MorphoFiltered-Gemini at MWE-2026 PARSEME 2.0 Subtask 1: Tackling LLM Overgeneration via Universal POS-based Constraints

Irina Moise* **and Sergiu Nisioi***
Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest
moiseirina42@gmail.com
sergiu.nisioi@unibuc.ro

Abstract

This paper describes **MorphoFiltered-Gemini**, a system submitted to the PARSEME 2.0 Shared Task (Scholivet et al., 2025), subtask 1 on MWE identification, covering all 17 target languages. The system combines LLM-based predictions generated via the Gemini API with a morphological post-filter designed to reduce false positives. Rather than optimizing peak performance on individual languages, our approach prioritizes cross-lingual stability and precision. As a result, the system exhibits a balanced performance across languages and MWE categories, achieving the highest Shannon evenness score among all submissions.

1 Introduction

Multiword expressions (MWEs) are "word combinations that exhibit lexical, syntactic, semantic, pragmatic, and/or statistical idiosyncrasies" (Baldwin and Kim, 2010) . For example, *a big fish* refers to an important person. Similar phenomena appear across languages, including light verb constructions (to *grant rights*), adjectival idioms (to be *on cloud nine*), fixed adpositional phrases (*on behalf of*), pronoun idioms (*each other*) and many others. The constituent words of multi-word expressions are common, and their combination behaves as a single unit.

PARSEME 2.0 addresses the automatic identification of MWEs in running text in a multilingual setting. Unlike previous PARSEME shared tasks, which focused only on verbal MWEs, this edition (Savary and Ramisch, 2025) extends the task to all syntactic types (verbal; nominal; adjectival and adverbial; functional). Systems must identify MWEs across 17 languages: Dutch, Egyptian (ca. 2700-2000 BC), French, Georgian, Greek (Ancient), Greek (Modern), Hebrew, Japanese, Latvian, Persian, Polish, Brazilian Portuguese, Romanian, Serbian, Slovene, Swedish, and Ukrainian.

MWEs may be continuous or discontinuous, and they can overlap, further complicating automatic detection. Systems are evaluated both at the token level and at the MWE level, where a correct prediction requires identifying all components of an expression. As a result, partial matches are penalized, making the task sensitive to recall and span boundary errors.

While morphologically constrained approaches often achieve high precision, they typically suffer from poor recall and limited generalization. In contrast, large language models generalize well but tend to accept too many candidates, leading to many false positives.

In this work, we explore a hybrid approach that combines the strengths of both paradigms. MorphoFiltered-Gemini[1] relies on Gemini 2.0 Flash-Lite (DeepMind, 2025) to generate MWE predictions using prompting, followed by morphological post-processing, applied selectively. The goal is not to maximize performance on specific languages, but to study whether minimal linguistic constraints can stabilize LLM behavior.

Our contributions are threefold: (i) we propose a LLM-based pipeline for multilingual MWE identification, (ii) we investigate the effects of few-shot prompting across diverse languages, and (iii) we introduce a morphological filtering strategy that improves precision without relying on language-specific training.

2 Experiments

We evaluated several strategies to understand the trade-offs between recall-oriented LLM predictions and precision-oriented filtering.

As an initial baseline, we implemented a purely rule-based system relying on morphological patterns. Although this approach achieved a token-based F1 of 0.152 and performed reasonably

*Corresponding authors.

[1] https://github.com/irinamoise/PARSEME

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 196–202

well for a small number of languages (e.g. Romanian and Persian), it failed to generalize, yielding extremely low scores for others such as Georgian. This confirmed the poor recall of rule-based methods in a multilingual setting.

We then applied Gemini as a post-processing validator on those candidates. Although the LLM consistently rejected around 30% of proposed MWEs, this strategy led to negligible improvements. Due to the very low recall of the rule-based system, only a small fraction of true MWEs were ever presented to the LLM, making it impossible to recover missing predictions.

Subsequently, Gemini was adopted as the primary predictor. Using separate prompts for detection and classification improved recall. Few-shot prompting with examples from the training sets and negative constraints yielded substantial gains for some languages, but caused severe degradation for others, including complete failure for Egyptian.

Finally, we introduced a lightweight morphological post-filter applied to LLM-generated predictions. This filter consistently lowers the number of false positives and leads to more stable performance across languages.

3 System Architecture and Methodology

The overall architecture of our system is illustrated in Figure 1 (see Appendix A).

The system follows a five-stage pipeline:

1. Preprocessing: Conversion of PARSEME CUPT annotations to BIO tags.
2. MWE Detection: Span prediction using Gemini 2.0 Flash Lite via batch prompting.
3. MWE Classification: Assignment of MWE categories to detected spans.
4. Morphological Post-Processing: Removal of unlikely or low-confidence MWEs using POS-based and single-token filters.
5. Format Conversion: Reconstruction of CUPT-formatted outputs from BIO predictions.

3.1 Preprocessing and BIO Representation

PARSEME annotations are provided in the CUPT format (Ramisch, 2018), an extension of Universal Dependencies (UD) (Nivre, 2020). While CUPT offers rich annotation capabilities, it is not directly suitable for LLM-based processing. We therefore convert annotations into standard BIO tags, which provide a simplified, token-level representation of MWE spans. This conversion simplifies span reconstruction from LLM outputs.

A limitation of this representation is that it does not adequately capture discontinuous MWEs or cases where tokens participate in multiple overlapping MWEs. As a consequence, such expressions are often misinterpreted or collapsed into incomplete spans during BIO-based processing, which leads to zero scores for discontinuous MWEs in the official evaluation. These structural conflicts, including overlapping and discontinuous configurations, are not yet resolved in the current system.

3.2 LLM-based MWE Detection and Classification

Batch Prompting for Detection MWE detection is performed using Gemini 2.0 Flash-Lite in batches of ten sentences. The prompt includes a general definition of MWEs, language-specific examples, and token-indexed sentences to ensure span identification (see appendix B.1). For a subset of languages (EL, FA, FR, SV), the prompts uses an improved strategy: the dictionary exemples are substituted by 10 full phrases with MWE examples extracted from the training data. An additional prompt containing negative constraints (anti-examples) explicitly defines undesirable outputs, thereby narrowing the solution space and ensuring the model adheres to specific stylistic and structural boundaries.

The output format is strictly constrained to token indices corresponding to predicted MWE spans, or a special NONE marker when no MWE is detected.

MWE Category Classification Detected MWEs are passed to a second prompt that assigns MWE categories based on generic category descriptions (see appendix B.2). Separating detection and classification makes the model focus on span identification independently of category semantics, which improved the results.

Few-Shot Prompting Strategies We experimented with few-shot prompting (Brown et al., 2020) using 4-5 examples of MWEs from a small dictionary and with full sentences from the training sets that contained MWEs (see appendix B.3). While the second strategy produced good results for a few languages, it led to severe degradation for others. These observations motivated the final

system design, which applies the last method selectively rather than uniformly across all languages.

3.3 Morphological Post-Processing

POS Pattern Filtering LLM predictions frequently include false positives. To adjust this effect, we applied a morphological filter that removes unlikely POS patterns, such as (DET, NOUN) or (ADJ, NOUN), while preserving high-precision constructions such as (VERB, NOUN), (ADP, NOUN). See appendix C and D.

This filter is applied to 11 languages (EGY, EL, FA, FR, GRC, NL, PL, RO, SR, SV, UK) for which empirical evaluation showed consistent precision gains (see appendix F).

Single-Token MWE Filtering Gemini occasionally predicts single tokens as MWEs, despite PARSEME annotations treating most single-token expressions as non-MWEs. To address this issue, we introduced a single-token filter that removes isolated MWE predictions (see appendix E). This filter proved beneficial for EL, FA, GRC, KA, and LV, and was therefore retained for these languages in the final system.

Language-Specific Exclusions For Hebrew, Japanese, and Slovene, morphological filtering consistently degraded performance. These languages were therefore excluded from postprocessing. This suggests that POS-based constraints are less reliable for languages with complex morphology, logographic writing systems, or ambiguous UD tagsets.

3.4 Caching and Seen/Unseen Analysis

To reduce redundant LLM calls, the system implements a caching mechanism that stores previously predicted MWE spans. This mechanism primarily benefits MWEs that appear multiple times across training and development data.

Evaluation results confirm a substantial performance gap between MWEs seen during training/development and unseen expressions. For MWEs identical to those in the training/development data, the system achieves high precision (P=75.06, R=15.91, F1=26.25). Variant forms of seen MWEs also benefit from caching and contextual similarity (P=54.54, R=12.00, F1=19.67). In contrast, unseen MWEs exhibit much lower scores (P=8.59, R=9.21, F1=8.89). When aggregating all seen-in-traindev MWEs, performance remains substantially higher (P=73.08, R=14.86, F1=24.70), highlighting the role of memorization and context reuse in LLM-based systems.

4 Results

Overall Performance In the official evaluation (Ramisch, 2025), **MorphoFiltered-Gemini** was submitted for all 17 languages. The system did not achieve top overall F1 scores, but it ranks third in token-based precision and exhibited a clear gap between token-based and MWE-based performance, as shown in Table 1. This discrepancy reflects the conservative behavior induced by morphological filtering and the strict nature of MWE-level evaluation, where partially correct predictions are penalized.

Token-Based vs MWE-Based Evaluation High token-based precision indicates accurate boundary detection for individual tokens, even when full MWEs are not perfectly reconstructed. In contrast, MWE-based evaluation requires exact span matches, penalizing conservative systems and those unable to represent discontinuous expressions.

Diversity Metrics The system achieves the highest Shannon evenness score among all submissions, as shown in Table 2, indicating balanced performance across languages and MWE categories. Unlike systems exhibiting strong performance peaks for a small subset of languages, **MorphoFiltered-Gemini** avoids extreme failures and maintains a stable cross-lingual behavior.

4.1 Error Samples

We discuss a few prediction errors in French, Portuguese and Romanian with the goal of presenting the current limitations of our system.

In French, the system frequently misses light verb constructions (LVC.full) when they are discontinuous. For example, in *"Éric Halphen reçoit à son cabinet un coup de fil anonyme"*(Éric Halphen receives an anonymous phone call at his office), the MWE *reçoit ... coup de fil* is not detected because the verb and the noun are separated by other words. Similar issues occur for adverbial idioms (AdvID) like *avec vigueur* in *"avec une vigueur accrue"*(with increased vigor), which are often interpreted as free prepositional phrases.

False positives mostly involve grammatical constructions that resemble MWEs but are actually compositional. In Romanian, a peposition and an infinitive marker *de a* are incorrectly predicted as

System	#Langs	Global MWE-based				Global Token-based			
		P	R	F1	Rank	P	R	F1	Rank
MorphoFiltered-Gemini	17/17	20.95	14.50	17.14	7	34.14	24.20	28.32	5

Table 1: General Ranking

System	#Langs	Richness		Shannon-Weaver Entropy		Shannon-Evenness	
		Value	Rank	Value	Rank	Value	Rank
MorphoFiltered-Gemini	17/17	56.53	7	3.71	5	0.97	1

Table 2: Diversity Ranking

an AdpID in *"șansa de a vedea clar"* (the chance to see clearly), although it is just a syntactic pattern. In Portuguese, standard contractions such as *de o* (of the) and *em o* (usually *in the*, here meaning *at the*) are over-generated as MWEs , as in *"No final do quarto ano"* (At the end of the fourth year), even though they reflect regular morphology rather than idiomatic usage. All in all, false positives tend to arise from surface patterns that look fixed, while false negatives are mainly caused by discontinuity, reflexive clitics, and complex adpositional structures.

5 Conclusion

The results suggest that combining general-purpose LLM predictions with minimal linguistic post-processing yields balanced evaluation outcomes, but with limitations. Future work includes extending the representation to support discontinuous MWEs and exploring adaptive filtering strategies that better balance recall and precision across languages.

Acknowledgements

This work was supported by the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology), and by the Romanian National Research Council (CNCS) through the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI) under grant PN-IV-P2-2.1-TE-2023-2007 InstRead.

References

Timothy Baldwin and Su Nam Kim. 2010. *Multiword Expressions*. Taylor and Francis.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Google DeepMind. 2025. Gemini 2.0 flash-lite: Cost-efficient multimodal reasoning. Google Developers Blog. Released February 5, 2025.

Joakim Nivre. 2020. Universal Dependencies v2: An Evergreen Corpus for Cormpus-based Linguistics and NLP. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043.

Carlos Ramisch. 2018. The CUPT format for MWE annotation in Universal Dependencies. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Information Extraction (LAW-MWE-2018)*, pages 220–231.

Carlos Ramisch. 2025. PARSEME 2.0 shared task subtask 1: Detailed results. https://gitlab.com/parseme/sharedtask-data/-/blob/master/2.0/subtask1/Detailed_results.md.

Agata Savary and Carlos Ramisch. 2025. PARSEME 2.0 shared task guidelines.

Manon Scholivet, Takuya Nakamura, Agata Savary, Éric Bilinski, and Carlos Ramisch. 2025. Parseme 2.0 shared task on identification and paraphrasing of multiword expressions. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*.

A System Pipeline

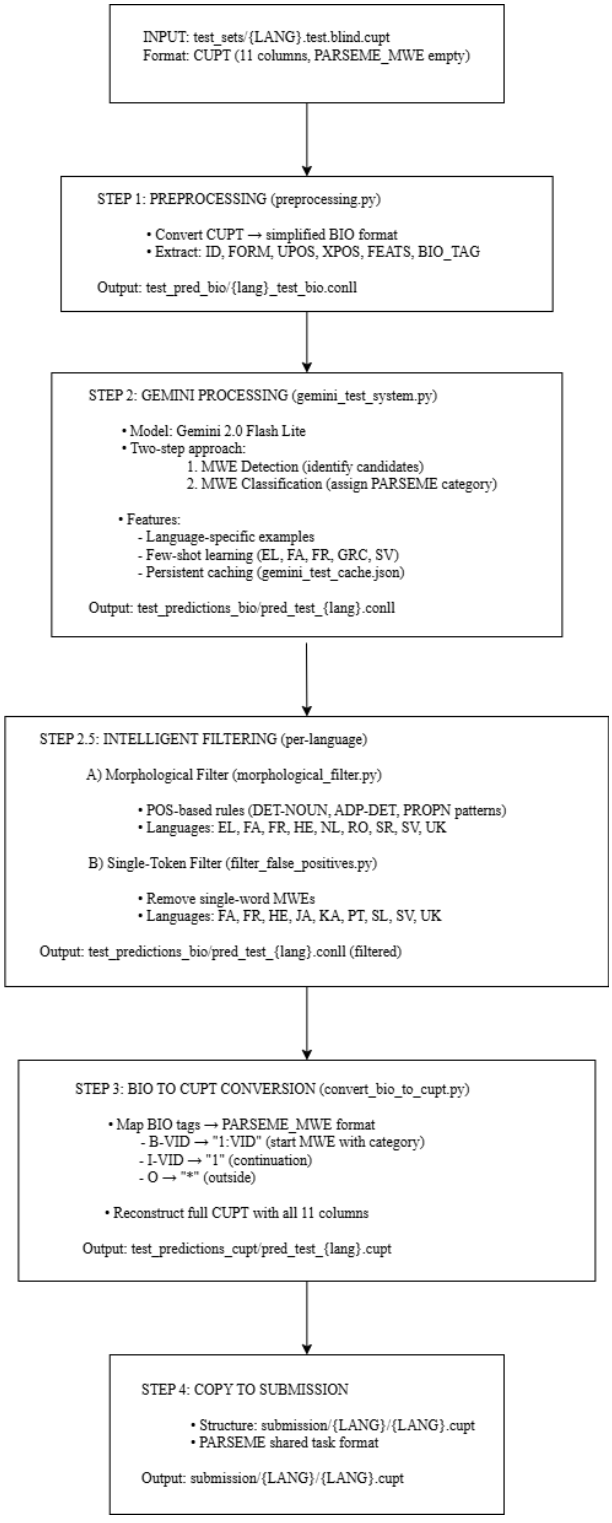

Figure 1: Detailed MWE Prediction Pipeline

B Prompting

B.1 Detailed MWE Detection Prompt

You are an expert annotator for the PARSEME shared task on multiword expression (MWE) identification. {strategy_note}

Task: Extract multiword expressions (MWEs) from sentences in {lang_info['name']}.

A MWE is a sequence of 2+ tokens that form a linguistic unit. Include:

- verbal MWEs
- nominal MWEs
- adjectival and adverbial MWEs
- functional MWEs

{few_shot_text}
{negative_examples}

CRITICAL SPAN RULES (follow exactly):

1. *Be MINIMAL: prefer SHORTER spans over longer ones.*
2. *For IAV: ONLY verb + preposition, NEVER include objects.*
   CORRECT: "depend on" | WRONG: "depend on him"
3. *For VPC: ONLY verb + particle, NEVER include objects.*
   CORRECT: "give up" | WRONG: "give up hope"
4. *DROP trailing punctuation ALWAYS.*
   CORRECT: "look up" | WRONG: "look up ."
5. *Use EXACT surface forms from input (copy verbatim).*

Format: [N]: MWE1 | MWE2 | NONE
Now analyze these NEW sentences:

Data Serialization and Batching Strategy: Sentences are processed in batches of 10 to ensure contextual consistency. Each sentence is serialized as a space-separated string of tokens, prefixed by a unique numerical identifier `[i]`. Then, the prompt concludes with a structural constraint:*"Output (one line per sentence, list only MWE surface forms separated by |)"*.

B.2 MWE Category Classification Prompt

Classify the category of this multiword expression (MWE) in {language}.

MWE: "{mwe_text}"

Categories (choose the MOST specific match):

VERBAL CATEGORIES (verb is present):

- VID: Verbal Idiom - idiomatic/figurative meaning
- LVC.full: Full Light Verb Construction - verb has little meaning, noun carries semantics
- LVC.cause: Causative Light Verb - causes state/action
- IAV: Inherently Adpositional Verb - verb REQUIRES preposition/particle (VERB+PREP only, 2-3 tokens)
- IRV: Inherently Reflexive Verb
- VPC.full/cause/semi: Verb-Particle Constructions
- MVC: Multi-Verb Construction

NON-VERBAL ID CATEGORIES (no verb):

- NID: Nominal Idiom

- AdvID: Adverbial Idiom (HINT: If single compound word functioning as adverb → AdvID)
- AdjID: Adjectival Idiom

FUNCTIONAL MWES:

- AdpID: Prepositional Idiom
- DetID: Determiner Idiom
- ConjID
- PronID, IntjID: Other idiom types

COMPOSITE (mix of categories):

- NV.VID: Nominal + Verbal Idiom
- AV.: Adverbial + Verbal constructions*

Respond with ONLY the category code (e.g., "VID", "LVC.full", "IAV", "AdvID"). Use lowercase for compound categories: "LVC.full" not "LVC.FULL".

B.3 Contextual Guidance: Few-Shot and Language Examples

To stabilize the LLM's predictions across diverse languages, the system injects language-specific examples and negative constraints into the prompt.

B.3.1 Positive Few-Shot Examples from Training Sets

For selected languages (SV, EL, FA, FR), the prompt includes full-sentence examples to illustrate span boundaries and category assignments. Examples:

Swedish (SV):
Sentence: *Disibodenbergklostret upplöstes och förföll i ruiner till följd av reformationen .*
MWEs: ***upplöstes | till följd av***
Categories: VPC.semi, AdpID

Portuguese (PT):*
Sentence: *A relatoria caiu com o ministro Gilmar Mendes , por meio de sorteio eletrônico .*
MWEs: ***por meio de***
Categories: AdpID
**Note: While PT did not use the full "improved" strategy in the final submission, it served as a development baseline.*

B.3.2 Negative Constraints (Anti-Examples)

To combat over-generation and the LLM's tendency to label compositional phrases as MWEs, the following explicit exclusions are included in the prompt:

IMPORTANT - NOT MWEs (do NOT extract these):

Simple noun phrases: "the book", "big house", "my friend"
Adjective + noun (compositional): "red car", "happy person"
Verb + full object: "read the book", "eat an apple"
Preposition + full noun phrase: "in the morning", "on the table"
Determiner + noun: "a cat", "the dog"

B.3.3 Language-Specific Prototype Examples

A small dictionary with a few examples for almost every language was created. For languages where full few-shot sentences were not used, the system provides prototypical MWE examples from that dictionary:

Dutch (NL):
"plaats vinden" (to take place) – light verb construction
"van tevoren" (beforehand) – fixed adverbial
"in orde" (in order) – fixed expression
"op de hoogte" (informed) – idiomatic phrase

Romanian (RO):
"de asemenea" (also/furthermore) – fixed adverbial
"în sfârșit" (finally) – temporal idiomatic expression
"cu toate că" (although) – compound conjunction
"a avea loc" (to take place) – verb with participle
"pe de altă parte" (on the other hand) – adverbial phrase

C Universal POS Filtering Patterns

The following POS-based constraints are applied to the LLM outputs to eliminate sequences that are unlikely to be Multiword Expressions. These rules prioritize precision by filtering common compositional or functional patterns:

`(DET, NOUN / ADJ / VERB):` Filters standard noun phrases and nominalized adjectives (e.g., *"the house"*).

`(ADJ, NOUN / NOUN, ADJ):` Eliminates simple compositional adjective-noun combinations unless previously seen as lexicalized.

`(DET, NUM / NUM, DET):` Removes articles combined with cardinal numbers.

`(PRON, VERB / VERB, PRON):` Filters subject-verb or verb-object pairs lacking idiomatic or reflexive properties.

`(CCONJ, NOUN / VERB):` Prunes spans starting or ending with coordinating conjunctions.

`(PUNCT, ANY / ANY, PUNCT):` Removes spans containing leading or trailing punctuation noise.

`(AUX, NOUN / ADJ):` Eliminates copular constructions (e.g., *"is good"*) that do not constitute MWEs.

D High-Precision Preservation Patterns

To maintain recall for core MWE categories, the filter is configured to bypass ("whitelist") the following high-confidence linguistic structures:

`(ADP, NOUN) and (ADP, DET, NOUN)` : Reliable prepositional idioms.

`(NOUN, ADP, NOUN)` : Common nominal compounds (e.g., *"horário de folga"*).

`(VERB, NOUN / ADP)` : Core verbal MWE types such as Light Verb Constructions (LVCs) and Inherently Adpositional Verbs (IAVs).

`(ADV, ADV / ADP)` : Multi-word adverbs and compound prepositions.

E Structural Pruning Rules

In addition to POS tagging, the system enforces three rigid structural constraints to refine the span boundaries:

1. **Length Check**: Predictions with fewer than 2 tokens are discarded (unless cached).
2. **Punctuation Only**: Spans consisting solely of non-alphanumeric characters are removed.
3. **Lexical Density**: Spans composed entirely of function words (`DET`, `PRON`) are eliminated.

F Ablation Study: Impact of Morphological Filtering

To quantify the impact of the POS-based constraints and single-token filters, we conducted an ablation study the development data at some point. We have since fixed some logic errors, therefore some languages who had bad scores are now benefiting from morphological filtering and vice versa. We compare the *Baseline LLM Pipeline* (prompting only) against the *Filtered Pipeline* (LLM + Morphological Post-processing).

	LLM		Filtered Pipeline	
Lang	**MWE-F1**	**Token-F1**	**MWE-F1**	**Token-F1**
EGY	0.1053	0.1026	0.1053	0.1026
EL	0.1452	0.3183	**0.2222**	0.3170
FA	0.1831	0.2898	**0.3470**	**0.4855**
FR	0.1056	0.1794	0.0594	0.1283
HE	0.0101	0.0134	0.0058	0.0095
JA	0.0603	0.2700	0.0571	0.2657
KA	0.0362	0.0531	0.0285	0.0488
NL	0.1721	0.3048	**0.3636**	**0.4762**
PL	0.0451	0.0584	**0.0452**	**0.0586**
PT	0.0857	0.1500	**0.0896**	**0.1558**
RO	0.0149	0.0218	0.0146	0.0215
SL	0.0431	0.0849	0.0370	0.0819
SR	0.0382	0.0525	0.0382	0.0525
SV	0.1792	0.2992	**0.1821**	**0.3002**
UK	0.0795	0.1026	0.0794	0.1027

Table 3: Scores of the filtering strategy

Analysis of Results:
A delicate aspect of our final system configuration was deciding to apply morphological filtering even to languages with marginal performance decreases during the abation study. The peformance is highly sensitive to the outputs of the LLM for that particular run (identical prompts can yeld varying levels of noise across different executions. We believe linguistic stability provided by the filters is more valuable for overall system robustness than hyper-optimizing for a limited data set.