Semantic Stars at MWE-2026 PARSEME 2.0 Subtask 2: Alternative Approaches for MWE Paraphrasing

Elif Bayraktar[1], Vedat Doğancan[1], Muhammed A. Gümüş[1], Nusret Ali Kızılaslan[1],

[1]Istanbul Technical University, Istanbul, Turkey,

Correspondence:{bayraktare24, dogancan23, gumus24, kizilaslan25 }@itu.edu.tr

Abstract

This paper describes the system submitted by Semantic Stars Team for Subtask 2 of the PARSEME 2.0 shared task (Paraphrasing Multiword Expressions). Our approach addresses the challenge of paraphrasing sentences containing MWEs such that the MWE is removed while the original meaning and grammatical structure are preserved. The paper describes multiple distinct approaches powered by open-weight Large Language Models (LLMs), each employing a combination of different techniques such as prompting, multi-agent pipelines and classical NLP methods. Four distinct methods are tested on the test data in French, including a fifth one combining the results from the first four. We tested with several different open-weight LLMs including Llama3.1:8b, Qwen3:8b and gpt-oss-120b and were able to achieve significant improvements over the baseline, securing the first place on the shared task leader board.

1 Introduction

Automatic paraphrasing of Multi-Word Expressions (MWEs), is a challenging task in Natural Language Processing (NLP). MWEs require precise identification of idiomatic boundaries (Savary et al., 2017) and the generation of semantic equivalents that preserve meaning and grammatical closeness. In French, with its discontinuous dependencies and strict morphological constraints, purely stochastic approaches often struggle to maintain grammatical and semantic integrity.

In this paper, we explore multiple paraphrasing methods to address these challenges within the PARSEME 2.0 shared task (Scholivet et al., 2026). We test different approaches, including detailed descriptive paraphrasing (Approach 1), minimal replacements (Approach 2), multi-stage LLM pipelines (Approach 3), and neuro-symbolic hybrid systems (Approach 4). Each method leverages distinct strategies to tackle the task, from rule-based validation to advanced LLM-driven generation. Additionally, we experimented with a combined approach that selects the best output from these methods, although it did not outperform the individual approaches. The results can be seen in Table 1. The following sections detail the architecture of each method and their respective performances.

2 Related Works

Multiword Expression (MWE) processing is a long-standing yet prevalent field in NLP. MWE processing encompasses challenging tasks such as discovery, identification, and paraphrasing of MWEs. It is clear that the academic literature on NLP has a continued interest in this topic. Surveys and review studies were conducted to explore the challenges, discover subtasks related to MWE processing and examine methods developed for modeling MWEs (Constant et al., 2017; Sag et al., 2002; Villavicencio and Idiart, 2019).

MWE paraphrasing, which is our focus, remains one of the most important tasks within the scope of MWE processing. For instance, Yimam et al. (2016) aimed to examine the importance of context for the paraphrase ranking task using a binary classification approach with K-Nearest Neighbors (kNN) and a learning-to-rank approach with the LambdaMART algorithm. Barančíková and Kettnerová (2018) used word embeddings to paraphrase Czech verbal MWEs using single verbs. A machine translation experiment was conducted comparing sentences from both versions, and the results showed that the performance of the approach is promising. Zhou et al. (2022) proposed an approach for the Idiomatic Sentence Paraphrasing (ISP) task without strong supervision. This study addressed the challenge of data scarcity in the MWE paraphrasing domain. Wada et al. (2023) proposed an unsupervised approach to MWE paraphrasing, which includes clustering sentences that

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 187–195

contain MWEs using the DBSCAN algorithm and utilizing pre-trained LLMs to generate paraphrases for MWEs. Barreiro and Mota (2023) addressed the problem of multilingual paraphrasing of MWEs and developed a tool named CLUE-Aligner to facilitate the alignment of non-contiguous multiword units. Liu et al. (2025) addressed the impact of MWEs on machine translation quality. Their results showed that the presence of MWEs negatively affects translation performance. To mitigate this, they proposed an LLM-based system that paraphrases MWEs into their literal counterparts using few-shot prompting before performing the machine translation task.

Ultimately, MWE paraphrasing for a wide range of languages remains a current and significant challenge in NLP. Following the objectives of the PARSEME 2.0 shared task, this study aims to develop a methodology to improve the MWE paraphrasing performance in French sentences.

3 Methodology

We developed 4 different methods to paraphrase the expressions. Here, we describe the two that achieved the highest BERTScores during automated evaluation, since these were the only ones submitted to the official leader board (Manon Scholivet, 2025). The other two are briefly mentioned in section 3.3. We also developed a fifth (combined) approach aimed at combining the best results from each of the methods.

3.1 Paraphraser 1

This approach addresses the challenge of removing idiomaticity while preserving meaning through a multi-agent pipeline powered by open-weight Local LLMs. The system employs a **"Generate-Validate-Fix"** architecture. It first generates a candidate paraphrase, then subjects it to a rigorous hybrid evaluation consisting of deterministic string-matching constraints and a semantic LLM judge. If a candidate fails, a specialized "Fixer" agent is activated.

We benchmarked several local models (including Llama3.2:3b, Gemma3:4b, Qwen3:8b and DeepSeek-r1:7b) and selected **Qwen3:8b** for its optimal balance of reasoning capability and resource efficiency as well as its multilingual abilities (Grattafiori et al., 2024; Team et al., 2025; Yang et al., 2025; Guo et al., 2025) .

The system moves beyond simple prompting by using an agentic workflow implemented in Python using the Ollama framework(Ollama Team) and LangChain Library(LangChain Team). The architecture consists of three distinct phases.

3.1.1 Phase 1: Initial Generation

The first stage employs a standard LLM agent prompted with strict constraints. We utilize a zero-shot prompting strategy. The prompt explicitly instructs the model to replace the MWE while keeping the rest of the sentence structure as close to the original as possible.

3.1.2 Phase 2: Hybrid Validation

Prior work has demonstrated that large language models often fail to follow negated instructions (e.g., "do not use word X") or strict negative constraints (Jang et al., 2023). Although LLMs excel at judging semantic content, we observed that they may still produce exactly the expression they are instructed to avoid. Therefore, we implement a hybrid validator:

Deterministic Validator (Classical NLP) We use Python-based string manipulation to enforce the "Elimination" criterion. We normalize both the MWE and the candidate prediction (lowercasing, punctuation removal). If the target MWE tokens appear in the prediction or if the prediction is identical to the source, the candidate is immediately marked as a failure ($Score = 0$) without consuming GPU resources for semantic checking.

Semantic Judge (LLM Agent) If the deterministic checks pass, a second LLM agent evaluates the candidate. This agent uses chain-of-thought and few-shot techniques to judge the results from the previous step. It is prompted to act as a "Linguistic Evaluator" checking two specific criteria: 1. **Meaning Preservation:** Does the paraphrase convey the exact sense of the original? 2. **Grammatical Closeness:** Does the paraphrase maintain the original morphological features (tense, number)?

3.1.3 Phase 3: The Fixer

If a candidate fails validation, it is passed to a "Fixer" agent. Unlike the initial generator, the Fixer receives the specific reason for failure (e.g., *"Failure: MWE (multiword expression) tokens still present"*). The Fixer uses detailed instructions as well as the data from the previous steps to correct the errors in the paraphrased sentence.

Example Expression Corrected by the Fixer While the original sentence was (fra) *Le **point de vue** de le réalisateur* (lit. 'The point of view of the director'), the initial generator produced a grammatically and idiomatically flawed substitute: (fra) *La **vue** de le réalisateur* (lit. 'The view of the director'). This was flagged by the *Semantic Judge* since, the use of *vue* (physical sight) is a non-idiomatic calque for "opinion" in this context. The *Fixer* agent successfully resolved the issue producing: (fra) *La **vision** du réalisateur* (lit. 'The vision of the director'). This result is more correct as it selects a semantically appropriate synonym (*vision*) and restores grammatical validity.

3.1.4 Design Decisions: Explainer Tool

We also experimented with a Retrieval Augmented Generation (RAG) approach using Wikipedia definitions, but it often introduced noise such as disambiguation errors. We found that 7B+ models performed better with internal knowledge. Although, there is potential to improve the RAG approach in future work, we chose to replace it with an internal "Explainer" mechanism in the Judge Agent (Section 3.1.2).

3.1.5 Parallelization

To optimize the performance, we implemented a thread-based parallelization strategy. We enabled concurrent request processing (`OLLAMA_NUM_PARALLEL=2`) and managed VRAM usage by dynamically limiting the context window.

The next approach consists of three main phases: MWE extraction, semantic replacement generation with similarity validation, and sentence transformation. We employ a pipeline that leverages LLMs (Qwen3:8b, gpt-oss-120b (OpenAI et al., 2025)) for linguistic processing and embedding models for semantic similarity verification.

3.1.6 Phase 1: MWE Extraction

The first phase involves identifying and extracting the target MWE from the input sentence. In the provided dataset, MWEs are marked using double square brackets (e.g., `[[expression here]]`). We employ regular expression pattern matching to extract these marked expressions from the raw text. This extraction step isolates the multiword expression that will subsequently be replaced with a semantically equivalent shorter form.

3.1.7 Phase 2: Single-Word Equivalent Generation with Semantic Validation

The second phase constitutes the core of our methodology: generating a semantically appropriate single-word replacement for the extracted MWE. This phase employs an iterative refinement process guided by semantic similarity measures.

Initial Replacement Generation We prompt the LLM to generate a single-word French equivalent for the given MWE. The model is instructed through a system prompt to analyze the multiword expression and output only a single-word replacement that preserves the original meaning. Few-shot examples are provided to guide the model's behavior.

Semantic Similarity Verification To ensure that the generated replacement maintains semantic closeness to the original MWE, we employ cosine similarity as our semantic equivalence metric. This verification step is crucial for filtering out replacements that, while grammatically valid, may drift from the intended meaning of the original expression.

Cosine Similarity Computation. We compute the cosine similarity between two vectors as:

$$\text{sim}(\mathbf{e}_{\text{mwe}}, \mathbf{e}_{\text{rep}}) = \frac{\mathbf{e}_{\text{mwe}} \cdot \mathbf{e}_{\text{rep}}}{\|\mathbf{e}_{\text{mwe}}\| \times \|\mathbf{e}_{\text{rep}}\|} \quad (1)$$

This score ranges from -1 to 1, with higher values indicating greater similarity.

Threshold-Based Validation. We establish an empirically determined similarity threshold of $\tau = 0.50$ to determine whether a proposed replacement is semantically acceptable. If $\text{sim}(\mathbf{e}_{\text{mwe}}, \mathbf{e}_{\text{rep}}) \geq \tau$, the replacement is accepted. Otherwise, the system initiates the iterative refinement process described in the following paragraph. This threshold was chosen empirically, taking into account the typical length of a multi-word expression relative to the surrounding sentence and the fact that BERTScore will be used for evaluation. Since BERTScore compares predictions to both a "minimal" (close to the source) and a "creative" (more divergent) reference and selects the closest match, using cosine similarity in our system primarily ensures that replacements stay close to the minimal reference, with iterative refinement applied only when similarity falls below the threshold.

Table 1: Automated & Manual evaluation scores for PARSEME 2.0 Subtask 2 French Test Set.

Approach	Ave. BERTScore	Manual Score	Diversity of New Words		
			Richness	Evenness	Entropy
Paraphraser 2	**93.90**	64.82	236	0.83	4.54
Paraphraser 1	89.46	**79.25**	**456**	0.90	5.48
Paraphraser 4	88.76	–	345	0.91	5.33
Combined	87.10	–	415	0.94	5.64
Paraphraser 3	85.72	–	399	**0.95**	**5.72**
Baseline	77.55	72.70	326	0.92	5.33

Note: The baseline model is gpt-oss-120b. Evenness and Entropy refer to Shannon Evenness and Shannon-Weaver Entropy, respectively. Manual evaluation was omitted for three approaches as they were not officially submitted to the leaderboard.

Iterative Refinement Process If the initial similarity score falls below the threshold, we initiate an iterative refinement process. The system maintains a blacklist of previously rejected candidates along with their similarity scores. In subsequent iterations, the LLM is prompted to generate alternative replacements while explicitly excluding blacklisted expressions. This process continues for a maximum of three iterations. Throughout this process, we track the best candidate encountered (i.e., the one with the highest similarity score). If no candidate exceeds the threshold after all iterations, we select the best candidate observed during the refinement process.

3.1.8 Phase 3: Sentence Transformation

The final phase transforms the original sentence by substituting the MWE with the validated single-word equivalent. We leverage the LLM, providing it with:

- The original sentence containing the MWE
- The identified multiword expression
- The validated single-word replacement

The model is instructed to perform the substitution while maintaining grammatical correctness. The prompt explicitly requests only the transformed sentence without additional explanations or formatting.

3.2 Combined Approach

Paraphrasers 3 and 4 utilize different strategies for paraphrasing. **Approach 3** is a multi-stage LLM-based pipeline that combines prompt engineering, POS-based constraints, and rule-based validation to generate and score paraphrase candidates. **Approach 4** integrates morpho-syntactic analysis and semantic enrichment with LLM generation and post-processing to preserve grammatical structure and figurative meaning.

In addition to these methods, we implemented a combined approach that aimed to leverage the strengths of each strategy. This approach used an LLM (gpt-oss-120b) as a selector to choose the optimal output from the candidates generated by the other methods. However, despite this, the combined approach performed worse, revealing some biases in the selection process (such as favoring or punishing approaches, depending on the order in which they are presented in the prompt). This suggests that the selection mechanism has room for improvement.

4 Results

We evaluated our system using the official PARSEME 2.0 metrics alongside a manual evaluation by language experts. The automated metrics include BERTScore (Zhang et al., 2019), Richness, Shannon Evenness, and Shannon-Weaver Entropy for diversity.

As we can see in Table 1, all of our experimental alternatives exceeded the BERTScore baseline of 77.55. On the official PARSEME 2.0 shared task leaderboard for French, our Paraphraser 2 approach (93.90) ranked first among all participants in the automated evaluation, while our Paraphraser 1 approach (79.25) secured first place overall in the manual scoring track. This divergence between automated and manual success is one of the key findings of our study, highlighting the importance of metric sensitivity in MWE paraphrasing.

Qualitative Analysis and Strategy Divergence Through manual inspection, we found that this divergence is rooted in the distinct paraphrasing strategies employed. Paraphraser 2 (Minimal) ex-

cels at lexical compression and identifying direct synonyms, which yields a high BERTScore due to its proximity to the source text. However, despite utilizing a much larger 120B parameter backbone ($gpt - oss - 120$), it is prone to contextual overgeneralization, for instance, reducing the technical NMWE *système d'information* (lit.'information system') to the overly broad *informatique* (lit.'IT').

In contrast, Paraphraser 1 (Explanatory) utilizes a multi-agent pipeline powered by a smaller 8B model ($Qwen3$) to generate descriptive, context-aware expansions. While this strategy results in a slightly lower BERTScore and higher verbosity, manual evaluation confirms it is superior in the majority of the complex cases. It preserves domain-specific precision and idiomaticity (e.g., *année dédiée à l'international* (lit.'the international dedicated year')) where the Minimal approach often produces literal calques or incorrect substitutions.

These empirical findings, further illustrated via comparative traces in Appendices B–E, suggest that the Explanatory approach is better suited for maintaining the semantic non-compositionality inherent in complex MWEs. Although our initial attempts at a combined ensemble did not yield immediate performance gains, the complementary nature of these strategies(lexical brevity versus semantic precision) suggests that a more refined categorical selection process based on MWE type remains a promising area for future optimization.

5 Conclusion

We presented several strategies for the PARSEME 2.0 French paraphrasing task, each integrating external validation—such as string matching and rule-based filtering—to overcome the limitations of LLMs in adhering to strict linguistic constraints. This hybrid approach, merging generative power with symbolic verification, proved essential for maintaining semantic precision and output quality.

Paraphraser 2 ranked first in automated scoring (93.90), yet Paraphraser 1 secured first place in the manual track (79.25). This "metric sensitivity gap" indicates that while automated scores favor lexical compression, human experts prioritize the semantic precision and explanatory depth necessary for non-compositional MWEs. The superior manual performance of the 8B approach over the 120B model suggests that targeted multi-agent architectures outperform raw parameter scale for complex linguistic tasks.

Although our evaluation focused on French, the majority of our pipeline components are language-agnostic, suggesting potential applicability to other languages addressed by the PARSEME 2.0 shared task. Future work could explore the cross-lingual transfer of these methods.

Discussion

Recent mechanistic studies have identified an "ironic rebound" effect in LLMs, where explicitly naming a forbidden word paradoxically primes the model to generate that very token (Mann et al., 2025). This phenomenon (where the instruction's activation of a concept is systematically stronger than its suppression signal) is a primary cause of negative constraint failure (Rana, 2026). While our methodology naming the target MWE within prompt delimiters (e.g., [[...]]) theoretically risks this priming, our system mitigates this through external string-matching constraints and iterative verification.

Furthermore, the task of paraphrasing itself may offer a **generalizable solution to the ironic rebound problem**. Future work could explore whether paraphrasing architectures such as ours could provide a generalizable solution to the ironic rebound problem. By transforming a negative constraint into a positive generative goal (a "generative pivot"), hybrid pipelines that combine neural generation with classical NLP methods may offer a robust pathway for bypassing the internal priming failures inherent in current LLM architectures.

Limitations

The combined approach underperformed due to biases in the LLM selection process, affecting its ability to consistently choose the best paraphrase. Additionally, the LLM-based multi-agent framework, unlike traditional rule-based NLP methods, can struggle with stability and precision, as it relies on the quality of underlying models, which may introduce errors or inconsistencies.

Acknowledgments

The authors used AI-assisted editing tools, to improve spelling, grammar, clarity, and readability during part of the manuscript preparation. After using the tools, the authors carefully reviewed and edited the text and take full responsibility for the final content.

References

Petra Barančíková and Václava Kettnerová. 2018. Paraphrases of verbal multiword expressions: The case of czech light verbs and idioms. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, volume 2, page 35. Language Science Press.

Anabela Barreiro and Cristina Mota. 2023. A multilingual paraphrasary of multiwords. In *Proceedings of the 1st International Workshop on Multilingual, Multimodal and Multitask Language Generation*, pages 47–56.

Matthieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: multiword expression processing: a survey. *Computational Linguistics*, 43(4):837–892.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*, volume 203 of *Proceedings of Machine Learning Research*, pages 52–62. PMLR.

LangChain Team. Langchain. `https://www.langchain.com/`. Framework for developing applications powered by large language models.

Linfeng Liu, Saptarshi Ghosh, and Tianyu Jiang. 2025. Evaluating the impact of verbal multiword expressions on machine translation. *arXiv preprint arXiv:2508.17458*.

Logan Mann, Nayan Saxena, Sarah Tandon, Chenhao Sun, Savar Toteja, and Kevin Zhu. 2025. Don't think of the white bear: Ironic negation in transformer models under cognitive load. *ArXiv*, abs/2511.12381.

Agata Savary Eric Bilinski Carlos Ramisch Manon Scholivet, Takuya Nakamura. 2025. PARSEME 2.0: Shared task on idiomaticity and multiword expressions.

Ollama Team. Ollama. `https://ollama.com/`. Local inference framework for large language models.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b and gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Shailesh Rana. 2026. Semantic gravity wells: Why negative constraints backfire. *Preprint*, arXiv:2601.08070.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *International conference on intelligent text processing and computational linguistics*, pages 1–15. Springer.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and 1 others. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th workshop on multiword expressions (MWE 2017)*, pages 31–47.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Pais. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Aline Villavicencio and Marco Idiart. 2019. Discovering multiword expressions. *Natural Language Engineering*, 25(6):715–733.

Takashi Wada, Yuji Matsumoto, Timothy Baldwin, and Jey Han Lau. 2023. Unsupervised paraphrasing of multiword expressions. *arXiv preprint arXiv:2306.01443*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Seid Muhie Yimam, Héctor Martínez Alonso, Martin Riedl, and Chris Biemann. 2016. Learning paraphrasing for multi-word expressions. In *MWE 2016-Multiword Expression Workshop 2016*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2022. Idiomatic expression paraphrasing without strong supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11774–11782.

A System Trace Example

The following trace demonstrates Paraphraser 1 ("Generate-Validate-Fix") resolving a combined lexical and grammatical failure.

Original Sentence: (fra) *Le **point de vue** de le réalisateur*
(lit. 'The point of view of the director')
'The director's perspective'.

Initial Prediction: (fra) *La **vue** de le réalisateur*
(lit. 'The view of the director').

Validator Feedback: *"Failure: The prediction contains a grammatical error in 'de le réalisateur' which should be 'du réalisateur'. Additionally, 'vue' is non-idiomatic for this context."*

Final Prediction: (fra) *La **vision** du réalisateur*
(lit. 'The vision of the director')
'The director's vision'.

B Comparative Trace: Expansion vs. Overgeneralization

This example illustrates the risk of overgeneralization in the minimalist approach (Approach 2) compared to the context-aware generation of Approach 1.

Original Sentence: (fra) *Malone L' année 1996 a été... d' **année internationale** pour...*
(lit. '...of the international year for...')

Paraphraser 1 (Explanatory): (fra) *...d'**année dédiée à l'international** pour...* (lit. '...of the international dedicated year for...')
Verdict: Passed first try. The agent expanded the MWE into an idiomatic descriptive phrase that preserves the UN's formal register.

Paraphraser 2 (Minimal): (fra) *...d'**Internationalisation** pour...*
(lit. '...of internationalization for...')
Verdict: Semantic failure. *Internationalisation* is a process, not a designation for a specific year, leading to a loss of core meaning.

C Comparative Trace: Stylistic Convergence

This example demonstrates a case where both the Minimalist and Explanatory approaches produce semantically accurate and idiomatic results.

Original Sentence: (fra) *...avec comme **personnage principal** le meilleur matador...*
(lit. '...with as main character the best matador...')

Paraphraser 1 (Explanatory): (fra) *...avec comme **figure centrale** le meilleur matador...*
(lit. '...with the best matador as the central figure...')
Verdict: High semantic similarity. The choice of *figure centrale* is idiomatic and provides a smooth, literary flow.

Paraphraser 2 (Minimal): (fra) *...avec comme **protagoniste** le meilleur matador...*
(lit. '...with as protagonist the best matador...')
Verdict: High semantic similarity. This shows the strength of Paraphraser 2's iterative search when a direct, single-word synonym is available. Note: capitalization was manually corrected from the system output.

D Comparative Trace: Granularity and Specificity

This case illustrates the difference between descriptive expansion and collective abstraction.

Original Sentence: (fra) *...issus de les **familles nobles** de la ville.*
(lit. '...descended from the noble families of the city.')

Paraphraser 1 (Explanatory): (fra) *...issus des **familles de la noblesse** de la ville.* (lit. '...descended from the noble families (or families of the nobility) of the city.')
Verdict: High similarity. It maintains the plural focus on family units.

Paraphraser 2 (Minimal): (fra) *...issus de **la Noblesse** de la ville.*
(lit. '...descended from the Nobility of the

city.')
Verdict: Highly idiomatic but less specific. By substituting a collective noun for a plural MWE, it shifts the focus from individual families to the social class as a whole.

E Comparative Trace: Technical and Domain-Specific Precision

This trace demonstrates the failure of minimalist constraints when handling specialized terminology.

Original Sentence: (fra) *Le Forbin a été équipé d' un **système d' information**...*
(lit. 'The Forbin was equipped with an information system...')

Paraphraser 1 (Explanatory): (fra) *...équipé d'un **système de gestion d'information**...* (lit. '...equipped with an information management system...')
Verdict: High semantic closeness. While slightly more descriptive than the original, it preserves the technical and operational register required for the maritime context.

Paraphraser 2 (Minimal): (fra) *...équipé d'**Informatique**...*
(lit. '...equipped with IT...')
Verdict: Low semantic closeness. *Informatique* is too broad; it fails to capture the notion of a specific operational management system, rendering the sentence vague and non-idiomatic.

F Prompt Templates

This appendix provides the prompt templates used for Methods 1 and 2.

Paraphraser 1

Paraphraser 1 utilizes a multi-agent approach to ensure semantic depth and grammatical correctness through a feedback loop.

`generation_template`

```
"""
You are a linguistic expert.
Task: Paraphrase the sentence to remove the
    idiomatic nature of the specified
    expression.

Input Sentence: "{sentence}"
Expression to replace: "{mwe}"

Instructions:
1. Replace the expression '{mwe}' with a
    literal, non-idiomatic equivalent.
2. Keep the rest of the sentence structure as
    close to the original as possible.
3. Ensure the output is in the SAME LANGUAGE as
     the input.
4. Output ONLY the new sentence text with no
    introduction or quotes.
"""
```

`judge_template`

```
"""
You are a Linguistic Evaluator.
Task: Evaluate a Paraphrase based on strict
    criteria.

Original: "{original}"
Prediction: "{prediction}"
MWE Removed: "{mwe}"

Evaluate against these criteria:
1. MEANING: What is the meaning of the
    multiword expression (MWE) in the sentence?
     Is the sense preserved?
2. GRAMMATICALITY: Is the prediction
    grammatically correct (no spelling/grammar
    errors)?
3. GRAMMATICAL CLOSENESS: Does it keep the
    exact Tense, Mood, and Number of the
    original?
   - Example Fail: Original is "He was walking"
     (Past Continuous), Prediction is "He walks
    " (Present). -> FAIL.

Respond in JSON only.
{format_instructions}
"""
```

`fixer_template`

```
"""
You are a Linguistic Correction Expert.
A system attempted to paraphrase a sentence but
     failed evaluation.

Data:
- Original Sentence: "{original}"
- MWE to Remove: "{mwe}"
- Failed Prediction: "{prediction}"
- Failure Reason: "{reason}"

Instructions:
1. Write a NEW paraphrase.
2. ELIMINATION: Do NOT use the MWE "{mwe}".
3. MEANING: Keep the exact meaning.
4. GRAMMATICAL CLOSENESS: You MUST match the
    Tense, Mood, and Number of the original
    sentence.
5. You MUST preserve the exact meaning of the
    original
6. Output ONLY the new sentence.
7. Original sentence MUST BE PRESERVED other
    than PARAPHRASING of the indicated MWE (
    multi word expression).
"""
```

Paraphraser 2

Paraphraser 2 focuses on high-precision lexical substitution using a larger model and an iterative blacklist mechanism.

FRENCH_MWE_TRANSLATION

```
"""
You are a tool that converts French multiword
    expressions into single-word equivalents.
    Briefly analyze the multiword expression
and convert them into a single word. You should
     ONLY OUTPUT THE NEW EXPRESSION. NOTHING
    ELSE. Here are some examples:

Input: Gravir les chelons
Output: Progresser

Input: Faire attention
Output: Surveiller

DO NOT RETURN THE INPUT ITSELF! YOU SHOULD GIVE
     A NEW EXPRESSION THAT HAS THE SAME MEANING
     AS THE ORIGINAL ONE. IT SHOULD BE A SINGLE
     WORD
"""
```

new_mwe_prompt

```
def new_mwe_prompt(blacklist: dict):
    blacklist_str = ""

    for word, similarity in blacklist.items():
        blacklist_str += f"\n- {word}"

    return f"""
    You are a tool that converts French
     multiword expressions into single-word
     equivalents. Briefly analyze the multiword
     expression
    and convert them into a single word
     expression. You should ONLY OUTPUT THE NEW
     EXPRESSION. NOTHING ELSE. Here are some
     examples:

    Input: Gravir les chelons
    Output: Progresser

    Input: Faire attention
    Output: Surveiller

    DO NOT RETURN THESE EXPRESSIONS:
    {blacklist_str}
    """
```

minimal_sentence_prompt

```
def minimal_sentence_prompt(sentence, mwe,
    new_exp):
    return f"""
You are a French text transformation tool. Your
     task is to replace a multiword expression
    in a sentence with a new expression.

INSTRUCTIONS:
- Replace the specified multiword expression
    with the new expression
- Keep the grammer correctly.
- Output ONLY the transformed sentence
- Do not include explanations, quotes, or
    additional text

EXAMPLE:
Sentence: Les lves de la classe de CE1 de l'
    cole Notre-Dame ont pris part  les
    festivits.
Multiword expression: pris part
New expression: particip
Output: Les lves de la classe de CE1 de l'cole
    Notre-Dame ont particip aux festivits.

YOUR TASK:
Sentence: {sentence}
Multiword expression: {mwe}
New expression: {new_exp}

JUST OUTPUT THE TRANSFORMED SENTENCE ONLY:
"""
```