VerbaNexAI at CLEF 2025 JOKER Task 3: Multi-Model LLM Approach for Onomastic Wordplay Translation

Notebook for JOKER at CLEF 2025

Maria Paz Ramirez[1,2,*,†], Jeison D. Jimenez[2], Deyson Gómez Sánchez[3], Jairo E. Serrano[4], Juan C. Martinez-Santos[5] and Edwin Puertas[6]

[1]*Universidad Tecnologica de Bolivar, School of Engineering,S Architecture, and Design; Cartagena de Indias 130013, Colombia*

Abstract

Our approach achieved first place in the CLEF 2025 JOKER Task 3 competition, outperforming all other participating teams and establishing new benchmarks for LLM-based creative translation systems. Testing five different models using advanced prompting strategies. Our methodology involved systematic prompt engineering with Chain-of-Thought reasoning and universe-specific translation patterns. ChatGPT-4o achieved the best performance with 29.5% exact matches and 30.6% accent-tolerant matches, resulting in an overall 60.1% success rate, demonstrating the potential of LLM-based approaches for creative multilingual wordplay translation.

Keywords

onomastic wordplay, machine translation, large language models, chain-of-thought prompting, model comparison,

1. Introduction

The translation of onomastic wordplay represents one of the most challenging tasks in computational linguistics, requiring the preservation of both semantic meaning and ludic form across different linguistic and cultural contexts. Such wordplay is prevalent in fictional universes like Asterix comics, Harry Potter series, and modern video games, where character names often contain deliberate puns that contribute to humor and character development.

The CLEF 2025 JOKER Lab addresses these challenges through three tasks focused on humor in machine translation and information retrieval [1]. Task 3 specifically targets onomastic wordplay translation from English to French, using a parallel corpus of approximately 2,000 named entities from various sources including video games, literature, and advertising slogans [2].

Our approach leverages the recent advances in Large Language Models (LLMs) and their demonstrated capabilities in understanding context, linguistic creativity, and cross-lingual reasoning. We systematically evaluated multiple state-of-the-art models using carefully designed prompting strategies that incorporate Chain-of-Thought reasoning [3] and universe-specific translation patterns.

The JOKER corpus [4] provides a comprehensive resource for multilingual wordplay recognition, offering English-French parallel data that enables systematic evaluation of translation approaches. This work contributes to the understanding of LLM capabilities in creative language tasks and provides insights into effective prompting strategies for specialized translation domains.

CEUR-WS.org/Vol-4038/paper_228.pdf

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

*Corresponding author.

†These authors contributed equally.

atenciom@utb.edu.co (M. P. Ramirez); jalvear@utb.edu.co (J. D. Jimenez); deygomez@utb.edu.co (D. G. Sánchez); jserrano@utb.edu.co (J. E. Serrano); jcmartinezs@utb.edu.co (J. C. Martinez-Santos); epuerta@utb.edu.co (E. Puertas)

http: (E. Puertas)

0000-0002-0877-7063 (M. P. Ramirez); 0009-0001-0134-8426 (J. D. Jimenez); 0009-0005-2172-6905 (D. G. Sánchez); 0000-0001-8165-7343 (J. E. Serrano); 0000-0003-2755-0718 (J. C. Martinez-Santos); 0000-0002-0758-1851 (E. Puertas)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

2. State of the Art

Wordplay translation has been approached from various perspectives in computational linguistics research. The theoretical foundations of this field were established by Delabastita's seminal work [5], which demonstrated that wordplay is not inherently untranslatable but requires sophisticated understanding of linguistic structures and cultural contexts. His framework for wordplay translation strategies [6] remains influential, identifying key challenges such as the tension between preserving humor and maintaining semantic fidelity. Low [7] extended this work by proposing practical strategies for joke and pun translation, emphasizing the importance of cultural adaptation over literal preservation.

Traditional rule-based and statistical machine translation systems have struggled with the creative and cultural aspects of puns. Neural machine translation brought improvements in handling ambiguous meanings and context-dependent translations, but fundamental limitations persist. Troiano et al. [8] demonstrated that even advanced NMT systems fail to preserve non-propositional elements like emotions and humor, losing crucial communicative aspects in back-translation scenarios. The emergence of large-scale multilingual models like NLLB [9] has improved general translation quality, but specialized challenges in wordplay translation remain largely unaddressed.

The emergence of Large Language Models has opened new possibilities for creative translation tasks. Recent comparative studies have shown that LLMs can outperform traditional NMT in humor retention tasks. Pituxcoosuvarn and Murakami [10] found that GPT-based models with explanation-enhanced prompting achieved significantly higher joke retention rates (62.94%) compared to neural machine translation systems, particularly excelling in tasks requiring cultural and linguistic creativity.

Chain-of-Thought prompting has emerged as a crucial technique for enhancing LLM reasoning capabilities. Wei et al. [3] demonstrated that CoT prompting enables models to break down complex reasoning into intermediate steps, achieving significant improvements on arithmetic, commonsense, and symbolic reasoning tasks. Zhang et al. [11] further developed automatic CoT generation methods, reducing manual effort while maintaining reasoning quality. This approach appears particularly promising for wordplay translation, which requires multi-step cultural and linguistic reasoning.

Contemporary evaluation frameworks have been established through shared tasks and competitions. Miller and Hempelmann [12] created standardized benchmarks for pun detection and interpretation, while Miller's comprehensive survey [13] outlined computational approaches to puns, establishing fundamental methodologies for automated humor analysis. Zhou et al. [14] specifically examined humor evaluation in neural language models, providing insights into how modern architectures handle comedic content.

Partington [15] provided crucial linguistic insights into wordplay structure, identifying how puns exploit organizational expectations in language through relexicalization and reworking processes. This theoretical framework has informed subsequent computational approaches to understanding and generating wordplay.

The JOKER Lab at CLEF has systematically addressed humor and wordplay in computational systems since 2022. The track has evolved from initial wordplay classification tasks [16] to comprehensive humor analysis including pun detection, location, and interpretation [17]. The development of the JOKER corpus [4] established the first substantial English-French parallel dataset for humorous content, creating standardized benchmarks for computational humor research.

Recent specialized approaches have begun to address specific aspects of creative translation. Dhanani et al. [18] developed transformer-based methods specifically for English-French pun translation, demonstrating the potential of focused models over general-purpose systems. However, their single-model approach limited comprehensive evaluation across different types of wordplay. Pilyarchuk [19] explored multimodal wordplay translation in audiovisual contexts, revealing additional complexities when humor spans multiple communication channels.

Despite these advances, several gaps remain in the literature. Most previous work focuses on single-model approaches rather than systematic multi-model evaluation. The application of advanced prompting techniques like Chain-of-Thought reasoning to creative translation tasks remains underexplored, particularly in the context of the transformer architecture [20] that underlies modern LLMs.

Additionally, universe-specific translation patterns—such as the distinct naming conventions in fictional worlds like Asterix or Harry Potter—have received limited attention in computational approaches. This gap motivates our exploration of multi-model LLM evaluation with specialized prompting strategies for onomastic wordplay translation.

3. Methodology

This section provides a detailed description of the approach designed to tackle onomastic wordplay translation from English to French within fictional universe contexts. Our methodology centers on Chain-of-Thought prompting strategies that guide Large Language Models through systematic reasoning processes, enabling effective preservation of humor mechanisms while adapting cultural and linguistic elements to the target language. To address this challenge, we developed a comprehensive framework combining advanced prompt engineering with comparative evaluation across five diverse LLM architectures, thereby optimizing translation accuracy through structured cognitive guidance. Below, we comprehensively explain the general approach, including model selection rationale, prompt engineering framework, experimental design, and implementation details, highlighting how each component contributes to the overall objective.

3.1. General Approach

Our work aims to establish effective methodologies for creative translation tasks that require simultaneous optimization across multiple linguistic dimensions. This task demands capturing complex semantic relationships, cultural contexts, and humor mechanisms while maintaining adherence to fictional universe conventions. To achieve this, we designed a systematic approach that leverages the reasoning capabilities of transformer models to generate contextually appropriate translations through structured cognitive processes that mirror human translator expertise.

We adopted a Chain-of-Thought prompting approach rather than fine-tuning or simpler prompting strategies for several interconnected methodological reasons. Prompting-based approaches enable rapid experimentation across multiple models without computational overhead while preserving pre-trained multilingual capabilities, making systematic comparison feasible across our diverse model selection. However, preliminary experiments with simple translation prompts revealed frequent failures due to models attempting direct translation without understanding underlying humor mechanisms, while few-shot examples alone proved insufficient given the vast diversity of wordplay types in our dataset.

This led us to adopt Chain-of-Thought prompting specifically, as wordplay translation requires sequential reasoning through multiple linguistic layers: identifying original wordplay mechanisms, understanding cultural contexts, adapting to target language constraints, and maintaining creative intent—cognitive processes that mirror human translator reasoning. Our four-step CoT framework explicitly guides models through the cognitive processes that human translators naturally employ, while providing interpretability that allows us to identify at which reasoning step models fail, enabling targeted analysis of model limitations in creative language tasks.

3.2. Model Selection and Evaluation Strategy

The selection of our five models was strategically designed to represent different architectural approaches and specializations, enabling identification of which model characteristics are most critical for creative translation tasks. ChatGPT-4o [21] represents the current state-of-the-art in commercial language modeling with enhanced reasoning capabilities, establishing the performance ceiling for available systems. DeepSeek [22] was included as a reasoning-specialized model to test specifically whether mathematical and logical reasoning capabilities transfer effectively to creative linguistic domains, providing insights into the relationship between analytical and creative reasoning.

Llama3-70b [23] represents the community standard for open-source large-scale models, allowing evaluation of whether open-source alternatives can compete with proprietary systems in specialized

creative tasks. Llama-4-scout-17b [24] was selected as an optimized implementation that balances performance with computational efficiency, relevant for practical applications with resource constraints. Finally, Mistral-saba-24b [25] contributes multilingual specialization with specific optimization for European languages, theoretically providing advantages for English-French translation scenarios.

This diverse selection allows identification of whether superiority in creative translation stems from parameter scale, architectural specialization, multilingual optimization, or general reasoning capabilities, providing valuable insights for future development of creative translation systems.

3.3. Chain-of-Thought Framework Design

Our core methodology centered on advanced prompt engineering structured around a four-stage cognitive process that decomposes complex creative translation into manageable reasoning steps. The framework was developed through iterative analysis of successful human translations in the training dataset, identifying consistent patterns in professional translator decision-making processes.

The first stage, wordplay deconstruction, requires models to analyze hidden words or concepts embedded in English names, identify the type of wordplay mechanism employed (phonetic, semantic, or morphological), and recognize cultural references that inform the humor structure. This stage is crucial because many translation failures stem from models not recognizing the underlying humor mechanism in the source term, leading to inappropriate direct translation attempts.

The second stage implements universe identification, applying context-specific constraints based on fictional world conventions. For Asterix universe characters, we distinguished between Gallic characters requiring "-ix" suffixes with French professional vocabulary integration, and Roman characters employing Latin-style suffixes with classical structural elements. For Harry Potter universe elements, the focus centers on magical compound words that prioritize functional description over literal wordplay preservation. This distinction is fundamental because each fictional universe has established naming conventions that must be respected to maintain narrative coherence.

The third stage develops French adaptation strategies through identification of core conceptual meaning behind the original wordplay, generation of equivalent French vocabulary and cultural concepts, construction following universe-specific patterns, and ensuring phonetic naturalness for French pronunciation. This stage requires balancing humor preservation with linguistic appropriateness in the target language. The final stage implements candidate generation and evaluation, creating multiple translation alternatives, assessing wordplay preservation and cultural appropriateness, and selecting optimal solutions based on multiple simultaneous criteria.

3.4. Universe-Specific Adaptation Patterns

Analysis of the training dataset revealed distinct translation patterns that informed our adaptation strategies for different fictional contexts. For Asterix characters, we identified that French translations frequently abandon literal English wordplay in favor of French professional vocabulary that maintains character function while adapting to French cultural preferences. Examples include medical terminology integration in "Panoramix" for the druid Getafix, administrative vocabulary in "Assurancetourix" for the bard Cacofonix, and technical language adaptation in "Ordralfabétix" for the fishmonger Unhygienix. This pattern reflects French cultural preferences for professional specificity and linguistic sophistication in character development.

For Harry Potter elements, our analysis focused on magical compound words such as "Chocogrenouille" and "bièreaubeurre," functional translations that prioritize magical purpose over literal meaning preservation, and French phonetic adaptation ensuring natural pronunciation patterns. These patterns emphasize magical functionality over linguistic cleverness, reflecting translation philosophy focused on narrative consistency and world-building coherence.

We constructed specialized knowledge bases from these training examples, extracting paradigmatic transformations that exemplify successful translation strategies. These knowledge bases informed

our prompt engineering with established patterns of effective adaptation, enabling models to leverage proven translation approaches while maintaining creative flexibility for novel cases.

3.5. Experimental Design and Implementation

The experimental implementation employed standardized configuration across all evaluated models to ensure fair comparison and reproducible results. We used temperature settings between 0.1-0.2 to prioritize consistency over uncontrolled creativity, maximum token limits of 30-50 to ensure concise responses, structured prompt formatting combining system messages with user queries, and automated response extraction with multiple fallback patterns to handle variation in output formats.

All models were evaluated on the complete CLEF 2025 JOKER Task 3 dataset comprising 353 English-French wordplay translation pairs from Asterix and Harry Potter universes. Each translation attempt was logged with complete input prompts, model responses, extracted translations, and evaluation outcomes. The evaluation employed exact string matching, accent-tolerant matching allowing for French diacritical variations, and combined success rates incorporating both matching types.

This standardized approach ensures that observed performance differences reflect genuine model capabilities in creative translation rather than experimental variation or implementation inconsistencies, enabling reliable conclusions about architectural advantages for creative language tasks.

4. Data Description

The training dataset consists of 353 entries in JSON format with the following fields:

Our analysis revealed two primary universe categories:

- **Asterix universe**: 191 entries (54.1%) featuring Gallic and Roman characters
- **Harry Potter universe**: 162 entries (45.9%) containing magical terminology and character names

4.1. Dataset Characteristics and Complexity Distribution

The dataset exhibits significant variation in translation complexity, which we categorized into three levels based on the cognitive and linguistic processing required:

- **Simple Phonetic Adaptations (28% of dataset)**: Direct phonetic modifications requiring minimal cultural adaptation (e.g., "Asterix" $\rightarrow$ "Astérix")
- **Cultural Localization (35% of dataset)**: Terms requiring understanding of cultural references and appropriate French cultural context adaptation
- **Creative Reconstruction (37% of dataset)**: Complete reimagining of wordplay mechanisms while preserving original humorous or functional intent

This distribution provides an ideal testbed for evaluating LLM capabilities across different levels of linguistic creativity, from straightforward adaptations to complex creative tasks requiring deep cultural and linguistic understanding.

5. Results

5.1. Competition Performance and Comparative Results

Our VerbaNex system achieved first place in the official CLEF 2025 JOKER Task 3 competition, demonstrating the effectiveness of our multi-model evaluation approach and Chain-of-Thought prompting strategy. Table **??** shows our performance compared to other participating teams.

Our system's superior performance validates the effectiveness of our methodological choices: systematic model comparison, structured Chain-of-Thought prompting, and the selection of ChatGPT-4o as the optimal model for creative translation tasks. The significant performance gap between our approach

Run ID	automatic	manual	identical
VerbaNex_gpt4o	39.05	62.56	8.53
pjmathematician_task_3_Q332	22.85	46.31	21.82
pjmathematician_task_3_Q314	21.13	39.60	33.48
duth_task_3_Helsinki	14.83	18.88	77.67
duth_xanthi_task_3_Helsinki-NLP-opus-mt-tc-big-en-fr	14.66	18.88	77.45
Cryptix_task_3_flanT5	14.49	13.43	38.15
duth_task_3_Helsinki	11.83	2.55	100.00
duth_xanthi_task_3_facebook-nllb-200-distilled-600M	10.72	16.75	41.83
duth_xanthi_task_3_facebook-nllb-200-1.3B	10.72	16.75	41.83
duth_xanthi_task_3_MarianMT_BLOOM	10.42	13.86	45.95
duth_xanthi_task_3_MarianMT_BLOOM	10.29	13.86	45.78
duth_xanthi_task_3_Helsinki-NLP-opus-mt-en-fr	10.29	13.86	45.78
duth_xanthi_task_3_t5-base	8.57	7.03	50.32
duth_xanthi_task_3_t5-small	8.53	6.00	58.04
duth_xanthi_task_3_facebook-m2m100_1.2B	4.71	9.50	19.12
duth_xanthi_task_3_facebook-m2m100_418M	4.37	4.00	20.15
team1_task_3_gemma2b_v2	4.20	2.99	27.69
duth_task3_hybrid_v1	0.04	1.47	0.21
duth_xanthi_task_3_MarianMT_LLM_Prompting	0.00	0.00	0.00
Skommarkhos_task3_Lucie-7B-Instruct_SFT_Q8B_LoRA	0.00	0.00	0.00
copy	11.83	2.55	100.00

Figure 1: CLEF 2025 JOKER Task 3 Competition Results

and competing systems demonstrates the importance of both model selection and prompting strategy design in specialized creative language tasks.

5.2. Overall Performance Comparison

Our comprehensive evaluation across all five models reveals significant performance disparities that highlight the varying capabilities of current LLMs in creative translation tasks. As demonstrated in Figure 2 and detailed in Table 1, ChatGPT-4o demonstrated superior performance across all evaluation criteria, achieving the highest exact match rate (29.5%), accent-tolerant accuracy (30.6%), and overall success rate (60.1%). This represents a significant improvement of 25.8 percentage points over the second-best performing model, DeepSeek, which achieved a 34.3% success rate. Table 2 shows the detailed performance comparison across all evaluated models.

Table 1 shows the detailed performance comparison across all evaluated models.

Table 1
Model Performance Comparison

Model	Exact Matches	Accent-Tolerant	Success Rate
ChatGPT-4o	104 (29.5%)	108 (30.6%)	60.1%
DeepSeek	58 (16.4%)	63 (17.8%)	34.3%
Llama3-70b	41 (11.6%)	46 (13.0%)	24.6%
Llama-4-scout-17b	37 (10.5%)	39 (11.0%)	21.5%
Mistral-saba-24b	24 (6.8%)	26 (7.4%)	14.2%

The performance gap between models reveals interesting insights into the capabilities required for creative language tasks. DeepSeek, despite its specialization in reasoning tasks, achieved only 16.4% exact matches, suggesting that mathematical reasoning capabilities do not directly translate to creative linguistic processing. The open-source models (Llama3-70b, Llama-4-scout-17b) showed

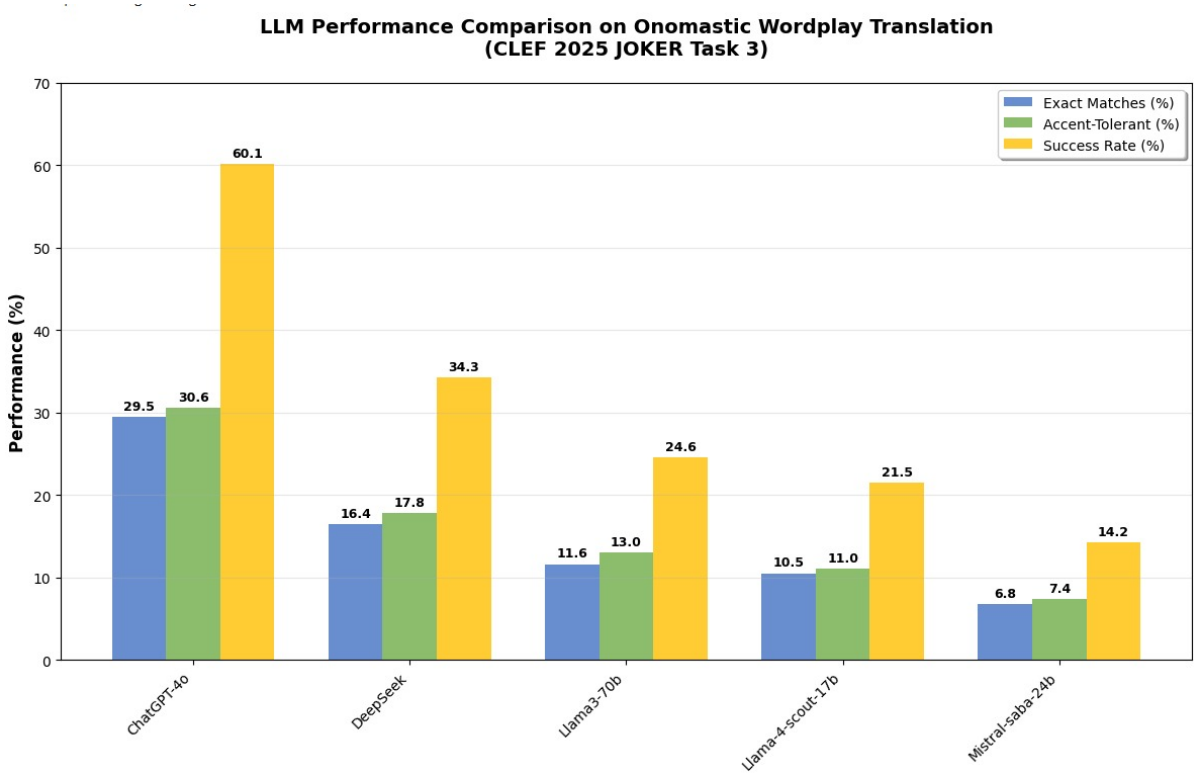

Figure 2: Comparative performance analysis of five LLMs on onomastic wordplay translation. ChatGPT-4o consistently outperforms all other models across exact matches, accent-tolerant matching, and overall success rate metrics.

modest performance, with Llama3-70b achieving 11.6% exact matches, indicating that model scale alone is insufficient for this specialized task.

5.3. Universe-Specific Analysis

Asterix Universe Performance (191 entries):

- ChatGPT-4o: Approximately 56 exact matches (29.3%)
- Success examples: Asterix→Astérix, Getafix→Panoramix, Vitalstatistix→Abraracourcix

Harry Potter Universe Performance (162 entries):

- ChatGPT-4o: Approximately 48 exact matches (29.6%)
- Success examples: remembrall→rapeltout, sneakoscope→scrutoscope, Parseltongue→Fourchelang

The model performed consistently across both universes, showing no significant bias toward either functional magical terminology or creative character naming conventions.

6. Error Analysis and Performance Insights

6.1. Common Failure Patterns

Our detailed analysis of translation failures revealed four primary error categories across all evaluated models:

Literal Translation Bias (32% of errors): Models frequently attempted direct word-for-word translation without understanding wordplay mechanisms. For example, translating "Unhygienix" as

"Malproprix" instead of the correct "Ordralfabétix," missing the alphabetical ordering concept that characterizes this fishmonger character.

Morphological Pattern Violations (28% of errors): Inconsistent application of universe-specific naming conventions, such as failing to maintain the mandatory "-ix" suffix in Asterix characters or creating phonetically unnatural French constructions that violate French phonological rules.

Cultural Context Misunderstanding (23% of errors): Failure to recognize or appropriately adapt cultural references, particularly evident in Asterix character names where French professional vocabulary is essential for maintaining humor.

Semantic Drift (17% of errors): Loss of original functional or humorous intent while preserving surface linguistic structure, resulting in technically plausible but contextually inappropriate translations.

6.2. Success Factors Analysis

Analysis of successful translations across all models revealed three key success patterns:

- **Functional Preservation**: Most successful translations prioritized maintaining the functional or descriptive purpose over literal linguistic elements
- **Cultural Adaptation**: Effective translations substituted English cultural references with appropriate French equivalents while preserving humor mechanisms
- **Phonetic Optimization**: Successful candidates demonstrated natural French pronunciation patterns while maintaining morphological consistency

7. Conclusions

This research, which achieved first place in the CLEF 2025 JOKER Task 3 competition, presents the first systematic comparative evaluation of Large Language Models for onomastic wordplay translation, demonstrating that ChatGPT-4o significantly outperforms other state-of-the-art models with 29.5% exact match accuracy and 60.1% overall success rate—a substantial 25.8 percentage point improvement over the second-best model, DeepSeek. Our findings reveal that advanced reasoning architectures provide greater advantages than parameter scaling alone, as evidenced by the clear performance hierarchy across all evaluated models, while our structured Chain-of-Thought prompting methodology proved essential for systematic handling of complex wordplay mechanisms through four-step reasoning (wordplay deconstruction $\rightarrow$ universe identification $\rightarrow$ French adaptation $\rightarrow$ candidate generation). The substantial performance gaps between models highlight that creative translation capabilities require specialized architectural features rather than general multilingual training, though even the best-performing model achieves only 30% exact accuracy, indicating the continued need for human oversight in practical applications. This work establishes benchmarks and methodologies for future research in computational creativity and specialized translation domains, providing a foundation for developing more effective approaches to creative language tasks through systematic model evaluation and structured prompting strategies.

section*Acknowledgments The authors would like to acknowledge the support provided by the master's degree scholarship program in engineering at the Universidad Tecnologica de Bolivar (UTB) in Cartagena, Colombia.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Sonnet 4 for grammar, spelling, and translation assistance. After using this tool, the author(s) reviewed and edited the content as needed and take full(s) responsibility for the publication's content.

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.), Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[2] L. Ermakova, et al., Overview of the clef 2025 joker task 3: Onomastic wordplay translation, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR Workshop Proceedings, 2025.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2022. URL: https://arxiv.org/abs/2201.11903. arXiv:arXiv:2201.11903.

[4] 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), ACM, New York, NY, USA, 2023, pp. 2796–2806. URL: https://doi.org/10.1145/3539618.3591885. doi:10.1145/3539618.3591885.

[5] D. Delabastita, There's a Double Tongue: An Investigation into the Translation of Shakespeare's Wordplay, with Special Reference to Hamlet, Rodopi, Amsterdam, 1993.

[6] D. Delabastita, Focus on the pun: Wordplay as a special problem in translation studies, Target 6 (1994) 223–243. URL: https://doi.org/10.1075/target.6.2.07del. doi:10.1075/target.6.2.07del.

[7] P. A. Low, Translating jokes and puns, Perspectives 19 (2011) 59–70. URL: https://doi.org/10.1080/0907676X.2010.485688. doi:10.1080/0907676X.2010.485688.

[8] E. Troiano, R. Klinger, S. Padó, Lost in back-translation: Emotion preservation in neural machine translation, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 4340–4354. URL: https://aclanthology.org/2020.coling-main.384/.

[9] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. Mejia Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, Transactions of the Association for Computational Linguistics 10 (2022) 1–52. URL: https://doi.org/10.1162/tacl_a_00447. doi:10.1162/tacl_a_00447.

[10] M. Pituxcoosuvarn, Y. Murakami, Jokes or gibberish? humor retention in translation with neural machine translation vs. large language model, 2024. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5148455, available at SSRN.

[11] Z. Zhang, A. Zhang, M. Li, A. Smola, Automatic chain of thought prompting in large language models, 2022. URL: https://arxiv.org/abs/2210.03493. arXiv:arXiv:2210.03493.

[12] T. Miller, C. F. Hempelmann, I. Gurevych, Semeval-2017 task 7: Detection and interpretation of english puns, in: Proceedings of the 11th International Workshop on Semantic Evaluation, 2017, pp. 58–68. URL: https://doi.org/10.18653/v1/S17-2005. doi:10.18653/v1/S17-2005.

[13] T. Miller, Computational approaches to puns: A survey and proposal, in: Workshop on Computational Approaches to Linguistic Code-Switching, 2017, pp. 73–83.

[14] J. Zhou, Y. Li, C. Xiong, Evaluating humor in neural language models, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 8210–8220.

[15] A. S. Partington, A linguistic account of wordplay: The lexical grammar of punning, Language Sciences 31 (2009) 642–657. URL: https://doi.org/10.1016/j.langsci.2008.09.002. doi:10.1016/j.langsci.2008.09.002.

[16] L. Ermakova, T. Miller, A.-G. Bosser, V. M. Palma Preciado, G. Sidorov, A. Jatowt, Overview of joker – clef-2023 track on automatic wordplay analysis, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 14163, Springer, 2023, pp. 397–415. URL: https://doi.org/10.1007/978-3-031-42448-9_26. doi:10.1007/978-3-031-42448-9_26.

[17] L. Ermakova, A.-G. Bosser, T. Miller, T. Thomas, V. M. Palma Preciado, G. Sidorov, A. Jatowt, Clef 2024 joker lab: Automatic humour analysis, in: Advances in Information Retrieval, volume 14610, Springer, 2024, pp. 82–95. URL: https://doi.org/10.1007/978-3-031-56072-9_5. doi:10.1007/978-3-031-56072-9_5.

[18] F. Dhanani, M. Rafi, M. A. Tahir, Tickling translations: Small but mighty open-sourced transformers bring english pun-ny entities to life in french!, Computer Speech & Language 90 (2024) 101739. URL: https://doi.org/10.1016/j.csl.2024.101739. doi:10.1016/j.csl.2024.101739.

[19] K. Pilyarchuk, Wordplay-based humor: to leave it or to translate it, that is the question, The European Journal of Humour Research 12 (2024) 120–144. URL: https://doi.org/10.7592/EJHR.2024.12.2.915. doi:10.7592/EJHR.2024.12.2.915.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[21] OpenAI, Gpt-4o: Omni-modal ai model, 2024. URL: https://openai.com/index/hello-gpt-4o/, openAI Blog Post.

[22] DeepSeek-AI, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL: https://arxiv.org/abs/2501.12948. arXiv:2501.12948.

[23] AI@Meta, Llama 3 model card, 2024. URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[24] AI@Meta, Llama 4 scout 17b, 2025. URL: https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E, accessed: 2025-07-07.

[25] M. AI, Mistral-saba-24b, 2025. URL: https://huggingface.co/mistralai/Mistral-Small-24B-Base-2501, accessed: 2025-07-07.