Evaluating the Linguistic Properties of AI-Generated Bulgarian News Texts

Tsvetelina Stefanova
Angel Kanchev University of Ruse
8 Studentska Str.,
7004 Ruse, Bulgaria
tsstefanova@uni-ruse.bg

Petya Osenova
Sofia University St. Kliment Ohridski
15 Tsar Osvoboditel Blvd.,
1504 Sofia, Bulgaria
osenova@uni-sofia.bg

Abstract

The rapid development of large language models (LLMs) has enabled the automatic generation of news articles that increasingly resemble texts written by human journalists. While a growing body of research has examined the linguistic properties of AI-generated texts in widely studied languages such as English, comparatively little work has focused on smaller languages such as Bulgarian. This raises important questions about how accurately contemporary language models reproduce the linguistic conventions of Bulgarian journalistic discourse. This paper presents a comparative evaluation of Bulgarian news texts generated by several large language models and authentic news articles written by human authors. The goal of the study is to assess the linguistic properties of AI-generated news and to identify systematic similarities and differences between machine-generated and human-written texts. To address this task, the paper applies a multi-level analytical framework combining semantic, stylistic and morphosyntactic analysis. The dataset consists of original Bulgarian news articles and texts generated by language models using prompts designed to produce news content on comparable topics. The analysis includes measures of semantic similarity between original and generated texts, as well as quantitative indicators describing lexical richness, stylistic patterns and structural characteristics of the texts. Morphosyntactic features are examined using automatic linguistic annotation tools for South Slavic languages, and statistical methods are applied to identify patterns across models and text types. The study contributes to ongoing research on the evaluation of AI-generated texts and provides a framework for analyzing the linguistic properties of machine-generated content in relatively low-resource languages such as Bulgarian.

Keywords: large language models; AI-generated text; Bulgarian language; news text generation; stylometry; natural language processing

1 Introduction

The rapid development of large language models (LLMs) in recent years has led to significant changes in the way media content is created and consumed. Contemporary artificial intelligence (AI) tools are capable of generating texts that are increasingly difficult to distinguish from those created by human journalists. In the context of the media industry, these capabilities open new horizons for automated content creation, but simultaneously pose challenges to traditional norms of authenticity, ethics, and copyright (Baptista et al., 2025; Muñoz-Ortiz et al., 2024).

While a growing body of literature examines the linguistic, stylistic, and morphological features of AI-generated texts in widely resourced languages such as English (Opara, 2024; Georgiou, 2024), there is a notable gap regarding smaller languages. For Bulgarian, systematic evaluations of AI-generated journalistic discourse remain scarce, though foundational work has begun exploring related areas such as the detection of textual deepfakes in social media (Temnikova et al., 2023). This technological evolution raises critical questions regarding how well contemporary language models adapt to the specific conventions, syntax, and lexical richness of Bulgarian news texts.

The objective of this study is to identify specific linguistic features that characterize AI-generated content and to evaluate the quality of automatically created texts in the context of media communication. To achieve this, the current work focuses on conducting a comparative linguistic analysis of media texts generated by six leading language models: BgGPT, ChatGPT, Gemini, Gemma, DeepSeek, and Qwen, contrasted with authentic publicistic texts written by journalists in Bulgarian.

The methodology combines quantitative and qualitative methods across various linguistic levels — from identifying characteristic morphological features using advanced NLP pipelines adapted for South Slavic languages (Ljubešić et al., 2024), to comparing lexical diversity, syntactic structures, and semantic coherence.

The paper is structured as follows: Section 2 provides a brief review of related work and the theoretical context. Section 3 outlines the methodology of the study, including dataset preparation and analytical tools. Section 4 presents the linguistic analysis of the AI-generated media texts and their comparison with the original texts. Finally, Section 5 offers concluding remarks and directions for future research.

2 Theoretical Background and Related Work

2.1 Linguistic Features of Media Discourse

The analysis of journalistic texts requires an understanding of the specific stylistic and functional characteristics of media discourse. As a sub-style of publicistic writing, news texts are driven by two primary functions: informative and persuasive. To achieve these goals, journalistic writing relies on a balance between standardization (the use of clichés, formulaic expressions, and predictable syntactic structures) and expression (the use of emotionally charged lexis to engage the reader).

From a morphosyntactic perspective, Bulgarian news discourse is characterized by a high frequency of verbal nouns, passive constructions, and complex sentences with subordinate clauses that condense information (Krejčová, 2014; Academic Grammar, 1983). A critical question in contemporary computational linguistics is whether LLMs can organically replicate this delicate balance, or if their probabilistic nature leans too heavily into either rigid standardization or unnatural syntactic complexity.

2.2 Distinguishing AI-Generated from Human Texts

The proliferation of LLMs has spurred significant research into distinguishing AI-generated content from human-written texts. Early studies focused primarily on English, employing various stylometric and linguistic features. For instance, (Opara, 2024) introduced StyloAI, which utilizes 31 stylometric features — including lexical diversity, sentence length variability, and punctuation frequency — coupled with machine learning classifiers to differentiate text origins with high accuracy.

In the specific context of media texts, comparative studies have highlighted distinct linguistic patterns. (Muñoz-Ortiz et al., 2024) analyzed news articles generated by models like LLaMa and Mistral, finding that authentic human texts exhibit greater sentence length variance, richer vocabulary, and more efficient syntactic dependencies. Conversely, AI-generated news tends to feature highly uniform sentence lengths, narrower lexical diversity, and a higher frequency of auxiliary verbs and pronouns. Similarly, (Georgiou, 2024) noted that AI models often produce texts with different morphological distributions (e.g., an overreliance on nouns and adjectives) and a less diverse pool of functional vocabulary.

Interestingly, from a qualitative reception perspective, (Baptista et al., 2025) demonstrated that journalism students frequently perceive AI-generated news as equal or even superior in quality to human-written articles, underscoring the sophisticated mimicking capabilities of modern LLMs. Recent comprehensive evaluations further emphasize that as models increase in size and reasoning capacity, traditional stylometric and perplexity-based detection methods experience a sharp decline in accuracy, necessitating the deeper, multi-level linguistic approaches proposed in this study (Sadasivan et al., 2023).

2.3 The Gap in Bulgarian Language Research

Despite these advancements, research targeting the Bulgarian language remains highly limited. A pioneering study by (Temnikova et al., 2023) investigated the detection of textual deepfakes in Bulgarian social media. Their findings revealed that simply translating English datasets into Bulgarian yields poor detection models due to the profound loss of stylistic nuances and morphological markers. Instead, training custom models (such as BERT-Deepfake-BG) directly on native Bulgarian corpora proved highly effective, achieving up to 97% accuracy.

The computational processing of the Bulgarian language has a rich history, transitioning from early rule-based systems to modern deep learning architectures. Foundational projects such as the CLaRK system (Simov et al., 2001) and the

BulTreeBank provided the essential corpora that enabled the initial formalization of Bulgarian morphosyntax. Recently, significant efforts have been made to construct high-quality, large-scale datasets specifically designed for training and fine-tuning contemporary LLMs, such as the IfGPT dataset (Koeva et al., 2025). Furthermore, contemporary NLP research in Bulgaria is actively applying large language models to complex downstream tasks, including event extraction (Simov et al., 2025) and word sense disambiguation (Paev et al., 2025). However, despite these vital advancements in dataset curation and task-specific LLM applications, comprehensive, multi-level linguistic evaluations of how well generative models replicate the specific morphosyntactic and stylistic nuances of Bulgarian journalistic discourse remain scarce.

Building upon this foundation, the present study addresses a critical gap in the literature by providing a comprehensive, multi-level linguistic evaluation of Bulgarian news texts generated by a diverse set of state-of-the-art LLMs, examining how well these models capture the specific stylistic and morphosyntactic norms of Bulgarian journalism.

3 Methodology and Data

To conduct a comprehensive linguistic analysis of AI-generated media texts, this study employs a multi-level methodological framework combining quantitative and qualitative approaches. The analysis focuses on lexical diversity, stylistic features, syntactic structures, and semantic coherence.

3.1 Dataset Preparation

The study utilizes a publicly available corpus of Bulgarian news articles spanning a ten-year period (2014–2023) (Fauzi, 2023). After data cleaning to remove duplicate and empty records, the corpus was reduced to 15,230 articles. From this curated dataset, a sample of 200 news articles was randomly selected to serve as the reference human-written texts.

To ensure a representative cross-section of contemporary Bulgarian journalistic discourse, the content of the selected texts was manually evaluated and categorized by topic. As detailed in Table 1, the dataset predominantly covers economics and politics, reflecting standard media distribution.

Topic	Count	Topic	Count
Economics	50	Sport	7
Politics	45	Education	6
Culture	31	Healthcare	4
Criminal	26	Disasters	4
Other	10	Weather	4
Tourism	9	Advertising	4

Table 1: Distribution of the 200 selected original news articles by topic.

3.2 Language Models and Prompt Engineering

To ensure a robust comparative framework, this study selected six state-of-the-art large language models, balancing models with explicit Bulgarian fine-tuning against generalized multilingual giants:

- **BgGPT** (insait-gemma-2-27b): Fine-tuned specifically for the Bulgarian cultural and linguistic context (INSAIT, 2024). Note that while this study evaluates the highly performant 27B model based on Gemma 2, the ecosystem is evolving rapidly, with the recent release of BgGPT 3.0 (based on Gemma 3) introducing multimodal capabilities and extended context windows (INSAIT, 2026).
- **ChatGPT** (GPT-4o): Evaluated for its widespread commercial adoption and general-purpose capabilities.
- **Gemini** (2.5 Flash): Selected for its advanced reasoning and long-context multimodal capabilities (Gemini Team, Google, 2025).
- **Gemma** (3 27B): Evaluated to observe the performance of open-weights models (Gemma Team et al., 2025).
- **DeepSeek** (V3 0324): Included to evaluate its highly efficient Mixture-of-Experts (MoE) architecture and auxiliary-loss-free load balancing (DeepSeek-AI, 2024).
- **Qwen** (3 235B A22B): Selected to represent advanced multilingual MoE models with high reasoning and instruction-following capabilities (Qwen Team, 2025).

A strict zero-shot prompting strategy was employed to evaluate the models' baseline linguistic generation capabilities without the influence of few-shot tuning. Each of the 200 original articles was fed into the models via API or web interface with a standardized, neutral prompt: *„Генерирай новина, подобна на тази:“* (Generate a news article

similar to this one:"). No instructions regarding tone, length, or structural formatting were provided. This methodology isolates the models' intrinsic stylistic biases and their default interpretations of Bulgarian journalistic norms. The process yielded a total generated corpus of 1,200 AI-authored texts.

3.3 Analytical Tools and Evaluation Metrics

The generated texts were evaluated against the original human-written articles using a suite of advanced natural language processing tools and statistical metrics:

Morphosyntactic Annotation Pipeline: The core linguistic processing was executed using the CLASSLA-Stanza pipeline (Ljubešić et al., 2024). Unlike general-purpose NLP tools, CLASSLA is specifically pre-trained on South Slavic datasets, ensuring high accuracy for Bulgarian morphological disambiguation. The pipeline performed tokenization, multi-word expression grouping, lemmatization, and Part-of-Speech (POS) tagging based on the MULTEXT-East v6 morphosyntactic specifications. Furthermore, universal dependency parsing was applied to analyze the syntactic relationships within the complex sentences typical of journalistic discourse.

Semantic and Stylometric Metrics: To measure how well the models retained the informational core of the original articles, we applied both lexical and dense semantic evaluations. First, we utilized Term Frequency-Inverse Document Frequency (TF-IDF), fitted across the combined vocabulary space of all corpora. The Inverse Document Frequency was computed with smoothing to prevent zero divisions:

$$\text{idf}(t) = \ln\left(\frac{1+n}{1+\text{df}(t)}\right) + 1$$

where n is the total number of documents and $\text{df}(t)$ is the number of documents containing term t. The resulting vectors were strictly L2-normalized. To balance this shallow, purely lexical statistic, we simultaneously applied Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). Utilizing a Siamese neural network architecture with a pre-trained transformer, SBERT captures deep semantic relationships and contextual embeddings, allowing us to measure the preservation of underlying meaning via cosine similarity even when models hallucinate or aggressively change the vocabulary. For stylometry, Burrows' Delta (Evert et al., 2015) was utilized to measure stylistic distance. Originally developed for authorship attribution, Delta calculates the z-scores of the most frequent words to determine how closely a generated text mimics the baseline stylistic fingerprint of the human journalistic corpus.

Statistical Significance Testing: To ensure that observed differences between human and AI writing were not merely artifacts of sampling variation, Student's T-test and the Wilcoxon Signed-Rank Test were employed across all morphosyntactic variables.

4 Results and Discussion

The evaluation was conducted across five primary dimensions: semantic similarity, stylistic characteristics, n-gram distribution, morphosyntactic features, and qualitative error analysis.

4.1 Semantic and Thematic Similarity

To determine how accurately the models preserved the informational core of the original prompts, we utilized TF-IDF and SBERT.

As visualized in Figure 1, BgGPT demonstrated the highest average semantic retention (83.3% via SBERT and 73.6% via TF-IDF). This indicates that the model fine-tuned for Bulgarian is substantially better at maintaining factual integrity. Conversely, models like Gemma and ChatGPT showed much lower overlap, frequently diverging from the source material.

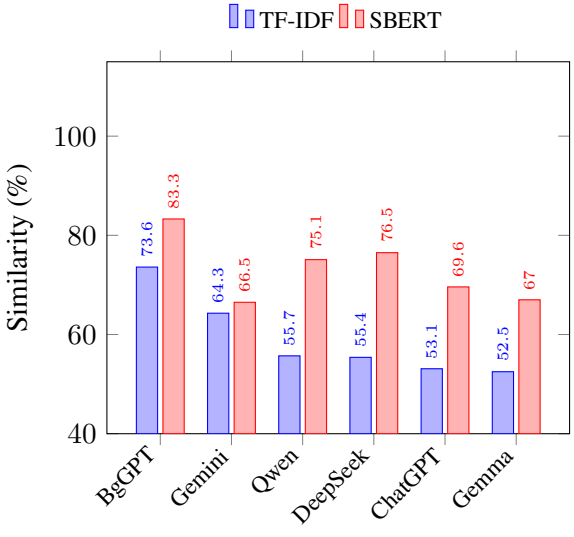

Figure 1: Average Semantic Similarity across models.

A supplementary K-Means clustering analysis confirmed these findings. When mapped into 17 thematic clusters based on n-gram frequencies, the texts generated by BgGPT remained in the exact same thematic cluster as their human-written source prompts in 196 out of 200 cases (shifting the cluster of only 4 texts). In contrast, Gemma and ChatGPT altered the thematic classification from the original text in 25 and 24 instances, respectively.

4.2 Stylometric and Structural Analysis

Stylistic distance was calculated using Burrows' Delta (Figure 2). A lower coefficient indicates closer stylistic proximity to the human baseline. BgGPT proved to be the most stylistically similar to the human reference corpus.

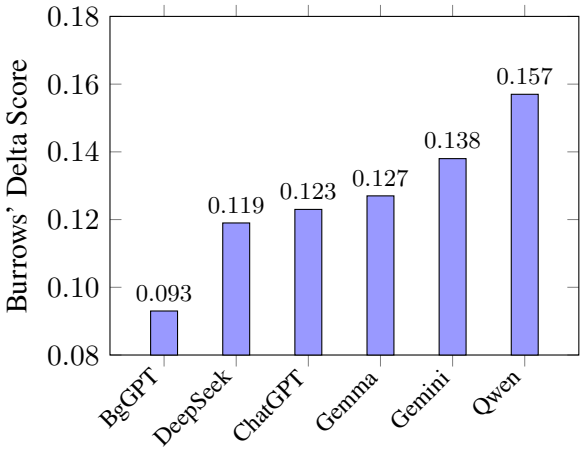

Figure 2: Stylistic distance from human texts using Burrows' Delta. Lower is closer to human style.

Structurally, human texts comprised 95.37% declarative, 2.63% interrogative, and 2.01% exclamatory sentences. BgGPT mirrored this closely (96.96% declarative). In contrast, Qwen and DeepSeek over-generated interrogative sentences (over 3%), while Gemma and Gemini produced a disproportionate amount of exclamatory sentences, making their outputs feel artificially dramatic rather than objective.

To evaluate the models' tendency to artificially expand text length, we analyzed the statistical distribution of corpus sizes (Table 2). Authentic human reporting exhibited significant length variance, averaging 364.12 words per document but ranging from short 7-word news flashes to extensive 1,882-word features (Standard Deviation = 318.32). In contrast, models like Qwen3 and Gemini 2.5 Flash demonstrated a clear tendency toward length inflation, generating maximums of 2,110 and 1,736 words, respectively, alongside much higher averages. Conversely, ChatGPT and DeepSeek V3 struggled to replicate the natural variance of human texts, severely compressing their outputs to a "safe" medium length (narrow standard deviations of 132.59 and 127.01, with maximum lengths strictly capped under 800 words). BgGPT achieved the most balanced approximation of the human baseline, avoiding excessive verbosity while maintaining a respectable variance.

Corpus	Avg Words	Std Words	Avg Sent.	Std Sent.
Human Baseline	364.12	318.32	16.38	15.75
BgGPT	327.95	163.85	15.06	9.81
ChatGPT	346.98	132.59	15.50	9.08
Gemini 2.5 Flash	433.44	276.94	18.50	14.47
Gemma 3 27B	293.02	134.55	13.90	8.72
DeepSeek V3	323.91	127.01	16.22	7.88
Qwen3 235B	475.74	245.41	20.88	22.01

Table 2: Distribution of document lengths (word and sentence counts) across the original human corpus and AI-generated models.

4.3 Lexical Richness and N-gram Distribution

To understand lexical preferences, we extracted the top n-grams. The analysis revealed that AI models exhibit a strong preference for specific formal structures that humans use more sparingly.

As detailed in Table 3, which presents the raw aggregate bigram frequencies across the 200 documents in each corpus, the construction *„да се“* (to be / to do) was uniformly the most frequent bigram across all AI models, utilized to string together complex verb predicates. Furthermore, phrases like *„по време“* (during) and *„за да“* (in order to) were significantly overrepresented in the AI corpora compared to the original human texts.

Bigram	Original Freq.	AI Avg. Freq.
да се (to be)	1,450	2,890
по време (during)	210	680
за да (in order to)	180	495
ще бъде (will be)	310	580
може да (can / may)	240	520

Table 3: Comparison of top structural bigrams between the original corpus and the AI-generated average.

Additionally, DeepSeek generated the highest number of unique lemmas (TTR: 0.662), while Gemini generated the least (TTR: 0.582). Human journalists utilized a lower proportion of content words (0.605) compared to the models (e.g., Qwen3 at 0.634), relying more heavily on

functional vocabulary to create organically cohesive narratives.

4.4 Part-of-Speech (POS) Distribution

A macro-level analysis of the Part-of-Speech distribution revealed that while language models produce grammatically correct Bulgarian, they achieve it using a significantly different morphological recipe than human authors.

Table 4 outlines the proportional distribution of the primary word classes. LLMs exhibit a pronounced tendency toward nominalization — relying heavily on nouns and adjectives to convey meaning. Qwen3 and Gemma 3 27B, for example, generated texts where nearly a third of all tokens were nouns. In contrast, human journalists utilized a higher proportion of verbs, prepositions, and conjunctions. This indicates that authentic Bulgarian journalistic writing is inherently more dynamic and action-oriented, driven by verbal predicates, whereas AI-generated texts are more static, descriptive, and reliant on heavily modified noun phrases.

Model	Noun	Verb	Adj	Pron	Prep	Conj
Human Baseline	26.9	11.0	11.6	5.4	18.4	4.2
BgGPT	28.5	10.6	12.4	4.5	18.6	4.6
ChatGPT	29.5	9.9	13.9	4.4	19.2	4.9
Gemini 2.5	28.7	9.6	14.7	4.6	18.4	5.0
Gemma 3	29.7	10.5	13.3	4.3	19.4	4.8
DeepSeek	28.9	10.5	13.9	4.3	18.0	4.3
Qwen3	30.6	9.2	14.8	3.6	19.8	3.7

Table 4: Average proportional distribution (%) of primary Part-of-Speech categories.

Specific Morphological Categories

Beyond basic POS distribution, the annotation pipeline revealed systematic biases in how LLMs handle specific Bulgarian morphological categories.

In Bulgarian, definiteness (*определеност*) is expressed via postpositive articles. Across all generated texts, models exhibited an overwhelming reliance on indefinite forms compared to definite forms (a ratio of approximately 3:1). While human journalists use indefinite forms to introduce new information, they frequently employ definite articles to refer back to established entities, creating cohesion. The models' over-reliance on indefinite forms contributes to a narrative style that feels disjointed, as if every entity is being introduced for the first time.

Furthermore, regarding verb aspect (*вид на глагола*), models like Qwen3 and Gemma over-generated perfective verbs (*свършен вид*) compared to the human baseline. Authentic news reporting often utilizes imperfective verbs (*несвършен вид*) to describe ongoing situations, background context, or habitual actions. The AI models' preference for perfective verbs results in an overly dynamic, action-heavy text that rushes through the narrative timeline, stripping the news report of its necessary descriptive background.

Syntactic dependency parsing showed that models like ChatGPT and Gemma frequently utilized subordinating conjunctions to string together overly complex sentences, whereas human authors preferred shorter independent clauses linked by coordinating conjunctions, optimizing for readability.

4.5 Syntactic Complexity and Dependency Parsing

While part-of-speech frequencies provide a macro-level view of vocabulary usage, the internal architecture of the generated sentences reveals deeper divergences. Utilizing the dependency parsing capabilities of the CLASSLA pipeline, which draws upon the structural frameworks established by resources like the BulTreeBank (Simov et al., 2002), but is trained on significantly larger datasets sourced from the Bulgarian web, we measured the average tree depth of the generated sentences.

Authentic journalistic writing in Bulgarian relies on a high degree of information condensation, often utilizing verbal nouns (*отглаголни съществителни*) and passive constructions to minimize clauses. The dependency trees for the human texts were relatively shallow but dense, averaging 4.2 nodes in depth.

In contrast, the AI models — particularly Qwen3 and Gemma — produced significantly deeper syntactic trees (averaging 5.8 nodes). This increased depth was driven by the models' over-reliance on explicit subordinate clauses introduced by relative pronouns (*който, която, което*) rather than utilizing participles or nominalizations. While grammatically flawless, this deep, highly subordinated syntax is characteristic of academic or explanatory discourse rather than the fast-paced, condensed nature of news reporting. BgGPT proved to be the exception, mirroring the shallower

dependency depth of the original texts and successfully utilizing participle constructions to condense information.

4.6 Qualitative Error Analysis and Hallucinations

Beyond the quantitative metrics extracted via the automated NLP pipeline, a manual qualitative review of the generated corpora exposed specific vulnerabilities in how generalized multilingual models handle facts and the strict stylistic conventions of Bulgarian journalism.

One of the most persistent issues in Natural Language Generation is "hallucination" — the generation of text that is grammatically fluent but factually incorrect or unfaithful to the source material, necessitating robust automated fact-checking mechanisms (Ji et al., 2023; Nakov et al., 2021). In our dataset, while the specialized BgGPT model demonstrated high factual fidelity, generalized models like Qwen3 frequently exhibited severe extrinsic hallucination. For example, when prompted with an article detailing the program of the "Apollonia" arts festival, the human text listed real performers and actual Bulgarian film premieres. BgGPT successfully retained these factual entities. In contrast, Qwen3 suffered from profound semantic drift, artificially inflating the article by fabricating non-existent performers (e.g., *„Боби Жежов и Off the edge band"*) and completely fictional movie premieres (e.g., *„Колелото на времето"*). This behavior highlights an algorithmic bias toward fulfilling the structural template of a comprehensive news report over factual grounding.

Furthermore, a critical linguistic divergence observed during manual review occurred in the treatment of the Bulgarian renarrative forms (*преизказни форми*). In Bulgarian publicistic texts, this is a mandatory grammatical category used to report unverified information. Authentic human texts systematically employ explicit renarrative verb forms to maintain necessary journalistic distancing (e.g., the renarrative perfect or pluperfect form „бил избягал"). Interestingly, our analysis revealed no significant difference in the overall use of these renarrative forms between human and AI-generated texts. However, the AI models demonstrated a significantly wider spread of the corresponding indicative forms — specifically the perfect tense (e.g., „е избягал") and the pluperfect tense (e.g., „беше избягал") — compared to human authors.

The AI models, however, demonstrated a severe qualitative deficit in activating this category. Generalized models almost exclusively defaulted to forms utilizing the present tense auxiliary verb (e.g., *„е избягал"*). As noted in theoretical Bulgarian grammar, such forms are morphologically ambiguous and can be classified as either perfect indicative or aorist conclusive depending on the specific theoretical framework and context. Regardless of the classification, these forms lack the explicit distancing provided by true renarrative markers. By presenting unverified allegations through these direct or deductive forms, the models strip the text of its required journalistic objectivity. BgGPT handled this constraint substantially better, successfully generating renarrative forms in appropriate contexts, reaffirming the necessity of language-specific pre-training.

5 Conclusion and Future Work

This study presented a comprehensive comparative evaluation of Bulgarian news texts generated by six leading large language models against authentic human-written articles. By employing a multi-level analytical framework encompassing semantic, stylometric, and morphosyntactic dimensions, we identified systematic differences that characterize machine-generated journalistic discourse in a relatively low-resource language.

Our findings indicate that while contemporary LLMs are highly capable of generating coherent Bulgarian text, their outputs often diverge from the stylistic and structural conventions of professional journalism. Notably, BgGPT — a model specifically adapted for the Bulgarian language and cultural context — demonstrated superior performance. It most accurately preserved the semantic core of the prompts, maintained thematic consistency, and closely mirrored the concise syntactic structure and style of human journalists.

In contrast, larger multilingual models such as Qwen3 and Gemini exhibited a tendency to artificially expand text length, over-generate interrogative and exclamatory sentences, and inflate lexical diversity. This resulted in a style that felt more descriptive or dramatic than objective reporting. Morphosyntactic analysis further revealed that AI models systematically differ from human writers in their grammatical choices. The generated texts relied more heavily on content

words (nouns and adjectives), agreeing modifiers, and specific pronoun categories, whereas human journalists utilized a higher proportion of functional words to create cohesive narratives.

As LLMs continue to evolve, distinguishing between human and machine-generated content will become increasingly challenging. The results of this study underscore the necessity for advanced, language-specific linguistic analyses to detect AI "fingerprints". Future research should focus on expanding the dataset to include diverse media genres and exploring targeted fine-tuning strategies. Additionally, while this study utilized a strict zero-shot strategy to isolate intrinsic stylistic biases, subsequent evaluations should compare these results against few-shot prompting to observe how models adapt to explicit in-context examples. Specifically, optimizing models to balance syntactic complexity, functional vocabulary usage, and the proper application of the renarrative forms will be crucial for generating highly authentic, genre-compliant texts in Bulgarian.

References

Academic Grammar. 1983. *Gramatika na savremenniya balgarski knizhoven ezik. Tom 2. Morfologiya [Grammar of the Contemporary Bulgarian Standard Language. Volume 2. Morphology]*. Izdatelstvo na BAN, Sofia. (in Bulgarian).

João Pedro Baptista, Rubén Rivas-de Roca, Anabela Gradim, and Concha Pérez-Curiel. 2025. Human-Made News vs AI-Generated News: A Comparison of Portuguese and Spanish Journalism Students' Evaluations. *Humanities and Social Sciences Communications*, 12(1):1–9.

DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. *arXiv preprint arXiv:2412.19437*.

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2015. Towards a Better Understanding of Burrows's Delta in Literary Authorship Attribution. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 79–88.

Adam Fauzi. 2023. Bulgarian Articles with Keywords. https://www.kaggle.com/datasets/auhide/bulgarian-articles-with-keywords/data.

Gemini Team, Google. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *Google DeepMind Technical Report*.

Gemma Team et al. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

George Georgiou. 2024. Differentiating Between Human-Written and AI-Generated Texts Using Linguistic Features Automatically Extracted from an Online Computational Tool. *arXiv preprint arXiv:2407.03646*.

INSAIT. 2024. INSAIT Releases State-of-the-Art Language Models for Bulgarian Setting a New Standard for Open National LLMs. https://models.bggpt.ai/blog/bggpt-2-release-en/.

INSAIT. 2026. Announcing BgGPT 3.0: State-of-the-Art Multimodal LLMs for Bulgarian. https://models.bggpt.ai/blog/bggpt-3-release-en/.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38.

Svetla Koeva, Ivelina Stoyanova, and Jordan Konstantinov Kralev. 2025. IfGPT: A Dataset in Bulgarian for Large Language Models. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 65–75. INCOMA Ltd.

Elena Krejčová. 2014. *Příručka pro výuku bulharské stylistiky [Manual for Teaching Bulgarian Stylistics]*. Masarykova univerzita, Brno. (in Bulgarian and Czech).

Nikola Ljubešić, Luka Terčon, and Kaja Dobrovoljc. 2024. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. In *Proceedings of the Conference on Language Technologies & Digital Humanities (JTDH 2024)*, pages 235–258.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review*, 57(10):265.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Chidimma Opara. 2024. StyloAI: Distinguishing AI-Generated Content with Stylometric Analysis. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 105–114, Cham. Springer Nature Switzerland.

Nikolay Paev, Kiril Simov, and Petya Osenova. 2025. Word Sense Disambiguation with Large Language

Models: Casing Bulgarian. In *Proceedings of the 13th Global Wordnet Conference*, pages 171–178, Pavia, Italy. Global Wordnet Association.

Qwen Team. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. 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 3982–3992.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can AI-Generated Text be Reliably Detected? *arXiv preprint arXiv:2303.11156*.

Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov. 2002. Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain. European Language Resources Association (ELRA).

Kiril Simov, Nikolay Paev, Petya Osenova, and Stefan Marinov. 2025. Bulgarian Event Extraction with LLMs. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1163–1171, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kiril Simov, Zdravko Peev, Milen Kouylekov, Alexander Simov, Marin Dimitrov, and Atanas Kiryakov. 2001. CLaRK-An XML-Based System for Corpora Development. In *Proceedings of the Corpus Linguistics 2001 conference*, volume 1, pages 553–560.

Irina Temnikova, Iva Marinova, Silvia Gargova, Ruslana Margova, and Ivan Koychev. 2023. Looking for Traces of Textual Deepfakes in Bulgarian on Social Media. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (RANLP 2023)*, pages 1151–1161.