Testing LLMs on JOKER Task 3: Onomastic Wordplay Translation and Task 4: Humour Generation

UBO at CLEF 2026

Liana Ermakova, Yaël Naud, Lilian Binet and Baptiste Dumas

HCTI, University of Brest, France

Abstract

This paper presents the UBO participation in JOKER at CLEF 2026 for Task 3 on onomastic wordplay translation and Task 4 on humour generation. We evaluate prompt-based large language models (Gemini, Qwen, Llama, Gemma, and Mistral) without task-specific fine-tuning. For Task 3, we translate English named-entity wordplay into French using contextual descriptions of the source named entities. Gemini produced the highest number of genuine alternative translations among all systems. For Task 4, we generate puns from structured context and, for French, additionally explore the generation of contrepèteries (spoonerisms). Beyond the official evaluation, we provide an in-depth manual analysis of Gemini-generated puns and contrepèteries, highlighting the current capabilities and limitations of LLMs for multilingual computational humour.

Keywords

onomastic wordplay, machine translation, humour generation, Large Language Models, LLM

1. Introduction

Wordplay is one of the most challenging aspects of natural language processing, as it relies on lexical ambiguity, phonetic similarity, cultural knowledge, and language-specific linguistic phenomena. While recent large language models (LLMs) have demonstrated impressive performance on a wide range of multilingual language generation tasks, their ability to understand, translate, and generate humour remains limited, particularly when preserving or creating wordplay.

The JOKER Lab at CLEF 2026 provides a shared evaluation framework for multilingual computational humour, including tasks on humour retrieval, wordplay translation, and humour generation [1]. In this paper, we describe the participation of the UBO team in Task 3 on onomastic wordplay translation [2] and Task 4 on humour generation [3]. Both tasks investigate the ability of LLMs to produce creative outputs without task-specific fine-tuning, making them an ideal testbed for evaluating the current capabilities of foundation models. We did not submit runs for Task 1 on humor retrieval [4] and Task 2 on pun translation [5].

Our submissions rely exclusively on prompt engineering applied to several state-of-the-art LLMs, namely Gemini, Qwen, Llama, Gemma, and Mistral. For Task 3, we translate English named-entity wordplay into French by providing the models with contextual descriptions of the source named entities. For Task 4, we generate puns in English, French, and Spanish from structured contextual information and additionally investigate the generation of French *contrepèteries*, a language-specific form of spoonerism that remains particularly challenging for current LLMs.

Beyond the official evaluation, we perform a qualitative manual analysis of the French outputs produced by Gemini to better understand the strengths and limitations of LLMs in generating different forms of humour. Our analysis highlights recurring error patterns and sheds light on the remaining challenges of multilingual humour generation, particularly for linguistically constrained forms of wordplay.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

liana.ermakova@univ-brest.fr (L. Ermakova)

https://www.joker-project.com (L. Ermakova)

0000-0002-7598-7474 (L. Ermakova)

The remainder of this paper is organized as follows. Section 2 describes our methodology for Task 3 on onomastic wordplay translation and Task 4 on humour generation. Section 3 presents the official evaluation results together with a manual analysis of the French jokes generated by Gemini. Finally, Section 4 concludes the paper and discusses future research directions.

2. Methodology

This section describes the methodology adopted for the UBO submissions to JOKER at CLEF 2026. We first present our approach to Task 3 on onomastic wordplay translation, followed by our methodology for Task 4 on humour generation. For all subtasks, we evaluated the following commercial and open-source LLMs:

- gemini-3-flash-preview[1];
- Qwen3-4B-Instruct-2507 [6];
- Mistral-7B-Instruct-v0.3 [7];
- google/gemma-2-9b-it [8];
- Meta-Llama-3.1-8B-Instruct [9].

All open models were used via the *transformers*[2] Python library.

2.1. Task 3: Onomastic Wordplay Translation from English into French

JOKER Task 3 focuses on translating English onomastic wordplay (i.e., name-based wordplay) into French [2]. The benchmark comprises 353 training pairs and 2, 607 test instances covering various types of wordplay, including puns, portmanteaux, alliterations, and anagrams, collected from fictional universes such as literature, films, and video games. Each instance was accompanied by a description providing a short context for the named entity, which is often necessary to recognize, understand, and translate the underlying wordplay:

```
[
  {
    "id":"en_2",
    "en":"Obelix",
    "description":"Obélix is one of the main characters in the Asterix comics. He
    ↪ is a strong, large, and good-natured Gaul who, unlike other villagers,
    ↪ has permanent superhuman strength due to accidentally drinking a magic
    ↪ potion as a child. Obélix is best friends with Asterix, and together they
    ↪ embark on numerous adventures. Known for his love of wild boar, Obélix's
    ↪ loyalty, humor, and appetite make him a beloved character in the series",
    "fr":"Obélix"
  }
]
```

We submitted 5 runs to Task 3, one per model. We provided the models with the English named entities together with their corresponding descriptions using the following prompt:

```
f"Translate the following English named entity into French.
This is a localization task, not a literal translation. Use the description to
↪ infer the meaning, symbolism, personality, and any hidden wordplay in the
↪ original name.
```

[1] https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview
[2] https://huggingface.co/docs/transformers

```
If the name contains a pun, portmanteau, joke, double meaning, or sound-based
↪ wordplay, create a French name that produces a similar effect for a
↪ French-speaking audience, even if the wording differs significantly from the
↪ original.
Output only the final French name.
Named entity: \"{ENTITY}\"
Description: \"{DESCRIPTION}\""
```

2.2. Task 4: Humour Generation

JOKER Task 4 aims to generate a short, natural-sounding humorous text from a structured description of a pun, including the pun word, its two intended senses, and optional supporting keywords [3]. Below is an example of a training instance in French:


```
{
  "id": "fr_0001",
  "lang": "fr",
  "pun_word": "barbec-ours",
  "sense_a": "barbecue",
  "sense_b": "bear",
  "keywords": ["BBQ", "cookout", "grill", "grizzly", "brown bear", "cub",
  ↪ "..."],
  "hint": "barbec-ours / barbecue + ours ~ <côtelette; grizzly; cuit>",
  "reference_joke": "Comment on cuit des côtelette de grizzly ? Au barbec-ours.",
  "reference_en": "What do you call grizzly's ribs ? A bearbecue."
}
```


For Task 4, we generated puns in English, French, and Spanish, as well as French *contrepèteries*. We submitted 20 runs in total.

2.2.1. Pun Generation in English, Spanish, and French

For each of the three languages, we submitted five runs, each based on a different LLM backbone prompted to generate a pun from the provided context (`pun_word` / `sense_a` / `sense_b`), resulting in a total of 15 runs. We applied the following prompt:

```
f"You are an expert pun writer in {LANGUAGE}.
Given the context below, generate one original, witty pun in {LANGUAGE} inspired
↪ by the main topic, concepts, or vocabulary in the context.
Rules:
- Make the {LANGUAGE} pun natural and funny.
- Keep it to one sentence.
- Avoid generic jokes.
- Do not include quotation marks, introductions, explanations, or any text other
↪ than the pun itself.
Context: {CONTEXT}"
```

2.2.2. Contrepèterie Generation in French

For French, we additionally submitted five runs (one per LLM backbone) prompting the models to generate a *contrepèterie* from the given `pun_word`. A *contrepèterie* is a common form of French wordplay in which swapping sounds, letters, or syllables between words reveals a hidden humorous meaning, e.g. *Quel beau **p**lan de **f**oire ! — Quel beau **f**lan de **p**oire !* (*What a great fair plan! — What a beautiful pear flan!*)

We used the following prompt:

Table 1
Official results for Task 3: Onomastic wordplay translation. **Score** is the official score; **Refs, Alter,** and **Copy** denote the percentages of reference matches, valid alternative translations, and untranslated outputs, respectively. **#runs** is the number of runs submitted by each team.

#	team	#runs	Score	Rank	Refs	Alter	Copy	run_id
1	UBO	5	71.88	1	44.11	45.37	8.20	gemini-3-flash
7	UBO	5	17.25	43	7.08	10.54	24.83	llama
10	**Baseline**	1	13.10	46	13.66	0.00	100.00	identity_run
13	UBO	5	4.47	50	3.54	0.32	4.35	Qwen3-4B
15	UBO	5	2.24	55	2.42	1.60	6.31	mistral
16	UBO	5	1.92	56	1.35	0.64	0.15	gemma

```
f"Based on the provided context, generate **exactly one authentic
French contrepèterie**.
Return **only** the following three fields, in this exact format:
Contrepèterie: <original French sentence>
Sens 1: <literal, innocent meaning of the original sentence>
Sens 2: <hidden meaning revealed by the contrepèterie>
Requirements:
* The contrepèterie must be a genuine and natural-sounding French expression.
* The original sentence must be grammatically correct and idiomatic French.
* The hidden meaning must result from a valid contrepèterie (phoneme, syllable,
↪ or letter transposition commonly accepted in French wordplay).
* The two meanings must be clearly different and understandable.
* Prefer witty, playful, or mildly risky humor rather than vulgar, offensive,
↪ discriminatory, or insulting content.
* Do not explain the transposition mechanism.
* Do not add commentary, analysis, notes, introductions, or formatting
beyond the three required fields.
* Generate exactly one contrepèterie and nothing else.
Context: {CONTEXT}"
```

3. Results

3.1. Results of Task 3: Onomastic Wordplay Translation

System outputs were evaluated using exact matching against reference translations and complemented by expert human assessment to account for valid alternative translations [2]. Table 1 presents the official results for Task 3. The identity baseline, which simply copies the English proper names, achieves 13% accuracy against the references. Our runs based on direct prompting of open LLMs performed worse than this baseline in terms of matching references. Llama left 25% of source names untranslated, whereas Mistral, Qwen, and Gemma translated nearly all names but with limited success according to both reference matching and manual evaluation.

The Gemini-based submission was our best-performing run, achieving an accuracy nearly four times that of the identity baseline in terms of reference matching. According to both the percentage of successful alternative translations and the official ranking based on reference matches and successful alternative translations, Gemini outperformed all other systems. However, even the best submissions achieved less than a 50% success rate in terms of genuine translations different from the JOKER references.

Table 2
CLEF 2026 Task 4: Official LLM Arena Results for English (top), Spanish (middle), and French (bottom). Bradley-Terry leaderboard from the 3-judge LLM ensemble (Groq `openai/gpt-oss-20b` + `deepseek-chat` + `models/gemini-2.5-flash-lite`, `majority` aggregation) with 100-round bootstrap 95% CIs. n_b counts head-to-head battles per run on the sparse Arena sample.

	Rank	Run	BT	CI low	CI high	n_b
English	6	`gemini-3_pun`	1,125	1,084	1,168	326
	71	`qwen_pun`	995	953	1,039	299
	87	`mistral_pun`	903	869	937	343
	89	`gemma_pun`	895	851	929	360
	92	`llama_pun`	816	765	858	316
Spanish	4	`gemini-3_pun`	1,351	1,310	1,396	363
	19	`gemma_pun`	1,125	1,084	1,172	322
	38	`qwen_pun`	1,100	1,072	1,144	366
	93	`llama_pun`	1,019	984	1,063	360
	106	`mistral_pun`	987	945	1,025	332
French	7	`gemini-3-flash_pun`	1,279	1,234	1,341	362
	50	`qwen-3_pun`	1,130	1,092	1,165	410
	55	`llama-3_pun`	1,036	989	1,077	342
	59	`gemma-3_pun`	1,022	973	1,062	350
	60	`mistral-3_pun`	1,021	982	1,052	325
	117	`mistral_contrepetrie`	682	630	719	378
	119	`llama_contrepetrie`	662	613	703	348
	120	`gemini-3-flash_contrepetrie`	655	601	706	331
	121	`qwen-3_contrepetrie`	644	577	694	356
	123	`gemma_contrepetrie`	618	564	661	355

3.2. Results of Task 4: Humour Generation

Task 4 submissions were ranked using the JOKER Lab Arena [3]. For each test instance, pairs of generated outputs were compared by an ensemble of LLM judges in a pairwise evaluation protocol. The pairwise preferences were then aggregated into a Bradley-Terry leaderboard with bootstrap confidence intervals, computed independently for each language. Table 2 official results of our runs for English, Spanish, and French subtasks, respectively [3].

Across all three languages, the Gemini-generated puns ranked among the top-performing systems. In contrast, puns generated by the open-source models generally received lower rankings. For French, all *contrepèterie* submissions were ranked near the bottom of the leaderboard, with the Gemini-generated *contrepèteries* performing worse than the pun-generation submissions of all our models. A possible explanation is that *contrepèteries* are a French-specific form of wordplay whose recognition and interpretation may be challenging for LLM-based judges. Consequently, the automatic evaluation may underestimate the quality of valid *contrepèteries* compared with more conventional forms of humour.

3.3. Manual Analysis of Generated Wordplay in French

Two annotators, Master's students in translation at the University of Brest, French native speakers, manually analyzed the French Gemini outputs for both pun and *contrepèterie* generation. Each instance was evaluated by a single annotator. The annotators were provided with the `pun_word`, `sense_a`, `sense_b`, and the corresponding generated wordplay. For *contrepèteries*, they were additionally given the two intended senses as explained by Gemini. For each instance, the annotators answered the following questions:

- Is it a wordplay? (Yes/No/Other)
- Is it a pun? (Yes/No/Other)

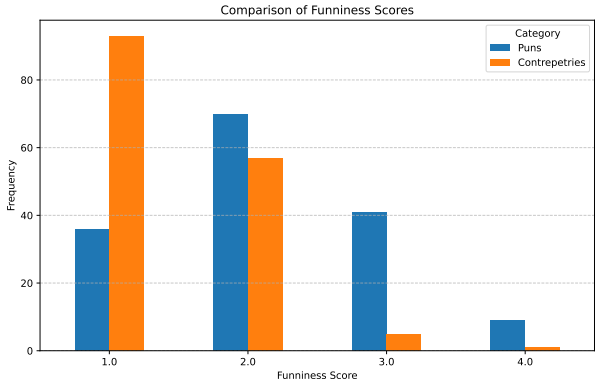

Figure 1: Comparison of the funniness scores of puns and contrepèteries generated by Gemini

Table 3
Distribution of Humour Types Produced by Gemini in the Pun Generation Task

Type	Wordplay	Pun	Contrepèterie	Offensive
Yes	152	126	—	1
No	1	28	156	154
Other	3	2	—	1

Table 4
Distribution of the Relevance of Gemini-Generated Puns to the Provided Context

Relevance	pun_word	sense_a	sense_b
Yes	119	134	124
Partially	29	9	6
No	8	13	15

- Is it a contrepèterie? (Yes/No/Other)
- Is it offensive? (Yes/No/Other)
- Is the text related to `pun_word`? (Yes/No/Partially/Other)
- Is the text related to `sense_a`? (Yes/No/Partially/Other)
- Is the text related to `sense_b`? (Yes/No/Partially/Other)
- Rate the funniness on the Likert scale from 1 (not funny) to 5 (very funny).

They could also add a comment in a free form.

Figure 1 compares the funniness scores of the puns and *contrepèteries* generated by Gemini. Overall, the generated jokes were generally perceived as not funny, with *contrepèteries* receiving particularly low ratings: approximately half were assigned the lowest possible funniness score.

3.3.1. Analysis of the Generated Puns

We manually analyzed 156 puns generated by Gemini. Among them, almost all (152 instances) were considered to be wordplay and 126 were classified as puns (see Table 3). The cases classified as "Other" were close to being wordplay but did not fully meet the criteria. Almost all generated wordplay were considered to be non-offensive. The majority of the generated jokes were considered to be relevant to the given context (see Table 4).

One limitation of the AI-generated puns is that they are often overly verbose or include unnecessary details, which can reduce their humorous effect. For example, with the *sense_a* “alaitchante” and *sense_b* “alléchante”, Gemini generated the joke *“La carrière d’une vache soprano est forcément alaitchante”*. In this example, describing the cow as a soprano does not add to the humour and seems unrelated to the

Table 5
Distribution of Humour Types Produced by Gemini in the Contrepètries Generation Task

Type	Wordplay	Pun	Contrepèterie	Offensive
Yes	87	22	79	37
No	67	133	75	116
Other	2	1	2	3

Table 6
Distribution of the Relevance of Gemini-Generated Contrepètries to the Provided Context

Relevance	pun_word	sense_a	sense_b
Yes	39	42	32
Partially	7	14	9
No	110	100	105

wordplay. It can also be seen in the contrepèterie *"Serge-Jean a le goût du blanc."* The fact that the man is named "Serge-Jean" does not bring anything humorous. Sometimes, the joke does not work because the generated text is not idiomatic. In *"Difficile pour un moteur de tourner rond quand sa cylindrée est restée bloquée au cube"*, the intended worpdlay "tourner en rond" and "cube" is recognizable. However, the expression "rester au cube" has no established meaning in French, making the punchline difficult to interpret.

3.3.2. Analysis of the Generated Contrepèterie

We manually analyzed 156 puns generated by Gemini. The statistics is given in Table 5. Among the generated text, 55% were considered to be wordplay with half of texts judged as contrepèteries. The majority of the generated texts were not judged offensive. The cases classified as "Other" were close to being contrepèteries but did not fully meet the criteria. For offensiveness, the option "Other" was selected when a potential resolution of a contrepèterie would have been offensive, but the generated contrepèterie did not fully work. However, 70% were judged unrelated to the given context (see Table 6).

For *contrepèteries*, Gemini struggles to generate valid instances. Although the generated sentences often resemble genuine *contrepèteries*, attempting to resolve the intended sound or letter transpositions frequently fails to produce a grammatically or semantically valid sentence in French. As a result, many outputs do not satisfy the linguistic constraints required for authentic *contrepèteries*. This limitation is illustrated by the example *"Goûtez-moi cette tomate"*. Applying the intended sound transposition yields *"Toutez-moi cette gomate "*, which is neither grammatically nor semantically valid in French. Since *toutez* and *gomate* are not French words, the generated output does not constitute a valid *contrepèterie*. Despite these limitations, Gemini occasionally generates valid *contrepèteries*.

Another interesting observation is that Gemini occasionally reproduces well-known *contrepèteries* instead of generating new ones. For example, the famous *contrepèterie* « folle de la messe » appears nine times, despite being unrelated to the corresponding input *"Cedi, " "il en va de suie", "anagram", "folie", "cinq", "formes", "dong", "obsolètement", and "SI | Venn"*. Similarly, *"glisser dans la piscine"* is generated four times for the unrelated *"précipitation", "Venn", "transparent", "White | Bluecorp"*. These examples suggest that, rather than conditioning on the provided input, the model sometimes defaults to memorized, well-known *contrepèteries*.

4. Conclusions

This paper presented the UBO participation in JOKER at CLEF 2026 for Task 3 on onomastic wordplay translation and Task 4 on humour generation. We evaluated several state-of-the-art LLMs using prompt-based approaches without task-specific fine-tuning. For wordplay translation, we investigated the use

of contextual descriptions of source named entities, while for humour generation we explored both pun generation in English, Spanish, and French as well as the generation of French *contrepèteries*.

The official evaluation and our manual analysis show that current LLMs still struggle with the linguistic constraints underlying wordplay. While direct prompting of open LLMs did not outperform the identity baseline for the onomastic wordplay translation task, the Gemini-based submission achieved nearly four times the baseline accuracy, with results comparable to the best-performing systems. Llama left 25% of source names untranslated, while Mistral, Qwen, and Gemma rarely did so but achieved limited success. Gemini was the best-performing model, outperforming all systems in both the official ranking and the generation of genuine alternative translations. Nevertheless, the best-performing systems remained below a 50% success rate for the cases requiring a genuine translation.

Across all three languages, Gemini-generated puns ranked among the top-performing systems, whereas the open-source models consistently ranked lower. In contrast, French *contrepèteries* were ranked near the bottom, possibly reflecting the difficulty of automatically evaluating this language-specific form of wordplay. Overall, the generated jokes received low funniness ratings, with *contrepèteries* performing particularly poorly. Generating authentic French *contrepèteries* remains a difficult task, requiring phonological and lexical manipulations that are not yet reliably mastered by existing models.

Our manual analysis revealed two notable limitations. First, AI-generated puns are often overly verbose or include unnecessary details, which can reduce their humorous effect. Second, Gemini occasionally reproduces well-known *contrepèteries* instead of generating new ones, suggesting that the model sometimes defaults to memorized examples rather than conditioning on the provided input.

A limitation of this work is that all submissions rely on a single prompt for each subtask and language, without exploring alternative prompting strategies. Moreover, the manual analysis focuses on the outputs of a single model (Gemini) and is therefore intended to provide qualitative insights rather than a comprehensive comparison across all evaluated LLMs. Since each instance was assessed by a single annotator, the manual evaluation may be affected by subjective judgments.

Future work will investigate prompt optimization techniques, including multi-step prompting and self-refinement, as well as extend our study to additional types of humour and wordplay and specialized evaluation methods for computational humour.

Acknowledgments

We thank the track and task organizers for their amazing service and effort in making realistic benchmarks available for analyzing and processing humorous text.

This project is funded by the National Research Agency (ANR-19-GURE-0001) under the program "*Investissements d'avenir*" integrated into France 2030.

Declaration on Generative AI

During the preparation of this work, the authors used *ChatGPT* and *Grammarly* in order to: **Grammar and spelling check** and **Paraphrase and reword**. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Ermakova, et al., Overview of the CLEF 2026 JOKER track: Humor detection, search, and translation, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, 2026.

[2] L. Ermakova, et al., Overview of the CLEF 2026 JOKER Task 3: Onomastic Wordplay Translation from English to French, in: [10], 2026.
[3] I. Kuzmin, L. Ermakova, A.-G. Bosser, J. Kamps, Overview of the CLEF 2026 JOKER Task 4: Humor Generation, in: [10], 2026.
[4] P. Vachharajani, A. Mukherjee, J. Singh, K. Tewari, S. Pal, L. Ermakova, J. Kamps, Overview of the CLEF 2026 JOKER Task 1: Humor-Aware Information Retrieval in English and Hinglish, in: [10], 2026.
[5] L. Ermakova, et al., Overview of the CLEF 2026 JOKER Task 2: Pun Translation from English to French, in: [10], 2026.
[6] Qwen Team, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.
[7] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.
[8] Gemma Team, Gemma (2024). URL: https://www.kaggle.com/m/3301. doi:10.34740/KAGGLE/M/3301.
[9] S. Weerawardhena, P. Kassianik, B. Nelson, B. Saglam, A. Vellore, A. Priyanshu, S. Vijay, M. Aufiero, A. Goldblatt, F. Burch, E. Li, J. He, D. Kedia, K. Oshiba, Z. Yang, Y. Singer, A. Karbasi, Llama-3.1-foundationai-securityllm-8b-instruct technical report, 2025. URL: https://arxiv.org/abs/2508.01059. arXiv:2508.01059.
[10] E. Sánchez Salido, A. Barrón Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.