0.7844 ↑8%	0.9017 ↑24%
Llama 3.3 70B	0.7408	0.7166 ↓3%	0.4556 ↓38%	0.7915 ↑7%	0.9071 ↑22%
Gemma 3 27b it	0.7360	0.7396 ↑0%	0.6069 ↓18%	0.7843 ↑7%	0.8562 ↑16%
Gemini 2.0 Flash	0.7683	0.7571 ↓1%	0.6767 ↓12%	0.7784 ↑1%	0.8019 ↑4%
Generating model: Llama 3.3 70B					
GPT 4.1 Mini	0.7246	0.6889 ↓5%	0.4794 ↓34%	0.7546 ↑4%	0.9017 ↑24%
Llama 3.3 70B	0.7408	0.6740 ↓9%	0.4635 ↓37%	0.7603 ↑3%	0.9091 ↑23%
Gemma 3 27b it	0.7360	0.7203 ↓2%	0.6311 ↓14%	0.7732 ↑5%	0.8562 ↑16%
Gemini 2.0 Flash	0.7683	0.7477 ↓3%	0.6986 ↓9%	0.7774 ↑1%	0.8019 ↑4%
Generating model: Gemma 3 27b it					
GPT 4.1 Mini	0.7246	0.7138 ↓1%	0.3651 ↓50%	0.8805 ↑22%	0.9017 ↑24%
Llama 3.3 70B	0.7408	0.7248 ↓2%	0.3562 ↓52%	0.8858 ↑20%	0.9091 ↑23%
Gemma 3 27b it	0.7360	0.7491 ↑2%	0.5907 ↓20%	0.8453 ↑15%	0.8562 ↑16%
Gemini 2.0 Flash	0.7683	0.7704 ↑0%	0.6577 ↓14%	0.7971 ↑4%	0.8019 ↑4%
Generating model: Gemini 2.0 Flash					
GPT 4.1 Mini	0.7246	0.7378 ↑2%	0.4660 ↓36%	0.8774 ↑21%	0.9017 ↑24%
Llama 3.3 70B	0.7408	0.7500 ↑1%	0.4496 ↓39%	0.8920 ↑20%	0.9091 ↑23%
Gemma 3 27b it	0.7360	0.7572 ↑3%	0.6227 ↓15%	0.8362 ↑14%	0.8562 ↑16%
Gemini 2.0 Flash	0.7683	0.7714 ↑0%	0.6849 ↓11%	0.7981 ↑4%	0.8019 ↑4%

Table 15: F_1 scores for persuasion detection on Russian data sample of Persuaficial. More specifically, on sample of Persuaficial generated using Russian texts from SemEval 2023 Task 3 dataset. The first column reports performance on SemEval 2023 Task 3 Russian human-annotated texts. The remaining columns show performance on LLM-generated Russian counterparts.

Model	Precision	Recall	F1	Accuracy
GPT 4.1 Mini – Paraphrasing				
RF	0.74	0.74	0.74	0.74
XGB	0.76	0.76	0.76	0.76
LGBM	0.76	0.76	0.76	0.76
GPT 4.1 Mini – Rewriting with Subtle Persuasive Effect				
RF	0.84	0.84	0.84	0.84
XGB	0.87	0.87	0.87	0.87
LGBM	0.86	0.86	0.86	0.86
GPT 4.1 Mini – Rewriting with Intensified Persuasive Effect				
RF	0.83	0.83	0.83	0.83
XGB	0.84	0.84	0.84	0.84
LGBM	0.86	0.85	0.85	0.85
GPT 4.1 Mini – Open-ended				
RF	0.97	0.97	0.97	0.97
XGB	0.97	0.97	0.97	0.97
LGBM	0.98	0.98	0.98	0.98

Table 16: Classification performance of detecting GPT-4.1-Mini-generated persuasive texts versus human-written persuasive texts using linguistic-feature representations. Each experiment reflects a different AI-generation strategy and uses data combined from three English datasets.

J Linguistic Differences between Human and Machine Generated Persuasive Texts - Additional Results

Tables 17 - 32 report the top features that most strongly distinguish AI-generated persuasive texts from human-written persuasive texts. We provide 16 tables in total, reflecting Cohen's d effect sizes computed separately for each generating model and for each generation setting: Paraphrasing, Rewriting subtle persuasion, Rewriting intensified persuasion, and Open-ended generation.

Stylometric Feature	Cohen's d	Sig.
GPT 4.1 Mini - Paraphrasing		
L_CONT_T	0.5054	✓
L_CONT_A	0.4590	✓
SY_INV_PATTERNS	-0.4542	✓
LTOKEN_RATIO_LEM	0.3713	✓
L_FUNC_A	-0.3619	✓
L_PUNCT_DASH	0.3378	✓
ST_REPET_WORDS	-0.3280	✓
L_PUNCT_SEMC	-0.2656	✓
L_PUNCT_COM	0.2298	✓
ASM	-0.2196	✓
VT_MIGHT	0.1995	✓
L_LINKS	-0.1954	✓
L_ADJ_POSITIVE	0.1723	✓
CDS	-0.1719	✓
PS_AGREEMENT	-0.1667	✓
L_ADV_SUPERLATIVE	0.1664	✓
L_ADV_COMPARATIVE	0.1592	✓
POS_ADV	0.1583	✓
PS_CAUSE	-0.1537	✓
PS_TIME	-0.1532	✓

Table 17: Top 20 linguistic features for AI-generated persuasive text with *Paraphrasing* generation approach and GPT 4.1 Mini model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
GPT 4.1 Mini - Rewriting with Subtle Persuasive Effect		
L_CONT_T	0.7036	✓
L_CONT_A	0.6222	✓
VT_MIGHT	0.6188	✓
SY_INV_PATTERNS	-0.6072	✓
L_PLURAL_NOUNS	0.5444	✓
L_FUNC_A	-0.5293	✓
SY_NARRATIVE	0.5242	✓
LTOKEN_RATIO_LEM	0.4993	✓
ST_REPET_WORDS	-0.4319	✓
POS_PRO	-0.4307	✓
L_YOU_PRON	-0.4245	✓
G_ACTIVE	-0.4232	✓
G_FUTURE	-0.4015	✓
L_SECOND_PERSON_PRON	-0.3997	✓
VT_FUTURE_SIMPLE	-0.3949	✓
CDS	-0.3837	✓
SENT_D_PP	0.3689	✓
L_PUNCT	-0.3401	✓
VT_MAY	0.3378	✓
SENT_ST_DIFFERENCE	-0.3208	✓

Table 18: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Subtle Persuasive Effect* generation approach and GPT 4.1 Mini model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
GPT 4.1 Mini - Rewriting with Intensified Persuasive Effect		
L_CONT_T	0.7685	✓
L_CONT_A	0.7654	✓
L_PUNCT_DASH	0.7243	✓
L_FUNC_A	-0.5702	✓
LTOKEN_RATIO_LEM	0.5075	✓
L_ADV_SUPERLATIVE	0.5033	✓
L_ADV_COMPARATIVE	0.4892	✓
SY_INV_PATTERNS	-0.4888	✓
POS_ADV	0.4766	✓
ST_REPET_WORDS	-0.4037	✓
L_ADJ_POSITIVE	0.3884	✓
POS_ADJ	0.3670	✓
ASM	-0.3241	✓
L_ADV_POSITIVE	0.3237	✓
PS_CAUSE	-0.3067	✓
L_PUNCT_COM	0.3045	✓
L_PUNCT_SEMC	-0.2626	✓
POS_PRO	-0.2415	✓
SENT_D_ADVP	0.2385	✓
L_NOUN_PHRASES	0.2333	✓

Table 19: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Intensified Persuasive Effect* generation approach and GPT 4.1 Mini model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
GPT 4.1 Mini - Open-ended		
L_CONT_T	1.6204	✓
L_CONT_A	1.4300	✓
LTOKEN_RATIO_LEM	1.3172	✓
SENT_D_NP	1.0532	✓
ST_REPET_WORDS	-0.9950	✓
L_PUNCT_DASH	0.9885	✓
L_FUNC_A	-0.8585	✓
SY_IMPERATIVE	0.8376	✓
POS_NOUN	0.8076	✓
POS_ADJ	0.7841	✓
L_ADJ_POSITIVE	0.7816	✓
SY_INV_PATTERNS	-0.7496	✓
L_LINKS	-0.7474	✓
VF_INFINITIVE	0.7252	✓
G_PAST	-0.6604	✓
G_ACTIVE	-0.6565	✓
L_FIRST_PERSON_SING_PRON	-0.6425	✓
L_I_PRON	-0.6425	✓
SENT_D_VP	0.6275	✓
VT_PAST_SIMPLE	-0.5774	✓

Table 20: Top 20 linguistic features for AI-generated persuasive text with *Open-ended* generation approach and GPT 4.1 Mini model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Llama - Paraphrasing		
SENT_ST_WPERSENT	0.7055	✓
SENT_ST_DIFFERENCE	-0.6979	✓
L_PUNCT_COM	0.6659	✓
SY_INV_PATTERNS	-0.6146	✓
L_CONT_T	0.5438	✓
L_CONT_A	0.4861	✓
G_ACTIVE	-0.4821	✓
L_PUNCT_DOT	-0.4509	✓
ASM	-0.4192	✓
POS_PRO	-0.4055	✓
FOS_FRONTING	0.3884	✓
L_ADJ_POSITIVE	0.3844	✓
L_PUNCT_SEMC	-0.3839	✓
L_YOU_PRON	-0.3761	✓
L_FUNC_A	-0.3469	✓
SY_SUBORD_SENT	0.3354	✓
L_PUNCT	-0.3339	✓
POS_PREP	0.3256	✓
VT_PRESENT_SIMPLE	-0.3229	✓
L_SECOND_PERSON_PRON	-0.3228	✓

Table 21: Top 20 linguistic features for AI-generated persuasive text with *Paraphrasing* generation approach and Llama 3.3 70B model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Llama - Rewriting with Subtle Persuasive Effect		
SY_INV_PATTERNS	-0.7470	✓
G_ACTIVE	-0.7152	✓
SENT_ST_WPERSENT	0.6961	✓
L_CONT_T	0.6226	✓
SENT_D_PP	0.6153	✓
L_ADJ_POSITIVE	0.6145	✓
POS_PREP	0.6086	✓
L_YOU_PRON	-0.5966	✓
POS_NOUN	0.5896	✓
L_PUNCT_COM	0.5848	✓
L_SECOND_PERSON_PRON	-0.5764	✓
POS_ADJ	0.5595	✓
SY_NARRATIVE	0.5419	✓
L_CONT_A	0.5413	✓
POS_PRO	-0.5400	✓
SENT_ST_DIFFERENCE	-0.5382	✓
SY_SUBORD_SENT	0.5373	✓
L_PUNCT	-0.5195	✓
ASM	-0.4923	✓
L_PLURAL_NOUNS	0.4688	✓

Table 22: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Subtle Persuasive Effect* generation approach and Llama 3.3 70B model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Llama - Rewriting with Intensified Persuasive Effect		
SENT_ST_WPERSENT	0.8019	✓
L_ADJ_POSITIVE	0.7196	✓
L_PUNCT_COM	0.7055	✓
L_CONT_T	0.6589	✓
POS_ADJ	0.6478	✓
L_CONT_A	0.6349	✓
G_ACTIVE	-0.6171	✓
SENT_ST_DIFFERENCE	-0.6110	✓
L_PUNCT_DOT	-0.5876	✓
SY_INV_PATTERNS	-0.5135	✓
ASM	-0.4697	✓
L_FUNC_A	-0.4692	✓
L_FUNC_T	-0.4456	✓
FOS_FRONTING	0.4390	✓
POS_PRO	-0.3973	✓
SENT_D_VP	0.3895	✓
L_PUNCT_SEMC	-0.3806	✓
CDS	-0.3720	✓
L_YOU_PRON	-0.3670	✓
L_NOUN_PHRASES	0.3630	✓

Table 23: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Intensified Persuasive Effect* generation approach and Llama 3.3 70B model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Llama - Open-ended		
VF_INFINITIVE	1.3397	✓
SY_IMPERATIVE	1.2406	✓
G_ACTIVE	-0.9149	✓
SENT_D_VP	0.8331	✓
L_PUNCT_DOT	-0.8088	✓
G_PAST	-0.7696	✓
L_OUR_PRON	0.7642	✓
L_LINKS	-0.7474	✓
VT_PAST_SIMPLE	-0.6985	✓
SY_EXCLAMATION	0.6879	✓
SY_INV_PATTERNS	-0.6857	✓
L_CONT_T	0.6780	✓
SY_COORD_SENT	0.6774	✓
L_PUNCT	-0.6493	✓
SENT_D_NP	0.6396	✓
L_FIRST_PERSON_SING_PRON	-0.6230	✓
L_I_PRON	-0.6230	✓
L_WE_PRON	0.6190	✓
L_IT_PRON	0.5548	✓
VT_MUST	0.5078	✓

Table 24: Top 20 linguistic features for AI-generated persuasive text with *Open-ended* generation approach and Llama 3.3 70B model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemma - Paraphrasing		
L_CONT_T	0.7526	✓
L_CONT_A	0.6659	✓
L_FUNC_A	-0.6181	✓
SY_INV_PATTERNS	-0.6133	✓
L_ADV_SUPERLATIVE	0.4274	✓
ST_REPET_WORDS	-0.4178	✓
L_ADV_COMPARATIVE	0.4141	✓
LTOKEN_RATIO_LEM	0.4081	✓
PS_CONDITION	-0.3995	✓
ASM	-0.3761	✓
CDS	-0.3606	✓
POS_ADV	0.3605	✓
L_ADJ_POSITIVE	0.3407	✓
PS_AGREEMENT	-0.3212	✓
L_PUNCT_COM	0.3172	✓
PS_CAUSE	-0.2988	✓
POS_ADJ	0.2872	✓
POS_PRO	-0.2805	✓
G_ACTIVE	-0.2710	✓
SENT_ST_WPERSENT	0.2582	✓

Table 25: Top 20 linguistic features for AI-generated persuasive text with *Paraphrasing* generation approach and Gemma 3 27b it model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemma - Rewriting with Intensified Persuasive Effect		
L_CONT_T	1.0424	✓
L_CONT_A	1.0122	✓
L_FUNC_A	-0.9718	✓
L_ADJ_POSITIVE	0.9024	✓
POS_ADJ	0.8267	✓
PS_CONDITION	-0.6942	✓
PS_CAUSE	-0.6071	✓
L_FUNC_T	-0.6021	✓
SY_INV_PATTERNS	-0.5923	✓
G_ACTIVE	-0.5882	✓
L_NOUN_PHRASES	0.5486	✓
POS_PRO	-0.5389	✓
L_ADV_SUPERLATIVE	0.5243	✓
L_ADV_COMPARATIVE	0.5002	✓
ASM	-0.4781	✓
CDS	-0.4777	✓
L_SINGULAR_NOUNS	0.4707	✓
POS_NOUN	0.4570	✓
L_YOU_PRON	-0.4543	✓
SENT_D_NP	0.4475	✓

Table 27: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Intensified Persuasive Effect* generation approach and Gemma 3 27b it model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemma - Rewriting with Subtle Persuasive Effect		
L_CONT_T	1.0245	✓
L_CONT_A	1.0176	✓
L_FUNC_A	-0.8850	✓
POS_PRO	-0.7267	✓
SY_INV_PATTERNS	-0.7119	✓
L_PLURAL_NOUNS	0.6762	✓
POS_NOUN	0.6717	✓
SENT_D_PP	0.6401	✓
L_ADJ_POSITIVE	0.6191	✓
G_ACTIVE	-0.5837	✓
L_SECOND_PERSON_PRON	-0.5783	✓
L_YOU_PRON	-0.5736	✓
SY_NARRATIVE	0.5666	✓
POS_ADJ	0.5620	✓
CDS	-0.5438	✓
L_FUNC_T	-0.4852	✓
VF_INFINITIVE	-0.4846	✓
ASM	-0.4624	✓
L_THEY_PRON	-0.4486	✓
POS_CONJ	-0.4429	✓

Table 26: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Subtle Persuasive Effect* generation approach and Gemma 3 27b it model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemma - Open-ended		
SY_IMPERATIVE	1.7354	✓
L_CONT_T	1.4021	✓
L_CONT_A	1.2523	✓
SENT_D_NP	1.1601	✓
PS_CAUSE	-1.1529	✓
VF_INFINITIVE	1.0451	✓
L_FUNC_A	-0.9623	✓
PS_CONDITION	-0.8951	✓
POS_NOUN	0.8730	✓
ST_REPET_WORDS	-0.8055	✓
LTOKEN_RATIO_LEM	0.7851	✓
L_LINKS	-0.7474	✓
SY_INV_PATTERNS	-0.7306	✓
L_SINGULAR_NOUNS	0.6881	✓
SY_EXCLAMATION	0.6825	✓
SENT_D_VP	0.6631	✓
POS_PREP	-0.6571	✓
L_I_PRON	-0.6461	✓
L_FIRST_PERSON_SING_PRON	-0.6461	✓
L_NOUN_PHRASES	0.6320	✓

Table 28: Top 20 linguistic features for AI-generated persuasive text with *Open-ended* generation approach and Gemma 3 27b it model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemini - Paraphrasing		
L_CONT_A	0.7275	✓
L_CONT_T	0.7074	✓
SY_INV_PATTERNS	-0.5802	✓
L_FUNC_A	-0.5371	✓
L_PUNCT_COM	0.4880	✓
LTOKEN_RATIO_LEM	0.4192	✓
L_ADJ_POSITIVE	0.4104	✓
ST_REPET_WORDS	-0.3981	✓
POS_ADJ	0.3677	✓
L_NOUN_PHRASES	0.3255	✓
ASM	-0.2996	✓
PS_CONDITION	-0.2850	✓
PS_CAUSE	-0.2759	✓
L_POSSESSIVES	0.2719	✓
POS_PREP	-0.2305	✓
CDS	-0.2242	✓
L_PUNCT	0.2218	✓
PS_AGREEMENT	-0.2169	✓
L_PLURAL_NOUNS	0.2053	✓
L_THEIR_PRON	0.2050	✓

Table 29: Top 20 linguistic features for AI-generated persuasive text with *Paraphrasing* generation approach and Gemini 2.0 Flash model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemini - Rewriting with Subtle Persuasive Effect		
L_CONT_A	0.9216	✓
L_CONT_T	0.9163	✓
L_FUNC_A	-0.7257	✓
SY_INV_PATTERNS	-0.7182	✓
POS_NOUN	0.5946	✓
G_ACTIVE	-0.5892	✓
POS_PRO	-0.5535	✓
VT_MIGHT	0.5222	✓
CDS	-0.5034	✓
SENT_D_PP	0.4976	✓
L_YOU_PRON	-0.4908	✓
LTOKEN_RATIO_LEM	0.4813	✓
L_PLURAL_NOUNS	0.4761	✓
L_ADJ_POSITIVE	0.4695	✓
L_SECOND_PERSON_PRON	-0.4555	✓
ASM	-0.4517	✓
L_PUNCT_COM	0.4347	✓
ST_REPET_WORDS	-0.4182	✓
POS_ADJ	0.4165	✓
SY_NARRATIVE	0.4079	✓

Table 30: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Subtle Persuasive Effect* generation approach and Gemini 2.0 Flash model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemini - Rewriting with Intensified Persuasive Effect		
L_CONT_A	1.0170	✓
L_CONT_T	0.9948	✓
L_FUNC_A	-0.8186	✓
L_ADJ_POSITIVE	0.7827	✓
POS_ADJ	0.7239	✓
L_PUNCT_COM	0.5901	✓
L_NOUN_PHRASES	0.5756	✓
SY_INV_PATTERNS	-0.5355	✓
LTOKEN_RATIO_LEM	0.4892	✓
L_ADV_SUPERLATIVE	0.4824	✓
PS_CONDITION	-0.4649	✓
L_ADV_COMPARATIVE	0.4629	✓
ASM	-0.4428	✓
ST_REPET_WORDS	-0.4386	✓
PS_CAUSE	-0.4330	✓
POS_ADV	0.4270	✓
G_ACTIVE	-0.4252	✓
POS_PRO	-0.3773	✓
L_FUNC_T	-0.3605	✓
PS_AGREEMENT	-0.3525	✓

Table 31: Top 20 linguistic features for AI-generated persuasive text with *Rewriting with Intensified Persuasive Effect* generation approach and Gemini 2.0 Flash model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.

Stylometric Feature	Cohen's d	Sig.
Gemini - Open-ended		
L_CONT_T	1.5669	✓
L_CONT_A	1.3747	✓
SY_IMPERATIVE	1.3186	✓
LTOKEN_RATIO_LEM	1.1953	✓
SY_EXCLAMATION	1.1092	✓
VF_INFINITIVE	1.0846	✓
ST_REPET_WORDS	-1.0032	✓
SENT_D_VP	0.9291	✓
L_FUNC_A	-0.8914	✓
POS_NOUN	0.8779	✓
L_NOUN_PHRASES	0.8190	✓
PS_CAUSE	-0.7878	✓
PS_CONDITION	-0.7832	✓
G_ACTIVE	-0.7769	✓
L_LINKS	-0.7474	✓
L_SINGULAR_NOUNS	0.7331	✓
SY_INV_PATTERNS	-0.6693	✓
G_PAST	-0.6464	✓
POS_PREP	-0.6279	✓
L_I_PRON	-0.6236	✓

Table 32: Top 20 linguistic features for AI-generated persuasive text with *Open-ended* generation approach and Gemini 2.0 Flash model vs. human-written persuasive texts (three samples of human datasets combined vs. AI counterparts). Cohen's d sorted by absolute value.