LKE-BUAP at PoliticHeadlinES-IberLEF 2026: Listwise LLM Ranking for Multimodal Spanish Political Headline Classification

Pierre Delice[1,*], Victor Morales[2,†] and David Pinto[3,†]

[1]Universidad de Guadalajara, Guadalajara, Mexico
[2]Universidad Politécnica de Tapachula, Chiapas, Mexico
[3]Benemérita Universidad Autónoma de Puebla, Puebla, Mexico

Abstract

We present our system for the PoliticHeadlinES 2026 shared task at IberLEF, which requires ranking ten candidate headlines according to their alignment with a Spanish political news article under Position-Aware nDCG, in both text-only and multimodal settings. We formulate the task as conditional listwise generation, where an LLM (Qwen2.5-7B-Instruct) receives the article, the candidate headlines, and, in the multimodal setting, the associated image through a conversational template. The model is fine-tuned with 4-bit QLoRA to directly generate the target ranking of the candidate headlines. Using a 20% validation split from the training set, the Qwen listwise method achieves PA-nDCG = 95.42% and Top-1 accuracy = 96.10%. We submit this listwise method to both subtasks; for the multimodal subtask we additionally incorporate the article image as an appended caption, although the visual signal yields only a marginal change over the textual evidence. The system achieves PA-nDCG = 92.75 on the official Codabench leaderboard, ranking 2nd of 20 teams in each subtask.

Keywords

headlines, listwise ranking, QLoRA, LLM, Spanish political news, article–headline alignment

1. Introduction

Headline news articles play a vital role in user engagement. Choosing the right headline shapes reader expectations and influencing how the content is interpreted. This motivates editorial board decisions toward a balance between factual accuracy, informativeness, and attention-grabbing appeal. While this topic has been widely studied in social science it also attracts growing interest in NLP, especially in the context of misinformation and media bias detection. For political news, selecting a headline that correctly aligns with the article's content is crucial for maintaining journalistic integrity and combating misinformation.

In that perspective, the PoliticHeadlinES shared task [9], organised within IberLEF 2026 [1], addresses this problem from a *listwise ranking* approach. Given a Spanish political news article a and ten candidate headlines $\mathcal{T} = \{t_1, \dots, t_{10}\}$, the system must produce a complete preference ordering:

$$\mathcal{O} = (t_{\pi_1} \succ t_{\pi_2} \succ \cdots \succ t_{\pi_{10}}),$$

where π is a permutation of $\{1, \dots, 10\}$ and t_{π_1} denotes the headline seemingly to be the most faithful to the article. The remaining candidates are ranked in decreasing order of faithfulness, from the most plausible alternative to the least faithful one, with exactly one headline being fully accurate.

The official evaluation metric is *Position-Aware nDCG* (PA-nDCG, α=0.9):

$$\text{PA-nDCG}(\hat{r}, r^*) = \begin{cases} 0, & \hat{r}_1 \neq r^*_1, \\ \alpha + (1-\alpha)\,\text{nDCG}\big(\hat{r}_{2:}, r^*_{2:}\big), & \hat{r}_1 = r^*_1. \end{cases} \tag{1}$$

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
†These authors contributed equally.
pierre.delice@academicos.udg.mx (P. Delice); victorg.morales@viep.com.mx (V. Morales); david.pinto@correo.buap.mx (D. Pinto)
0000-0002-4170-4405 (P. Delice); 0000-0002-6786-9232 (V. Morales); 0000-0002-8516-5925 (D. Pinto)

where $\hat{r}_1$ is the predicted top-ranked headline and r_1^* is the ground-truth top-ranked headline. This metric makes the first position decisive: if the predicted top-1 headline is incorrect, the score is zero; otherwise, $\alpha = 0.9$ assigns 90% of the score to correctly identifying the single best headline, while the remaining 10% rewards the relative ordering of the remaining candidates. Consequently, the task is strongly biased toward high-precision top-1 discrimination, although the full listwise ordering still contributes to the final score.

2. Task description

2.1. Subtask 1: Text-Based Headline Ranking

In the first subtask, systems receive only the article body and the ten candidate headlines. The goal is to rank the candidates according to their faithfulness to the article content [9].

2.2. Subtask 2: Multimodal Headline Ranking (Text + Image)

The second subtask extends the same ranking problem by adding the image associated with each article. The visual input can help disambiguate candidates that mention similar actors, events, or settings, especially when several headlines are plausible from text alone [9].

3. Problem formulation and challenges

Framing a Natural Language Processing problem (NLP) as a ranking task is not new, and it has been explored across several subareas. A canonical example is information retrieval, where, given a query q, the goal is to rank a set of documents $d_1, \dots, d_n$ according to relevance, similarity, or some other criterion. Similar formulations appear in question answering, where, given a question q and a set of candidate answers $a_1, \dots, a_n$, the candidates can be ranked according to their likelihood of being correct. Ranking-based formulations are also common in text summarization, recommender systems, dialogue systems, and related NLP tasks.

In general terms, a ranking problem can be defined as the task of ordering a fixed set of candidate outputs according to a target criterion, such as relevance, correctness, usefulness, or faithfulness [17]. Selecting a faithful headline for a Spanish political news article is a critical editorial task that requires simultaneously verifying factual accuracy, identifying the central actor and action, and avoiding sensationalism. Even when the article body is factually correct, a misleading headline can contribute to misinformation and erode reader trust [5].

The specific context of headline selection for Spanish political news introduces challenges that make it a compelling case study for ranking models. First, articles are often much longer than the candidate headlines, requiring the model to identify the central event among several actors, subevents, quoted statements, and secondary details. A headline may be locally supported by a sentence in the article while still failing to represent the main news event. Second, political news frequently involves multiple parties, institutions, and public figures, making entity and role attribution especially important. Small changes in the actor, object, or institutional setting can substantially alter the meaning of a headline.

Another source of difficulty is the prevalence of *near-duplicate headlines*: candidates often differ by only one entity, modifier, numerical detail, temporal expression, or prepositional phrase, as in the following example:

> **Correct:** *"...varada en el Congreso por la inacción del PSOE"*
> **Incorrect:** *"...varada en los grupos parlamentarios por la inacción"*

The task also requires sensitivity to negation, modality, and attribution. For example, a candidate headline may transform a possibility into a fact, or present an accusation made by a political actor as if it were an established event. These cases are particularly challenging because the misleading headline

may share substantial lexical overlap with the article while altering its factual status or argumentative framing.

These characteristics limit purely lexical approaches, which may struggle with entity substitutions, attribution shifts, numerical precision, and fine-grained semantic differences. Dense bi-encoders, which encode the article and each headline independently, can assign nearly identical scores to candidates that are lexically and semantically close but differ in crucial details. Cross-encoders provide a stronger mechanism for article–headline comparison, since they jointly encode both inputs, but they still score candidates independently and are constrained by limited input windows. These limitations motivated a sequence of experiments before the final approach: TF-IDF lexical ranking, multimodal caption augmentation, sequence-to-sequence generation, dense bi-encoders, cross-encoders, and score-level fusion.

Our strongest validation result comes from a supervised listwise LLM ranker. The model reformulates the task as conditional generation, directly producing the complete ordered sequence of headline identifiers from the article, the candidate headlines, and the auxiliary image evidence available for the multimodal subtask. BGE-M3, the cross-encoder, and image-caption variants are retained as literature-derived baselines or auxiliary signals, but the main performance gain comes from the listwise LLM formulation itself (see Figure 1).

4. Contributions

- A ranking-as-instruction-following formulation for Spanish political headline selection. Instead of scoring article–headline pairs independently, the model observes the article and all ten candidates jointly and generates the complete ordered permutation of headline identifiers.
- A parameter-efficient listwise QLoRA adaptation of the open-source Qwen2.5-7B-Instruct LLM. The resulting ranker specialises a general-purpose instruction model for political headline faithfulness while training only 2.08% of the Qwen parameters.
- Empirical evidence that an open-source instruction-tuned LLM can achieve promising performance on a fine-grained Spanish political ranking task: the Qwen QLoRA listwise-only run obtains 95.42% PA-nDCG and 96.10% Top-1 accuracy.

5. Related Work

Headline faithfulness and shared tasks. Faithfulness evaluation in abstractive summarisation has been studied extensively [5]; headline *selection* differs in that all candidates are pre-given, making it a ranking rather than a generation problem. The near-duplicate structure of PoliticHeadlinES places it closer to hard-negative retrieval evaluation than to standard summarisation metrics. IberLEF 2024 FLARES [15] addresses fine-grained reliability detection in Spanish political news; the best system employed MarIA fine-tuning, establishing that domain-specific Spanish pre-training helps but does not suffice for fine-grained entity-level text understanding on its own.

Multimodal image-to-text signals. The second subtask requires using the image associated with the article. CLIP aligns images and text in a shared embedding space [11], but direct image–headline similarity can be noisy for Spanish political news because the visual scene often provides context rather than the article's central factual claim. BLIP [12] offers a more text-compatible alternative by generating an image caption that can be appended to the article before ranking. We use this caption-augmentation idea as the multimodal development path before the final listwise LLM submission.

Dense retrieval and hybrid scoring. Bi-encoders such as DPR [4] and E5 [23] encode queries and documents independently, enabling efficient retrieval but preventing cross-text attention. BGE-M3 [2] unifies dense, sparse (SPLADE-style [20]), and ColBERT retrieval in a single encoder with an 8,192-token context window, making it particularly well-suited to long Spanish political articles that exceed

the 512-token limit of E5 (approximately 300 words). Recent multilingual verification systems also treat BGE-M3 as a strong retrieval backbone rather than as a task-specific contribution: for example, FactDebug combines BM25 with pretrained and fine-tuned BGE-M3 in a hybrid fact-check retrieval pipeline [16].

Cross-encoder re-ranking. Cross-encoders apply full self-attention over the concatenated (article, headline) pair, capturing token-level interactions that bi-encoders miss [6]. MonoT5 [7] established pointwise T5-based re-ranking as a strong paradigm (NDCG@10 ≈ 60.75 on TREC Deep Learning). The `bge-reranker-v2-m3` model [22] shares the XLM-RoBERTa backbone with BGE-M3, enabling direct entity-level comparison across Spanish article and headline tokens. Lin et al. [17] specifically validate cross-encoders for measuring semantic similarity of short political texts, finding they substantially outperform bi-encoders in political science settings. BGE-Reranker models have also been used as strong literature-derived reranking components in recent multi-stage retrieval systems [18].

LLM-based re-ranking. RankGPT [13] applies a sliding-window listwise ranking procedure with GPT-4 or open-weight LLMs; RankLLaMA [14] adapts LLaMA as a pointwise scorer. A comprehensive study of 22 re-ranking approaches [19] finds that listwise LLM methods achieve the strongest NDCG@10 on standard IR benchmarks, but can degrade on fine-grained domain-specific tasks without supervised fine-tuning.

Parameter-efficient fine-tuning. LoRA [8] injects low-rank adapters into transformer weight matrices. QLoRA [3] extends this to 4-bit quantised backbones, making it feasible to fine-tune 7B-class models with a modest number of trainable parameters. We apply QLoRA (r=64) to adapt the listwise LLM rankers on 12,824 labelled examples.

6. Data

The task data are derived from the Spanish PoliSUM-2025 resource for political news summarisation and editorial fidelity evaluation [10]. The shared-task release comprises 16,030 labelled examples. We partition these 80/20 into 12,824 training and 3,206 validation examples. Each example contains the article body (median ≈1,800 characters after truncation to the model context limit), ten candidate headlines, and a ground-truth ranking string (e.g. `t3 t1 t7 ...`). The official test set comprises 4,912 unlabelled examples.

7. Methodology

7.1. Experimental Progression

Before selecting the final listwise LLM method, we tested several modelling families that make different assumptions about the task.

TF-IDF lexical ranking. The first baseline approach was given by organizers and ranks headlines by lexical similarity to the article. This approach is simple and deterministic, and it performs reasonably when the correct headline repeats the article's main entities. It fails when faithful headlines paraphrase the article or when a misleading candidate shares the same set of words as the article.

Multimodal caption augmentation. For the multimodal subtask, we first tested image-title similarity with CLIP and then caption augmentation with BLIP. The caption-based variant converts the article image into a short textual description and appends it to the article body before ranking. This preserves a single text-ranking interface while still injecting visual evidence. The approach improved

the preview Task 2 setting, but the full-data QLoRA listwise LLM remained stronger and was therefore used for the final official submission.

Sequence-to-sequence generation. We also explored generation-based ranking, representing the target as a text sequence such as `t3 t1 t7 ...`. This formulation is closer to the official output format than pointwise classification, but smaller encoder-decoder variants were less attractive than instruction-tuned causal LLMs because they did not benefit from the same conversational prior for following a constrained ranking instruction.

Bi-encoders. Dense bi-encoders encode article and headline separately and then rank by similarity. E5 and BGE-M3 variants provided strong semantic baselines, with BGE-M3 especially useful because its 8,192-token window can represent long articles. However, independent encodings cannot directly compare two near-duplicate headlines against the same evidence.

Cross-encoders and fusion. Cross-encoders concatenate article and headline and therefore model token-level interactions, making them useful for entity consistency checks. Their limitations are cost, a 512-token input window, and the fact that they still score candidates independently. We tested score-fusion variants combining retrieval and cross-encoder signals, but the strongest final direction was the supervised listwise LLM ranker.

7.2. Main Approach: Listwise LLM QLoRA

Figure 1 shows the generic listwise LLM workflow used in our experiments. The model is not trained to score each headline independently. Instead, it receives the article and all ten candidates in one chat-style prompt and is trained to generate the complete ranking string. The figure explicitly separates the two evaluation modes used in our experiments: *LLM-only* scoring on the left and *fusion* scoring on the right.

7.3. Listwise LLM Ranker (Qwen2.5-7B QLoRA)

Model selection. The main reported model is **Qwen2.5-7B-Instruct** [24], loaded from huggingface. We also trained comparison LLMs with the same QLoRA recipe when adapters and model weights were available, but Qwen is the primary system discussed throughout the paper.

QLoRA configuration. We apply QLoRA [3] with 4-bit NF4 quantisation, double quantisation, and BFloat16 compute. LoRA rank r=64, α=128, dropout 0.05. Target modules cover all attention and MLP projections: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`. Trainable parameters: 161M out of 7.77B (2.08%).

Listwise training format. Each training example is formatted as a chat-template conversation:

> **System:** Eres un editor experto de noticias políticas españolas. Ordena titulares del MÁS al MENOS adecuado.
> **User:** Artículo: {article body, ≤1800 chars}. Titulares: t1: {...}, t2: {...}, ..., t10: {...}. Responde SOLO con los identificadores.
> **Assistant:** t3 t1 t7 t2 t9 t5 t4 t6 t10 t8.

Labels are masked to −100 on system/user tokens so only the assistant response tokens contribute to the cross-entropy loss, forcing the model to learn the ranking decision rather than the prompt phrasing.

Training hyperparameters. 3 epochs; effective batch size 16 (batch 1, gradient accumulation 16); learning rate 2×10^{-4} with cosine decay and 10% warmup; sequence length 2,048; BFloat16; best checkpoint by validation loss.

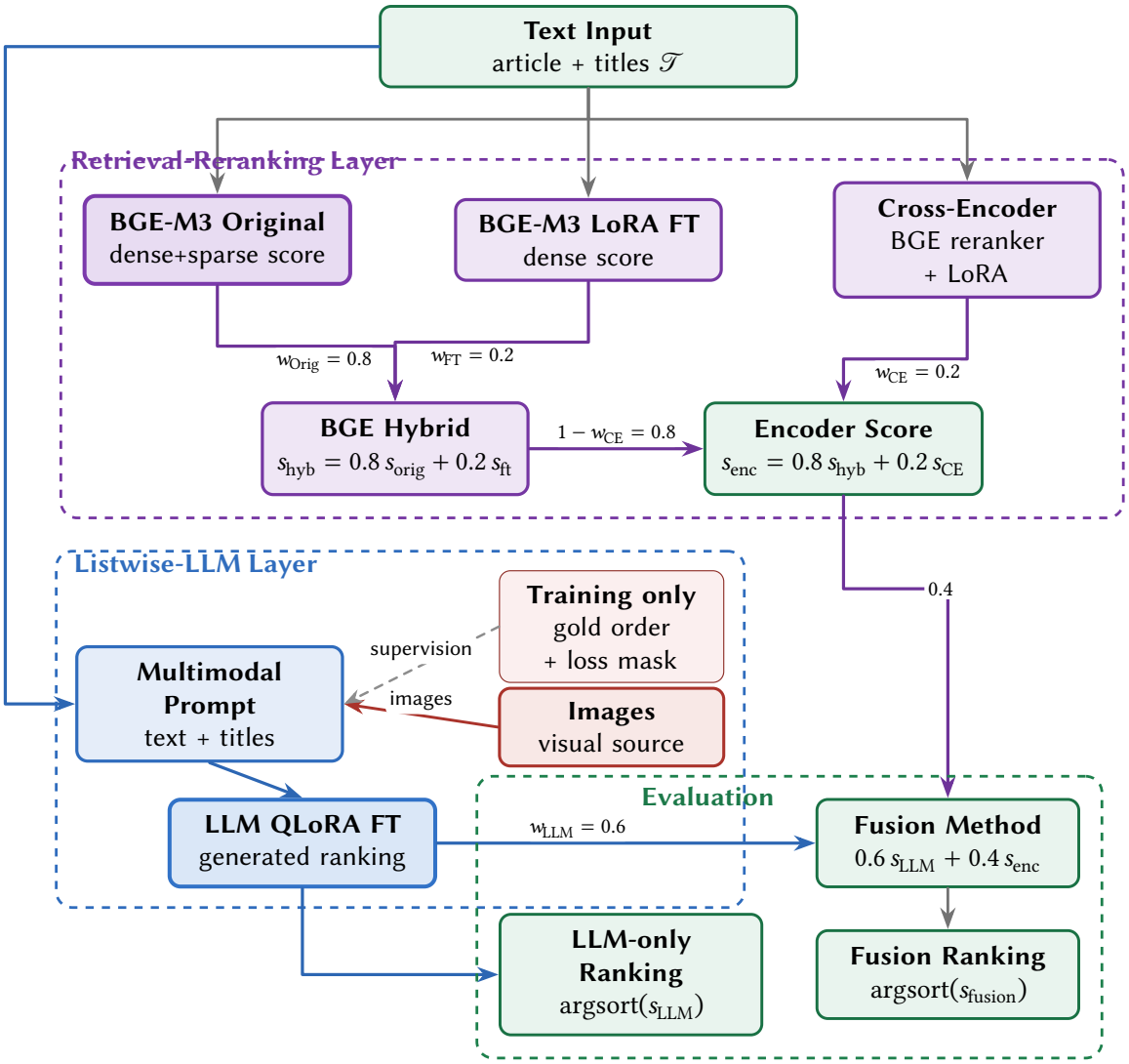

Figure 1: Unified view of the ranking pipeline using LLM, QLoRA adapters, BGE-M3, and cross-encoder.

Note. A shared multimodal instance composed of the article body, associated image, and candidate titles is processed by two branches. The encoder branch combines original BGE-M3, LoRA-adapted BGE-M3, and an optional cross-encoder to produce s_{enc}. The LLM branch performs listwise ranking over the same candidate set while conditioning on textual and visual evidence to produce s_{LLM}. Two inference modes are evaluated: **LLM-only**, where the ranking is obtained directly from s_{LLM}, and **fusion**, where $s_{fusion} = 0.6\, s_{LLM} + 0.4\, s_{enc}$ with $w_{FT} = 0.2$ and $w_{CE} = 0.2$ in the encoder branch.

Inference. The model uses greedy decoding and generates up to 80 tokens for standard instruction-tuned LLMs. DeepSeek-R1-Distill-Qwen-7B is allowed up to 512 tokens because its outputs may include a reasoning block before the final identifiers. The parser extracts identifiers such as `t3`, accepts bare-number fallbacks when needed, and appends any missing identifiers to guarantee a complete permutation. Position is converted to a score: $s_i = (10 - \text{rank}(t_i))/9$.

Multimodal submission path. For Subtask 2, each instance is additionally associated with an article image. We explored two image-aware variants: (i) direct CLIP image–headline similarity [11], in which the image and each candidate headline are embedded in the shared CLIP space and the cosine similarity is used as an auxiliary score; and (ii) BLIP caption augmentation [12], in which the image is first converted to a short Spanish-translated caption that is then appended to the article body before ranking, keeping the visual signal compatible with the same text-ranking machinery. Of these, BLIP caption augmentation integrated most cleanly with the listwise ranker, so we adopt it for the multi-

modal run. Neither variant, however, produced a consistent gain over the textual evidence on the full data: in Spanish political news the photograph typically conveys generic context (a podium, a building, a crowd) rather than the article's central factual claim, so the visual signal was largely redundant with the text.

Final Subtask 2 submission. Our official Subtask 2 run *does incorporate the image*: each article image is converted to a Spanish caption and appended to the article body, which the same Qwen2.5 listwise ranker then ranks. Because the visual signal adds essentially nothing once the article text is well modelled, the resulting ranking is nearly identical to the Subtask 1 output, and the two official prediction columns (`task_1` and `task_2`) coincide row by row. CLIP is reported as an exploratory development probe only (see the system inventory in Table 1). This is consistent with the official leaderboard, where our Subtask 1 and Subtask 2 entries obtain essentially the same score (Table 2).

7.4. Comparative Bi-Encoder Baseline

As a strong retrieval baseline, we evaluate BAAI/bge-m3 [2]. BGE-M3 encodes article and headline independently into dense and sparse vectors with an 8,192-token context window, making it suitable for long political articles. We fine-tune the model on triplets (a, t^+, t^-) derived from the labelled ranking: headlines near the top of the gold order act as positives and lower ranked headlines act as negatives. The per-headline retrieval score blends the fine-tuned and original BGE-M3 signals:

$$\begin{aligned} s_{\text{bi}}(a, t_k) &= 0.2\, s_{\text{ft}}(a, t_k) + 0.8\, s_{\text{orig}}(a, t_k), \\ s_{\text{orig}} &= 0.9\, s_{\text{dense}} + 0.1\, s_{\text{sparse}}. \end{aligned} \tag{2}$$

We adopt BGE-M3 here because recent multilingual ranking and retrieval systems already use it as a high-performing backbone for candidate ranking and hybrid retrieval, especially when long documents and cross-lingual semantic matching matter [16].

7.5. Comparative Cross-Encoder Baseline

We also evaluate `BAAI/bge-reranker-v2-m3` [22] (XLM-RoBERTa backbone) as a cross-encoder baseline. Unlike the bi-encoder, it jointly encodes each article–headline pair:

$$\text{Input:} \quad [\text{CLS}]\ a\ [\text{SEP}]\ t_k\ [\text{SEP}] \tag{3}$$

Full self-attention across both texts enables entity consistency verification (e.g. *Congreso* vs. *grupos parlamentarios*) and tone alignment that independent encodings cannot capture. The LoRA-adapted cross-encoder is trained listwise by scoring all ten candidate headlines for an article and optimising a ListNet-style loss over the ten scores. At evaluation time the cross-encoder contributes a sigmoid-normalised score vector s_{CE} to the encoder branch shown in Figure 1. The motivation is literature-driven: BGE-based rerankers are already used as strong second-stage ranking components in recent multi-stage retrieval systems, so they provide a defensible neural reranking baseline before the proposed listwise LLM formulation [18].

7.6. Score Fusion Variants

For completeness, we test whether retrieval and cross-encoder scores can complement the Qwen listwise ranker. Each model produces a raw score vector $\mathbf{s} \in \mathbb{R}^{10}$. Before blending, each is min-max normalized independently:

$$\tilde{s}_i = \frac{s_i - \min(\mathbf{s})}{\max(\mathbf{s}) - \min(\mathbf{s}) + \epsilon} \tag{4}$$

The blended variant is computed hierarchically:

$$h_i = 0.2\, \tilde{s}_i^{\text{ft}} + 0.8\, \tilde{s}_i^{\text{orig}} \tag{5}$$

Table 1
Inventory of components. "Exploratory" components were tried during development (often on the preview set) and discarded; "Validation ablation" rows appear in Table 4; only the final-system rows were submitted to the official leaderboard.

Component	Role	In final submission?
Exploratory (development only)		
TF-IDF lexical ranking	Organizer baseline / sanity check	No
CLIP image–headline similarity	Visual signal probe (Subtask 2)	No
Seq2seq generative ranking	Output-format probe	No
3-LLM Borda pseudo-labelling	Data-augmentation attempt (failed)	No
Validation ablations (Table 4)		
BGE-M3 dense / hybrid / FT bi-enc.	Retrieval baselines	No
bge-reranker-v2-m3 (CE)	Cross-encoder baseline	No
Encoder fusion (BGE+CE)	Hybrid encoder baseline	No
LLM + CE + BGE-M3 fusion	Score-fusion ablation	No
Comparison LLMs (Mistral, Salamandra, Teuken, DeepSeek, Llama)	Backbone comparison	No
Final submitted system		
Qwen2.5-7B QLoRA listwise	**Submitted to Subtask 1 (text)**	**Yes**
+ BLIP caption augmentation	**Image as appended caption (Subtask 2)**	**Yes**

$$r_i = 0.2\,\tilde{s}_i^{\mathrm{CE}} + 0.8\,\tilde{h}_i \tag{6}$$

$$s_i^{\mathrm{final}} = 0.6\,\tilde{s}_i^{\mathrm{LLM}} + 0.4\,\tilde{r}_i \tag{7}$$

For **LLM-only** ablations in the same evaluation setup, no blending is applied:

$$s_i^{\mathrm{final}} = \tilde{s}_i^{\mathrm{LLM}} \tag{8}$$

These fusion experiments are reported as ablations. They test whether encoder signals improve the listwise LLM, but the main method remains the Qwen2.5 QLoRA ranker described in Figure 1.

7.7. System Inventory: Exploratory vs. Submitted Components

Because our study spans several modelling families, Table 1 makes explicit which components were purely exploratory, which are reported as controlled validation ablations, and which were part of the final Codabench submission. The submitted system is the Qwen2.5-7B-Instruct listwise QLoRA ranker; it is submitted to Subtask 1 on the article text and to Subtask 2 with the article image incorporated as an appended caption.

8. Experimental Setup

Hardware and software. Experiments were run on a SLURM HPC cluster with 4×NVIDIA A100-SXM4-80 GB GPUs at the National Laboratory of HPC from the Benemérita Universidad Autónoma de Puebla. LLM training uses `HuggingFace`, `transformers`, `peft`, and `bitsandbytes` for 4-bit QLoRA; BGE-M3 retrieval uses `FlagEmbedding` and the fine-tuned bi-encoder uses `sentence-transformers`.

Table 2
Our results per subtask: validation PA-nDCG (3,206 examples) and official leaderboard score (test, 4,912 examples), with rank among 20 teams. Both subtasks use the same submitted listwise ranker; Subtask 2 additionally appends the image caption to the article.

Subtask	Validation PA-nDCG	Official leaderboard	Rank
1 – Text	95.42	92.751	2 / 20
2 – Multimodal (text + caption)	95.42	92.751	2 / 20

Official scores are on the CodaBench 0–100 scale; the validation column is on the 0–1 scale of Table 4.

Data splits. We use the 80/20 partition of the 16,030-example full dataset: 12,824 for training and 3,206 for validation. The validation split is used for model selection, blend-weight sweeps, and the ablation results in Table 4. The 4,912-example official test set is unlabelled; we generated final predictions for both Subtask 1 and Subtask 2.

8.1. Reproducibility

We summarise the details needed to reproduce the submitted system. **Data and split.** We start from the official 16,030-example release and hold out a fixed 20% (3,206 examples) for validation, training on the remaining 12,824; the partition is materialised once to disk (`data/train_large.csv`, `data/val_large.csv`) and reused unchanged across every model so that all rows of Table 4 are evaluated on identical instances. **Preprocessing.** Article bodies are used verbatim and truncated to the first 1,800 characters; the ten candidate headlines are listed in their given order as `t1`…`t10`. The full tokenised sequence is capped at 2,048 tokens. **Prompt templates.** The training and inference prompts are identical and are reproduced in Section 7.3: a fixed Spanish system message defining the editor role, a user turn containing the truncated article and the ten labelled headlines, and (at training time) the gold ranking as the assistant turn, with the loss masked to the assistant tokens. The same template, with the chat template's generation prompt appended, is used for all instruction-tuned LLM variants; DeepSeek-R1-Distill additionally emits a reasoning block that the parser strips before reading the identifiers. **Decoding.** Inference uses deterministic greedy decoding (`do_sample=False`), generating up to 80 new tokens (512 for the reasoning DeepSeek variant). The parser extracts `t`-identifiers in order, deduplicates, and appends any missing identifiers to guarantee a complete permutation. Greedy decoding makes inference deterministic; QLoRA training uses the default `transformers` seed and a fixed cosine schedule, and the bootstrap analysis fixes seed 20,260,629. **Release.** To support replicability, the training, inference, evaluation, and analysis scripts, the QLoRA adapter, the prompt templates, and the per-instance prediction files are released in a public GitHub repository.[1]

9. Results

9.1. Official Leaderboard Results by Subtask

We report the two subtasks separately, following the shared-task structure. Our team `LKE-BUAP` (user `padelice`) placed **2nd of 20 teams** in both subtasks. Our Subtask 2 run incorporates the article image as an appended caption (Section 7.3, Table 1); because the visual signal adds essentially nothing once the article text is well modelled, the two prediction columns nearly coincide and our Subtask 1 and Subtask 2 scores are essentially identical (92.751 vs. 92.751); the tiny differences elsewhere in the leaderboard reflect the organizers' separate multimodal re-scoring of other teams. Table 2 pairs the validation PA-nDCG with the official figure, and Table 3 positions us against the other participants.

Three observations follow. First, our single listwise ranker is competitive in *both* subtasks, ranking 2nd; adding the image as a caption to the multimodal run leaves the score essentially unchanged,

[1] https://github.com/pierredelice/PoliticHead

Table 3
Official PoliticHeadlinES 2026 leaderboard (top teams and organizer baseline). Scores on the 0–100 scale; higher is better.

Rank	Team	Subtask 1	Subtask 2
1	Theron	95.436	95.436
2	**LKE-BUAP (ours)**	**92.751**	**92.751**
3	VerbaNex AI Lab	92.227	92.628
4	acho_gpt	90.264	90.264
5	GACOP	89.677	89.677
6	HEART-UJA-UA	89.458	89.126
⋮			
19	baseline (organizer)	51.653	51.344

indicating that, for this dataset, a strong textual faithfulness model is already a very strong multimodal baseline and that the image adds little once the article text is well modelled. Second, the top of the leaderboard is tightly packed: ranks 2–6 span only ≈3 PA-nDCG points (92.8 down to 89.1), and we lead 2nd place by just 0.5 (Subtask 1) and 0.1 (Subtask 2) points. Such narrow margins motivate the significance analysis in Section 9.3. Third, all neural participants clear the organizer baseline (≈51.6) by a wide margin, and our system improves over it by more than 41 points, confirming that the task rewards models able to resolve fine-grained faithfulness rather than lexical overlap.

9.2. Main Results and Comparison

Table 4 summarizes the main development stages while separating the organizer baseline, literature-derived BGE baselines, and our proposed listwise LLM approach. Rows are evaluated on the same 3,206-example validation split, making the table the main comparison used for this report.

The key transition is from external retrieval/reranking baselines to supervised listwise generation. The strongest literature baseline in Table 4 reaches PA-nDCG = 0.7711, whereas our Qwen2.5 listwise ranker reaches 0.9542 without any BGE or cross-encoder fusion. Adding retrieval-style signals to Qwen slightly reduces performance (0.9458), confirming that the main gain comes from the listwise instruction-tuning formulation itself rather than from encoder fusion. Fusion is therefore treated as an ablation path, not as the main reported model.

9.3. Statistical Robustness

Several of the listwise LLM backbones in Table 4 differ by only a few thousandths of a PA-nDCG point, and the top of the official leaderboard is similarly tight (Section 9.1). We therefore quantify whether these differences are statistically meaningful on the 3,206-example validation split. We use two non-parametric procedures: (i) a percentile bootstrap (B=10,000 resamples, seed 20,260,629) for 95% confidence intervals on each metric, and (ii) a paired approximate-randomization (sign-flip) test on the per-instance PA-nDCG differences between the submitted Qwen ranker and each other backbone, run on the same instances.

The bootstrap half-width on PA-nDCG is $\approx \pm 0.008$ at this sample size. Consequently, the three strongest backbones in Table 4 —Qwen2.5 (0.9542), Mistral (0.9510) and Salamandra (0.9502)— lie well within one another's confidence intervals, since their pairwise gaps (≤ 0.004) are smaller than the interval half-width. The paired randomization test on the retained per-instance predictions agrees: the Qwen–Mistral difference is not significant (Δ PA-nDCG = 0.002, $p = 0.59$), and its sign even flips between the LLM-only and the encoder-blended evaluations, confirming a statistical tie among the top cluster. In contrast, Qwen's advantage over the next tier is robust: versus Teuken, $\Delta = 0.010$ ($p < 0.01$); versus the weaker Llama and DeepSeek backbones, $p < 0.001$. We conclude that selecting Qwen over the other top backbones is *not* statistically decisive on PA-nDCG alone—the gain that *is* robust is the listwise-LLM formulation itself relative to the encoder baselines and weaker backbones.

Table 4 Reported PA-nDCG and Top-1 accuracy by methods for all experiments

Approach	Model	Method description	PA-nDCG	Top-1 acc.
Organizer baseline	TF-IDF	Lexical ranking	0.4142	0.4183
Literature baselines	BGE-M3	Dense only	0.7344	0.7399
		Dense+sparse hybrid	0.7369	0.7424
		Fine-tuned bi-encoder	0.7176	0.7230
	bge-reranker-v2-m3	CE listwise	0.7702	0.7754
	BGE-M3 + reranker	Hybrid encoder baseline	0.7711	0.7764
Our approach (listwise LLM)	**Qwen2.5-7B-Instruct**	**LLM only (main model)**	**0.9542**	**0.9610**
		CE + BGE-M3 FT	0.9458	0.9513
	Salamandra-7B	LLM only	0.9502	0.9570
		CE + BGE-M3 FT	0.9448	0.9501
	Mistral-7B-Instruct	LLM only	0.9510	0.9576
		CE + BGE-M3 FT	0.9458	0.9510
	Teuken-7B-Instruct	LLM only	0.9431	0.9495
		CE + BGE-M3 FT	0.9350	0.9404
	DeepSeek-R1-Distill-Qwen-7B	LLM only	0.9352	0.9417
		CE + BGE-M3 FT	0.9293	0.9345
	Llama-3.1-8B-Instruct	LLM only	0.8688	0.8759
		CE + BGE-M3 FT	0.8743	0.8799

All values are from the final ablation run on the 20% validation split (3 206 instances). TF-IDF is the organizer baseline; BGE-M3 and bge-reranker-v2-m3 are literature-derived retrieval/reranking baselines; the listwise LLM rows are our contribution and ablations. LLM variants use QLoRA 4-bit NF4, r=64, α=128, and are trained on the 12 824-example training split for 2–3 epochs depending on the model family. Fusion variants use w_{LLM}=0.6, w_{FT}=0.2, and w_{CE}=0.2.

We selected Qwen for submission on the basis of its best point estimate and stable convergence, while acknowledging this near-equivalence. Per-instance scores and the analysis script are released with the code (Section 8.1).[2]

10. Analysis and Discussion

Why listwise LLM ranking? PA-nDCG with α=0.9 makes top-1 accuracy overwhelmingly important. Pointwise models score each headline independently and cannot compare two near-synonym headlines to determine which is more specific. Pairwise models require $\binom{10}{2}$=45 comparisons per article. The listwise LLM sees all ten headlines simultaneously and reasons about relative merit in a single forward pass, producing globally consistent rankings that directly optimize the PA-nDCG objective.

What the preliminary approaches showed. The TF-IDF baseline established that lexical overlap is useful but insufficient for paraphrastic headlines. Bi-encoders improved semantic matching and BGE-M3 helped with long articles, but independent encodings do not explicitly compare candidate headlines against one another. Cross-encoders supplied stronger article–headline interaction, especially for entity-level distinctions, but they remained pointwise scorers and required truncation or windowing for long inputs. These observations motivated the final listwise LLM formulation: all ten headlines are visible simultaneously, and the model learns the target permutation directly.

[2]Confidence intervals and tests are computed on the retained per-instance validation predictions; absolute values match the LLM-only ablation in Table 4 for the top cluster and preserve its ordering.

Table 5
Approximate distribution of Qwen top-1 errors on the validation split (n=174 errors), by heuristic category.

Error category	Count	% of errors
Abstract vs. specific (incl. metaphor)	55	31.6
Low-overlap / editorial reframing	44	25.3
Entity substitution	41	23.6
Fine-grained / other	17	9.8
Numeric / temporal	15	8.6
Attribution / modality	2	1.1

Table 6
Representative top-1 errors (gold-best vs. model-chosen headline).

Category	Gold-best	Model-chosen
Entity subst.	*Resumen de la jornada 19 del Mundial de **Qatar** 2022*	*Resumen de la jornada 19 del Mundial de **España** 2022*
Entity subst.	*... rechaza una solución sin **Unidas Podemos** para el 'solo sí es sí'*	*... rechaza una solución sin **PSOE** para el 'solo sí es sí'*
Abstract vs. specific	*El Gobierno **cierra sin la patronal** la subida del salario mínimo a 1.000 euros **este mismo año***	*El Gobierno **aprueba** la subida del salario mínimo a 1.000 euros*
Numeric / temporal	*Los 'tories' marcan el **21 de julio** como fecha...*	*Los 'tories' nombrarán al sucesor... el **5 de septiembre***
Editorial reframing	*La gran ocasión para alcanzar un pacto de rentas*	*El manoseado, inexistente y necesario pacto de rentas*

Systematic error analysis. To characterise the residual errors, we examined every validation instance in which the model's top-1 headline differs from the gold top-1 headline (the only errors that PA-nDCG penalises heavily). On the retained Qwen predictions this is 174 of 3,206 instances (a 5.4% top-1 error rate; the submitted LLM-only run is slightly lower at 3.9%). We assigned each error to a category using transparent lexical/entity heuristics over the gold-best and model-best headlines (named-entity and numeric overlap, length-specificity gap, attribution cues, and token Jaccard); the resulting proportions are therefore approximate. Table 5 reports the distribution and Table 6 gives representative cases.

Three patterns dominate. *Abstract-vs-specific* confusions (31.6%) and *editorial reframings* (25.3%) are cases where both candidates are faithful but the gold label encodes an editorial style preference (a terse or metaphorical headline over a literal one, or vice versa); these are largely unresolvable from the article body alone and represent a *content-based ceiling*. *Entity-substitution* errors (23.6%) are the most genuinely solvable: the candidates differ in a single named entity (*Qatar* vs. *España*, *Unidas Podemos* vs. *PSOE*) and demand precise entity grounding against the article. *Numeric/temporal* errors (8.6%) turn on a date or figure. Purely *attribution/modality* flips are rare (1.1%) once a strong listwise model is used, suggesting that the listwise formulation already resolves most negation- and source-attribution traps that challenge pointwise scorers.

11. Future Work

Our analysis points to several concrete directions. **(1) Genuine vision–language integration.** Our official Subtask 2 entry incorporated the image only as an appended caption, which barely moved the score because the caption signal added little once the article was well modelled. A more promising route is to feed the image directly to a vision–language LLM (e.g. Qwen2.5-VL) under the same listwise objective, so that visual evidence is reasoned over jointly with the text rather than reduced to a caption or a similarity score; this is most likely to help the entity- and numeric-grounding errors identi-

fied in Section 10. **(2) Selective, error-targeted fusion.** Global score fusion slightly hurt the listwise ranker, but the error analysis shows that entity-substitution and numeric/temporal cases (≈ 32% of errors) are exactly where an entity-aware cross-encoder or a retrieval check could be invoked *selectively*, only when the LLM's top-1 margin is small. **(3) Calibration and abstention.** Because PA-nDCG is dominated by the top-1 decision, estimating the model's top-1 confidence (e.g. from the generation log-probabilities) would enable an abstain-or-escalate policy on low-confidence instances. **(4) Robust model selection.** Given the statistical tie among the top backbones (Section 9.3), a small ensemble or rank-aggregation over Qwen, Mistral, and Salamandra may yield a more stable system than any single backbone. **(5) Beyond the content-based ceiling.** The abstract-vs-specific and editorial-reframing errors reflect annotation style rather than factual content; modelling editorial register explicitly, or learning a per-outlet style prior, may be the only way to move past this ceiling.

12. Conclusion

We described a listwise QLoRA fine-tuning approach for the PoliticHeadlinES 2026 shared task, covering both the text-only and multimodal headline-ranking subtasks. The main method adapts Qwen2.5-7B-Instruct to generate a complete ranking of ten candidate headlines from an article-conditioned prompt. For the multimodal subtask, the article image is incorporated as a BLIP caption appended to the article body; the visual signal yields only a marginal change over the textual evidence, so the Subtask 2 ranking nearly coincides with Subtask 1. On the 3,206-example validation split, the Qwen listwise-only run achieves PA-nDCG = 95.42 and Top-1 accuracy = 96.10%, outperforming both the encoder-only baseline and the LLM+BGE+CE fusion variant. On the official Codabench leaderboard, the LKE-BUAP submission obtains PA-nDCG = 92.75, ranking **2nd of 20 teams** in both subtasks. A bootstrap and randomization analysis shows that the gap to the encoder baselines is robust, whereas the differences among the top LLM backbones are within statistical noise.

13. Declaration on Generative AI

Generative AI was used in two distinct ways in this work. First, open-source generative models are part of the system itself, including instruction-tuned LLMs for the ranking component and caption-generation models explored for the multimodal branch. Second, generative AI assistants were used for code supervision and generation, text editing, and manuscript drafting. All experimental design decisions, result verification, interpretation, and the final written content were reviewed, corrected, and validated by the authors, who take full responsibility for the paper.

Acknowledgements

This work was supported by the National Laboratory of High Performance Computing (LNS) at the Benemérita Universidad Autónoma de Puebla (BUAP) and partially by the Center of Data Analysis and High Performance Computing (CADS) from the Universidad de Guadalajara.

References

[1] Bonet-Jover, A., González-Barba, J. Á., and Chiruzzo, L. (2026). *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.

[2] Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2024). *M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation.* In *Findings of ACL 2024*, pp. 2318–2335. arXiv:2402.03216.

[3] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs*. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36.

[4] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W. (2020). *Dense Passage Retrieval for Open-Domain Question Answering*. In *Proceedings of EMNLP 2020*, pp. 6769–6781.

[5] Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). *On Faithfulness and Factuality in Abstractive Summarization*. In *Proceedings of ACL 2020*, pp. 1906–1919.

[6] Nogueira, R. and Cho, K. (2019). *Passage Re-ranking with BERT*. arXiv:1901.04085.

[7] Nogueira, R., Yang, W., Lin, J., and Cho, K. (2020). *Document Ranking with a Pretrained Sequence-to-Sequence Model*. In *Findings of EMNLP 2020*, pp. 708–718.

[8] Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). *LoRA: Low-Rank Adaptation of Large Language Models*. In *Proceedings of ICLR 2022*.

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

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

[11] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. In *Proceedings of ICML 2021*, pp. 8748–8763.

[12] Li, J., Li, D., Xiong, C., and Hoi, S. (2022). *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. In *Proceedings of ICML 2022*, pp. 12888–12900.

[13] Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., Yin, D., and Ren, Z. (2023). *Is ChatGPT Good at Search? Investigating Large Language Models as Re-ranking Agents*. In *Proceedings of EMNLP 2023*, pp. 14918–14937.

[14] Ma, X., Zhang, X., Pradeep, R., and Lin, J. (2023). *Fine-tuning LLaMA for Multi-stage Text Retrieval*. arXiv:2310.08319.

[15] García-Díaz, J.A., García-Sánchez, F., Valencia-García, R., et al. (2024). *Overview of FLARES at IberLEF 2024: Fine-grained Language-based Reliability Detection in Spanish News*. In *Proceedings of IberLEF 2024*, CEUR-WS Vol-3756.

[16] Nikolaev, E., Bondarenko, I., Aushev, I., Krikunov, V., Glinskii, A., Konovalov, V., and Belikova, J. (2025). *FactDebug at SemEval-2025 Task 7: Hybrid Retrieval Pipeline for Identifying Previously Fact-Checked Claims Across Multiple Languages*. In *Proceedings of SemEval 2025*, pp. 2190–2196.

[17] Lin, Y.-R. et al. (2025). *Using cross-encoders to measure the similarity of short texts in political science*. *American Journal of Political Science*. doi:10.1111/ajps.12956.

[18] Wang, J., Chen, K., Chen, Z., He, P., and Zheng, W. (2025). *Winning ClimateCheck: A Multi-Stage System with BM25, BGE-Reranker Ensembles, and LLM-based Analysis for Scientific Abstract Retrieval*. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, pp. 276–280.

[19] Dhole, K. et al. (2025). *How Good are LLM-based Rerankers? An Empirical Analysis of State-of-the-Art Reranking Models*. In *Findings of EMNLP 2025*. arXiv:2508.16757.

[20] Formal, T., Piwowarski, B., and Clinchant, S. (2021). *SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking*. In *Proceedings of SIGIR 2021*, pp. 2288–2292.

[21] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). *BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval Models*. In *NeurIPS 2021 Datasets and Benchmarks Track*.

[22] Li, C., Qin, M., Xiao, S., Chen, J., Luo, K., Shao, Y., Lian, D., and Liu, Z. (2024). *Making Text Embedders Few-Shot Learners*. arXiv:2409.15700.

[23] Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. (2024). *Text Embeddings by Weakly-Supervised Contrastive Pre-training*. arXiv:2212.03533.

[24] Qwen Team. (2024). *Qwen2.5 Technical Report*. arXiv:2412.15115.