

Benchmarking Mathematical Reasoning in a Low-Resource Language: Structured Prompting and Evaluation in Basque

Anonymous EMNLP submission

Abstract

LLMs have shown impressive performance on tasks requiring complex reasoning, but most evaluations focus exclusively on English. This work investigates how well LLMs perform mathematical reasoning in low-resource languages, using Basque as a primary case study. To support this analysis, we introduce **MASEU**, a benchmark designed to evaluate reasoning in Basque across arithmetic, algebraic, and logical tasks, and assess both existing open models and newly trained systems. We address three key questions: how well LLMs support Basque in reasoning tasks, to what extent English in prompts can improve results, and the effect of continued pretraining in Basque. To explore these aspects, we use a prompting strategy adapted for mathematical reasoning (*DUP prompting*), which allows for more precise experimentation across zero-shot and few-shot settings, providing insights into how multilingual models handle reasoning tasks in under-represented languages.

1 Introduction

Large Language Models (LLMs) have led to large improvements on complex reasoning tasks, ranging from deductive reasoning (Saparov and He, 2023; Stechly et al., 2025) to mathematical reasoning (Patel et al., 2021; Shi et al., 2023). Much of this improvement stems from in-context learning techniques, e.g., few-shot learning and novel reasoning techniques such as Chain-of-Thought (CoT) (Wei et al., 2022) or Tree-of-Thought (ToT) (Yao et al., 2023) prompting. While these task-agnostic reasoning techniques have already shown improvements for mathematical reasoning, other approaches have instead tried to incorporate the idiosyncrasies of mathematics into the reasoning strategy. Deep Understanding Prompting (DUP) (Zhong et al., 2024), for example, provides a targeted prompting strategy adapted for mathematical reasoning.

Mathematical reasoning (Roy and Roth, 2015; Hendrycks et al., 2021), provides a robust benchmark for evaluating structured problem-solving and logical inference in LLMs, as answers are consistent and do not require human evaluation. The task is often presented as math word problems (MWP), where a complex situation is provided, followed by a specific mathematical question. The model must then return the correct value.

However, an LLM’s reasoning ability is often tied to its language capabilities (Shi et al., 2023), and models generally achieve stronger performance on math reasoning in high-resource languages, such as English. While translating data into English can help, it also may introduce noise, and therefore it would be preferable for models to have similar reasoning abilities directly in low-resource languages.

Although mathematical reasoning has been previously studied in low-resource languages (Shi et al., 2023; Huang et al., 2023; Qin et al., 2023; Huang et al., 2024), there is little research into the effects of language-specific pretraining and the combination of languages inside of task-specific prompting techniques, such as DUP. To address these concerns, we present a novel mathematical reasoning dataset in Basque (MASEU), as well as a targeted evaluation of how effectively LLMs can perform mathematical reasoning on this dataset. Specifically, we ask three principal research questions:

RQ1: How well can LLMs perform complex mathematical reasoning tasks in a few-shot scenario for a low-resource language like Basque?

RQ2: Do models in Basque see performance gains from including English into the intermediate mathematical reasoning steps?

RQ3: Does continued pretraining improve in-language reasoning abilities?

The first focuses on evaluating how effectively current LLMs, including both general-purpose and Basque-adapted models, support complex reasoning tasks in Basque. The second examines whether selectively incorporating English into prompts can improve model performance, given its overwhelming presence in pretraining data. The third offers a vision of how much the English bias can be mitigated through continued pretraining. Together, these questions aim to advance our understanding of language-specific and cross-linguistic factors that influence mathematical reasoning in low-resource settings.

To support our first research question, *MASEU* (§3) provides a controlled benchmark for evaluating mathematical reasoning in Basque. Covering arithmetic, algebraic, and logical problems, *MASEU* offers a carefully curated collection of examples that ensure both linguistic accuracy and mathematical rigor. This dataset fills a critical gap by providing a reliable, controlled setting for assessing the reasoning capabilities of LLMs in a low-resource language context.

Similarly, for the second research question, we adopt *Deep Understanding Prompting* (*DUP*) (Zhong et al., 2024) as a framework for structured multilingual prompting. Building on the foundations of Chain-of-Thought (CoT) prompting, *DUP* offers a more systematic approach by explicitly decomposing the reasoning process into modular steps: extracting the main question, isolating relevant facts, and generating a solution.

Finally, to answer the third research question, we compare 8B and 70B multilingual models (Llama 3.1) with Latxa—Llama 3.1 with continuous pretraining in Basque—as well as similar size reasoning models (Qwen 2.5).

2 Related Work

Mathematical reasoning has become a key capability for evaluating LLMs, reflecting their capacity for structured problem-solving, inference, and symbolic computation. Early research in this area focused on mathematical word problems (MWP), particularly on equation parsing and symbolic logic in arithmetic and algebraic contexts (Roy and Roth, 2015; Koncel-Kedziorski et al., 2015), laying the foundation for logic-based reasoning frameworks.

A major shift occurred with the introduction of in-context learning through large-scale models (Brown et al., 2020), which enabled few-shot

prompting strategies. This, along with advances in transfer learning and instruction tuning (Raffel et al., 2020; Ouyang et al., 2022; Sanh et al., 2022), expanded the adaptability of LLMs across tasks. However, standard prompting approaches often underperform in multi-step reasoning, motivating new methods such as *Scratchpad* (Nye et al., 2022) and program synthesis (Odena et al., 2021) that guide intermediate computation.

Chain-of-Thought (CoT) prompting (Wei et al., 2022) emerged as a breakthrough, significantly improving performance on complex reasoning tasks. Follow-up strategies, including *Math-Prompter* (Imani et al., 2023) and Cooperative Reasoning (CoRe) (Zhu et al., 2023), enhanced robustness through structured reasoning and verification loops. Recently, Deep Understanding Prompting (*DUP*) (Zhong et al., 2024) introduced a three-stage pipeline, core question extraction, information selection, and answer generation, achieving state-of-the-art results on *GSM8K* (Cobbe et al., 2021).

Despite these advances, most techniques have focused on high-resource languages, especially English. Existing multilingual benchmarks (Hu et al., 2020; Ruder et al., 2021; Conneau et al., 2020) often assess only basic reasoning skills, leaving open the question of how well LLMs perform in more complex multilingual contexts (Ponti et al., 2020). This gap is particularly concerning given the stark disparities in language coverage across the field, where a vast number of the world’s languages, including those spoken by millions, lack basic NLP tools and datasets. Recent surveys highlight that NLP progress has been overwhelmingly concentrated on a small subset of the world’s languages, while many others, particularly from regions such as sub-Saharan Africa and Southeast Asia, remain significantly underrepresented in both research attention and resource availability (Joshi et al., 2020).

Recent work has highlighted how English often serves as a high-performing pivot language due to pretraining biases (Zhao et al., 2021; Winata et al., 2021), and strategies such as Chain-of-Thought prompting have shown promising cross-lingual transfer. Building on this, prompting designs explicitly tailored for multilingual reasoning have emerged. Cross-Lingual Prompting (CLP) introduces alignment mechanisms that map reasoning chains between source and target languages, while Cross-Lingual Self-Consistent Prompting (CLSP) ensembles reasoning paths across multiple lan-

ID	Question	Answer
21	English: An oil pipe in the sea broke . Before engineers started to fix the pipe , 6522 liters of oil leaked into the water . While the engineers worked , the pipe leaked 5165 liters of oil . In all , how many liters of oil leaked into the water ? Basque: Itsasoko petrolio-hodi bat hautsi da . Ingeniariek petrolio-hodia konpontzen hasi baino lehen, 6522 litro olio isuri egin dira uretara . Langileek lan egiten zuten bitartean, petrolio-hodiak 5165 litro isuri ditu . Guztira, zenbat litro olio isuri dira uretara ?	11687
706	English: A book is on sale for 10 % off . If the regular price is 27.9 dollars , what is the sale price ? Basque: Liburu bat salgai dago % 10 ko deskontuarekin . Precio arrunta 27.9 dolar bada, zein da salmenta prezioa ?	25.11
1651	English: For Halloween Emily received 5 pieces of candy from neighbors and 13 pieces from her older sister . If she only ate 9 pieces a day , how long would the candy last her ? Basque: Halloweenerako Emilyk 5 gozoki jaso zituen auzokideetatik eta 13 bere ahizpa nagusiarengandik . Bakarrik 9 karamelo jaten bazituen egunero, zenbat denbora iraungo dizkiote berari gozokiak ?	2

Table 1: Illustrative examples from the *MASEU* dataset. Each example includes a MWP originally written in English, followed by its corresponding translation into Basque. This bilingual presentation highlights the semantic and structural consistency preserved throughout the translation process.

guages, yielding state-of-the-art results on MGSM, XNLI, and PAWS-X (Qin et al., 2023). In contrast, Cross-Lingual Thought Prompting (XLT) uses structured, language-agnostic templates to encourage generalization without tuning (Huang et al., 2023). These developments underscore the growing importance of prompt structure in bridging performance gaps between high- and low-resource languages.

Building on this line of work, our study applies DUP prompting to evaluate state-of-the-art LLMs in minority languages, focusing exclusively on Basque. By leveraging recent developments in multilingual benchmarks and structured reasoning, we aim to examine how well these models generalize across languages, how effectively they adapt to low-resource settings, and whether their performance remains consistent when confronted with linguistic variation.

In doing so, this research contributes to the broader goal of promoting linguistic inclusivity within multilingual language modeling. While recent models increasingly emphasize built-in reasoning capabilities (DeepSeek-AI et al., 2025; OpenAI, 2025; Team, 2024), we do not evaluate these systems directly. Instead, we adopt a structured reasoning approach that enables fine-grained control over language use within the prompt. This allows us to better isolate and analyze the effects of multilingual prompt composition, particularly in low-resource settings like Basque, offering a clearer understanding of how language choice influences reasoning fidelity and generalization.

3 MASEU Dataset

To enable more reliable and linguistically faithful evaluation of mathematical reasoning in low-resource languages, we introduce **MASEU**, a new

dataset specifically constructed for Basque. The design of MASEU is grounded in the need for higher-quality resources that go beyond simple language understanding and address reasoning complexity in a structured, multilingual setting. It is based on a manually curated subset of the *mawps-asdiv-a_svamp* corpus (Patel et al., 2021), which merges three well-established benchmarks in the domain of MWPs: *MAWPS* (Koncel-Kedziorski et al., 2016), *ASDiv-A* (Miao et al., 2020), and *SVAMP* (Patel et al., 2021). These datasets were selected for their diversity in reasoning types, consistent structure, and pedagogical value, making them particularly suitable for testing LLM performance in multilingual and instructional contexts.

MASEU comprises 195 development and 1584 test entries, all carefully translated into Basque with close attention to both mathematical fidelity and linguistic naturalness. We employed a controlled translation process to preserve the original intent, difficulty level, and logical structure of each problem, as illustrated by the examples shown in Table 1. This ensures that the Basque version reflects idiomatic usage while maintaining conceptual equivalence, allowing for robust reasoning evaluation without introducing semantic drift.

The primary motivation for MASEU stems from the limitations of existing multilingual benchmarks, particularly MGSM (Shi et al., 2023). While MGSM plays a vital role in evaluating cross-lingual transfer and reasoning by aligning GSM8K with ten typologically diverse languages, its Basque translation (Baucells et al., 2025) includes only 250 test instances. As a result, this dataset could suffer from higher variance and be less predictive of true differences between models. MASEU’s larger scale allows for more stable experimentation, and its linguistic coverage ensures better representation of the

Dataset	# Dev	# Test	Langs.
MASEU	195	1584	en, eu
MGSM	8	250	bn, de, en, es, eu, fr, ja, ru, sw, te, th, zh

Table 2: Datasets used in the experiments, along with the corresponding number of evaluation examples. MGSM is the multilingual GSM8K subset; MASEU is our newly translated dataset.

complexities inherent in the language. In combination with MGSM, it enables both high-level cross-lingual comparisons and deep, language-specific analysis, providing a more complete testbed for multilingual reasoning in LLMs, as summarized in Table 2.

4 Experiments

We evaluate a set of state-of-the-art, publicly available instruction-tuned language models, selected for their complementary strengths and relevance to multilingual reasoning. *Llama3.1* (Grattafiori et al., 2024), in its 8B and 70B versions, serves as a strong open-source baseline, widely used in the community and representative of general-purpose models without explicit language adaptation. *Qwen2.5* (Yang et al., 2024), available in 7B and 72B variants, has shown highly competitive performance in multilingual and reasoning benchmarks, making it a suitable point of comparison for evaluating generalization across languages.

We also include the *Latxa models* (Etxaniz et al., 2024; Sainz et al., 2025), a family derived from Llama3.1 and explicitly adapted for Basque. Available in 8B and 70B configurations, these models allow us to isolate the impact of language-specific pretraining. The smaller Latxa model is optimized for low-resource conditions, while the larger variant is designed for robust multilingual reasoning with a focus on Basque. This selection enables a detailed analysis of how model scale, multilingual capacity, and linguistic specialization affect performance in a low-resource language setting.

4.1 DUP Prompting and Multilingual Adaptation

To support controlled evaluation of mathematical reasoning in multilingual contexts, we adopt the *Deep Understanding Prompting (DUP)* (Zhong et al., 2024) framework. Rather than focusing on

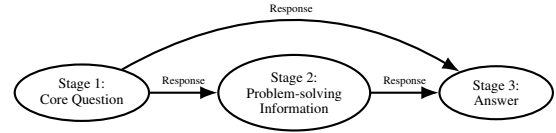


Figure 1: Reasoning flow across DUP stages. The response from Stage 1 is reused in Stages 2 and 3, and the response from Stage 2 is also used in Stage 3.

Stage 1: Reveal the Core Question

MWP: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

Please extract the core question, only the most comprehensive and detailed one!

A: How many hours a week does John spend taking care of dogs?

Stage 2: Extract the Problem-solving Information

MWP: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

Note: Please extract the question-solving information related to the problem (How many hours a week does John spend taking care of dogs?), only extract the most useful information, and list them one by one!

A: 1. John takes care of 10 dogs. 2. Each dog takes 0.5 hours a day to walk and take care of their business. 3. There are 7 days in a week.

Stage 3: Generate and Extract the Answer

MWP: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

Hint: 1. John takes care of 10 dogs. 2. Each dog takes 0.5 hours a day to walk and take care of their business. 3. There are 7 days in a week. How many hours a week does John spend taking care of dogs?

Please understand the Hint and question information, then solve the question step by step and show the answer.

A: 35

Figure 2: Schematic of the three-phase DUP prompting framework. Input MWP and generated responses use standard font, templates are in bold, and prior-stage responses are italicized. The three stages are: (i) identifying the core question, (ii) extracting problem-solving information, and (iii) generating the final answer based on earlier stages.

DUP as a reasoning enhancement strategy, we employ its structured three-stage format-composed of core question extraction, identification of key problem-solving information, and answer generation, as a tool for precise manipulation of language within different parts of the prompt. The flow of responses between stages is depicted in Figure 1. As shown in the diagram, the response generated in Stage 1 (Core Question) is propagated forward into both Stage 2 (Problem-solving Information) and Stage 3 (Answer), while the response from Stage 2 is likewise used as input for Stage 3. Figure 2 shows an example of the three-phase DUP prompting framework. The sequential reuse in DUP allows us to systematically vary the language used for each stage’s prompt, response, and previous

Lang.	MWP	Template	Response
Basque	EUS	EUS	EUS
Mixed (T)	EUS	EN	EUS
Mixed (T, R)	EUS	EN	EN
English	EN	EN	EN

Table 3: Language configurations in DUP prompting. Each row shows the language used for the MWP, template and model response. The Basque configuration uses Basque throughout, Mixed (T) uses English templates, Mixed (T, R) also uses English for responses, English is fully in English.

response components, such as mixing Basque and English across phases, and analyze how these multilingual configurations affect reasoning accuracy and transfer. Combined with the formal language assignments in Table 3, this flow diagram offers a clear operational view of DUP’s structure and its experimental flexibility in multilingual settings.

Although originally proposed as a zero-shot technique, we extend DUP to one-shot and few-shot settings by providing additional in-context examples within each phase. This modification allows us to evaluate whether contextual supervision improves reasoning performance without altering the underlying structure. All few-shot prompts follow the original DUP templates to ensure consistency. Examples of these prompts are included in Appendix A. We also compare DUP with standard prompting baselines under equivalent shot conditions. This design allows us to evaluate whether the structured reasoning provided by DUP offers consistent advantages across different supervision levels. A complete breakdown of results is provided in Appendix B.

To analyze how language influences reasoning performance, we use DUP as a controlled framework that separates problem understanding from answer generation. All Basque prompts were translated directly from English using a carefully designed procedure to preserve both mathematical structure and linguistic fidelity. In addition to a fully Basque setting, we evaluate two mixed-language configurations: Mixed (T), where the template (T) is in English and the model’s output remains in Basque, and Mixed (T, R), where both the template and response (T, R) are in English while the original MWP is still presented in Basque.

These configurations, summarized in Table 3,

allow us to isolate the role of language at different stages of the reasoning process, including task interpretation and solution formulation. If Mixed (T) outperforms Mixed (T, R), it may indicate that generating responses in the target language supports better task alignment. Conversely, stronger performance in Mixed (T, R) could reflect a preference for conducting all reasoning steps in English, the model’s dominant training language. A fully English setting is also included to test whether full alignment with the pretraining language yields performance gains, offering a reference point for evaluating the effect of language mixing.

Altogether, these multilingual configurations support a more nuanced analysis of how language composition affects reasoning in low-resource contexts like Basque. Example prompts for all variants are available in Appendix C.

5 Results

In this section, we present the results of our evaluation of LLMs on mathematical reasoning tasks in Basque, revisiting the main research questions.

RQ1: How well can LLMs perform mathematical reasoning tasks in Basque?

Table 4 presents the main results under few-shot DUP prompting, highlighting clear differences in Basque performance across models. These disparities are shaped by model scale, language specialization, dataset properties, and pretraining strategies.

Across both datasets, the Latxa models, specifically adapted for Basque, consistently achieve strong performance. On MASEU, Latxa:8B reaches 81.63, outperforming similarly sized general-purpose models like Llama3.1:8B (55.37) and Qwen2.5:7B (48.48).

Performance trends differ between datasets. MASEU generally yields higher scores than MGSM, particularly for Basque. This may be due to differences in problem formulation, linguistic coverage, or domain diversity. MGSM, derived from GSM8K, tends to follow a more uniform structure and reasoning style, while MASEU draws from MAWPS, ASDiv-A, and SVAMP, which offer a broader range of linguistic expressions and problem types. This variety may better align with the reasoning capabilities of multilingual models, contributing to improved performance.

Interestingly, despite lacking Basque adaptation, Llama3.1:70B achieves the top score on MASEU.

	Model	Basque	Mixed (T) Δ	Mixed (T, R) Δ	English Δ
MASEU	Latxa:8B	81.63	-6.69	+0.06	-2.59
	Llama3.1:8B	55.37	+2.65	+17.55	+32.95
	Qwen2.5:7B	48.48	+1.52	-4.42	+37.75
	Latxa:70B	88.32	+1.89	+3.85	+8.33
	Llama3.1:70B	89.65	-0.44	-0.06	+4.48
	Qwen2.5:72B	83.65	-4.86	-9.91	+7.01
MGSM	Latxa:8B	46.80	-8.80	+9.60	+25.20
	Llama3.1:8B	22.40	-4.40	+16.00	+50.80
	Qwen2.5:7B	16.40	-0.40	+3.20	+61.20
	Latxa:70B	71.20	+6.40	+11.20	+22.00
	Llama3.1:70B	66.00	+2.40	+7.60	+20.00
	Qwen2.5:72B	51.60	-4.80	+2.80	+35.60

Table 4: Accuracy using few-shot DUP prompting on MASEU, MGSM. Scores are shown for the Basque setting along with deltas (Δ) for adding English in the template (Mixed(T)), the template and response (Mixed(T, R)), and including all parts in English. The best result for each model size and dataset is underlined.

This may be attributed to its scale, generalization capacity, and effective in-context learning when provided with sufficient examples. However, Latxa:70B’s near-parity indicates the value of continued training on low-resource languages.

Model performance also reflects differences in pretraining. Although Qwen2.5 targets multilingual and mathematical capabilities, its math training appears to be primarily focused on English and Chinese¹, which may limit its effectiveness in other languages such as Basque. Meanwhile, Llama3.1, despite officially supporting fewer languages, benefits from broader general-domain exposure.

In summary, language-specific adaptation strongly benefits smaller models in low-resource settings, while scale and in-context learning help large generalist models remain competitive. These results show the importance of balanced multilingual pretraining and dataset design for reliable reasoning in underrepresented languages like Basque.

RQ2: What is the impact of English in mixed-language prompts?

Table 4 reveals a nuanced, model-dependent impact of introducing English into prompts for mathematical reasoning in Basque. General-purpose models such as Llama3.1:8B and Qwen2.5:7B show dramatic improvements when switching entirely to English prompts. On MASEU, Llama3.1:8B gains +32.95 points and Qwen2.5:7B gains +37.75,

while on MGSM, their gains reach +50.80 and +61.20, respectively. These results highlight the strong dependency of generalist models on English, their dominant pretraining language, especially in reasoning-intensive tasks. Larger models such as Llama3.1:70B and Qwen2.5:72B also show big improvements when using English prompts in MGSM, but they benefit less on MASEU, with more modest gains of +4.48 and +7.01, respectively. This discrepancy may reflect differences in dataset construction, with MGSM aligning more closely with the English-centric patterns found in pretraining corpora, whereas MASEU, with its broader linguistic and reasoning diversity, poses more distinct challenges that cannot be fully mitigated by simply switching to English.

The Mixed (T) configuration, where only the template is in English, yields inconsistent results. On MASEU, Llama3.1:8B improves by +2.65, while Qwen2.5:7B gains +1.52, and Latxa:70B by +1.89. However, some models regress, such as Qwen2.5:72B with -4.86. No clear trend emerges, suggesting that partial English prompting without accompanying English output provides limited or unpredictable benefit.

By contrast, the Mixed (T, R) setup, where both template and response are in English, yields more consistent improvements. On MGSM, Llama3.1:8B gains +16.00, Latxa:70B +11.20, and Qwen2.5:72B +2.80. This indicates that shifting both input and output into English can

¹<https://qwenlm.github.io/blog/qwen2.5-math/>

help bridge model limitations in handling low-resource languages, particularly for MGSM. However, on MASEU, the gains are generally smaller, and in some cases even negative, as seen with Qwen2.5:72B at -9.91.

There is also a large jump between the results of the Mixed (T, R) setup and those using fully English prompts, on all models except Latxa:8B. This suggests a strong pretraining bias toward English for reasoning, possibly due to math pretraining data being predominantly English-based. This effect is especially pronounced in MGSM, where performance gains from fully English prompts are highest, indicating that models are overfit to English-domain math reasoning tasks.

RQ3: What is the impact continued pretraining in Basque?

In Basque the best result overall comes from the much larger Llama3.1:70B at 89.65, closely followed by Latxa:70B at 88.32, but the strong performance of the smaller Latxa model highlights the effectiveness of continued pretraining. Notably, continued pretraining in Basque does not harm and seems to enhance performance in English. The Latxa:70B model achieves the best scores in both MASEU and MGSM for English configurations, outperforming even models that were not exposed to Basque. This suggests that further adaptation in a low-resource language can improve overall reasoning robustness, potentially due to richer semantic alignment or stronger cross-lingual transfer capacity enabled by instruction tuning.

6 Analysis of Prompting Strategy

In this section we perform an ablation analysis of the use of few-shot strategies and DUP prompting.

Few-shot prompting: Across both datasets, we observe that increasing the number of contextual examples leads to consistent improvements, with Few-Shot prompting outperforming the other configurations. On MASEU, this trend is evident in Figure 3 (a similar trend is found on MGSM in Table 10 in the appendix), where each additional example improves model performance, particularly for smaller or less specialized models in Basque. Zero-Shot setups generally lag behind, while One-Shot prompts provide modest but consistent gains. The advantage of Few-Shot settings is especially notable when working with languages like Basque, which are considered low-resource due to their

Lang.	Model	non-DUP	DUP Δ
Basque	Latxa:8B	74.59	+2.29
	Llama3.1:8B	46.02	+4.85
	Qwen2.5:7B	42.80	+1.31
	Latxa:70B	85.17	+0.82
	Llama3.1:70B	82.17	+4.25
	Qwen2.5:72B	72.90	+6.38
English	Latxa:8B	87.46	-9.38
	Llama3.1:8B	86.42	-0.16
	Qwen2.5:7B	87.19	-2.13
	Latxa:70B	91.66	+4.52
	Llama3.1:70B	91.77	+1.25
	Qwen2.5:72B	86.21	+3.98

Table 5: Model-wise comparison of non-DUP vs. DUP prompting on Basque and English, (micro) averaged across the MASEU and MGSM datasets in the Few-Shot setting. The results highlight how DUP affects multilingual mathematical reasoning across different model scales and languages.

limited presence in pretraining corpora and scarce availability of labeled data. In such cases, the inclusion of multiple examples offers crucial guidance that helps compensate for the model’s weaker baseline familiarity with the language.

DUP prompting: We extend our analysis by evaluating DUP prompting in the Few-Shot setting. Table 5 shows the difference of using DUP micro-averaged on both datasets, and separated by language.

In the Basque setting, all models show improvements with DUP prompting. For example, Latxa:8B gains +2.29 and Llama3.1:8B improves by +4.85, suggesting that step-by-step prompting helps disambiguate problems in morphologically rich languages. In contrast, in English, only larger models benefit from DUP, while smaller ones experience performance drops. The most pronounced decline is seen in Latxa:8B, which loses 9.38 points. This may reflect difficulties small models have with intermediate reasoning steps, possibly due to limited capacity or weaker alignment between decomposition strategies and English output formats.

Overall, Table 5 and Table 7 confirm that DUP’s benefits depend on model scale, language, and base performance. Its effectiveness is especially clear in low-resource languages like Basque and for models with weaker baseline reasoning. Additional discussion and examples are provided in Appendix B.

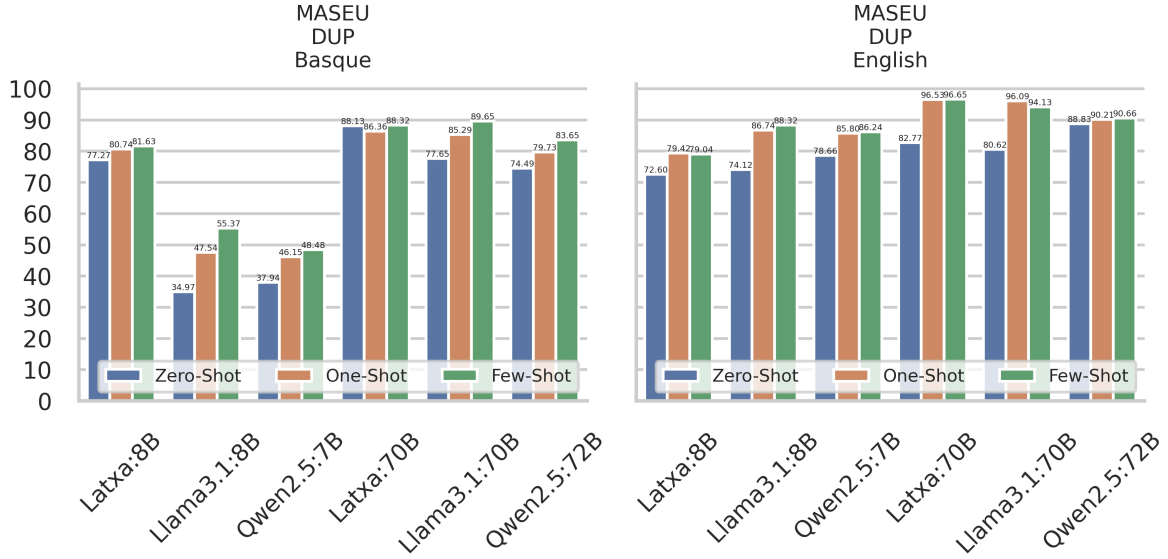


Figure 3: Comparative evaluation of *Zero-Shot*, *One-Shot*, and *Few-Shot* prompting techniques, in conjunction with the *DUP* prompting strategy, on the MASEU dataset. Results are reported for both Basque and English to assess multilingual reasoning performance.

7 Conclusion

In this work we have introduced MASEU, a new benchmark designed to evaluate how current large language models handle mathematical reasoning in Basque, a low-resource language. We then used this new resource together with the Basque and English portion of MGSM to analyze 1) how well LLMs perform mathematical reasoning tasks in Basque and 2) what the impact is of English in mixed-language prompts. We experiment with three diverse LLM model families (Latxa, Llama, Qwen) and two parameter classes (7-8 billion and 70-72 billion), as well as four prompting strategies - using all Basque, introducing English into the template (Mixed(T)), introducing English in the template and response (Mixed(T, R)), and only English.

Results across both MASEU and the MGSM dataset reveal that all models, including those specifically trained for Basque, perform better in English than in Basque, underscoring the ongoing challenges in providing equitable multilingual support.

Despite these challenges, small language-specific models like Latxa:8B significantly outperform general-purpose models of comparable size, highlighting the effectiveness of explicit linguistic adaptation. At larger scales, general-purpose models tend to show substantial improvements, suggest-

ing that increased model capacity can help bridge the gap, even in the absence of targeted language-specific training.

Interestingly, when solving Basque problems, most models perform better when intermediate reasoning steps are carried out in English, particularly in the Mixed(T, R) setting, which pairs Basque problem statements with English reasoning. This suggests that models may be overfitting to pre-training distributions, where mathematical reasoning data is predominantly in English. As a result, English-based reasoning likely aligns better with learned patterns, improving clarity and structure even in multilingual tasks. Additionally, we find that DUP prompting offers greater gains for larger models, implying that a certain capacity threshold is needed to fully exploit this technique.

Overall, our findings show that multilingual mathematical reasoning remains highly sensitive to both language and model architecture. While generalist LLMs still struggle with Basque, our results emphasize the value of targeted adaptation, as well as thoughtful prompt design, in narrowing the performance gap for underrepresented languages.

Limitations

This study focuses on evaluating mathematical reasoning in LLMs under multilingual and low-resource language conditions, with Basque as the main case study. While Basque provides a strong

testbed due to its typological uniqueness and under-representation, results may not generalize to other low-resource languages with different grammatical or orthographic properties. Future work should extend this investigation to a broader set of languages to better understand the generalizability of cross-lingual reasoning performance.

Another limitation concerns the dataset construction process. While the MASEU dataset was carefully translated from established English benchmarks to preserve mathematical content and linguistic naturalness, the translations were not conducted by professional linguists. This may introduce stylistic biases or limit exposure to more diverse or regionally specific variants of Basque. As such, conclusions drawn from this resource should be viewed in light of its translation-based design.

Additionally, all evaluated models are publicly available instruction-tuned LLMs. While this ensures reproducibility and accessibility, we do not include proprietary systems such as GPT-4, Claude, or Gemini, which may exhibit different behavior, particularly in multilingual settings. Our results should therefore be interpreted as representative of open-source model capabilities rather than the current upper bounds in the field.

Finally, although our experiments explore various in-context learning configurations and multilingual setups, we limit our analysis to a single structured prompting approach inspired by staged reasoning methods such as DUP. Future work could compare this structured setup with other prompting paradigms or incorporate model fine-tuning to better understand how to support low-resource languages in complex reasoning tasks.

Ethical considerations

Our study involves no human annotation or data collection from human participants. All datasets used in this work are derived from existing public benchmarks or generated through the controlled application of language models. In particular, the new MASEU dataset was created by translating entries from publicly available English-language datasets into Basque using careful, manual processes guided by linguistic and mathematical fidelity. No personal, private, or sensitive information is included in the dataset or any of the prompts. As such, we believe that the data used in this work poses no foreseeable risk of harm or negative societal impact.

All language models evaluated in this paper are publicly available and accessed via platforms such as the Hugging Face Hub². These include both general-purpose multilingual models³ and models specifically adapted for Basque⁴. We have complied with the licenses and usage policies associated with each model, and all experiments are conducted for research purposes only. No proprietary or closed-access models are used in this study, ensuring the reproducibility of our findings.

Throughout our experimental setup, we use structured prompting techniques to examine multilingual reasoning performance. These techniques are designed to be neutral in tone and content, and we do not prompt models to produce or analyze material that could raise ethical concerns, such as hate speech, discrimination, or misinformation. Given the focus on mathematical reasoning and the nature of the linguistic data, we consider the ethical risks associated with this work to be minimal.

References

- Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. [IberoBench: A benchmark for LLM evaluation in Iberian languages](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

²<https://huggingface.co/>

³<https://huggingface.co/collections/meta-llama/llama-31-669fc079a0c406a149a5738f>, <https://huggingface.co/collections/Qwen/qwen25-66e81a666513e518adb90d9e>

⁴<https://huggingface.co/collections/HiTZ/latxa-instruct-682f356091452b0028380804>

691	DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang,	Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate	746
692	Junxiao Song, et al. 2025. Deepseek-r1: Incentiviz-	Kushman, and Hannaneh Hajishirzi. 2016. MAWPS:	747
693	ing reasoning capability in llms via reinforcement	A math word problem repository . In <i>Proceedings of</i>	748
694	learning . <i>arXiv preprint arXiv:2501.12948</i> .	<i>the 2016 Conference of the North American Chapter</i>	749
		<i>of the Association for Computational Linguistics: Hu-</i>	750
695	Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Ald-	<i>man Language Technologies</i> , pages 1152–1157, San	751
696	abe, German Rigau, Eneko Agirre, Aitor Ormazabal,	Diego, California. Association for Computational	752
697	Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An	Linguistics.	753
698	open language model and evaluation suite for Basque .		
699	In <i>Proceedings of the 62nd Annual Meeting of the</i>	Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su.	754
700	<i>Association for Computational Linguistics (Volume 1:</i>	2020. A diverse corpus for evaluating and developing	755
701	<i>Long Papers</i>), pages 14952–14972, Bangkok, Thai-	English math word problem solvers . In <i>Proceedings</i>	756
702	land. Association for Computational Linguistics.	<i>of the 58th Annual Meeting of the Association for</i>	757
		<i>Computational Linguistics</i> , pages 975–984, Online.	758
703	Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri,	Association for Computational Linguistics.	759
704	Abhinav Pandey, and et al. Abhishek Kadian. 2024.		
705	The llama 3 herd of models .	Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari,	760
		Henryk Michalewski, Jacob Austin, David Bieber,	761
706	Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul	David Dohan, Aitor Lewkowycz, Maarten Bosma,	762
707	Arora, Steven Basart, Eric Tang, Dawn Song, and	David Luan, Charles Sutton, and Augustus Odena.	763
708	Jacob Steinhardt. 2021. Measuring mathematical	2022. Show your work: Scratchpads for intermediate	764
709	problem solving with the math dataset. <i>NeurIPS</i> .	computation with language models .	765
710	Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham	Augustus Odena, Charles Sutton, David Martin Do-	766
711	Neubig, Orhan Firat, and Melvin Johnson. 2020.	han, Ellen Jiang, Henryk Michalewski, Jacob Austin,	767
712	Xtreme: A massively multilingual multi-task bench-	Maarten Paul Bosma, Maxwell Nye, Michael Terry,	768
713	mark for evaluating cross-lingual generalization. In	and Quoc V. Le. 2021. Program synthesis with large	769
714	<i>Proceedings of ICML</i> .	language models. In <i>n/a</i> , page n/a, n/a. N/a.	770
715	Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin	OpenAI. 2025. Openai o3: Enhanced reasoning capa-	771
716	Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not	bilities in large language models . Accessed on May	772
717	all languages are created equal in LLMs: Improv-	13, 2025.	773
718	ing multilingual capability by cross-lingual-thought		
719	prompting . In <i>Findings of the Association for Com-</i>	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	774
720	<i>putational Linguistics: EMNLP 2023</i> , pages 12365–	Carroll Wainwright, Pamela Mishkin, Chong Zhang,	775
721	12394, Singapore. Association for Computational	Sandhini Agarwal, Katarina Slama, Alex Ray, John	776
722	Linguistics.	Schulman, Jacob Hilton, Fraser Kelton, Luke Miller,	777
		Maddie Simens, Amanda Askell, Peter Welinder,	778
723	Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and	Paul F Christiano, Jan Leike, and Ryan Lowe. 2022.	779
724	Fei Yuan. 2024. Mindmerger: Efficiently boosting	Training language models to follow instructions with	780
725	LLM reasoning in non-english languages . In <i>The</i>	human feedback . In <i>Advances in Neural Information</i>	781
726	<i>Thirty-eighth Annual Conference on Neural Informa-</i>	<i>Processing Systems</i> , volume 35, pages 27730–27744.	782
727	<i>tion Processing Systems</i> .	Curran Associates, Inc.	783
728	Shima Imani, Liang Du, and Harsh Shrivastava. 2023.	Arkil Patel, Satwik Bhattamishra, and Navin Goyal.	784
729	MathPrompter: Mathematical reasoning using large	2021. Are NLP models really able to solve simple	785
730	language models . In <i>Proceedings of the 61st An-</i>	math word problems? In <i>Proceedings of the 2021</i>	786
731	<i>Annual Meeting of the Association for Computational</i>	<i>Conference of the North American Chapter of the</i>	787
732	<i>Linguistics (Volume 5: Industry Track)</i> , pages 37–	<i>Association for Computational Linguistics: Human</i>	788
733	42, Toronto, Canada. Association for Computational	<i>Language Technologies</i> , pages 2080–2094, Online.	789
734	Linguistics.	Association for Computational Linguistics.	790
735	Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika	Edoardo Maria Ponti, Goran Glavaš, Olga Majewska,	791
736	Bali, and Monojit Choudhury. 2020. The state and	Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020.	792
737	fate of linguistic diversity and inclusion in the NLP	XCOPA: A multilingual dataset for causal common-	793
738	world . In <i>Proceedings of the 58th Annual Meeting of</i>	sense reasoning . In <i>Proceedings of the 2020 Con-</i>	794
739	<i>the Association for Computational Linguistics</i> , pages	<i>ference on Empirical Methods in Natural Language</i>	795
740	6282–6293, Online. Association for Computational	<i>Processing (EMNLP)</i> , pages 2362–2376, Online. As-	796
741	Linguistics.	sociation for Computational Linguistics.	797
742	Rik Koncel-Kedziorski, Hannaneh Hajishirzi, and	Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang,	798
743	Ashish Sabharwal. 2015. Parsing algebraic word	and Wanxiang Che. 2023. Cross-lingual prompt-	799
744	problems into equations. <i>Transactions of the Associ-</i>	ing: Improving zero-shot chain-of-thought reasoning	800
745	<i>ation for Computational Linguistics</i> , 3:585–597.	across languages . In <i>Proceedings of the 2023 Con-</i>	801
		<i>ference on Empirical Methods in Natural Language</i>	802

A Extending DUP Prompting to Multi-Shot Settings

This appendix presents the complete results from our experiments extending the DUP framework beyond its original zero-shot formulation to include one-shot and few-shot configurations. The primary goal of these experiments is to examine how varying levels of in-context supervision influence the effectiveness of DUP’s structured reasoning stages when applied to multilingual mathematical reasoning tasks. By comparing zero-shot, one-shot, and few-shot prompts containing two examples, we aim to assess how limited in-context supervision influences model performance, particularly in low-resource language settings.

This appendix presents the complete results from our experiments extending DUP prompting beyond the original zero-shot formulation to include One-Shot and Few-Shot configurations. The goal of these experiments is to evaluate how increasing in-context supervision interacts with the structured reasoning stages of DUP prompting in multilingual mathematical tasks.

To support transparency, reproducibility, and a deeper understanding of our experimental design, we provide concrete examples of the exact prompts used across different configurations. Specifically, Figures 4, 5, and 6 display the prompt templates and responses for the three DUP stages, *Reveal the Core Question*, *Extract the Problem-Solving Information*, and *Generate and Extract the Answer*, using two representative few-shot examples in Basque. Their English counterparts are shown in Figures 7, 8, and 9, providing a parallel illustration of how in-context examples are incorporated across stages in a high-resource language. These side-by-side examples highlight how structured prompting can be consistently applied across languages, allowing for controlled comparison of reasoning behaviors under different linguistic and supervision conditions.

This approach has been systematically applied in both the primary experiments, which focus on assessing mathematical reasoning in a minority language, Basque, and in complementary evaluations conducted in English. The consistent application of DUP across languages and supervision levels allows us to probe the interaction between linguistic representation, prompt structure, and model inference capabilities, and to assess the scalability of DUP as a multilingual prompting strategy.

Q: Rogerrek 5 teniseko pilota ditu. Teniseko piloten 2 pote gehiago erosi ditu. Pote bakoitzak 3 teniseko pilota ditu. Zenbat teniseko pilota ditu orain?
Mesedez, atera galdera nagusia, bakarrik osatuena eta zehatzena!
A: Zenbat teniseko pilota ditu orain?

Q: Bederatzi ordenagailu zeuden zerbitzari-gelan. Beste bost ordenagailu instalatu ziren egun bakoitzean, astelehenetik ostegunera. Zenbat ordenagailu daude orain zerbitzari-gelan?
Mesedez, atera galdera nagusia, bakarrik osatuena eta zehatzena!
A: Zenbat ordenagailu daude orain zerbitzari-gelan?

Q: Johnnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?
Mesedez, atera galdera nagusia, bakarrik osatuena eta zehatzena!
A: Zenbat ordu ematen ditu astean txakurrak zaintzen?

Figure 4: *Reveal the Core Question* stage using two illustrative examples in Basque. The template is highlighted in bold to indicate its structure and function within the instruction.

Q: Rogerrek 5 teniseko pilota ditu. Teniseko piloten 2 pote gehiago erosi ditu. Pote bakoitzak 3 teniseko pilota ditu. Zenbat teniseko pilota ditu orain?
Oharra: Mesedez, atera galdera ebazteko informazioa (Zenbat teniseko pilota ditu orain?), atera bakarrik informazio erabilgarriena, zerrendatu banan-banan!
A: 1. Roger 5 pilotarekin hasi zen. 2. 2 pote, bakoitza 3 teniseko pilotakoa, 6 teniseko pilota dira. 3. $5 + 6 = 11$.

Q: Bederatzi ordenagailu zeuden zerbitzari-gelan. Beste bost ordenagailu instalatu ziren egun bakoitzean, astelehenetik ostegunera. Zenbat ordenagailu daude orain zerbitzari-gelan?
Oharra: Mesedez, atera galdera ebazteko informazioa (Zenbat ordenagailu daude orain zerbitzari-gelan?), atera bakarrik informazio erabilgarriena, zerrendatu banan-banan!
A: 1. 4 egun daude astelehenetik ostegunera. 2. 5 ordenagailu gehitu ziren egunero. 3. Horrek esan nahi du guztira $4 * 5 = 20$ ordenagailu gehitu zirela. 4. Hasieran 9 ordenagailu zeuden, beraz orain $9 + 20 = 29$ ordenagailu daude.

Q: Johnnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?
Oharra: Mesedez, atera galdera ebazteko informazioa (Zenbat ordu ematen ditu astean txakurrak zaintzen?), atera bakarrik informazio erabilgarriena, zerrendatu banan-banan!
A: 1. Johnnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. 3. Horrek esan nahi du egunero $10 * 0.5 = 5$ ordu behar dituela. 4. Astean 7 egun daude, beraz, $5 * 7 = 35$ ordu behar ditu astean.

Figure 5: *Extract the Problem-Solving Information* stage using two illustrative examples in Basque. The template is highlighted in bold to indicate its structure and function within the instruction, while the question inherited from the *Reveal the Core Question* stage is shown in italics.

Q: Rogerrek 5 teniseko pilota ditu. Teniseko piloten 2 pote gehiago erosi ditu. Pote bakoitzak 3 teniseko pilota ditu. Zenbat teniseko pilota ditu orain?

Aholkua: 1. Roger 5 pilotarekin hasi zen. 2. 2 pote, bakoitza 3 teniseko pilotakoa, 6 teniseko pilota dira. 3. $5 + 6 = 11$. Zenbat teniseko pilota ditu orain?

Mesedez, ulertu Aholkua eta galderaren informazioa, ondoren, galdera pausoz pauso ebatzi eta erantzuna erakutsi.

A: 11

Q: Bederatzi ordenagailu zeuden zerbitzari-gelan. Beste bost ordenagailu instalatu ziren egun bakoitzean, astelehenetik ostegunera. Zenbat ordenagailu daude orain zerbitzari-gelan?

Aholkua: 1. 4 egun daude astelehenetik ostegunera. 2. 5 ordenagailu gehitu ziren egunero. 3. Horrek esan nahi du guztira $4 * 5 = 20$ ordenagailu gehitu zirela. 4. Hasieran 9 ordenagailu zeuden, beraz orain $9 + 20 = 29$ ordenagailu daude. Zenbat ordenagailu daude orain zerbitzari-gelan?

Mesedez, ulertu Aholkua eta galderaren informazioa, ondoren, galdera pausoz pauso ebatzi eta erantzuna erakutsi.

A: 29

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Aholkua: 1. Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. 3. Horrek esan nahi du egunero $10 * 0.5 = 5$ ordu behar dituela. 4. Astean 7 egun daude, beraz, $5 * 7 = 35$ ordu behar ditu astean. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Mesedez, ulertu Aholkua eta galderaren informazioa, ondoren, galdera pausoz pauso ebatzi eta erantzuna erakutsi.

A: 35

Figure 6: *Generate and Extract the Answer* stage with two Basque examples. The template is highlighted in bold to indicate its structure and function within the instruction, while context from the *Reveal the Core Question* and *Extract Problem-Solving Information* stages are shown in italics.

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Please extract the core question, only the most comprehensive and detailed one!

A: How many tennis balls does he have now?

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

Please extract the core question, only the most comprehensive and detailed one!

A: How many computers are now in the server room?

Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

Please extract the core question, only the most comprehensive and detailed one!

A: How many hours a week does he spend taking care of dogs?

Figure 7: *Reveal the Core Question* stage using two illustrative examples in English. The template is highlighted in bold to indicate its structure and function within the instruction.

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Note: Please extract the question-solving information related to the problem (How many tennis balls does he have now?), only extract the most useful information, and list them one by one!

A: 1. Roger started with 5 balls. 2. 2 cans of 3 tennis balls each is 6 tennis balls. 3. $5 + 6 = 11$.

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

Note: Please extract the question-solving information related to the problem (How many computers are now in the server room?), only extract the most useful information, and list them one by one!

A: 1. There are 4 days from monday to thursday. 2. 5 computers were added each day. 3. That means in total $4 * 5 = 20$ computers were added. 4. There were 9 computers in the beginning, so now there are $9 + 20 = 29$ computers.

Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

Note: Please extract the question-solving information related to the problem (How many hours a week does he spend taking care of dogs?), only extract the most useful information, and list them one by one!

A: 1. John takes care of 10 dogs. 2. Each dog takes 0.5 hours a day to walk and take care of their business. There are 7 days in a week.

Figure 8: *Extract the Problem-Solving Information* stage using two illustrative examples in English. The template is highlighted in bold to indicate its structure and function within the instruction, while the question inherited from the *Reveal the Core Question* stage is shown in italics.

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Hint: 1. Roger started with 5 balls. 2. 2 cans of 3 tennis balls each is 6 tennis balls. 3. $5 + 6 = 11$. How many tennis balls does he have now?

Please understand the Hint and question information, then solve the question step by step and show the answer.

A: 11

Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?

Hint: 1. There are 4 days from monday to thursday. 2. 5 computers were added each day. 3. That means in total $4 * 5 = 20$ computers were added. 4. There were 9 computers in the beginning, so now there are $9 + 20 = 29$ computers. How many computers are now in the server room?

Please understand the Hint and question information, then solve the question step by step and show the answer.

A: 29

Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

Hint: 1. John takes care of 10 dogs. 2. Each dog takes 0.5 hours a day to walk and take care of their business. There are 7 days in a week. How many hours a week does he spend taking care of dogs?

Please understand the Hint and question information, then solve the question step by step and show the answer.

A: 35

Figure 9: *Generate and Extract the Answer* stage with two English examples. The template is highlighted in bold to indicate its structure and function within the instruction, while context from the *Reveal the Core Question* and *Extract Problem-Solving Information* stages are shown in italics.

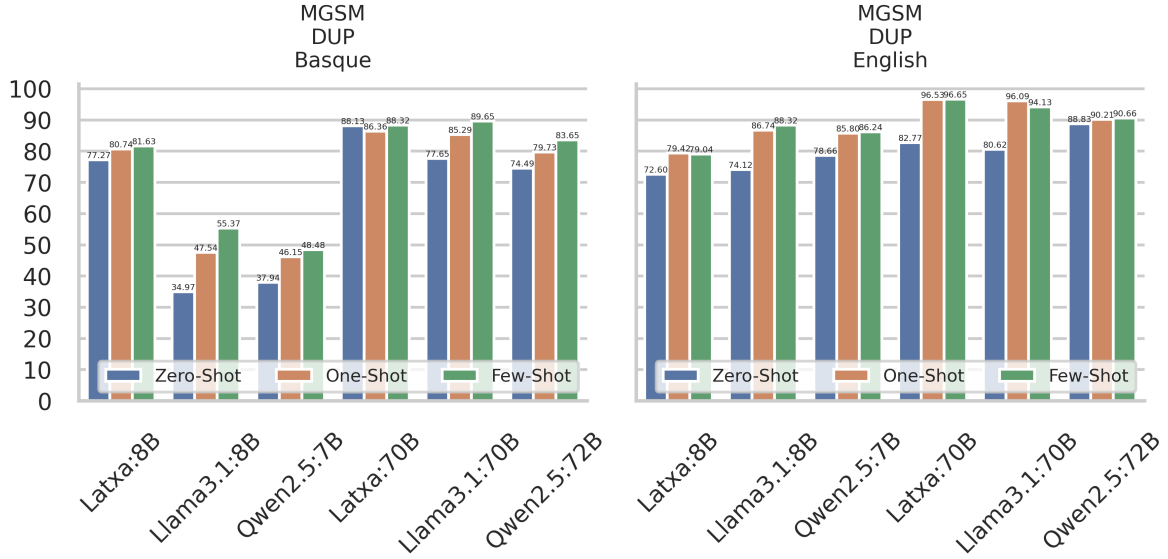


Figure 10: Comparative evaluation of Zero-Shot, One-Shot, and Few-Shot prompting techniques, in conjunction with the DUP prompting strategy, on the MGSM dataset. Results are reported for both Basque and English to assess multilingual reasoning performance.

The performance results for each model and shot configuration on the MGSM benchmark are visualized in Figure 10, while Table 6 reports the full detailed metrics for both MGSM and MASEU. These results enable a detailed comparison of model behavior under varying levels of supervision, helping to assess the combined impact of prompt structure and contextual information on reasoning accuracy across languages. Together, these materials provide a more complete picture of how DUP can be adapted for more flexible and effective use in few-shot prompting scenarios.

B Detailed Results on DUP vs. Standard Prompting

This appendix provides the complete set of experimental results comparing DUP and non-DUP prompting strategies across all evaluated models, benchmarks, and language configurations in the few-shot setting. Figures 11 and 12 visualize performance on the MASEU and MGSM benchmarks, respectively, highlighting how multilingual reasoning is affected by structured versus conventional prompting techniques.

These visualizations enable a detailed comparison of prompting effects, revealing how reasoning quality is shaped by prompt structure and model type. The results highlight the particular effectiveness of structured guidance in low-resource contexts like Basque. In such settings, well-designed

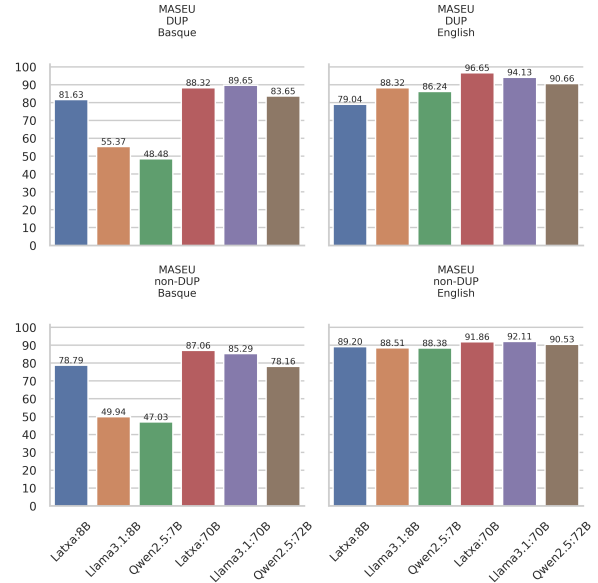


Figure 11: Performance comparison of DUP and non-DUP prompting on the Basque subset of the MASEU dataset under Few-Shot settings. The figure shows how structured prompting using DUP affects model performance across different architectures and scales, highlighting the consistent benefits of DUP in low-resource language contexts.

prompts help guide the model’s output more effectively, improving both accuracy and interpretability. This underscores the importance of prompt formulation in adapting general-purpose models to underrepresented languages.

Dataset	Technique	Model	Zero-Shot	One-Shot (Δ)	Few-Shot (Δ)
MASEU	DUP	Latxa:8B	73.55	80.21 (+6.66)	79.33(+5.78)
		Llama3.1:8B	49.67	61.93(+12.26)	68.66 (+18.99)
		Qwen2.5:7B	48.67	55.43(+6.76)	57.20 (+8.53)
		Latxa:70B	86.68	91.12(+4.44)	91.84 (+5.16)
		Llama3.1:70B	80.08	89.70(+9.62)	90.64 (+10.56)
		Qwen2.5:72B	77.91	81.31(+3.40)	81.71 (+3.80)
	non-DUP	Latxa:8B	81.16	82.36(+1.20)	84.00 (+2.84)
		Llama3.1:8B	66.88	65.62(−1.26)	69.22 (+2.34)
		Qwen2.5:7B	61.68	65.25(+3.57)	67.70 (+6.02)
		Latxa:70B	89.27	88.63(−0.64)	89.46 (+0.19)
		Llama3.1:70B	84.09	87.56(+3.47)	88.70 (+4.61)
		Qwen2.5:72B	77.94	82.16(+4.22)	84.34 (+6.40)
MASEU	DUP	Latxa:8B	53.00	55.50 (+2.50)	53.30(+0.30)
		Llama3.1:8B	26.00	36.60(+10.60)	38.00 (+12.00)
		Qwen2.5:7B	30.70	31.40(+0.70)	32.40 (+1.70)
		Latxa:70B	75.60	79.10(+3.50)	81.10 (+5.50)
		Llama3.1:70B	51.90	68.70(+16.80)	73.50 (+21.60)
		Qwen2.5:72B	55.10	58.50(+3.40)	60.00 (+4.90)
	non-DUP	Latxa:8B	60.60	61.60(+1.00)	62.20 (+1.60)
		Llama3.1:8B	50.60	49.00(−1.60)	47.20(−3.40)
		Qwen2.5:7B	44.80	48.00 (+3.20)	47.80(+3.00)
		Latxa:70B	82.00	83.60 (+1.60)	81.80(−0.20)
		Llama3.1:70B	74.00	74.00(0.00)	76.00 (+2.00)
		Qwen2.5:72B	40.00	44.20(+4.20)	49.20 (+9.20)
AVG			57.43	61.06(+3.63)	62.51 (+5.08)

Table 6: Accuracy results of different prompting techniques, Zero-Shot, One-Shot, and Few-Shot, evaluated on the MASEU and MGSM datasets, using both DUP and non-DUP prompting strategies across various language settings.

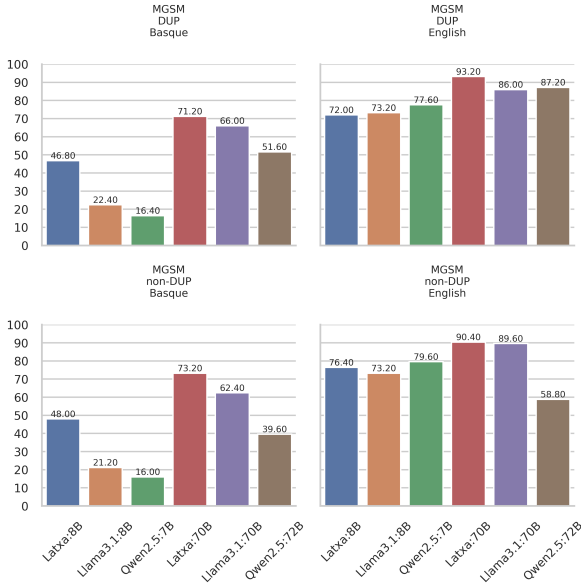


Figure 12: Performance comparison of DUP and non-DUP prompting on the Basque subset of the *MGSM* dataset under Few-Shot settings. The figure shows how structured prompting using DUP affects model performance across different architectures and scales, highlighting the consistent benefits of DUP in low-resource language contexts.

To complement the visual summaries, the full set of disaggregated accuracy scores, covering all combinations of model, dataset, prompting strat-

egy, and language configuration, is provided in Table 7. This tabular data supports detailed analysis and offers a complete accounting of experimental outcomes for all evaluated settings.

C Language Key Configurations in DUP Prompting

This appendix provides a detailed overview of the multilingual prompting configurations used to evaluate the impact of language mixing within the DUP prompting framework. Specifically, we present the variants in which Basque and English are combined in different components of the prompt, including the MWP input, the template instructions, and the expected model responses. These configurations are designed to isolate the effect of language usage within structurally guided reasoning tasks and to better understand how models handle multilingual contexts, particularly when switching between a low-resource language such as Basque and a high-resource language like English.

The appendix includes representative prompt examples for each configuration, as shown in the accompanying figures. These examples illustrate how each stage of the DUP strategy, core question extraction, problem-solving information identification, and final answer generation, was constructed

Language	Shots	Model	non-DUP	DUP	Language	Shots	Model	non-DUP	DUP
Basque	Zero-Shot	Latxa:8B	73.67	77.27	Basque	Zero-Shot	Latxa:8B	44.40	47.60
		Llama3.1:8B	45.39	34.97			Llama3.1:8B	28.40	17.60
		Qwen2.5:7B	36.68	37.94			Qwen2.5:7B	11.60	14.40
		Latxa:70B	85.35	88.13			Latxa:70B	70.40	75.60
		Llama3.1:70B	80.18	77.65			Llama3.1:70B	64.00	46.40
		Qwen2.5:72B	67.80	74.49			Qwen2.5:72B	26.80	42.80
	One-Shot	Latxa:8B	75.95	80.74		One-Shot	Latxa:8B	48.40	47.60
		Llama3.1:8B	41.10	47.54			Llama3.1:8B	19.20	21.60
		Qwen2.5:7B	42.61	46.15			Qwen2.5:7B	17.20	14.00
		Latxa:70B	84.85	86.36			Latxa:70B	75.60	73.20
		Llama3.1:70B	84.53	85.29			Llama3.1:70B	62.40	59.60
		Qwen2.5:72B	74.56	79.73			Qwen2.5:72B	35.20	48.00
	Few-Shot	Latxa:8B	78.79	81.63		Few-Shot	Latxa:8B	48.00	46.80
		Llama3.1:8B	49.94	55.37			Llama3.1:8B	21.20	22.40
		Qwen2.5:7B	47.03	48.48			Qwen2.5:7B	16.00	16.40
		Latxa:70B	87.06	88.32			Latxa:70B	73.20	71.20
		Llama3.1:70B	85.29	89.65			Llama3.1:70B	62.40	66.00
		Qwen2.5:72B	78.16	83.65			Qwen2.5:72B	39.60	51.60
English	Zero-Shot	Latxa:8B	88.64	72.60	English	Zero-Shot	Latxa:8B	76.80	62.40
		Llama3.1:8B	88.38	74.12			Llama3.1:8B	72.80	45.60
		Qwen2.5:7B	86.68	78.66			Qwen2.5:7B	78.00	75.20
		Latxa:70B	93.18	82.77			Latxa:70B	93.60	76.40
		Llama3.1:70B	88.01	80.62			Llama3.1:70B	84.00	63.60
		Qwen2.5:72B	88.07	88.83			Qwen2.5:72B	53.20	80.80
	One-Shot	Latxa:8B	88.76	79.42		One-Shot	Latxa:8B	74.80	68.00
		Llama3.1:8B	90.15	86.74			Llama3.1:8B	78.80	68.40
		Qwen2.5:7B	87.88	85.80			Qwen2.5:7B	78.80	74.80
		Latxa:70B	92.42	96.53			Latxa:70B	91.60	90.40
		Llama3.1:70B	90.59	96.09			Llama3.1:70B	85.60	79.60
		Qwen2.5:72B	89.77	90.21			Qwen2.5:72B	53.20	86.00
	Few-Shot	Latxa:8B	89.20	79.04		Few-Shot	Latxa:8B	76.40	72.00
		Llama3.1:8B	88.51	88.32			Llama3.1:8B	73.20	73.20
		Qwen2.5:7B	88.38	86.24			Qwen2.5:7B	79.60	77.60
		Latxa:70B	91.86	96.65			Latxa:70B	90.40	93.20
		Llama3.1:70B	92.11	94.13			Llama3.1:70B	89.60	86.00
		Qwen2.5:72B	90.53	90.66			Qwen2.5:72B	58.80	87.20

Table 7: Comparison of DUP and non-DUP prompting performance across Zero-Shot, One-Shot, and Few-Shot settings. Results are shown for the MASEU dataset (right) and the MGSM dataset (left).

under varying language conditions. Together, they offer insight into how linguistic composition within a prompt may influence reasoning performance, and serve as a reference for reproducibility and further analysis of language-sensitive prompting strategies.

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?
Mesedez, atera galdera nagusia, bakarrik osatuena eta zehatzena!
A: Zenbat ordu ematen ditu astean txakurrak zaintzen?

Figure 13: *Reveal the Core Question* stage, in which the MWP, the template, highlighted in bold, and the corresponding response are all entirely expressed in Basque.

C.1 Basque

In the following prompts, the text is presented entirely in Basque. For this case, the translated version of the original dataset has been used. The language model is able to recognize the linguistic context and continues the interaction by responding in the same language.

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Oharra: Mesedez, atera galdera ebazteko informazioa (*Zenbat ordu ematen ditu astean txakurrak zaintzen?*), atera bakarrik informazio erabilgarriena, zerrendatu banan-banan!

A: 1. Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. 3. Horrek esan nahi du egunero $10 * 0.5 = 5$ ordu behar dituela. 4. Astean 7 egun daude, beraz, $5 * 7 = 35$ ordu behar ditu astean.

Figure 14: *Extract the Problem-solving Information* stage, in which the MWP, the template, highlighted in bold, the core question from the *Reveal the Core Question* stage, in italics, and the corresponding response are all entirely expressed in Basque.

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Aholkua: 1. *Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. 3. Horrek esan nahi du egunero $10 * 0.5 = 5$ ordu behar dituela. 4. Astean 7 egun daude, beraz, $5 * 7 = 35$ ordu behar ditu astean. Zenbat ordu ematen ditu astean txakurrak zaintzen?*

Mesedez, ulertu Aholkua eta galderaren informazioa, ondoren, galdera pausoz pauso ebatzi eta erantzuna erakutsi.

A: 35

Figure 15: *Generate and Extract the Answer in Basque* stage in Basque, in which the MWP, the template, highlighted in bold, the core question from the *Reveal the Core Question* stage, in italics, and the information from the *Extract Problem-Solving Information* stage, also in italics, are all entirely expressed in Basque.

C.2 Mixed with English Templates

In the following prompts, it can be observed that the templates used are written in English, whereas both the original problem statement and the model's responses are provided in Basque. The model is explicitly instructed regarding the language to be used for each component of the interaction, including the input elements and the expected output.

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Please extract the core question, only the most comprehensive and detailed one!

The provided sentence is in Basque.

Please, output the answer in Basque.

A: Zenbat ordu ematen ditu astean txakurrak zaintzen?

Figure 16: *Reveal the Core Question* stage, in which both the MWP and the corresponding response are expressed in Basque, while the template, highlighted in bold, is presented in English.

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Note: Please extract the question-solving information related to the problem (*Zenbat ordu ematen ditu astean txakurrak zaintzen?*), only extract the most useful information, and list them one by one!

The provided sentence is in Basque and question is in Basque.

Please, output the answer in Basque.

A: 1. Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. 3. Horrek esan nahi du egunero $10 * 0.5 = 5$ ordu behar dituela. 4. Astean 7 egun daude, beraz, $5 * 7 = 35$ ordu behar ditu astean.

Figure 17: *Extract the Problem-solving Information* stage, in which the MWP, the core question from the *Reveal the Core Question* stage, in italics, and the corresponding response are expressed in Basque, while the template, highlighted in bold, is presented in English.

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Hint: 1. *Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. 3. Horrek esan nahi du egunero $10 * 0.5 = 5$ ordu behar dituela. 4. Astean 7 egun daude, beraz, $5 * 7 = 35$ ordu behar ditu astean. Zenbat ordu ematen ditu astean txakurrak zaintzen?*

Please understand the Hint and question information, then solve the question step by step and show the answer.

The provided sentence is in Basque and the question and the numbered steps are in Basque.

A: 35

Figure 18: *Generate and Extract the Answer* stage, in which the MWP, the core question from the *Reveal the Core Question* stage, in italics, and the information from the *Extract Problem-Solving Information* stage, also in italics, are expressed in Basque, while the template, highlighted in bold, is presented in English.

C.3 Mixed with English Templates and Responses

In the following prompts, the use of English extends from the templates to the generated responses, with only the original problem statement preserved in Basque. The model is explicitly instructed about the language to be used for each component of the task, including both the input segments and the expected output.

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Please extract the core question, only the most comprehensive and detailed one!

The provided sentence is in Basque.

Please, output the answer in English.

A: How many hours a week does he spend taking care of dogs?

Figure 19: *Reveal the Core Question* stage, in which the MWP is expressed in Basque, while the template, highlighted in bold, and the corresponding response are presented in English.

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Note: Please extract the question-solving information related to the problem (*How many hours a week does he spend taking care of dogs?*), only extract the most useful information, and list them one by one! The provided sentence is in Basque and question is in English. Please, output the answer in English.

A: 1. John takes care of 10 dogs. 2. Each dog takes 0.5 hours a day to walk and take care of their business. There are 7 days in a week.

Figure 20: *Extract the Problem-solving Information* stage, in which the MWP is expressed in Basque, while the template, highlighted in bold, the core question from the *Reveal the Core Question* stage, in italics, and the corresponding response are presented in English.

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Hint: 1. John takes care of 10 dogs. 2. Each dog takes 0.5 hours a day to walk and take care of their business. There are 7 days in a week. *How many hours a week does he spend taking care of dogs?*

Please understand the Hint and question information, then solve the question step by step and show the answer.

The provided sentence is in Basque and the question and the numbered steps are in English.

A: 35

Figure 21: *Generate and Extract the Answer* stage, in which the MWP is expressed in Basque, while the template, highlighted in bold, the core question from the *Reveal the Core Question* stage, in italics, and the information from the *Extract Problem-Solving Information* stage, also in italics, are presented in English.

C.4 English

In the following prompts, the text is presented entirely in English. For this case, the English version of the dataset is used. The language model recognizes the input language and continues the interaction accordingly, generating responses in the same language.

Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

Please extract the core question, only the most comprehensive and detailed one!

A: How many hours a week does he spend taking care of dogs?

Figure 22: *Reveal the Core Question* stage, in which the MWP, the template, highlighted in bold, and the corresponding response are all entirely expressed in English.

Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

Note: Please extract the question-solving information related to the problem (*How many hours a week does he spend taking care of dogs?*), only extract the most useful information, and list them one by one!

A: 1. John takes care of 10 dogs. 2. Each dog takes 0.5 hours a day to walk and take care of their business. There are 7 days in a week.

Figure 23: *Extract the Problem-solving Information* stage, in which the MWP, the template, highlighted in bold, the core question from the *Reveal the Core Question* stage, in italics, and the corresponding response are all entirely expressed in English.

Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

Hint: 1. John takes care of 10 dogs. 2. Each dog takes 0.5 hours a day to walk and take care of their business. There are 7 days in a week. *How many hours a week does he spend taking care of dogs?*

Please understand the Hint and question information, then solve the question step by step and show the answer.

A: 35

Figure 24: *Generate and Extract the Answer in Basque* stage in Basque, in which the MWP, the template, highlighted in bold, the core question from the *Reveal the Core Question* stage, in italics, and the information from the *Extract Problem-Solving Information* stage, also in italics, are all entirely expressed in English.

D Prompts

Following the prompts presented in Appendix C, this section provides the complete prompts along with their corresponding responses for the example MWPs across the three stages of the DUP technique. Additionally, variations introduced by the use of different languages are incorporated. These examples illustrate the full procedural application of the DUP methodology, demonstrating how language shifts and iterative prompt-response interactions influence the structure and resolution of MWPs. This detailed presentation aims to offer a clearer understanding of the linguistic and methodological adaptations involved in the DUP strategy across multilingual contexts.

D.1 Reveal the Core Question

D.1.1 Basque

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Mesedez, atera galdera nagusia, bakarrik osatuen eta zehatzenena!

1092	A: Zenbat ordu ematen ditu astean txakurrak zaintzen?	kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?	1134
			1135
1093	D.1.2 Mixed with English Templates	Oharra: Mesedez, atera galdera ebazteko informazioa (Zenbat ordu ematen ditu astean txakurrak zaintzen?), atera bakarrik informazio erabilgarriena, zerrendatu banan-banan!	1136
1094			1137
1095			1138
1096	Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?	A: 1. Johnek 10 txakur zaintzen ditu.	1139
1097		2. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko.	1140
1098		3. Horrek esan nahi du egunero $10 \times 0.5 = 5$ ordu behar dituela.	
1099		4. Astean 7 egun daude, beraz, $5 \times 7 = 35$ ordu behar ditu astean.	
1100	Please extract the core question, only the most comprehensive and detailed one!		
1101	The provided sentence is in Basque.		
1102	Please, output the answer in Basque.		
1103	A: Zenbat ordu ematen ditu astean txakurrak zaintzen?		
1104			
		D.2.2 Mixed with English Templates	1141
			1142
1105	D.1.3 Mixed with English Templates and Responses		1143
1106			1144
1107		Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?	1145
1108			1146
1109	Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?	Note: Please extract the question-solving information related to the problem (Zenbat ordu ematen ditu astean txakurrak zaintzen?), only extract the most useful information, and list them one by one!	1147
1110		The provided sentence is in Basque and question is in Basque.	1148
1111		Please, output the answer in Basque.	1149
1112			1150
1113	Please extract the core question, only the most comprehensive and detailed one!		1151
1114	The provided sentence is in Basque.		1152
1115	Please, output the answer in English.		1153
1116	A: How many hours a week does John spend taking care of dogs?	A: 1. Johnek 10 txakur zaintzen ditu.	1154
1117		2. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko.	1155
		3. Horrek esan nahi du egunero $10 \times 0.5 = 5$ ordu behar dituela.	1156
		4. Astean 7 egun daude, beraz, $5 \times 7 = 35$ ordu behar ditu astean.	
1118	D.1.4 English		
1119			
1120			
1121	Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?	D.2.3 Mixed with English Templates and Responses	1157
1122			1158
1123			1159
1124			1160
1125	Please extract the core question, only the most comprehensive and detailed one!	Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?	1161
1126			1162
1127	A: How many hours a week does John spend taking care of dogs?	Note: Please extract the question-solving information related to the problem (How many hours a week does John spend taking care of dogs?), only extract the most useful information, and list them one by one!	1163
		The provided sentence is in Basque and question	1164
1128	D.2 Extract the Problem-solving Information		1165
1129	D.2.1 Basque		1166
1130			1167
1131			1168
1132	Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren		1169
1133			1170

is in English.

Please, output the answer in English.

A: 1. John takes care of 10 dogs.
2. Each dog takes 0.5 hours a day to walk and take care of their business.
3. There are 7 days in a week.

D.2.4 English

Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

Note: Please extract the question-solving information related to the problem (How many hours a week does John spend taking care of dogs?), only extract the most useful information, and list them one by one!

A: 1. John takes care of 10 dogs.
2. Each dog takes 0.5 hours a day to walk and take care of their business.
3. There are 7 days in a week.

D.3 Generate and Extract the Answer

D.3.1 Basque

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Aholkua: 1. Johnek 10 txakur zaintzen ditu. 2. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. 3. Horrek esan nahi du egunero $10 * 0.5 = 5$ ordu behar dituela. 4. Astean 7 egun daude, beraz, $5 * 7 = 35$ ordu behar ditu astean.

Zenbat ordu ematen ditu astean txakurrak zaintzen?

Mesedez, ulertu Aholkua eta galderaren informazioa, ondoren, galdera pausoz pauso ebatzi eta erantzuna erakutsi.

A: 35

D.3.2 Mixed with English Templates

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren

kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Hint: 1. Johnek 10 txakur zaintzen ditu. 2. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. 3. Horrek esan nahi du egunero $10 * 0.5 = 5$ ordu behar dituela. 4. Astean 7 egun daude, beraz, $5 * 7 = 35$ ordu behar ditu astean.

Zenbat ordu ematen ditu astean txakurrak zaintzen?

Please understand the Hint and question information, then solve the question step by step and show the answer.

The provided sentence is in Basque and the question and the numbered steps are in Basque.

A: 35

D.3.3 Mixed with English Templates and Responses

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurrak zaintzen?

Hint: 1. John takes care of 10 dogs. 2. Each dog takes 0.5 hours a day to walk and take care of their business. 3. There are 7 days in a week.

How many hours a week does John spend taking care of dogs?

Please understand the Hint and question information, then solve the question step by step and show the answer.

The provided sentence is in Basque and the question and the numbered steps are in English.

A: 35

D.3.4 English

Q: John takes care of 10 dogs. Each dog takes .5 hours a day to walk and take care of their business. How many hours a week does he spend taking care of dogs?

Hint: 1. John takes care of 10 dogs. 2. Each dog takes 0.5 hours a day to walk and take care of their business. 3. There are 7 days in a week.

How many hours a week does John spend taking care of dogs?

Please understand the Hint and question information, then solve the question step by step and show the answer.

A: 35

E Results

The complete set of experimental results for establishing the performance benchmarks of the various tested LLMs is presented in this section. These experiments were conducted using DUP technique, which incorporates iterative prompt-response mixing and integrates language variations throughout its stages, and the standard non-DUP prompting approach. The results provide a comprehensive overview of the models' mathematical reasoning capabilities across both high-resource and low-resource languages. This extensive evaluation serves as a foundational reference for understanding the interplay between model scale, language specialization, and prompting methodology in multilingual mathematical reasoning tasks.

E.1 MASEU

Model	Basque	Mixed (T) Δ	Mixed (T, R) Δ
Latxa:8B	77.27	-3.66	-6.57
Llama3.1:8B	34.97	-5.30	+24.94
Qwen2.5:7B	37.94	+2.21	0.00
Latxa:70B	88.13	+1.01	-1.45
Llama3.1:70B	77.65	+1.96	+4.80
Qwen2.5:72B	74.49	+2.08	-2.78

Table 8: Results of performance metrics of DUP prompting with Zero-Shot for MASEU with different language configurations.

Model	Basque	Mixed (T) Δ	Mixed (T, R) Δ
Latxa:8B	80.74	-1.07	+0.25
Llama3.1:8B	47.54	-2.97	+21.34
Qwen2.5:7B	46.15	+3.79	-6.31
Latxa:70B	86.36	+3.28	+5.56
Llama3.1:70B	85.29	+2.71	+4.10
Qwen2.5:72B	79.73	+1.89	-6.06

Table 9: Results of performance metrics of DUP prompting with One-Shot for MASEU with different language configurations.

Model	Basque	Mixed (T) Δ	Mixed (T, R) Δ
Latxa:8B	81.63	-6.69	+0.06
Llama3.1:8B	55.37	+2.65	+17.55
Qwen2.5:7B	48.48	+1.52	-4.42
Latxa:70B	88.32	+1.89	+3.85
Llama3.1:70B	89.65	-0.44	-0.06
Qwen2.5:72B	83.65	-4.86	-9.91

Table 10: Results of performance metrics of DUP prompting with Few-Shot for MASEU with different language configurations.

Model	Basque	English Δ
Latxa:8B	77.27	-4.67
Llama3.1:8B	34.97	+39.14
Qwen2.5:7B	37.94	+40.72
Latxa:70B	88.13	-5.37
Llama3.1:70B	77.65	+2.97
Qwen2.5:72B	74.49	+14.33

Table 11: Results of performance metrics of DUP prompting with Zero-Shot for MASEU with Basque and English languages.

Model	Basque	English Δ
Latxa:8B	80.74	-1.33
Llama3.1:8B	47.54	+39.20
Qwen2.5:7B	46.15	+39.65
Latxa:70B	86.36	+10.16
Llama3.1:70B	85.29	+10.80
Qwen2.5:72B	79.73	+10.48

Table 12: Results of performance metrics of DUP prompting with One-Shot for MASEU with Basque and English languages.

Model	Basque	English Δ
Latxa:8B	81.63	-2.59
Llama3.1:8B	55.37	+32.95
Qwen2.5:7B	48.48	+37.75
Latxa:70B	88.32	+8.33
Llama3.1:70B	89.65	+4.48
Qwen2.5:72B	83.65	+7.01

Table 13: Results of performance metrics of DUP prompting with Few-Shot for MASEU with Basque and English languages.

Model	Basque	English Δ
Latxa:8B	73.67	+14.96
Llama3.1:8B	45.39	+42.99
Qwen2.5:7B	36.68	+50.00
Latxa:70B	85.35	+7.83
Llama3.1:70B	80.18	+7.83
Qwen2.5:72B	67.80	+20.27

Table 14: Results of performance metrics of non-DUP prompting with Zero-Shot for MASEU with Basque and English languages.

Model	Basque	English Δ
Latxa:8B	75.95	+12.82
Llama3.1:8B	41.10	+49.05
Qwen2.5:7B	42.61	+45.27
Latxa:70B	84.85	+7.58
Llama3.1:70B	84.53	+6.06
Qwen2.5:72B	74.56	+15.21

Table 15: Results of performance metrics of non-DUP prompting with One-Shot for MASEU with Basque and English languages.

Model	Basque	English Δ
Latxa:8B	78.79	+10.42
Llama3.1:8B	49.94	+38.57
Qwen2.5:7B	47.03	+41.35
Latxa:70B	87.06	+4.80
Llama3.1:70B	85.29	+6.82
Qwen2.5:72B	78.16	+12.37

Table 16: Results of performance metrics of non-DUP prompting with Few-Shot for MASEU with Basque and English languages.

Model	Basque	Mixed (T) Δ	Mixed (T, R) Δ
Latxa:8B	47.60	-0.80	+7.60
Llama3.1:8B	17.60	-6.80	+12.40
Qwen2.5:7B	14.40	+0.40	+4.00
Latxa:70B	75.60	-1.60	+0.80
Llama3.1:70B	46.40	-6.40	+11.20
Qwen2.5:72B	42.80	+6.40	+4.80

Table 17: Results of performance metrics of DUP prompting with Zero-Shot for MGSM with different language configurations.

Model	Basque	English Δ
Latxa:8B	47.60	+14.80
Llama3.1:8B	17.60	+28.00
Qwen2.5:7B	14.40	+60.80
Latxa:70B	75.60	+0.80
Llama3.1:70B	46.40	+17.20
Qwen2.5:72B	42.80	+38.00

Table 20: Results of performance metrics of DUP prompting with Zero-Shot for MGSM with Basque and English languages.

Model	Basque	English Δ
Latxa:8B	44.40	+32.40
Llama3.1:8B	28.40	+44.40
Qwen2.5:7B	11.60	+66.40
Latxa:70B	70.40	+23.20
Llama3.1:70B	64.00	+20.00
Qwen2.5:72B	26.80	+26.40

Table 23: Results of performance metrics of non-DUP prompting with Zero-Shot for MGSM with Basque and English languages.

Model	Basque	Mixed (T) Δ	Mixed (T, R) Δ
Latxa:8B	47.60	+0.80	+10.40
Llama3.1:8B	21.60	-1.20	+14.40
Qwen2.5:7B	14.00	+4.00	+4.80
Latxa:70B	73.20	+2.00	+4.40
Llama3.1:70B	59.60	+6.80	+9.60
Qwen2.5:72B	48.00	+2.00	+2.00

Table 18: Results of performance metrics of DUP prompting with One-Shot for MGSM with different language configurations.

Model	Basque	English Δ
Latxa:8B	47.60	+20.40
Llama3.1:8B	21.60	+46.80
Qwen2.5:7B	14.00	+60.80
Latxa:70B	73.20	+17.20
Llama3.1:70B	59.60	+20.00
Qwen2.5:72B	48.00	+38.00

Table 21: Results of performance metrics of DUP prompting with One-Shot for MGSM with Basque and English languages.

Model	Basque	English Δ
Latxa:8B	48.40	+26.40
Llama3.1:8B	19.20	+59.60
Qwen2.5:7B	17.20	+61.60
Latxa:70B	75.60	+16.00
Llama3.1:70B	62.40	+23.20
Qwen2.5:72B	35.20	+18.00

Table 24: Results of performance metrics of non-DUP prompting with One-Shot for MGSM with Basque and English languages.

Model	Basque	Mixed (T) Δ	Mixed (T, R) Δ
Latxa:8B	46.80	-8.80	+9.60
Llama3.1:8B	22.40	-4.40	+16.00
Qwen2.5:7B	16.40	-0.40	+3.20
Latxa:70B	71.20	+6.40	+11.20
Llama3.1:70B	66.00	+2.40	+7.60
Qwen2.5:72B	51.60	-4.80	+2.80

Table 19: Results of performance metrics of DUP prompting with Few-Shot for MGSM with different language configurations.

Model	Basque	English Δ
Latxa:8B	46.80	+25.20
Llama3.1:8B	22.40	+50.80
Qwen2.5:7B	16.40	+61.20
Latxa:70B	71.20	+22.00
Llama3.1:70B	66.00	+20.00
Qwen2.5:72B	51.60	+35.60

Table 22: Results of performance metrics of DUP prompting with Few-Shot for MGSM with Basque and English languages.

Model	Basque	English Δ
Latxa:8B	48.00	+28.40
Llama3.1:8B	21.20	+52.00
Qwen2.5:7B	16.00	+63.60
Latxa:70B	73.20	+17.20
Llama3.1:70B	62.40	+27.20
Qwen2.5:72B	39.60	+19.20

Table 25: Results of performance metrics of non-DUP prompting with Few-Shot for MGSM with Basque and English languages.