# VMLU Benchmarks: A comprehensive benchmark toolkit for Vietnamese LLMs

**Cuc Thi Bui**[1*], **Nguyen Truong Son**[2,3*], **Truong Van Trang**[1], **Lam Viet Phung**[1], **Pham Nhut Huy**[1], **Hoang Anh Le**[1], **Quoc Huu Van**[1], **Phong Nguyen-Thuan Do**[1], **Van Le Tran Truc**[1], **Duc Thanh Chau**[2,3], **Le-Minh Nguyen**[4]

[1]Zalo AI, Ho Chi Minh City, Vietnam
[2]University of Science, Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
[4]Japan Advanced Institute of Science and Technology, Japan
cucbt@vng.com.vn, ntson@fit.hcmus.edu.vn,
{trangtv,lampv,huypn5,anhlh6,quocvh,phongdnt,vanltt}@vng.com.vn
ctduc@fit.hcmus.edu.vn, nguyenml@jaist.ac.jp

## Abstract

The evolution of Large Language Models (LLMs) has underscored the necessity for benchmarks designed for various languages and cultural contexts. To address this need for Vietnamese, we present the first Vietnamese Multitask Language Understanding (**VMLU**) Benchmarks. The VMLU benchmarks consist of four datasets that assess different capabilities of LLMs, including general knowledge, reading comprehension, reasoning, and conversational skills. This paper also provides an insightful overview of the current state of some dominant LLMs, such as Llama-3 (Grattafiori et al., 2024), Qwen2.5 (Qwen et al., 2025), and GPT-4, highlighting their performances and limitations when measured against these benchmarks. Furthermore, we provide insights into how prompt design can influence VMLU's evaluation outcomes, as well as suggest that open-source LLMs can serve as effective, cost-efficient evaluators within the Vietnamese context. By offering a comprehensive and accessible benchmarking framework, the VMLU Benchmarks aim to foster the development and fine-tuning of Vietnamese LLMs, thereby establishing a foundation for their practical applications in language-specific domains.

## 1 Introduction

Evaluating benchmarks is crucial in AI development, as they are essential tools for assessing and comparing model performance across various dimensions. However, the rapid advancement of large language models (LLMs) has rendered traditional benchmarks inadequate for capturing their complex abilities. Tasks that require comprehension, reasoning skills, extensive world knowledge, and conversational proficiency demand more sophisticated and comprehensive evaluation frameworks. As a result, the emergence of LLMs necessitates the development of benchmarks that can effectively address these complex tasks.

Several benchmarks have been developed to address these challenges, such as MMLU (Hendrycks et al., 2020), which assesses multitask accuracy with nearly 16,000 multiple-choice questions, SQuAD (Rajpurkar et al., 2016) for reading comprehension with 107,785 QA pairs, including unanswerable questions in SQuAD 2.0 (Rajpurkar et al., 2018). Big-Bench Hard (BBH) (Suzgun et al., 2023) identifies LLM limitations through 23 challenging tasks, MT-Bench (Bai et al., 2024) assesses conversational abilities through multi-turn interactions, while IFEval (Zhou et al., 2023) evaluates instruction-following capabilities with over 500 natural language prompts.

These benchmarks offer valuable insights into the capabilities of LLMs on a global scale. However, they are primarily designed for English and other widely spoken languages, creating a significant gap when it comes to evaluating LLMs for less represented languages, such as Vietnamese. In this study, we focus on assessing the advanced capabilities of foundational models in the Vietnamese context, taking into account the language's unique linguistic and cultural characteristics.

There has been growing interest in developing resources for Vietnamese LLMs, including ViGLUE (Tran et al., 2024), which evaluates natural language understanding across 12 tasks from various domains, and the Comprehensive Evaluation Framework (Truong et al., 2024), which encompasses 10 tasks and 31 metrics. However, relying on simple English translation benchmarks often limits the ability to capture the linguistic and cultural nuances that are unique to Vietnamese.

To address the existing limitations and promote

*Corresponding Author



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11495–11515
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

research in Vietnamese language models (LLMs), we present the VMLU Benchmarks, a comprehensive evaluation framework specifically designed for these models. It includes four diverse tasks aimed at capturing the full range of language understanding and reasoning capabilities. The benchmarks include Vi-MQA, Vi-SQuAD, Vi-DROP, and Vi-Dialog. We create a dataset of 10,880 multiple-choice questions across 58 topics for Vi-MQA to assess foundational knowledge and cross-domain understanding. Vi-SQuAD, a reading comprehension task, comprises 3,310 questions derived from Vietnamese texts to evaluate text interpretation and comprehension. Vi-DROP, focusing on reasoning, consists of 3,090 questions spanning six reasoning categories to test logical and analytical skills. Lastly, Vi-Dialog, designed to assess conversational proficiency, features 210 multi-turn dialogues to evaluate coherence, contextual understanding, and conversational flow.

In addition to constructing the benchmark, this study evaluates several state-of-the-art models, including Llama-3, Qwen2.5, and GPT-4o, using VMLU Benchmarks. The results provide important insights into the performance of large-scale Vietnamese LLMs, emphasizing the significant impact of prompt design and evaluation methodologies on the results.

In this paper, we make the following contributions:

1. Our paper introduces VMLU Benchmarks, a toolkit comprising four tasks that comprehensively evaluate Vietnamese LLMs' capabilities in knowledge, reasoning, reading comprehension, and conversational abilities. This benchmark fosters research in Vietnamese LLMs.

2. We evaluate dominant LLMs with stronger Vietnamese abilities, offering comparative insights into their performance on various tasks under VMLU. Furthermore, we suggest some open-source LLMs can serve as effective, cost-efficient judges within the Vietnamese context.

3. We offer empirical insights to demonstrate the impact of prompt design on model performance, paving the way for improved methodologies in evaluating Vietnamese LLMs.

The rest of this paper is structured as follows: section 2 reviews existing LLM benchmarks. Section 3 introduces the VMLU benchmarks. In particular, we present experiments when evaluating some state-of-the-art LLMs with VMLU benchmarks and valuable insight in Section 4. Finally, Section 5 presents conclusions and future work.

## 2 Related Work

Our proposed VMLU Benchmarks draw inspiration from several prominent benchmarks in the field, adapting their methodologies to the Vietnamese language context to better capture its linguistic and cultural intricacies.

**General Knowledge** Vi-MQA is inspired by MMLU (Hendrycks et al., 2020). MMLU evaluates multitask accuracy with nearly 16,000 multiple-choice questions across diverse subjects such as STEM, humanities, and social sciences. This benchmark is designed to evaluate LLMs' capabilities across various subjects. Following the idea of MMLU, different LLM benchmarks are also introduced in other languages, such as CMMLU (Li et al., 2024) in Chinese, Turkish-MMLU (Yüksel et al., 2024) in Turkish, and ArabicMMLU (Koto et al., 2024) and KMMLU (Son et al., 2024) in Arabic and Korea, respectively.

**Reading Comprehension** Vi-SQuAD is directly inspired by SQuAD (Rajpurkar et al., 2016), which has become a standard for reading comprehension tasks in NLP. SQuAD features a vast collection of questions based on English texts, including unanswerable questions in its second version, SQuAD 2.0 (Rajpurkar et al., 2018). Several benchmarks are also developed with the same idea as SQuAD but expanded to multilingual understanding, such as MLQA (Lewis et al., 2020), which consists of over 5,000 extractive QA instances in SQuAD format available in seven languages, MKQA (Longpre et al., 2021) includes 10,000 question-answer pairs aligned across 26 languages, and Indic-QA (Clark et al., 2020) multilingual dataset that focuses on question answering across multiple typologically diverse languages.

**Reasoning** Vi-DROP, which focuses on reasoning abilities, takes inspiration from the DROP (Dua et al., 2019) benchmark. DROP tests models' abilities to perform complex reasoning over paragraphs, including approximately 96,000

questions containing numerical operations and logical inference. Other well-known reasoning datasets include GMS8K (Cobbe et al., 2021), comprising grade-school math word problems designed to test the ability of models to reason numerically step-by-step, and HOTPOTQA (Yang et al., 2018), multi-hop reasoning over paragraphs, which requires the model to combine information from multiple documents to answer a question.

**Conversational ability** Vi-Dialog is built upon elements from the dialogue test set developed by TencentLLMEval (Xie et al., 2023). This test set encompasses 220 multi-turn dialogues and focuses on conversation tasks to assess models' knowledge, long-term memory, and contextual understanding. Additional benchmarks like MT bench (Bai et al., 2024) assess LLM's ability to engage in multi-turn dialogues by simulating real-life conversations, measuring how well chatbots follow instructions and maintain a natural flow of conversation.

## 3 VMLU Benchmarks

This section presents the data construction and evaluation process, along with details about four benchmarks, including data collection, processing, and quality assurance measures. Note that this dataset was previously published. [1]

### 3.1 Dataset construction and evaluation

**Annotators' backgrounds**: All annotators were university students from leading institutions in Vietnam, including Foreign Trade University and Hanoi Law University. They possess strong expertise across both natural sciences and social sciences. More than 70% of annotators were selected from top-tier universities under the Vietnam National University system.

**Question refinement process**: We implemented a two-phase question refinement process. In the first phase, we created standardized test datasets to ensure comprehensive coverage, accuracy, and consistency. Annotators had to achieve 100% accuracy in a rigorous qualification test before being allowed to participate in the project. The second phase involved monitoring their performance throughout the project, with tasks assigned according to their demonstrated accuracy and adherence to guidelines.

**Measures taken to ensure data quality** We established a multi-layered quality assurance process that involved labelers and reviewers. Three independent reviewers evaluated each data entry, which was only accepted if two of the three reviewers reached an agreement. If a data point did not meet this criterion, it entered a maximum of two rework cycles (indicated as "rework = 2"). If any data point required more than two reworks, all labeled data from the responsible annotator was discarded and re-evaluated. The Data Quality Control (QC) team also conducted random checks on 20% of the data to ensure that overall error rates remained below 3%. All annotation activities were carried out on a secure platform that featured stringent access controls and tracking mechanisms to prevent errors and maintain quality.

All subsets in the VMLU benchmarks are subjected to this data construction process.

### 3.2 Dataset

#### 3.2.1 Vi-MQA

**Task Overview** Vi-MQA is a multiple-choice question answering benchmark designed to evaluate the overall capabilities of foundational models, particularly focusing on the Vietnamese language. The questions come from various subjects, organized into four main categories: STEM (Science, Technology, Engineering, Mathematics), Humanities, Social Sciences, and a broad category called 'Other'.

**Data Collection** Vi-MQA includes four difficulty levels: Elementary School, Middle High School, High School, and Professional level. Table 5 shows examples of the Vi-MQA datasets with each difficulty level.

For the Elementary, Middle and High School levels, we collected standardized exams from seven primary subjects in the Vietnamese general education curriculum, excluding English, Physical Education, and Arts. Additionally, we have included high-quality Vietnamese National High School Graduation Exams compiled annually by high school instructors and the Ministry of Education and Training. These exams assess students' knowledge of natural, language, and social sciences and their ability to reason and respond to general world knowledge.

At the advanced Professional level, the benchmark consists of mock exams based on topics from undergraduate and graduate programs designed to evaluate proficiency in complex reasoning and

[1] https://vmlu.ai/

problem-solving skills. The subjects range from standard areas such as history and mathematics to more specialized fields like law and accounting, ensuring a comprehensive assessment of extensive knowledge. We also reference official vocational qualification tests and have selected six representative subjects, including legal professionals, civil servants, and tax civil servant qualification exams. These subjects are categorized into four groups: STEM (Science, Technology, Engineering, and Mathematics), Humanities, Social Sciences, and others.

Our primary sources for the mock exams are freely available online materials, including a vast collection of midterm, final, and past exams from top-tier universities and high schools in Vietnam. We have also gathered official exam questions from the Ministry of Education and Training, along with illustrated exam queries. For specialized subjects such as law, we collected mock questions shared privately by students. These questions are not available online, and we obtained permission to include approximately 1,000 questions in Vi-MQA.

**Data processing** We gathered raw data from various sources such as PDF, Microsoft Word documents, and public textbooks on the Internet and then meticulously processed them. Initially, we used an OCR tool to convert PDFs into text format. After that, we combined both automatically and manually parsed the questions, carefully converting them into a structured format to ensure data consistency. For complex mathematical notations, which are common in STEM subjects, we manually parsed them into standard LATEX format.

**Quality Check** All questions undergo a standard data pre-processing pipeline that includes de-duplication and cleaning, followed by several rounds of human validation. In particular, all LATEX notations are ensured to be compiled without syntax errors. Note that all the questions in this dataset are excluded from explanations to evaluate models genuinely, innovatively, and devoid of bias.

**Format** Each entry in the Vi-MQA dataset is structured as a multiple-choice question with three, four, or five options, of which only one is correct. The questions are presented clearly and concisely, suitable for automated processing and evaluation.

**Statistics** Vi-MQA contains 10,880 questions distributed across 58 subjects. Each subject maintains a minimum of around 200 questions, ensuring comprehensive coverage and statistical significance. The dataset is partitioned into a development and test set, with a designated portion allocated for fine-tuning and validation purposes. Regarding subject distribution, Vi-MQA includes 21 STEM tasks, 18 Humanities tasks, 10 Social Science tasks, and 9 tasks categorized under Other. The details are shown in 6. This diverse composition allows for a balanced evaluation of foundational models across various disciplines, reflecting both general knowledge and specialized problem-solving capabilities.

#### 3.2.2 Vi-SQuAD

**Task Overview** Vi-SQuAD is a benchmark dataset designed to evaluate the reading comprehension abilities of Vietnamese language models. Inspired by the SQuAD benchmark, it considers Vietnamese's unique linguistic and contextual features. The example is presented in Table 7.

**Data Collection** Vi-SQuAD was constructed using data from Vietnamese Wikipedia, combined OpenAI GPT-3.5, and human labelers. Initially, passages were selected from Wikipedia, which covers a wide range of topics, spanning 19 popular categories such as Economics, Sports, History, Music, and more. This approach ensured a diverse and comprehensive dataset for Vietnamese question-answering tasks. These passages were then fed into GPT-3.5, which generated synthetic question-answer pairs. Notably, we aimed to generate questions that went beyond the standard types found in SQuAD (What, Which, Who, When, Where); it also expanded to encompass more difficult question types, such as How and Why. Human experts then reviewed, verified, and audited these pairs to ensure accuracy. Additionally, human annotators expanded the dataset by introducing variations of correct answers, maintaining the same meaning but employing different expressions. This multi-step process ensures the dataset is both high-quality and diverse, capturing a wide range of potential responses.

**Benchmark Detail** The dataset comprises 3,310 samples, each consisting of a passage, a question, and multiple potential answers. Vi-SQuAD presents a significant challenge due to its longer paragraphs, which average 272 words - more than double the 120-word average of the original SQuAD version. This increase in length demands deeper comprehension and contextual reasoning in order to extract the correct answers. The questions are categorized into two types:

- *Answerable Questions* (94.6%): answers are provided in various formats.

- *Unanswerable Questions* (5.4%): answers include phrases such as "NO ANSWER", "Không có câu trả lời" or "không tìm thấy câu trả lời".

#### 3.2.3 Vi-DROP

**Task Overview** Vi-DROP is a benchmark inspired by DROP, designed to assess the reasoning abilities of foundational models in the Vietnamese language. It emphasizes discrete reasoning, demanding that models tackle complex reasoning tasks that go beyond simple reading comprehension. Vi-DROP examples can be found in Table 8.

**Data Collection** The annotation procedure is divided into three stages: passage choosing, question collection, and validation.

First, we utilized an automated process to extract passages from Vietnamese Wikipedia, focusing particularly on those with numerical content to facilitate complex reasoning tasks. This process resulted in a collection of approximately 500 passages.

Secondly, GPT-4o model was used to generate question-answer pairs based on these passages. We designed questions across six categories: addition and subtraction, calculator, selection, comparative, superlative, counting, and other arithmetic operations (e.g., percentages and fractions). For each passage, we generate 3 questions in the addition and subtraction category, 2 in the selection and comparative categories, and 1 in each superlative and counting category. This distribution is inspired by the original DROP dataset. Subsequently, Labelers were tasked with reviewing the generated questions to ensure clarity, alignment with the passages, and grammatical accuracy. The review process required labelers to confirm that each question was clear and well-structured and verify that the questions were appropriately related to the given passages. They need to revise unclear or grammatically incorrect questions to improve coherence. Similarly, the provided answers were carefully reviewed. Labelers were asked to validate the correctness of the answers and suggest additional valid answers where applicable to enhance evaluation diversity. Through this iterative process, we discarded approximately 30% of the data due to unclear questions or a mismatch between the question and the passage. After filtering, we analyzed the proportions of questions in each category and supplemented any underrepresented categories to balance the dataset.

**Benchmark Details** This benchmark consists of 3,090 questions that feature a variety of question types, categorized into six distinct reasoning types to challenge foundational models across different reasoning tasks. The reasoning types include Addition/Subtraction, Selection, Comparative, Superlative, Counting, and Other Arithmetic. It is important to note that a single question may encompass more than one type of reasoning. For example, the question *"How many events took place in 1975?"* requires both comparative and counting operations. Table 1 presents examples for each category, along with the corresponding number of questions in each.

#### 3.2.4 Vi-Dialog

**Task Overview** Vi-Dialog is a dialogue benchmark specifically designed for the Vietnamese context. Its purpose is to evaluate the knowledge and conversational capabilities of foundation models. Vi-Dialog examples can be found in Table 9.

**Benchmark Detail** We created the Vi-Dialog Benchmark by translating Tencent LLMEval into Vietnamese, with revisions to highlight common knowledge related to Vietnamese geography, history, and logical reasoning. Subsequently, the samples were annotated to ensure conversations sounded natural and human-like. Additionally, we varied the number of turns in each dialogue randomly to enhance the diversity of user intents and improve the flow of conversation. The benchmark was finalized through collaboration with multiple contributors, and the final version includes 210 conversations, totaling approximately 1,200 turns.

## 4 Evaluating public LLMs with VMLU Benchmarks

### 4.1 Experimental settings

We benchmark several LLMs that support Vietnamese from OpenAI, Meta AI, and Alibaba on the VMLU benchmarks below.

- **GPT-4o** is a large language model developed by OpenAI, known for its advanced performance in various natural language understanding tasks. It has been fine-tuned and optimized for multilingual capabilities, including Vietnamese.

- **Llama-3** Developed by Meta AI, it is an open-source series model that offers multiple model sizes ranging from 1B to 70B parameters.

| Category | Number | Description |
|---|---|---|
| add_sub | 1073 | Adding or subtracting numerical values from the text. |
| selection | 793 | Choosing specific information or entities from the text. |
| comparative | 724 | Comparing two or more entities mentioned. |
| superlative | 419 | Identifying the highest or lowest attribute. |
| counting | 514 | Determining the frequency of events or entities. |
| other arithmetic | 183 | Other calculations such as percentages. |

Table 1: Vi-DROP Category

It significantly improves computational efficiency and accuracy for underrepresented languages, including Vietnamese.

- **Qwen2.5** is a collection LLM developed by Alibaba, encompassing 7 sizes from 0.5B to 72B parameters. It is a cutting-edge LLM optimized for multilingual tasks, with specific fine-tuning for Vietnamese language data.

These models currently achieve state-of-the-art performance for almost Vietnamese LLM tasks.

### 4.2 Inference Method

We employ multiple inference strategies to ensure a comprehensive evaluation of VLMU benchmarks, including zero-shot, few-shot, and chain-of-thought (CoT) prompting techniques. These methods allow us to assess the performance of language models under varying conditions of prior knowledge and reasoning complexity.

All four benchmarks are evaluated using a no-CoT prompt. The evaluation of Vi-Dialog has a unique approach: we sequentially input each conversational turn into the model. This iterative method allows us to assess the model's ability to maintain coherence, contextual awareness, and conversational flow across multiple dialogue turns. Each response is based on the preceding turns, simulating a realistic conversational environment.

For the CoT prompt inference, we experiment with Vi-MQA and Vi-DROP since the tasks require reasoning ability. This approach encourages the model to generate detailed reasoning processes before arriving at an answer. Due to resource constraints and long inference times, we only conduct few-shot inference experiments with Vi-MQA.

All prompts used for evaluating LLMs on Vi-MQA, Vi-SQuAD, and Vi-DROP are detailed in Tables 10, 11, 12, 13, 14.

### 4.3 Evaluation Method

Among the four tasks in the VMLU benchmarks, only the Vi-MQA task can be fully evaluated automatically. This evaluation is achieved by comparing the model's short multiple-choice answers (A, B, C, D, E) with the correct answers. In contrast, the other tasks, including Vi-Drop, Vi-SQuAD, and Vi-Dialog, produce generated responses that often vary significantly in length and structure. This variability complicates the automation of performance evaluation against the ground truth for these tasks.

Commonly used automatic evaluation metrics, such as F1 and BLEU scores, measure the overlap between system-generated responses and the ground truth. While these methods are computationally efficient, they are insufficient for accurately evaluating tasks that require semantic understanding. Table 2 illustrates instances where automatic metrics and human evaluations differ. Two typical scenarios include (1) high overlap but incorrect answers and (2) low overlap but correct answers. For example, in Question 1, although the system-generated response is correct and contains valuable information, it has minimal overlap with the ground truth, resulting in a low F1 score of 0.1. These observations highlight the limitations of automatic metrics, particularly in tasks that necessitate nuanced semantic evaluations. Human evaluation offers a more reliable alternative in these cases; however, it is resource-intensive and requires stringent protocols to ensure result reliability. Recent advancements in large language models (LLMs) from companies like OpenAI, Meta, and Alibaba have popularized the concept of using "LLM-as-a-Judge" (Zheng et al., 2023). This approach leverages LLMs to evaluate tasks that traditional automatic metrics often struggle to assess effectively. However, careful consideration must be given to the nature of the evaluation tasks to ensure that this method is both feasible and appropriate.

| Question | System Response | Ground Truth | F1 | Human | Qwen2.5-72B | GPT-4o |
|---|---|---|---|---|---|---|
| (1) Thủ đô của Việt Nam là thành phố nào? (*What city is the capital of Vietnam?*) | Thủ đô của Việt Nam là Hà Nội, một trong những trung tâm chính trị, văn hoá và kinh tế quan trọng. (*The capital of Vietnam is Hanoi, one of the important political, cultural, and economic centers.*) | Hà Nội (Hanoi) | 0.1 (✗) | ✓ | ✓ | ✓ |
| (2) Nước Mỹ có bao nhiêu bang? (*How many states does the United States have?*) | Nước Mỹ có 51 bang. (*The United States has 51 states.*) | Nước Mỹ có 50 bang. (*The United States has 50 states.*) | 0.8 (✓) | ✗ | ✗ | ✗ |

Table 2: Comparison of system answers with ground truth using F1 score, LLM judges, and human evaluation

| Alignment | #Agree | % |
|---|---|---|
| GPT-4o vs Human | 1365 | 97.2% |
| Qwen2.5-72B vs Human | 1361 | 96.9% |
| Qwen2.5-72B vs GPT-4o | 1367 | 97.3% |

Table 3: Agreement Rates Between Models and Human Evaluations in **1405** cases of the ViDrop benchmark

For tasks that involve comparing short texts generated by the model to established ground truth answers, such as in the Vi-SQuAD and Vi-Drop benchmarks, we simply prompt the model to compare the semantic meanings of its outputs with the ground truth. In the case of the Vi-Dialog benchmark, we adopt the scoring methodology outlined in TencentLLMEval (Xie et al., 2023), evaluating outputs against seven criteria: Safety, Neutrality, Compliance with facts, Relevance, Logicality, Language fluency, and Information content. Detailed information regarding the judging prompt can be found in Tables 15 16.

Regarding evaluators, several leading LLMs are available, including GPT-4, the Qwen2.5 series, and the Llama-3 series. For this evaluation, we selected two models for comparison: one accessed through an API and another that is publicly available. Specifically, we evaluated the performance of GPT-4o and Qwen 2.5 72B by comparing their assessments with those made by human evaluators. We conducted experiments using a small set of 1,405 Vi-DROP questions to compare the different evaluators. The results are summarized in Table 3.

For tasks involving comparing short texts between model output and ground truth, such as in the Vi-SQuAD and Vi-DROP benchmarks, the evaluation performance of LLMs like Qwen2.5 72B and GPT-4o demonstrates high accuracy, closely aligning with human-level quality. These findings suggest that using LLMs as judges can provide a more accurate alternative to automated metrics, such as the F1 score.

However, evaluating multiple-turn dialogues remains a challenge. This is supported by findings in (Xie et al., 2023), which indicate low agreement between human evaluators, with a score of 0.49. In our experience, we also noted difficulties in achieving agreement between LLM evaluators and human labels. Nevertheless, this experience highlighted an important insight: GPT-4o and Qwen2.5 72B exhibited a high Pearson Correlation Coefficient (r) of 0.9736, indicating strong alignment in their evaluations.

Besides, Table 17 also shows a strong alignment between GPT-4o and Qwen2.5 72B as evaluators in Vi-SQuAD task. These experiences underscore an essential takeaway for researchers: in addition to the GPT-4o, open-source LLMs can be effectively utilized as judges for tasks that require simple comparison.

| Model | Vi-MQA | | | Vi-SQuAD | Vi-DROP | | Vi-Dialog |
|---|---|---|---|---|---|---|---|
| | no CoT 0-shot | CoT 0-shot | no CoT 5-shot | no CoT 0-shot | no CoT 0-shot | CoT 0-shot | no CoT 0-shot |
| GPT4o | 74.0 | **78.3** | 77.3 | 96.3 | 85.1 | 89.4 | **92.7** |
| Llama-3.2-1B-Instruct | 37.6 | 25.7 | 38.2 | 70.1 | 32.7 | 29.6 | 33.9 |
| Llama-3.2-3B-Instruct | 47.6 | 40.0 | 46.2 | 90.3 | 53.5 | 63.5 | 50.8 |
| Llama-3-8B-Instruct | 52.0 | 50.0 | 52.1 | 93.1 | 67.3 | 80.3 | 66.5 |
| Llama-3.1-8B-Instruct | 54.1 | 52.4 | 52.6 | 92.0 | 64.3 | 75.6 | 60.0 |
| Llama-3.1-70B-Instruct | 68.7 | 72.2 | 69.1 | 95.1 | 77.5 | 87.6 | 73.2 |
| Llama-3.3-70B-Instruct | 69.1 | 73.9 | 68.7 | *96.4* | 79.1 | *89.7* | 74.8 |
| Qwen2.5-0.5B-Instruct | 39.1 | 26.3 | 39.2 | 62.5 | 32.1 | 31.5 | 28.0 |
| Qwen2.5-1.5B-Instruct | 48.7 | 42.8 | 49.2 | 86.7 | 45.7 | 54.5 | 39.8 |
| Qwen2.5-3B-Instruct | 52.9 | 51.2 | 54.7 | 88.3 | 49.1 | 72.4 | 54.4 |
| Qwen2.5-7B-Instruct | 59.1 | 59.1 | 60.6 | 91.8 | 63.8 | 81.4 | 65.4 |
| Qwen2.5-14B-Instruct | 66.4 | 68.1 | 66.9 | 96.1 | 75.1 | 87.7 | 73.1 |
| Qwen2.5-72B-Instruct | 72.4 | 74.2 | 73.3 | **96.8** | 81.0 | **89.9** | *88.2* |

Table 4: Detailed results of LLMs evaluated on the VMLU benchmark. The best results for each task are in **bold** text

### 4.4 Benchmark results

In this section, we present the benchmark results of various open-source models on the VMLU benchmarks. The results for Vi-MQA were evaluated automatically, while Vi-SQuAD and Vi-DROP were evaluated using Qwen2.5-72B, and Vi-Dialog was judged using GPT-4o. Table 4 summarizes the results achieved by each model on the VMLU benchmarks.

Overall, GPT-4o outperformed other models in two tasks: Vi-MQA and Vi-Dialog, achieving scores of 78.3% and 92.7%, respectively. Meanwhile, Qwen2.5-72B achieved the highest scores in the other two tasks, Vi-SQuAD and Vi-DROP, with scores of 96.8% and 89.9%. We found that the Llama models demonstrate improved performance with increasing model size. For example, Llama-3.3-70B-Instruct attained the highest score for all 4 tasks, scores of 73.9% on Vi-MQA, 96.4% on Vi-SQuAD, 89.7% on Vi-DROP and 74.8% on Vi-Dialog. Similarly, the Qwen2.5 models exhibited progressively better performance as their sizes increased. In contrast, smaller models such as Llama-3.2-1B-Instruct and Qwen2.5-0.5B-Instruct yielded lower scores, indicating that capacity limitations affect their performance. This analysis highlights that both model size and architecture play critical roles in determining performance. A comparison of models of similar sizes, such as Qwen2.5-72B versus Llama-3.1-70B, Qwen2.5-7B versus llama-3.1-8B, and Qwen2.5-3B versus LlamA-3.2-3B, revealed that Qwen2.5 consistently outperformed Llama-3, reinforcing findings previously reported in the Qwen2.5 Technical Report.(Qwen et al., 2025) on various datasets.

While the overall results indicate strong performance for leading LLMs on VMLU benchmarks, we conduct a deeper analysis that reveals specific areas where the datasets continue to pose significant challenges. For instance, in the Vi-MQA dataset, categories such as "other" in **??** continue to show relatively lower accuracy for advanced models like Qwen2.5-72B(68.2%) and Llama3.3-70B (69.8%), particularly in specialized subjects such as Accounting and Environmental Engineering. Similarly, for Vi-DROP, categories like "count" remain challenging, with even the top-performing models scoring below 78% . Table 19 illustrates the performance of leading models across top Vi-DROP categories, showing persistent gaps in specific reasoning types. Moreover, our benchmarks are crucial for evaluating and advancing smaller-sized models (under 7B parameters). These models are vital for practical deployment due to their efficiency and resource-friendliness, and their current underperformance on our benchmarks underscores the need for continued research and development in this area.

Through these experiences, we observed that several factors, such as CoT prompting and few-

shot inference, could affect the evaluation results.

**CoT Prompting** For the Vi-MQA dataset, using CoT prompt resulted in improved performance, particularly for larger models with over 14 billion parameters. Notably, the Llama-3.3-70B-Instruct model showed the most significant improvement, increasing its performance from 69.1% to 73.8% (+4.7%). Upon analyzing the Vi-MQA results by category, we found that the increase in performance primarily came from STEM subjects, which require more reasoning steps. Detailed breakdowns of the results can be found in Table 18. In contrast, smaller models saw a decline in performance; for example, the Qwen2.5-0.5B-Instruct model only improved by 12.8%. Regarding the Vi-DROP dataset, we observed a substantial improvement in results with CoT prompting compared to no-CoT prompting, except for some models that were too small. Notably, Qwen2.5-3B-Instruct experienced the highest increase proportion, from 49.1% to 72.4% (+23.3%). Other models also displayed substantial increases, with around 10%. The results indicate that CoT prompting effectively enhances benchmark performance, particularly for tasks that require the models' reasoning abilities.

**Few-shot inference** In the experiment 5-shot on the Vi-MQA dataset, the Qwen2.5 series demonstrates consistent improvement over the zero-shot baseline. Specifically, Qwen2.5-3B-Instruct achieves the greatest improvement to 54.7%, which is an increase of 1.8% compared to its zero-shot performance. In contrast, the Llama3 model shows some fluctuations in its results. For instance, Llama-3.2-1B-Instruct increases from 37.6% to 38.2% (+0.6%), while the performance of Llama-3.1-8B-Instruct declines from 54.1% to 52.6% (-1.5%). Overall, these results suggest that 5-shot prompting does not offer advantages across the Llama-3 series, but it positively impacts the performance of Qwen2.5 models on the Vi-MQA dataset.

## 5 Conclusion and Future works

The VMLU Benchmarks introduced in this study provide a comprehensive evaluation framework for assessing the capabilities of Vietnamese large language models (LLMs). By covering four critical tasks, this benchmark effectively measures LLMs' performance across domains such as general knowledge, reading comprehension, reasoning, and conversational abilities. Our evaluation of state-of-the-art models, including GPT-4o, Llama-3, and Qwen2.5, highlights the strengths and limitations of these models in the Vietnamese context. The results emphasize the importance of prompt design in influencing model performance, offering valuable insights for future improvements. We are confident that our benchmark dataset and analytical insights will empower researchers to evaluate and design Vietnamese LLMs effectively.

## Limitations

Despite the comprehensive nature of the VMLU Benchmarks in evaluating Vietnamese LLMs across multiple dimensions, several limitations persist. First, the benchmark primarily focuses on four task types and does not include other essential tasks, such as summarization or instruction following, which could provide a more complete assessment. Second, while practical, the reliance on LLMs for evaluation may introduce biases and inconsistencies when compared to human evaluation methods. Lastly, assessing Vi-Dialog remains a challenge, as current methods struggle to accurately capture conversational quality, coherence, and understanding of context in a reliable and scalable way.

## Ethics Statement

All data in the VMLU benchmarks has been sourced from public resources that allow for redistribution. Additionally, all test instances in the VMLU benchmarks have undergone a thorough review to ensure the exclusion of any examples that raise ethical concerns.

## References

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 7421–7454. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, et al. 2024. The llama 3 herd of models.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Qwen, :, 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, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Minh-Nam Tran, Phu-Vinh Nguyen, Long Nguyen, and Dien Dinh. 2024. ViGLUE: A Vietnamese general language understanding benchmark and analysis of Vietnamese language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4174–4189, Mexico City, Mexico. 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.

Shuyi Xie, Wenlin Yao, Yong Dai, Shaobo Wang, Donlin Zhou, Lifeng Jin, Xinhua Feng, Pengzhi Wei, Yujie Lin, Zhichao Hu, Dong Yu, Zhengyou Zhang, Jing Nie, and Yuhong Liu. 2023. Tencentllmeval: A hierarchical evaluation of real-world capabilities for human-aligned llms.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Senel, Anna Korhonen, and Hinrich Schuetze. 2024. TurkishMMLU: Measuring massive multitask language understanding in Turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7035–7055, Miami, Florida, USA. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models.

# A Dataset Details

## A.1 Vi-MQA Examples

| Level | Subject | Question |
|---|---|---|
| Elementary School | Chemistry | Tính chất nào sau đây không phải là tính chất của thủy tinh chất lượng cao:<br>A. Rất trong<br>B. Bền, khó vỡ<br>C. Chịu được nóng, lạnh<br>D. Dễ cháy<br>Đáp án: D |
| Middle School | Geography | Việc phát triển nông-lâm-thủy sản tạo cơ sở nguyên liệu cho ngành phát triển công nghiệp nào?<br>A. Công nghiệp năng lượng<br>B. Công nghiệp chế biến lương thực thực phẩm<br>C. Công nghiệp hóa chất<br>D. Công nghiệp sản xuất vật liệu xây dựng<br>Đáp án: B |
| High School | History | Sự kiện nào sau đây đã tạo ra một cơ chế giải quyết các vấn đề liên quan đến hòa bình và an ninh ở châu Âu?<br>A. Định ước Henxinki (08/1975)<br>B. Liên Xô và Mỹ ký Hiệp định hạn chế vũ khí tiến công chiến lược<br>C. Mỹ và Liên Xô tuyên bố chấm dứt Chiến tranh lạnh<br>D. Hiệp định về những cơ sở của quan hệ giữa Đông Đức và Tây Đức<br>Đáp án: A |
| University | Clinical Pharmacology | Khái niệm DƯỢC LỰC HỌC:<br><br>A. Động học của sự hấp thu, phân phối, chuyển hóa và thải trừ thuốc<br>B. Nghiên cứu tác động của thuốc trên cơ thể sống<br>C. Nghiên cứu về tác động của cơ thể đến thuốc<br>D. Là môn khoa học nghiên cứu về thuốc<br>Đáp án: B |

Table 5: Vi-MQA Examples

## A.2 Vi-MQA Category

Table 6: Statistic Vi-MQA benchmark categories

| Id | Subject | Category | # Questions |
|---|---|---|---|
| 01 | Elementary Mathematics | STEM | 200 |
| 02 | Elementary Science | STEM | 200 |
| 03 | Middle School Biology | STEM | 188 |
| 04 | Middle School Chemistry | STEM | 200 |
| 05 | Middle School Mathematics | STEM | 119 |
| 06 | Middle School Physics | STEM | 200 |
| 07 | High School Biology | STEM | 200 |

Table 6: Statistic Vi-MQA benchmark categories

| Id | Subject | Category | # Questions |
|---|---|---|---|
| 08 | High School Chemistry | STEM | 200 |
| 09 | High School Mathematics | STEM | 163 |
| 10 | High School Physics | STEM | 200 |
| 11 | Applied Informatics | STEM | 200 |
| 12 | Computer Architecture | STEM | 200 |
| 13 | Computer Network | STEM | 197 |
| 14 | Discrete Mathematics | STEM | 182 |
| 15 | Electrical Engineering | STEM | 194 |
| 16 | Introduction to Chemistry | STEM | 197 |
| 17 | Introduction to Physics | STEM | 191 |
| 18 | Introduction to Programming | STEM | 197 |
| 19 | Metrology Engineer | STEM | 155 |
| 20 | Operating System | STEM | 200 |
| 21 | Statistics and Probability | STEM | 192 |
| 22 | Middle School Civil Education | Social Science | 196 |
| 23 | Middle School Geography | Social Science | 162 |
| 24 | High School Civil Education | Social Science | 200 |
| 25 | High School Geography | Social Science | 179 |
| 26 | Business Administration | Social Science | 192 |
| 27 | Ho Chi Minh Ideology | Social Science | 197 |
| 28 | Macroeconomics | Social Science | 200 |
| 29 | Microeconomics | Social Science | 200 |
| 30 | Principles of Marxism and Leninism | Social Science | 200 |
| 31 | Sociology | Social Science | 196 |
| 32 | Elementary History | Humanity | 195 |
| 33 | Middle School History | Humanity | 200 |
| 34 | Middle School Literature | Humanity | 192 |
| 35 | High School History | Humanity | 200 |
| 36 | High School Literature | Humanity | 200 |
| 37 | Administrative Law | Humanity | 100 |
| 38 | Business Law | Humanity | 197 |
| 39 | Civil Law | Humanity | 200 |
| 40 | Criminal Law | Humanity | 180 |
| 41 | Economic Law | Humanity | 178 |
| 42 | Education Law | Humanity | 183 |
| 43 | History of World Civilization | Humanity | 200 |
| 44 | Idealogical and Moral Cultivation | Humanity | 200 |
| 45 | Introduction to Laws | Humanity | 139 |
| 46 | Introduction to Vietnam Culture | Humanity | 200 |
| 47 | Logic | Humanity | 192 |
| 48 | Revolutionary Policy of the Vietnamese Communist Party | Humanity | 200 |
| 49 | Vietnamese Language and Literature | Humanity | 192 |
| 50 | Accountant | Other | 186 |
| 51 | Clinical Pharmacology | Other | 200 |
| 52 | Environmental Engineering | Other | 189 |
| 53 | Internal Basic Medicine | Other | 189 |
| 54 | Preschool Pedagogy | Other | 112 |
| 55 | Tax Accountant | Other | 192 |

Table 6: Statistic Vi-MQA benchmark categories

| Id | Subject | Category | # Questions |
|---|---|---|---|
| 56 | Tax Civil Servant | Other | 189 |
| 57 | Civil Servant | Other | 189 |
| 58 | Driving License Certificate | Other | 189 |

### A.3 Vi-SQuAD Examples

Table 7: "Examples of Answerable and Unanswerable Questions in the Vi-SQuAD Dataset"

**Category: Answerable question**

**Passage:**
VinFast VF 6 là mẫu xe ô tô thông minh chạy động cơ điện phân khúc D cỡ nhỏ được phát triển, giới thiệu năm 2022, phân phối ra thị trường năm 2023 bởi VinFast, thành viên của Tập đoàn Vingroup. Cuối tháng 9, VinFast chính thức giới thiệu bản thương mại của chiếc Suv/ Cuv hạng B VF 6 với giá niêm yết 675 triệu cho bản Eco pin thuê; giá mua pin và giá chênh cho bản Plus đều là +90 triệu, có dải đèn LED tạo hình cánh chim đặc trưng của VinFast, với 5 lựa chọn màu ngoại thất và 2 màu nội thất. VF 6 cũng áp dụng bảo hành chính hãng 7 năm hoặc 160.000 km (tùy theo điều kiện nào đến trước); chính sách hỗ trợ đặc biệt cho các sự cố phát sinh do lỗi của nhà sản xuất, gây bất tiện cho người dùng; hay cam kết giá mua lại ô tô điện đã qua sử dụng sau 5 năm.
**Question:**
Có bao nhiêu lựa chọn màu ngoại thất cho mẫu xe VF6?

**Possible answers:**
{'5', '5 lựa chọn màu ngoại thất', 'với 5 lựa chọn màu ngoại thật', 'Có 5 lựa chọn màu ngoại thất cho mẫu xe VF 6', '5 lựa chọn' }

**Category: Unanswerable Question**

**Passage:**
Du lịch Paris là một trong những ngành kinh tế quan trọng không chỉ của thành phố Paris mà còn cả nước Pháp vì Paris được mệnh danh là trung tâm châu Âu và cũng là niềm tự hào của Pháp. Với vị trí địa lý, trung tâm chính trị, kinh tế, văn hóa đã giúp Paris trở thành một điểm đến hấp dẫn từ rất lâu trong lịch sử. Cuối thế kỷ 19 và đầu thế kỷ 20, thành phố đã nhiều lần tổ chức các triển lãm thế giới, đánh dấu cho việc ngành du lịch bắt đầu trở nên quan trọng đối với nền kinh tế thành phố. Trong thời kỳ phồn vinh này, Paris cũng đã xây dựng thêm nhiều công trình, khách sạn, cửa hàng... góp phần cho sự phát triển của du lịch thành phố ngày nay.
Đón khoảng 30 triệu du khách mỗi năm, Paris là một trong những điểm đến thu hút nhất. Bên cạnh du lịch giải trí, thành phố còn là địa điểm thường xuyên của các hội nghị, cũng là nơi tổ chức nhiều hội chợ, triển lãm quan trọng. Những công trình kiến trúc nổi tiếng, các bảo tàng với những hiện vật giá trị, các khu phố in đậm dấu ấn lịch sử, văn hóa, những trung tâm mua sắm... tất cả đã khiến du khách không ngừng tìm đến với "kinh đô ánh sáng". Những công trình, địa điểm vùng ngoại ô cùng góp phần làm Paris thêm phần hấp dẫn.
Ngành du lịch thành phố hiện nay cũng phải đối mặt với sự canh tranh từ nhiều đô thị lớn khác, đặc biệt là London và Roma. Nhiều khách du lịch đánh giá Paris là một thành phố đắt đỏ và kém hiếu khách. Mặc dù vậy, trong một cuộc điều tra của Văn phòng du lịch Paris vào mùa hè năm 2008, hầu như tất cả các du khách được hỏi đều cho biết họ sẽ quay lại thành phố này trong tương lai.
**Question:**
Paris có những công trình kiến trúc nổi tiếng nào?

**Possible answers:**
{'NO ANSWER'}

### A.4 Vi-DROP Examples

Table 8: Six categories of Vi-DROP and examples

**Category: Add_Sub**

**Passage:**
Anh phát hiện thêm 5.296 ca nhiễm nCoV, nâng tổng số lên 166.441. Nước này ghi nhận 26.097 ca tử vong, tăng 4.419 trường hợp so với một ngày trước, tương đương 17%, sau khi đưa số người chết trong viện dưỡng lão và những nơi khác vào thống kê. (vnexpress). Đức báo cáo thêm 1.462 ca nhiễm và 125 ca tử vong do nCoV, nâng ca nhiễm và tử vong lên lần lượt 161.197 và 6.405.(vnexpress). Nga báo cáo thêm 5.841 ca nhiễm và 105 trường hợp tử vong, nâng số ca nhiễm và tử vong cả nước lên lần lượt 99.399 và 972. (vnexpress)

**Question:**
Tính tổng số ca nhiễm mới được báo cáo trong một ngày từ ba quốc gia này.

**Possible answers:**
["12599 ca", "12599 ca nhiễm"]

**Category: Selection**

**Passage:**
Thông tin nhân khẩu Có 287.012 hộ, trong đó 29,60% có trẻ em dưới 18 tuổi sống chung với họ, 45,20% là đôi vợ chồng sống với nhau, 14,70% có nữ hộ và không có chồng, và 36,20% là không lập gia đình. 30,50% hộ gia đình đã được tạo ra từ các cá nhân và 10,30% có người sống một mình 65 tuổi hoặc lớn tuổi hơn. Cỡ hộ trung bình là 2,37 và cỡ gia đình trung bình là 2,97. Trong dân số quận đã được trải ra với 24,30% dưới độ tuổi 18, 8,90% 18-24, 30,40% 25-44, 22,80% từ 45 đến 64, và 13,50% từ 65 tuổi trở lên người. Độ tuổi trung bình là 37 năm. Đối với mỗi 100 nữ có 91,60 nam giới. Đối với mỗi 100 nữ 18 tuổi trở lên, đã có 87,60 nam giới.

**Question:**
Có bao nhiêu phần trăm dân số từ 25- 44 tuổi?

**Possible answers:**
[“30,40%”, “30.40”,” “30,4 phần trăm”, “Có 30,4 phần trăm dân số từ 25-44 tuổi”]

**Category: Comparative**

**Passage:**
Theo Thống kê Dân số Hoa Kỳ năm 2000, quận có dân số 845.303 người, 346.790 gia hộ, và 212.582 gia đình sống trong quận hạt. Mật độ dân số là 801 người/km$^2$ (2.075 người/mi$^2$). Có 373.393 đơn vị nhà ở với mật độ trung bình 354 nhà/km$^2$ (917 nhà/mi$^2$). Trong quận này, 72,93% dân số là người da trắng, 23,43% là người da đen hay Mỹ gốc Phi, 0,18% là người Mỹ bản thổ, 1,61% là người gốc Á, 0,03% là người gốc Thái Bình Dương, 0,51% từ các chủng tộc khác, và 1,32% từ hai hoặc nhiều chủng tộc. 1,13% dân số là người Hispanic hay Latino thuộc một chủng tộc nào. (Xem Chủng tộc và dân tộc trong Thống kê Dân số Hoa Kỳ.)

**Question:**
Số lượng đơn vị nhà ở với mật độ trung bình 354 nhà/km$^2$ cao hay thấp hơn 400.000?

**Possible answers:**
[“thấp hơn”, “Thấp hơn”, “400.000”, “thấp hơn”, “400.000”]

**Category: Superlative**

**Passage:**
Anh phát hiện thêm 5.296 ca nhiễm nCoV, nâng tổng số lên 166.441. Nước này ghi nhận 26.097 ca tử vong, tăng 4.419 trường hợp so với một ngày trước, tương đương 17%, sau khi đưa số người chết trong viện dưỡng lão và những nơi khác vào thống kê. (vnexpress) Đức báo cáo thêm 1.462 ca nhiễm và 125 ca tử vong do nCoV, nâng ca nhiễm và tử vong lên lần lượt 161.197 và 6.405.(vnex- press) Nga báo cáo thêm 5.841 ca nhiễm và 105 trường hợp tử vong, nâng số ca nhiễm và tử vong cả nước lên lần lượt 99.399 và 972. (vnexpress)

**Question:**

Quốc gia nào có tỉ lệ tăng số ca tử vong cao nhất trong ngày?

**Possible answers:**
["Anh"]

**Category: Counting**

**Passage:**
Đơn vị hành chính 02 thôn tên: Thọ Sơn, 5 buôn. Buôn Krông thành lập năm 1992, diện tích tự nhiên 400,3 ha, dân số 91 hộ 421 khẩu, có 2 dân tộc anh em sinh sống Êđê và Tày. Sản xuất nông nghiệp lúa nước và ngô lai là 2 nguồn thu nhập chính của Buôn. Buôn Krang diện tích tự nhiên 2.270,04 ha, dân số 194 hộ 933 khẩu, chủ yếu đồng bào người Êđê, sản xuất lương thực chủ yếu cây lương thực ngắn ngày và cà phê là nguồn thu nhập chính. Buôn Kmăl diện tích tự nhiên 1.175 ha, dân số 214 hộ 1.106 khẩu, chủ yếu đồng bào người Êđê.

**Question:**
Có bao nhiêu dân tộc được nhắc đến trong đoạn văn?

**Possible answers:**
["2 dân tộc", "2"]

**Category: Other Arithmetic**

**Passage:**
Năm 2020, dân số toàn huyện Hồng Dân là 112.548 người, trong đó, dân số thành thị là 11.537 người chiếm 10,25%, dân số nông thôn là 101.011 người chiếm 89,75%. Theo thống kê ngày 1 tháng 11 năm 2021, dân số huyện Hồng Dân là 113.351 người, trong đó: dân số thành thị là 11.616 (10,25%), dân số nông thôn là 101.735 người (89,75%). Tổng dân số năm 2020 toàn huyện là 112.548 người, chiếm 12,32% dân số toàn tỉnh, mật độ dân số 265 người/km$^2$. Xã có dân số cao nhất là Ninh Hoà với 19.255 người (chiếm 17,11% dân số toàn huyện), Ninh Quới A 16.041 người (chiếm 14,25%), thấp nhất là xã Ninh Thạnh Lợi A 9.671 người (chiếm 8,59%).

**Question:**
Phần trăm dân số thành thị chênh lệch so với dân số nông thôn ở huyện Hồng Dân năm 2021 là bao nhiêu phần trăm?

**Possible answers:**
["79,5%"]

### A.5 Vi-Dialog Examples

| **Sample 1 – Dialog requires information shared in the earlier conversation.** |
|---|
| Question: Tôi muốn chơi trò chơi điện tử ở nhà với bạn bè vào cuối tuần. Bạn có đề xuất nào về trò chơi phù hợp để chơi nhiều người không? |
| Question: Trò chơi thứ nhất bạn vừa đề cập được sản xuất vào năm nào và tên công ty sản xuất là gì? |
| Question: Đối với trò chơi thứ hai mà bạn đề cập, công ty sản xuất có trụ sở chính ở quốc gia nào? |
| Question: Công ty này còn có những trò chơi nổi tiếng nào khác? |
| Question: Super Mario Bros và Mario Kart, sự khác biệt và mối liên hệ giữa hai trò chơi này là gì? |
| **Sample 2 – Dialog requires reasoning ability** |
| Question: Doanh thu phòng vé của Lật Mặt 2 đạt bao nhiêu? |
| Question: Doanh thu phòng vé cao hơn hay thấp hơn Lật Mặt? |
| Question: Tại sao? |
| Question: Bạn có thể giới thiệu thêm cho tôi một số phim theo thể loại này được không? |

Table 9: Vi-Dialog Example

# B Prompt Inference

## B.1 Vi-MQA

### B.1.1 Prompt inference Vi-MQA non-CoT, 0-shot

```
Trả lời câu hỏi trắc nghiệm.
Câu hỏi: {question}
Các lựa chọn:
{context}

Đáp án đúng là:
```

Table 10: Prompt Inference Vi-MQA Zero Shot

### B.1.2 Prompt inference Vi-MQA CoT, 0-shot

```
#TASK: Trả lời câu hỏi trắc nghiệm bằng cách suy nghĩ từng bước (step by step) sau đó đưa ra đáp án.
# Instruction:
- Phải chọn đúng 1 đáp án duy nhất: A hoặc B, C, D, E.
- Dòng cuối cùng bắt buộc trả lời theo định dạng Đáp án đúng là: A/B/C/D/E

# Output format:
Output:
Bước 1: ...
...
Bước 2: ...
...
Đáp án đúng là: A/B/C/D/E
# Input:
Câu hỏi: question

Các lựa chọn:
{context}

# Output:
```

Table 11: Prompt Inference Vi-MQA CoT

### B.2 Vi-SQuAD

Table 12: Prompt Inference Vi-SQuAD

Hãy trả lời câu hỏi dựa vào nội dung đoạn văn / văn bản. Yêu cầu:
- Câu trả lời phải rất ngắn gọn được trích ra từ văn bản. Chỉ sử dụng thông tin trong văn bản được cung cấp.
- Với những câu hỏi về số liệu, người, thời gian, ... thì câu trả lời có thể là một từ hoặc cụm từ.
- Nếu trong văn bản không có câu trả lời, hoặc nội dung đoạn văn không liên quan thì output 'NO ANSWER'.

Văn bản:
{context}

Câu hỏi: {question}
Câu trả lời:

### B.3 Vi-DROP

#### B.3.1 Prompt inference Vi-DROP CoT

Table 13: Prompt Inference ViDrop (CoT)

# Task:
Dựa vào đoạn văn, hãy trả lời câu hỏi sau bằng cách đưa ra lời giải thích cụ thể từng bước. Nếu là so sánh hơn kém, hơn nhất, hãy đưa ra các biểu thức so sánh mới kết luận.

# Output format:
Bước 1:
...
Bước 2:
...
Kết luận:
...

#### B.3.2 Prompt inference Vi-DROP no-CoT

#Task:
Dựa vào đoạn văn, hãy trả lời câu hỏi sau. Chỉ cung cấp câu trả lời ngắn gọn không giải thích gì thêm.
# Input:
Đoạn văn:
====
{context}
====
Câu hỏi: question
#Output:

Table 14: Prompt Inference Vi-DROP Without CoT

# C Prompt Evaluate

## C.1 Vi-SQuAD and Vi-DROP.

Table 15: Prompt evaluation Vi-SQuAD and Vi-DROP

Dưới đây là thông tin trả lời của ứng viên (candidate) và danh sách đáp án chấp nhận của chuyên gia (expert) cho tác vụ hỏi đáp.
=====
<question>
{QUESTION}
</question>

<expert>
{GOLDS}
</expert>

<candidate>
{ANSWER}
</candidate>
=====
Hãy đánh giá độ chính xác của đáp án của ứng viên bằng so sánh với một trong những đáp án đúng của chuyên gia. Nếu câu trả lời có ý nghĩa giống với với một trong những đáp án thì output Đúng, ngược lại output Sai.
- Chỉ ouput Đúng/Sai

OUTPUT:

## C.2 Vi-Dialog

Table 16: Prompt evaluation Vi-Dialog using GPT-4o

[Dialog]
{dialog}

[Evaluation Criteria]
The following are 7 evaluation rules, with decreasing priority:
a. Safety: No pornography, no politics, no violence or illegal content
b. Neutrality: No prejudice, racial discrimination, or subjective bias
c. Compliance with facts: Not against the truth, common sense
d. Relevance: The content of the answer matches the user's question
e. Logicality: No contradictions, coherence
f. Language fluency: Clear description, no typos, no grammar errors, and understandable
g. Information content: No omission of key points, reasoning process for arithmetic problems, irrelevant content will be deducted points

Note:
- Answers that violate rules a/b/c/d will be scored between 1 and 3 points, which are low scores.
- Answers that are correct but violate rules e/f/g will be scored between 4-7, which are medium scores.
- Only answers that are correct and meet the above 7 evaluation criteria can score 8 points or more, which are high scores.
[Evaluation Criteria End]

[Output Rules]
Please strictly follow the requirements below:
- The first line outputs a paragraph of text, explaining the detailed reasons for scoring the answer.
- The second line outputs a number, representing the assistant's score. Please strictly rate the model's answer according to the scoring range of 1 to 10, and the number can only be a positive integer between 1 and 10, such as output: 5, decimals such as 5.5 cannot appear Please strictly output the above two lines of content in accordance with the above regulations, separated by a single newline character between each line.
- The third line produce a list of errors, which present whether the dialogue violate the rules.
The expected output format should be:
Feedback: reason for the score
Score: your score for the model
Errors:
- Safety: Yes/No
- Neutrality: Yes/No
- Compliance with facts: Yes/No
- Relevance: Yes/No
- Logicality: Yes/No
- Language fluency: Yes/No
- Information content: Yes/No
[Output Rules End]

Please output your judgment by Vietnamese language

## D Comparison of LLMs for LLM-as-a-Judge

| | **Evaluator** | | | | | |
|---|---|---|---|---|---|---|
| **Model** | **Qwen** | ***Rank*** | **GPT-4o** | ***Rank*** | **GPT-4o-mini** | ***Rank*** |
| GPT-4o | 96,3 | 2 | 96,1 | 2 | 96,5 | 2 |
| Llama-3.2-1B-Instruct | 70,1 | 12 | 67,9 | 12 | 71,5 | 12 |
| Llama-3.2-3B-Instruct | 90,3 | 9 | 89,4 | 9 | 91,0 | 9 |
| Llama-3-8B-Instruct | 93,1 | 6 | 92,6 | 6 | 93,6 | 6 |
| Llama-3.1-8B-Instruct | 92,0 | 7 | 91,3 | 7 | 93,6 | 7 |
| Llama-3.1-70B-Instruct | 95,1 | 5 | 94,3 | 5 | 96,1 | 5 |
| Llama-3.3-70B-Instruct | 95,8 | 4 | 95,5 | 3 | 96,4 | 3 |
| Qwen2.5-0.5B-Instruct | 62,5 | 13 | 60,3 | 13 | 64,5 | 13 |
| Qwen2.5-1.5B-Instruct | 86,7 | 11 | 85,6 | 11 | 87,9 | 11 |
| Qwen2.5-3B-Instruct | 88,3 | 10 | 86,9 | 10 | 89,7 | 10 |
| Qwen2.5-7B-Instruct | 91,8 | 8 | 90,7 | 8 | 93,0 | 8 |
| Qwen2.5-14B-Instruct | 96,1 | 3 | 95,5 | 3 | 96,2 | 4 |
| Qwen2.5-72B-Instruct | 96,8 | 1 | 96,9 | 1 | 97,2 | 1 |

Table 17: Evaluation of the Vi-SQuAD Task Using Different LLMs as Evaluator (Qwen: Qwen2.5-72B, GPT-4o: gpt-4o-2024-08-06 , GPT-4o-Mini: gpt-4o-mini-2024-07-18

## E Breakdown comparison between CoT and Non-CoT prompting techniques in Vi-MQA

| | Qwen2.5-72B-Instruct | | | Llama-3.3-70B-Instruct | | |
|---|---|---|---|---|---|---|
| **subject** | **CoT** | **No CoT** | **Gap** | **CoT** | **No CoT** | **Gap** |
| **TOTAL** | 74,2 | 72,4 | +1,8 | 73,9 | 69,1 | +4,8 |
| **By category** | | | | | | |
| **STEM** | **76,2** | **71,7** | **+4,5** | **75,2** | **65,6** | **+9,5** |
| HUMANITY | 72,9 | 72,4 | +0,5 | 72,7 | 70,8 | +2,0 |
| SOCIAL SCIENCE | 78,5 | 77,7 | +0,7 | 77,2 | 75,9 | +1,3 |
| OTHER | 67,1 | 68,2 | -1,1 | 69,8 | 66,5 | +3,3 |
| **STEM's subjects** | | | | | | |
| Statistics and probability | 86,2 | 64,9 | +21,3 | 81,6 | 51,2 | +30,5 |
| Computer architecture | 86,7 | 73,9 | +12,8 | 81,7 | 67,8 | +13,9 |
| Middle school physics | 85,6 | 75,0 | +10,6 | 77,2 | 66,1 | +11,1 |
| Middle school mathematics | 69,4 | 59,3 | +10,2 | 70,4 | 53,7 | +16,7 |
| Introduction to physics | 77,5 | 67,6 | +9,8 | 73,4 | 59,0 | +14,5 |
| High school physics | 71,7 | 63,9 | +7,8 | 75,0 | 56,7 | +18,3 |
| Elementary mathematics | 90,6 | 84,4 | +6,1 | 85,6 | 68,3 | +17,2 |
| High school mathematics | 71,0 | 68,2 | +2,7 | 66,2 | 52,7 | +13,5 |
| Computer network | 83,2 | 81,0 | +2,2 | 81,6 | 70,4 | +11,2 |

Table 18: Breakdown Comparison between CoT and No-CoT Prompting techniques on Vi-MQA task of Qwen2.5-72B-Instruct and Meta-Llama-3.3-70B-Instruct

## F Breakdown of the top category evaluation result in Vi-DROP benchmark

| | Category | Number | GPT4o | Qwen2.5-72B | Llama-3.3-70B |
|---|---|---|---|---|---|
| 1 | add_sub | 830 | 90.0 | 91.5 | 91.7 |
| 2 | selection | 682 | 94.4 | 94.3 | 95.6 |
| 3 | comparison | 559 | 95.9 | 96.1 | 95.4 |
| 4 | count | 430 | **76.9** | **77.4** | **75.1** |
| 5 | comparison1 | 331 | 92.2 | 91.2 | 92.2 |

Table 19: Breakdown of the top category evaluation result in Vi-DROP benchmark