pjmathematician at the CLEF 2025 JOKER Lab Tasks 1, 2 & 3: A Unified Approach to Humour Retrieval and Translation using the Qwen LLM Family

Notebook for the JOKER Lab, CLEF 2025

Poojan Vachharajani[1]

[1]*Netaji Subhas University of Technology, New Delhi, India*

Abstract

This paper details the participation of the pjmathematician team in all three tasks of the JOKER 2025 track: Humour-aware Information Retrieval, Pun Translation, and Onomastic Wordplay Translation. Our approach uniformly leverages the Qwen family of large language models (LLMs), applying distinct strategies tailored to each task's unique challenges. For Task 1, we implemented a two-stage process involving an LLM-based joke filter and explainer followed by a dense retriever, achieving a MAP@1000 of 0.3501 for English and 0.4221 for Portuguese on the test set. For Task 2, we fine-tuned various Qwen models to translate puns from English to French, with our best model (Qwen2.5-14B) attaining a BLEU score of 0.379. For Task 3, we explored zero-shot prompting for the complex task of translating onomastic wordplay, where our Qwen3-32B model achieved an exact match accuracy of 0.22. These results demonstrate the versatility of modern LLMs in handling nuanced and creative language tasks, highlighting the effectiveness of filtering for humour-aware retrieval, fine-tuning for pun translation, and the challenges in zero-shot onomastic translation.

Keywords

Computational Humor, Wordplay, Information Retrieval, Machine Translation, Large Language Models,

1. Introduction

The automatic analysis of wordplay and humour presents significant challenges for natural language processing, requiring a deep understanding of semantics, cultural context, and linguistic creativity [1]. The JOKER track at CLEF 2025 continues to foster research in this area through a set of shared tasks focused on the retrieval, translation, and interpretation of humorous texts [2].

Our team, pjmathematician, participated in all three tasks:

- **Task 1: Humour-aware Information Retrieval**: Retrieving documents that are both relevant to a query and contain wordplay, for English and Portuguese [3].
- **Task 2: Translation of Puns**: Translating English puns into French while preserving the humorous mechanism [4].
- **Task 3: Onomastic Wordplay Translation**: Translating meaningful names from English to French, often from fictional works [5].

The recent advancements in Large Language Models (LLMs) offer powerful new tools for tackling these complex tasks. Our work investigates the capabilities of the Qwen family of models [6, 7] across the JOKER challenges. We hypothesize that by employing tailored strategies—retrieval augmentation, fine-tuning, and sophisticated zero-shot prompting—we can effectively address the nuances of each task. This paper describes our methods, presents our experimental results, and discusses the findings from applying these models to the JOKER 2025 datasets [8].

CLEF 2025 Working Notes, September 9 – 12, 2025, Madrid, Spain

pjmathematician@gmail.com (P. Vachharajani)

2. Related Work

The work presented in this paper intersects with several key areas of computational linguistics and information retrieval: computational humor, machine translation of creative language, and the application of Large Language Models (LLMs) to these nuanced tasks. Our research is situated within the context of the JOKER shared task at CLEF, which specifically encourages interdisciplinary work on the automatic analysis of humor and wordplay [1].

2.1. Humour-aware Information Retrieval

Task 1 extends a line of research focused on retrieving documents that are not only topically relevant but also possess a specific stylistic quality—in this case, humor. The JOKER track has been instrumental in creating reusable datasets to foster research in this niche. Previous approaches in humor detection and retrieval have ranged from methods using classic machine learning with engineered features to deep learning models. More recently, the advent of LLMs has introduced new possibilities, with research exploring their use for humor detection and explanation [9, 10]. Our two-stage pipeline, which uses a larger LLM for deep content analysis (identifying and explaining jokes) before passing documents to a specialized dense retriever, contributes a novel, structured approach to this retrieval problem, moving beyond simple classification to a more integrated retrieval framework.

2.2. Translation of Wordplay and Puns

The translation of humor, particularly puns, is a classic and formidable challenge in both human and machine translation. The difficulty lies in the fact that puns often rely on language-specific phonological or semantic ambiguities that rarely have direct equivalents in a target language [11]. The work of Delabastita, who proposed a typology of pun translation strategies (e.g., pun-to-pun, pun-to-non-pun), provides a crucial theoretical framework for this task [12]. While traditional machine translation has proven impervious to wordplay, recent studies have begun to explore how LLMs can be prompted to handle such creative translations, though challenges remain [13, 14]. Our approach for Task 2 builds on this by not just prompting but fine-tuning Qwen models with a specific ‘pun->pun‘ instructional strategy. This aligns with findings that fine-tuning can adapt LLMs for specialized tasks, including creative text generation and improved translation quality [6, 15].

2.3. Onomastic Wordplay Translation

Task 3, the translation of meaningful names (onomastic wordplay), is a specialized subset of literary translation [16]. Such names are common in fiction, from classic literature to modern series like *Harry Potter*, and their translation is crucial for preserving characterization and literary effect [17]. Translation strategies for names can range from direct copying to substitution with a functional or creative equivalent in the target language [18]. This task is particularly challenging for automated systems due to its high demand for cultural and contextual knowledge and creative adaptation. Our use of a sophisticated zero-shot prompt with a powerful LLM for this task explores the current limits of these models in handling highly creative and context-dependent translation without specific fine-tuning, a domain where LLMs have shown promise but also limitations [19].

2.4. Large Language Models for Creative Tasks

Underpinning our entire approach is the transformative capability of modern LLMs. While early research focused on humor detection with smaller models, today’s large-scale models have demonstrated remarkable, if imperfect, abilities in generation and nuanced understanding [20]. Research has shown that prompting and fine-tuning are effective methods for steering LLMs towards specific, often creative, goals [6]. Our work systematically applies these LLM-centric techniques—zero-shot prompting, explanation generation, and instruction-based fine-tuning—across the three distinct JOKER

tasks, demonstrating their flexibility and providing insights into which strategy is best suited for each type of humor processing challenge.

3. Approach

We utilized various models from the Qwen series, applying a specific methodology for each task. All experiments were conducted using the official JOKER 2025 datasets.

3.1. Task 1: Humour-aware Information Retrieval

Our approach for this task was a two-step pipeline: (1) content analysis and filtering, and (2) dense retrieval.

3.1.1. Content Analysis and Filtering

First, we processed the entire document corpus using Qwen3-14B and Qwen3-32B models to identify potential jokes and generate explanations. This step aimed to create a richer representation of each document by explicitly flagging humorous content. We used the prompt shown in Listing 1. This generated a ‘isJoke‘ flag and a textual ‘explanation‘ for each document, which were used in subsequent steps.

3.1.2. Dense Retrieval

We used Qwen3-4B and Qwen3-8B embedding models as our retrievers [7]. We experimented with several configurations:

- **Joke Filtering**: We ran experiments on the full dataset, as well as on subsets filtered by the ‘isJoke‘ flag from Qwen3-14B and Qwen3-32B.
- **Document Representation**: We indexed documents using either their original text or their text augmented with the generated ‘explanation‘.
- **Query Prompt**: We tested two retrieval prompts: a default query (‘’Given a web search query, retrieve relevant passages that answer the query’‘) and a humour-specific query (‘’Given a query, retrieve relevant jokes related to the query’‘).

3.2. Task 2: Translation of Puns

For this task, we adopted a fine-tuning approach to teach the models the creative art of pun translation. We used LoRA to fine-tune several Qwen models on the provided English-French parallel corpus. The core of our approach was the instructional prompt shown in Listing 2, which guided the model to focus on the ‘pun->pun‘ translation strategy. We experimented with different model sizes and LoRA parameters (rank ‘r‘ and alpha ‘a‘).

3.3. Task 3: Onomastic Wordplay Translation

Given the highly creative and context-dependent nature of onomastic wordplay, we opted for a zero-shot prompting approach with the powerful Qwen3-14B and Qwen3-32B models. The prompt (Listing 3) was carefully designed to be a comprehensive guide, instructing the model on various translation principles, including when to keep names identical, perform direct translations, or pursue creative adaptations based on the provided context (‘description‘).

4. Results

This section presents the results for each task. Our runs were submitted with the team name pjmathematician.

4.1. Task 1: Humour-aware Information Retrieval

We evaluated our retrieval configurations on the English training data to identify the best setup. As shown in Table 1, using a 'jokequery' prompt and filtering the corpus with a joke detector consistently outperformed other methods. The combination of the Qwen3-8B retriever with the Qwen3-14B filter and 32B explanation yielded the highest MAP@1000 score of 0.48.

Table 1
Task 1 ablation study results on English training data (MAP@1000).

Retriever	Filter	Explanation	MAP@1000
Qwen3-8B	14B filter	32B explanation	**0.48**
Qwen3-4B	14B filter	32B explanation	0.47
Qwen3-8B	32B filter	32B explanation	0.46
Qwen3-4B	14B filter	14B explanation	0.45
Qwen3-4B	No filter	14B explanation	0.44
Qwen3-8B	No filter	32B explanation	0.32

For our final submission on the test set (Table 2), we selected the best-performing configurations. For English, augmenting documents with explanations generated by a 14B model and filtering with a 32B model performed best, achieving a MAP of 0.3501. For Portuguese, where filtering was not applied, the approach was still effective, achieving a MAP of 0.4221. This indicates that augmenting documents with semantic explanations of humor is a viable strategy even without a pre-filtering step.

Table 2
Task 1 final results on test data (MAP@1000).

Language	Retriever	Filter	Explanation	MAP@1000
English	Qwen3-8B	32B filter	14B explanation	**0.3501**
English	Qwen3-8B	14B filter	14B explanation	0.3486
English	Qwen3-8B	No filter	32B explanation	0.3438
Portuguese	Qwen3-4B	No filter	32B explanation	**0.4221**
Portuguese	Qwen3-4B	No filter	14B explanation	0.4217

4.2. Task 2: Translation of Puns

The results of our fine-tuning experiments are shown in Table 3. The larger Qwen2.5-14B model, fine-tuned with a LoRA rank of 128, achieved the highest BLEU score of 0.379 on the test set. This demonstrates a clear correlation between model size and performance on this creative translation task. While BLEU is a limited metric for wordplay, these scores indicate that the fine-tuned models learned to generate syntactically valid and semantically relevant translations.

Table 3
Task 2 results on the test set (BLEU score).

Model	LoRA Params (r, a)	BLEU
Qwen2.5 14B	128, 256	**0.379**
Qwen3 8B	64, 128	0.374
Qwen3 4B	64, 128	0.372

4.3. Task 3: Onomastic Wordplay Translation

The zero-shot results for onomastic translation are presented in Table 4. This proved to be the most challenging task. The larger Qwen3-32B model achieved a higher exact match accuracy of 0.22 compared to the 14B model. While these scores may seem low, the strictness of the exact match metric for such a creative task must be considered. Many generated translations were semantically plausible and creative, even if they did not perfectly match the reference.

Table 4
Task 3 results on the test set (Exact Match Accuracy).

Model	Accuracy
Qwen3-32B	**0.22**
Qwen3-14B	0.20

5. Conclusions

In our participation in the JOKER 2025 track, we systematically evaluated the Qwen family of models on three distinct humour-related tasks. Our findings are threefold. First, for humour-aware information retrieval, a two-stage approach of filtering and explaining content with a powerful LLM before passing it to a dense retriever is highly effective. Second, for creative translation tasks like puns, fine-tuning moderately sized LLMs with targeted instructional prompts yields strong results. Third, the zero-shot translation of highly idiosyncratic onomastic wordplay remains a significant challenge, though larger models show incremental progress.

Future work could explore more advanced RAG (Retrieval-Augmented Generation) frameworks for Task 1, combining retrieval with a generative re-ranking step. For Tasks 2 and 3, exploring few-shot in-context learning or more sophisticated fine-tuning methods could further improve creative generation capabilities.

Declaration on Generative AI

During the preparation of this work, the author(s) used the Qwen family of models (Qwen3-32B, Qwen3-14B, Qwen2.5-14B, Qwen3-8B, Qwen3-4B) in order to: Generate content, Analyze data. Specifically, the models were used to generate joke explanations for Task 1, produce French translations for Tasks 2 and 3, and to perform the initial humor classification in Task 1. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content. The design of the prompts and the experimental framework represents the core contribution of the author(s).

References

[1] L. Ermakova, R. Campos, A.-G. Bosser, T. Miller, Overview of the CLEF 2025 JOKER lab: Humour in machine, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[2] L. Ermakova, R. Campos, A.-G. Bosser, T. Miller, CLEF 2025 JOKER lab: Humour in the machine, in: C. Hauff, et al. (Eds.), Advances in Information Retrieval. ECIR 2025, volume 15576 of *Lecture Notes in Computer Science*, Springer, Cham, 2025. doi:10.1007/978-3-031-88720-8_59.

[3] L. Ermakova, R. Campos, A.-G. Bosser, T. Miller, Overview of the CLEF 2025 JOKER task 1: Humour-aware information retrieval, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes

of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[4] L. Ermakova, A.-G. Bosser, T. Miller, R. Campos, Overview of the CLEF 2025 JOKER task 2: Wordplay translation from english into french, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[5] L. Ermakova, A.-G. Bosser, T. Miller, R. Campos, Overview of the CLEF 2025 JOKER task 3: Onomastic wordplay translation, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.
[6] Qwen Team, Qwen3 technical report, 2024. arXiv:2405.09388.
[7] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, J. Zhou, Qwen3 embedding: Advancing text embedding and reranking through foundation models, arXiv preprint arXiv:2406.05176 (2024).
[8] L. Ermakova, A.-G. Bosser, A. Jatowt, T. Miller, The JOKER corpus: English-french parallel data for multilingual wordplay recognition, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), Association for Computing Machinery, 2023, pp. 2796–2806. doi:10.1145/3539618.3591885.
[9] A. Poddar, B. Ribeiro, D. McCallum, Augmenting large language models with humor theory to understand puns, 2024. URL: https://www.purdue.edu/gradschool/research/data-and-reports/gsc-repository/files/2024/oral-presentations/cs-poddar-aniruddha.pdf.
[10] T. Winters, et al., THInC: A theory-driven framework for computational humor detection, arXiv preprint arXiv:2402.00845 (2024).
[11] D. Delabastita, Wordplay and translation: A selective survey, in: Wordplay and Translation, St. Jerome, Manchester, 1996.
[12] F. Delzendehrooy, G. Koochacki, An investigation of pun translatability in english translations of sa'di's ghazals based on delabastita's proposed model, Advances in Language and Literary Studies 7 (2016) 259–276.
[13] T. Miller, A. Lüdeling, H. Göransson, Human-computer interaction in pun translation, in: Proceedings of the Workshop on Wordplay in Complete Works (Wordplay 2022), 2022, pp. 26–36.
[14] F. Regattin, Automatic translation and wordplay: An amateur's (playful) thoughts, in: Proceedings of the Workshop on Wordplay in Complete Works (Wordplay 2022), 2022, pp. 55–62.
[15] M. Ebrahimi, et al., The fine-tuning paradox: Boosting translation quality without sacrificing llm abilities, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023.
[16] L. Manini, Meaningful literary names: Their forms and functions, and their translation, The Translator 2 (1996) 161–178.
[17] T. Lappi, Translation of names in children's fantasy literature: Bringing the young reader into play, Perspectives: Studies in Translation Theory and Practice 13 (2005) 126–138.
[18] E. Crisafulli, The adequate translation as a methodological tool: The case of dante's onomastic wordplay in english, Target. International Journal of Translation Studies 13 (2001) 1–28.
[19] M. Bayer, et al., Optimising chatgpt for creativity in literary translation: A case study from english into dutch, chinese, catalan and spanish, arXiv preprint arXiv:2404.16709 (2024).
[20] O. Popova, M. Dadić, Ž. Agić, Exploring humor in natural language processing: A comprehensive review of joker tasks at clef symposium 2023, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.

A. Prompts

```
You are a humor analysis assistant with expertise in jokes and wordplay. When presented
    with text, evaluate it according to the following criteria and output your analysis
    in JSON format enclosed within <json></json> tags.

## Evaluation Criteria:
1. Determine if the text is a joke (True/False)
2. If it is a joke, identify the type (pun, one-liner, knock-knock, observational, dad
    joke, anti-joke, etc.)
3. Provide a 1-2 sentence explanation of the joke mechanics or wordplay
4. Rate the joke on a scale of 1-5 (1=barely amusing, 5=extremely funny)

If the text is not a joke, only provide the explanation of why it's not considered a
    joke.

## Rating Guidelines:
1 = Barely amusing, might elicit a slight acknowledgment
2 = Mildly funny, might cause a small smile
3 = Moderately funny, could provoke a chuckle
4 = Very funny, likely to cause laughter
5 = Extremely funny, potentially causing uncontrolled laughter

## Output Format:
For a joke:
<json>
{
  "isJoke": True,
  "type": "[joke type]",
  "explanation": "[1-2 sentence explanation]",
  "rating": [1-5]
}
</json>

For non-jokes:
<json>
{
  "isJoke": False,
  "explanation": "[explanation why it's not a joke]"
}
</json>

## Examples:

Input: "Why don't scientists trust atoms? Because they make up everything."

Output:
<json>
{
  "isJoke": True,
  "type": "Pun",
  "explanation": "This joke plays on the dual meaning of 'make up' - both as 'constitute
      /form' and 'fabricate/lie about'.",
  "rating": 3
}
</json>

Input: "The weather has been really nice lately."

Output:
```

```
<json>
{
  "isJoke": False,
  "explanation": "This is a straightforward observation about weather conditions with no
      humorous elements, punchline, or wordplay."
}
</json>
```

Listing 1: Prompt for Joke Analysis (Task 1)

```
You are an expert bilingual translator specializing in wordplay and humor across English
     and French. Your task is to translate English puns and jokes into French while
    preserving the humor mechanism of the original.

## Translation Guidelines:

1. Identify the wordplay mechanism in the English pun (homophony, polysemy, etc.)
2. Translate using Delabastita's  p u n pun  strategy - preserve both form and meaning of
    the original wordplay
3. The French translation must contain an equivalent pun that works in French
4. Prioritize maintaining the same semantic field as the original when possible
5. If a direct translation doesn't preserve the pun, find French words with similar
    ambiguity

Output directly and only the translated french pun.
```

Listing 2: Fine-tuning Prompt for Pun Translation (Task 2)

```
You are an expert translator specializing in onomastic wordplay translation from English
     to French. Your task is to translate names, phrases, and terms containing wordplay
    while preserving the humor, references, and creative essence of the original.

## Task Description
You will receive:
- **en**: The English name/phrase containing wordplay
- **description**: Context explaining the name, its meaning, and usage

You must provide:
- Only the French translation in a specific JSON format

## Translation Principles

1. **Understand the Wordplay First**
   - Identify the pun, allusion, or linguistic creativity in the original
   - Consider the character/object's traits, function, or role as described

2. **Translation Approaches (in order of preference):**

   a) **Keep Identical** when:
   - The name is Latin-based and works in both languages (e.g., "Vulnera Sanentur," "
       Felix Felicis")
   - The reference is universal or already established

   b) **Direct Translation** when:
   - The wordplay relies on descriptive meaning (e.g., "Vipertooth"     "Dent-de-vip re
       ")
   - Compound words where components can be directly translated

   c) **Creative Adaptation** when:
```

```
   - Cultural adaptation is needed for the pun to work
   - Suffix patterns matter (e.g., "-ix" in Asterix characters)
   - Sound patterns or pronunciation-based humor is involved
   - The original contains idioms or cultural references

   d) **Functional Equivalence** for:
   - Spells (focus on magical effect: "obliviate"     "oubliettes")
   - Acronyms (maintain humor while changing letters: "SPEW"     "S.A.L.E")
   - Location names (prioritize French geographic naming conventions)

3. **Consider Character Traits**
   - For character names, ensure the French version reflects their personality
   - If the name describes appearance/behavior, maintain this connection

4. **Maintain Cultural Elements**
   - For fantasy series (Harry Potter, Asterix), respect established translation
       patterns
   - Consider French humor styles and wordplay traditions

5. **Phonetic Considerations**
   - Ensure pronounceability in French
   - When appropriate, preserve sound similarities to the English version

## Output Format
Your response must follow this exact JSON format:

<json>
{
"fr": "your French translation here"
}
</json>

Provide only this JSON object with no additional text, explanations, or commentary.
```

Listing 3: Zero-shot Prompt for Onomastic Translation (Task 3)