

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

Iñigo Martínez-Criado¹, Aitor Soroa^{1,2}, Jeremy Barnes^{1,2}

¹University of the Basque Country UPV/EHU, ²HiTZ Center – Ixa
inigomartinez@gmail.com, {a.soroa, jeremy.barnes}@ehu.eus

Abstract

Large Language Models (LLMs) have shown impressive performance on tasks requiring complex reasoning, but most evaluations tend to focus on English and other high-resource languages. 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. We then use this dataset to address three key questions: 1) how well do LLMs support Basque in reasoning tasks, 2) to what extent can including English in prompts improve results, and 3) what is the effect of continued pretraining in Basque? To explore these aspects, we use prompting strategies adapted for mathematical reasoning, building upon the foundations of *CoT prompting* and one of its subsequent evolutions, *DUP prompting*, which together allow for more precise experimentation across zero-shot and few-shot settings, providing insights into how multilingual models handle reasoning tasks in underrepresented languages.

Keywords: Natural Language Processing, Large Language Models, Low-Resource Languages, Minority Languages, Mathematical Reasoning, Math Word Problems, Basque

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).

Mathematical reasoning (Roy and Roth, 2015; Hendrycks et al., 2021) in particular 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.

Much improvement in mathematical reasoning 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 demonstrated effectiveness for mathematical reasoning, other approaches have also aimed to better align the prompting process with the specific characteristics of mathematical problems. Among these, Deeply Understanding the Problems (DUP) (Zhong et al., 2024) offers a complementary strategy that can further support and refine reasoning in mathematical contexts.

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 within task-specific prompting techniques, such as CoT or DUP. To address these concerns, we present a novel mathematical reasoning dataset in Basque (MASEU¹), 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?

¹Available at <https://huggingface.co/datasets/inigomartinez/MASEU>.

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* (see Section §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 consider structured prompting strategies for multilingual reasoning, including both *Chain-of-Thought (CoT)* (Wei et al., 2022) and *Deeply Understanding the Problems (DUP)* (Zhong et al., 2024). We conduct a series of experiments applying both CoT (Kojima et al., 2022) and DUP frameworks, systematically decomposing the reasoning process into intermediate steps, and analyze the impact on the results.

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 (SP) approaches of-

ten underperform in multi-step reasoning, motivating new methods such as *Scratchpad* (Nye et al., 2022) and program synthesis (Austin 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. Subsequent work further refined CoT (Kojima et al., 2022), which extended it to zero-shot settings and inspired new prompting paradigms. Building upon this line of research, *Deeply Understanding the Problems (DUP)* (Zhong et al., 2024) introduced a more structured three-stage pipeline, comprising 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 CoT 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 languages, 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

ID	Question	Answer
21	<p>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?</p> <p>Basque: Itsasoko petrolio-hodi bat hautsi da. Ingeniariak petrolio-hodia konpontzen hasi baino lehen, 6522 litro olio isuri dira uretara. Langileek lan egiten zuten bitartean, petrolio-hodiak 5165 litro isuri ditu. Guztira, zenbat litro olio isuri dira uretara?</p>	11687
706	<p>English: A book is on sale for 10% off. If the regular price is 27.9 dollars, what is the sale price?</p> <p>Basque: Liburu bat salgai dago %10 eko deskontuarekin. Prezio arrunta 27.9 dolar bada, zein da salmenta prezioa?</p>	25.11
1651	<p>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?</p> <p>Basque: Halloweenerako Emilyk 5 gozoki jaso zituen auzokideengandik eta 13 bere ahizpa nagusiarengandik. Egunean 9 gozoki bakarrik jaten baditu, zenbat denbora iraungo diote gozokiek?</p>	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.

between high- and low-resource languages.

Building on this line of work, our study applies both CoT prompting and 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 by a single native speaker, without the use of any machine translation tools or automated assistance. The translator followed a fully manual, carefully controlled translation process to ensure both mathematical fidelity and linguistic naturalness, faithfully preserving the original intent, difficulty level, and logical structure of each problem, as illustrated by the examples shown in Table 1. This approach guarantees that the Basque version reflects idiomatic usage while maintaining conceptual equivalence, enabling 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 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

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 and evaluation sizes. MASEU contains 195 development and 1584 test instances per language (English, Basque); MGSM contains 8 development and 250 test instances per 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?
A: **Let's think step by step.**

Figure 1: CoT prompt, where the template, highlighted in bold, is added, which guides the model to reason through the problem step by step.

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. Structured Prompting and Multilingual Adaptation

We adopt two structured prompting strategies to support controlled evaluation of mathematical reasoning in multilingual contexts: *Chain-of-Thought (CoT)* (Kojima et al., 2022) and *Deeply Understanding the Problems (DUP)* (Zhong et al., 2024). In the CoT approach, we follow the strategy used in (Zhong et al., 2024), where we prepend the template "Let's think step by step." to the response. This way, we encourage the model to perform explicit reasoning, as shown in Figure 1.

Regarding DUP, we follow the three-stage pipeline, composed of core question extraction, identification of key problem-solving information, and answer generation, as depicted in Figure 2.

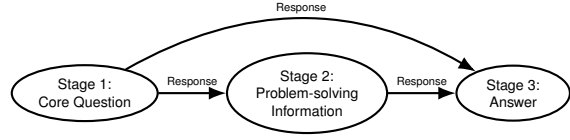


Figure 2: Reasoning flow across DUP stages. The response from Stage 1 is reused in Stages 2 and 3, and the response from Stage 2 is also used in Stage 3.

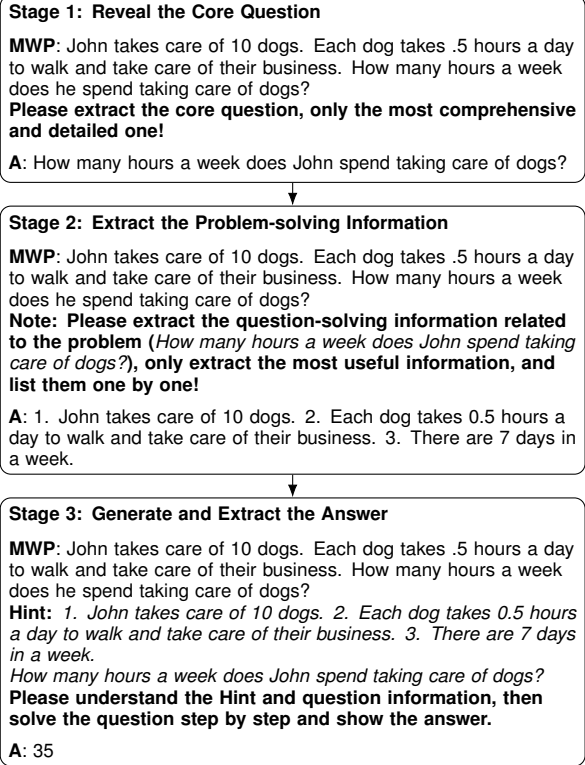


Figure 3: Schematic of the three-phase DUP prompting framework. Input MWP and generated responses use standard font, templates are in bold, and prior-stage responses are italicized. The three stages are: (i) identifying the core question, (ii) extracting problem-solving information, and (iii) generating the final answer based on earlier stages.

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 3 shows an example of the three-phase DUP prompting. Sequential reuse in DUP allows systematic variation of the language used in each prompt, response, and previous response, mixing Basque and English across phases to analyze how multilingual configurations affect reasoning accuracy and transfer. Combined with the language settings in Table 3, this flow diagram illustrates DUP's flexibility for multilingual experimentation.

Although originally designed for zero-shot use, we extend both CoT and DUP to one-shot and

Prompting	Lang.	MWP	Template	Response
CoT, DUP	Basque	EUS	EUS	EUS
	Mixed (T)	EUS	EN	EUS
	English	EN	EN	EN
DUP only	Mixed (T, R)	EUS	EN	EN

Table 3: Language configurations for CoT and DUP prompting. Each row indicates the language used for the MWP, template, and response. CoT covers Basque, Mixed (T), and English, while DUP additionally includes Mixed (T, R). In the Basque setting, all components are in Basque, Mixed (T) uses an English template, Mixed (T, R) applies English to both template and response, and English is used throughout.

few-shot settings by adding in-context examples within each prompt or stage. This enables evaluation of whether contextual supervision enhances reasoning without altering the structure. All few-shot prompts follow the standard CoT and DUP templates for consistency, and both techniques are compared with standard prompting (SP) under equivalent conditions to assess whether structured reasoning or step-by-step guidance provides consistent advantages.

To analyze how language influences reasoning, all Basque prompts were translated from English to preserve mathematical and linguistic fidelity. For CoT, we test fully Basque and Mixed setups, where the template (T) is in English. In DUP, we also include a Mixed (T, R) configuration, where both the template and responses (T, R) are in English while the original problem formulation remains in Basque (see Figures 1 and 3 for examples of templates and responses). This integrated design isolates the effect of language at different reasoning stages and enables direct comparison of CoT and DUP under identical multilingual conditions.

As summarized in Table 3, these configurations clarify how language choice impacts 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 results in Mixed (T, R) could reflect a preference for reasoning entirely in English, the dominant training language. A fully English setting is also included to test whether full alignment with the pretraining language yields further gains.

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 prompting for both CoT and DUP, highlighting differences in Basque performance across models. These disparities are influenced 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 80.87 under CoT and 81.63 under DUP (gains of +2.08 and +2.84 over SP), outperforming similarly sized general-purpose models such as Llama3.1:8B (+14.01 CoT, +5.43 DUP) and Qwen2.5:7B (+3.79 CoT, +1.45 DUP). These results confirm that even under standard prompting, Latxa’s Basque adaptation provides a clear advantage, confirming that LLMs can perform mathematical reasoning in Basque, particularly when the model is specialized for the language.

Performance trends differ between datasets. MASEU generally yields higher scores than MGSM, particularly for Basque. This may reflect either differences in intrinsic problem difficulty or differences in alignment with the models’ training distribution. 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 match patterns seen during multilingual pretraining, potentially contributing to improved performance for both CoT and DUP prompting. At the same time, we cannot rule out that certain MASEU problems are, on average, less structurally demanding than those in MGSM, which could also partially explain the performance gap.

Interestingly, Latxa:70B achieves the highest overall MASEU score under the DUP framework, reaching 92.17 (+5.1) in the Mixed (T, R) configuration, outperforming Llama3.1:70B, which peaks at 89.59 (+4.3). Under CoT, Latxa:70B also performs best (87.56 vs. 83.33), showing that continued Basque pretraining enhances the model’s ability to apply structured and step-by-step reasoning. These results highlight how linguistic adaptation can complement large-scale generalization, yielding stronger reasoning performance in low-resource contexts.

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 Chinese,² which may limit its effectiveness in other languages such as Basque. Meanwhile, Llama3.1,

²<https://qwenlm.github.io/blog/qwen2.5-math/>

	Model	SP	CoT		DUP			CoT	DUP
		Basque	Basque Δ	Mixed (T) Δ	Basque Δ	Mixed (T) Δ	Mixed (T, R) Δ	English	English
MASEU	Latxa:8B	78.79	+2.08	+3.53	+2.84	-3.85	+2.90	90.02	79.04
	Llama3.1:8B	49.94	+14.01	+14.45	+5.43	+8.08	+22.98	<u>90.97</u>	88.32
	Qwen2.5:7B	47.03	+3.79	+2.91	+1.45	+2.97	-2.97	89.14	86.23
	Latxa:70B	87.06	+0.50	-0.57	+1.26	+3.15	+5.11	90.34	<u>96.65</u>
	Llama3.1:70B	85.29	-1.96	-3.41	+4.36	+3.92	+4.30	89.90	94.13
	Qwen2.5:72B	78.16	+2.08	-1.14	+5.49	+0.63	-4.42	90.28	90.66
MGSM	Latxa:8B	48.00	-1.20	+3.60	-1.20	-10.00	+8.40	73.20	72.00
	Llama3.1:8B	21.20	+8.80	+9.20	+1.20	-3.20	+17.20	77.60	73.20
	Qwen2.5:7B	16.00	+3.20	+3.60	+0.40	0.00	+3.60	<u>80.00</u>	77.60
	Latxa:70B	73.20	-5.60	-3.60	-2.00	+4.40	+9.20	83.20	<u>93.20</u>
	Llama3.1:70B	62.40	-16.40	-18.40	+3.60	+6.00	+11.20	85.20	86.00
	Qwen2.5:72B	39.60	+14.80	+15.60	+12.00	+7.20	+14.80	82.40	87.20

Table 4: Accuracy using few-shot CoT and DUP prompting on MASEU and MGSM. For CoT, the Basque score is reported along with improvements or declines, with deltas, for templates that include English in the Mixed (T) setting. For DUP, the Basque score is reported together with deltas for Mixed (T), Mixed (T, R), and fully English configurations. Highest scores among small (7B–8B) and large (70B–72B) models for each Basque dataset are in **bold**. Best results in English are underlined.

despite officially supporting fewer languages, benefits from broader general-domain exposure.

In summary, structured prompting techniques provide consistent benefits for low-resource languages. Across both CoT and DUP, models with language-specific adaptation achieve higher accuracy in Basque, while larger generalist models benefit from scale and in-context learning. DUP further enhances performance through its modular reasoning pipeline and multilingual configurations. 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. Both CoT and DUP results show that general-purpose models such as Llama3.1:8B and Qwen2.5:7B benefit strongly from English inclusion, particularly in fully English prompts. On MASEU, Llama3.1:8B achieves a score of 90.97 under CoT and 88.32 under DUP, while Qwen2.5:7B reaches 89.14 and 86.23, respectively. Latxa:8B, which is already adapted to Basque, attains 90.02 under CoT and 79.04 under DUP. On MGSM, Llama3.1:8B scores 77.60 in CoT and 73.20 in DUP, Qwen2.5:7B scores 80.00 and 77.60, and Latxa:8B scores 73.20 and 72.00. These results highlight the strong dependency of generalist models on English, their dominant pretraining language, especially for reasoning-intensive tasks. Larger models, such as Llama3.1:70B and Qwen2.5:72B, show

consistent gains with English prompts, reaching 89.90 and 94.13 for Llama3.1:70B and 90.28 and 90.66 for Qwen2.5:72B on MASEU, and 85.20 and 86.00 for Llama3.1:70B and 82.40 and 87.20 for Qwen2.5:72B on MGSM, suggesting that dataset characteristics influence the benefit of English inclusion.

The Mixed (T) configuration, where only the template is in English, yields variable results across both prompting strategies. On MASEU, Llama3.1:8B gains +14.45 under CoT and +8.08 under DUP, Qwen2.5:7B gains +2.91 under CoT and +2.97 under DUP, and Latxa:70B exhibits minor changes with -0.57 under CoT and +3.15 under DUP. On MGSM, Llama3.1:8B gains +9.20 under CoT but experiences a decrease of -3.20 under DUP, Qwen2.5:7B gains +3.60 under CoT and remains unchanged under DUP, and Latxa:70B decreases by -3.60 under CoT but gains +4.40 under DUP. While mixed-language prompting may induce bilingual reasoning traces, our results do not reveal a consistent association between such patterns and correctness, suggesting that language alignment rather than code-switching per se may underlie the observed gains. This indicates that partial English prompting without accompanying English responses provides limited or inconsistent benefits, particularly for models adapted to Basque.

In DUP, the Mixed (T, R) setup, where both template and responses are in English, produces more consistent improvements. On MASEU, Llama3.1:8B gains +22.98, Latxa:70B gains +5.11, and Latxa:8B gains +2.90. On MGSM, Llama3.1:8B gains +17.20, Latxa:70B gains +9.20, and Latxa:8B gains +8.40. These results suggest that aligning both input and output languages can help generalist models compensate for limitations

in low-resource language reasoning, particularly in English-centric datasets like MGSM.

Finally, moving to fully English prompts generally produces the largest gains, except for Latxa:8B, which is already adapted to Basque. This pattern is observed in both CoT and DUP, reflecting the pretraining bias toward English and the predominance of English-based mathematical data. The effect is most pronounced on MGSM, where fully English prompts maximize performance, indicating that mixed-language interventions must consider both model pretraining and dataset characteristics to achieve optimal reasoning outcomes in low-resource languages.

RQ3: What is the impact continued pretraining in Basque?

In Basque, the best overall results are obtained under the DUP framework, with the much larger Llama3.1:70B reaching 89.65 and Latxa:70B following closely at 88.32, while under CoT their scores remain similarly close, 83.33 vs. 87.56. The strong performance of the smaller Latxa:8B model, surpassing Llama3.1:8B and Qwen2.5:7B across both prompting methods, 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 most cases across 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 may improve certain reasoning abilities, 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 few-shot strategies using both CoT and DUP prompting.

Few-shot prompting: 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 4, 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 limited presence in pretraining corpora and scarce availability of labeled data.

Lang.	Model	SP	CoT Δ	DUP Δ
Basque	Latxa:8B	63.40	+2.00	-0.16
	Llama3.1:8B	35.57	+11.62	+8.62
	Qwen2.5:7B	31.52	+3.37	+0.90
	Latxa:70B	80.13	-2.32	+3.52
	Llama3.1:70B	73.85	-10.05	+5.56
	Qwen2.5:72B	58.88	+7.84	+5.95
English	Latxa:8B	82.80	-1.19	-7.28
	Llama3.1:8B	80.86	+3.43	-0.10
	Qwen2.5:7B	84.00	+0.58	-2.08
	Latxa:70B	91.13	-4.36	+3.80
	Llama3.1:70B	90.86	-3.31	-0.79
	Qwen2.5:72B	74.67	+11.68	+14.27

Table 5: Model-wise comparison of SP, CoT, and DUP prompting on Basque and English, macro-averaged across the MASEU and MGSM datasets in the Few-Shot setting. The results highlight how different prompting strategies affect multilingual mathematical reasoning across model scales and languages.

In such cases, the inclusion of multiple examples offers crucial guidance that helps compensate for the model’s weaker baseline familiarity with the language.

Prompting strategies: We extend our analysis by comparing SP, CoT, and DUP in the Few-Shot setting. Table 5 reports results macro-averaged across both datasets, separated by language.

In Basque, CoT generally improves performance over SP, with gains varying across models. For example, Latxa:8B improves by +2.00 and Llama3.1:8B by +11.62, showing that structured reasoning guidance benefits models even in a low-resource language. DUP effects are more heterogeneous. Some models see further improvement, such as Qwen2.5:72B increasing by +5.95, while others, including Latxa:8B, experience slight declines of -0.16. This pattern suggests that deeper reasoning decomposition can enhance understanding for some architectures but may slightly challenge smaller or more specialized models.

In English, CoT produces modest improvements or small declines depending on the model, with Latxa:8B declining by -1.19 and Llama3.1:8B increasing by +3.43, indicating that structured reasoning guidance remains effective even in high-resource settings. DUP shows a similarly mixed picture, with larger models such as Latxa:70B and Qwen2.5:72B gaining +3.80 and +14.27 respectively, whereas smaller models, including Latxa:8B, decline by -7.28, suggesting that the addition of reasoning stages may impose challenges for models with limited capacity.

Overall, results confirm that continued pretraining in Basque enhances both CoT and DUP ef-

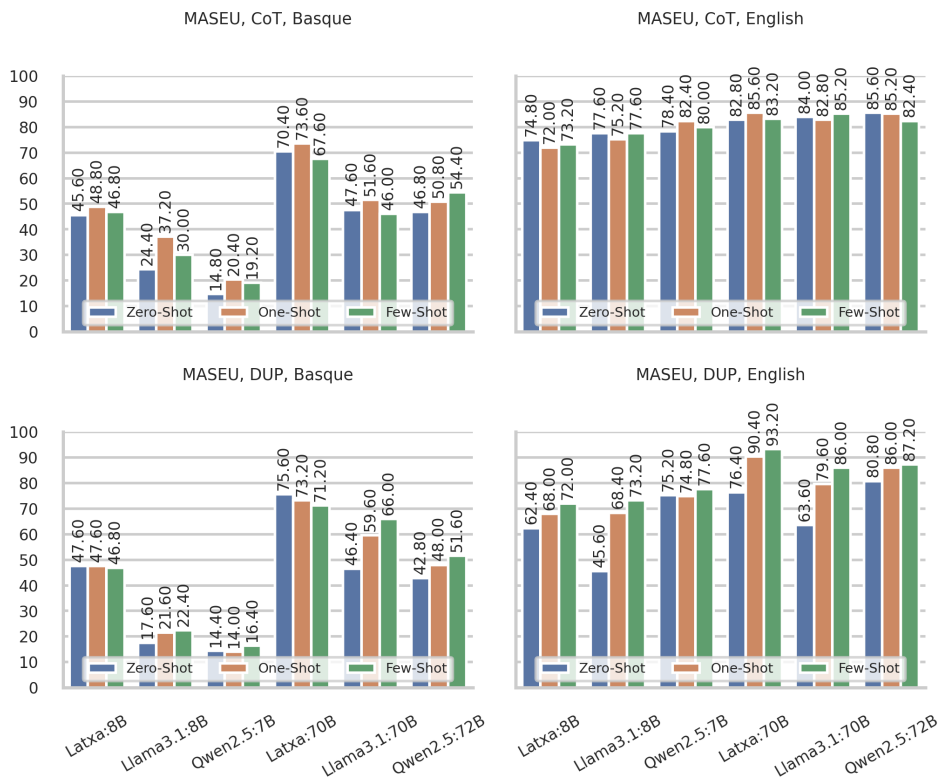


Figure 4: Comparative evaluation of *Zero-Shot*, *One-Shot*, and *Few-Shot* prompting techniques, in conjunction with the *CoT* and *DUP* prompting strategies, on the *MASEU* dataset. Results are reported for both Basque and English to assess multilingual reasoning performance.

fectiveness, and that explicit reasoning guidance is particularly advantageous in low-resource languages and for models with weaker baseline reasoning.

7. Conclusion

In this work we have introduced MASEU, a new benchmark designed to evaluate how large language models handle mathematical reasoning in Basque, a low-resource language. Using this resource along with the Basque and English portion of MGSM, we analyze 1) how well LLMs perform mathematical reasoning in Basque and 2) the impact of English in mixed-language prompts. We experiment with three model families (Ltxa, Llama, Qwen), two parameter classes (7-8 billion and 70-72 billion), and four prompting strategies: all Basque, Mixed(T), Mixed(T, R), and only English.

Results across both MASEU and MGSM reveal that all models, including those specifically trained for Basque, perform better in English than in Basque in this case study, underscoring the ongoing challenges in providing equitable multilingual support for low-resource languages.

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 show substantial improvements, suggesting that increased model capacity can help bridge the gap even without targeted language-specific training.

When solving Basque problems, most models perform better if the intermediate reasoning steps are in English, particularly in the Mixed(T, R) setting pairing Basque problem statements with English reasoning. This suggests that models may overfit to pretraining distributions, where mathematical reasoning data is mainly in English, so English-based reasoning likely aligns better with learned patterns, improving clarity and structure even in multilingual tasks. Moreover, both CoT and DUP prompting enhance reasoning performance, especially in Basque. CoT provides consistent moderate gains, while DUP delivers larger improvements for high-capacity models, indicating that explicit reasoning decomposition benefits from sufficient model depth and alignment.

Overall, our findings show that multilingual mathematical reasoning is highly sensitive to both language and model architecture in the Basque case. While generalist LLMs still struggle with Basque,

our results emphasize the value of targeted adaptation and thoughtful prompt design in narrowing the performance gap for underrepresented languages. Extending this analysis to typologically diverse low-resource languages remains an important direction for future work to assess generality of patterns.

Acknowledgments

This work is also supported by the Spanish Ministry of Science, Innovation and Universities (MCIN/AEI/10.13039/501100011033/) through DeepThought PID2024-159202OB-C21 and HumanAIZE AIA2025-163322-C61.

8. 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 underrepresentation, 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 two structured prompting approaches, CoT and DUP, inspired by staged reasoning methods. 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.

9. 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 Hub.³ These include both general-purpose multilingual models⁴ and models specifically adapted for Basque.⁵ 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.

10. Bibliographical References

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. 2021. [Program synthesis with large language models](#). *ArXiv*, abs/2108.07732.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez,

³<https://huggingface.co/>

⁴<https://huggingface.co/collections/meta-llama/llama-31-669fc079a0c406a149a5738f>, <https://huggingface.co/collections/Qwen/qwen25-66e81a666513e518adb90d9e>

⁵<https://huggingface.co/collections/HiTZ/latxa-instruct-682f356091452b0028380804>

- Anna Salles, Susana Sotelo Docio, Júlia Falção, Jose Javier Saiz, Robert 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 B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.
- 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](#). *ArXiv*, abs/2110.14168.
- 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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *ArXiv*, abs/2501.12948.
- Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. [Latxa: An open language model and evaluation suite for Basque](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. Abhishek Kadian. 2024. [The Llama 3 herd of models](#). *ArXiv*, abs/2407.21783.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). In *The Thirty-fifth Annual Conference on Neural Information Processing Systems*.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization](#). In *Proceedings of the 37th International Conference on Machine Learning*. JMLR.org.
- Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. [Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.
- Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. [Mindmerger: Efficiently boosting LLM reasoning in non-english languages](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Shima Imani, Liang Du, and Harsh Shrivastava. 2023. [MathPrompter: Mathematical reasoning using large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, Toronto, Canada. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.
- Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. [Parsing algebraic word prob-](#)

- lems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.
- Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. [MAWPS: A math word problem repository](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California. Association for Computational Linguistics.
- Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. [A diverse corpus for evaluating and developing English math word problem solvers](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, Online. Association for Computational Linguistics.
- Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2022. [Show your work: Scratchpads for intermediate computation with language models](#). *ArXiv*, abs/2112.00114.
- OpenAI. 2025. [Openai o3: Enhanced reasoning capabilities in large language models](#). Accessed on May 13, 2025.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. [Are NLP models really able to solve simple math word problems?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. [XCOPA: A multilingual dataset for causal commonsense reasoning](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.
- Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. [Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(1).
- Subhro Roy and Dan Roth. 2015. [Solving general arithmetic word problems](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.
- Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. [XTREME-R: Towards more challenging and nuanced multilingual evaluation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero, Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, Mikel Artetxe, and Aitor Soroa. 2025. [Instructing large language models for low-resource languages: A systematic study for basque](#). *ArXiv*, abs/2506.07597.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. [Multitask prompted training enables zero-shot task generalization](#). In *International Conference on Learning Representations*.

- Abulhair Saparov and He He. 2023. [Language models are greedy reasoners: A systematic formal analysis of chain-of-thought](#). In *The Eleventh International Conference on Learning Representations*.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. [Language models are multilingual chain-of-thought reasoners](#). In *The Eleventh International Conference on Learning Representations*.
- Kaya Stechly, Karthik Valmeekam, Atharva Gundawar, Vardhan Palod, and Subbarao Kambhampati. 2025. [Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens](#). *ArXiv*, abs/2505.13775.
- Qwen Team. 2024. [Qwq-32b-preview: Experimental text-based reasoning model](#). Accessed on May 13, 2025.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.
- Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. [Are multilingual models effective in code-switching?](#) In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153, Online. Association for Computational Linguistics.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. [Qwen2.5 technical report](#). *ArXiv*, abs/2412.15115.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. [A closer look at few-shot crosslingual transfer: The choice of shots matters](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics.
- Qihuang Zhong, Kang Wang, Ziyang Xu, Juhua Liu, Liang Ding, and Bo Du. 2024. [Achieving >97% on gsm8k: Deeply understanding the problems makes llms better solvers for math word problems](#). *Frontiers of Computer Science*, 20(1):2001310.
- Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Yongfeng Huang, Ruyi Gan, Jiaying Zhang, and Yujiu Yang. 2023. [Solving math word problems via cooperative reasoning induced language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4471–4485, Toronto, Canada. Association for Computational Linguistics.

Appendix 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 is to examine how varying levels of in-context supervision influence DUP’s structured reasoning stages in multilingual mathematical reasoning tasks, particularly in low-resource languages. By comparing zero-shot, one-shot, and few-shot prompts containing two examples, we aim to assess the impact of limited in-context supervision on model performance.

To support transparency and reproducibility, we provide concrete examples of the exact prompts used across configurations. Figures 5, 6, and 7 display the 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, with English counterparts in Figures 8, 9, and 10. These side-by-side examples illustrate how in-context examples are incorporated across stages and languages, highlighting structured prompting and enabling controlled comparisons of reasoning behaviors.

This approach has been systematically applied in both the primary experiments on mathematical reasoning in Basque and in complementary evaluations 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?

Figure 5: *Reveal the Core Question* stage with one example. 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.

Figure 6: *Extract the Problem-Solving Information* with one example. Template appears in bold, while context from *Reveal the Core Question* 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

Figure 7: *Generate and Extract the Answer* with one example. Template appears in bold, while context from *Reveal the Core Question* and *Extract the Problem-Solving Information* 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?

Figure 8: *Reveal the Core Question* stage with one example. 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.

Figure 9: *Extract the Problem-Solving Information* with one example. Template appears in bold, while context from *Reveal the Core Question* 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

Figure 10: *Generate and Extract the Answer* with one example. Template appears in bold, while context from *Reveal the Core Question* and *Extract the Problem-Solving Information* in italics.

The performance results for each model and shot configuration on the MGSM benchmark are visualized in Figure 11, 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.

Appendix B. Detailed Results on CoT, DUP, and Standard Prompting

This appendix provides the complete set of experimental results comparing CoT, DUP and SP

prompting strategies across all evaluated models, benchmarks, and language configurations in the few-shot setting. Figures 12 and 13 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 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.

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

Appendix C. Language Key Configurations in CoT and DUP Prompting

This appendix details the multilingual prompting configurations used to assess the impact of language mixing in CoT and DUP frameworks. We describe variants combining Basque and English across prompt components, including MWP input, template instructions, and model responses. In CoT, mixing is limited to the template, keeping reasoning in a single language, whereas DUP extends multilingual interaction to reasoning and response stages. These configurations isolate the effects of language use in structured reasoning, particularly when switching between a low-resource language such as Basque and a high-resource language like English.

The appendix also includes representative prompt examples for each configuration. The figures show how the three DUP stages, core question extraction, problem-solving information identification, and final answer generation, were implemented under different language conditions, clarifying how linguistic composition may influence reasoning performance and supporting reproducibility and further analysis of multilingual prompting.

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

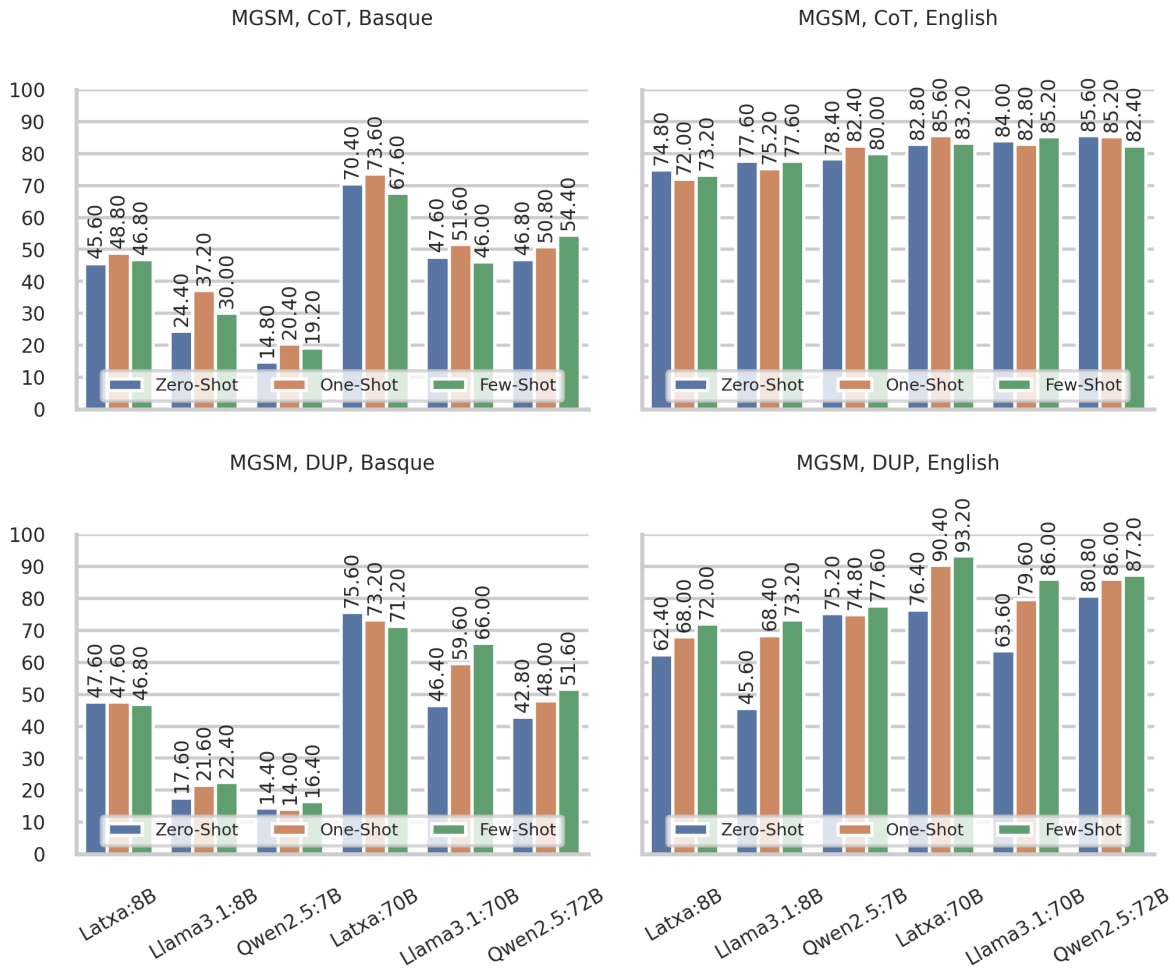


Figure 11: Comparative evaluation of Zero-Shot, One-Shot, and Few-Shot prompting techniques, in conjunction with the CoT and DUP prompting strategies, on the MGSM dataset. Results are reported for both Basque and English to assess multilingual reasoning performance.

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?
A: Urratsez urratseko arrazoiketa.

Figure 14: CoT prompt, where the MWP and the template, highlighted in bold, are all 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?
Mesedez, atera galdera nagusia, bakarrik osatuena eta zehatzena!
A: Zenbat ordu ematen ditu astean txakurrak zaintzen?

Figure 15: Reveal the Core Question stage, where the MWP, the template, highlighted in bold, and the response are all 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?
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 16: Extract the Problem-solving Information stage, where the MWP and template, highlighted in bold, the core question from the Reveal the Core Question stage, in italics, and the response are all in Basque.

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)
	CoT	Latxa:8B	77.69	83.57(+5.88)	84.41 (+6.72)
		Llama3.1:8B	63.22	70.83(+7.61)	73.10 (+9.88)
		Qwen2.5:7B	56.55	60.86(+4.31)	63.30 (+6.75)
		Latxa:70B	86.13	86.53(+1.60)	88.13 (+2.00)
		Llama3.1:70B	78.07	85.35 (+7.28)	85.04(+6.97)
		Qwen2.5:72B	78.07	81.99(+3.92)	82.51 (+4.44)
	SP	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)
	CoT	Latxa:8B	56.27	57.07(+0.80)	57.20 (+0.93)
		Llama3.1:8B	42.80	47.47 (+4.67)	46.00(+3.20)
		Qwen2.5:7B	36.27	40.80 (+4.53)	39.60+3.33
		Latxa:70B	75.60	78.00 (+2.40)	73.47(-2.13)
		Llama3.1:70B	60.27	62.27 (+2.00)	58.40(-1.87)
		Qwen2.5:72B	59.60	63.60(+4.00)	64.00 (+4.40)
	SP	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 CoT, DUP, and SP prompting strategies across various language settings.

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurak 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 txakurak zaintzen?
Mesedez, ulertu Aholkua eta galderaren informazioa, ondoren, galdera pausoz pauso ebatzi eta erantzuna erakutsi.
A: 35

Figure 17: *Generate and Extract the Answer* stage, where the MWP and 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 in Basque.

C.2. Mixed with English Templates

In the following prompts, the templates are written in English, while both the original problem state-

ment and the model's responses are in Basque. The model is explicitly instructed about the language of each input and output component. The *CoT* configuration is also evaluated under this mixed setting: although the template is in English, the expected output language is not specified, allowing the model to determine it from context.

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurak zaintzen?
A: Let's think step by step.

Figure 18: *CoT* prompt, where the MWP is in Basque, while the template, highlighted in bold, is in English.

Lang.	Shots	Model	SP	CoT	DUP	Lang.	Shots	Model	SP	CoT	DUP
Basque	Zero-Shot	Latxa:8B	73.67	72.10	77.27	Basque	Zero-Shot	Latxa:8B	44.40	45.60	47.60
		Llama3.1:8B	45.39	51.01	34.97			Llama3.1:8B	28.40	24.40	17.60
		Qwen2.5:7B	36.68	39.65	37.94			Qwen2.5:7B	11.60	14.80	14.40
		Latxa:70B	85.35	84.22	88.13			Latxa:70B	70.40	70.40	75.60
		Llama3.1:70B	80.18	72.85	77.65			Llama3.1:70B	64.00	47.60	46.40
	Qwen2.5:72B	67.80	73.23	74.49	Qwen2.5:72B		26.80	46.80	42.80		
	One-Shot	Latxa:8B	75.95	80.30	80.74		One-Shot	Latxa:8B	48.40	48.80	47.60
		Llama3.1:8B	41.10	60.23	47.54			Llama3.1:8B	19.20	37.20	21.60
		Qwen2.5:7B	42.61	48.61	46.15			Qwen2.5:7B	17.20	20.40	14.00
		Latxa:70B	84.85	84.72	86.36			Latxa:70B	75.60	73.60	73.20
		Llama3.1:70B	84.53	82.70	85.29			Llama3.1:70B	62.40	51.60	59.60
	Qwen2.5:72B	74.56	79.42	79.73	Qwen2.5:72B		35.20	50.80	48.00		
	Few-Shot	Latxa:8B	78.79	80.87	81.63		Few-Shot	Latxa:8B	48.00	46.80	46.80
		Llama3.1:8B	49.94	63.95	55.37			Llama3.1:8B	21.20	30.00	22.40
		Qwen2.5:7B	47.03	50.82	48.48			Qwen2.5:7B	16.00	19.20	16.40
Latxa:70B		87.06	87.56	88.32	Latxa:70B	73.20		67.60	71.20		
Llama3.1:70B		85.29	83.33	89.65	Llama3.1:70B	62.40		46.00	66.00		
Qwen2.5:72B	78.16	80.24	83.65	Qwen2.5:72B	39.60	54.40	51.60				
English	Zero-Shot	Latxa:8B	88.64	86.93	72.60	English	Zero-Shot	Latxa:8B	76.80	74.80	62.40
		Llama3.1:8B	88.38	88.26	74.12			Llama3.1:8B	72.80	77.60	45.60
		Qwen2.5:7B	86.68	87.12	78.66			Qwen2.5:7B	78.00	78.40	75.20
		Latxa:70B	93.18	89.20	82.77			Latxa:70B	93.60	82.80	76.40
		Llama3.1:70B	88.01	88.57	80.62			Llama3.1:70B	84.00	84.00	63.60
	Qwen2.5:72B	88.07	91.04	88.83	Qwen2.5:72B		53.20	85.60	80.80		
	One-Shot	Latxa:8B	88.76	89.02	79.42		One-Shot	Latxa:8B	74.80	72.00	68.00
		Llama3.1:8B	90.15	90.78	86.74			Llama3.1:8B	78.80	75.20	68.40
		Qwen2.5:7B	87.88	88.32	85.80			Qwen2.5:7B	78.80	82.40	74.80
		Latxa:70B	92.42	90.15	96.53			Latxa:70B	91.60	85.60	90.40
		Llama3.1:70B	90.59	90.59	96.09			Llama3.1:70B	85.60	82.80	79.60
	Qwen2.5:72B	89.77	91.10	90.21	Qwen2.5:72B		53.20	85.20	86.00		
	Few-Shot	Latxa:8B	89.20	90.03	79.04		Few-Shot	Latxa:8B	76.40	73.20	72.00
		Llama3.1:8B	88.51	90.97	88.32			Llama3.1:8B	73.20	77.60	73.20
		Qwen2.5:7B	88.38	89.14	86.24			Qwen2.5:7B	79.60	80.00	77.60
Latxa:70B		91.86	90.34	96.65	Latxa:70B	90.40		83.20	93.20		
Llama3.1:70B		92.11	89.90	94.13	Llama3.1:70B	89.60		85.20	86.00		
Qwen2.5:72B	90.53	90.28	90.66	Qwen2.5:72B	58.80	82.40	87.20				

Table 7: Comparison of SP, CoT and 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).

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

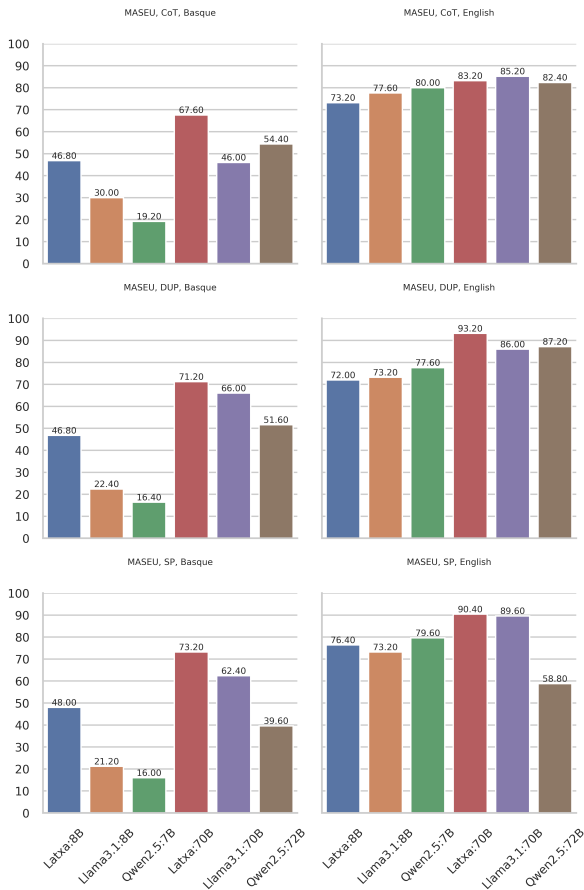


Figure 12: Performance comparison of CoT, DUP and SP prompting on the Basque subset of the *MASEU* dataset under Few-Shot settings. The figure illustrates how structured prompting strategies, particularly CoT and DUP, affect model performance across architectures and scales, highlighting their benefits in low-resource language contexts.

Q: Johnek 10 txakur zaintzen ditu. Egunean ordu-erdi behar du txakur bakoitza paseatzeko eta haren kontuez arduratzeko. Zenbat ordu ematen ditu astean txakurak 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 txakurak 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 21: *Generate and Extract the Answer* stage, where 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 all in Basque, while the template, highlighted in bold, is presented in English.

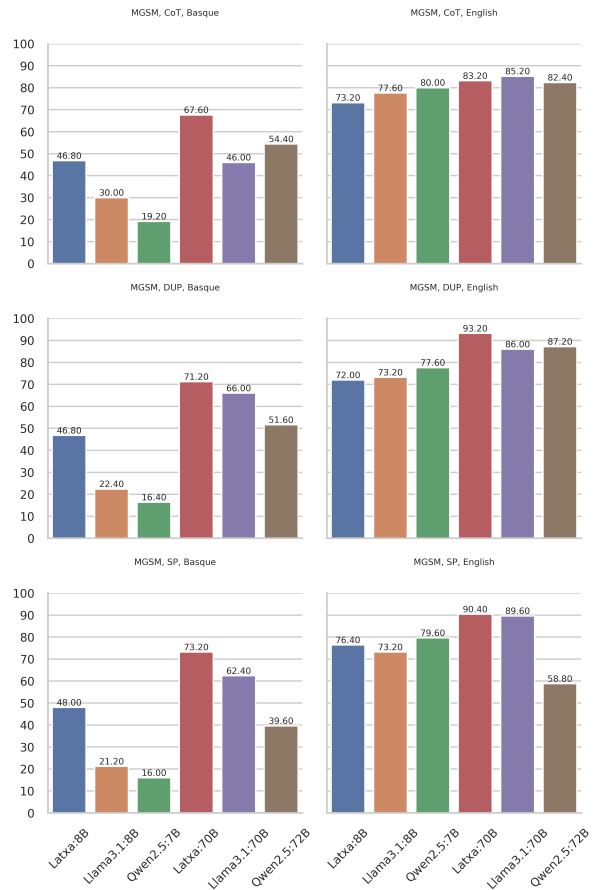


Figure 13: Performance comparison of CoT, DUP and SP prompting on the Basque subset of the *MGSM* dataset under Few-Shot settings. The figure illustrates how structured prompting strategies, particularly CoT and DUP, affect model performance across architectures and scales, highlighting their benefits in low-resource language contexts.

C.3. Mixed with English Templates and Responses

In the following prompts, English is used from templates to generated responses, while the original problem statement remains in Basque. The model is instructed on the language for each component, including input and 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 txakurak 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 22: *Reveal the Core Question* stage, where the MWP is in Basque, while the template, highlighted in bold, and the corresponding response are all 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 23: *Extract the Problem-solving Information* stage, where the MWP is 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 all 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 24: *Generate and Extract the Answer* stage, where the MWP is 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 all 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?
A: Let's think step by step.

Figure 25: *CoT* prompt, where the MWP and the template, highlighted in bold, are all 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?
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 26: *Reveal the Core Question* stage, where the MWP, the template, highlighted in bold, and the corresponding response are all 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 27: *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 28: *Generate and Extract the Answer* stage, where the MWP and 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 in English.

Appendix D. Prompts

Building on the prompts presented in Appendix C, this section provides the complete prompt-response examples for the MWPs. We first present the CoT prompts, consisting of a structured instruction for step-by-step reasoning, along with their outputs under different language configurations. We then include the full DUP sequences across its three stages, illustrating how each step builds on the previous one. These examples demonstrate the procedural implementation of CoT and DUP and show how language variation and iterative interactions affect the structure and resolution of MWPs across multilingual settings.

D.1. CoT

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?
A: Urratsez urratseko arrazoiketa.

A: Johnek 10 txakur zaintzen ditu. Bakoitzak 0.5 ordu behar ditu egunean. Hori $10 \times 0.5 = 5$ ordu da. Aste batean 7 egun daude. Beraz, astean $5 \times 7 = 35$ ordu behar ditu.

D.1.2. 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?

A: Let's think step by step.

A: Johnek 10 txakur zaintzen ditu. Bakoitzak 0.5 ordu behar ditu egunean. Hori $10 \times 0.5 = 5$ ordu da. Aste batean 7 egun daude. Beraz, astean $5 \times 7 = 35$ ordu behar ditu.

D.1.3. 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?

A: Let's think step by step.

A: To find out how many hours a week John spends taking care of dogs, we need to first calculate how many hours a day he spends taking care of dogs. Each dog takes 0.5 hours a day to walk and take care of their business. Since John takes care of 10 dogs, the total time spent per day is:
 $10 \text{ dogs} \times 0.5 \text{ hours/dog} = 5 \text{ hours/day}$ Since there are 7 days in a week, the total time spent per week is:
 $5 \text{ hours/day} \times 7 \text{ days/week} = 35 \text{ hours/week}$ So, John spends 35 hours a week taking care of dogs

D.2. Reveal the Core Question

D.2.1. 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 osatuena eta zehatzenena!

A: Zenbat ordu ematen ditu astean txakurrak zaintzen?

D.2.2. 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?

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?

D.2.3. 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?

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 John spend taking care of dogs?

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?

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?

D.3. Extract the Problem-solving Information

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?

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

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?

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

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?

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!

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.
3. There are 7 days in a week.

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?

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.4. Generate and Extract the Answer

D.4.1. 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.4.2. 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.4.3. 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.4.4. 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

Appendix 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 the CoT and DUP techniques, which apply structured reasoning prompts, with DUP further incorporating iterative prompt-response mixing and language variations across its stages, alongside the SP 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.

Model	Zero-Shot				One-Shot				Few-Shot			
	Basque	Mixed (T) Δ	Mixed (T, R) Δ	English Δ	Basque	Mixed (T) Δ	Mixed (T, R) Δ	English Δ	Basque	Mixed (T) Δ	Mixed (T, R) Δ	English Δ
SP Prompting												
Latxa:8B	73.67	-	-	+14.96	75.95	-	-	+12.82	78.79	-	-	+10.42
Llama3.1:8B	45.39	-	-	+42.99	41.10	-	-	+49.05	49.94	-	-	+38.57
Qwen2.5:7B	36.68	-	-	+50.00	42.61	-	-	+45.27	47.03	-	-	+41.35
Latxa:70B	85.35	-	-	+7.83	84.85	-	-	+7.58	87.06	-	-	+4.80
Llama3.1:70B	80.18	-	-	+7.83	84.53	-	-	+6.06	85.29	-	-	+6.82
Qwen2.5:72B	67.80	-	-	+20.27	74.56	-	-	+15.21	78.16	-	-	+12.37
CoT Prompting												
Latxa:8B	72.10	+1.96	-	+14.84	80.34	+1.07	-	+8.71	80.87	+1.45	-	+9.15
Llama3.1:8B	51.01	-0.63	-	+37.25	60.23	+1.26	-	+30.56	63.95	+0.44	-	+27.02
Qwen2.5:7B	39.65	+3.22	-	+47.47	48.61	-2.97	-	+39.71	50.82	-0.88	-	+38.32
Latxa:70B	84.22	+0.76	-	+4.99	84.72	0.00	-	+5.43	87.56	-1.07	-	+2.78
Llama3.1:70B	72.85	-0.06	-	+15.72	82.70	+0.06	-	+7.89	83.33	-1.45	-	+6.57
Qwen2.5:72B	72.23	-3.28	-	+91.04	79.42	-3.98	-	+11.68	80.24	-3.22	-	+10.04
DUP Prompting												
Latxa:8B	77.27	-3.66	-6.57	-4.67	80.74	-1.07	+0.25	-1.33	81.63	-6.69	+0.06	-2.59
Llama3.1:8B	34.97	-5.30	+24.94	+39.14	47.54	-2.97	+21.34	+39.20	55.37	+2.65	+17.55	+32.95
Qwen2.5:7B	37.94	+2.21	0.00	+40.72	46.15	+3.79	-6.31	+39.65	48.48	+1.52	-4.42	+37.75
Latxa:70B	88.13	+1.01	-1.45	-5.37	86.36	+3.28	+5.56	+10.16	88.32	+1.89	+3.85	+8.33
Llama3.1:70B	77.65	+1.96	+4.80	+2.97	85.29	+2.71	+4.10	+10.80	89.65	-0.44	-0.06	+4.48
Qwen2.5:72B	74.49	+2.08	-2.78	+14.33	79.73	+1.89	-6.06	+10.48	83.65	-4.86	-9.91	+7.01

Table 8: Results of performance metrics of SP, CoT and DUP prompting with Zero-Shot, One-Shot and Few-Shot for MASEU with different language configurations.

Model	Zero-Shot				One-Shot				Few-Shot			
	Basque	Mixed (T) Δ	Mixed (T, R) Δ	English Δ	Basque	Mixed (T) Δ	Mixed (T, R) Δ	English Δ	Basque	Mixed (T) Δ	Mixed (T, R) Δ	English Δ
SP Prompting												
Latxa:8B	44.40	-	-	+32.40	48.40	-	-	+26.40	48.00	-	-	+28.40
Llama3.1:8B	28.40	-	-	+44.40	19.20	-	-	+59.60	21.20	-	-	+52.00
Qwen2.5:7B	11.60	-	-	+66.40	17.20	-	-	+61.60	16.00	-	-	+63.60
Latxa:70B	70.40	-	-	+23.20	75.60	-	-	+16.00	73.20	-	-	+17.20
Llama3.1:70B	64.00	-	-	+20.00	62.40	-	-	+23.20	62.40	-	-	+27.20
Qwen2.5:72B	26.80	-	-	+26.40	35.20	-	-	+18.00	39.60	-	-	+19.20
CoT Prompting												
Latxa:8B	45.60	+2.80	-	+29.20	48.80	+1.60	-	+23.20	46.80	+4.80	-	+26.40
Llama3.1:8B	24.40	+2.00	-	+53.20	37.20	-7.20	-	+38.00	30.00	+0.40	-	+47.60
Qwen2.5:7B	14.80	+0.80	-	+63.60	20.40	-0.80	-	+62.00	19.20	+0.40	-	+60.80
Latxa:70B	70.40	+3.20	-	+12.40	73.60	+1.20	-	+12.00	67.60	+2.00	-	+15.60
Llama3.1:70B	47.60	+1.60	-	+36.40	51.60	+0.80	-	+31.20	46.00	-2.00	-	+39.20
Qwen2.5:72B	46.80	-0.40	-	+38.80	50.80	+4.00	-	+34.40	54.40	+0.80	-	+28.40
DUP Prompting												
Latxa:8B	47.60	-0.80	+7.60	+14.80	47.60	+0.80	+10.40	+20.40	46.80	-8.80	+9.60	+25.20
Llama3.1:8B	17.60	-6.80	+12.40	+28.00	21.60	-1.20	+14.40	+46.80	22.40	-4.40	+16.00	+50.80
Qwen2.5:7B	14.40	+0.40	+4.00	+60.80	14.00	+4.00	+4.80	+60.80	16.40	-0.40	+3.20	+61.20
Latxa:70B	75.60	-1.60	+0.80	+0.80	73.20	+2.00	+4.40	+17.20	71.20	+6.40	+11.20	+22.00
Llama3.1:70B	46.40	-6.40	+11.20	+17.20	59.60	+6.80	+9.60	+20.00	66.00	+2.40	+7.60	+20.00
Qwen2.5:72B	42.80	+6.40	+4.80	+38.00	48.00	+2.00	+2.00	+38.00	51.60	-4.80	+2.80	+35.60

Table 9: Results of performance metrics of SP, CoT and DUP prompting with Zero-Shot, One-Shot and Few-Shot for MGSM with different language configurations.