IPN at MWE-2026 PARSEME 2.0 Subtask 1: MWE Identification via Related Languages and Harnessing Thinking Mode

Anna Hülsing, Noah-Manuel Michael, Daniel Mora Melanchthon, Andrea Horbach
Kiel University, Germany
Leibniz Institute for Science and Mathematics Education, Kiel, Germany
huelsing@ipn.uni-kiel.de

Abstract

We present IPN, our system for Subtask 1 of the PARSEME 2.0 Shared Task, which targets the identification of MWEs in 17 languages. Overall, IPN outperformed a much larger-parameter baseline model, yet a performance gap to the top-performing systems remains. To better understand these results, we investigate QWEN3-32B's suitability for mono-, cross- and multilingual MWE identification. We also explore whether this model benefits from prepending automatically generated thinking data to the gold label during instruction-tuning. We find that target language data is vital for instruction-tuning. Prepending generated thinking data to a subset of the training data slightly improves performance for two out of three languages, but more detailed evaluation is required.

1 Introduction

This paper describes IPN, our system for Subtask 1 in the PARSEME 2.0 Shared Task (Scholivet et al., 2026), focusing on the identification of multiword expressions (MWEs) in 17 languages. An MWE is a combination of at least two lexemes that exhibits varying degrees of idiosyncrasy at least one linguistic level. Representative examples are *to pay a visit* (verbal MWE), *larger than life* (adjectival/adverbial), *hot dog* (nominal), or *every other* (as in *every other day*, functional).

In our system, we build upon recent developments in MWE detection by Ide et al. (2025) and use an open-weight model from the QWEN3 family (Qwen Team, 2025). We aim to maximize lexical diversity by sampling as many different MWE types as possible, and try to automatically enhance our training data to support the model's thinking process. For instruction-tuning, we use training data from multiple languages.

In this report, we describe our data preprocessing in Section 3 and our model in Section 4. We evaluate our experimental choices in Section 5 by addressing (1) the influence of the usage of synthetic thinking data for instruction-tuning and (2) the influence of the language(s) used for instruction-tuning on model performance. We do this by comparing four training data conditions: target language data, data only from languages related to the target language, data only from unrelated languages, and target data plus data from related languages.[1]

2 Background & Related Work

Instruction-Tuning. Savary et al. (2019) observe that unseen MWEs constitute the main source of errors and thus argue that MWE identification can be enhanced by large-coverage MWE lexicons. In a similar vein, Überrück-Fries et al. (2024) train a rule-based system that exploits MWEs extracted from a large online lexicon, and Tanner and Hoffman (2023) leverage encodings of WordNet sense definitions (Miller, 1994) in a BERT-based word-sense-disambiguation approach adapted to English MWE identification. However, Ide et al. (2025) outperform the system by Tanner and Hoffman (2023) by instruction-tuning QWEN2 on as little as 780 English training sentences. We examine whether QWEN is also effective for non-English languages.

Relatedness. Phylogenetic relatedness and geographical or cultural proximity are known to promote the sharing of MWEs across languages (Gluski, 1971; Strauss, 1994; Colson, 2008; Perepadia and Malakhova, 2023), although many expressions transcend these boundaries (Piirainen, 2012). This suggests potential for cross-lingual transfer. Swaminathan and Cook (2023) fine-tune BERT-based models on English and Portuguese and evaluate on Galician data from SemEval-2022 (Tayyar Madabushi et al., 2022). They find that fine-tuning on English does not surpass a majority-class

[1] The source code is available at https://github.com/AnHu2410/PARSEME_2_ipn.

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 177–186
March 28, 2026 ©2026 Association for Computational Linguistics

baseline, whereas fine-tuning on Portuguese – an Ibero-Romance language like Galician – does. Motivated by these findings, we examine how the relatedness of instruction-tuning data to the target language affects QWEN3's performance.

Thinking. Wang et al. (2025b) elicit critique from a teacher model for math problem solving and use this critique to fine-tune a student model, thereby improving its mathematical reasoning beyond standard SFT. Zhang et al. (2025) elicit explanations of why prompts are adversarial or benign and use these explanations to fine-tune a second model, achieving better adversarial-prompt classification. Similarly, we generate rule-based "thinking" data and use them to enhance MWE identification via instruction-tuning.

3 Data

To perform instruction-tuning efficiently while preserving as much coverage as possible, we apply the following filtering steps to reduce the size of the PARSEME 2.0 Shared Task training data: First, we collect all MWE types as sets of their lemmas. Then, for each MWE type we select one training sentence that contains it, while allowing a single sentence to serve as the representative example for multiple MWE types. This way, we sample all MWE types that occur in the training data at least once. Some occur more than once, if they happen to appear in sentences that were actually sampled for a different MWE. Additionally, we add up to 20 sentences that do not contain an MWE. The type-token ratio in the training datasets as well as the number of sentences that resulted from our filtering step is shown in Table 1. As for some languages there are substantially fewer sentences than for others, we decided to use the entire training data for languages where the filtering resulted in less than 400 sentences (EGY, EL, NL, PT).

Language	Tokens	Types	Sentences
EGY	117	66	79 (431)
EL	670	417	355 (1380)
FA	4415	1939	1010
FR	4604	2042	1230
HE	12286	4327	3480
JA	2664	1626	854
KA	1517	432	433
LV	2436	770	698
NL	95	88	75 (90)
PL	11946	3219	2915
PT	167	134	128 (421)
RO	57609	3263	2624
SL	5481	2027	1730
SR	8094	3457	2621
SV	2876	1118	957
UK	4929	2428	2021

Table 1: Number of MWE tokens, MWE types and number of sentences after filtering PARSEME 2.0 training data per language. For very small datasets, we used the entire training data (numbers of sentences in brackets). Included in *Sentences* are the sentences we added that did not contain MWEs.

4 Model

Ide et al. (2025) have demonstrated the effectiveness of instruction-tuning QWEN-2.5-72B-INSTRUCT for English MWE identification. We build on and extend their work by resorting to the more recent QWEN3 family, which covers 119 languages and dialects (Qwen Team, 2025; it is not mentioned which ones exactly). More precisely, we select QWEN3-32B, because in preliminary experiments this dense variant performed best under our computational constraints (one NVIDIA H100 with 18432 CUDA cores, 80GB GPU memory and 2.0TB/s memory bandwidth).

4.1 Model Input

Model input consists of the prompt in combination with one of the sentences from the filtered training data (see Table 1). The prompt is adapted from Ide et al. (2025), and consists of the following building blocks: a definition of what MWEs are, a congruency pointer instructing the model to make use of congruent MWEs (i.e. MWEs whose lexemes are one-to-one literal translations across two languages), and a format pointer. Each one of these blocks includes a reference example to guide the model (see complete prompt in Appendix A.1). The definition block differs slightly from the prompt from Ide et al. (2025). In particular, we removed the sentence "In other words, a semantically idiomatic [MWE] takes on a meaning that is unique to that combination of words", because we consider uniqueness of meaning to be a general property of lexical items and regard strict synonymy as rare. As the original prompt from Ide et al. (2025) was intended for MWE identification in English, we added the congruency pointer to exploit the potential of cross-lingually congruent MWEs. Finally, we opted for a lexeme-index format, e.g. "to and fro; 7,8,9 | break up; 12, 13". We deemed this to be more token-efficient than the format employed by Ide et al. (2025). The prompt-sentence combination was used as input for both instruction-tuning

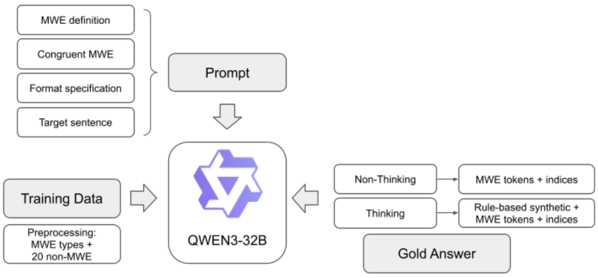

Figure 1: Overview over instruction-tuning procedure.

and inference.

4.2 Gold Thinking Content and Answer

Since QWEN3-32B supports thinking mode, we sought to exploit – or at least preserve – it's thinking capabilities by providing thinking content in the gold data used for instruction-tuning. To avoid the very time-consuming manual creation of thinking data, we employed a rule-based procedure to generate gold thinking content for each sentence in the filtered training data. We based this procedure on the thinking data that the model produced correctly without instruction-tuning (see Table 6 in Appendix A.2). Here, we identified the following Elements: 1) parsing of the sentence, 2) repetition of the MWE definition provided in the prompt, 3) discussion of potential MWE expressions, 4) repetition of the indices, and 5) declaration of the final answer.

Then, we mimicked the thinking of the non-instruction-tuned model for any given sentence in the training data: As a first step, we repeated the words and their indices from the gold data. In doing so, we combined the parsing of the sentence and the repetition of the indices (see Element 1 and 4). Secondly, we repeated the MWE definition from the prompt (see Element 2). As a third step, we discussed each MWE. To do this, we used a dictionary that we had constructed from all MWE categories occurring in the training data across all languages. For each category (e.g. AdvID), the dictionary stores a short paraphrase derived from the broad definitions in the PARSEME guidelines[2]. Examples of these paraphrases are given in Appendix A.5, Table 8. For each sentence, we retrieved the MWE category from the gold standard and, instead of providing a free-form explanation of why the expression is an MWE, we concatenated the lemmas of the MWE with the stored paraphrase of its category. After discussing all MWEs in the sentence we added the declaration of the final answer (Element 5). The complete thinking content was prepended to the gold answer, using QWEN3's formatting with thinking tags (<*think*> and </*think*>). For an example, see Appendix A.2.

This approach shows four main differences compared to the original thinking that would likely be avoided under manual annotation but arise in our automatic generation setup: Firstly, instead of mentioning which specific lexemes are e.g. non-compositional, the generated thinking mentions parts of speech such as "the verb is used non-compositionally". Secondly, our thinking does not explicitly refer to congruent MWEs; instead it gives examples of English MWEs of the same category from the PARSEME guidelines. These two differences could be optimized in a more elaborate setup. Thirdly, we do not discuss why certain words are *not* an MWE. As recall is often a problem with MWE identification (Ide et al., 2025), we deemed this acceptable. Fourthly, we did not include thinking tokens such as "Wait" as well as thought steps that are not needed for reaching the correct conclusion, as it was shown that these do not improve performance (Wang et al., 2025a; Li et al., 2025).

4.3 Instruction-tuning and Inference

Instruction tuning is a form of fine-tuning where, instead of optimizing on input–output pairs, one optimizes on instruction–input–output triples, so the model learns to respond to instructions rather than just map inputs to outputs. Figure 1 summarizes this triple – instruction (prompt), input (training data), and output (gold answer) – as used in our experiments and as described in the preceding sections. We instruction-tuned QWEN3-32B using LoRA adapters under 4-bit quantization. With the Unsloth library (Daniel Han and Unsloth Team, 2023), processing 100 sentences took approximately 3 minutes for the non-thinking and 4 minutes for the thinking variant. For VRAM-efficient inference, we used an 8-bit quantized version of QWEN3-32B. Inference for 100 sentences took approximately 30 minutes for the thinking variant and 3 minutes for the non-thinking variant. Inference with the non-instruction-tuned model took about 40 minutes for 100 sentences. Hyperparameters are shown in Appendix A.3.

[2]Version 2.0: `https://PARSEMEfr.lis-lab.fr/PARSEME-st-guidelines/2.0/`, accessed 20.11.2025

4.4 Postprocessing Model Output

During inference, the model mostly followed the format that we asked for in the prompt. However, in some cases, it made slight changes, such as adding newlines to its prediction. Also, in rare cases, the thinking exceeded our token limit, and thus no prediction was generated. Therefore, we deleted noise such as line breaks and replaced all unfinished thinking by None. As the indices were not always correct, we align each predicted item to the original CUPT file as follows: If the lexeme occurs only once, we directly tag this lexeme in the input cupt file. If it occurs multiple times, we select the index closest to the predicted one. The resulting indices are used to map the predicted MWEs to lexemes in the CUPT file.

5 Experiments and Results

For our PARSEME 2.0 submission, we used the thinking variant for all languages except those with the largest blind test sets (KA and LV), for which we employed the non-thinking variant to maintain feasible inference times. Moreover, for languages belonging to the same branch of a language family, we jointly trained a single model on all its languages (Germanic, Slavic, Romance); for all other languages we trained one model per language. In order to validate these methods, we performed two experiments: Firstly, we investigated whether instruction-tuning the model with automatically generated thinking content in addition to the gold answer boosts performance. Secondly, we explored how MWE identification performance on a target language varies when training on target-language data only, on same-branch data, on data from unrelated languages, and on target+branch data.

Influence of Generated Thinking: For gauging the influence of thinking content on our model's performance, we instruction-tune three models. For the first one, the gold answer consists only of the tokens annotated as being part of an MWE in the training data plus their indices (non-thinking variant). For the second one this gold answer was appended to automatically generated thinking content (thinking variant, see Section 4.2). For the third one, 20% of the gold answers were presented in the thinking and 80% in the non-thinking variant (20/80-variant), in an attempt to mitigate potential overfitting to the wording of the thinking content. As test languages, we selected Romanian, Slovene, and Swedish because they represent different branches of the Indo-European language family (Romance, Slavic, and Germanic branch). Again, we jointly instruction-tuned a single model for all languages of each branch.

Evaluation was carried out on the PARSEME 2.0 dev sets, where the dev set for Romanian was cut to 2,500 sentences to maintain feasible inference times. To account for variability in model outputs arising from hardware differences and from the use of a non-zero temperature (required for optimal thinking), we trained two models with different random seeds and report the mean and standard deviation (SD) across inference runs (see Table 2). For Romanian, the non-thinking variant shows the best performance, while for Slovene and Swedish the 20/80-variant shows the best performance. The second best performance differs across languages, suggesting that the relative effectiveness of thinking varies in a language-specific manner. A limited amount of generated thinking appears beneficial for this task, whereas fully thinking-augmented instruction-tuning does not outperform either the non-thinking or the 20/80 configuration. Since these experiments were conducted after the Shared Task, the official results could, in principle, be improved by using either the 20/80 or the non-thinking variant.

	thinking	**20/80**	**non-thinking**
RO	30.3 (± 0.9)	32.95 (± 0.85)	**35.8** (±0.3)
SL	23.0 (± 2.3)	**25.6** (± 1.0)	19.1 (±2.1)
SV	22.9 (± 1.0)	**24.7** (± 1.65)	24.4 (±0.5)

Table 2: Mean MWE-based F1 scores (± SD) for thinking, non-thinking and a mixed training regime per language (Romanian, Slovene, and Swedish).

Influence of Relatedness: We next examined how performance on a target language degrades when instruction-tuning is performed only on target-language data, which in our exemplary cases is Romanian and Slovene (*target*), on data from other languages in the respective language family branch (Romance, Slavic; *branch*), on data from unrelated languages (*unrelated*), and on data from the target language plus data from the same branch (*target+*), with evaluation always on the target language.[3] For the first three scenarios, we instruction-tuned on 1500 sentences that we randomly selected

[3]We did not consider the Germanic branch as PARSEME 2.0 only covers Swedish and Dutch, with Dutch including only 90 training and 10 dev sentences. This renders a meaningful comparison with other language families impossible. For all other targets, there was only a single language per branch.

train	RO			SL		
	precision	recall	F1	precision	recall	F1
target	20.6 (±0.3)	**41.2** (±0.7)	27.4 (±0.1)	12.7 (±1.9)	**34.5** (±6.3)	**18.6** (±2.9)
branch	13.1 (±0.5)	22.6 (±0.3)	16.6 (±0.5)	7.7 (±0.6)	26.4 (±1.3)	11.9 (±0.9)
unrelated	11.1 (±0.3)	20.4 (±0.6)	14.4 (±0.4)	8.0 (±0.1)	22.3 (±4.3)	11.7 (±0.5)
target+	24.5 (±1.6)	40.2 (±2.9)	**30.4** (±0.4)	12.6 (±0.1)	34.4 (±4.3)	18.4 (±1.2)
none	**32.6** (±0.0)	22.1 (±0.0)	26.3 (±0.0)	**15.8** (±0.0)	19.1 (±0.0)	17.3 (±0.0)

Table 3: Mean MWE-based precision, recall, and F1 (± standard deviation) for Romanian (RO) and Slovene (SL) under different instruction-tuning data conditions.

from our filtered datasets (see Table 1). For *target+*, we combined 1500 sentences from *target* and 1500 from *branch*. The exact numbers for each scenario are found in Appendix A.4. We perform these experiments in the non-thinking variant, as inference times are reduced and thinking does not seem to introduce an unmitigated benefit. Again, we train two models with different random seeds and report mean and SD, and evaluate on the PARSEME 2.0 dev set (we reduce the Romanian dev set to 2,500 sentences). In addition, we report the scores of the non-instruction-tuned model (*none*; we ran the same model twice, which resulted in the same scores). The results are shown in Table 3. Instruction-tuning on *target*- plus *branch*-language data yields the best F1 for Romanian and the second-highest F1 for Slovene; only 0.2 points below the *target* model. Given the relatively large standard deviations, the two Slovene results are likely not meaningfully different. These findings corroborate the method employed for the Shared Task. The worst performance is seen when instruction-tuning on unrelated languages, as expected. Instruction-tuning on non-target language data shows a worse performance than using the untuned model, which highlights the necessity of target language data for instruction-tuning.

5.1 Results in the Overall PARSEME 2.0 Context

In the official PARSEME 2.0 ranking, IPN ranked third overall with an MWE-based mean F1-score of 28.4 across all 17 languages tested in the blind scenario. However, this score is clearly below the scores obtained by the two top-performing systems (MTLB-STRUCT: F1=48.4; Sahara-Tokenizers: F1=57.3). Except for Egyptian, Ancient Greek, Georgian, and Latvian, IPN outperforms the GPT-OSS-120B baseline in MWE-based F1 for all languages. This suggests that, in medium- to high-resource settings, instruction-tuning reverses the performance gap between the 32B and 120B models for this task. The four underperforming languages are used by 0.1% or less of the websites as content language, which shows that even though the QWEN3 family covers 119 languages and dialects, low-resource languages still pose a problem.[4] Our model is marked by a high recall, which is often competitive with the top-performing models (e.g. in the mean global token-based setting, recall is less than one point below the top-performing system), but also by low precision. The high recall is induced by instruction-tuning: as Table 3 shows, recall consistently improves when performing instruction-tuning (the only exception being RO, where recall decreases when instruction-tuning on unrelated languages), while precision declines.

6 Conclusion

We draw two main conclusions from our experiments. First, prepending generated thinking to the gold answer appears helpful for two out of three languages, and only when applied to a subset of training instances, rather than uniformly to all examples. At the same time, the thinking variant requires roughly an order of magnitude more inference time than the non-thinking model, yielding a poor cost–benefit ratio. Future work should evaluate the thinking variant more deeply, e.g. by investigating different settings (80/20, 50/50), and by conducting ablation studies to test whether removing specific components of the thinking data improves performance. Second, only instruction-tuning on data containing target-language data (*target* or *target+*) reliably outperforms the untuned baseline, underscoring the importance of in-language supervision for MWE identification with QWEN3. For most languages, IPN outperformed a much larger-parameter baseline model, yet a substantial performance gap to the top-performing systems remains.

[4] https://w3techs.com/technologies/overview/content_language; date of access: February 2, 2026.

Limitations

With QWEN3 being an open-weight but not an open-source model, code and pretraining data cannot be accessed. This means that we do not know whether the data was obtained in an ethically responsible way. Also, potential biases in the model cannot easily be detected and remedied.
In both the Shared Task and our subsequent experiments, the system prompt was included in the user prompt rather than being provided in a dedicated system role; this was discovered only after the Shared Task, so our reported results may underestimate the performance of a correct system–user prompt configuration.

References

Jean-Pierre Colson. 2008. Cross-linguistic phraseological studies: An overview. In Sylviane Granger and Fanny Meunier, editors, *Phraseology: An Interdisciplinary Perspective*, pages 191–206. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Michael Han Daniel Han and Unsloth Team. 2023. Unsloth.

Jerzy Gluski, editor. 1971. *Proverbs: A Comparative Book of English, French, German, Italian, Spanish and Russian Proverbs with a Latin Appendix*. Elsevier Pub. Co., Amsterdam and New York.

Yusuke Ide, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. CoAM: Corpus of all-type multiword expressions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27004–27021, Vienna, Austria. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Language models can easily learn to reason from demonstrations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15979–15997, Suzhou, China. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Daria Perepadia and Yuliia Malakhova. 2023. Chinese and Japanese phraseologisms: A comparative aspect. *Philological Treatises*, 15(2):120–137.

Elisabeth Piirainen. 2012. *Widespread Idioms in Europe and Beyond: Toward a Lexicon of Common Figurative Units*, volume 5 of *International Folkloristics*. Peter Lang, New York.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

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.

Emanuel Strauss. 1994. *Dictionary of European Proverbs*. Routledge, London and New York.

Raghuraman Swaminathan and Paul Cook. 2023. Token-level identification of multiword expressions using pre-trained multilingual language models. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 1–6, Dubrovnik, Croatia. Association for Computational Linguistics.

Joshua Tanner and Jacob Hoffman. 2023. MWE as WSD: Solving multiword expression identification with word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 181–193, Singapore. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Till Überrück-Fries, Agata Savary, and Agnieszka Dryjańska. 2024. Sailing through multiword expression identification with Wiktionary and linguse: A case study of language learning. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 248–262, Rennes, France. LiU Electronic Press.

Chenlong Wang, Yuanning Feng, Dongping Chen, Zhaoyang Chu, Ranjay Krishna, and Tianyi Zhou. 2025a. Wait, we don't need to "wait"! removing thinking tokens improves reasoning efficiency. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7459–7482, Suzhou, China. Association for Computational Linguistics.

Yubo Wang, Ping Nie, Kai Zou, Lijun Wu, and Wenhu Chen. 2025b. Unleashing the reasoning potential of LLMs by critique fine-tuning on one problem. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3017–3027, Suzhou, China. Association for Computational Linguistics.

Yuyou Zhang, Miao Li, William Han, Yihang Yao, Zhepeng Cen, and Ding Zhao. 2025. Safety is not only about refusal: Reasoning-enhanced fine-tuning for interpretable LLM safety. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18727–18746, Vienna, Austria. Association for Computational Linguistics.

A Appendix

A.1 Prompt

Table 4 shows the complete prompt used in our experiments. Appended to the prompt is an example sentence from the training data.

A.2 Example for Original and Generated Thinking

Table 5 shows a complete example of the gold answer used for instruction-tuning in our thinking-variant, which consists of the automatically generated thinking and the solution. For the non-thinking variant, only the solution was shown to the model during instruction-tuning, and for the 20/80-variant, 20% of all instances were presentend with and 80% without the automatically generated thinking content.

Table 6 presents the original thinking for the sentence from Tables 4 and 5 of the pretrained, but non-instruction-tuned model. The following elements are found in both the original and the generated thinking: 1) parsing of the sentence and/or repetition of the indices ○ , 2) repetition of the MWE definition provided in the prompt △, 3) discussion of potential MWE expressions □, and 4) declaration of the final answer ⊞.

A.3 Hyperparameters

LoRA adapters were applied with rank r=16, α=32, and dropout set to 0. Inference was then carried out with the default hyperparameters (Qwen Team, 2025) for thinking mode (temperature: 0.6, top-p value: 0.95), and for non-thinking mode (temperature: 0.7, top-p value: 0.8, presence penalty:1.5), each with a top-k value of 20 and a maximum of 5000 tokens using VLLM (Kwon et al., 2023).

A.4 Relatedness: Numbers of Training Sentences

Instruction-tuning was conducted under four data conditions, with the first three of them comprising 1500 sentences and the last one comprising 3000 sentences. The sources and compositions of the training data are summarized in Table 7.

A.5 MWE Description Dictionary

Table 8 shows three examples of MWE categories, how they are described in the PARSEME annotation guidelines (`https://PARSEMEfr.lis-lab.fr/PARSEME-st-guidelines/2.0`, date of access: November 20th, 2025), and how they were transformed into paraphrases. The paraphrases were used as building blocks for the thinking content, more precisely for the discussion of potential MWE expressions (see □ in Table 5).

definition	You are a helpful system for identifying multiple-word expressions (MWEs). Identify all MWEs in the given sentence, and output their surface forms. Each sentence is a string of words delimited by". An MWE is defined as a sequence that satisfies the following three conditions. 1. It consists of multiple words that are always realized by the same lexemes. The individual lexemes cannot be replaced by synonyms without distorting the meaning of the expression as a whole or violating language conventions. 2. It displays semantic, lexical, or syntactic idiomaticity. Semantic idiomaticity occurs when the meaning of an expression cannot be explicitly derived from its components. Lexical idiomaticity occurs when one or more components of an expression are not used as stand-alone words in standard English. Syntactic idiomaticity occurs when the grammar of an expression cannot be derived directly from that of its components. For example, semantically idiomatic MWEs include 'break up', the lexically idiomatic include 'to and fro', and syntactically idiomatic MWEs include 'long time no see'. 3. It is not a multi-word named entity, i.e., a specific name of a person, facility, etc.
congruency pointer	Remember that you can identify congruent MWEs across different languages. For example, you can identify the Romanian MWE 'pur și simplu' because you know the English MWE 'pure and simple'. Similarly, the Portuguese MWE 'ter lugar' is easy to identify because of the English MWE 'take place'. And the French MWE 'feu de circulation' is easy to identify, because it is almost congruent to the English MWE 'traffic lights'.
format pointer	Respond by providing all tokens of the MWE, and their indices. If no MWE occurs, output 'None'. If there are multiple MWEs, separate them by |, for example 'to and fro; 7,8,9 | break up; 12, 13'. Sentence: Macron planeja implementar a reforma por meio de decreto presidencial, o que significa que vai saltar os legisladores.

Table 4: Prompt used in our experiments with example sentence from Portuguese training data. The English translation of the sentence is: "Macron plans to implement the reform by means of a presidential decree, which means he will bypass lawmakers."

thinking	<think> ◯ Okay, let's tackle this query. The user wants me to identify all multiple-word expressions (MWEs) in the given sentence. The sentence is split into words by newlines, so first I need to parse each line as a token. Let me list them out with their indices to keep track. The words are: 1 Macron 2 planeja 3 implementar 4 a 5 reforma 6 por 7 meio 8 de 9 decreto 10 presidencial 11 , 12 o 13 que 14 significa 15 que 16 vai 17 saltar 18 os 19 legisladores 20 . △ Now we apply the MWE criteria. The first condition is that the expression consists of multiple words that are always realized by the same lexemes. The individual lexemes cannot be replaced by synonyms without distorting the meaning of the expression as a whole or violating language conventions. The second condition is that it displays semantic, lexical, or syntactic idiomaticity. Semantic idiomaticity occurs when the meaning of an expression cannot be explicitly derived from its components. Lexical idiomaticity occurs when one or more components of an expression are not used as stand-alone words in standard English. Syntactic idiomaticity occurs when the grammar of an expression cannot be derived directly from that of its components. For example, semantically idiomatic MWEs include 'break up', the lexically idiomatic include 'to and fro', and syntactically idiomatic MWEs include 'long time no see'. The third condition is that it is not a multi-word named entity. □ Looking at the words, let's check for possible MWEs. por meio de is a multiword expression. More precisely, it is a conjunction idiom or ConjID. It is comparable to the English ConjID "in that" in the following sentence: "she was fortunate in that she had friends to help her". The words cannot be realized by synonymous lexemes without altering the meaning of the construction as a whole, which satisfies condition 1. The given expression fulfills condition 2, as the expression displays syntactic idiomaticity. Also, the meaning of the expression is non-compositional. The expression is a conjunction and therefore does not refer to a person or place, so it also meets condition 3. ⊞ After considering all possibilities, my final answer is: por meio de; 6,7,8 </think>
solution	por meio de; 6,7,8

Table 5: Example for automatically generated thinking data for the sentence shown and translated in Table 4, and the gold solution.

original thinking	<think> ◯ Okay, let's tackle this query. The user wants me to identify all MWEs in the given sentence. The sentence is in Portuguese, and the example mentions that MWEs can be identified across languages if they are congruent with English ones. First, I need to parse the sentence. The words are: Macron, planeja, implementar, a, reforma, por, meio, de, decreto, presidencial, , , o, que, significa, que, vai, saltar, os, legisladores, . Looking at each word, I need to check if any sequences meet the three conditions of MWEs. △ Let's go through the conditions again. Condition 1: The MWE must consist of multiple words that are always realized by the same lexemes. They can't be replaced by synonyms without changing the meaning or violating conventions. Condition 2: It has semantic, lexical, or syntactic idiomaticity. Semantic idioms can't be derived from their components. Lexical idioms have components not used as standalone words. Syntactic idioms have grammar not derived from components. Condition 3: Not a named entity. ☐ Now, checking the Portuguese words. Let's see if any phrases are known MWEs. For example, "por meio de" is a common phrase in Portuguese, meaning "by means of." In English, it's "by means of" or "through," so that's a congruent MWE. Let's check if "por meio de" is a fixed expression. Yes, it's a prepositional phrase that's used as a single unit. The individual words "por" and "meio" can be standalone, but together they form an MWE. However, does it have semantic idiomaticity? The meaning of "por meio de" is "by means of," which is the same as the literal translation. So maybe not semantic idiomaticity. But syntactically, maybe? Wait, the structure "por meio de" is a fixed prepositional phrase. In Portuguese, it's a common collocation, so it might be considered an MWE. However, the user's example for English includes "take place" as a syntactically idiomatic MWE. But in this case, "por meio de" is a standard prepositional phrase. Hmm. Another possible MWE is "saltar os legisladores." Wait, "saltar" can mean "to jump over" or "to skip." But "saltar os legisladores" might be a specific expression. However, I'm not sure if that's a standard MWE. Alternatively, "vai saltar os legisladores" – does that mean skipping the legislators? Not sure. Maybe not a standard MWE. What about "decreto presidencial"? That's a noun phrase meaning "presidential decree." Each word is standard, so probably not an MWE. Looking for other possibilities. "por meio de" is a common phrase, but I'm not sure if it's an MWE here. The user's example for "take place" is a syntactically idiomatic MWE. But "por meio de" is a standard prepositional phrase, so maybe not. Alternatively, maybe "por meio de" is considered an MWE because it's a fixed collocation. But I need to check if the components can be replaced by synonyms. If you say "por meio de" vs. "por meio de" – replacing with synonyms might not be possible without changing meaning. However, in Portuguese, "por meio de" is a fixed expression, so maybe it's an MWE. Another possibility is "o que significa que" – but that's a common structure, not an MWE. "O que significa que" is like "which means that" in English. But it's a standard grammatical structure, not an idiom. ◯ So, the only possible MWE here might be "por meio de" (indices 6,7,8). Let me check the indices again. The words are: 1: Macron 2: planeja 3: implementar 4: a 5: reforma 6: por 7: meio 8: de 9: decreto 10: presidencial 11: , 12: o 13: que 14: significa 15: que 16: vai 17: saltar 18: os 19: legisladores 20: . So "por meio de" is at positions 6,7,8. If that's an MWE, then that's the only one. But I need to confirm if it's considered an MWE. Since the user's example included "take place" as an MWE, which is a congruent phrase with "ter lugar" in Portuguese, which is "take place" in English. So "por meio de" is congruent with "by means of" in English, which is a standard MWE. Therefore, "por meio de" is an MWE here. Another possible MWE could be "saltar os legisladores," but I don't think that's a standard MWE. "Saltar" in this context might mean "to skip over," but without more context, it's hard to tell. However, "saltar" as a verb can be used in different ways, so it's not a fixed expression. ⊞ So the answer would be "por meio de; 6,7,8". </think>

Table 6: Example for original thinking data for the sentence shown and translated in Table 4 generated by non-instruction-tuned QWEN3-32B.

Lang.	Condition	Training data
RO	**target**	1,500 Romanian sentences.
	branch	1,500 Romance sentences (1,079 French + 421 Portuguese).
	unrelated	1,500 sentences from Swedish, Dutch, Egyptian, Modern Greek, Farsi, Hebrew, Japanese, Georgian, Latvian, Ukrainian, Serbian, Slovene, and Polish; per-language contributions capped at $\approx$117–118 sentences, Dutch contributing its full 90.
	target+branch	3,000 sentences: 1,500 Romanian (as in **target**) + 1,500 Romance (as in **branch**).
SL	**target**	1,500 Slovene sentences.
	branch	1,500 Slavic sentences (500 Ukrainian + 500 Serbian + 500 Polish).
	unrelated	1,500 sentences from Romanian, French, Portuguese, Swedish, Dutch, Egyptian, Modern Greek, Farsi, Hebrew, Japanese, Georgian, and Latvian; per-language contributions capped at $\approx$128–129 sentences, Dutch contributing its full 90.
	target+branch	3,000 sentences: 1,500 Slovene (as in **target**) + 1,500 Slavic (as in **branch**).

Table 7: Instruction-tuning data conditions for Romanian (RO) and Slovene (SL).

Type	Guideline Description	Building Block
LVC.full	LVCs in which the verb is semantically totally bleached, EN: to give a lecture.	*<MWE lemmas>* is a multiword expression. More precisely, it is a full light verb construction (LVC.full) because it is formed by a verb and a noun, where the verb is semantically fully bleached. A similar bleaching occurs in the English MWE *(to) give (a) lecture*. The words cannot be realized by synonymous lexemes without altering the meaning of the construction as a whole, which satisfies condition 1. The given expression fulfills condition 2, as the verb displays semantic idiomaticity: the verb does not have its original meaning anymore, instead the meaning is fully bleached. The expression does not refer to a specific person or place, so it also meets condition 3.
NV.LVC.full	Deverbal nominal stemming from an LVC.full (NV.LVC.full), EN: a decision maker - deriving from the LVC.full to make a decision.	*<MWE lemmas>* is a multiword expression. More precisely, it is a deverbal nominal stemming from a full light verb construction (NV.LVC.full). NV describes the fact that a noun (N) is derived from a verb (V), whereas LVC.full refers to the full light verb construction which is the basis for this multiword expression. An English example is *decision maker*, which stems from the LVC.full *(to) make (a) decision*. The words cannot be realized by synonymous lexemes without altering the meaning of the construction as a whole, which satisfies condition 1. The given expression fulfills condition 2, as the verb displays semantic idiomaticity: the deverbal noun does not have its original meaning anymore, instead the meaning is fully bleached. The expression does not refer to a specific person or place, so it also meets condition 3.
AdvID	Adverbial idiom (AdvID) – a universal category, caracterized by lexical, morphological or syntactic irregularity, EN: by and large.	*<MWE lemmas>* is a multiword expression. More precisely, it is an adverbial idiom (AdvID), a universal category characterized by lexical, morphological, or syntactic irregularity. It is comparable to English AdvIDs such as “on the whole” or “by and large.” The lexemes and order are fixed; substituting synonyms disrupts acceptability or the idiomatic meaning, which satisfies condition 1. It fulfills condition 2 through semantic idiomaticity (the expression is non-compositional). The expression does not denote a specific named entity, so it also meets condition 3.

Table 8: Broad MWE category descriptions for three exemplary MWE categories from the PARSEME 2.0 guidelines, and the building blocks made out of the descriptions for the generation of thinking content.