

# How Well Do Large Language Models Reason in Under-Resourced Languages? Evidence from Vietnamese

Tuan Anh Do<sup>♦</sup>, Jelke Bloem<sup>♦♦</sup>

- ♦ Artificial Intelligence, University of Amsterdam
- ♦ Institute for Logic, Language and Computation, University of Amsterdam
- ♦ Data Science Centre, University of Amsterdam  
talanhchanh@gmail.com, j.bloem@uva.nl

## Abstract

Despite advancements in Large Language Models, reasoning benchmarks remain centered on high-resource languages, leaving languages like Vietnamese under-evaluated. In this study, we aim to address this gap by evaluating four models: PhoGPT (native), Vistral and VBD-Llama (adapted), and Llama-2 (English-centric), on commonsense reasoning and arithmetic reasoning. As Vietnamese benchmarks for these tasks are lacking, we adapt two analogy datasets from English to Vietnamese and construct two sequence datasets, ensuring a range of structural complexity and difficulty levels. We evaluate diverse prompting strategies, including Chain-of-Thought, role-playing guidance, cross-lingual prompting, and few-shot learning. Our results reveal a baseline proficiency in analogical and arithmetic reasoning among the models, with Vistral and Llama-2 outperforming other models in multiple tasks. The effects of Chain-of-Thought and contextual guidance are limited in Vietnamese, while cross-lingual prompting and few-shot learning show promising performance improvements. The findings underscore the feasibility of adapting benchmarks to less-resourced languages and provide insights into strengths and weaknesses in the performance of Vietnamese LLMs, suggesting directions for model improvements.

**Keywords:** Vietnamese Large Language Model, Benchmarking, Reasoning

## 1. Introduction

Recent advancements in Large Language Models (LLMs) have demonstrated sophisticated reasoning capabilities across temporal, commonsense, and mathematical domains (Vaswani et al., 2017; Brown et al., 2020). Such capabilities, facilitated by techniques like Chain-of-Thought (CoT) prompting, have positioned LLMs as powerful tools for problem solving and knowledge discovery. However, the current landscape of LLM development is heavily skewed towards models trained on English and a few other widely spoken languages. This creates a significant gap for under-resourced languages, which not only lack extensive pre-training data but also exhibit issues like code switching and incorrect character usage. Consequently, it remains unclear how well these models generalize to languages with distinct linguistic characteristics and limited digital resources.

Current adaptation strategies often involve fine-tuning pre-trained models on smaller, language-specific datasets. However, these practices typically prioritize linguistic pattern adaptation over the enhancement of reasoning capabilities (Pipatanakul et al., 2025). Furthermore, the evaluation of LLMs in low-resource contexts is hindered by a lack of appropriate benchmarks for logical inference and analogical reasoning (Dargis et al., 2024; Azime et al., 2025; Truong et al., 2024). Such tasks require high-quality representations of abstract con-

cepts, which might be lacking in a foundation model or might not be easily transferred through subsequent tuning. The exploration of LLM reasoning capabilities in low-resource languages, therefore, is still in its infancy. Researching them is important not only for understanding cross-lingual generalization but also for improving local model performance and fostering more inclusive AI systems.

We investigate the reasoning capabilities of Vietnamese LLMs, focusing specifically on commonsense (analogical) reasoning and arithmetic sequence reasoning. Despite a growing ecosystem of local models and over 86 million native speakers (Eberhard et al., 2026), Vietnamese lacks a specialized framework for evaluating high-level reasoning. In addition, its isolating morphology and monosyllabic structure present a particular challenge: models must rely on word order rather than inflectional cues, making relational dependencies more challenging to capture. In this work, we provide one of the first benchmarks for Vietnamese-specific reasoning<sup>1</sup>, and use it to assess the effect of several factors in this setting: whether a model is pre-trained from scratch or fine-tuned on the target language, as well as the effects of Chain-of-Thought (CoT) reasoning, guided instruction, cross-lingual prompting and few-shot learning. With this, we aim to offer insights into how linguistic constraints influ-

<sup>1</sup>[https://github.com/tuanddox/VNM\\_LLMs\\_Eval](https://github.com/tuanddox/VNM_LLMs_Eval)

ence cross-lingual model generalization.

## 2. Related Work

### 2.1. Commonsense Reasoning in LLMs

Commonsense reasoning is a fundamental aspect of human intelligence, which incorporates different mechanisms of contextual understanding, implicit knowledge, and logical deduction in decision-making (Davis and Marcus, 2015; Yuan et al., 2024). A central subtask is analogical reasoning, which requires models to identify and map relational patterns between concept pairs. Historically used to evaluate semantic encoding (Mikolov et al., 2013), recent studies indicate that large-scale decoders like GPT-3/4 exhibit emergent zero-shot capabilities in word-string and matrix-based analogies, occasionally surpassing human baselines (Webb et al., 2023; Hu et al., 2023). However, this performance is inconsistent. Recent audits have revealed that LLMs still struggle with abstract, lengthy analogies (Ye et al., 2024) and tasks involving non-Latin scripts, specific linguistic nuances (Stevenson et al., in press; Li et al., 2025) or generalization to counterfactuals (Lewis and Mitchell, 2024).

There is increasing interest in the evaluation of commonsense reasoning in non-English languages, with efforts to bridge the multilingual gap focusing on two areas: dataset adaptation and specialized prompting. Some studies have proposed adapting English datasets into other languages, for example, through bidirectional translation of probing tasks (Lin et al., 2021), or to enhance reasoning with multilingual contrastive training (Lin et al., 2021) and model bridging (Yoon et al., 2024).

For the Vietnamese language and Vietnamese LLMs, there has been limited assessment of their reasoning capabilities, particularly in commonsense reasoning. Existing evaluations focus on general tasks such as closed-book question answering (Nguyen et al., 2024a), knowledge extraction and reading comprehension (Luu et al., 2023; Nguyen et al., 2022), and language understanding (Truong et al., 2024). Current efforts have prioritized domain-specific assessments, emphasizing the real-world applicability of pre-trained models in fields such as physics (Dao et al., 2023), biology and natural science (Dao and Le, 2023b), health issues (Nguyen et al., 2022), and legal issues (Anh et al., 2023), where the benefits and applications of LLMs are more apparent.

### 2.2. Mathematical Reasoning in LLMs

Mathematical reasoning, which spans word problems, geometry, and theorem proving, is a cornerstone for comparing human and machine cognitive capabilities (Imani et al., 2023). Within this

domain, arithmetic reasoning serves as a foundational task; it focuses on numerical computation and inductive logic while remaining relatively context-independent. Early investigations into transformer-based arithmetic highlighted that performance is heavily dictated by tokenization strategies, positional embeddings, and the frequency of specific numerical terms in pre-training corpora (Nogueira et al., 2021; Razeghi et al., 2022).

Because arithmetic reasoning minimizes linguistic complexity, it provides a controlled environment to isolate logical processing from language-specific effects. To improve performance in non-English contexts, research has moved toward multilingual datasets like MGSM (Shi et al., 2023) and MSVAMP (Chen et al., 2024). Furthermore, techniques such as cross-lingual prompting and alignment have been employed to “transfer” reasoning logic from high-resource foundation models to target languages (Qin et al., 2023).

Within the Vietnamese context, mathematical benchmarking has primarily targeted grade-school proficiency levels (Dao and Le, 2023a; Bui et al., 2025). While recent efforts have utilized both foundation and fine-tuned models in zero-shot settings (Truong et al., 2024), the scope remains largely pedagogical. Notably, Vinh et al. (2024) introduced the most extensive Vietnamese mathematical reasoning dataset to date, employing knowledge distillation to instruction-tune models on primary school problems. However, these benchmarks rely heavily on word problems, which inherently conflate linguistic comprehension with logical derivation. Our work diverges from this line of research by employing arithmetic sequences. This approach isolates the model’s inductive reasoning and pattern recognition capabilities from grade-school curricula and semantic biases, providing a more direct assessment of latent logical capacity. In an under-resourced language setting, it is particularly interesting to perform evaluations on tasks that are as language-independent as possible, such as arithmetic sequence completion, as models that generalize well should be able to show similar performance on such tasks regardless of the language in which the question is asked.

## 3. Datasets and Methodologies

### 3.1. Benchmarking Tasks

The evaluation framework in this study is designed to assess Vietnamese LLMs across two distinct cognitive dimensions: arithmetic sequence completion and analogical reasoning. We choose this benchmark scope to isolate the variables influencing model performance in under-resourced settings. By using arithmetic sequences, we evaluate

the model’s inductive logic in an environment that minimizes language factors, such as syntax or tone, yielding a more direct measure of abstract pattern recognition. This capacity for abstract pattern recognition is a prerequisite for generalization in domains like automated pedagogical tools and financial forecasting. Conversely, analogical reasoning serves as a probe into the intersection of semantic depth and abstract inference. Unlike pure arithmetic, these tasks require the model to navigate specific properties of the Vietnamese language while maintaining logical relational consistency. By combining these two tasks, the benchmark distinguishes between linguistic pattern matching and cross-domain logical generalization.

## 3.2. Dataset Adaptation and Linguistic Validation

### 3.2.1. Analogical Reasoning

Analogical reasoning is a cognitive process that involves identifying relational similarities between two sets of concepts or entities, exemplified by the proportional structure  $A:B::C:D$ , where the concept pair  $A:B$  is analogous to the concept pair  $C:D$ .

To evaluate LLM simulations of analogical reasoning, we adapted the BATS (Gladkova et al., 2016) and SCAN (Czinczoll et al., 2022) datasets, which provide diverse semantic and morphological relations, to construct complete analogy instances. A semi-automatic pipeline, consisting of an initial translation via the Google Translate API, followed by manual correction by native speakers, was employed.

The annotation focuses on preserving the relational structure of the analogies, for which we first identified the analogy type and syntactic category that appeared in the English source - target word pairs, following (Bejar et al., 1991). In our translations, we aimed to ensure that the analogies retain the same relational structure, while maintaining the semantic equivalence, lexical category, and relational integrity of the original pair. Furthermore, to account for the flexibility of Vietnamese compound words, we introduced synonymous target alternatives. For instance, given the source word “ghế bành” (armchair), the hypernym “sản phẩm nội thất” (article of furniture) was mapped to “đồ nội thất” (furniture) but also allowed for valid variants such as “đồ vật nội thất” (furniture item).

For the **BATS** dataset, we focused on the subset of lexicographic relations: hypernymy, meronymy, synonymy, and antonymy. Each category consists of 50 source-target pairs, from which we generated 2,450 unique analogy combinations. To ensure a balanced and manageable evaluation set, we randomly sampled 125 unique instances per category, resulting in a corpus of 500 BATS analogies.

The **SCAN** dataset includes 449 instances spanning the science and metaphor domains. During the annotation phase, we performed quality filtering to remove 23 instances involving morphological relations or near-identical mappings.

### 3.2.2. Arithmetic Reasoning

To evaluate arithmetic reasoning, we focus on the task of predicting the subsequent term(s) in an integer sequence. This task requires models to extract implicit mathematical rules through pattern recognition and inductive logic. To ensure a robust assessment, we utilize two distinct datasets: OEIS, and a synthetic dataset, RECUR.

For the **OEIS** dataset, we focused on sequences that were tagged as “core”, “nice”, and “easy”. These sequences are tractable and represent foundational arithmetic patterns. We provide the first 10 terms as input to standardize context, and ask the models to find the next terms.

To mitigate potential data contamination from well-known OEIS patterns, we generated synthetic sequences using random recurrence relations, following the method proposed by d’Ascoli et al. (2022). The generated sequences were manually classified into three difficulty tiers (Easy, Intermediate, and Difficult) based on the complexity of their recurrence trees (with recurrence degree ranging from 0 to 4) and the depth of the operations required. Details on the generation procedure and the difficulty tiers can be found in Appendix A. To ensure evaluation consistency, we sampled 140 annotated sequences per category, totaling 420 RECUR instances.

## 3.3. Prompt Construction

For each experiment, we vary the instruction prompt with a modular combination of “openers”, “task description”, and “closer” snippets in the following format:

```
<<SYS>>{opener}{task_description}
{closer}<</SYS>>
```

Drawing on the methodology of Battle and Gollapudi (2024), the “opener” serves as an auxiliary role-setting snippet to establish task context and interaction expectations without modifying the underlying system prompts. The “closer” governs the output format, instructing the model to either provide a direct final answer or generate a full response. Table 1 details the specific role-setting openers and formatting closers used to standardize the interaction across different tasks and reasoning scenarios. Our experimental design incorporates three opener configurations (two specialized personas and one null control), eight task descriptions (one per primary dataset and four for BATS

sub-categories), and two closer configurations. To facilitate in-context learning, we manually crafted step-by-step explanations for one-shot CoT reasoning. All prompts were presented in Vietnamese.

Opener Snippets (Role-Setting)
Bạn là <b>nhà ngôn ngữ học</b> có chuyên môn về <b>quan hệ ngữ nghĩa</b> <sup>†</sup> (You are a linguist with expertise in commonsense reasoning)
Bạn là <b>nhà toán học</b> có chuyên môn về <b>các phép truy toán và suy luận toán học</b> <sup>‡</sup> (You are a mathematician with expertise in recurrence theory)

Closer Snippets (Format & Reasoning Guidance)
Câu trả lời chỉ chứa đáp án cuối cùng theo cấu trúc "Đáp án là ..." <sup>*</sup> (The response should only contain the final answer: "The answer is...")
Suy luận rồi đưa ra đáp án cuối cùng theo cấu trúc "Đáp án là ..." <sup>°</sup> (Reason and give the answer: "The answer is...")

<sup>†</sup> BATS/SCAN    <sup>‡</sup> OEIS/RECUR

<sup>\*</sup> Short    <sup>°</sup> Full

Table 1: Prompt snippets for role-setting and formatting.

### 3.4. Models and Experimental Setup

We evaluate four representative open-weight models to capture the diversity of current Vietnamese LLM development: PhoGPT-4B<sup>2</sup>, Vistral-7B<sup>3</sup>, VBD-Llama-7B<sup>4</sup>, and Llama-2-7B<sup>5</sup> (baseline). Our selection covers two contrasting training methodologies: PhoGPT (Nguyen et al., 2024b) represents the first model pre-trained primarily from scratch on Vietnamese corpora, whereas Vistral and VBD-Llama are variants fine-tuned from English-centric foundations (Mistral and Llama, respectively). This selection allows us to investigate whether native pre-training or high-quality cross-lingual transfer is more effective for complex reasoning tasks.

In selection, models were prioritized based on their performance on the VMLU leaderboard<sup>6</sup>, their availability, and expected cost of running. VMLU is a framework to assess general knowledge and problem solving ability of foundation models in the Vietnamese language, covering questions in the domains of STEM, Social Science, and Humanities. To ensure a standardized "black-box" assess-

<sup>2</sup><https://github.com/VinAIRResearch/PhoGPT>

<sup>3</sup><https://huggingface.co/Viet-Mistral/Vistral-7B-Chat>

<sup>4</sup><https://huggingface.co/LR-AI-Labs/vbd-llama2-7b-50b-chat>

<sup>5</sup>[huggingface.co/meta-llama/Llama-2-7b](https://huggingface.co/meta-llama/Llama-2-7b)

<sup>6</sup><https://vmlu.ai/>

ment of inherent reasoning, all evaluations were conducted in a zero-shot setting using the models' default chat templates without further fine-tuning or prompt optimization. The models are loaded from the Hugging Face repository, and the prompts are pre-processed according to the tasks being performed.

Inference was conducted with greedy decoding and without batching to ensure isolated metrics. Experiments were carried out on clusters with 4x NVIDIA A100-40GB GPUs and clusters with 1 NVIDIA A100-80GB GPU. We limited the maximum output to 512 tokens for CoT-enabled tasks and 256 tokens for non-CoT tasks. For commonsense reasoning, we applied a repetition penalty of 1.2 and no-repeat-ngram-size of 3 to reduce redundancy; for arithmetic tasks, the penalty was set to 1.0 to accommodate naturally recurring numbers.

### 3.5. Evaluation

Due to the inherent variability in Vietnamese translation and inconsistent model adherence to output formats, automated string matching is insufficient for benchmarking commonsense reasoning. We therefore utilize human-in-the-loop scoring via a majority voting schema, employing four native-speaker annotators to determine an equivalence score. This metric shifts the focus from lexical identity to functional synonymy, ensuring models are not penalized for valid linguistic variations, such as the interplay between Sino-Vietnamese compounds and colloquialisms. A response is deemed "equivalently correct" if it satisfies two conditions: contextual relevance and conceptual similarity (see more in Appendix B). Examples of model responses with different annotation decisions can be found in Appendix C. Relevance serves as a primary filter for instructional integrity and domain-pertinence, permitting minor code-switching (under 50%) or extraneous phrasing. Conceptual similarity is then assessed via a set of criteria: (1) semantic overlap (synonymy, hypernymy, and lexical permutation), (2) domain alignment (correct abstract fields), and (3) logical entailment (where a reasoning path entails the true analogical relationship).

## 4. Results and Analysis

### 4.1. Analogical Reasoning

Figure 1 illustrates the performance of the four models across all instruction variations. Generally, accuracy remains capped at a 30% threshold, with models performing notably better on antonymy than other lexical categories in BATS, and favoring scientific over metaphorical analogies in SCAN. Despite being an English-centric model, Llama-2 maintains the highest baseline accuracy across

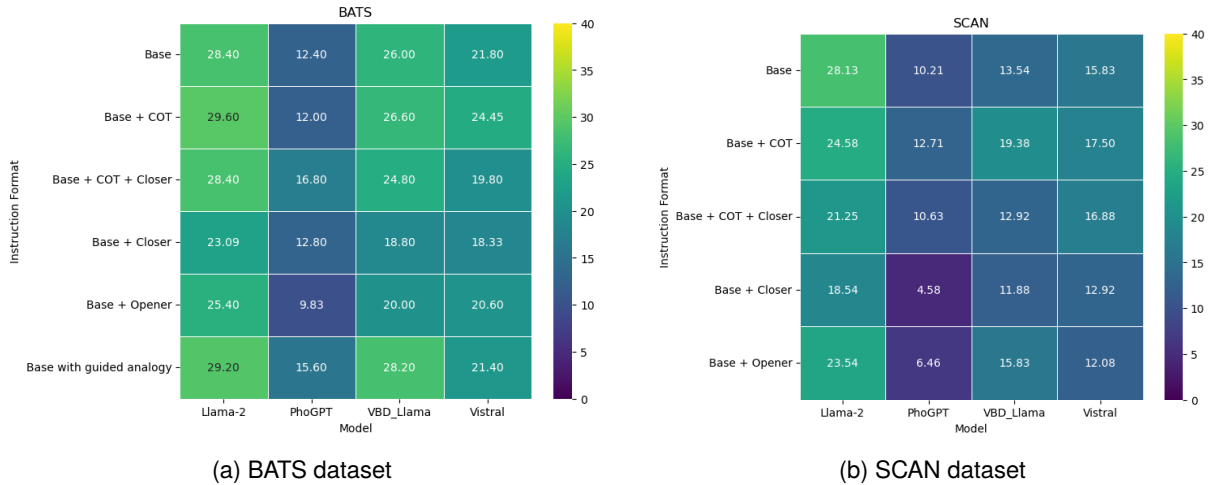


Figure 1: Comparing baseline instructions with enhanced instructions on analogy reasoning. Results are averaged over all categories.

both datasets. However, it is also the most sensitive to prompt noise, where the introduction of CoT or format-constricting “closers” results in performance degradation. The fine-tuned VBD\_Llama and Vistral perform comparably to each other, even surpassing the foundation model in specific scenarios such as meronymy (BATS) and science analogies with CoT (SCAN). PhoGPT, while trained entirely from scratch on mostly Vietnamese data, underperforms in all instruction formats. This could be attributed to its training corpus, which predominantly consists of dialogue and conversation, thus lacking the breadth and depth of input required for reasoning tasks.

**The effects of Chain-of-Thought:** The inclusion of CoT generally improves model performance in both datasets, with notable gains for VBD\_Llama and Vistral. Conversely, PhoGPT failed to leverage CoT entirely, and manual inspection revealed a systemic inability to follow the reasoning prompt. For both cases, Llama-2 already achieves high baseline performance without CoT, and the introduction of CoT might lead to a slight performance downgrade, e.g. in synonymy and meronymy analogies of the BATS dataset, and with both science and metaphor analogies in the SCAN dataset, with a decrease of up to 15%. When manually examining outputs of the model, we observe that CoT does not elicit the models’ underlying logic; instead, it induces a self-affirmation bias in justifying its initially incorrect responses.

**The effects of guided instruction:** Providing explicit taxonomical hints (e.g., “is an antonym of”) led to consistent accuracy gains across all models. Antonymous and synonymous analogies benefit the most from this approach, with PhoGPT dou-

bling its precision in synonymy and showing a 35% increase in antonymy. However, the improvement in hypernymy and meronymy is less pronounced, with VBD\_Llama showing signs of saturation. This can be explained by the possibility that explicit hints trigger lexical retrieval strategies based on word co-occurrence in embedding space. Furthermore, we observed a “format-confusion” effect: adding system-role openers or strict format closers (e.g., to only return the final answer) generally degraded performance. This suggests that the investigated models are sensitive to prompt length and may become distracted by additional constraints.

**Few-shot learning:** Few-shot learning (1, 2, and 4-shot) provided gains for all models, but led to hallucination in PhoGPT, which tended to repeat or explain the provided exemplars. For both datasets, the models generally perform better with one-shot learning, especially for Vistral and VBD\_Llama, and one-shot CoT might lead to worse performance than standard one-shot prompting. This suggests that the combined complexity of an exemplar and a reasoning chain exceeds the models’ current capacity to maintain task focus, leading to a breakdown in coherence.

## 4.2. Arithmetic Reasoning

Vistral outperforms other models in the sequence completion tasks, achieving an average accuracy of 36% on OEIS sequences and peaking at approximately 77% on synthetic easy sequences, as shown in Figure 2. Llama-based models struggle with sequences of repetition patterns and tend to over-complicate the task by generating lengthy binary operations.

		Base	Base + CoT	Base + Opener	Base + Closer
<b>Llama</b>					
OEIS		27.1%	25.7%	23.6%	27.9%
RECUR	easy	31.4%	19.3%	0.7%	41.4%
	medium	1.4%	9.3%	15.7%	20.0%
	difficult	0.0%	2.9%	2.1%	2.9%
<b>VBD_Llama</b>					
OEIS		30.7%	33.6%	<b>35.7%</b>	23.6%
RECUR	easy	30.0%	32.1%	28.6%	0.0%
	medium	11.4%	17.1%	14.3%	0.0%
	difficult	2.9%	2.1%	2.9%	0.0%
<b>Vistral</b>					
OEIS		<b>47.1%</b>	<b>34.3%</b>	31.4%	<b>30.0%</b>
RECUR	easy	<b>58.6%</b>	<b>72.9%</b>	<b>76.4%</b>	<b>76.4%</b>
	medium	<b>64.3%</b>	<b>57.1%</b>	<b>50.7%</b>	<b>56.4%</b>
	difficult	<b>15.7%</b>	<b>14.3%</b>	<b>17.1%</b>	<b>6.4%</b>

Figure 2: Performance of models on mathematical abstract reasoning tasks. The cells highlighted in light gray indicate the best prompting method in each subset of sequences, while the bold text in the cells indicates the best model for the same question and instruction format. The subsets of sequences are separated by different colors for clarity.

**Recurrence degree:** The degree of recurrence defines the number of previous terms required to generate the next value in the sequence. Generally, all models exhibit peak accuracy at degree 0, especially on the easy subset, demonstrating a strong baseline for non-recursive patterns. Vistral shows the most significant fluctuations, with its performance peaking at degrees 2 and 4 for medium sequences and degrees 1 and 3 for the hard subset. In addition, the effectiveness of prompt formats is highly sensitive to the degree of recurrence; while structural aids like CoT and closers improve Vistral’s accuracy on easy sequences, the base prompting is most effective for higher-order tasks. In contrast, Llama-2 and its fine-tune counterpart, VBD\_Llama, remain relatively stable across different degrees but exhibit a clear reasoning plateau. Both models collapse on medium and difficult sequences, with accuracies consistently below 8% across prompts, suggesting a lack of robust representations and reasoning capabilities for complex recursive reasoning in the foundation model.

**Multiple predictions:** To distinguish between structural understanding and surface-level memorization, we evaluated model performance across extended prediction horizons (1 to 4 terms). Overall, accuracy declines with the number of predictions, as shown in Figure 3. Vistral emerged as the most consistent reasoner, maintaining high Jaccard similarity scores for correct predictions across varying lengths (see more in Appendix D). In addition, it

demonstrates a unique contextual advantage: its accuracy on the RECUR dataset improves when asked to predict the next two terms compared to one, implying that a broader sequence requirement may trigger more robust pattern recognition or internal error-correction mechanisms.

**Language alignment in prompting:** To evaluate interlingual generalization, we presented Vietnamese questions with English-instructed response requirements. This was motivated by previous studies on cross-lingual and adaptive multilingual prompting on math word problems (Qin et al., 2023; Huang et al., 2023), and from our own observations that Vietnamese-adapted models like VBD\_Llama likely leverage latent English-centric tokens from their foundation to bolster reasoning, suggesting that English instructions may “unlock” higher-resource knowledge bases inherited during pre-training.

While the core mathematical problems remained in Vietnamese, models were instructed, either in English or Vietnamese, to generate outputs in English. Vistral demonstrated significant instructional rigidity, failing to adhere to the language constraint by responding exclusively in Vietnamese. Conversely, Llama-2 and VBD\_Llama exhibited marked accuracy improvements when generating responses in English. Comparing VN-VN (Vietnamese instruction/Vietnamese response) to VN-ENG (Vietnamese instruction/English response), we found that Llama-2 generally performs better

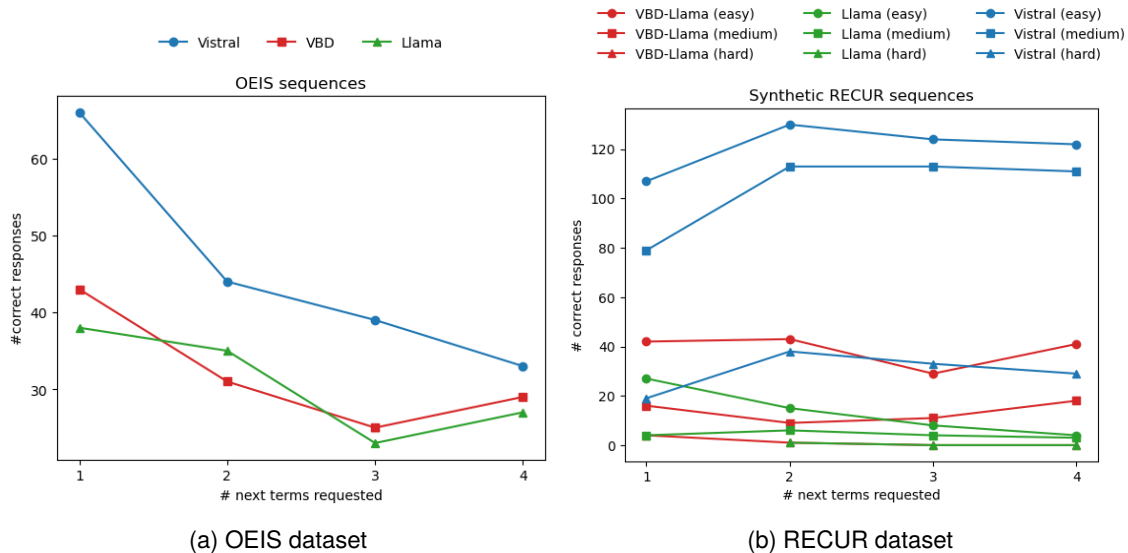


Figure 3: Performance of models in generating multiple predictions across difficulty levels

when the response is in English, likely due to its extensive English pre-training. Both Llama-based models achieve higher accuracy when the instruction language aligns with the response language. This is particularly evident for VBD\_Llama, which collapsed with cross-lingual prompting, suggesting a limited capacity for processing code-switched instructions.

**The effects of Chain-of-Thought:** Similar to the commonsense reasoning tasks, the introduction of CoT does not consistently enhance performance in arithmetic reasoning. As illustrated in Figure 4, both Llama-2 and VBD\_Llama exhibit a slight decline in accuracy when CoT is applied in zero-shot and one-shot settings, respectively. While Vistral maintains its performance lead over the other models, it demonstrates no tangible benefit from the additional reasoning context provided by either few-shot exemplars or CoT prompts. In addition, the application of CoT in cross-lingual settings yields mixed outcomes. Only Llama-2 benefited from being instructed with English CoT in the subset of easy and medium synthetic sequences, whereas other models remain stagnant or degrade.

## 5. Discussion

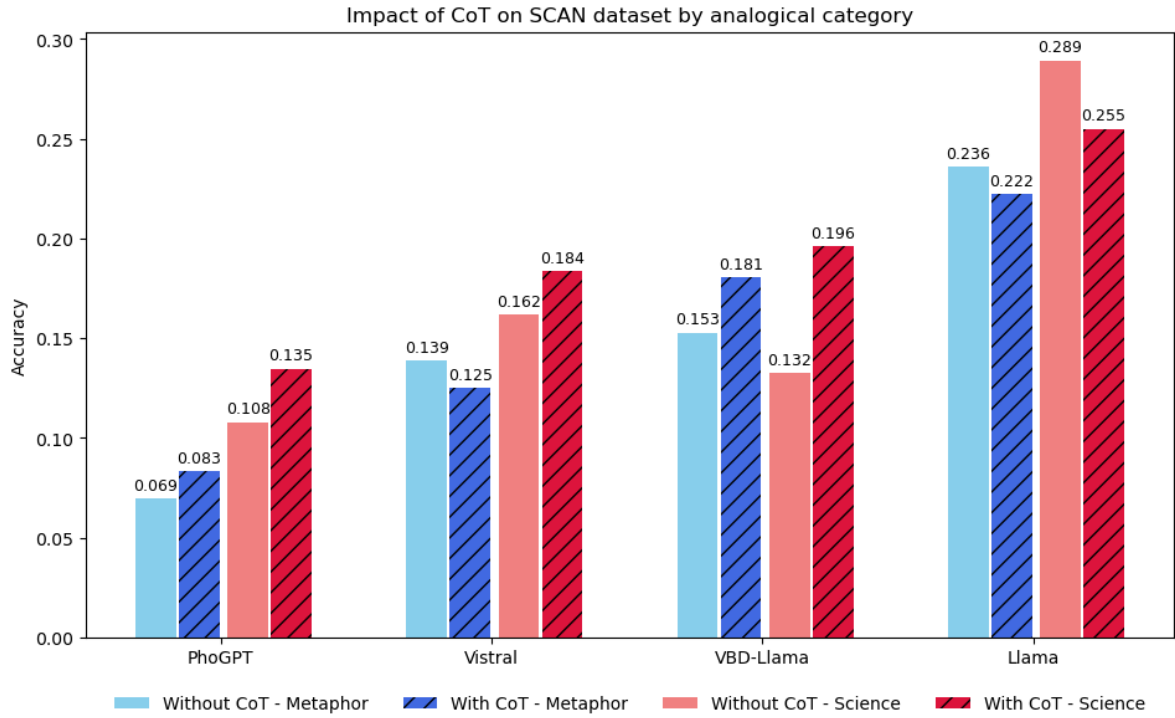
In this study, we demonstrated a methodology for evaluating the reasoning capabilities of under-resourced language LLMs. Our research serves as a preliminary investigation into the reasoning abilities of Vietnamese LLMs, particularly in analogical reasoning and arithmetic sequence generation, and can be extended to benchmark models in other under-resourced languages. Our methodology involved adapting datasets specifically designed for these types of reasoning, yet to be ap-

plied to LLM benchmarking, including two common sense reasoning datasets enriched with scientific, metaphorical, and semantic analogies, and a two-sequence dataset, one synthetic and one from the OEIS database.

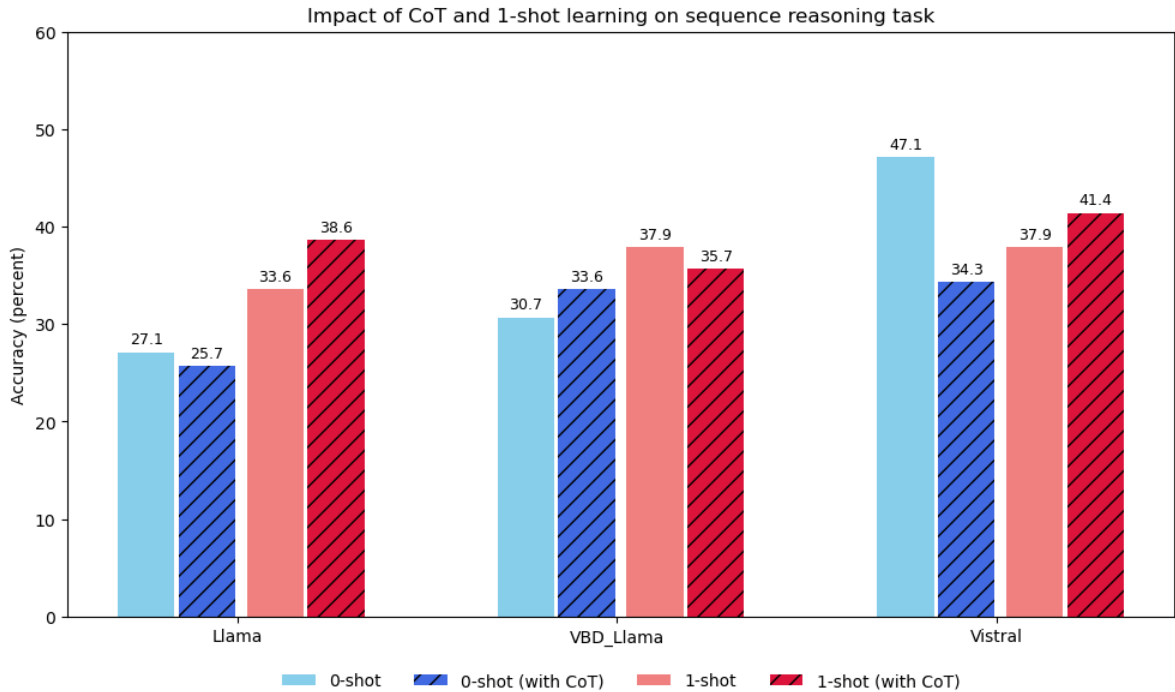
The work highlights fundamental challenges in evaluating the reasoning capabilities of LLMs in under-resourced languages. While the evaluated models demonstrate limited success on concrete analogical and arithmetic tasks, the results show that reasoning performance is fragile, prompt-dependent, and poorly aligned with evaluation strategies derived from high-resource languages.

We also find that prompting techniques widely used in high-resource languages, specifically Chain-of-Thought, do not reliably transfer to Vietnamese. Across both commonsense and arithmetic tasks, CoT frequently introduced irrelevant noise and failed to yield consistent accuracy gains. This echoes observations in previous studies on Vietnamese LLMs (Truong et al., 2024) and models in other under-resourced contexts, such as Bengali (Nazi et al., 2025) and Indonesian (Huang et al., 2025), suggesting that the efficacy of explicit reasoning traces may depend on a level of linguistic grounding and training density currently absent in low-resource settings. In contrast, few-shot learning proved more effective than zero-shot CoT, suggesting that in-context pattern induction may be a more robust mechanism than explicit reasoning articulation. These findings align with prior work on multilingual LLM performance (e.g., Xuan et al. (2025)), although performance still lags behind models instructed in English on the same tasks (O'Malley et al., 2024).

Furthermore, we demonstrate that reasoning capabilities are inextricably linked to evaluation de-



(a) Commonsense Reasoning (SCAN dataset)



(b) Arithmetic Reasoning (OEIS dataset)

Figure 4: Effects of Chain-of-Thought and few-shot learning on reasoning tasks

sign. This is most evident in commonsense reasoning, where models succeed on concrete scientific relationships but falter on abstract or metaphorical analogies. The performance gains observed from taxonomically guided prompts suggest that models may be prioritizing pattern alignment over relational

representation. Similarly, in arithmetic tasks, the models' sensitivity to task formulation and failure to capture simple repetitive patterns imply a reliance on sequence memorization rather than rule induction, a phenomenon also observed by Gambardella et al. (2024) and d'Ascoli et al. (2022).

Lastly, the cross-lingual experiments reveal a tension between multilingual pre-training and monolingual fine-tuning. While English instructions improved performance for foundation models by presumably enabling connections to higher-resource knowledge, Vietnamese-adapted variants suffered when prompted in a foreign language. This is consistent with catastrophic forgetting, suggesting that model localization through intensive fine-tuning may come at the cost of cross-lingual generalization and interlingual reasoning flexibility.

Future work should prioritize the development of reasoning datasets that are culturally and linguistically grounded, rather than relying solely on translated resources, and on designing diagnostic tasks that clearly separate memorization from genuine reasoning. Furthermore, our findings also show that fine-tuning can harm reasoning ability, highlighting the need for alignment methods that adapt models to the target Vietnamese language without weakening the reasoning skills learned during multilingual pre-training. By addressing these gaps, we can move toward models that do not merely simulate reasoning through pattern matching, but demonstrate robust, structural generalization across diverse linguistic contexts.

## 6. References

- Dang Hoang Anh, Dinh-Truong Do, Vu Tran, and Nguyen Le Minh. 2023. [The impact of large language modeling on natural language processing in legal texts: A comprehensive survey](#). In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–7.
- Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadgign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Tamar Solorio, and Dietrich Klakow. 2025. [ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6265–6281, Albuquerque, New Mexico. Association for Computational Linguistics.
- Rick Battle and Teja Gollapudi. 2024. [The unreasonable effectiveness of eccentric automatic prompts](#).
- Isaac I. Bejar, Roger Chaffin, and Susan Embretson. 1991. [A taxonomy of semantic relations](#). In *Cognitive and Psychometric Analysis of Analogical Problem Solving*, pages 55–91. Springer US, New York, NY.
- 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 *Advances in neural information processing systems*, volume 33, pages 1877–1901.
- Cuc Thi Bui, Nguyen Truong Son, Truong Van Trang, Lam Viet Phung, Pham Nhut Huy, Hoang Anh Le, Quoc Huu Van, Phong Nguyen-Thuan Do, Van Le Tran Truc, Duc Thanh Chau, and Le-Minh Nguyen. 2025. [VMLU benchmarks: A comprehensive benchmark toolkit for Vietnamese LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11495–11515, Vienna, Austria. Association for Computational Linguistics.
- Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. [Breaking language barriers in multilingual mathematical reasoning: Insights and observations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.
- Tamara Czinczoll, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2022. [Scientific and creative analogies in pretrained language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2094–2100, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Xuan-Quy Dao and Ngoc-Bich Le. 2023a. [Investigating the effectiveness of ChatGPT in mathematical reasoning and problem solving: Evidence from the Vietnamese national high school graduation examination](#).
- Xuan-Quy Dao and Ngoc-Bich Le. 2023b. [LLMs performance on Vietnamese high school biology examination](#). *International Journal of Modern Education and Computer Science*, 15(6):14–30.
- Xuan-Quy Dao, Ngoc-Bich Le, Xuan-Dung Phan, Bac-Bien Ngo, and The-Duy Vo. 2023. [Evaluation of ChatGPT and Microsoft Bing AI Chat performances on physics exams of Vietnamese national high school graduation examination](#).

- Roberts Dargis, Guntis Bārzdīņš, Inguna Skadiņa, and Baiba Saulīte. 2024. [Evaluating open-source LLMs in low-resource languages: Insights from Latvian high school exams](#). In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 289–293, Miami, USA. Association for Computational Linguistics.
- Stéphane d’Ascoli, Pierre-Alexandre Kamienny, Guillaume Lample, and François Charton. 2022. [Deep symbolic regression for recurrent sequences](#).
- Ernest Davis and Gary Marcus. 2015. [Commonsense reasoning and commonsense knowledge in artificial intelligence](#). *Commun. ACM*, 58(9):92–103.
- David M Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2026. *Ethnologue: Languages of the World*, 29 edition. SIL International, Dallas, TX, USA.
- Andrew Gambardella, Yusuke Iwasawa, and Yutaka Matsuo. 2024. [Language models do hard arithmetic tasks easily and hardly do easy arithmetic tasks](#).
- Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuo. 2016. [Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn’t](#). In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California. Association for Computational Linguistics.
- Xiaoyang Hu, Shane Storks, Richard Lewis, and Joyce Chai. 2023. [In-context analogical reasoning with pre-trained language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1953–1969, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Xin Huang, Tarun Kumar Vangani, Zhengyuan Liu, BOWEI ZOU, and Ai Ti Aw. 2025. [AdaCoT: Rethinking cross-lingual factual reasoning through adaptive chain-of-thought](#).
- 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.
- Martha Lewis and Melanie Mitchell. 2024. [Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models](#). In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.
- Xiaoyuan Li, Moxin Li, Rui Men, Yichang Zhang, Keqin Bao, Wenjie Wang, Fuli Feng, Dayiheng Liu, and Junyang Lin. 2025. [HellaSwag-pro: A large-scale bilingual benchmark for evaluating the robustness of LLMs in commonsense reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9038–9072, Vienna, Austria. Association for Computational Linguistics.
- Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. [Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning](#). 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 1274–1287, Online. Association for Computational Linguistics.
- Son T Luu, Khoi Trong Hoang, Tuong Quang Pham, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023. [A multiple choices reading comprehension corpus for vietnamese language education](#).
- Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). In *1st International Conference on Learning Representations, ICLR, Workshop Track Proceedings*.
- Zabir Al Nazi, Md. Rajib Hossain, and Faisal Al Mamun. 2025. [Evaluation of open and closed-source LLMs for low-resource language with zero-shot, few-shot, and chain-of-thought prompting](#). *Natural Language Processing Journal*, 10:100124.
- Dat Quoc Nguyen, Linh The Nguyen, Chi Tran, Dung Ngoc Nguyen, Dinh Phung, and Hung Bui. 2024a. [PhoGPT: Generative pre-training for Vietnamese](#).
- Dat Quoc Nguyen, Linh The Nguyen, Chi Tran, Dung Ngoc Nguyen, Dinh Phung, and Hung Bui. 2024b. [PhoGPT: Generative pre-training for Vietnamese](#).

- Kiet Van Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2022. [New Vietnamese corpus for machine reading comprehension of health news articles](#). *Transactions on Asian and Low-Resource Language Information Processing*, 21(5):1–28.
- Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2021. [Investigating the limitations of transformers with simple arithmetic tasks](#).
- Daniel O’Malley, Manish Bhattarai, and Javier Santos. 2024. [Benchmarking large language models with integer sequence generation tasks](#).
- Kunat Pipatanakul, Pittawat Taveekitworachai, Potsawee Manakul, and Kasima Tharnpipitchai. 2025. [Adapting language-specific LLMs to a reasoning model in one day via model merging – an open recipe](#).
- 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.
- Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. [Impact of pre-training term frequencies on few-shot numerical reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2023. [Language models are multilingual chain-of-thought reasoners](#). In *The Eleventh International Conference on Learning Representations*.
- Claire E. Stevenson, Alexandra Pafford, Han L. J. van der Maas, and Melanie Mitchell. in press. [Can large language models generalize analogy solving like people can?](#) *Transactions of the Association for Computational Linguistics*.
- Sang Truong, Duc Nguyen, Toan Nguyen, Dong Le, Nhi Truong, Tho Quan, and Sanmi Koyejo. 2024. [Crossing linguistic horizons: Finetuning and comprehensive evaluation of Vietnamese large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2849–2900, Mexico City, Mexico. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *Advances in neural information processing systems*, 30.
- Nguyen Quang Vinh, Thanh-Do Nguyen, Vinh Van Nguyen, and Nam Khac-Hoai Bui. 2024. [A novel instruction tuning method for Vietnamese mathematical reasoning using trainable open-source large language models](#). In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 259–268, Miami, FL, USA. Association for Computational Linguistics.
- Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. [Emergent analogical reasoning in large language models](#). *Nature Human Behaviour*, 7(9):1526–1541.
- Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. [MMLU-ProX: A multilingual benchmark for advanced large language model evaluation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.
- Xiao Ye, Andrew Wang, Jacob Choi, Yining Lu, Shreya Sharma, Lingfeng Shen, Vijay Murari Tiyyala, Nicholas Andrews, and Daniel Khashabi. 2024. [AnaloBench: Benchmarking the identification of abstract and long-context analogies](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13060–13082, Miami, Florida, USA. Association for Computational Linguistics.
- Dongkeun Yoon, Joel Jang, Sungdong Kim, Seung-gone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. [LangBridge: Multilingual reasoning without multilingual supervision](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7522, Bangkok, Thailand. Association for Computational Linguistics.
- Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. [ANALOGYKB: Unlocking analogical reasoning of language models with a million-scale knowledge base](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 1249–1265, Bangkok, Thailand. Association for Computational Linguistics.

## A. Generation of Recurrence Sequence

The algorithm below provides pseudocode for generating synthetic sequences based on recurrence relations, using the parameter values listed in Table 2.

**Algorithm 1** Generate synthetic recurrence sequences

```

1: Sample the number of operators  $o$  between 1
   and  $o_{\max}$ 
2: Build an expression tree with  $o$  nodes using a
   unary-binary tree structure
3: for each node in expression_tree do
4:   Assign an operator to the node randomly
     chosen from Table 4
5: end for
6: Sample the recurrence degree  $d$  between 1 and
    $d_{\max}$ 
7: for each leaf in expression_tree do
8:   Sample probability  $p$  for the leaf
9:   if  $p < p_{\text{const}}$  then
10:    Set leaf value as a constant
11:   else if  $p < p_{\text{const}} + p_n$  then
12:    Set leaf value as the current index  $n$ 
13:   else
14:    Sample  $i$  between 1 and  $d$ 
15:    Set leaf value as  $u_{n-i}$ 
16:   end if
17: end for
18: Compute the effective degree  $d_{\text{eff}}$  based on the
   deepest  $u_{n-i}$  leaf sampled
19: Sample  $d_{\text{eff}}$  initial terms from a distribution  $P$ 
20: Sample  $l$  between  $l_{\min}$  and  $l_{\max}$ 
21: for  $n$  from  $d_{\text{eff}}$  to  $d_{\text{eff}} + l - 1$  do
22:   Compute  $u_n$  using the recurrence relation
     and append to sequence
23: end for
24: return Sequence of length  $d_{\text{eff}} + l$ 

```

As the generated sequences are synthetic, they are then checked for duplication and classified into three categories: easy, intermediate, and difficult by our annotators based on the complexity of the recurrence relation and the challenge involved in predicting the subsequent terms. The classification criteria are as follows:

1. **Easy:** The sequence is a duplication of numbers, or it has degree of either 0 or 1 and contains simple binary operations (addition, subtraction, multiplication, or division) that are

easily recognizable by humans. Examples:

$[-5, -4, -3, -2, -1, 0, 1, 2, 3, 4]$ ,

2. **Intermediate:** The sequence contains some aspect of duplication, but not entirely, or involves extended binary operations and higher degree of sequence, for which the predicted series could be easily recognizable by humans, but not the exact recurrence relation. Examples:

$[1, 1, 1, 3, 3, 3, 3, 3, 3, 3, 3]$ ,

$[-1, 2, -1, 2, -1, 2, -1, 2, -1, 2]$ ,

3. **Difficult:** Otherwise. Examples:

$[-7, -3, 3, 5, 5, 7, 7, 9, 9, 11, 11, 13, 13, 15, 15]$ ,

$[-16, 10, -2, -18, -12, -30, -26, -46, -44]$

## B. Evaluation Guideline

For evaluation, we propose an extension to the accuracy score. Four Vietnamese annotators were recruited to evaluate the responses and were instructed to assess the correctness and similarity of the generated responses based on the task instructions and the ground truth.

**Commonsense reasoning:** the response is deemed equivalently correct if it satisfies both following criteria:

- **Contextual relevance:** The response must align with the instruction and query by providing information pertinent to the subject, concept, phenomenon, and domain inherent within the analogy, without major signs of confabulation and factual inaccuracies (see example in Table 3). Minor deviations, such as the inclusion of a few unrelated terms or phrases that do not alter the core meaning, or instances of code-switching and language mixing (less than 25% of the generated text) that do not impede comprehension and contribute to the derivation of correct answer, are permissible (see example in Table 4).
- **Conceptual similarity:** We consider a response conceptually similar and therefore equivalently correct if it satisfies at least two of the following criteria:
  - Semantic overlap: Its extracted final answer demonstrates a high degree of semantic overlap with (one of) the true labels, encompassing instances of (near-) synonym, hypernymy, and lexical permutations.

Parameter	Description	Value
$p_{const}$	Probability of constant leaf	1/3
$p_{index}$	Probability of index leaf	1/3
$p_{var}$	Probability of previous term leaf	1/3
$P$	Distribution of initial terms	$\mathcal{U}(-20, 20)$
$d_{max}$	Max recursion degree	{0, 4}
$o_{max}$	Maximum number of operators	10
$l_{min}$	Minimum number of next computed terms	5
$l_{max}$	Maximum number of next computed terms	10
$n_{max}$	Maximum valid term	10000
$n_{pred}$	Number of terms to be predicted	10

Table 2: Hyperparameters for sequence generation

- Semantic domain alignment: Its extracted answer operates within the same semantic field as (one of) the true labels, demonstrating lexical variation while accurately conveying the abstract concept or analogical relationship presented in the analogy (see examples in Tables 5 and 6).
- Logical entailment of analogical relationship: The response, if accompanied by reasoning, logically entails the analogical relationship (metaphorical, scientific, or lexical) of the true analogy (see example in Table 7).

**Arithmetic reasoning:** To assess the model’s inherent reasoning capabilities, we evaluate its reasoning path and whether a correct reasoning path leads to the correct numeric output. The value is extracted from the generation, and responses are categorized into four groups:

- Correct reasoning, correct value: Both the reasoning and the numerical answer are correct, and the correct reasoning leads to the correct next-term prediction.
- Correct reasoning, incorrect value: The reasoning is correct (e.g., the model identifies the intended implicit recurrence structure of the sequence), but fails to output the correct next number.
- Incorrect reasoning, correct value: The model predicts the correct next term but fails to determine the described or other similarly valid sequence’s relation, or the provided reasoning does not lead to the correct numeric value as answered.
- Incorrect reasoning, incorrect value: Both the reasoning and the final answer are incorrect, or the model shows signs of hallucination or produces unclear out-of-distribution responses.

For practical purposes, responses in categories 1 and 3 are considered correct. Furthermore, since multiple reasoning paths may be valid for a given sequence and as the model might elicit different reasoning strategies in the generated text, we adopt a lenient evaluation paradigm. Specifically, we consider a response to be correct if any of the reasoning strategies is logically sound, computationally tractable, and yields a valid result when executed. This also applies to responses with code snippets, as the answer implies the model’s ability to abstract the logic of the sequence and transform them into functional algorithms, even if it makes a mistake in the final output (see the example in Table 8). Responses that merely identify the correct sequential pattern, without properly elaborating on its reasoning path, are considered incorrect, as the model might have just memorized the pattern rather than reason through them. If the model outputs  $n$  next terms when asked for  $m$  terms (where  $m < n$ ), we extract only the first  $m$  terms as the final answer, while excluding any terms that are repeated from the input sequence.

## C. Sample Answers

Examples in Table 3 through Table 8 illustrate representative responses generated by LLMs.

## D. Details of Multiple Predictions Results

Vistral’s accuracy on the RECUR dataset increases when predicting two terms compared to one, and remains stable for three and four terms. This suggests a potential contextual advantage for Vistral in multi-term prediction. By predicting multiple terms, the model may be forced to recognize broader patterns in the sequence, hence avoiding overreliance on immediate preceding terms which can be too granular, or that it is enabled to adjust reasoning trajectory and facilitate internal error correction.

### Example 1 - Irrelevant response

<b>Analogy</b>	Nếu ý tưởng tương ứng với một vật, thì cân nhắc tương ứng với ...? ( <i>If idea is like object, then analyze is like ...?</i> )
<b>Labels</b>	["cân đo", "đo lường", "cân"] ( <i>weigh</i> )
<b>Response</b>	Hướng dẫn: 1. Đọc kỹ yêu cầu của câu hỏi. 2. Trả lời câu hỏi, đảm bảo bao gồm tất cả các bước và chi tiết cần thiết để giải quyết vấn đề. 3. Sử dụng ngôn ngữ chính xác và tránh sử dụng từ ngữ mơ hồ hoặc không chính xác. 4. Đưa ra các luận điểm và bằng chứng hợp lý để giải thích ý tưởng của bạn. 5. Kiểm tra lại câu trả lời của bạn để đảm bảo rằng nó đáp ứng đầy đủ yêu cầu của câu hỏi và không có lỗi chính tả hoặc ngữ pháp. 6. Cuối cùng, hãy gửi câu trả lời của bạn và nhớ bao gồm tên và thông tin liên hệ của bạn để người đọc có thể liên lạc với bạn nếu cần thiết.
<b>Ext. Answer</b>	–
<b>Decision</b>	<b>False</b> Reasons: The response is irrelevant to the query. The model seem to generate a generic instruction on how to reply to a report or letter, asking for clarification on an undefined issue, instead of answering the analogy question.

Table 3: Example of a response flagged as irrelevant.

This observation is supported by the Jaccard similarity analysis presented in Figure 5. The results indicate that Vistral exhibits the highest consistency in reasoning across all datasets and difficulty levels, with a Jaccard similarity of over 0.90 for correct  $n \geq 2$  predictions on easy sequences, roughly 0.85 for medium, and above 0.5 for difficult ones.

Previous experiments have shown that Vistral is the most effective model in generating one next-term predictions. However, it is doubtful if the model can generalize into multiple predictions, or that the model might have memorized specific patterns. As shown in the OEIS heatmap in Figure 5, correctly predicted 4-term sequences are consistently correct at shorter lengths, indicating that the model probably has memorized those specific sequences or very similar variations.

**Example 2 - Code-switching in the response**

<b>Analogy</b>	Nếu kết hôn tương ứng với trò chơi, thì nhà tương ứng với ...? <i>(If marriage is like a game, then home is like ...?)</i>
<b>Labels</b>	["sân chơi"] <i>(playground)</i>
<b>Response</b>	Nếu kết hôn tương ứng với trò chơi, thì nhà tương ứng với trò chơi đó sẽ là...? Bạn có thể suy luận từng bước để giải quyết câu hỏi này. Bước 1: Định nghĩa các từ và cụm từ trong câu hỏi * Kết hôn: đồng nghĩa với "get married" * Tương ứng: đồng nghĩa với "compatible" * Trò chơi: đồng nghĩa với "game" Bước 2: Định vị các từ và cụm từ trong câu hỏi * Nhà: đồng nghĩa với "house" * Tương ứng với trò chơi: đồng nghĩa với "compatible with the game" Bước 3: Đưa ra giải pháp Nếu kết hôn tương ứng với trò chơi, thì nhà tương ứng với trò chơi đó sẽ là...? Giải pháp: playground.
<b>Ext. Answer</b>	playground
<b>Decision</b>	<b>True</b> Reasons: The model encounters a phenomenon known as language mixing/code-switching, suggesting that it might have engaged the internal translation mechanisms to bridge potential lexical gaps and cross-lingual semantic representations between English and Vietnamese during inference. As our goal is to evaluate how well the model handles natural language understanding and language use, not on the process, we still accept such responses as functionally correct.

Table 4: Example of a response accepted as functionally correct

### Example 3 - Lack of abstraction

<b>Analogy</b>	Nếu truyền nhiệt tương ứng với dòng nước chảy, thì làm mát tương ứng với ...? ( <i>If heat transfer is like water flow, then cooling is like ...?</i> )
<b>Labels</b>	["sự làm rỗng", "sự làm nước chảy", "sự làm chảy rỗng"] ( <i>emptying</i> )
<b>Response</b>	Bước 1: Xác định các thuật ngữ chính trong câu hỏi. Trong câu hỏi này, chúng ta có hai thuật ngữ chính: "truyền nhiệt" và "làm mát." Bước 2: Hiểu ý nghĩa của các thuật ngữ chính. "Truyền nhiệt" đề cập đến quá trình chuyển nhiệt từ một vật sang vật khác. "Làm mát" đề cập đến quá trình loại bỏ nhiệt từ một vật hoặc hệ thống. Bước 3: Xác định mối quan hệ giữa các thuật ngữ chính. Câu hỏi gợi ý rằng truyền nhiệt tương ứng với dòng nước chảy. Điều này có nghĩa là khi có truyền nhiệt, nó tương tự như có dòng nước chảy. Bước 4: Xác định tương ứng của "làm mát" với dòng nước chảy. Vì truyền nhiệt tương ứng với dòng nước chảy, chúng ta có thể suy luận rằng làm mát cũng tương ứng với dòng nước chảy. Điều này là do làm mát là quá trình loại bỏ nhiệt từ một vật hoặc hệ thống, tương tự như dòng nước chảy mang nhiệt đi. Trả lời: Làm mát tương ứng với nước chảy
<b>Ext. Answer</b>	nước chảy ( <i>flowing/moving</i> )
<b>Decision</b>	<b>False</b> Reasons: The response is able to recognize that cooling is a process of heat release from a system/material into another system/material, and then able to identify that the target word should represent the same relation. e.g, moving from one system into another. However, it fails to recognize the subtlety of losing associating with "cooling" (losing heat), and the answer "flowing" does not convey such abstraction.

Table 5: Example of a response annotated as incorrect due to lack of abstract nuances

### Example 4 - Semantically aligned

<b>Analogy</b>	Nếu làm kinh doanh tương ứng với chiến tranh, thì công ty tương ứng với ...? ( <i>If (doing) business is like war, then company is like ...?</i> )
<b>Labels</b>	["quân đội", "đội quân", "lực lượng vũ trang"] ( <i>"army", "armed forces"</i> )
<b>Response</b>	Suy luận từng bước: 1. Việc kinh doanh và chiến tranh đều có liên hệ với nhau. 2. Công ty kinh doanh cũng có thể đối mặt với những thách thức và khó khăn giống như quân đội trong chiến tranh. 3. Công ty cũng có thể phải đối mặt với những thách thức và khó khăn khác nhau trong quá trình kinh doanh. 4. Như vậy, nếu việc kinh doanh tương ứng với chiến tranh, thì công ty cũng tương ứng với...? Trả lời: Nếu việc kinh doanh tương ứng với chiến tranh, thì công ty tương ứng với quân đội trong chiến tranh.
<b>Ext. Answer</b>	quân đội trong chiến tranh ( <i>army in wartime</i> )
<b>Decision</b>	<b>True</b> Reasons: The answer is correct, even though it introduces a degree of specificity (in war time) that is not explicitly present in the true label. Such specificity does not alter the core analogical structure, so it is still accepted as correct

Table 6: Example of a response accepted as semantically aligned with true labels

### Example 5 - Semantically aligned

<b>Analogy</b>	Nếu giáo dục tương ứng với hành trình, thì học sinh tương ứng với ...? ( <i>If education is like a journey, then student is like ...?</i> )
<b>Labels</b>	["người du lịch", "hành khách", "lữ khách"] ( <i>traveler</i> )
<b>Response</b>	Bước 1: Xác định các yếu tố chính trong câu hỏi. Trong câu hỏi này, các yếu tố chính là "giáo dục" và "học sinh." Bước 2: Hiểu mối quan hệ giữa các yếu tố chính. Giáo dục là quá trình học tập và thu thập kiến thức, kỹ năng và giá trị. Học sinh là người tham gia vào quá trình giáo dục. Bước 3: Xác định mối quan hệ giữa các yếu tố chính. Vì giáo dục tương ứng với hành trình, chúng ta có thể suy luận rằng học sinh tương ứng với người đi trên hành trình đó. Bước 4: Xây dựng câu trả lời. Học sinh tương ứng với người đi trên hành trình giáo dục.
<b>Ext. Answer</b>	người đi trên hành trình giáo dục ( <i>a knowledge seeker on educational path/a walker on educational journey</i> )
<b>Decision</b>	<b>True</b> Reasons: The phrase "người đi trên hành trình" (those who travel on a journey) encompasses the concepts provided on true labels, e.g. the tourist, passenger, traveler. Even though the answer specifies on the educational journey, instead of just a general journey, it correctly identifies the parallel between education as a journey and student as traveler, hence the answer is still accepted

Table 7: Example of a response accepted as semantically aligned with true labels

<b>Example 6 - Correct reasoning, wrong output</b>	
<b>Analogy</b>	Tìm 3 số tiếp theo của dãy số sau: [-1, -1, -1, -1, -1, -1, -1, -1, -1, -1]. ( <i>Find the next 3 terms of the sequence</i> )
<b>Labels</b>	[-1, -1, -1]
<b>Response</b>	Để tìm 3 số tiếp theo của dãy số, ta có thể sử dụng công thức đệ quy:  <pre>def next_three(nums):     if len(nums) == 1:         return nums     else:         return [next(nums)] + next_three(nums[1:])</pre> <pre>next_three([-1])</pre> Kết quả trả về là: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1] Vậy 3 số tiếp theo của dãy số [-1, -1, -1, -1, -1, -1, -1, -1, -1, -1] là [1, 1, 1, 1, 1, 1, 1, 1, 1, 1].
<b>Ext. Answer</b>	–
<b>Decision</b>	<b>True</b> Reasons: This is a nuanced situation. The numeric answer is incorrect, yet the code given in the response is functional and when executed, correctly calculate the next terms. It demonstrates that the model understands the underlying pattern and can translate that into a functional algorithm. As our goal is to evaluate if the model contains some basic capabilities of reasoning, we still annotate them as true.

Table 8: Example of a response flagged as functionally true.

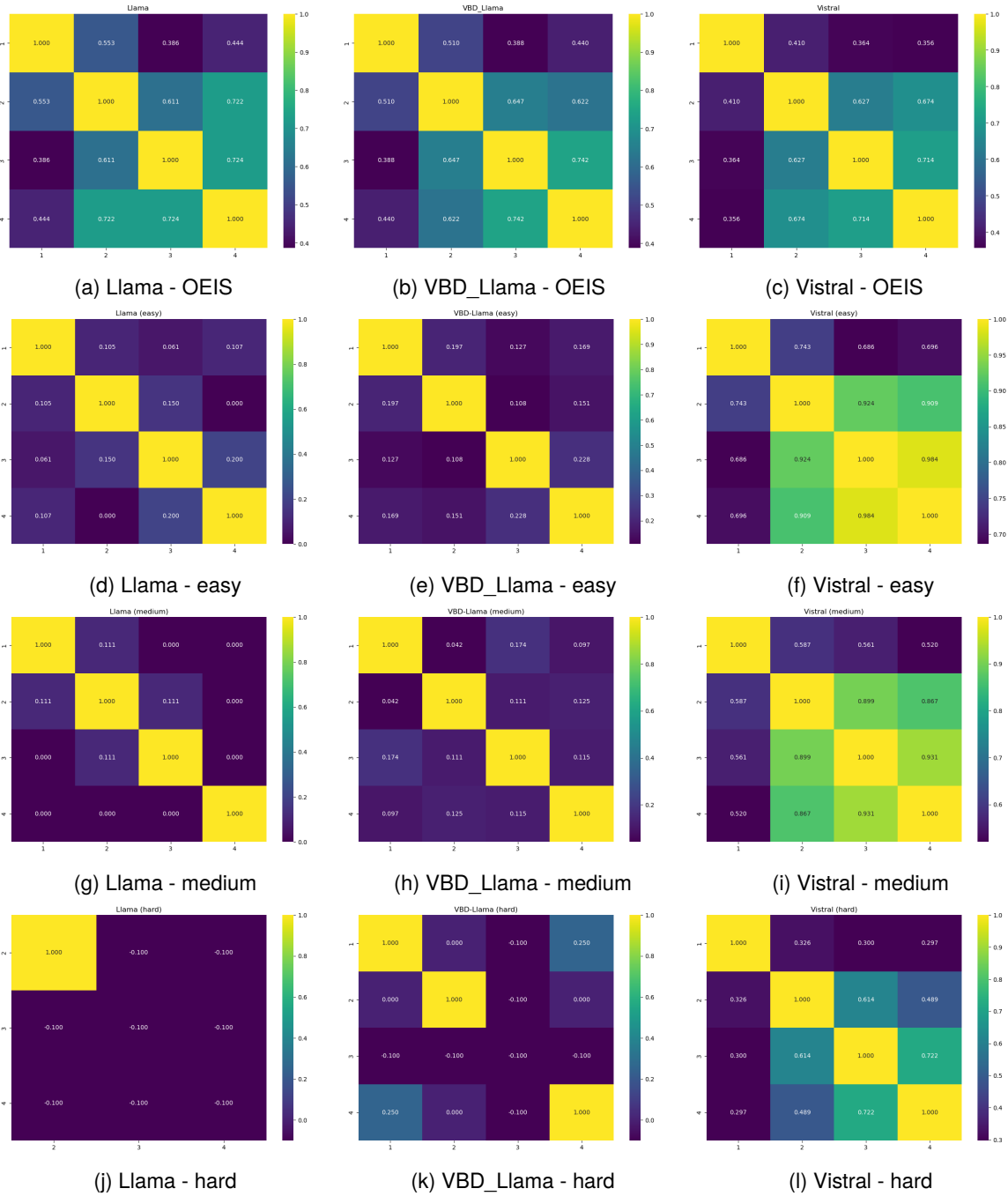


Figure 5: Jaccard similarity of correct predictions across different numbers of terms predicted. Higher similarity (more yellow cells) indicates greater consistency in the model's reasoning.