LYS-UDC at PoliticHeadlinES-IberLEF 2026: Ranking Pipeline for Spanish Political News Headline Classification

Claudia García-López[1,*,†], Roi Santos-Ríos[†] and Jesús Vilares[†]

Universidade da Coruña, CITIC, Departamento de Ciencias de la Computación y Tecnologías de la Información, Campus de Elviña s/n, 15071, A Coruña, Spain

Abstract

This work presents the contribution of our group, the Language and Information Society (LYS) research group, to the headline ranking in Spanish political news (PoliticHeadlinES) shared task. Due to time and computational power limitations, we have opted for a rapid-prototyping approach with simplicity as our premise. In both ranking subtasks, text-only and multimodal, our proposed small-size LLM-based solution was articulated through a highly structured prompt design and the strategic use of regular expressions to guarantee the clean extraction of predictions. To control the output of the model in this new application context, inference was configured with a temperature of zero, and the model was assigned the role of a political expert. News contents were processed applying strict length limitations to minimize the Lost-in-the-Middle phenomenon, also applying 4-bit quantization in the case of the multimodal subtask. Our results show a highly competitive performance, particularly within the small-sized LLM category.

Keywords

Political news, Ranking, Large Language Models, Quantization

1. Introduction

PoliticHeadlinES shared task [1] within IberLEF 2026 [2] deals with the automatic assignment of headlines to news articles from a pool of candidates, taking the Spanish PoliSUM-2025 dataset as basis [3]. While Large Language Models (LLM) excel at natural language generation, integrating them into an automated evaluation pipeline for a ranking-based task like this presents significant technical barriers. Their conversational nature frequently introduces unnecessary text that corrupts strict output formats. Furthermore, managing long context windows when processing extensive news articles can trigger the *lost-in-the-middle* phenomenon [4], thus degrading the ability of the model to follow strict formatting instructions.

To mitigate these issues, this work proposes a robust zero-shot extraction methodology executed across two phases: an initial API-based exploration for rapid prototyping, and a final local implementation using open-source models to ensure data privacy. By combining strict prompt engineering with classic pattern matching and numerical fallback mechanisms, we ensure the reliable extraction of valid rankings, allowing for precise and automated metric calculations.

2. Methodology

Due to strict time limitations (we joined the task at the last minute), our proposal opted for a rapid prototyping approach where we limited ourselves to those resources immediately available, with simplicity as our premise. Our methodology circumvents the need for exhaustive, computationally expensive fine-tuning by adopting a zero-shot inference paradigm [5]. We model the *Learning-to-Rank* (LTR) task [6] as a conditioned text generation problem. The architecture of our solution is

IberLEF 2026, September 2026, León, Spain

*Corresponding author.

†These authors contributed equally.

claudia.garcia.lopez@udc.es (C. García-López); roi.santos.rios@udc.es (R. Santos-Ríos); jesus.vilares@udc.es (J. Vilares)

https://dunque.github.io/ (R. Santos-Ríos); https://www.grupolys.org/~jvilares/ (J. Vilares)

0009-0009-0541-1232 (R. Santos-Ríos); 0000-0003-2941-1834 (J. Vilares)

designed around strict prompt engineering, contextual constraint management, and a highly resilient post-processing extraction pipeline.

2.1. Prompt Engineering and Task Formulation

To force the language models to act as deterministic classifiers rather than conversational agents, the inference parameters were strictly controlled. Temperature was set to zero across all experiments to eliminate token sampling randomness, ensuring that the model consistently selects the highest-probability output. As shown in Appendix A, where we include an example, the prompt structure was meticulously designed in three blocks:

1. **Persona and Role-Playing:** The system prompt assigns the model the persona of an *"expert in Spanish politics and journalistic editing"*. This semantic framing biases the latent space of our model towards a formal, analytical register, improving its ability to discern subtle editorial nuances between the candidate headlines.
2. **Contextual Input:** The prompt injects the news body followed by the ten enumerated headline candidates. To mitigate the *lost-in-the-middle* phenomenon, where LLM fail to attend to instructions in long contexts [4], the news bodies were dynamically truncated. So, in the case of the text-only subtask, they were limited to their first 2,000 characters, while in the multimodal one, they were restricted to 1,500 characters to leave context window space for image tokens.
3. **Structural Constraints:** The prompt requires that the output must be formed exclusively by a comma-separated list of numbers enclosed in square brackets. It strictly forbids greetings, introductions, or internal thought tags.

2.2. Preliminary tests: API-based Exploration and Prototyping

Initial development and validation were performed using the Groq API[1] for ultra-fast inference, working on a fixed random sample of 1,000 news articles drawn from the training corpora. In the case of the multimodal subtask, a custom Python function encoded local image files into Base64 strings, injecting them into the API payload as an `image_url`.

Regarding the models evaluated, for the text-only subtask, we tested `llama-3.1-8b-instant`,[2] `qwen3-32b`,[3] and `llama-3.3-70b-versatile`.[4] For the multimodal task, the vision-capable `meta-llama/llama-4-scout-17b-16e-instruct` model[5] was evaluated on a 500-article subset.

2.3. Local Implementation and Quantization

In the next phase of our work, we transitioned to a fully local environment using the Ollama inference engine,[6] thus ensuring data privacy and eliminated external rate limits.

Regarding the models evaluated in this new environment, due to our computational power limitations, this time we deployed purely textual architectures, including `MistralSmall-24BInstruct-2501` [7] and `microsoft/phi-4`[8], alongside the `google/gemma-4-E4B-it` model.[9] Additionally, to process multimodal tasks locally without out-of-memory errors, 4-bit quantization was implemented for the Gemma model [7]. This reduced the memory footprint while preserving the cross-attention mechanisms necessary to synthesize text and visual cues.

The hardware used for these experiments consisted of a server with a triple setup of NVIDIA RTX 6000 Ada Generation featuring 48GB of VRAM each. Still, only one card was used per experiment.

[1] https://console.groq.com/ (visited on May 2026).
[2] https://console.groq.com/docs/model/llama-3.1-8b-instant (visited on May 2026).
[3] https://console.groq.com/docs/model/qwen/qwen3-32b (visited on May 2026).
[4] https://console.groq.com/docs/model/llama-3.3-70b-versatile (visited on May 2026).
[5] https://console.groq.com/docs/model/meta-llama/llama-4-scout-17b-16e-instruct (visited on May 2026).
[6] https://ollama.com (visited on May 2026).
[7] https://ollama.com/library/mistral-small:24b-instruct-2501-q4_K_M (visited on May 2026).
[8] https://ollama.com/library/phi4 (visited on May 2026).
[9] https://ollama.com/library/gemma4:e4b-it-q4_K_M (visited on May 2026).

API Configuration	Acc@1	Hit@3	Mean nDCG	Mean MRR
Text: `llama-3.1-8b-instant`	0.66	0.88	0.8371	0.7844
Text: `qwen3-32b`	0.81	0.98	0.9156	0.8900
Text: `llama-3.3-70b-versatile`	0.84	0.99	0.9385	0.9170
Vision: `llama-4-scout` (Text Only)	0.88	1.00	0.9531	0.9367
Vision: `llama-4-scout` (Multimodal)	0.88	1.00	0.9544	0.9383

Table 1
Results using the Groq API.

2.4. Robust Parsing and Fallback Mechanisms

Finally, to guarantee seamless integration with automated scoring systems, our pipeline implements a two-tier pattern-based extraction layer:

1. **Primary Extraction:** The script applies the regular expression `r'\[(.*?)\]'` to parse the raw text output, isolating the content exclusively contained within square brackets and stripping conversational noise.
2. **Heuristic Fallback:** If the model ignores bracket formatting, a secondary mechanism utilizes the pattern `r'\d+'` to extract every isolated digit. The logic filters these to the 1-10 range and deduplicates them to construct a valid ranking. Finally, target headlines are also isolated using the regular expression `r'(\d+)'` for automated metric calculation.

3. Results

The evaluation of our proposal was conducted by measuring the models' ability to correctly identify and rank the original editorial headline against nine semantically challenging distractors [1]. To do this, performance was quantified through a suite of standard ranking metrics:

- *Accuracy Top-1 (Acc@1)*, to measure the frequency of the correct headline appearing in the first position.
- *Binary Success within the Top-3 (Hit@3)*, to assess its presence within the top three candidates.
- *Mean Normalized Discounted Cumulative Gain (Mean nDCG)*, to evaluate the overall quality of the ranked list.
- *Mean Reciprocal Rank (Mean MRR)*, to determine the average rank of the target headline.

Our experimental findings are presented in two stages, corresponding to the initial API-based validation and the final official local implementation.

3.1. Preliminary tests: API-based Validation

The results obtained in this first development phase, previously described in Section 2.2, are detailed in Table 1. As can be seen, there is a clear positive correlation between model scale and ranking performance in the textual task. The Accuracy Top-1 improves significantly from 66.00% with the 8-billion parameter `llama-3.1-8b`, to 84.00% with the 70-billion parameter `llama-3.3-70b`. Notably, the vision-instructed model (`llama-4-scout`) demonstrated exceptional semantic reasoning, achieving the highest overall performance (88.00% Accuracy) even when restricted to a text-only configuration. The inclusion of multimodal data for this model yielded identical Accuracy and Hit@3 metrics with a negligible variation in Mean MRR and nDCG, indicating that, for this specific 500-article sample, the truncated textual features were already sufficiently discriminative to resolve the distractors.

Local Configuration	nDCG@10
Text: Mistral-Small-24B	78.15%
Text: Phi-4	78.15%
Text: **Gemma-4-it**	**82.41%**
Multimodal: **Gemma-4-it (4-bit)**	**80.82%**

Table 2
Results using the local implementation (including our official submissions to the task, in bold).

3.2. Local Implementation

These second set of results, shown in Table 2, corresponds to our local implementation, as described in Section 2.3. As can be seen, within the purely textual approaches, `Mistral-Small-24B` and `Phi-4` exhibited identical behavior across the main metrics, both achieving an nDCG@10 of 78.15%. In contrast, `Gemma-4-it` demonstrated superior textual discrimination, reaching 82.41% and establishing itself as our strongest setup overall. Notably, the multimodal implementation `Gemma-4-it`, despite incorporating visual context, yielded a slightly lower nDCG@10 of 80.82%.

This 1.59-point performance gap between the multimodal (80.82%) and text-only (82.41%) configurations of `Gemma-4-it` points to a concrete limitation of the deployment setup rather than an inherent weakness of multimodal reasoning. Specifically, the 4-bit quantization required to fit the vision-capable model within available memory introduces precision loss that appears to offset the informational gain provided by the visual signal. This interpretation is supported by the API-based exploratory results in Table 1, where `llama-4-scout` achieves virtually identical performance in text-only and multimodal settings on the same sample and without quantization, suggesting that visual context need not be detrimental when model precision is preserved. Future work should therefore prioritize higher-precision quantization schemes (e.g., 8-bit) or architectures with a lower memory footprint to isolate the true contribution of visual cues.

Comparing these local results against the API-based validation in Table 1, a direct comparison is not straightforward given that the API experiments were conducted on a fixed sample of 1,000 articles, whereas the local evaluation reflects the full official dataset. Nevertheless, a consistent trend is observed: performance scales with model size across both evaluation settings, and `Gemma-4-it` achieves a notably strong result relative to its size, outperforming both `Mistral-Small-24B` and `Phi-4` despite operating within the same hardware constraints.

3.3. Official Results

Our official submissions to the PoliticHeadlinES task correspond to those obtained with our local setup using `Gemma-4-it`. It must be noted that, despite being a rapid prototype subject to various constraints in terms of time and available computing resources, the solution presented by our team (LYS-UDC) achieved a competitive standing, ranking 9th (out of 21 participant teams) in the text-only task with a score of 0.824, and ranking 10th in the multimodal task with a score of 0.808.

Furthermore, it is worth noting that our system relies exclusively on locally deployed models under strict computational constraints (limited hardware, 4-bit quantization for the multimodal subtask, and a strict time budget) yet achieves an nDCG@10 of 0.824 in the text-only task and 0.808 in the multimodal task. This suggests that the pipeline design (particularly the prompt structure, truncation strategy, and fallback extraction mechanisms) compensates substantially for the reduction in raw model capacity, making zero-shot inference with small locally deployed LLMs a viable and competitive approach for this task.

3.4. Cross-Model Disagreement Analysis

To further characterize the behavior of our local pipeline, we analyzed the degree of agreement between the three text-only configurations at the top-1 prediction level. The three models (`Gemma-4-it`,

Pair	Mean τ	Median τ
Gemma-4-it vs. Mistral-Small-24B	0.516	0.556
Gemma-4-it vs. Phi-4	0.615	0.644
Mistral-Small-24B vs. Phi-4	0.519	0.556
Gemma-4-it text vs. Gemma-4-it multimodal	0.609	0.644

Table 3
Mean and median Kendall's tau between full rankings of each model pair.

`Mistral-Small-24B`, and `Phi-4`) agree on the same top-ranked headline in 72.3% of cases (3,552 out of 4,912 articles). Pairwise, `Gemma-4-it` and `Phi-4` show the highest agreement (82.8%), followed by `Mistral-Small-24B` and `Phi-4` (80.5%), while `Gemma-4-it` and `Mistral-Small-24B` agree least frequently (78.6%). In only 2.8% of articles do all three models predict a different top-1 headline, suggesting that the task presents a clear majority signal in most cases. Interestingly, in 8.2% of articles `Mistral-Small-24B` and `Phi-4` converge on the same prediction while `Gemma-4-it` diverges, which may reflect architectural differences in how the models handle the structured prompt format rather than genuine semantic disagreement.

Regarding the text-only versus multimodal comparison for `Gemma-4-it`, both configurations agree on the top-1 headline in 83.2% of cases. When they disagree (16.8% of articles), the text-only top-1 candidate is ranked second by the multimodal model in 12.5% of cases, and third or lower in only 4.3%. This indicates that the visual signal rarely causes drastic changes in the ranking. Instead, the small shifts it introduces appear to be, on balance, counterproductive under 4-bit quantization.

To complement this top-1 analysis, we computed Kendall's tau between the full rankings of each model pair, measuring agreement across all ten candidate positions. As shown in Table 3, the three text-only models exhibit moderate pairwise agreement, with mean tau values ranging from 0.516 (`Gemma-4-it` vs. `Mistral-Small-24B`) to 0.615 (`Gemma-4-it` vs. `Phi-4`). This indicates that, beyond the first position, the models diverge considerably in how they order the remaining candidates, suggesting that the task presents genuine ambiguity for positions 2–10. Notably, the agreement between `Gemma-4-it` in its text-only and multimodal configurations (tau = 0.609) is comparable to the inter-model agreement, which further supports the interpretation that 4-bit quantization introduces ranking noise of a similar magnitude to that caused by switching to a different model architecture.

4. Conclusions

The PoliticHeadlinES shared task has provided the research community with comprehensive insights into the capabilities of current state-of-the-art Large Language Models for structured ranking tasks. Regarding our participation, our dual-phase approach successfully established a robust methodology that allowed us to transition from rapid API-based prototyping to privacy-preserving local deployment without sacrificing performance.

Our experimental results confirm a significant correlation between model parameter count and the ability to resolve complex semantic distractors: while smaller models like `Llama 3.1 8B` provided a viable baseline, the 70B variant and specialized vision architectures like `Llama 4 Scout` exhibited superior performance, reaching up to 88.00% Top-1 accuracy. This suggests that the dense semantic representations found in larger models are critical for distinguishing the subtle editorial nuances characteristic of political journalism.

Technically, we have demonstrated that the inherent "chattiness" of generative models can be successfully managed through a combination of strict zero-temperature deterministic inference and a resilient pattern-based extraction layer. The implementation of a heuristic-based fallback mechanism proved to be a critical component for production-level stability, ensuring that even when models failed to follow complex formatting instructions, valid ranking data could be recovered.

Furthermore, the transition to a local environment using `Ollama` demonstrated that high performance ranking is achievable on moderate-power hardware through the use of quantization techniques. How-

ever, the results also revealed an important limitation: the 4-bit quantization applied to `Gemma-4-it` for the multimodal subtask introduced sufficient precision loss to offset the potential benefit provided by visual context, thus resulting in a slightly lower nDCG@10 (80.82%) compared to its full-precision textual counterpart (82.41%). This finding highlights that quantization is a critical factor when deploying multimodal models, and suggests that future work should explore higher-precision quantization schemes or more efficient architectures to fully exploit the complementary information provided by visual cues in political news classification.

5. Future Work

Several directions emerge naturally from the limitations identified in this work. First, the performance gap between the API-based exploration and the local implementation suggests that scaling up the locally deployed models (or relaxing the computational constraints through access to higher-end hardware) could yield substantial improvements without changes to the pipeline design itself.

Second, the multimodal results point to quantization as a key bottleneck. Future work should evaluate higher-precision schemes (e.g., 8-bit quantization) or architectures with a lower memory footprint that can process visual inputs without sacrificing model precision, in order to isolate the true contribution of visual context to the ranking task.

Third, the cross-model disagreement analysis reveals that the three local models diverge considerably beyond the first ranked position, with mean Kendall's tau values between 0.516 and 0.615. This opens the door to ensemble approaches that aggregate the rankings of multiple models, which could reduce individual model noise and improve overall nDCG@10.

Finally, the truncation strategy applied to mitigate the lost-in-the-middle phenomenon introduces its own information loss, particularly for articles where the disambiguating content appears late in the body. Future work could explore the use of automatic summarization techniques [8] rather than simply cutting at a fixed character limit.

Source code

The source code of our pipeline is freely available at https://github.com/claudiagarcialopez/PoliticHeadlinES_LYS-UDC, released under the GNU General Public License v3.0 (GPLv3).

Acknowledgments

We acknowledge grants LATCHING (PID2023-147129OB-C21), funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU; and CIDMEFEO, funded by the Spanish National Statistics Institute (INE); as well as funding by Xunta de Galicia (ED431C 2024/02), and Centro de Investigación de Galicia (CITIC). CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (ED431G 2023/01).

Declaration on Generative AI

During the preparation of this work, the authors used Gemini in order to: verify the correctness of the text translation and to assist with LaTeX coding. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Gómez-Navalón, T. Bernal-Beltrán, R. Pan, F. García-Sánchez, J. A. García-Díaz, R. Valencia-García, Overview of PoliticHeadlinES at IberLEF 2026: Multimodal Headline Ranking in Spanish Political News, Procesamiento del Lenguaje Natural 77 (2026).

[2] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.

[3] R. Pan, T. Bernal-Beltrán, J. Gómez-Navalón, J. A. García-Díaz, R. Valencia-García, Spanish PoliSUM-2025: Evaluating Stylistic and Editorial Fidelity in Abstractive Summarization of Political News, Procesamiento del Lenguaje Natural 76 (2026) 177–189.

[4] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the Middle: How Language Models Use Long Contexts, Transactions of the Association for Computational Linguistics 12 (2024) 157–173. URL: https://aclanthology.org/2024.tacl-1.9/. doi:10.1162/tacl_a_00638.

[5] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large Language Models are Zero-Shot Reasoners, in: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022), Curran Associates Inc., Red Hook, NY, USA, 2022, p. 1613. doi:10.52202/068431-1613.

[6] J. Lee, G. Bernier-Colborne, T. Maharaj, S. Vajjala, Methods, Applications, and Directions of Learning-to-Rank in NLP Research, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1900–1917. URL: https://aclanthology.org/2024.findings-naacl.123/. doi:10.18653/v1/2024.findings-naacl.123.

[7] S. Ma, H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, F. Wei, The Era of 1-Bit LLMs: All Large Language Models are in 1.58 Bits, CoRR abs/2402.17764 (2024). URL: https://doi.org/10.48550/arXiv.2402.17764. doi:10.48550/ARXIV.2402.17764. arXiv:2402.17764.

[8] Y. Zhang, H. Jin, D. Meng, J. Wang, J. Tan, A comprehensive survey on automatic text summarization with exploration of LLM-based methods, Neurocomputing 663 (2026) 131928. URL: https://www.sciencedirect.com/science/article/pii/S0925231225026001. doi:https://doi.org/10.1016/j.neucom.2025.131928.

A. A Prompt Example

```
Eres un experto en política española. Tu tarea es ordenar 10 titulares
del más probable al menos probable basándote en el CUERPO de la noticia y en la
    IMAGEN adjunta (si la hay).

INSTRUCCIÓN ESTRICTA: Responde EXCLUSIVAMENTE con una lista de números
separados por comas y encerrados entre corchetes. No añadas saludos,
ni introducciones, ni etiquetas de pensamiento.

Ejemplo de respuesta válida: [3, 5, 1, 10, 2, 4, 8, 7, 6, 9]
```

Listing 1: System prompt for the multimodal subtask.