Theron at PoliticHeadlinES-IberLEF 2026: In-Context Rule Learning with GPT-5.5 for Multimodal Spanish Political Headline Ranking

Le Thanh Tung[1], Le Nguyen Quynh Nhu[1] and Nguyen Thanh Tuan[1,*]

[1]University of Information Technology, Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam

Abstract

This paper describes the participation of team **Theron** in the PoliticHeadlinES shared task at IberLEF 2026, whose goal is to rank ten candidate headlines for a given Spanish political news article, where exactly one headline is correct and the remaining nine are strong, topically related distractors. We focus on **Task 2** (multimodal: text + image), which determines the official leaderboard. Our approach is based on *in-context rule learning*: we sample roughly 3,000 training instances and prompt a large language model to induce a compact set of explicit ranking rules that capture how the gold headline differs from distractors. These induced rules are then injected as a system prompt into a multimodal LLM (GPT-5.5), which receives the article body, the associated image, and the ten candidate headlines and produces the final ranking as a strict permutation of `t1–t10`. We compare this prompt-based approach against a focused suite of text-only baselines, including BM25, two multilingual cross-encoder rerankers (`BAAI/bge-reranker-v2-m3` and `jinaai/jina-reranker-v3`), zero-shot multilingual NLI based on `MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7`, and three additional text-only LLM rankers (GPT-4.1-mini, GPT-5.2, GPT-5.4-mini) sharing the same induced-rule prompt. The rule-grounded multimodal LLM substantially outperforms all text-only baselines and achieves **the first place** on the official Task 2 leaderboard. We release our inference code and prompts to facilitate reproducibility.

Keywords

Headline ranking, Multimodal NLP, Spanish political news, Large language models, In-context rule learning, Cross-encoder rerankers, IberLEF 2026

1. Introduction

In the modern news ecosystem, the headline is often the only piece of information a reader sees before deciding to click, share, or skip an article. Headlines compress an entire story into a short, attention-grabbing sentence and therefore play a central role in how political information circulates and how public opinion is shaped [1]. Beyond their journalistic function, automatically selecting an appropriate headline is also a relevant problem for Natural Language Processing (NLP) and Information Retrieval (IR), and is closely related to learning-to-rank in text [2].

The PoliticHeadlinES shared task at IberLEF 2026 [3, 4] formulates this problem as a constrained ranking problem on Spanish political news. For each article, ten candidate headlines are provided, of which exactly one is the original headline and the remaining nine are carefully selected distractors that share entities, topics, or stylistic patterns with the article. The shared task is organised into two tasks:

- **Task 1 (text-only)**: rank the candidate headlines using only the article body and the candidates.
- **Task 2 (multimodal)**: in addition to text, an image associated with the article is provided, and the official leaderboard is computed on this setting.

The presence of strong distractors makes the task substantially harder than headline classification or generation: many distractors are perfectly fluent, topically coherent, and could plausibly headline a

IberLEF 2026, September 2026, León, Spain

*Corresponding author

23521740@gm.uit.edu.vn (L. T. Tung); 23521123@gm.uit.edu.vn (L. N. Q. Nhu); 23521724@gm.uit.edu.vn (N. T. Tuan)

similar article. Identifying the correct headline therefore requires fine-grained semantic comparison between text, headline frames, named entities, and visual content.

In this paper we describe the system submitted by team **Theron**, which centres on a single design choice: instead of fine-tuning, we use a frontier multimodal LLM (`GPT-5.5`) and inject *induced ranking rules* into the system prompt. The rules are not hand-written but obtained automatically by sampling about 3,000 training instances and asking an LLM to summarise the regularities that distinguish the gold headline from the nine distractors. We then evaluate this approach against a focused set of text-only baselines (BM25, two multilingual cross-encoder rerankers, multilingual NLI, and three text-only LLM rankers — GPT-4.1-mini, GPT-5.2, and GPT-5.4-mini — all sharing the same induced-rule prompt). Our main findings are:

1. Pure lexical retrieval (BM25) [5] and even strong multilingual cross-encoders are clearly insufficient on this task: distractors share too much surface and topical content with the gold headline.
2. Zero-shot LLM ranking already outperforms all classical baselines, but is sensitive to the prompt [6]: providing induced, dataset-grounded rules gives a consistent and large improvement over generic "rank by relevance" prompts.
3. Adding the image via the multimodal endpoint of GPT-5.5 yields a further improvement on Task 2, especially on instances where two headlines share the same political topic but disagree on the depicted event, person, or scene.
4. Within the LLM family, the gain comes primarily from the induced rules rather than from any specific backbone, while capacity still matters: GPT-5.5 is the strongest text-only ranker and GPT-4.1-mini the weakest.

The resulting system ranks **1st** on the official Task 2 leaderboard.

2. Related Work

2.1. Headline Ranking and Selection

Headline-related research in NLP has traditionally focused on *generation* (producing a headline from an article) and *classification* tasks such as clickbait detection. In contrast, headline *ranking*, as introduced by PoliticHeadlinES, is closer to a learning-to-rank formulation [2] and dense passage retrieval, particularly because candidate headlines often exhibit very high lexical and topical overlap.

The PoliticES [7] family of shared tasks previously explored Spanish political language through ideology detection, whereas PoliticHeadlinES extends this direction toward multimodal headline selection. A closely related benchmark is the Spanish PoliSUM-2025 task [8], which evaluates editorial fidelity in abstractive summarisation of political news and similarly emphasizes preserving the journalistic framing of the original article.

2.2. Lexical Retrieval and Cross-encoder Reranking

BM25 [5] remains a strong lexical baseline for short-document ranking tasks [9]. Beyond classical bag-of-words approaches, modern transformer-based rerankers [10] have become standard for refining a small candidate pool.

Cross-encoder rerankers, including monoT5-style models [11] and the more recent BGE family of multilingual rerankers [12], jointly encode the query and candidate headline as a single sequence and produce sharper relevance estimates than bi-encoders, albeit at a higher computational cost. In this work, we adopt `BAAI/bge-reranker-v2-m3` [12] and `jinaai/jina-reranker-v3` [13] as zero-shot text-only baselines.

2.3. Zero-shot NLI as Ranking

A natural zero-shot baseline for headline selection is to cast article–headline pairs as premise–hypothesis pairs and use multilingual natural language inference (NLI) models trained on MultiNLI [14] and XNLI-style data [15] to estimate entailment probabilities. Recent multilingual NLI checkpoints based on mDeBERTa-v3 [16], such as the publicly released `MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7` [17], achieve strong zero-shot text classification across many languages. Although this formulation is intuitive, political headlines are often stylistically compressed or editorialised, meaning they do not strictly entail the article content in practice.

2.4. LLMs as Rankers

Large language models have recently demonstrated strong zero-shot ranking capabilities when candidate documents are provided in-context [18]. Modern frontier models such as GPT-5 additionally support multimodal inputs, making them particularly suitable for multimodal headline ranking scenarios in which accompanying article images can help disambiguate otherwise topically similar headlines.

Our work follows this direction while focusing specifically on *rule-grounded* prompting strategies derived from training data [19], rather than relying solely on manually engineered prompts.

3. Task and Dataset Descriptions

3.1. Task formulation

For each article a_i, the organisers provide:

- the full article body text(a_i);
- an associated image img(a_i) (used in Task 2);
- ten candidate headlines $H_i = \{h_i^1, \ldots, h_i^{10}\}$, of which exactly one is the gold headline and nine are distractors.

A system must output a complete permutation π_i of the tokens `t1–t10` representing the ranking of H_i from most to least likely to be the correct headline. The submission file `results.csv` contains the columns `id`, `task_1`, and `task_2`, each task storing the predicted ranking as a space-separated permutation. Per the official guidelines, if a system only addresses one task, the same ranking is reported for the other.

3.2. Dataset

The PoliticHeadlinES corpus consists of Spanish political news articles collected from online media. Each instance has a unique MD5 identifier, an associated illustrative image (used in Task 2), and exactly ten candidate headlines indexed `t1–t10`, where exactly one is the gold headline written by the publisher and the other nine are strong, topically related distractors. Distractors are deliberately chosen to share entities, events, or framing with the article, which prevents trivial keyword-based solutions and forces models to reason about subtle differences in central entities, granularity, and headline frame. Figure 1 shows a representative instance to illustrate the structure of the data and the difficulty of the task.

The corpus is released as two files: a labelled **train** set (`train.csv`) and an unlabelled **test** set (`test_public.csv`) whose gold rankings are evaluated only on the official CodaLab platform. Table 1 reports the size of each split. Recall that every instance contributes ten candidate headlines, so the number of (article, headline) pairs is ten times the number of instances.

We tokenise `article_body` and each candidate headline by whitespace and report the resulting word-count distributions. Headlines are short and remarkably uniform across splits: the median headline has 14 words, the 95th percentile is 21 words, and the longest headline has 50 words. Article bodies, in contrast, span almost four orders of magnitude. Table 2 summarises their distribution. The corpus

Table 1
Sizes of the train and test splits used in this work. Each instance contains 10 candidate headlines, so the number of (article, headline) pairs equals ten times the number of instances.

Split	# instances	# pairs	Use
train	16,030	160,300	labelled training pool
test	4,912	49,120	official unlabelled test

Table 2
Article body word-length distribution per split. The first block reports the percentage of instances in each length bucket; the second block reports summary statistics. Lengths are computed by whitespace tokenisation of `article_body`.

Bucket (words)	train	test
< 50	8.9%	0.0%
50–99	0.0%	0.0%
100–199	1.0%	0.9%
200–499	20.5%	20.9%
500–999	46.6%	48.5%
1,000–1,999	21.1%	27.1%
$\geq$ 2,000	1.9%	2.6%
median	686	765
mean	751	856
$p95$	1,525	1,677
max	14,257	8,217

is dominated by articles in the 200–2,000 word range ($\sim$88% of every split), with a long tail of very long articles (up to 14,257 words in train) and a non-trivial fraction ($\sim$ 9%) of very short stubs (< 50 words) in train, presumably caused by truncated scrapes or aggregator pages; the test set is visibly cleaner, with no sample shorter than 134 words and a slightly higher mean length than train. This skew motivates two design choices in our pipeline: (i) we do *not* truncate article bodies, so long-form context is preserved, and (ii) we keep the induced rule prompt short enough to leave a generous token budget for the article.

4. Methodology

Our system is centred on a multimodal LLM ranker driven by an *induced* system prompt. The pipeline proceeds in three stages: (i) iterative rule induction from the training data, illustrated in Figure 2; (ii) multimodal ranking with GPT-5.5; and (iii) post-processing and validation of the output ranking.

4.1. Stage 1: Iterative in-context rule induction

Stage 1 builds a textual *rule bank* R that captures, in natural language, the heuristics by which the gold headline can be discriminated from its nine distractors. Rather than extracting rules from a large batch of examples in a single shot, we grow R *incrementally, one training sample at a time*, through the closed feedback loop depicted in Figure 2. We initialise R_0 with a short seed (a one-line instruction to "rank the ten candidates by how likely each is to be the gold headline") and iterate over a curated subset of approximately 3,000 instances drawn from `train.csv`. The choice of 3,000 instances was determined empirically: we ran the induction loop with 1,000, 2,000, 3,000, 4,000, and 5,000 samples and observed that the rule bank produced at 3,000 samples was identical to those produced at 4,000 and 5,000, indicating that convergence had already been reached; we therefore use 3,000 as the stopping point to minimise API cost. At step t, given the current rule bank R_t, we perform:

Associated image
(`739f65dd. . . 1266d`)

ID: `9450d2e1. . . 22a8e`

Article body (excerpt): "Hasta hace unos años estábamos acostumbrados a que la selección que innovaba futbolísticamente era Holanda, pero ahora es España", señala Ramon Besa… "Nadie juega como juega España. Al oír las declaraciones de Tata Martino y otros seleccionadores, se están fijando cómo juega"… [Morata, Luis Enrique, Japón, Costa Rica–Alemania, Mundial].

Candidate headlines:

`t1`	Los seleccionadores de otros países se han dado cuenta: "Nadie juega como *Japón*"
`t2`	El Govern estudiará excepciones a la tasa turística…
`t3`	Díaz Ayuso, un golpe de mano en tres pasos
`t4`	Jorge Vilda tras la victoria de España: "El terremoto…"
`t5`	Los planes del carbón de Zapatero, un pozo sin fondo…
`t6`	Un concejal de Vox compara la bandera LGTBI+…
`t7`	Un partido sin etiquetas inquieta a los demócratas de EE UU
`t8`	Así será el sorteo de los grupos del Mundial de Qatar…
`t9`	Levante y Valladolid empatan en un partido intrépido
`t10`	**[GOLD]** Los seleccionadores de otros países se han dado cuenta: "Nadie juega como *España*"

Gold ranking (`y_true`): `t10 t1 t4 t8 t5 t3 t9 t2 t7 t6`

Figure 1: A representative training instance from PoliticHeadlinES. The gold headline (`t10`) and the hardest distractor (`t1`) share an identical frame and differ only in the central entity (*España* vs. *Japón*); other distractors mix same-topic (sports/World Cup) and off-topic (Spanish politics) headlines, illustrating why simple keyword-overlap signals are insufficient.

1. **Sample.** Present a single instance to the LLM as one input, consisting of the article body, the associated image, the ten labelled candidate headlines `t1–t10`, and the current rule bank R_t injected as the system prompt.
2. **LLM inference.** The LLM, conditioned on R_t, ranks the ten candidates and emits a predicted permutation $\hat{\pi}_t$ of `t1–t10`.
3. **Compare against ground truth.** We retrieve the gold permutation $\pi_t^\star$ from `y_true` and compute the rank position of the gold headline and the set of high-confidence wrong placements made by the model.
4. **LLM error analysis.** A second LLM call (the *analyst*) inspects the discrepancy between $\hat{\pi}_t$ and $\pi_t^\star$ together with the article, the candidates, and R_t. Its task is to attribute the mistake to a *missing* or *imprecise* rule and to propose a concrete edit: ADD, REFINE, or MERGE. New or refined rules must be (a) *generalisable*, i.e. not specific to one entity, topic, or article, and (b) *compatible* with the existing rules in R_t.
5. **Rule-bank update.** The proposed edit is applied to obtain R_{t+1}. A lightweight consistency check rejects edits that are trivially redundant or that contradict a rule already in the bank, in which case $R_{t+1} = R_t$.

Because R is updated after *every* sample, the LLM at step $t{+}1$ already benefits from the lesson learned at step t. Empirically the bank converges quickly: after the first few hundred samples the rate of accepted edits drops sharply and the remaining samples mostly reinforce existing rules rather than introduce new ones. We freeze the bank once the acceptance rate falls below a small threshold; the

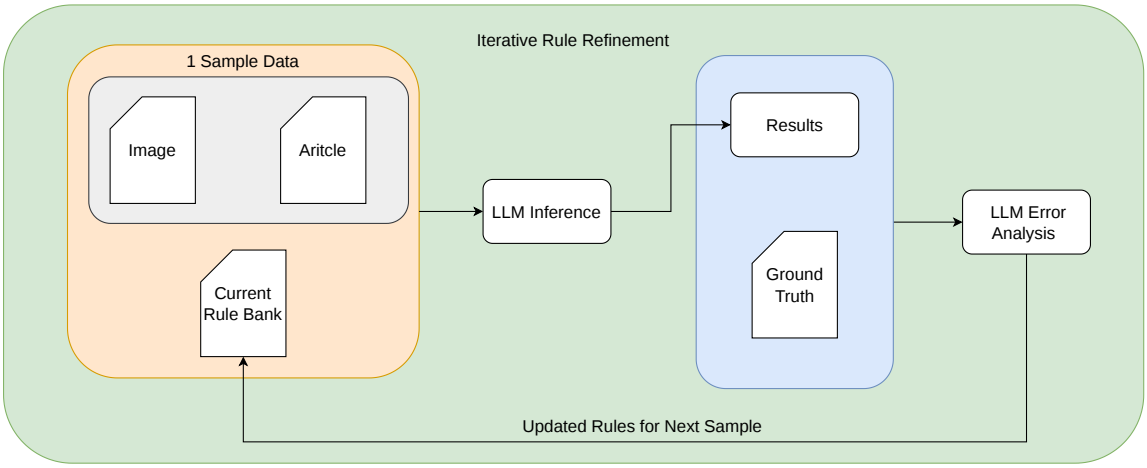

Figure 2: Iterative rule-refinement loop used in Stage 1. For each training instance, an LLM ranks the ten candidate headlines conditioned on the article body, the associated image, and the *current rule bank*. An error-analysis pass compares the predicted ranking against the ground-truth ranking and proposes an update to the rule bank, which is fed back into the next sample. The loop produces the compact natural-language rule set used as the system prompt in Stage 2.

resulting rule bank $R^\star$ is reproduced verbatim in the released code and, in summary, encodes nine rules covering:

1. closeness of wording, structure, and distinctive phrases;
2. tolerance for one-word noise if the headline frame is preserved;
3. priority on shared central entities (person, party, organisation, location, legal case, event, number);
4. event vs. topic vs. broad-field granularity;
5. resistance to single-keyword overlap;
6. using the image as a tie-breaker rather than a primary signal;
7. preference for natural, complete real-news headlines when all candidates are distant;
8. penalisation of truncated, generic, or article-type-mismatched headlines;
9. preservation of article type (news vs. analysis vs. sports vs. entertainment).

The frozen rule bank $R^\star$ is used as the system prompt for Stage 2.

4.2. Stage 2: Multimodal ranking with GPT-5.5

At inference time, for each test article we build a structured user message containing the cleaned article body and the ten candidate headlines labelled `t1–t10`. When available, the article image is loaded from disk, base64-encoded into a data URL, and attached as an `input_image` part with `detail=low`. The system prompt is the induced rule set from Stage 1.

We call the OpenAI Responses API with `model=gpt-5.5` and a strict JSON schema that forces the output to be of the form

```
{ "ranking": ["t...", "t...", ..., "t..."] }
```

with exactly ten unique items. Because GPT-5 family models do not honour a custom `temperature`, we rely on the deterministic decoding of the served model. Long article bodies are truncated only by collapsing whitespace; we do not perform topical or sentence-level pruning. Token budgets are bounded by `max_output_tokens=2000`.

4.3. Stage 3: Output validation and repair

Even with a strict JSON schema, the produced permutation must be checked: we (i) lowercase and normalise tokens (e.g., `T1` → `t1`, common mis-spellings of `t10`), (ii) detect near-permutations with one duplicate and one missing token and repair them by replacing the second occurrence of the duplicate, and (iii) raise an error and retry if the result is not a valid permutation of `t1`–`t10`. Up to three retries are performed per instance with exponential backoff. The final ranking is written to both `task_1` and `task_2` columns, since the same multimodal model is used for both tasks.

5. Experimental Setup

5.1. Data preprocessing

All systems share the same minimal text preprocessing pipeline applied to every instance before scoring. The `article_body` and each of the ten candidate headlines are decoded as UTF-8, normalised to lowercase, stripped of leading and trailing whitespace, and have all internal sequences of whitespace (spaces, tabs, newlines) collapsed into a single space. Common HTML artefacts that survive in a small fraction of scrapes (` `, stray `<br>` fragments, double-encoded quotes) are likewise removed. We deliberately do *not* apply any topical or sentence-level truncation, do not strip stop words, and do not lemmatise; preserving the full article body is important because gold headlines often reuse rare entities or distinctive phrases that appear only deep into the article. For Task 2, the article image is loaded from disk, re-encoded to JPEG, and embedded as a base64 data URL; we keep the image at its original aspect ratio and rely on the LLM's vision adapter for any further resizing.

All Hugging Face baselines (BM25, dense embeddings, cross-encoders, and the NLI ranker) are run with their default tokenisers and `fp16` inference on a single NVIDIA Tesla T4 GPU (16 GB). LLM-based rankers are called through their respective vendor APIs, so they incur no local GPU cost.

5.2. Baselines

To contextualise our multimodal LLM ranker, we implement a focused set of text-only baselines covering three classical families — lexical retrieval [9], cross-encoder reranking [10], and NLI-based ranking — plus three additional text-only LLM rankers. All systems produce a complete permutation of `t1`–`t10` per article and are scored with the official PoliticHeadlinES script.

5.2.1. Lexical baseline

As a sanity check we include **BM25** [5], the canonical lexical retrieval algorithm. For each article we tokenise the body and the ten candidate headlines with a Spanish-friendly whitespace tokeniser and rank the candidates by their BM25 score against the article. Scores are computed on the fly per article over its ten candidates, so no global corpus statistics leak between instances.

5.2.2. Cross-encoder rerankers

In contrast to bi-encoders, cross-encoders jointly tokenise the article and a candidate headline as a single sequence, which yields sharper relevance estimates at higher computational cost. We evaluate two multilingual zero-shot cross-encoders without any fine-tuning on PoliticHeadlinES:

- `BAAI/bge-reranker-v2-m3`, a widely used multilingual cross-encoder reranker that is competitive across languages including Spanish;
- `jinaai/jina-reranker-v3`, a recent multilingual reranker from the Jina family, included as a stronger and more recent point of comparison.

Both models follow the monoT5-style paradigm of casting reranking as a relevance-scoring problem [11]. For each article we score every (article, headline) pair independently and rank the ten candidates by descending relevance score.

5.2.3. NLI-based ranking

A natural zero-shot baseline for headline selection is to cast each (article, headline) pair as a (premise, hypothesis) pair and use a multilingual NLI model trained on MultiNLI [14] and XNLI-style data [15] to estimate the entailment probability $p(\text{ENTAIL} \mid \text{article}, \text{headline})$. Concretely, we use `MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7` [17], a multilingual mDeBERTa-v3 [16] checkpoint fine-tuned on a large mixture of NLI corpora and known to transfer well to Spanish in a zero-shot setting. The ten candidates are ranked in decreasing order of the entailment probability. This formulation is intuitive but brittle in our setting because political headlines are stylistically compressed or editorialised and rarely strictly entail the body, so NLI is expected to underperform when used in isolation.

5.2.4. LLM-based ranking (text-only)

We further evaluate four LLM rankers, all sharing the same input/output JSON-schema interface and all conditioned on the same induced-rule prompt $R^{\star}$ produced by Stage 1. The four backbones span a recent generation of OpenAI models that differ in capacity and modality support:

- **GPT-4.1-mini (text-only).** A compact, lower-cost model that serves as a sanity check on whether the induced rules already deliver competitive ranking on a smaller backbone.
- **GPT-5.2 (text-only).** An early frontier-class GPT-5 variant; included to study how rule conditioning interacts with model scale.
- **GPT-5.4-mini (text-only).** A later GPT-5 variant with improved reasoning, kept text-only to isolate the contribution of the visual modality.
- **GPT-5.5 (text + image).** Our final system. The same induced-rule prompt is used as the system prompt and the article image is additionally attached as an `input_image` part with `detail=low`. This is the only LLM ranker that consumes the image.

The image-only difference between GPT-5.4-mini (text) and GPT-5.5 (text + image) is by design and lets us read off the contribution of the visual modality directly. For all four LLM rankers we use the OpenAI Responses API, with a strict JSON schema enforcing a 10-element permutation of `t1`–`t10`, no custom temperature, and `max_output_tokens=2000`. Long-running jobs are sharded into N disjoint partitions and consolidated at the end, so that runs are resumable on errors or rate limits.

5.3. Evaluation

Submissions are evaluated independently for Task 1 and Task 2 by the official PoliticHeadlinES scoring script [3]. The metric reflects how high the gold headline appears in the predicted permutation: a system that places the gold candidate at rank 1 scores higher than one that places it at rank 5, irrespective of the relative order of the nine distractors. Higher is better. All numbers reported in this paper come from the official CodaLab platform, evaluated on the unlabelled `test_public.csv` split.

5.4. Results

Table 3 reports the official Task 2 leaderboard scores of all systems. Several clear patterns emerge.

First, the lexical baseline (BM25) [5] performs very poorly, scoring only 0.164. This was expected: distractors are deliberately constructed to share a large fraction of the article's vocabulary with the gold headline, so simple term-frequency overlap carries almost no discriminative signal. Zero-shot multilingual NLI is markedly stronger than BM25 (0.366) but still well below the rerankers, confirming that strict entailment is an imperfect proxy for headline appropriateness: editorialised, stylistically compressed headlines often do not literally entail the article body even when they are correct.

Second, the two multilingual cross-encoder rerankers form a substantially stronger baseline tier. `BAAI/bge-reranker-v2-m3` reaches 0.707 and `jinaai/jina-reranker-v3` reaches 0.738, more than four times the BM25 score and roughly twice the NLI score. This +0.031 gap between the

Table 3
Task 2 results on the official test leaderboard. Higher is better.

System	Official test
BM25	0.164
`BAAI/bge-reranker-v2-m3`	0.707
`jinaai/jina-reranker-v3`	0.738
Multilingual NLI (mDeBERTa-v3, XNLI-style)	0.366
GPT-4.1-mini (text-only, induced rules)	0.882
GPT-5.2 (text-only, induced rules)	0.918
GPT-5.4-mini (text-only, induced rules)	0.904
GPT-5.5 (text + image, induced rules)	**0.954**

two rerankers shows that more recent multilingual rerankers do help on Spanish political news, but it also suggests that purely retrieval-style supervision saturates well below the LLM regime: cross-encoders can localise lexical and semantic overlap but struggle with the fine-grained "which event/which person/which framing" decisions that dominate the hard cases of PoliticHeadlinES.

Third, the four induced-rule LLM rankers all outperform every non-LLM baseline by a wide margin. The smallest backbone, **GPT-4.1-mini**, already scores 0.882 ($+0.144$ over the strongest cross-encoder), which strongly suggests that the bulk of the improvement comes from the induced ranking rules rather than from raw model scale. Within the GPT-5 family, GPT-5.2 (0.918) slightly outperforms GPT-5.4-mini (0.904) — the latter being a smaller, distilled variant — indicating that, at fixed prompt, capacity still matters but is not strictly monotonic in version number.

Finally, our full system, **GPT-5.5 with text + image and induced rules**, achieves the best score of $\mathbf{0.954}$, an absolute improvement of $+0.036$ over the best text-only LLM (GPT-5.2) and $+0.216$ over the strongest non-LLM baseline (`jina-reranker-v3`). The gap between GPT-5.5 (multimodal) and the induced-rule text-only LLMs isolates the contribution of the visual modality on Task 2: visual cues are particularly helpful on cases where two headlines share the same political topic but disagree on the depicted event, person, or scene, where text alone is genuinely ambiguous. This is the configuration submitted to the official leaderboard, on which our team **Theron** ranks first.

To better understand the residual error, we inspect a small sample of failure cases of our final system. Errors typically fall into three categories: (i) the gold headline is itself ambiguous or noisy (e.g., contains formatting artefacts), and the model prefers a more natural distractor; (ii) two headlines describe the same event with almost identical wording, and the model picks the wrong one for stylistic reasons; (iii) the image is generic (a portrait, a logo, an empty hemicycle) and provides no disambiguating signal. Cases of type (iii) suggest that a more discriminative use of the image, e.g. via an explicit visual entity recognition step, could yield further gains.

6. Conclusion and Future Works

We described **Theron**'s submission to the PoliticHeadlinES shared task at IberLEF 2026. Our system uses a frontier multimodal LLM (`GPT-5.5`) prompted with a small set of induced ranking rules obtained from $3{,}000$ training examples, together with the article body, ten candidate headlines, and (for Task 2) the article image. The same model is used for both tasks, with the image only attached on Task 2. Despite its simplicity, the system substantially outperforms a focused suite of text-only baselines, including BM25, two multilingual cross-encoder rerankers (`BAAI/bge-reranker-v2-m3` and `jinaai/jina-reranker-v3`), an mDeBERTa-v3 multilingual NLI ranker, and three text-only LLM rankers (GPT-4.1-mini, GPT-5.2, GPT-5.4-mini) sharing the same induced-rule prompt, and ranks **first** on the official Task 2 leaderboard.

In future work we plan to explore: (i) explicit named-entity and event extraction over both the article and the image to feed the LLM with structured candidates; (ii) lightweight fine-tuning of multilingual

cross-encoders on the PoliticHeadlinES training data, since most public rerankers are not specialised for Spanish political news; and (iii) more principled use of the image, including visual entity linking via vision–language pre-training [20] and OCR for embedded text in news photos.

Acknowledgements

The authors thank the PoliticHeadlinES organisers for setting up the shared task and providing the evaluation infrastructure. This work was supported by the University of Information Technology, Vietnam National University Ho Chi Minh City.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] E. Ifantidou, Newspaper headlines, relevance and emotive effects, Journal of Pragmatics 218 (2023) 17–30.
[2] J. Lee, G. Bernier-Colborne, T. Maharaj, S. Vajjala, Methods, applications, and directions of learning-to-rank in NLP research, in: Findings of the Association for Computational Linguistics: NAACL 2024, 2024, pp. 1900–1917.
[3] 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).
[4] 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.
[5] S. Robertson, H. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends in Information Retrieval 3 (2009) 333–389.
[6] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, ACM Computing Surveys 55 (2023) 1–35.
[7] J. A. García-Díaz, S. M. Jiménez-Zafra, M. T. Martín-Valdivia, F. García-Sánchez, L. A. Ureña-López, R. Valencia-García, Overview of PoliticES at IberLEF 2023: Political ideology detection in Spanish texts, Procesamiento del Lenguaje Natural 71 (2023) 409–416.
[8] 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.
[9] T.-Y. Liu, Learning to Rank for Information Retrieval, Foundations and Trends in Information Retrieval 3 (2009) 225–331.
[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, pp. 4171–4186.
[11] R. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document Ranking with a Pretrained Sequence-to-Sequence Model, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 708–718.
[12] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 2318–2335.

[13] S. Sturua, I. Mohr, M. K. Akram, M. Günther, B. Wang, M. Krimmel, F. Wang, G. Mastrapas, A. Koukounas, N. Wang, H. Xiao, jina-embeddings-v3: Multilingual Embeddings with Task LoRA, arXiv preprint arXiv:2409.10173 (2024).

[14] A. Williams, N. Nangia, S. R. Bowman, A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2018, pp. 1112–1122.

[15] A. Conneau, R. Rinott, G. Lample, A. Williams, S. R. Bowman, H. Schwenk, V. Stoyanov, XNLI: Evaluating Cross-lingual Sentence Representations, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (2018).

[16] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations (ICLR), 2023.

[17] M. Laurer, W. van Atteveldt, A. Casas, K. Welbers, Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI, Political Analysis 32 (2024) 84–100.

[18] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020).

[19] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, Large language models are human-level prompt engineers, in: The Eleventh International Conference on Learning Representations, 2023.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, International Conference on Machine Learning (2021).