Corpus-Linguists' Little Helpers? Evaluating LLMs for Linguistic Annotation: The Case of Sensationalist Headlines Corpus

Petra Bago, Virna Karlić
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{pbago, vkarlic}@ffzg.unizg.hr

Abstract
Manual annotation of pragmastylistic features in sensationalist media is a resource-intensive bottleneck for corpus-based research, particularly for lower-resource languages. This paper evaluates whether Large Language Models (LLMs) can reliably automate this process. We benchmark two proprietary models, OpenAI's GPT-5 and Google's Gemini 2.5 Pro, on annotating eight sensationalist linguistic and orthographic features within a corpus of 508 Serbian celebrity magazine headlines. Our methodology involves a systematic comparison of five prompting strategies: zero-shot, few-shot (1, 3, and 5 examples), and chain-of-thought. Results demonstrate that LLMs can achieve high alignment with a manually curated gold standard, reaching a peak macro-F1 score of 98.76%. Notably, the most effective and cost-efficient configuration was GPT-5 using a simple zero-shot prompt. Qualitative error analysis reveals that remaining inaccuracies are systematic, primarily involving pragmatic conventions, discourse scope, and quoted speech. We conclude that LLMs are viable for first-pass annotation of well-defined features in Serbian, though implicit and genre-dependent cues require further study. To support reproducibility and future research on underrepresented languages, we provide our full prompting setup, evaluation procedures, and a detailed cost comparison.

Keywords: Large Language Models (LLMs), Automated Annotation, Sensationalism, Prompting Strategies, Underrepresented Languages, Serbian Language

1. Introduction

The proliferation of digital media has fundamentally altered how information is presented and consumed, leading to the rise of sensationalist language in headlines designed to capture audience attention in a crowded online space. Analyzing these pragmastylistic features is crucial for understanding modern media discourse, but this has traditionally relied on manual corpus annotation—a process that is labor-intensive, time-consuming, and a significant bottleneck for large-scale research. The foundational linguistic work on which this study is based, for example, required expert annotators to manually code hundreds of headlines for 13 distinct features, highlighting the need for more efficient methods (Karlić and Bartol, 2024).

Recent work suggests that large language models (LLMs) can support or automate parts of annotation workflows, often at a substantially lower cost than manual labeling, although performance depends on task design, prompting, and the degree of linguistic explicitness in the target categories (Gilardi et al., 2023; Tan et al., 2024; Brown et al., 2020; Wei et al., 2022). However, such evaluations remain much rarer for underrepresented languages than for English, and especially rare for fine-grained corpus-linguistic annotation tasks. Our case study addresses this gap by evaluating two proprietary LLMs (OpenAI's GPT-5 and Google's Gemini 2.5 Pro) on the annotation of the eight sensationalist language features in Serbian headlines. We compare five prompting strategies, analyze recurring error patterns, and discuss cost and reproducibility. In this way, the paper contributes both an empirical benchmark for Serbian and a compact methodological reference point for similar work on low-resource or underrepresented languages.

More specifically, our contributions are threefold: (i) we test whether highly explicit sensationalist language features can be annotated reliably in Serbian using LLMs; (ii) we compare prompt designs under identical evaluation conditions; and (iii) we identify which error types still require human review, thereby clarifying the realistic role of LLMs in corpus annotation workflows.

2. Related Work

The present study builds directly upon the corpus-pragmatic research of Karlić and Bartol (2024), who investigated the features of sensationalist language in Serbian digital media. Their work established a foundational, manually annotated corpus for this task, consisting of 508 headlines collected from the online edition of Serbia's *Hello!* magazine during March and April 2024, identifying 13 relevant features of sensationalist language. We use that corpus as a gold standard for testing whether LLMs can support this type of linguistic annotation in Serbian. This automation addresses a key challenge in natural language processing (NLP): traditional data annotation is labor-intensive and costly (Gilardi et al., 2023; Tan et al., 2024). The emergence of LLMs presents a cost-effective solution, with studies showing that models like ChatGPT can significantly outperform human crowd-workers on annotation tasks at a fraction of the cost (Gilardi et al., 2023).

The effectiveness of LLMs is highly dependent on the prompting methodology. The paradigm of few-shot learning, popularized by Brown et al. (2020), showed that LLMs can perform new tasks by conditioning on a few examples without gradient

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 33–41
11–12 May 2026.

updates. However, Min et al. (2022) later found that the correctness of the input-label pairs in these examples is less important than their ability to specify the label space, input distribution, and output format. For more complex tasks, chain-of-thought (CoT) prompting elicits intermediate reasoning steps, either through few-shot exemplars (Wei et al., 2022) or simple zero-shot triggers like “Let’s think step by step” (Kojima et al., 2022). This technique has since been refined with methods like self-consistency (Wang et al., 2023) and used to augment knowledge for smaller models (Wu et al., 2023).

These methods are increasingly being applied to complex tasks such as pragmatic and stylistic analyses of hate speech and stance detection (Albladi et al., 2025). However, research highlights substantial challenges, as LLM performance is highly dependent on the degree of textual explicitness—models tend to struggle when the stance is implied or the target is not explicitly mentioned (Li and Conrad, 2024; Akash, Fahmy, and Trabelsi, 2025).

Considering the advantages and limitations of these methods, our case study addresses a gap at the intersection of these fields by systematically evaluating prompting strategies for pragmastylistic language feature annotation, using sensationalist headlines in Serbian, a language still underrepresented in LLM evaluation. The analysis focuses on explicit (graphostylistic and morphosyntactic) features of sensationalist language, leaving more implicit categories for future research.

3. Corpus and Gold Standard Annotation

The dataset used in this study is a specialized corpus of 508 headlines collected from the online edition of *Hello!*, which is Serbia's most-read celebrity magazine. All headlines were published in the "Celebrity News" section during a two-month period covering March and April 2024. The language of the corpus is Serbian. This source was deemed highly suitable for the analysis of sensationalist language, as the celebrity news genre is characterized by the prominent use of such discourse strategies. The original research for which the corpus was developed focused on analyzing the prevalence and function of 13 categories of sensationalist language features, which can be organized into four groups: graphostylistic, morphosyntactic, lexico-semantic and pragmastylistic features. This study focuses on the first two groups, while the remaining two will be addressed in future research.

To establish a reliable benchmark for our automated annotation experiments, we utilize the manually annotated version of this corpus as our gold standard. The creation of this gold standard was a rigorous, multi-step process. The annotation was performed by two expert linguists, who are the authors of the original study on this corpus. Each author first annotated the entire corpus independently. Subsequently, their annotations were systematically compared to identify any differences. All discrepancies between the two annotators were then resolved through discussion until a final consensus was reached for every headline. This consensus-based approach ensures a high-quality and reliable dataset for evaluating the performance of our LLM-based methods.

While the original annotation scheme covered 13 pragmastylistic categories, the present study uses a subset of eight features that can be operationalized primarily through orthographic and morphosyntactic cues: (1) `Question`, (2) `Exclamation Mark`, (3) `Ellipsis`, (4) `Second-Person ('you'-form)`, (5) `First-Person Plural ('we'-form)`, (6) `Imperative`, (7) `Superlative`, and (8) `Writing in Caps`.

This restriction is deliberate: our aim is not to claim full automation of sensationalism analysis, but to test the upper bound of LLM performance on features with comparatively clear formal correlates. At the same time, the corpus available for this study is text-only. Visual cues such as font size, boldface, or page placement may also contribute to sensationalist framing, but they are not represented in the present dataset and therefore fall outside the scope of the current evaluation.

4. LLM-based Annotation Methodology

To assess whether LLMs can support annotation of sensationalist language features in Serbian, we compared two models under multiple prompting conditions against the gold standard described above.

4.1 Models

We selected two state-of-the-art proprietary models for our experiments: **OpenAI's GPT-5** and **Google's Gemini 2.5 Pro**. Both models were accessed via their respective official APIs. To investigate the impact of output variability on performance, we tested **Gemini 2.5 Pro** at three different **temperature settings**: **0.2** (more deterministic), **0.5** (balanced), and **0.8** (more creative). For all experiments involving **GPT-5**, a **constant default temperature** was used as the model does not have the option to regulate the temperature. We limit the comparison to these two proprietary models because they were available within the project constraints and allowed consistent API-based evaluation. Newer proprietary families and open-source alternatives are important comparators, but a broader benchmark is left for future work.

4.2 Prompting Strategies

We compare five prompting strategies. In all cases, models were instructed to output a structured JSON format to enable automated evaluation.

4.2.1 Zero-Shot (ZS)

In this configuration, the model receives no in-context examples. The prompt provides a persona ("You are an expert linguist"), a clear objective, detailed definitions for each of the eight linguistic features, and instructions for the required JSON output format. This strategy tests the model's foundational, pre-trained knowledge of linguistic concepts.

4.2.2 Few-Shot (FS)

This strategy builds on the zero-shot prompt by including a set of correct headline-annotation pairs before presenting the target headline. The goal is to provide the model with in-context examples of the desired output. We tested three variants:

- **Few-Shot (n=1)**: One example is provided.
- **Few-Shot (n=3)**: Three examples are provided.
- **Few-Shot (n=5)**: Five examples are provided.

4.2.3 Chain-of-Thought (CoT)

This is an advanced prompting technique designed to encourage a more deliberative reasoning process. The CoT prompt explicitly instructs the model to first generate a step-by-step analysis of the headline for each feature in a "reasoning" field before producing the final boolean annotations in an "annotation" field. This method tests whether forcing the model to articulate its reasoning process improves its accuracy on potentially linguistically ambiguous cases.

For reproducibility, the prompt family remained constant across models and conditions: the same feature definitions, output schema, and evaluation headlines were used throughout, with only the number of in-context examples or the reasoning instruction changed between conditions. The exact prompts used in the experiments are provided in Appendix A.

4.3 Experimental Setup and Evaluation

Our experimental setup consisted of processing each of the 508 headlines from our gold standard corpus with every model configuration. This resulted in 5 distinct runs for GPT-5 and 15 runs for Gemini 2.5 Pro (5 strategies × 3 temperatures), for a total of 20 complete annotation sets.

The model-generated annotations were then automatically compared against the gold standard. Given the imbalanced distribution of features in the corpus, we report precision, recall, and F1-score for each feature:

- **Precision**: The proportion of positive identifications that were actually correct.
- **Recall**: The proportion of actual positives that were correctly identified.
- **F1-Score**: The harmonic mean of precision and recall, providing a single score that balances both metrics.

For overall comparison, we report micro- and macro-averaged F1-scores:

- **Micro-Averaged F1-Score** aggregates decisions across all eight features and gives more weight to frequent categories.
- **Macro-Averaged F1-Score** averages feature-wise F1-scores and therefore weights frequent and infrequent categories equally. We use this measure as the primary comparison metric because it better reflects balanced performance across both common and rare features.

5. Results

This section presents the quantitative results of our experiments. We evaluated a total of 20 model configurations against a gold standard corpus of 508 headlines. Our primary metric for comparing overall performance is the macro-averaged F1-score, which accounts for class imbalance across the eight annotated features.

5.1 Overall Performance

Both models demonstrated exceptionally high performance, with the top configurations achieving macro-F1 scores approaching 99%.

The **highest-performing** configuration overall was OpenAI's **GPT-5 using a zero-shot prompt**, which achieved a **macro-F1 score of 98.76%**.

Google's **Gemini 2.5 Pro** also performed strongly, with its **best results** obtained using few-shot prompting. The peak performance for Gemini 2.5 Pro was a **macro-F1 score of 98.69%**, achieved by several configurations, including **few-shot (n=3 & n=5) at a temperature of 0.5** and **few-shot (n=5) at a temperature of 0.8**. A summary of model performance across all experiments, along with associated API costs, is presented in Table 1.

5.2 Impact of Prompting Strategies

The effectiveness of the five prompting strategies varied between the two models, as detailed in Table 1.

For **GPT-5**, performance was remarkably stable across all prompting methods, with macro-F1 scores varying by less than 0.1%. Surprisingly, the simplest **zero-shot strategy yielded the highest score (98.76%)**. The addition of examples in few-shot settings and the guided reasoning of CoT did not lead to improvements, with the lowest score being a still-high 98.67% for few-shot (n=5) and CoT.

For **Gemini 2.5 Pro**, there was a clearer, albeit small, benefit to using few-shot examples. The zero-shot approach peaked at a macro-F1 of 98.59%, while the **few-shot strategies** consistently performed better, reaching a maximum of **98.69% with both n=3 and n=5 examples**. The temperature parameter did not show a monotonic effect; however, the **model's peak performance for most strategies** was often achieved at a **moderate temperature of**

0.5. Average precision, recall and macro F1-score per model across all 20 experiments are presented in Appendix B.

5.3 Feature-Level Performance

While overall performance was high, the models' ability to correctly identify features varied. Table 2 shows the average F1-score for each of the eight features, calculated across all 20 experimental runs.

Features defined by simple, unambiguous orthographic rules were the easiest for the models to annotate. **`Ellipsis`** was identified perfectly in all runs, achieving a **100.00% average F1-score**, followed closely by **`Exclamation Mark`** at **99.80%**.Conversely, features requiring a degree of contextual interpretation proved more challenging.

The **most difficult feature** to annotate was **`Writing in Caps`** used for emphasis, with the **lowest average F1-score of 96.05%**. An analysis of the best-performing run (GPT-5 zero-shot) reveals that this was largely due to lower recall (95.47%) compared to precision (97.51%), indicating the model failed to identify some true instances of the feature. The second most challenging category was **`Second-Person ('you'-form)`** used for addressing the reader, with an **average F1-score of 98.03%**. All other features were annotated with an average F1-score above 98.4%. Average precision, recall, and F1-Score per feature across all 20 experiments are presented in Appendix C.

Model	**Best Configuration**	**Macro-F1**	**Micro-F1**	**Total Cost**
GPT-5	ZS	**98.76%**	**98.65%**	$18.20 ($3.64 per experiment)
	FS (n=1)	98.68%	98.55%	
	FS (n=3)	98.73%	98.62%	
	FS (n=5)	98.67%	98.52%	
	CoT	98.67%	98.55%	
Gemini 2.5 Pro	ZS (t=0.2)	98.55%	98.43%	$86.29 ($5.75 per experiment, 58% more expensive)
	ZS (t=0.5)	98.56%	98.43%	
	ZS (t=0.8)	98.59%	98.47%	
	FS (n=1, t=0.2)	98.67%	98.55%	
	FS (n=1, t=0.5)	98.62%	98.50%	
	FS (n=1, t=0.8)	98.55%	98.43%	
	FS (n=3, t=0.2)	98.58%	98.45%	
	FS (n=3, t=0.5)	**98.69%**	**98.57%**	
	FS (n=3, t=0.8)	98.57%	98.45%	
	FS (n=5, t=0.2)	98.62%	98.50%	
	FS (n=5, t=0.5)	**98.69%**	**98.57%**	
	FS (n=5, t=0.8)	**98.69%**	**98.57%**	
	CoT (t=0.2)	98.63%	98.50%	
	CoT (t=0.5)	98.68%	98.57%	
	CoT (t=0.8)	98.60%	98.47%	

Table 1: Summary of Model Performance Across All Experiments: Micro- and Macro-F1 Scores.

Feature	**Average F1-Score**
`Ellipsis`	100.00%
`Exclamation Mark`	99.80%
`Question`	99.23%
`Imperative`	99.06%
`Superlative`	98.50%
`First-Person Plural ('we'-form)`	98.42%
`Second-Person ('you'-form)`	98.03%
`Writing in Caps`	96.05%

Table 2: Average F1-score per feature across all 20 experiments, ranked from highest to lowest.

Overall, these results suggest that the current task is highly tractable for LLMs when the target categories are formally explicit and text-internal. They should not be interpreted as evidence that all sensationalist phenomena can be annotated equally well.

6. Discussion and Error Analysis

The results of our experiments show that modern LLMs can perform this restricted linguistic annotation tasks with very high agreement with a consensus gold standard. The **98.76% macro-F1 score of GPT-5 in a zero-shot setting** is particularly notable because it was obtained without task-specific fine-tuning or extensive prompt engineering. This section interprets these findings, analyzes the remaining error types, and discusses methodological implications.

6.1 The Efficacy of Prompting Strategies

A key finding of this study is the remarkable success of the **zero-shot strategy with GPT-5**. Contrary to the common assumption that more complex prompting yields better results, our experiments showed that providing in-context examples (few-shot) or forcing a step-by-step reasoning process (CoT) did not improve, and in some cases slightly hindered, performance compared to the simple zero-shot prompt. This suggests that for features with clear orthographic and morphosyntactic definitions, GPT-5's extensive pre-training has already encoded a deep, implicit understanding of these linguistic rules. In this context, additional examples may have constrained rather than helped the model, perhaps because the task itself was already sufficiently explicit.

Interestingly, **Gemini 2.5 Pro showed a slight but consistent benefit from few-shot prompting**, with its peak performance of 98.69% macro-F1 achieved with three or five examples. This divergence suggests that different model architectures may rely on in-context learning to different degrees for this type of task. The lack of a clear trend related to the temperature setting further indicates that for this classification task, model creativity was less important than its core linguistic knowledge.

6.2 Qualitative Error Analysis

While overall accuracy was high, a qualitative analysis of the errors provides invaluable insight into the current limitations of LLMs in understanding pragmatic and/or stylistic nuance. We identified three primary categories of recurring errors.

6.2.1 Convention vs. Emphasis (`Writing in Caps`)

The most challenging feature, with the lowest average F1-score of 96.05%, was `Writing in Caps`. The models were explicitly instructed to annotate this feature only when capital letters were used for emphasis. The majority of errors were false positives where the model correctly identified a capitalized word but failed to understand its conventional, non-emphatic function. In headlines such as, *Više ne želi da se krije, bez griže savesti! Toni Bijelić uslikan ispred hotela **FOTO*** (EN: *He no longer wants to hide, without a shred of guilt! Toni Bijelić photographed in front of the hotel **PHOTO***), the model incorrectly flagged the headline for using capital letters for emphasis. This pattern repeated consistently for journalistic conventions like FOTO (EN: PHOTO), VIDEO, and ANKETA (EN: POLL). This suggests that the models can detect orthographic salience more easily than pragmatic function, and may mistake genre-specific conventions for stylistic emphasis. A more explicit exclusion rule in the prompt would likely reduce this error type.

6.2.2 Failure of Discourse Scope: The "Quote Bleed" Problem

A significant number of errors for features involving person and mood (`First-Person Plural ['we'-form]`, `Second-Person ['you'-form]`, `Imperative`) stemmed from the model's inability to distinguish between the headline's narrative voice and the voice of a person being quoted, although the instructions clearly specified that these features should be annotated only when the pronoun or verb form directly reflects the headline's own address or stance. For example, in the headline *"Zorica Brunclik: Moj bivši muž ne zna kako mu izgledaju unuci, a Sari **smo** nedavno **proslavili** punoletstvo"* (EN: *Zorica Brunclik: My ex-husband doesn't even know what his grandchildren look like, and **we recently celebrated** Sara's coming of age.*), the model incorrectly labeled it as `First-Person Plural ('we'-form)`. It correctly identified the first-person plural verb form but failed to recognize that it was part of a direct quote and not an address from the headline's author. Similarly, in ***Pitajte** Tonija, meni je to svakako kompliment...* (EN: ***Ask** Toni, to me, that's a compliment anyway...*), the imperative "Pitajte" was misattributed to the headline's voice. This demonstrates a weakness in parsing nested discourse structures and correctly attributing linguistic features within their proper scope.

6.2.3 Formal Cues Overruling Pragmatic Intention (`Question`)

Several errors in the `Question` category occurred in headlines whose (elliptical) structure could be interpreted as interrogative because they contain a question word—even though they lack a question mark and serve the communicative function of announcing content. For instance, in the headline *Dvorac, luksuzni automobili, televizija... **Šta** Dragana Mirković dobija posle razvoda od Tonija* (EN: *A castle, luxury cars, a TV network... **What** Dragana Mirković gets after divorcing Toni*), the model flagged it as a question. Human annotators, on the other hand, did not interpret such cases as questions, as they considered the overall meaning of the headline and understood it as an announcement (in the sense of "Here's what Dragana Mirković gets…"). These results are unsurprising because the category sits at the boundary between formal marking and communicative intention. More explicit prompt wording could likely reduce this error type, but some ambiguity is inherent to the annotation scheme itself.

6.3 Practical and Methodological Implications

The study offers several practical takeaways for corpus linguistics, especially for work on underrepresented languages where annotation resources are limited.

First, the extremely high accuracy of the best model configurations suggests that LLMs can be reliably used for a **"first-pass" annotation of large corpora**, leaving human experts to focus their efforts on reviewing and correcting the small percentage of errors, particularly for more ambiguous features. This human-in-the-loop approach could drastically reduce the time and resources required for corpus creation.

Second, the **cost comparison between models** reveals that price does not necessarily correspond to performance. For instance, while GPT-5 achieved top-tier accuracy at a cost of **$3.64** per experiment, Gemini 2.5 Pro incurred a cost of **$5.75** per experiment (about **58% more expensive**) for comparable results. This finding highlights that the most costly model is not always the most effective, and that even simpler prompting strategies can yield optimal outcomes. Consequently, researchers should rely on empirical testing rather than assuming a direct correlation between model price, prompt complexity, and quality of output.

Finally, we compared the **costs of LLM annotation** with those of **human annotation**. Although the human annotation task in this study involved 13 categories—making direct cost equivalence difficult—the contrast remains striking. Assuming that a human annotator requires approximately **one minute per title**, and that annotation costs in Croatia range from **€13 (≈$15) to €25 (≈$29) per hour**, LLM-based annotation (using GPT-5) proves to be **35 to 58 times more cost-effective**, and over **four times faster**. Specifically, manual annotation would require roughly **8.5 hours**, whereas the best LLM model configuration completed the task in just **2 hours**. These estimates should be interpreted cautiously because the current study covers only 8 of the 13 original categories and because expert review remains necessary for ambiguous cases. A full cost comparison should therefore be revisited once the complete annotation scheme has been tested.

From a reproducibility perspective, the main lesson is that strong results on this task depend less on highly elaborate prompting than on clear category definitions, consistent output formatting, and transparent evaluation against a stable gold standard.

From an ethical and community perspective, LLM-assisted annotation should be understood as a way to support expert work on Serbian language resources (and other underrepresented languages), not replace it. Human oversight remains essential, particularly when categories depend on discourse interpretation, media conventions, or culturally specific cues.

7. Conclusion & Future Work

7.1 Conclusion

This study set out to evaluate the efficacy of state-of-the-art LLMs for the automated annotation of sensationalist language features in Serbian news headlines. Our findings show that both GPT-5 and Gemini 2.5 Pro can perform this task with very high agreement with a manually curated gold standard.

Our main contribution is an empirical benchmark for Serbian that compares models, prompting strategies, and costs under the same evaluation setup. We demonstrated that for the eight well-defined features under investigation, the most effective and cost-efficient method was **GPT-5 in a simple zero-shot configuration**, which achieved a top **macro-F1 score of 98.76%**. This counter-intuitive result suggests that for tasks with clear orthographic and morphosyntactic criteria, the models' extensive pre-training may be more reliable than limited in-context examples. Our detailed error analysis revealed that the models' few mistakes were not random but systematic, stemming from a difficulty in parsing pragmatic intent (e.g., journalistic conventions versus stylistic emphasis) and complex discourse structures (e.g., text within direct quotes). Overall, our work confirms that LLMs are a powerful and reliable tool for accelerating linguistic corpus annotation for Serbian and potentially other underrepresented languages, offering a practical pathway for researchers to scale their analyses, while also clarifying where expert review remains necessary.

7.2 Future Work

Building on the findings of this study, we propose several avenues for future research:

Expanding Feature Complexity: A logical next step is to apply our methodology to the more subjective and semantically complex features from the original annotation scheme, such as "expressive lexicon," "set phrases," and "information gaps". This would test the limits of the models' nuanced linguistic understanding.

Cross-Lingual and Cross-Genre Analysis: To test the generalizability of our findings, this methodology could be replicated on corpora from other (underrepresented) languages and diverse media genres (e.g., political news, social media, scientific articles).

Benchmarking Open-Source Models: Future comparisons should include newer proprietary models and strong open-source alternatives in order to establish a more comprehensive cost-performance benchmark.

Prompt and Guideline Refinement: Future work could mitigate the identified error patterns by adding explicit exclusion rules, counter-examples, and clearer instructions about quotation scope and genre-specific conventions.

Multimodal Extension: Since sensationalism is also expressed visually, future work should examine whether typography, font emphasis, image-headline pairing, or page placement improve the annotation of categories that are difficult to resolve from plain text alone.

8. Limitations

While our findings demonstrate the high potential of LLMs for linguistic annotation, the scope of this study has several limitations that should be considered when interpreting the results.

First, our analysis is constrained by its **data and genre scope**. The study utilizes a corpus of sensationalist headlines from a single source, the Serbian *Hello!* magazine, over a two-month period. The linguistic patterns and model performance observed may not generalize to other genres (e.g., political news, scientific articles) or other languages outside the South Slavic family.

Second, the **task complexity was intentionally limited**. We focused on eight features with clear orthographic or morphosyntactic definitions, which represent a subset of the 13 sensationalist language features identified in the source research. More subjective and semantically nuanced categories, such as "expressive lexicon" and "set phrases," were excluded and would likely present a greater challenge to automated annotation. Furthermore, our error analysis revealed that some annotation guidelines, like the distinction between conventional and emphatic capitalization, possess inherent ambiguity that creates a ceiling for model performance.

Third, the **methodological scope** was focused on two proprietary models. The results do not extend to the performance of open-source alternatives or to newer proprietary systems, which may offer different cost-performance profiles. Our exploration of prompting, while systematic, was not exhaustive, and other advanced prompt engineering techniques could potentially yield different outcomes.

Finally, although the gold standard is expert-based and consensus-derived, the paper does **not report a separate inter-annotator agreement statistic** for the subset used here. The results should therefore be interpreted as agreement with this gold standard rather than as a full substitute for independent expert annotation.

9. Acknowledgments

This research was supported by the European Union under the Next Generation EU program, as part of the project "Social Construction of Abstract Meanings Through Reading – SKAZ" (581-990-1036). Additional support was provided by the Faculty of Humanities and Social Sciences, University of Zagreb, through the institutional research project "Contemporary South Slavic languages: (socio)pragmatic analyses 2" (11-937-1034).

10. Bibliographical References

Akash, A. U., Fahmy, A., & Trabelsi, A. (2025). Can Large Language Models Address Open-Target Stance Detection? In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds), *Findings of the Association for Computational Linguistics: ACL 2025* (pp. 971–985). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-acl.54

Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Rahgouy, M., Raychawdhary, N., Marghitu, D., & Seals, C. (2025). Hate Speech Detection Using Large Language Models: A Comprehensive Review. *IEEE Access*, *13*, 20871–20892. https://doi.org/10.1109/ACCESS.2025.3532397

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 1877–1901.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Karlić, V., & Bartol, T. (2026). Sex, lies & intensifiers: a corpus-pragmatic analysis of sensationalist language features in magazine headlines. *13. međunarodni interdisciplinarni simpozijum "Susret kultura"*. Novi Sad: Filozofski fakultet. (in press)

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 22199–22213.

Li, M., & Conrad, F. (2024). *Advancing Annotation of Stance in Social Media Posts: A Comparative Analysis of Large Language Models and Crowd Sourcing* (No. arXiv:2406.07483). arXiv. https://doi.org/10.48550/arXiv.2406.07483

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11048–11064. https://doi.org/10.18653/v1/2022.emnlp-main.759

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., Karami, M., Li, J., Cheng, L., & Liu, H. (2024). Large Language Models for Data Annotation and Synthesis: A Survey. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

Processing (pp. 930–957). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.54

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). *Self-Consistency Improves Chain of Thought Reasoning in Language Models* (No. arXiv:2203.11171). arXiv. https://doi.org/10.48550/arXiv.2203.11171

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 24824–24837.

Wu, D., Zhang, J., & Huang, X. (2023). Chain of Thought Prompting Elicits Knowledge Augmentation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 6519–6534). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.408

11. Appendix

11.1 Appendix A: Prompting Strategies

All experimental runs used the **Base System Instruction** below. Specific strategies (Zero-Shot, Few-Shot, Chain-of-Thought) modified only the final block of the prompt.

11.1.1 Base System Instruction (Common to all)

```
##Goal
You are an expert linguist acting as a precise annotator. Your task is to analyze the provided news headline and, using the definitions below, determine the presence or absence of each of the 8 specified sensationalist linguistic features.

## Language Note
The headline is in Serbian (Latin alphabet). Your analysis must be based on the Serbian text. Your entire output must be in English.

## Feature Definitions
Analyze the headline for the presence of the following 8 features:
1. **Question:** The headline is grammatically a question and/or ends with a question mark.
2. **Exclamation Mark:** The headline contains one or more exclamation marks.
3. **Ellipsis:** The headline contains an ellipsis ('...').
4. **Second-Person ('you'-form):** The headline directly addresses the reader using the pronoun 'you' and/or a verb in the second person.
5. **First-Person Plural ('we'-form):** The headline uses the pronoun 'we' and/or a verb in the first person plural.
6. **Imperative:** The headline contains a command verb.
7. **Superlative:** The headline uses the superlative form or an adjective/adverb with the prefix 'pre-'.
8. **Writing in Caps:** The headline contains one or more words written entirely in capital letters for emphasis.

## Output Format
Your entire response must be a single, valid JSON object and nothing else. This object must contain exactly one top-level key: "annotation".
- The value for "annotation" must be a nested JSON object containing the 8 boolean features using **exactly** the strings "true" or "false" as values.
```

11.1.2 Strategy-Specific Variations

Strategy	Appended Instruction / Examples Block
ZS	`Now, analyze the following headline and provide the complete JSON object containing only the final annotation.`

FS	[Base Instruction] + N examples in the following format:
	`{`
	`"headline": "ČESTITAMO! Nevena Božović u osmom mesecu trudnoće",`
	`"annotation": {`
	`"Question": "false",`
	`"Exclamation Mark": "true",`
	`"Ellipsis": "false",`
	`"Second-Person ('you'-form)": "false",`
	`"First-Person Plural ('we'-form)": "true",`
	`"Imperative": "false",`
	`"Superlative": "false",`
	`"Writing in Caps": "true"`
	`}`
	`}`
	Note: FS1 used Example 1; FS3 used Examples 1–3; FS5 used Examples 1–5.

CoT	**Modified Output Format:**
	`## Output Format`
	`Your entire response must be a single, valid JSON object and nothing else. This object must contain exactly two top-level keys: "reasoning" and "annotation".`
	`- The value for the "reasoning" key must be a string containing your concise, bulleted-list thought process (keep it short).`
	`- The value for the "annotation" key must be a nested JSON object containing the 8 boolean features using **exactly** the strings "true" or "false" as values.`

Example provided:

```
<headline_example>
ČESTITAMO! Nevena Božović u osmom
mesecu trudnoće
</headline_example>

<json_example>
{{
  "reasoning": "Thought:\\n-
Question: Not grammatically a
question nor no question mark. ->
false\\n- Exclamation Mark:
Contains '!'. -> true\\n-
Ellipsis: No '...'. -> false\\n-
Second-Person ('you'-form): No
direct address. -> false\\n-
First-Person Plural ('we'-form):
Contains 'ČESTITAMO' (We
congratulate). -> true\\n-
Imperative: No command verb. ->
false\\n- Superlative: No
superlative form nor
adjective/adverb with prefix 'pre-
'. -> false\\n- Writing in Caps:
Contains 'ČESTITAMO'. -> true",
  "annotation": {{
    "Question": "false",
    "Exclamation Mark": "true",
    "Ellipsis": "false",
    "Second-Person ('you'-form)":
"false",
    "First-Person Plural ('we'-
form)": "true",
    "Imperative": "false",
    "Superlative": "false",
    "Writing in Caps": "true"
  }}
}}
</json_example>
```

11.1.3 Few-Shot Exemplars (Abbreviated)

Ex 1: "ČESTITAMO! Nevena Božović u osmom mesecu trudnoće" (Labels: Exclamation, 1st Plural, All Caps)

Ex 2: "ZAVIRITE U LUKSUZNI DOM MAJE ŠUPUT: Jedna od najprimamljivijih prostorija ženama je njen ormar" (Labels: 2nd Person, Imperative, Superlative, All Caps)

Ex 3: "'KAKO IZDRŽATI 7 DANA BEZ...?': Nikolina Pišek započela zanimljiv izazov" (Labels: Question, Ellipsis, All Caps)

Ex 4: "Sačekajte sa planiranjem godišnjeg odmora: najiščekivaniji festival otkrio je detalje programa pod zvezdama" (Labels: 2nd Person, Imperative, Superlative)

Ex 5: "Žestok okršaj u Đokovićevom boksu: Dadilja u Parizu bolje obučena od Jelene? ANKETA" (Labels: Question, All Caps)

11.2 Appendix B: Summary of Model Performance

Average precision, recall and macro F1-score per model across all 20 experiments.

Model	Configuration	Avg. Precision	Avg. Recall	Avg. Macro-F1
GPT-5	ZS	99.04%	98.65%	98.76%
	FS (n=1)	99.00%	98.55%	98.68%
	FS (n=3)	99.01%	98.62%	98.73%
	FS (n=5)	99.01%	98.53%	98.67%
	CoT	98.96%	98.55%	98.67%
Gemini 2.5 Pro	ZS (t=0.2)	98.87%	98.42%	98.55%
	ZS (t=0.5)	98.89%	98.43%	98.56%
	ZS (t=0.8)	98.90%	98.47%	98.59%
	FS (n=1, t=0.2)	98.98%	98.55%	98.67%
	FS (n=1, t=0.5)	98.92%	98.50%	98.62%
	FS (n=1, t=0.8)	98.86%	98.43%	98.55%
	FS (n=3, t=0.2)	98.90%	98.45%	98.58%
	FS (n=3, t=0.5)	98.98%	98.57%	98.69%
	FS (n=3, t=0.8)	98.89%	98.45%	98.57%
	FS (n=5, t=0.2)	98.02%	98.50%	98.62%
	FS (n=5, t=0.5)	98.99%	98.57%	98.69%
	FS (n=5, t=0.8)	98.97%	98.57%	98.69%
	CoT (t=0.2)	98.95%	98.50%	98.63%
	CoT (t=0.5)	98.96%	98.57%	98.68%
	CoT (t=0.8)	98.92%	98.48%	98.60%

11.3 Appendix C: Average Score per Feature

Average precision, recall, and F1-score per feature across all 20 experiments, ranked from highest to lowest F1-score.

Feature	Avg. Precision	Avg. Recall	Avg. F1-Score
Ellipsis	100.00%	100.00%	100.00%
Exclamation Mark	99.81%	99.80%	99.80%
Question	99.27%	99.22%	99.23%
Imperative	99.20%	99.02%	99.06%
Superlative	98.69%	98.45%	98.50%
First-Person Plural ('we'-form)	98.85%	98.25%	98.42%
Second-Person ('you'-form)	98.25	97.96%	98.03%
Writing in Caps	97.51%	95.44%	96.05%