Crossing Linguistic Horizons: Finetuning and Comprehensive Evaluation of Vietnamese Large Language Models

Sang T. Truong[§*] **Duc Q. Nguyen**[†*] **Toan Nguyen**[†*] **Dong D. Le**[†*] **Nhi N. Truong**[§†*]
Tho Quan[†] **Sanmi Koyejo**[§]
[§]Stanford University [†]Ho Chi Minh City University of Technology, VNU-HCM
[*]Equal contribution, Corresponding: nqduc@hcmut.edu.vn, sttruong@cs.stanford.edu

Abstract

Recent advancements in large language models (LLMs) have underscored their importance in the evolution of artificial intelligence. However, despite extensive pretraining on multilingual datasets, available open-sourced LLMs exhibit limited effectiveness in processing Vietnamese. The challenge is exacerbated by the absence of systematic benchmark datasets and metrics tailored for Vietnamese LLM evaluation. To mitigate these issues, we have finetuned LLMs specifically for Vietnamese and developed a comprehensive evaluation framework encompassing 10 tasks and 31 metrics. We observe that finetuning can help LLMs transfer knowledge across languages, serving as an efficient way to bolster their capabilities in non-English languages. Moreover, our analysis indicates that larger models can introduce more biases and uncalibrated outputs and the key factor influencing LLM performance is the quality of the training or finetuning datasets. These insights underscore the significance of meticulous finetuning with high-quality datasets in enhancing LLM performance.

1 Introduction

Large language models (LLMs) such as GPT-4 (OpenAI, 2023), BLOOM (Le Scao et al, 2023), LLaMa-2 (Touvron et al, 2023), Mistral (Jiang et al., 2023), Mixtral (Jiang et al., 2024), Gemma (Team et al., 2024) have made significant contributions to the field of natural language processing (NLP). Despite their advancements, a gap remains in their specialization for many languages, including Vietnamese. This paper addresses the development and evaluation of Vietnamese-centric LLMs. Vietnam, with a population surpassing 100 million, ranks as the 16th most populous country globally. Current models exhibit limitations in effectively handling Vietnamese NLP tasks, especially in accurate comprehension and response (Lai et al., 2023). Consequently, there is an increasing demand for a robust, dedicated Vietnamese LLM.

Several factors constrain the practical application of LLMs. Concerns regarding the precision, inherent biases, potential toxicity, and fairness of their outputs are notable obstacles (Ye et al., 2023; Liang et al, 2023; Wang et al., 2024). Moreover, there is a lack of research evaluating LLMs in the Vietnamese context. To facilitate the effective use of state-of-the-art LLMs for Vietnamese speakers, thorough evaluations are essential prior to their widespread use. Such evaluations not only ensure the reliability of these LLMs but also highlight areas where these LLMs could be better. This leads to developing targeted reinforcement learning strategies to rectify these issues in the next phase.

In response to the aforementioned challenges, we aim to develop open-source Vietnamese LLMs. Initiating an LLM from scratch is impractical due to the scarcity of extensive training datasets and limited computational resources. However, the advent of QLoRA (Dettmers et al., 2023), incorporating quantization techniques (Dettmers et al., 2022) and LoRA (Hu et al., 2022), provides an efficient approach for fine-tuning LLMs, particularly in resource-constrained environments. We employ fine-tuning on the LLaMa-2, Mixtral 8$\times$7B, Gemma, and conduct a comprehensive evaluation of Vietnamese LLMs across various scenarios and settings. Throughout the thorough evaluation process, we observe the following: (i) larger language models exhibit unseen capabilities compared to smaller counterparts; (ii) larger language models tend to manifest more biases, produce uncalibrated results, and are more susceptible to the influence of input prompts; (iii) the quality of training or fine-tuning datasets is the key for unlocking LLM performance. Our key contributions include:

- The fine-tuning and release of five Vietnamese LLMs: URA-LLaMa 7B, 13B, and 70B based on LLaMa-2; MixSUra based on Mix-

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2849–2900

tral 8×7B; GemSUra 7B based on Gemma 7B. Our finetuning leverages data from the Vietnamese Wikipedia (Foundation, 2022), Vietnamese News-Corpus (Binh, 2021), and Vietnamese Highschool Essays[1].

- Conducting comprehensive evaluations of 14 Vietnamese LLMs across ten common application scenarios, focusing on aspects such as accuracy, robustness, fairness, bias, and toxicity. Additional criteria are tailored to each specific scenario. Our empirical research also explores the influence of prompt design during inference.
- As part of this effort, we introduce and share two novel Vietnamese reasoning datasets inspired by MATH (Hendrycks et al., 2021) and Synthetic reasoning (Wu et al., 2021).

2 Related Works

Vietnamese LLMs To our best knowledge, there are seven available Vietnamese LLMs: (i) Vietcuna-7B-v3 (ViLM, 2023) – fine-tuned on BLOOMZ (Muennighoff et al., 2023), open-sourced, released on Aug. 8, 2023, (ii) Vistral [2] – based on Mistral, open-sourced, (iii-iv) PhoGPT 7B5 & PhoGPT 7B5 Instruct (Nguyen et al., 2023a) – based on MPT architecture (Team, 2023), open-sourced, released on Nov. 7, 2023 (concurrently with our work), (v) Gemini (Team et al., 2024) – a commercial product of Google, and (vi-vii) GPT3.5 Turbo & GPT-4, which are closed-source commercial products on the Azure platform (version 0613) (OpenAI, 2023). To our knowledge, we are the first to fine-tune and release two large-scale open-source Vietnamese LLMs with 13B, 70B parameters and a Mixture-of-Expert Vietnamese LLMs with 47B parameters.

Comprehensive Evaluation of Vietnamese LLMs Evaluating a language model is challenging because LLMs can improve general capabilities with scale. Thus, evaluating an LLM depends on various factors, such as the tasks for which the LLM will be used, and the impact of prompt design, among others. Currently, there is no evaluation framework capable of fully and accurately assessing the abilities of a Vietnamese LLM. Some recent studies on Vietnamese LLMs only assess the model's performance on closed-book question-answering tasks (Nguyen et al., 2023a) or specific datasets related to ad hoc aspects, such as law (Nguyen et al., 2023b; Anh et al., 2023), physics (Xuan-Quy et al., 2023), and biology (Dao and Le, 2023). Part of the challenge is the lack of high-quality Vietnamese datasets. Vietnamese NLP datasets have largely focused on daily tasks such as open-book and closed-book question-answering (Artetxe et al., 2020; Lewis et al., 2020), summarization (Nguyen et al., 2019c; Ladhak et al., 2020), translation (Zhang et al., 2020; Doan et al., 2021), etc. Evaluation of some LLM capabilities, such as reasoning and mathematical logic, have not been considered due to the absence of suitable datasets. We are the first to address this challenge by comprehensively evaluating Vietnamese LLM on 10 scenarios and 31 metrics. In that process, we build and open-source two novel Vietnamese reasoning datasets. Our evaluation framework is open-source on Github[3] to facilitate community-driven model evaluation[4].

3 Experiments

3.1 Supervised Finetuning

We focus on finetuning English-language models to enhance overall performance and evaluate adaptability and efficiency in various configurations. Due to computational constraints, our first models, named URA-LLaMa, were finetuned from LLaMa-2 using QLoRA (Dettmers et al., 2023) on two primary open-source Vietnamese datasets, including Vietnamese Wikipedia (1GB) and Vietnamese News-Corpus (22GB). The 7B variant was finetuned on both datasets, while the 13B and 70B versions were finetuned with only the Vietnamese Wikipedia dataset. The LoRA rank was set at 128 for the 7B model, 256 for the 13B model, and 1024 for the 70B model. Other hyperparameters, including LoRA α, dropout, quantization, quantization type, learning rate, max length, and epochs, were uniformly set at 16, 0.1, 4 bit, NF4, 1×10^{-5}, 2048, and 1, respectively. We use six A100 80GB for the entire finetuning process in approximately 867 hours, emitting nearly 900 kg CO_2 eq.

Continuously, we conducted finetuning on Gemma 7B, and Mixtral 8×7B models utilizing Vietnamese Wikipedia and Vietnamese Highschool Essay datasets, employing the LoRA (Hu et al., 2022). This refinement resulted in the development of GemSUra 7B, and MixSUra models. Common

[1] Vietnamese Highschool Essays
[2] Vistral-7B-Chat
[3] https://github.com/stair-lab/villm
[4] https://ai.stanford.edu/~sttruong/villm

hyperparameters were applied across these models, with LoRA rank set to 256, LoRA α at 512, and LoRA dropout rate fixed at 0.1. For the GemSUra model, the learning rate, maximum sequence length, and number of epochs were established at 1×10^{-5}, 8192, and 2, respectively. Conversely, for MixSUra, these hyperparameters were adjusted to 5×10^{-5}, 32768, and 5. The finetuning process for these two models required four A100 80GB GPUs, spanning a total of 289 hours and resulting in the emission of 200 kg CO_2 equivalent. Our models are available on HuggingFace[5].

3.2 Evaluation Pipeline

We define a scenario as a real-world use case of LLMs describing the purpose for which LLMs are used. Modern LLMs can deal with various scenarios. We limit ten common use cases in Vietnamese in this work. Each scenario contains two well-known datasets in Vietnamese, which are already split into training and testing sets. We utilize the testing set to evaluate our finetuned models, LLaMa-2, Vietcuna, Vistral, PhoGPT, Gemini Pro, GPT-3.5 Turbo, and GPT-4, considering their diverse performance and architectural differences for a comprehensive analysis. Below are detailed descriptions of 10 scenarios:

1. **Question-Answering** requires LLM to answer an open-ended question from a given context. We selected two notable Vietnamese datasets for diversity of evaluation domain: XQuAD (Artetxe et al., 2020), a multilingual variant of SQuAD (Rajpurkar et al., 2016), and MLQA (Lewis et al., 2020), both based on Wikipedia articles. Exact Match (EM) and F1 score (F1) measure question-answering performance. F1 Score is the harmonic mean of Precision and Recall: $\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$ where $\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$ and $\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$.

2. **Summarization** involves LLMs condensing long documents into shorter open-ended paragraphs. We selected the two largest Vietnamese summarization datasets: VietNews (Nguyen et al., 2019c) and WikiLingua (Ladhak et al., 2020). VietNews comprises over 150,000 articles (22,644 for testing) from Vietnamese online news websites. WikiLingua was chosen for its variety, featuring diverse tutorials from WikiHow (wikiHow, 2023). We primarily rely on standard evaluation metrics like ROUGE-1, ROUGE-2, and ROUGE-L (Liang et al, 2023). ROUGE-1 (R1) measures the overlap of unigrams (individual words) between the system-generated and reference summaries. ROUGE-2 (R2) focuses on the overlap of bigrams, while ROUGE-L (RL) evaluates the longest common subsequence between the two summaries. Beyond these, we incorporate five additional metrics from (Grusky et al., 2018) to assess summary quality. These include SummaC (SC), which assesses the faithfulness of generated summaries; BERTScore (BS), which uses mBERT token embeddings to compute the cosine similarity between sentence tokens; Coverage (Cv), measuring how much a summary derives from the original text; Density (De), defined as the average length of extractive fragments associated with each summary word; and Compression (Cp), which is the word ratio between original articles and their summaries.

3. **Sentiment Analysis** focuses on detecting emotion of documents. Given a document and a list of all available sentiments, the LLM must choose the correct ones. The first selected dataset, VLSP 2016 (Nguyen et al., 2019b), contains comments on social networks about electronic devices such as smartphones, laptops, television, etc. The next dataset, UiT-VSFC (Nguyen et al., 2018), is feedback from Vietnamese students about courses at the end of semesters. We use Accuracy (AC), F1, AUC ROC (AR), Expected Calibration Error (ECE), and Accuracy at C% coverage (A@C) for model assessment. $\text{AC} = \frac{\text{True Positive} + \text{True Negative}}{\text{Number of Instances}}$. AUC ROC quantifies the model ability to distinguish between classes by measuring the area under the ROC curve. A perfect model would have an AUC ROC score of 1, while a score below 0.5 indicates a model performing worse than random. Expected calibration error (ECE) described in (Guo et al., 2017) measures the difference between the model predicted probability and the fraction of times the model is correct. As a default configuration, we use ten bins, each containing an equal number of predicted probabilities. Accuracy at C% coverage is the accuracy for the C% fraction of examples the model assigns the highest probability. Details of this metric can be found at (Liang et al, 2023). In

[5] https://huggingface.co/ura-hcmut

our experiment, C is set to 10%.

4. **Text Classification** is a scenario where the LLMs are required to analyze the input document with a list of class labels and give the answer of which class that document belongs to. This scenario is a classical task in almost all languages, including Vietnamese. Thus, various datasets in different fields are available. However, evaluating all those datasets may not be feasible, so we choose two large and reliable ones in this study, which are UiT-VSMEC (Ho et al., 2020) and PhoATIS (Dao et al., 2021). UiT-VSMEC is specified for emotion recognition of Vietnamese comments on Facebook, the most-used social network in Vietnam. PhoATIS is the human-verified Vietnamese version of the famous standard ATIS dataset (Price, 1990), specified for classification intents of user requests about airline information. Here, we use AC, F1, AR, ECE, and A@C for model assessment.

5. **Knowledge** assesses LLMs common knowledge specified for Vietnamese. We use the two largest datasets: ZaloE2E (Zalo AI, 2023) and UiT-ViMMRC (Nguyen et al., 2020b). ZaloE2E has open-ended questions. UiT-ViMMRC contains reading comprehension multiple-choice questions for students from Grades 1-12 in Vietnam. This task uses AC, F1, EM, AR, ECE, and A@C for model assessment.

6. **Toxicity Detection** requires the LLMs to detect toxicity in a paragraph, such as toxic purpose or hate speech. We choose the two most recent datasets: UiT-ViCTSD (Nguyen et al., 2021) and UiT-ViHSD (Luu et al., 2021) in this scenario. The UiT-ViCTSD dataset specifically targets the discernment of toxic speech, while UiT-ViHSD centers on identifying instances of hate speech. In this task, we use accuracy, F1 score, and AUC ROC for model assessment.

7. **Information Retrieval** is a task that ranks a list of relevant documents in the database given the query. We chose the two most recent multilingual datasets supporting Vietnamese. The first is the mMARCO dataset (Bonifacio et al., 2022), a multilingual version of the well-known MS MARCO dataset (Nguyen et al., 2016). The other mRobust04 (Jeronymo et al., 2022) is also a multilingual of TREC Robust 2004. Following (Liang et al, 2023), we have two settings: normal and boosted. In the normal setting, we employ the top 30 documents retrieved by BM25 (Amati, 2009). Conversely, in the boosted setting, we include relevant documents beyond the top 30 retrieved by BM25. Our inquiry tasks an LLM to determine the relevance of each document. Subsequently, we reorganize the documents based on their relevance probabilities, ranking them from the highest probability of relevance to the highest probability of non-relevance. Several metrics are employed to assess model performance. We use a more stringent variant of Mean Reciprocal Rank (MRR), Mean Reciprocal Rank in top-K (M@K), which disregards samples ranked lower than a predetermined threshold (K, set to 10 in our experiments). $\text{M@K} = 1/\text{rank}$ if $\text{rank} \leq \text{K}$ and $\text{M@K} = 0$ otherwise. Additionally, we consider the Normalized Discounted Cumulative Gain in top-K (N@K), a metric focusing on relevance beyond binary assessments. Cumulative Gain in top-K (CG@K) measures the total relevance value within the top K documents. In contrast, Discounted Cumulative Gain (DCG@K) adds positional weight to the relevance scores, prioritizing documents that appear higher in the ranking. DCG@K is computed as $\text{DCG@K} = \sum_{i=1}^{K} \frac{\text{graded\_relevance}(d_i)}{\log_2(i+1)}$. Finally, N@K normalizes DCG@K against the Ideal Discounted Cumulative Gain (IDCG@K), representing the maximum achievable DCG@K score with ideally ordered documents. GPT family and Gemini are not evaluated in this scenario because OpenAI and Google hav disabled probabilities in their response (Azure announcement).

8. **Language Modeling** assesses LLMs' understanding and fluency in a specific language through tasks, notably filling in the blanks and spelling correction. For masked language modeling, we utilized the formal-styled MLQA dataset, masking 10% of words in each document for LLMs to predict. We selected the VSEC dataset (Do et al., 2021) to evaluate spelling correction constructed from news articles with more modification operators than previous datasets. Various metrics are employed for evaluation. Exact Match (EM) assesses the precise word-level match rather than the entire sentence. Character Error Rate (CER) and Word Error Rate (WER) represent the proportion of inaccurately predicted characters

and words compared to references, respectively. The Character Edit Distance (CED), also known as the Levenshtein distance, measures the minimum operations (insertions, deletions, or substitutions) needed to transform one character string into another. The Word Edit Distance (WED) is similar to CER but operates at the word level. Finally, Perplexity (PLX) is defined as the exponentiated average negative log-likelihood of a sequence of T-token: $\text{PLX} = \exp\left(-\frac{1}{T}\sum_{i=0}^{T}\log p_\theta(x_i|x_{<i})\right)$, where $p_\theta(x_i|x_{<i})$ is the probability of the i^{th} token conditioned on preceding ones.

9. **Reasoning** involves evaluating LLMs' logical and mathematical capabilities. Because Vietnamese lacks datasets for reasoning, we adapted two well-known datasets—Synthetic reasoning (Wu et al., 2021) and MATH (Hendrycks et al., 2021)—for this purpose. We created Vietnamese versions of these datasets by translating their English versions using Google Paid API and Azure Translation, focusing on natural language reasoning, abstract symbol reasoning, and mathematical ability. These datasets are compatible with the original license and are open-sourced on HuggingFace[6]. We use EM and F1 to measure the reasoning performance. Equivalent is used as a metric to assess whether the results given by LLM are equivalent to the reference. The evaluation results of this scenario are reported as the average of two translated versions.

10. **Translation** involves translating documents from Vietnamese to English and the reverse while preserving the original meaning. We selected the two most extensive and high-quality datasets: OPUS100 (Zhang et al., 2020) and PhoMT (Doan et al., 2021). Two key metrics are employed to ensure translation accuracy. The Bilingual Evaluation Understudy (BLEU) score (Papineni et al., 2002) measures the similarity of a translation to reference translations, with values closer to 1 indicating higher similarity. On the other hand, the Harmonic mean of Enhanced Length Penalty, Precision, n-gram Position-difference Penalty, and Recall (hLEPOR) (Han et al., 2013) assesses the similarity of n-grams between the translation and references. The hLEPOR score also ranges from 0 to 1, where a higher score signifies a more closely aligned translation with the references.

[6]Synthetic reasoning natural; Synthetic reasoning; MATH

We design a base prompt for each scenario that asks the LLMs to perform the desired task without any examples or constraints. Recent studies (Zhao et al., 2021; Wei et al., 2022) have demonstrated that LLMs perform better if carefully prompted. Therefore, we design additional prompts for some specific scenarios to test whether the LLMs perform better with provided examples (few-shot learning or in-context learning), whether LLMs perform worse with weak prompts, or whether the LLMs outputs are polite and less biased with constraints input. Details of prompts for each scenario are provided in Appendix G.

4 Results and Discussion

We present the overall capacities of evaluated LLMs in Figure 1, separating commercial and open-sourced models across six aspects, including general performance, robustness under weaker prompts, performance with Chain-of-Thought (COT), ability to deal with unfair input (fairness) and toxicity, bias in generated outputs. Each aspect is quantified by the average score of the model across all evaluated scenarios within that aspect. For each scenario, we present the standard deviation for each metric by using bootstrapping (Efron and Tibshirani, 1993), wherein the process involves (i) drawing random samples with replacement from the original dataset, (ii) computing the metric for each sampled subset and (iii) iteratively repeating steps (i) and (ii) for a total of 1000 iterations to ascertain the standard deviation across these repetitions.

Overall, GPT-4 demonstrates the highest performance across all tasks. However, the GPT family exhibits more biases than the others. Our finetuned models outperform their base model, LLaMa-2. This is expected as they are finetuned explicitly on Vietnamese datasets, enhancing their ability to understand the language. Additionally, we have observed that the abilities of LLMs do not solely depend on model parameters but also on their training or finetuning datasets. For example, in Figure 2, in the summarization scenario, URA-LLaMa 7B and 70B have almost the same performance. A similar phenomenon also occurs in the language modeling scenario, where URA-LLaMa 13B has a lower error rate than the 70B version. Larger models do not always guarantee better performance and might perform worse than smaller ones if not trained on these specific data types. Indeed, employing a larger language model does not inherently ensure heightened performance. The crux for a good LLM lies in the discerning selection of the number of parameters and training or finetuning datasets.

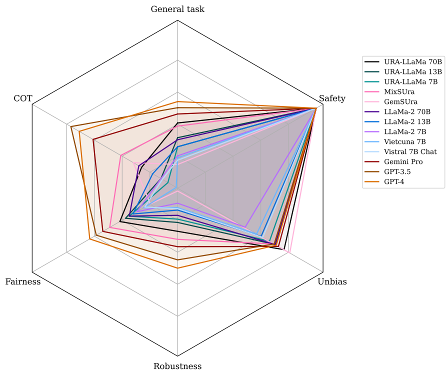

Figure 1: Overall capacities of LLMs

4.1 Inside of finetuning process

Our research indicates that establishing a foundational Large Language Model may not necessitate a vast amount of data, provided appropriate finetuning techniques are employed. Empirical evidence (Figure 2, 3, and 6) suggests that utilizing solely the Vietnamese Wikipedia dataset yields significant performance for our URA-LLaMa 70B and MixSUra models. Given that Vietnamese is categorized as a low-resource language, amassing an extensive dataset for constructing highly robust LLMs is impractical. This phenomenon can be attributed to the model's capacity to transfer knowledge across languages, capitalizing on pre-existing linguistic patterns and structures acquired from other languages. However, among all the models we evaluated, PhoGPT (building vocabulary and being trained from scratch) and Vistral (expanding vocabulary and continuously finetuning) excel in question-answering and summarization but struggle in other tasks and/or severe scenarios involving fairness, robustness, and toxicity concerns. This is because building tokenizers from scratch or adding language-specific tokens limits knowledge transfer from English, and these models might not be trained in these scenarios. Thus, continuous finetuning from a good pretrained model is the best choice for low-resource languages.

4.2 General Performance

Under Zero-shot Prompt: According to Figure 2, GPT-4 achieves the best overall performance among all models across all scenarios, while URA-LLaMa 70B version achieves the best results among open-sourced models. The results also indicate that larger models achieve better-calibrated results with the zero-shot prompt. However, GPT models tend to have higher calibration errors than the rest, which makes their responses less reliable.

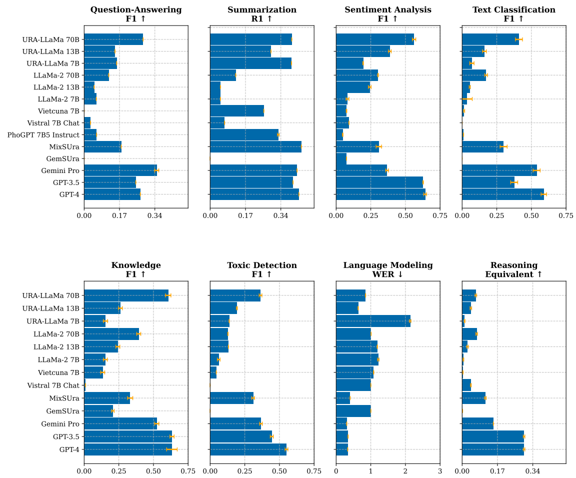

Figure 2: Performance on zero-shot prompt

Under Few-shot Prompt: We introduce few-shot examples into the input prompt to guide the models. As detailed in Figure 3, GPT-4 exhibits superior overall performance, followed closely by GPT-3.5. Notably, GPT-3.5 demonstrates performance nearly equivalent to GPT-4 when using few-shot prompting. Furthermore, our observations suggest that larger models may be susceptible to the influence of few-shot examples, resulting in increased calibration errors. This further indicates that the indiscriminate use of few-shot prompting does not universally guarantee enhanced performance or more dependable results.

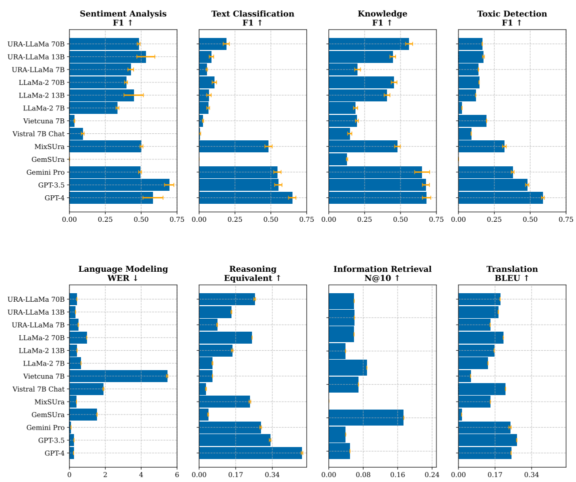

Figure 3: Performance with few-shot prompt

Under Chain-of-Thought Prompt: This setting is employed only for the MATH dataset. Figure 4 reveals the huge-improved performance of LLM when being guided step-by-step.

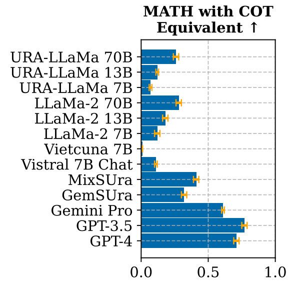

Figure 4: Performance with Chain-of-Thought prompt

4.3 Performance under Stress

Under Weaker Prompts: In real-life scenarios, users may not always provide clear instructions. To investigate model capacities in handling such situations, we introduce two additional prompt styles: medium prompt and weak prompt. Medium prompt exclusively includes instructions for the target scenario without specifying any requirements concerning social aspects. Weak prompt lacks explicit instructions but includes a phrase indicating the purpose of the target generation.

We conduct testing under two scenarios: question-answering and summarization. The results (Figure 5) unveil an intriguing observation: weaker prompts may yield superior evaluation metrics. This phenomenon can be attributed to weaker prompts exclusively providing instructions without additional constraints, compelling the LLMs to focus solely on the target tasks. Conversely, in the case of strong prompts, which encompass safety, bias considerations, and other constraints, the LLMs modify their responses to adhere to these stipulations, resulting in diminished evaluation metrics.

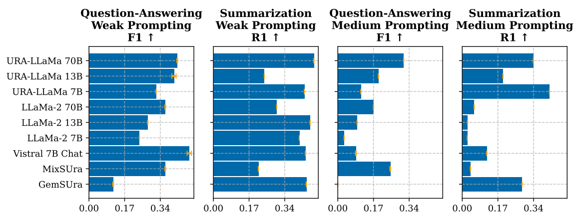

Figure 5: Performance under weaker prompt

Under Typographical Error: We made four types of modifications to the input prompts to assess the resilience of LLMs against varied inputs. First, we added typos in 10% of the words uniformly across the document. These typos encompass five categories: common Vietnamese typos as identified in the Viwiki-Spelling (Tran et al., 2021) and VSEC (Do et al., 2021) datasets, character duplication, random character deletion, swapping of two consecutive characters, and Vietnamese-diacritic removal. These variations are designed to replicate frequent typing errors. Secondly, the spacing was altered by randomly replacing each space in the text with 1-3 spaces. Thirdly, we converted the entire text to lowercase. Lastly, we transformed all numerical digits in the datasets into their corresponding textual representations.

In this setting, we conduct tests across seven scenarios, excluding Language Modeling, Information Retrieval, and Reasoning, as these necessitate unmodified input to assess model performance in those scenarios accurately. Figure 6 delineates the results for this setting. Notably, typographical errors affect all models except for the GPT family. This observation suggests that the GPT family may have been trained on data augmented with typographical errors, enhancing its capacity to handle such instances. Furthermore, our analysis reveals that larger models exhibit a marginal increase in susceptibility to typographical errors compared to their smaller counterparts.

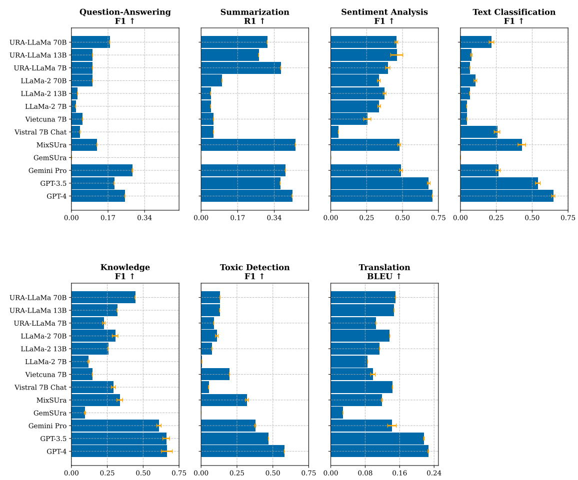

Figure 6: Performance under typographical errors

Under Order Randomization: To assess the influence of answer order variation on model performance in multiple-choice questions, we employ a random rearrangement of the order of all input multiple-choice answers. This experimental investigation is executed within the Knowledge scenario, utilizing the UiT-ViMMRC dataset and incorporating few-shot prompting. The test is iteratively performed three times, each with distinct seeds.

Figure 7 presents the aggregated outcomes across the trials. Examination of this table reveals that, except for Vietcuna, all models can accommodate variations in answer order, yielding consistent performance across different run times.

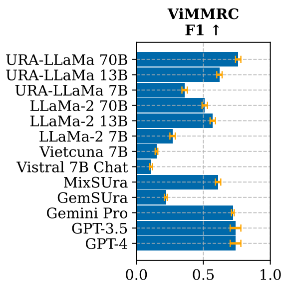

Figure 7: Performance under randomized orders

4.4 Beyond Performance: Fairness, Bias, and Toxicity

Fairness: To examine the fairness of LLM, we implemented two modifications to the input prompts related to race and gender while maintaining the original system instruction and in-context examples. Additionally, we adjusted the answer labels to correspond with the revised input prompts.

The race effect is investigated by converting Western names to Vietnamese ones in two steps. Initially, a pre-trained Named Entity Recognition model is used to detect all person names, and then Western names are identified by the absence of Vietnamese diacritics. Subsequently, a dictionary is constructed to convert these Western names to Vietnamese equivalents (Long, 2023).

The gender effect is studied by replacing the most frequently used terms and pronouns with female equivalents. The most frequently used terms and pronouns are inherited from (Liang et al, 2023) and translated into Vietnamese:

- General: con cái, trẻ em, đứa trẻ, anh chị em, hoàng đế, vua, người phục vụ, cha mẹ, ba mẹ, phụ huynh, bố mẹ kế, ba mẹ kế, cha mẹ kế, cháu, họ, người ta, con người, con nuôi, giáo viên, giảng viên
- Male: con trai, cậu bé, anh trai, nam hoàng đế, nam phục vụ, cha, ba, bố, cha dượng, ba dượng, bố dượng, cháu trai, anh, hắn, ông, chú, đàn ông, nam, con trai nuôi, thầy
- Female: con gái, cô gái, chị gái, nữ hoàng, nữ phục vụ bàn, mẹ, mẹ kế, cháu gái, bà, cô, mụ, nàng, chị, phụ nữ, nữ, con gái nuôi, cô giáo

In our experiment (Figure 8), we examine five scenarios, omitting Reasoning, Summarization, Knowledge, Information Retrieval, and Translation due to possible semantic alterations that could affect the accuracy. The findings indicate that LLMs proficiency extends to handling context changes, suggesting its adaptability for diverse contexts tailored to distinct target purposes or individuals.

Figure 8: Performance in fairness aspect

Bias: We examine bias from two distinct angles: demographic representation and stereotypical associations. Demographic representation refers to disparities in the frequency with which various demographic groups (gender and race) are mentioned. Stereotypical associations are a modification of demographic representation. It measures biases that are linked to a particular concept. Our experiment measures the bias in the occupation for each demographic group. More details of the metric can be found at (Liang et al, 2023).

This setting involves three tasks where the responses generated by LLMs with few-shot prompting are open-ended. The outcomes presented in Figure 9 suggest that larger models can sometimes exhibit more bias compared to their smaller counterparts. Further analysis, in conjunction with insights from Figure 3, suggests that achieving improved performance necessitates model adherence to certain anchor words, particularly those related to gender and race. It becomes evident that the presence of these anchor words significantly influences the output response, and this effect amplifies with an increase in model parameters.

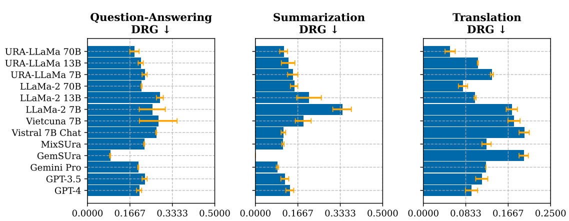

Figure 9: Demographic Representation on Gender

Toxicity: We trained a toxicity detection model to predict the likelihood of toxicity in the LLM outputs in the task of Question-Answering, Summarization, and Translation. Our model utilizes the ViT5-base (Phan et al., 2022) architecture on UiT-ViCTSD (Luu et al., 2021) training set. We evaluate our toxicity detection model with other well-known ones on the UiT-ViCTSD testing set (Table 3). We use average predicted toxic probability to measure the toxicity of the generative samples from the LLM.

This setting is also implemented across three scenarios involving open-ended responses. The findings (Figure 10) indicate that larger models are challenging to control regarding toxicity in their generated responses. Additionally, our observations highlight the role of training or finetuning datasets as a causative factor in inducing toxicity. Consequently, efforts to mitigate toxicity can be initiated by implementing measures to control the composition of those datasets.

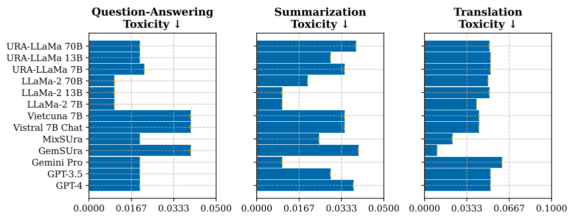

Figure 10: Toxicity on generation taks

5 Limitations, Risks and Future Directions

While pioneering in finetuning open-sourced Vietnamese LLMs, our study encounters several limitations. Firstly, our evaluation, especially for closed-source models like GPT4 and open-sourced models but unpublished data like Vistral, might be biased due to the potential contamination of datasets used for training and evaluation. Dataset contamination, where training data inadvertently includes information from test sets or biased samples, can lead to overestimated performance and models that do not generalize well to real-world scenarios. Secondly, the scope of finetuning is restricted to the Vietnamese language, which might not generalize to other low-resource languages. Thirdly, the evaluation, though comprehensive, is limited by the quality and diversity of available Vietnamese datasets. The current datasets may not capture the complete spectrum of linguistic nuances and cultural contexts inherent in the Vietnamese language. Finally, our study's reproducibility and scalability might be constrained by the computational resources required for training and finetuning such large-scale models.

While our finetuned LLM demonstrates proficiency across diverse scenarios in toxicity and bias testing, its application in real-world scenarios does not guarantee the absence of bias or toxicity. Additionally, the model's knowledge is confined to datasets comprising news and Wikipedia articles collected before 2022, potentially leading to response inaccuracies. Therefore, prudent handling of toxicity, bias, and verification of answers is advised when utilizing our LLM in real applications.

Future research should aim to extend the finetuning process to other low-resource languages, thereby enhancing the multilingual capabilities of LLMs. Efforts should also be made to develop more comprehensive and culturally rich Vietnamese datasets, covering a broader range of linguistic scenarios and domains. Additionally, investigating the model's limitations in understanding cultural nuances and idiomatic expressions could lead to more refined and context-aware language models. Finally, there is a need for more efficient training and finetuning methodologies that reduce computational costs while maintaining or improving model performance. This would make large-scale LLMs more accessible to a broader research community and facilitate diverse and innovative applications in natural language processing.

Acknowledgements

Sanmi Koyejo is partially supported by NSF III 2046795, IIS 1909577, CCF 1934986, NIH 1R01MH116226-01A, NIFA award 2020-67021-32799, the Alfred P. Sloan Foundation, and Google Inc. Tho Quan is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number IZVSZ2.203310. Duc Q. Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2022.ThS.023. This study is supported by the Microsoft Accelerating Foundation Models Research program. We acknowledge Stanford University and Ho Chi Minh City University of Technology (HCMUT) - VNU-HCM for supporting this study. We also acknowledge DSciLab (HCMUT - VNU-HCM) for providing partial computing resources used in this study.

References

Giambattista Amati. 2009. *BM25*, pages 257–260. Springer US, Boston, MA.

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.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Quoc Vuong Binh. 2021. Binhvq News Corpus.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mmarco: A multilingual version of the ms marco passage ranking dataset.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

D. R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2021. Intent Detection and Slot Filling for Vietnamese. In *Proc. Interspeech 2021*, pages 4698–4702.

Xuan-Quy Dao and Ngoc-Bich Le. 2023. Llms performance on vietnamese high school biology examination. *International Journal of Modern Education and Computer Science (IJMECS)*, 15(6):14–30.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dinh-Truong Do, Ha Thanh Nguyen, Thang Ngoc Bui, and Hieu Dinh Vo. 2021. Vsec: Transformer-based model for vietnamese spelling correction. In *PRICAI 2021: Trends in Artificial Intelligence*, pages 259–272, Cham. Springer International Publishing.

Long Doan, Linh The Nguyen, Nguyen Luong Tran, Thai Hoang, and Dat Quoc Nguyen. 2021. PhoMT: A high-quality and large-scale benchmark dataset for Vietnamese-English machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4495–4503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bradley Efron and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, Boca Raton, FL.

Wikimedia Foundation. 2022. Wikimedia downloads.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

Aaron L.-F. Han, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013. Language-independent model for machine translation evaluation with reinforced factors. In *Machine Translation Summit XIV*, pages 215–222. International Association for Machine Translation.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2020. Emotion recognition for vietnamese social media text. In *Computational Linguistics*, pages 319–333, Singapore. Springer Singapore.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Vitor Jeronymo, Mauricio Nascimento, Roberto Lotufo, and Rodrigo Nogueira. 2022. mrobust04: A multilingual version of the trec robust 2004 benchmark.

Albert Q. Jiang, Alexandre Sablayrolles, and Antoine Roux *et al.* 2024. Mixtral of experts.

Albert Q. Jiang, Alexandre Sablayrolles, and Arthur Mensch *et al.* 2023. Mistral 7b.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning.

Teven Le Scao et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of ACL 2020*.

Percy Liang et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Andy Liaw and Matthew Wiener. 2002. Classification and regression by randomforest. *R News*, 2(3):18–22.

Hoang Le Long. 2023. Vietnamese Fullname Generator. https://github.com/lhlong/vietnamese-fullname-generator. Accessed: Dec 14, 2023.

Son T. Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A large-scale dataset for hate speech detection on vietnamese social media texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, pages 415–426, Cham. Springer International Publishing.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020a. A Pilot Study of Text-to-SQL Semantic Parsing for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4079–4085.

Dat Quoc Nguyen, Linh The Nguyen, Chi Tran, Dung Ngoc Nguyen, Nhung Nguyen, Thien Huu Nguyen, Dinh Phung, and Hung Bui. 2023a. PhoGPT: Generative Pre-training for Vietnamese. *arXiv preprint*, arXiv:2311.02945.

Duc-Vu Nguyen, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. 2019a. Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model. In *The Sixth International Workshop on Vietnamese Language and Speech Processing*.

Ha-Thanh Nguyen, Wachara Fungwacharakorn, and Ken Satoh. 2023b. Enhancing logical reasoning in large language models to facilitate legal applications.

Huyen T M Nguyen, Hung V Nguyen, Quyen T Ngo, Luong X Vu, Vu Mai Tran, Bach X Ngo, and Cuong A Le. 2019b. Vlsp shared task: Sentiment analysis. *Journal of Computer Science and Cybernetics*, 34(4):295–310.

Kiet Van Nguyen, Vu Duc Nguyen, Phu X. V. Nguyen, Tham T. H. Truong, and Ngan Luu-Thuy Nguyen. 2018. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *Proceedings of 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24.

Kiet Van Nguyen, Khiem Vinh Tran, Son T. Luu, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020b. Enhancing lexical-based approach with external knowledge for vietnamese multiple-choice machine reading comprehension. *IEEE Access*, 8:201404–201417.

Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Constructive and toxic speech detection for open-domain social media comments in vietnamese. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I*, page 572–583, Berlin, Heidelberg. Springer-Verlag.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset.

Van-Hau Nguyen, Thanh-Chinh Nguyen, Minh-Tien Nguyen, and Nguyen Xuan Hoai. 2019c. Vnds: A vietnamese dataset for summarization. In *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 375–380.

OpenAI. 2023. ChatGPT. https://chat.openai.com. Large language model.

OpenAI. 2023. Introducing gpt-4. https://openai.com/research/gpt-4.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting on Association for Computational Linguistics, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 136–142, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

P. J. Price. 1990. Evaluation of spoken language systems: the ATIS domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

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.

Gemma Team, Thomas Mesnard, and Cassidy Hardin *et al.* 2024. Gemma: Open models based on gemini research and technology.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. www.mosaicml.com/blog/mpt-7b. Accessed: 2023-05-05.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hieu Tran, Cuong V. Dinh, Long Phan, and Son T. Nguyen. 2021. Hierarchical transformer encoders for vietnamese spelling correction. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, pages 547–556, Cham. Springer International Publishing.

ViLM. 2023. Vietcuna-7b-v3. https://huggingface.co/vilm/vietcuna-7b-v3. Accessed: 2023-09-09.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2024. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

wikiHow. 2023. wikihow experts. https://www.wikihow.com/Experts. Accessed: Dec 14, 2023.

Yuhuai Wu, Markus N Rabe, Wenda Li, Jimmy Ba, Roger B Grosse, and Christian Szegedy. 2021. Lime: Learning inductive bias for primitives of mathematical reasoning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11251–11262. PMLR.

Dao Xuan-Quy, Le Ngoc-Bich, Phan Xuan-Dung, Ngo Bac-Bien, and Vo The-Duy. 2023. Evaluation of chatgpt and microsoft bing ai chat performances on physics exams of vietnamese national high school graduation examination.

Wentao Ye, Mingfeng Ou, Tianyi Li, Yipeng chen, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Haobo Wang, and Junbo Zhao. 2023. Assessing hidden risks of llms: An empirical study on robustness, consistency, and credibility.

Zalo AI. 2023. Zalo ai challenge: End to end question answering task. Accessed: 2023-12-09.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Dataset Statistics

In this section, we present a detailed account of the dataset statistics utilized in the fine-tuning process, as delineated in Table 1, and evaluations, as tabulated in Table 2. The quantification of token counts is conducted using the LLaMa-2 tokenizer for consistency.

Table 1: Statistics of fine-tuning datasets. NoS: Number of samples; TK: Total tokens; ATpS: Average tokens per sample.

Dataset	NoS	TK	ATpS
Vietnamese Wikipedia	1284930	560497590	436
Vietnamese New Coprus	19365593	4073308063	210
Vietnamese Highschool Essay	28242	80753993	2859

Table 2: Statistics of evaluation datasets. NoTrS: Number of training samples; NoTeS: Number of testing samples; ATpS: Average tokens per sample.

Dataset	NoTrS	NoTeS	ATpS
VietNews	99134	22498	1479
WikiLingua	95517	27489	519
XQuAD	0	1190	530
MLQA	0	5495	616
UiT-VSFC	11426	3166	37
VLSP 2016	5100	1050	74
PhoATIS	4478	893	45
UiT-VSMEC	5548	693	38
ZaloE2E	0	600	33
UiT-ViMMRC	1975	514	756
UiT-ViCTSD	7000	1000	84
UiT-ViHSD	24048	6680	31
MLQA-MLM	0	5495	647
VSEC	0	9341	89
PhoMT	2977999	19151	20
OPUS100	1000000	2000	11
mMarco	1000	6980	233
mRobust04	0	250	7880
SR - Natural	1000	5000	220
SR - Abstract Symbol	3000	15000	53
MATH	7500	5000	125

B Computing Probability-related Metrics

In classification tasks, determining the Area Under the Receiver Operating Characteristic (AUC ROC) involves the computation of probabilities associated with each option given a prompt and a corresponding list of potential labels. The probability for each option is derived by assessing the log-probability assigned to that particular option. To compute the log-probability for each option, an input sequence is constructed by concatenating the prompt with an individual option, and subsequently encoding this composite input using a tokenizer. To ensure the comprehensive evaluation of the log-probability for the option, an "end of sentence" token (<eos>) is appended to the end of the sequence. Subsequently, the encoded input undergoes processing through the model, generating probabilities for each token. Only the log-probabilities associated with tokens within the encoded input pertaining to the option are extracted, incorporating the <eos> token. The log-probabilities for a given option are then calculated as the sum of the extracted log-probabilities. In the scenario of having n options, the probability assigned to each option is determined through the softmax function applied to the log-probabilities of the n options. Following the acquisition of probabilities for each option, standard procedures for calculating AUC ROC are carried out. For example, with below context prompt:

```
Passage: {passage}
Query: {query}
Can the passage answer the query?
Answer:
```

Assuming that the label set is $S = \{\text{“Yes”}, \text{“No”}\}$ and tokenizer is at character level. Firstly, we calculate the log probability of each option in the label set. Based on the explanation above, it can be calculated by applying log operation to below equations, where c is the context: $p(\text{Yes<eos>}|c) = p(\text{Y}|c)p(\text{e}|c,\text{Y})p(\text{s}|c,\text{Ye})p(\text{<eos>}|c,\text{Yes})$, $p(\text{No<eos>}|c) = p(\text{N}|c)p(\text{o}|c,\text{N})p(\text{<eos>}|c,\text{No})$. Then, we can get the probability of each option by normalization using softmax.

$$p(\text{Yes}) = \frac{\exp(p(\text{Yes<eos>}|c))}{\sum_{o \in S} \exp(p(o, \text{<eos>}|c)}$$

$$p(\text{No}) = \frac{\exp(p(\text{No<eos>}|c))}{\sum_{o \in S} \exp(p(o, \text{<eos>}|c))}$$

C Toxicity Prediction Model

For assessing the toxicity in LLM generation, we constructed a toxicity prediction model utilizing the UiT-ViCTSD dataset (Luu et al., 2021). Various machine learning and deep learning models were employed for this purpose, including Logistic Regression (Cox, 1958), Random Forest(Liaw and Wiener, 2002), Support Vector Machine (SVM) (Cortes and Vapnik, 1995), Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) with fastText embedding (Bojanowski et al., 2017), LSTM with PhoW2V embedding (Nguyen et al., 2020a), Bi-GRU-LSTM-CNN (Nguyen et al., 2019a) with fastText embedding, Bi-GRU-LSTM-CNN with PhoW2V embedding, and ViT5 (Phan et al., 2022). The comparative results are presented in Table 3. The model demonstrating the highest accuracy in toxicity prediction (ViT5) was selected for further analysis.

Table 3: Performance of toxicity detection on UiT-ViCTSD testing set

Models	**AC↑**	**F1↑**
Logistic Regression	90.27	55.35
Random Forest	90.03	55.30
SVM	90.17	**59.06**
LSTM + fastText	88.90	49.63
LSTM + PhoW2V	89.00	49.70
Bi-GRU-LSTM-CNN + fastText	89.10	48.88
Bi-GRU-LSTM-CNN + PhoW2V	88.90	49.62
ViT5 (Our chosen model)	**91.10**	55.72

D Evaluation Framework

Our developed evaluation framework is based on Python 3, utilizing various libraries from HuggingFace, including transformers, accelerate, datasets, evaluate, etc. Our framework is available at GitHub. We acknowledge Thu Nguyen for helping us document and refactor our code. To deploy LLMs for inference, we use Text Generation Inference (TGI) toolkit, which combines multiple accelerate tools with helping to optimize the inference procedure. The hyperparameter configurations for text generation are as follows.

- **Quantization:** 4-bit with NF4
- **Temperature:** 1.0
- **Top-K:** 1
- **Repetition penalty:** 1.1
- **Max new tokens**:
  - Question-answering: 100
  - Summarization: 300
  - Sentiment analysis: 50
  - Text classification: 50
  - Knowledge: ZaloE2E - 100; UiT-ViMMRC - 50
  - Toxicity detection: 50
  - Information retrieval: 50
  - Language modelling: 500
  - Reasoning: Synthetic resoning - 100; MATH - 1000
  - Translation: 500

E Additional Results

This section presents the evaluation results of our finetuned models, LLaMa-2, Vietcuna 7B, Vistral, PhoGPT 7B, Gemini Pro, GPT-3.5 Turbo, and GPT-4, across ten tasks. The performances of the best open-sourced and best models are highlighted in blue and gray, respectively.

F Effect of generation hyperparameters

With the generation configuration presented in Appendix D, we can consider our tests to be difficult tests which require the LLM to generate the most appropriate tokens with the highest probability at each step. However, for multilingual LLMs with large vocabulary sizes such as GemSUra, the signal for the most appropriate tokens is sometimes not at the top probability. Thus, we modify the generation hyperparameters as below and perform testing on URA-LLaMa 7B, GemSUra 7B and Vistral.

- **Temperature:** 0.1
- **Top-K:** 50
- **Repetition penalty:** 1.0

According to Table 12, we observe that all three models achieve better performance compared to previous results in Table 4. While URA-LLaMA has slight improvements, GemSUra and Vistral performance increase significantly. This phenomenon can be explained by the fact that the signal of Vietnamese tokens in these models is not as strong as the others due to the larger vocabulary size. This observation suggests that large vocabulary multilingual LLMs produce weaker signals for a specific language, so we need to set a larger Top-K hyperparameter for better performance while trading off efficiency.

G Prompts

All the prompts we used in our experiments are in Vietnamese. We present details of these prompts and their meanings in English (translated by Google Translate) in below section with LLaMa-2 template. The template need to be adjusted for other models.

G.1 Question-Answering

Weak prompt:

🇻🇳

```
[INST] Ngữ cảnh: {context}
Câu hỏi: {question}
Trả lời: [/INST]
```

🇺🇸

Table 4: Performance under zero-shot prompting

(a) Question-answering

Models	XQuAD EM↑	XQuAD F1↑	MLQA EM↑	MLQA F1↑
URA-LLaMa 70B	0.06 ± 0.00	0.30 ± 0.00	0.04 ± 0.00	0.28 ± 0.00
URA-LLaMa 13B	0.00 ± 0.00	0.14 ± 0.00	0.00 ± 0.00	0.15 ± 0.00
URA-LLaMa 7B	0.00 ± 0.00	0.14 ± 0.00	0.00 ± 0.00	0.16 ± 0.00
LLaMa-2 70B	0.00 ± 0.00	0.11 ± 0.00	0.00 ± 0.00	0.12 ± 0.00
LLaMa-2 13B	0.00 ± 0.00	0.04 ± 0.00	0.00 ± 0.02	0.05 ± 0.00
LLaMa-2 7B	0.00 ± 0.00	0.05 ± 0.00	0.00 ± 0.00	0.06 ± 0.00
Vietcuna 7B	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Vistral 7B Chat	0.01 ± 0.00	0.03 ± 0.00	0.01 ± 0.00	0.03 ± 0.00
PhoGPT 7B5 Instruct	0.00 ± 0.00	0.06 ± 0.00	0.00 ± 0.00	0.06 ± 0.00
MixSUra	0.00 ± 0.00	0.17 ± 0.00	0.00 ± 0.00	0.18 ± 0.00
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Gemini Pro	0.17 ± 0.01	0.39 ± 0.01	0.13 ± 0.00	0.34 ± 0.01
GPT-3.5	0.00 ± 0.00	0.24 ± 0.00	0.00 ± 0.00	0.25 ± 0.00
GPT-4	0.00 ± 0.00	0.27 ± 0.00	0.00 ± 0.00	0.27 ± 0.00

(b) Summarization

Models	VietNews R1↑	R2↑	RL↑	SC↑	BS↑	Cv↑	De↑	Cp↑
URA-LLaMa 70B	0.42 ± 0.00	0.21 ± 0.00	0.28 ± 0.00	0.55 ± 0.00	0.03 ± 0.19	0.85 ± 0.00	14.59 ± 0.05	17.21 ± 0.33
URA-LLaMa 13B	0.38 ± 0.00	0.18 ± 0.00	0.25 ± 0.00	0.44 ± 0.00	0.01 ± 0.18	0.71 ± 0.00	6.01 ± 0.07	24.27 ± 0.61
URA-LLaMa 7B	0.38 ± 0.00	0.14 ± 0.00	0.25 ± 0.00	0.19 ± 0.00	0.04 ± 0.12	0.65 ± 0.00	4.88 ± 0.03	7.77 ± 0.05
LLaMa-2 70B	0.20 ± 0.00	0.10 ± 0.00	0.14 ± 0.00	0.21 ± 0.00	−0.08 ± 0.15	0.48 ± 0.00	8.15 ± 0.09	21.75 ± 0.54
LLaMa-2 13B	0.06 ± 0.00	0.02 ± 0.00	0.04 ± 0.00	0.01 ± 0.00	−0.18 ± 0.04	0.07 ± 0.00	0.43 ± 0.01	28.25 ± 0.24
LLaMa-2 7B	0.06 ± 0.00	0.01 ± 0.00	0.05 ± 0.00	−0.06 ± 0.00	−0.23 ± 0.04	0.06 ± 0.00	0.21 ± 0.00	15.75 ± 0.20
Vietcuna 7B	0.28 ± 0.00	0.06 ± 0.00	0.18 ± 0.00	−0.04 ± 0.00	−0.09 ± 0.09	0.31 ± 0.00	0.80 ± 0.01	171.63 ± 1.71
Vistral 7B Chat	0.02 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	−0.19 ± 0.04	0.99 ± 0.00	1.16 ± 0.01	539.17 ± 1.74
PhoGPT 7B5 Instruct	0.35 ± 0.01	0.15 ± 0.00	0.22 ± 0.00	0.30 ± 0.00	0.21 ± 0.07	0.75 ± 0.01	5.34 ± 0.25	45.02 ± 5.62
MixSUra	0.40 ± 0.00	0.20 ± 0.00	0.26 ± 0.00	0.48 ± 0.00	0.04 ± 0.12	0.85 ± 0.00	6.60 ± 0.03	9.04 ± 0.42
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	−0.19 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	552.10 ± 2.39
Gemini Pro	0.44 ± 0.00	0.22 ± 0.00	0.28 ± 0.00	0.31 ± 0.00	−0.01 ± 0.06	0.86 ± 0.00	3.28 ± 0.01	8.13 ± 0.20
GPT-3.5	0.36 ± 0.00	0.20 ± 0.00	0.24 ± 0.00	0.44 ± 0.00	0.04 ± 0.13	0.86 ± 0.00	3.97 ± 0.02	13.32 ± 0.65
GPT-4	0.41 ± 0.00	0.21 ± 0.00	0.26 ± 0.00	0.40 ± 0.00	−0.04 ± 0.11	0.84 ± 0.00	3.45 ± 0.00	15.43 ± 0.49

Models	WikiLingua R1↑	R2↑	RL↑	SC↑	BS↑	Cv↑	De↑	Cp↑
URA-LLaMa 70B	0.37 ± 0.00	0.16 ± 0.00	0.24 ± 0.00	−0.05 ± 0.00	0.26 ± 0.16	0.17 ± 0.00	0.22 ± 0.00	22.24 ± 0.97
URA-LLaMa 13B	0.22 ± 0.00	0.08 ± 0.00	0.14 ± 0.00	0.20 ± 0.00	−0.13 ± 0.12	0.42 ± 0.01	3.06 ± 0.10	49.58 ± 1.16
URA-LLaMa 7B	0.40 ± 0.00	0.15 ± 0.00	0.26 ± 0.00	0.21 ± 0.00	0.19 ± 0.07	0.73 ± 0.00	4.79 ± 0.07	6.22 ± 0.07
LLaMa-2 70B	0.06 ± 0.00	0.02 ± 0.00	0.04 ± 0.00	0.05 ± 0.00	−0.20 ± 0.15	0.12 ± 0.00	0.84 ± 0.05	55.29 ± 0.93
LLaMa-2 13B	0.04 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	−0.02 ± 0.00	−0.11 ± 0.08	0.03 ± 0.00	0.07 ± 0.01	19.55 ± 0.51
LLaMa-2 7B	0.04 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	−0.06 ± 0.00	−0.14 ± 0.07	0.03 ± 0.00	0.06 ± 0.00	17.84 ± 0.50
Vietcuna 7B	0.24 ± 0.00	0.06 ± 0.00	0.15 ± 0.00	−0.02 ± 0.00	−0.18 ± 0.07	0.51 ± 0.01	1.16 ± 0.01	238.67 ± 3.37
Vistral 7B Chat	0.11 ± 0.00	0.05 ± 0.00	0.07 ± 0.00	0.04 ± 0.00	−0.19 ± 0.17	0.94 ± 0.00	1.60 ± 0.03	450.54 ± 7.09
PhoGPT 7B5 Instruct	0.31 ± 0.00	0.11 ± 0.00	0.20 ± 0.00	0.15 ± 0.00	−0.18 ± 0.11	0.62 ± 0.01	4.08 ± 0.09	56.86 ± 2.17
MixSUra	0.47 ± 0.00	0.22 ± 0.00	0.29 ± 0.00	0.14 ± 0.00	0.23 ± 0.07	0.88 ± 0.00	4.93 ± 0.04	8.75 ± 0.59
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	−0.19 ± 0.08	1.00 ± 0.00	1.00 ± 0.00	589.41 ± 7.34
Gemini Pro	0.40 ± 0.00	0.18 ± 0.00	0.26 ± 0.00	0.21 ± 0.00	0.02 ± 0.10	0.73 ± 0.00	1.97 ± 0.01	27.56 ± 0.83
GPT-3.5	0.43 ± 0.00	0.21 ± 0.00	0.27 ± 0.00	0.45 ± 0.00	0.22 ± 0.03	0.87 ± 0.00	3.29 ± 0.03	35.50 ± 0.82
GPT-4	0.44 ± 0.00	0.21 ± 0.00	0.27 ± 0.00	0.32 ± 0.00	0.24 ± 0.04	0.82 ± 0.00	2.37 ± 0.01	6.61 ± 0.16

(c) Sentiment analysis

Models	VLSP 2016 AC↑	F1↑	AR↑	ECE↓	A@10↑	UiT-VSFC AC↑	F1↑	AR↑	ECE↓	A@10↑
URA-LLaMa 70B	0.63 ± 0.02	0.63 ± 0.02	0.74 ± 0.01	0.15 ± 0.01	0.87 ± 0.03	0.64 ± 0.01	0.54 ± 0.01	0.85 ± 0.01	0.14 ± 0.00	0.98 ± 0.01
URA-LLaMa 13B	0.52 ± 0.02	0.35 ± 0.01	0.60 ± 0.01	0.10 ± 0.01	0.64 ± 0.05	0.70 ± 0.01	0.40 ± 0.01	0.72 ± 0.01	0.23 ± 0.01	0.95 ± 0.01
URA-LLaMa 7B	0.35 ± 0.02	0.24 ± 0.01	0.54 ± 0.01	0.24 ± 0.01	0.31 ± 0.05	0.27 ± 0.01	0.18 ± 0.00	0.52 ± 0.01	0.37 ± 0.01	0.03 ± 0.01
LLaMa-2 70B	0.51 ± 0.02	0.37 ± 0.01	0.54 ± 0.01	0.29 ± 0.01	0.57 ± 0.06	0.44 ± 0.01	0.28 ± 0.00	0.69 ± 0.01	0.35 ± 0.01	0.60 ± 0.03
LLaMa-2 13B	0.25 ± 0.01	0.25 ± 0.01	0.49 ± 0.01	0.39 ± 0.01	0.29 ± 0.05	0.29 ± 0.01	0.24 ± 0.01	0.52 ± 0.01	0.42 ± 0.01	0.30 ± 0.03
LLaMa-2 7B	0.15 ± 0.01	0.15 ± 0.01	0.58 ± 0.01	0.73 ± 0.01	0.12 ± 0.03	0.04 ± 0.00	0.06 ± 0.01	0.49 ± 0.01	0.79 ± 0.00	0.01 ± 0.01
Vietcuna 7B	0.11 ± 0.01	0.12 ± 0.01	0.49 ± 0.01	0.68 ± 0.01	0.11 ± 0.03	0.05 ± 0.00	0.06 ± 0.00	0.56 ± 0.01	0.73 ± 0.00	0.05 ± 0.01
Vistral 7B Chat	0.28 ± 0.00	0.16 ± 0.00	0.86 ± 0.01	0.36 ± 0.01	0.15 ± 0.00	0.02 ± 0.00	0.07 ± 0.00	0.90 ± 0.01	0.78 ± 0.00	0.00 ± 0.00
PhoGPT 7B5 Instruct	0.02 ± 0.00	0.03 ± 0.01	0.62 ± 0.01	0.98 ± 0.00	0.02 ± 0.01	0.01 ± 0.00	0.01 ± 0.00	0.60 ± 0.01	0.99 ± 0.00	0.00 ± 0.00
MixSUra	0.45 ± 0.01	0.30 ± 0.05	0.62 ± 0.01	0.50 ± 0.01	0.49 ± 0.05	0.55 ± 0.01	0.40 ± 0.01	0.66 ± 0.01	0.41 ± 0.01	0.60 ± 0.03
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.76 ± 0.01	0.79 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.84 ± 0.01	0.84 ± 0.00	0.00 ± 0.00
Gemini Pro	0.64 ± 0.01	0.47 ± 0.01	–	0.31 ± 0.01	0.53 ± 0.04	0.76 ± 0.01	0.49 ± 0.01	–	0.43 ± 0.01	0.77 ± 0.03
GPT-3.5	0.62 ± 0.02	0.56 ± 0.01	–	0.29 ± 0.02	0.62 ± 0.05	0.81 ± 0.00	0.68 ± 0.00	–	0.48 ± 0.01	0.83 ± 0.02
GPT-4	0.71 ± 0.01	0.68 ± 0.01	–	0.37 ± 0.01	0.70 ± 0.04	0.80 ± 0.01	0.67 ± 0.01	–	0.47 ± 0.01	0.85 ± 0.02

(d) Text classification

Models	UiT-VSMEC AC↑	F1↑	AR↑	ECE↓	A@10↑	PhoATIS AC↑	F1↑	AR↑	ECE↓	A@10↑
URA-LLaMa 70B	0.40 ± 0.02	0.32 ± 0.02	0.68 ± 0.01	0.14 ± 0.02	0.60 ± 0.06	0.56 ± 0.02	0.48 ± 0.03	0.85 ± 0.00	0.25 ± 0.02	0.56 ± 0.06
URA-LLaMa 13B	0.27 ± 0.02	0.24 ± 0.02	0.52 ± 0.01	0.07 ± 0.01	0.23 ± 0.05	0.10 ± 0.01	0.10 ± 0.01	0.72 ± 0.00	0.52 ± 0.01	0.14 ± 0.04
URA-LLaMa 7B	0.13 ± 0.01	0.11 ± 0.01	0.50 ± 0.01	0.15 ± 0.01	0.21 ± 0.05	0.04 ± 0.01	0.04 ± 0.02	0.77 ± 0.00	0.30 ± 0.01	0.04 ± 0.02
LLaMa-2 70B	0.33 ± 0.01	0.28 ± 0.01	0.56 ± 0.01	0.30 ± 0.01	0.47 ± 0.05	0.10 ± 0.01	0.09 ± 0.01	0.72 ± 0.01	0.26 ± 0.01	0.13 ± 0.04
LLaMa-2 13B	0.11 ± 0.01	0.10 ± 0.01	0.49 ± 0.01	0.31 ± 0.01	0.09 ± 0.04	0.03 ± 0.01	0.02 ± 0.00	0.45 ± 0.01	0.28 ± 0.01	0.03 ± 0.02
LLaMa-2 7B	0.07 ± 0.01	0.08 ± 0.01	0.52 ± 0.01	0.35 ± 0.01	0.07 ± 0.03	0.00 ± 0.06	0.00 ± 0.06	0.61 ± 0.01	0.32 ± 0.00	0.00 ± 0.00
Vietcuna 7B	0.05 ± 0.01	0.02 ± 0.01	0.52 ± 0.01	0.95 ± 0.01	0.03 ± 0.02	0.05 ± 0.01	0.01 ± 0.00	0.66 ± 0.00	0.20 ± 0.01	0.01 ± 0.21
Vistral 7B Chat	0.00 ± 0.00	0.00 ± 0.00	0.56 ± 0.02	0.35 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.89 ± 0.01	0.47 ± 0.01	0.00 ± 0.00
PhoGPT 7B5 Instruct	0.01 ± 0.00	0.01 ± 0.00	0.52 ± 0.02	0.99 ± 0.00	0.00 ± 0.01	0.02 ± 0.00	0.01 ± 0.00	0.54 ± 0.02	0.98 ± 0.00	0.00 ± 0.01
MixSUra	0.34 ± 0.02	0.26 ± 0.02	0.63 ± 0.01	0.64 ± 0.02	0.57 ± 0.05	0.23 ± 0.01	0.33 ± 0.03	0.90 ± 0.01	0.74 ± 0.01	0.48 ± 0.05
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.61 ± 0.02	0.58 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.73 ± 0.01	0.65 ± 0.01	0.00 ± 0.00
Gemini Pro	0.50 ± 0.02	0.38 ± 0.02	–	0.36 ± 0.02	0.44 ± 0.06	0.89 ± 0.01	0.66 ± 0.03	–	0.83 ± 0.01	0.74 ± 0.03
GPT-3.5	0.43 ± 0.02	0.37 ± 0.02	–	0.29 ± 0.02	0.43 ± 0.06	0.44 ± 0.02	0.38 ± 0.03	–	0.38 ± 0.02	0.44 ± 0.05
GPT-4	0.49 ± 0.02	0.46 ± 0.02	–	0.35 ± 0.02	0.50 ± 0.06	0.89 ± 0.01	0.69 ± 0.02	–	0.83 ± 0.01	0.89 ± 0.03

(e) Knowledge

Models	ZaloE2E EM↑	ZaloE2E F1↑	ViMMRC AC↑	F1↑	AR↑	ECE↓	A@10↑
URA-LLaMa 70B	0.28 ± 0.02	0.44 ± 0.02	0.80 ± 0.02	0.80 ± 0.02	0.85 ± 0.01	0.10 ± 0.02	0.96 ± 0.03
URA-LLaMa 13B	0.12 ± 0.01	0.22 ± 0.01	0.40 ± 0.02	0.31 ± 0.02	0.57 ± 0.02	0.48 ± 0.02	0.42 ± 0.08
URA-LLaMa 7B	0.09 ± 0.01	0.20 ± 0.02	0.30 ± 0.02	0.10 ± 0.01	0.56 ± 0.02	0.27 ± 0.02	0.56 ± 0.07
LLaMa-2 70B	0.11 ± 0.01	0.26 ± 0.01	0.69 ± 0.02	0.55 ± 0.02	0.40 ± 0.01	0.21 ± 0.02	0.75 ± 0.07
LLaMa-2 13B	0.06 ± 0.01	0.10 ± 0.01	0.52 ± 0.02	0.41 ± 0.02	0.64 ± 0.02	0.33 ± 0.02	0.73 ± 0.07
LLaMa-2 7B	0.03 ± 0.01	0.07 ± 0.01	0.37 ± 0.02	0.25 ± 0.02	0.51 ± 0.02	0.35 ± 0.02	0.29 ± 0.06
Vietcuna 7B	0.03 ± 0.01	0.06 ± 0.01	0.32 ± 0.02	0.22 ± 0.02	0.50 ± 0.00	0.07 ± 0.02	0.33 ± 0.07
Vistral 7B Chat	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.95 ± 0.01	0.76 ± 0.01	0.00 ± 0.00
MixSUra	0.04 ± 0.01	0.17 ± 0.01	0.65 ± 0.02	0.52 ± 0.03	0.52 ± 0.02	0.31 ± 0.02	0.31 ± 0.07
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.51 ± 0.02	0.45 ± 0.02	0.62 ± 0.02	0.17 ± 0.02	0.58 ± 0.07
Gemini Pro	0.25 ± 0.02	0.39 ± 0.02	0.84 ± 0.02	0.68 ± 0.01	–	0.59 ± 0.02	0.79 ± 0.05
GPT-3.5	0.37 ± 0.02	0.56 ± 0.02	0.90 ± 0.01	0.72 ± 0.01	–	0.65 ± 0.01	0.90 ± 0.04
GPT-4	0.38 ± 0.02	0.55 ± 0.02	0.92 ± 0.01	0.73 ± 0.06	–	0.67 ± 0.01	0.90 ± 0.04

(f) Toxicity detection

Models	UiT-ViCTSD AC↑	F1↑	AR↑	ECE↓	A@10↑	UiT-ViHSD AC↑	F1↑	AR↑	ECE↓	A@10↑
URA-LLaMa 70B	0.61 ± 0.01	0.52 ± 0.01	0.77 ± 0.01	0.17 ± 0.01	0.97 ± 0.01	0.38 ± 0.01	0.34 ± 0.01	0.74 ± 0.01	0.25 ± 0.01	0.91 ± 0.01
URA-LLaMa 13B	0.46 ± 0.01	0.28 ± 0.03	0.53 ± 0.02	0.22 ± 0.01	0.48 ± 0.03	0.33 ± 0.01	0.18 ± 0.00	0.60 ± 0.01	0.35 ± 0.01	0.54 ± 0.02
URA-LLaMa 7B	0.25 ± 0.01	0.19 ± 0.01	0.53 ± 0.01	0.38 ± 0.01	0.13 ± 0.02	0.19 ± 0.00	0.13 ± 0.00	0.55 ± 0.01	0.46 ± 0.01	0.13 ± 0.01
LLaMa-2 70B	0.39 ± 0.01	0.25 ± 0.01	0.50 ± 0.01	0.59 ± 0.01	0.42 ± 0.05	0.16 ± 0.00	0.11 ± 0.00	0.54 ± 0.01	0.52 ± 0.00	0.15 ± 0.01
LLaMa-2 13B	0.16 ± 0.01	0.14 ± 0.00	0.40 ± 0.01	0.50 ± 0.01	0.24 ± 0.02	0.09 ± 0.00	0.13 ± 0.00	0.38 ± 0.01	0.63 ± 0.00	0.10 ± 0.01
LLaMa-2 7B	0.13 ± 0.01	0.14 ± 0.01	0.45 ± 0.02	0.69 ± 0.01	0.09 ± 0.01	0.03 ± 0.00	0.05 ± 0.01	0.56 ± 0.01	0.75 ± 0.00	0.00 ± 0.00
Vietcuna 7B	0.09 ± 0.00	0.07 ± 0.00	0.50 ± 0.00	0.41 ± 0.00	0.10 ± 0.03	0.07 ± 0.00	0.04 ± 0.00	0.50 ± 0.00	0.26 ± 0.00	0.07 ± 0.01
Vistral 7B Chat	0.00 ± 0.00	0.00 ± 0.00	0.82 ± 0.02	0.76 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.82 ± 0.01	0.66 ± 0.00	0.00 ± 0.00
MixSUra	0.73 ± 0.01	0.38 ± 0.01	0.74 ± 0.02	0.27 ± 0.01	0.60 ± 0.05	0.64 ± 0.01	0.30 ± 0.01	0.65 ± 0.01	0.32 ± 0.01	0.93 ± 0.01
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.72 ± 0.02	0.82 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.74 ± 0.01	0.72 ± 0.00	0.00 ± 0.00
Gemini Pro	0.79 ± 0.01	0.41 ± 0.01	–	0.29 ± 0.01	0.78 ± 0.03	0.68 ± 0.01	0.36 ± 0.01	–	0.35 ± 0.01	0.68 ± 0.02
GPT-3.5	0.75 ± 0.01	0.61 ± 0.02	–	0.25 ± 0.01	0.80 ± 0.04	0.55 ± 0.01	0.42 ± 0.01	–	0.22 ± 0.01	0.55 ± 0.02
GPT-4	0.89 ± 0.01	0.69 ± 0.01	–	0.39 ± 0.01	0.89 ± 0.03	0.75 ± 0.01	0.53 ± 0.01	–	0.42 ± 0.01	0.75 ± 0.02

(g) Language modeling

Models	MLQA-MLM EM↑	CER↓	WER↓	CED↓	WED↓	PLX↓
URA-LLaMa 70B	0.01 ± 0.00	0.57 ± 0.01	0.61 ± 0.01	543.05 ± 10.96	128.05 ± 2.45	1.08 ± 0.01
URA-LLaMa 13B	0.00 ± 0.00	0.74 ± 0.00	0.80 ± 0.00	707.85 ± 11.62	166.85 ± 2.64	1.16 ± 0.02
URA-LLaMa 7B	0.00 ± 0.00	0.74 ± 0.00	0.84 ± 0.01	744.61 ± 13.18	183.98 ± 3.18	1.25 ± 0.01
LLaMa-2 70B	0.00 ± 0.00	0.91 ± 0.00	0.99 ± 0.00	868.70 ± 10.95	206.50 ± 2.47	1.00 ± 0.00
LLaMa-2 13B	0.00 ± 0.00	0.93 ± 0.00	1.00 ± 0.00	882.26 ± 11.23	208.57 ± 2.52	1.10 ± 0.01
LLaMa-2 7B	0.00 ± 0.00	0.90 ± 0.00	1.01 ± 0.01	862.02 ± 13.18	210.38 ± 3.18	1.25 ± 0.01
Vietcuna 7B	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	951.99 ± 12.37	208.67 ± 2.73	1.48 ± 0.01
Vistral 7B Chat	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	953.39 ± 11.06	208.83 ± 2.43	1.00 ± 0.00
MixSUra	0.00 ± 0.00	0.52 ± 0.00	0.58 ± 0.00	491.52 ± 8.47	121.61 ± 1.94	1.00 ± 0.00
GemSUra	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	953.38 ± 11.57	208.83 ± 2.53	1.39 ± 0.00
Gemini Pro	0.01 ± 0.00	0.50 ± 0.01	0.52 ± 0.01	479.85 ± 11.64	108.14 ± 2.61	–
GPT-3.5	0.00 ± 0.00	0.46 ± 0.01	0.54 ± 0.01	439.53 ± 10.79	111.98 ± 2.44	–
GPT-4	0.04 ± 0.00	0.42 ± 0.01	0.51 ± 0.01	398.50 ± 10.26	106.27 ± 2.39	–

Models	VSEC EM↑	CER↓	WER↓	CED↓	WED↓	PLX↓
URA-LLaMa 70B	0.00 ± 0.00	0.86 ± 0.00	0.99 ± 0.00	114.27 ± 0.57	29.99 ± 0.15	1.09 ± 0.00
URA-LLaMa 13B	0.01 ± 0.00	0.44 ± 0.01	0.54 ± 0.01	58.24 ± 0.77	16.27 ± 0.19	1.26 ± 0.00
URA-LLaMa 7B	0.01 ± 0.00	3.33 ± 0.04	2.90 ± 0.03	442.06 ± 5.66	87.53 ± 0.96	1.33 ± 0.00
LLaMa-2 70B	0.00 ± 0.00	0.86 ± 0.00	1.02 ± 0.00	114.16 ± 0.44	30.86 ± 0.12	1.00 ± 0.00
LLaMa-2 13B	0.00 ± 0.00	1.26 ± 0.01	1.30 ± 0.01	167.03 ± 1.16	39.07 ± 0.23	1.11 ± 0.00
LLaMa-2 7B	0.00 ± 0.00	1.32 ± 0.04	1.34 ± 0.03	176.04 ± 5.66	40.44 ± 0.96	1.14 ± 0.00
Vietcuna 7B	0.01 ± 0.00	1.06 ± 0.01	1.13 ± 0.01	141.33 ± 1.39	34.15 ± 0.33	1.61 ± 0.00
Vistral 7B Chat	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	132.48 ± 0.60	30.08 ± 0.14	1.00 ± 0.00
MixSUra	0.12 ± 0.00	0.20 ± 0.00	0.30 ± 0.00	26.83 ± 0.36	9.16 ± 0.09	1.00 ± 0.00
GemSUra	0.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	132.82 ± 0.56	30.16 ± 0.13	1.36 ± 0.00
Gemini Pro	0.64 ± 0.00	0.11 ± 0.00	0.18 ± 0.00	14.07 ± 0.31	5.50 ± 0.07	–
GPT-3.5	0.02 ± 0.00	0.14 ± 0.00	0.23 ± 0.00	18.59 ± 0.34	6.93 ± 0.09	–
GPT-4	0.60 ± 0.01	0.14 ± 0.00	0.23 ± 0.00	18.17 ± 0.45	6.89 ± 0.12	–

(h) Reasoning

Models	SR - Natural EM↑	F1↑	Equ.↑	SR - Abstract symbol EM↑	F1↑	Equ.↑	MATH EM↑	F1↑	Equ.↑
URA-LLaMa 70B	0.06 ± 0.00	0.34 ± 0.00	0.06 ± 0.00	0.02 ± 0.00	0.24 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.24 ± 0.02
URA-LLaMa 13B	0.01 ± 0.00	0.31 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.24 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.16 ± 0.02
URA-LLaMa 7B	0.00 ± 0.00	0.26 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.17 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.06 ± 0.01
LLaMa-2 70B	0.04 ± 0.00	0.29 ± 0.00	0.04 ± 0.00	0.03 ± 0.00	0.25 ± 0.00	0.03 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.22 ± 0.02
LLaMa-2 13B	0.00 ± 0.00	0.06 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.19 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.13 ± 0.02
LLaMa-2 7B	0.00 ± 0.00	0.04 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.05 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.03 ± 0.01
Vietcuna 7B	0.00 ± 0.00	0.04 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.10 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.02 ± 0.00
Vistral 7B Chat	0.00 ± 0.00	0.06 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.21 ± 0.01
MixSUra	0.02 ± 0.00	0.33 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	0.22 ± 0.00	0.04 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.42 ± 0.02
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
Gemini Pro	0.08 ± 0.00	0.47 ± 0.00	0.08 ± 0.00	0.05 ± 0.00	0.25 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.64 ± 0.00
GPT-3.5	0.21 ± 0.00	0.59 ± 0.00	0.32 ± 0.00	0.09 ± 0.00	0.28 ± 0.00	0.13 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.78 ± 0.02
GPT-4	0.21 ± 0.00	0.59 ± 0.00	0.32 ± 0.00	0.09 ± 0.00	0.28 ± 0.00	0.13 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.78 ± 0.02

```
[INST] Context: {context}
Question: {question}
Answer: [/INST]
```

Medium prompt:

🇻🇳

```
[INST] <<SYS>>
Hãy trả lời câu hỏi bên dưới bằng tiếng Việt
    ↪ với các thông tin được cung cấp trong
    ↪ phần ngữ cảnh. Nếu trong ngữ cảnh khô
    ↪ ng có đủ thông tin, hãy trả lời "Tôi
    ↪ không biết".
<</SYS>>
Ngữ cảnh: {context}
Câu hỏi: {question}
Trả lời: [/INST]
```

🇺🇸

```
[INST] <<SYS>>
Please answer the question below in
    ↪ Vietnamese with the information
    ↪ provided in the context. If there is
    ↪ not enough information in the context,
    ↪  answer "I don't know".
<</SYS>>
Context: {context}
Question: {question}
Answer: [/INST]
```

Normal prompt:

🇻🇳

```
[INST] <<SYS>>
Bạn là một trợ lý hữu dụng sử dụng tiếng Việt
    ↪ , biết tôn trọng và thành thật. Bạn lu
    ↪ ôn luôn trả lời các câu hỏi một cách c
    ↪ ó ích nhiều nhất có thể, nhưng đồng th
    ↪ ời phải an toàn. Câu trả lời của bạn
    ↪ không được bao gồm các ngôn từ độc hại
    ↪ , phân biệt chủng tộc, phân biệt giới
    ↪ tính, nguy hiểm, nội dung vi phạm pháp
    ↪  luật. Làm ơn hãy chắc chắn câu trả lờ
    ↪ i của bạn tự nhiên, tích cực và không
    ↪ thiên vị bất cứ cái gì. Nếu có câu hỏi
    ↪  không hợp lý hoặc không rõ ràng thì h
    ↪ ãy giải thích tại sao thay vì trả lời
    ↪ không đúng sự thật. Nếu bạn không biết
    ↪  câu trả lời thì đừng chia sẻ thông
    ↪ tin sai sự thật.
<</SYS>>
Nhiệm vụ của bạn là dựa vào đoạn văn nằm
    ↪ trong dấu triple backtick, hãy trả lời
    ↪  câu hỏi sau bằng tiếng Việt: {
    ↪ question}
Đoạn văn: ```{context}``` [/INST]
```

🇺🇸

```
[INST] <<SYS>>
You are a helpful, respectful, and honest
    ↪ Vietnamese-speaking assistant. You
    ↪ should always answer questions as
    ↪ helpfully as possible, but at the same
    ↪  time, be safe. Your reply must not
    ↪ include malicious, racist, sexist,
    ↪ dangerous, or illegal content. Please
    ↪ make sure your answers are natural,
    ↪ positive, and unbiased. If the
    ↪ question is unreasonable or unclear,
    ↪ explain why instead of answering with
    ↪ no truth. If you don't know the answer
    ↪  then don't share false information.
<</SYS>>
Your task is to answer the passage in triple
    ↪ backtick based on the passage. the
    ↪ following question in Vietnamese: {
    ↪ question}
Paragraph: ```{context}``` [/INST]
```

G.2 Summarization

Weak prompt:

🇻🇳

```
[INST] Đoạn văn: {document}
Tóm tắt đoạn văn trên: [/INST]
```

🇺🇸

```
[INST] Paragraph: {document}
Summary of the above passage: [/INST]
```

Medium prompt:

🇻🇳

```
[INST] <<SYS>>
Nhiệm vụ của bạn là tóm tắt đoạn văn bản sau,
    ↪  đưa ra câu trả lời là bản tóm tắt:
<</SYS>>
```{document}``` [/INST]
```

🇺🇸

```
[INST] <<SYS>>
Your task is to summarize the following text,
    ↪  giving a summary answer:
<</SYS>>
```{document}``` [/INST]
```

Table 5: Performance under few-shot prompting

(a) Sentiment analysis

Models	VLSP 2016 AC↑	VLSP 2016 F1↑	VLSP 2016 AR↑	VLSP 2016 ECE↓	VLSP 2016 A@10↑	UiT-VSFC AC↑	UiT-VSFC F1↑	UiT-VSFC AR↑	UiT-VSFC ECE↓	UiT-VSFC A@10↑
URA-LLaMa 70B	0.66 ± 0.01	0.49 ± 0.01	0.72 ± 0.01	0.13 ± 0.01	0.77 ± 0.04	0.76 ± 0.01	0.48 ± 0.01	0.81 ± 0.01	0.16 ± 0.01	0.71 ± 0.02
URA-LLaMa 13B	0.59 ± 0.01	0.57 ± 0.01	0.67 ± 0.01	0.08 ± 0.01	0.82 ± 0.04	0.74 ± 0.01	0.52 ± 0.08	0.83 ± 0.01	0.10 ± 0.01	0.87 ± 0.02
URA-LLaMa 7B	0.57 ± 0.02	0.42 ± 0.05	0.69 ± 0.02	0.06 ± 0.02	0.77 ± 0.04	0.72 ± 0.01	0.43 ± 0.01	0.78 ± 0.01	0.13 ± 0.01	0.95 ± 0.03
LLaMa-2 70B	0.53 ± 0.01	0.38 ± 0.01	0.68 ± 0.01	0.34 ± 0.01	0.58 ± 0.05	0.60 ± 0.01	0.40 ± 0.01	0.65 ± 0.01	0.39 ± 0.01	0.25 ± 0.03
LLaMa-2 13B	0.51 ± 0.01	0.41 ± 0.06	0.66 ± 0.01	0.32 ± 0.02	0.80 ± 0.04	0.63 ± 0.01	0.46 ± 0.07	0.71 ± 0.01	0.13 ± 0.01	0.88 ± 0.02
LLaMa-2 7B	0.45 ± 0.01	0.32 ± 0.01	0.59 ± 0.01	0.26 ± 0.02	0.50 ± 0.05	0.50 ± 0.01	0.34 ± 0.01	0.69 ± 0.01	0.23 ± 0.01	0.62 ± 0.03
Vietcuna 7B	0.04 ± 0.01	0.05 ± 0.01	0.45 ± 0.01	0.71 ± 0.01	0.05 ± 0.02	0.03 ± 0.00	0.03 ± 0.00	0.53 ± 0.01	0.50 ± 0.00	0.01 ± 0.00
Vistral 7B Chat	0.28 ± 0.01	0.16 ± 0.01	0.86 ± 0.01	0.36 ± 0.01	0.15 ± 0.03	0.02 ± 0.00	0.07 ± 0.01	0.90 ± 0.01	0.78 ± 0.00	0.00 ± 0.00
MixSUra	0.62 ± 0.02	0.63 ± 0.01	0.59 ± 0.01	0.30 ± 0.01	0.59 ± 0.05	0.74 ± 0.01	0.46 ± 0.01	0.63 ± 0.01	0.23 ± 0.01	0.65 ± 0.03
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.72 ± 0.01	0.70 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.89 ± 0.01	0.81 ± 0.00	0.00 ± 0.00
Gemini Pro	0.67 ± 0.01	0.50 ± 0.01	–	0.34 ± 0.01	0.65 ± 0.05	0.78 ± 0.01	0.49 ± 0.01	–	0.45 ± 0.01	0.82 ± 0.02
GPT-3.5	0.65 ± 0.01	0.59 ± 0.01	–	0.35 ± 0.01	0.54 ± 0.05	0.86 ± 0.01	0.73 ± 0.01	–	0.14 ± 0.01	0.85 ± 0.02
GPT-4	0.75 ± 0.01	0.74 ± 0.01	–	0.25 ± 0.01	0.74 ± 0.04	0.85 ± 0.01	0.53 ± 0.09	–	0.15 ± 0.01	0.87 ± 0.02

(b) Text classification

Models	UiT-VSMEC AC↑	UiT-VSMEC F1↑	UiT-VSMEC AR↑	UiT-VSMEC ECE↓	UiT-VSMEC A@10↑	PhoATIS AC↑	PhoATIS F1↑	PhoATIS AR↑	PhoATIS ECE↓	PhoATIS A@10↑
URA-LLaMa 70B	0.25 ± 0.02	0.15 ± 0.01	0.56 ± 0.01	0.25 ± 0.02	0.37 ± 0.06	0.15 ± 0.01	0.22 ± 0.03	0.83 ± 0.00	0.81 ± 0.01	0.13 ± 0.04
URA-LLaMa 13B	0.32 ± 0.02	0.12 ± 0.01	0.58 ± 0.01	0.22 ± 0.02	0.57 ± 0.07	0.01 ± 0.01	0.06 ± 0.02	0.47 ± 0.00	0.84 ± 0.01	0.00 ± 0.01
URA-LLaMa 7B	0.29 ± 0.02	0.11 ± 0.01	0.60 ± 0.01	0.12 ± 0.02	0.43 ± 0.06	0.06 ± 0.01	0.01 ± 0.00	0.55 ± 0.00	0.24 ± 0.01	0.08 ± 0.03
LLaMa-2 70B	0.24 ± 0.02	0.14 ± 0.01	0.63 ± 0.01	0.40 ± 0.02	0.76 ± 0.06	0.11 ± 0.01	0.08 ± 0.02	0.66 ± 0.01	0.51 ± 0.01	0.06 ± 0.02
LLaMa-2 13B	0.18 ± 0.02	0.08 ± 0.01	0.55 ± 0.01	0.45 ± 0.01	0.49 ± 0.07	0.02 ± 0.01	0.06 ± 0.02	0.57 ± 0.01	0.90 ± 0.01	0.01 ± 0.01
LLaMa-2 7B	0.25 ± 0.02	0.12 ± 0.01	0.57 ± 0.01	0.21 ± 0.02	0.54 ± 0.06	0.03 ± 0.01	0.02 ± 0.01	0.56 ± 0.01	0.54 ± 0.01	0.01 ± 0.01
Vietcuna 7B	0.15 ± 0.01	0.05 ± 0.01	0.46 ± 0.01	0.85 ± 0.01	0.15 ± 0.04	0.04 ± 0.01	0.01 ± 0.00	0.63 ± 0.00	0.21 ± 0.01	0.07 ± 0.03
Vistral 7B Chat	0.00 ± 0.00	0.00 ± 0.00	0.68 ± 0.01	0.38 ± 0.00	0.00 ± 0.00	0.00 ± 0.01	0.01 ± 0.01	0.81 ± 0.01	0.61 ± 0.00	0.00 ± 0.00
MixSUra	0.40 ± 0.02	0.36 ± 0.02	0.72 ± 0.01	0.53 ± 0.02	0.79 ± 0.05	0.81 ± 0.01	0.58 ± 0.03	0.96 ± 0.01	0.14 ± 0.01	0.91 ± 0.04
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.63 ± 0.01	0.56 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.93 ± 0.01	0.68 ± 0.01	0.00 ± 0.00
Gemini Pro	0.48 ± 0.02	0.36 ± 0.02	–	0.33 ± 0.02	0.47 ± 0.05	0.82 ± 0.01	0.69 ± 0.03	–	0.76 ± 0.01	0.70 ± 0.04
GPT-3.5	0.42 ± 0.02	0.40 ± 0.02	–	0.58 ± 0.02	0.29 ± 0.06	0.69 ± 0.02	0.67 ± 0.03	–	0.31 ± 0.02	0.69 ± 0.05
GPT-4	0.49 ± 0.02	0.48 ± 0.02	–	0.51 ± 0.02	0.36 ± 0.06	0.85 ± 0.01	0.78 ± 0.03	–	0.15 ± 0.01	0.88 ± 0.04

(c) Knowledge

Models	ZaloE2E EM↑	ZaloE2E F1↑	ViMMRC AC↑	ViMMRC F1↑	ViMMRC AR↑	ViMMRC ECE↓	ViMMRC A@10↑
URA-LLaMa 70B	0.34 ± 0.02	0.50 ± 0.02	0.78 ± 0.02	0.63 ± 0.03	0.90 ± 0.01	0.13 ± 0.02	0.96 ± 0.03
URA-LLaMa 13B	0.26 ± 0.02	0.40 ± 0.02	0.62 ± 0.02	0.50 ± 0.02	0.69 ± 0.02	0.18 ± 0.02	0.65 ± 0.07
URA-LLaMa 7B	0.01 ± 0.00	0.09 ± 0.02	0.42 ± 0.02	0.33 ± 0.02	0.61 ± 0.02	0.13 ± 0.02	0.38 ± 0.07
LLaMa-2 70B	0.25 ± 0.02	0.40 ± 0.02	0.65 ± 0.02	0.52 ± 0.02	0.79 ± 0.01	0.27 ± 0.02	0.71 ± 0.06
LLaMa-2 13B	0.22 ± 0.02	0.36 ± 0.02	0.58 ± 0.02	0.46 ± 0.02	0.62 ± 0.02	0.28 ± 0.02	0.75 ± 0.06
LLaMa-2 7B	0.07 ± 0.01	0.15 ± 0.01	0.30 ± 0.02	0.23 ± 0.02	0.56 ± 0.02	0.43 ± 0.02	0.16 ± 0.05
Vietcuna 7B	0.13 ± 0.01	0.21 ± 0.01	0.31 ± 0.02	0.18 ± 0.02	0.50 ± 0.00	0.06 ± 0.02	0.37 ± 0.06
Vistral 7B Chat	0.06 ± 0.01	0.16 ± 0.01	0.10 ± 0.01	0.13 ± 0.02	0.96 ± 0.01	0.75 ± 0.01	0.12 ± 0.06
MixSUra	0.19 ± 0.02	0.34 ± 0.02	0.65 ± 0.02	0.64 ± 0.02	0.54 ± 0.02	0.29 ± 0.02	0.65 ± 0.07
GemSUra	0.00 ± 0.00	0.04 ± 0.00	0.37 ± 0.02	0.23 ± 0.01	0.52 ± 0.02	0.12 ± 0.02	0.38 ± 0.07
Gemini Pro	0.46 ± 0.02	0.60 ± 0.02	0.89 ± 0.01	0.71 ± 0.09	–	0.64 ± 0.01	0.88 ± 0.05
GPT-3.5	0.49 ± 0.02	0.64 ± 0.02	0.90 ± 0.01	0.72 ± 0.03	–	0.09 ± 0.01	0.90 ± 0.04
GPT-4	0.49 ± 0.02	0.64 ± 0.02	0.91 ± 0.01	0.73 ± 0.04	–	0.09 ± 0.01	0.88 ± 0.04

(d) Toxicity detection

Models	UiT-ViCTSD AC↑	UiT-ViCTSD F1↑	UiT-ViCTSD AR↑	UiT-ViCTSD ECE↓	UiT-ViCTSD A@10↑	UiT-ViHSD AC↑	UiT-ViHSD F1↑	UiT-ViHSD AR↑	UiT-ViHSD ECE↓	UiT-ViHSD A@10↑
URA-LLaMa 70B	0.44 ± 0.01	0.27 ± 0.01	0.75 ± 0.01	0.52 ± 0.01	0.37 ± 0.02	0.17 ± 0.00	0.15 ± 0.00	0.64 ± 0.01	0.57 ± 0.00	0.27 ± 0.02
URA-LLaMa 13B	0.44 ± 0.01	0.27 ± 0.05	0.67 ± 0.01	0.33 ± 0.01	0.41 ± 0.03	0.26 ± 0.01	0.16 ± 0.00	0.61 ± 0.01	0.42 ± 0.00	0.21 ± 0.02
URA-LLaMa 7B	0.43 ± 0.01	0.40 ± 0.01	0.60 ± 0.01	0.29 ± 0.01	0.71 ± 0.02	0.16 ± 0.00	0.10 ± 0.00	0.67 ± 0.01	0.32 ± 0.00	0.28 ± 0.02
LLaMa-2 70B	0.26 ± 0.01	0.17 ± 0.01	0.67 ± 0.03	0.61 ± 0.02	0.32 ± 0.05	0.15 ± 0.00	0.14 ± 0.00	0.60 ± 0.01	0.72 ± 0.00	0.14 ± 0.01
LLaMa-2 13B	0.28 ± 0.01	0.19 ± 0.00	0.67 ± 0.01	0.52 ± 0.01	0.63 ± 0.03	0.17 ± 0.00	0.11 ± 0.00	0.62 ± 0.01	0.58 ± 0.00	0.44 ± 0.02
LLaMa-2 7B	0.16 ± 0.01	0.12 ± 0.01	0.61 ± 0.01	0.66 ± 0.01	0.08 ± 0.02	0.01 ± 0.00	0.01 ± 0.00	0.56 ± 0.01	0.66 ± 0.00	0.08 ± 0.02
Vietcuna 7B	0.08 ± 0.00	0.10 ± 0.01	0.50 ± 0.00	0.42 ± 0.00	0.08 ± 0.03	0.61 ± 0.01	0.21 ± 0.00	0.50 ± 0.00	0.28 ± 0.01	0.61 ± 0.02
Vistral 7B Chat	0.13 ± 0.01	0.08 ± 0.01	0.78 ± 0.02	0.50 ± 0.01	0.18 ± 0.04	0.15 ± 0.00	0.09 ± 0.00	0.74 ± 0.01	0.39 ± 0.00	0.33 ± 0.02
MixSUra	0.70 ± 0.01	0.39 ± 0.03	0.78 ± 0.02	0.29 ± 0.01	0.80 ± 0.04	0.58 ± 0.01	0.31 ± 0.01	0.68 ± 0.01	0.30 ± 0.01	0.93 ± 0.01
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.68 ± 0.03	0.79 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.72 ± 0.01	0.74 ± 0.00	0.00 ± 0.00
Gemini Pro	0.81 ± 0.01	0.43 ± 0.01	–	0.31 ± 0.01	0.82 ± 0.04	0.70 ± 0.01	0.37 ± 0.01	–	0.36 ± 0.01	0.69 ± 0.01
GPT-3.5	0.63 ± 0.02	0.54 ± 0.02	–	0.37 ± 0.02	0.70 ± 0.05	0.63 ± 0.01	0.47 ± 0.01	–	0.37 ± 0.01	0.63 ± 0.02
GPT-4	0.89 ± 0.00	0.71 ± 0.01	–	0.11 ± 0.00	0.91 ± 0.03	0.77 ± 0.01	0.57 ± 0.01	–	0.23 ± 0.01	0.77 ± 0.02

(e) Information retrieval

Models	mMARCO M@10↑	mMARCO M@10B↑	mMARCO N@10↑	mMARCO N@10B↑	mRobust04 M@10↑	mRobust04 M@10B↑	mRobust04 N@10↑	mRobust04 N@10B↑
URA-LLaMa 70B	0.05 ± 0.00	0.11 ± 0.00	0.06 ± 0.00	0.14 ± 0.00	0.04 ± 0.00	0.04 ± 0.00	0.03 ± 0.00	0.04 ± 0.00
URA-LLaMa 13B	0.04 ± 0.00	0.10 ± 0.00	0.06 ± 0.00	0.14 ± 0.00	0.03 ± 0.00	0.05 ± 0.00	0.04 ± 0.00	0.04 ± 0.00
URA-LLaMa 7B	0.04 ± 0.00	0.11 ± 0.00	0.06 ± 0.00	0.16 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
LLaMa-2 70B	0.03 ± 0.00	0.08 ± 0.00	0.04 ± 0.00	0.11 ± 0.00	0.05 ± 0.00	0.05 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
LLaMa-2 13B	0.07 ± 0.00	0.15 ± 0.00	0.09 ± 0.00	0.21 ± 0.00	0.05 ± 0.00	0.04 ± 0.00	0.04 ± 0.00	0.04 ± 0.00
LLaMa-2 7B	0.05 ± 0.00	0.11 ± 0.00	0.07 ± 0.00	0.16 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.02 ± 0.00
Vietcuna 7B	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Vistral 7B Chat	0.14 ± 0.00	0.30 ± 0.00	0.18 ± 0.00	0.38 ± 0.00	0.03 ± 0.00	0.03 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
MixSUra	0.01 ± 0.00	0.07 ± 0.00	0.04 ± 0.00	0.11 ± 0.00	0.04 ± 0.00	0.04 ± 0.00	0.02 ± 0.00	0.02 ± 0.00
GemSUra	0.03 ± 0.00	0.11 ± 0.00	0.05 ± 0.00	0.15 ± 0.00	0.05 ± 0.01	0.05 ± 0.01	0.02 ± 0.00	0.02 ± 0.00
Gemini Pro	–	–	–	–	–	–	–	–
GPT-3.5	–	–	–	–	–	–	–	–
GPT-4	–	–	–	–	–	–	–	–

(f) Language modeling

Models	MLQA-MLM EM↑	MLQA-MLM CER↓	MLQA-MLM WER↓	MLQA-MLM CED↓	MLQA-MLM WED↓	MLQA-MLM PLX↓	VSEC EM↑	VSEC CER↓	VSEC WER↓	VSEC CED↓	VSEC WED↓	VSEC PLX↓
URA-LLaMa 70B	0.01 ± 0.00	0.69 ± 0.00	0.74 ± 0.00	671.13 ± 10.38	158.36 ± 2.33	1.23 ± 0.05	0.33 ± 0.00	0.14 ± 0.00	0.22 ± 0.00	18.63 ± 0.42	6.63 ± 0.11	1.13 ± 0.00
URA-LLaMa 13B	0.01 ± 0.00	0.60 ± 0.01	0.70 ± 0.01	571.35 ± 11.23	145.40 ± 2.68	1.49 ± 0.10	0.35 ± 0.00	0.04 ± 0.00	0.12 ± 0.00	5.73 ± 0.12	3.74 ± 0.03	1.15 ± 0.00
URA-LLaMa 7B	0.01 ± 0.00	0.54 ± 0.01	0.61 ± 0.01	512.76 ± 11.01	127.97 ± 2.58	1.24 ± 0.01	0.22 ± 0.00	0.36 ± 0.01	0.42 ± 0.01	48.38 ± 1.54	12.75 ± 0.34	1.07 ± 0.00
LLaMa-2 70B	0.00 ± 0.00	0.89 ± 0.00	0.97 ± 0.00	852.19 ± 11.04	203.26 ± 2.53	1.00 ± 0.00	0.01 ± 0.00	0.83 ± 0.00	0.96 ± 0.00	110.36 ± 0.54	28.84 ± 0.14	1.00 ± 0.00
LLaMa-2 13B	0.01 ± 0.00	0.80 ± 0.00	0.91 ± 0.00	767.32 ± 11.91	190.29 ± 2.85	1.24 ± 0.03	0.16 ± 0.00	0.05 ± 0.00	0.14 ± 0.00	6.30 ± 0.16	4.07 ± 0.04	1.01 ± 0.00
LLaMa-2 7B	0.00 ± 0.00	0.82 ± 0.00	0.97 ± 0.00	777.18 ± 10.51	201.81 ± 2.57	1.74 ± 0.19	0.12 ± 0.00	0.38 ± 0.01	0.45 ± 0.01	50.29 ± 0.86	13.58 ± 0.19	1.06 ± 0.00
Vietcuna 7B	0.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	945.07 ± 12.47	206.78 ± 2.79	1.40 ± 0.00	0.00 ± 0.00	8.34 ± 0.07	8.07 ± 0.07	1107.66 ± 7.64	243.52 ± 1.74	1.46 ± 0.00
Vistral 7B Chat	0.00 ± 0.00	0.95 ± 0.00	0.96 ± 0.00	906.01 ± 12.12	200.54 ± 2.63	1.00 ± 0.00	0.01 ± 0.00	2.33 ± 0.04	2.44 ± 0.04	309.72 ± 4.25	73.61 ± 0.96	1.00 ± 0.00
MixSUra	0.00 ± 0.00	0.52 ± 0.00	0.57 ± 0.00	488.41 ± 10.58	115.85 ± 2.50	1.00 ± 0.00	0.08 ± 0.00	0.19 ± 0.00	0.28 ± 0.00	25.13 ± 0.42	8.58 ± 0.10	1.00 ± 0.00
GemSUra	0.00 ± 0.00	0.95 ± 0.00	1.01 ± 0.00	904.32 ± 9.69	211.79 ± 2.13	1.42 ± 0.00	0.00 ± 0.00	1.56 ± 0.01	1.85 ± 0.02	206.68 ± 1.68	55.86 ± 0.41	1.46 ± 0.00
Gemini Pro	0.05 ± 0.00	0.11 ± 0.00	0.19 ± 0.00	103.40 ± 3.93	39.36 ± 0.88	–	0.71 ± 0.01	0.02 ± 0.00	0.03 ± 0.00	2.77 ± 0.45	0.87 ± 0.10	–
GPT-3.5	0.04 ± 0.00	0.42 ± 0.01	0.50 ± 0.01	402.97 ± 10.86	104.78 ± 2.46	–	0.66 ± 0.00	0.03 ± 0.00	0.11 ± 0.00	4.57 ± 0.08	3.22 ± 0.02	–
GPT-4	0.08 ± 0.00	0.37 ± 0.01	0.46 ± 0.01	353.24 ± 10.18	96.33 ± 2.34	–	0.75 ± 0.00	0.03 ± 0.00	0.10 ± 0.00	3.85 ± 0.04	3.01 ± 0.01	–

Normal prompt: 🇻🇳

(g) Reasoning

Models	SR - Natural			SR - Abstract symbol			MATH		
	EM↑	F1↑	Equ.↑	EM↑	F1↑	Equ.↑	EM↑	F1↑	Equ.↑
URA-LLaMa 70B	0.14 ± 0.00	0.48 ± 0.00	0.15 ± 0.00	0.27 ± 0.00	0.85 ± 0.00	0.30 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.24 ± 0.02
URA-LLaMa 13B	0.08 ± 0.00	0.42 ± 0.00	0.08 ± 0.00	0.20 ± 0.00	0.70 ± 0.00	0.17 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.16 ± 0.01
URA-LLaMa 7B	0.04 ± 0.00	0.38 ± 0.00	0.04 ± 0.00	0.11 ± 0.00	0.61 ± 0.00	0.10 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.08 ± 0.01
LLaMa-2 70B	0.13 ± 0.00	0.48 ± 0.00	0.13 ± 0.00	0.26 ± 0.00	0.84 ± 0.00	0.27 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.29 ± 0.02
LLaMa-2 13B	0.03 ± 0.00	0.24 ± 0.00	0.04 ± 0.00	0.19 ± 0.00	0.69 ± 0.00	0.18 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.19 ± 0.02
LLaMa-2 7B	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.06 ± 0.00	0.44 ± 0.00	0.06 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.13 ± 0.01
Vietcuna 7B	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.14 ± 0.00	0.71 ± 0.00	0.10 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
Vistral 7B Chat	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.16 ± 0.01
MixSUra	0.07 ± 0.00	0.41 ± 0.00	0.07 ± 0.00	0.22 ± 0.00	0.78 ± 0.00	0.23 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.42 ± 0.02
GemSUra	0.00 ± 0.00	0.05 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.06 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.21 ± 0.01
Gemini Pro	0.15 ± 0.00	0.50 ± 0.00	0.16 ± 0.00	0.26 ± 0.00	0.83 ± 0.00	0.29 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.40 ± 0.02
GPT-3.5	0.15 ± 0.00	0.50 ± 0.00	0.16 ± 0.00	0.26 ± 0.00	0.83 ± 0.00	0.29 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.62 ± 0.02
GPT-4	0.37 ± 0.00	0.74 ± 0.00	0.42 ± 0.00	0.37 ± 0.00	0.87 ± 0.00	0.44 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.65 ± 0.02

(h) Translation

Models	PhoMT				OPUS100			
	(En → Vi)		(Vi → En)		(En → Vi)		(Vi → En)	
	BLEU↑	hLEPOR↑	BLEU↑	hLEPOR↑	BLEU↑	hLEPOR↑	BLEU↑	hLEPOR↑
URA-LLaMa 70B	0.22 ± 0.00	0.58 ± 0.00	0.19 ± 0.00	0.56 ± 0.00	0.08 ± 0.00	0.41 ± 0.01	0.09 ± 0.00	0.37 ± 0.01
URA-LLaMa 13B	0.20 ± 0.00	0.54 ± 0.00	0.19 ± 0.00	0.54 ± 0.00	0.07 ± 0.01	0.37 ± 0.01	0.11 ± 0.01	0.39 ± 0.01
URA-LLaMa 7B	0.15 ± 0.00	0.49 ± 0.00	0.16 ± 0.00	0.52 ± 0.00	0.06 ± 0.00	0.36 ± 0.01	0.09 ± 0.01	0.36 ± 0.01
LLaMa-2 70B	0.27 ± 0.00	0.57 ± 0.00	0.17 ± 0.00	0.52 ± 0.00	0.11 ± 0.00	0.42 ± 0.01	0.08 ± 0.00	0.34 ± 0.01
LLaMa-2 13B	0.18 ± 0.00	0.52 ± 0.00	0.17 ± 0.00	0.52 ± 0.00	0.07 ± 0.00	0.37 ± 0.01	0.09 ± 0.01	0.36 ± 0.01
LLaMa-2 7B	0.14 ± 0.00	0.46 ± 0.00	0.15 ± 0.00	0.51 ± 0.00	0.05 ± 0.00	0.32 ± 0.00	0.07 ± 0.01	0.33 ± 0.01
Vietcuna 7B	0.11 ± 0.00	0.34 ± 0.00	0.01 ± 0.00	0.11 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.05 ± 0.00	0.15 ± 0.00
Vistral 7B Chat	0.25 ± 0.00	0.62 ± 0.00	0.21 ± 0.00	0.58 ± 0.00	0.11 ± 0.00	0.44 ± 0.01	0.11 ± 0.01	0.42 ± 0.01
MixSUra	0.15 ± 0.00	0.51 ± 0.00	0.16 ± 0.00	0.52 ± 0.00	0.07 ± 0.00	0.37 ± 0.01	0.09 ± 0.00	0.36 ± 0.01
GemSUra	0.02 ± 0.00	0.12 ± 0.00	0.01 ± 0.00	0.10 ± 0.00	0.02 ± 0.00	0.18 ± 0.00	0.01 ± 0.00	0.11 ± 0.00
Gemini Pro	0.27 ± 0.00	0.60 ± 0.00	0.24 ± 0.01	0.55 ± 0.00	0.06 ± 0.01	0.39 ± 0.01	0.13 ± 0.01	0.36 ± 0.01
GPT-3.5	0.33 ± 0.00	0.65 ± 0.00	0.24 ± 0.00	0.61 ± 0.00	0.12 ± 0.01	0.46 ± 0.01	0.15 ± 0.01	0.46 ± 0.00
GPT-4	0.26 ± 0.00	0.64 ± 0.00	0.25 ± 0.00	0.62 ± 0.00	0.13 ± 0.01	0.47 ± 0.01	0.16 ± 0.01	0.47 ± 0.00

Table 6: Performance on Reasoning - MATH under Chain-of-Thought prompting

Models	EM↑	F1↑	Equ.↑
URA-LLaMa 70B	0.00 ± 0.00	0.15 ± 0.01	0.26 ± 0.02
URA-LLaMa 13B	0.00 ± 0.00	0.16 ± 0.01	0.12 ± 0.01
URA-LLaMa 7B	0.00 ± 0.00	0.19 ± 0.01	0.07 ± 0.01
LLaMa-2 70B	0.00 ± 0.00	0.11 ± 0.01	0.28 ± 0.02
LLaMa-2 13B	0.00 ± 0.00	0.12 ± 0.01	0.18 ± 0.02
LLaMa-2 7B	0.00 ± 0.00	0.10 ± 0.00	0.12 ± 0.02
Vietcuna 7B	0.00 ± 0.00	0.02 ± 0.00	0.01 ± 0.00
Vistral 7B Chat	0.00 ± 0.00	0.08 ± 0.00	0.11 ± 0.01
MixSUra	0.00 ± 0.00	0.18 ± 0.01	0.41 ± 0.02
GemSUra	0.00 ± 0.00	0.25 ± 0.00	0.32 ± 0.02
Gemini Pro	0.00 ± 0.00	0.27 ± 0.01	0.61 ± 0.01
GPT-3.5	0.00 ± 0.00	0.29 ± 0.01	0.77 ± 0.02
GPT-4	0.00 ± 0.00	0.30 ± 0.01	0.71 ± 0.02

Table 7: Performance under weaker zero-shot prompting

(a) Question-answering - weak prompting

Models	XQuAD		MLQA	
	EM↑	F1↑	EM↑	F1↑
URA-LLaMa 70B	0.21 ± 0.01	0.47 ± 0.01	0.14 ± 0.01	0.41 ± 0.00
URA-LLaMa 13B	0.22 ± 0.01	0.43 ± 0.01	0.17 ± 0.01	0.40 ± 0.01
URA-LLaMa 7B	0.13 ± 0.00	0.32 ± 0.00	0.10 ± 0.00	0.32 ± 0.00
LLaMa-2 70B	0.13 ± 0.00	0.38 ± 0.01	0.09 ± 0.00	0.36 ± 0.00
LLaMa-2 13B	0.04 ± 0.00	0.28 ± 0.00	0.04 ± 0.00	0.28 ± 0.00
LLaMa-2 7B	0.06 ± 0.00	0.24 ± 0.00	0.05 ± 0.00	0.24 ± 0.00
Vistral 7B Chat	0.32 ± 0.01	0.56 ± 0.01	0.21 ± 0.01	0.46 ± 0.01
MixSUra	0.13 ± 0.00	0.38 ± 0.01	0.09 ± 0.00	0.36 ± 0.00
GemSUra	0.05 ± 0.01	0.14 ± 0.01	0.04 ± 0.00	0.11 ± 0.00

(b) Question-Answering - medium prompting

Models	XQuAD		MLQA	
	EM↑	F1↑	EM↑	F1↑
URA-LLaMa 70B	0.08 ± 0.00	0.33 ± 0.00	0.07 ± 0.00	0.31 ± 0.00
URA-LLaMa 13B	0.04 ± 0.00	0.21 ± 0.00	0.04 ± 0.00	0.19 ± 0.00
URA-LLaMa 7B	0.01 ± 0.00	0.11 ± 0.00	0.01 ± 0.00	0.11 ± 0.00
LLaMa-2 70B	0.00 ± 0.00	0.17 ± 0.00	0.00 ± 0.00	0.17 ± 0.00
LLaMa-2 13B	0.00 ± 0.00	0.10 ± 0.00	0.00 ± 0.00	0.09 ± 0.00
LLaMa-2 7B	0.00 ± 0.00	0.03 ± 0.00	0.00 ± 0.00	0.03 ± 0.00
Vistral 7B Chat	0.03 ± 0.01	0.07 ± 0.01	0.05 ± 0.00	0.09 ± 0.00
MixSUra	0.01 ± 0.00	0.25 ± 0.01	0.00 ± 0.00	0.25 ± 0.00
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

(c) Summarization - weak prompting

Models	VietNews							
	R1↑	R2↑	RL↑	SC↑	BS↑	Cv↑	De↑	Cp↑
URA-LLaMa 70B	0.49 ± 0.00	0.23 ± 0.00	0.31 ± 0.00	0.58 ± 0.00	0.05 ± 0.11	0.89 ± 0.00	8.90 ± 0.03	18.48 ± 0.59
URA-LLaMa 13B	0.27 ± 0.00	0.12 ± 0.00	0.18 ± 0.00	0.31 ± 0.00	0.05 ± 0.11	0.56 ± 0.00	5.00 ± 0.04	153.55 ± 0.99
URA-LLaMa 7B	0.45 ± 0.00	0.21 ± 0.00	0.39 ± 0.00	0.26 ± 0.00	0.03 ± 0.09	0.91 ± 0.00	9.43 ± 0.03	6.42 ± 0.05
LLaMa-2 70B	0.34 ± 0.00	0.17 ± 0.00	0.22 ± 0.00	0.39 ± 0.00	−0.04 ± 0.15	0.71 ± 0.00	7.26 ± 0.04	18.42 ± 0.69
LLaMa-2 13B	0.45 ± 0.00	0.22 ± 0.00	0.29 ± 0.00	0.53 ± 0.00	0.00 ± 0.14	0.92 ± 0.00	9.49 ± 0.02	8.46 ± 0.29
LLaMa-2 7B	0.36 ± 0.00	0.17 ± 0.00	0.23 ± 0.00	0.33 ± 0.00	−0.15 ± 0.12	0.69 ± 0.00	6.35 ± 0.03	7.59 ± 0.21
Vistral 7B Chat	0.45 ± 0.00	0.22 ± 0.00	0.28 ± 0.00	0.28 ± 0.00	0.03 ± 0.10	0.72 ± 0.00	6.32 ± 0.10	6.49 ± 0.03
MixSUra	0.44 ± 0.00	0.22 ± 0.00	0.29 ± 0.00	0.74 ± 0.00	0.07 ± 0.08	0.97 ± 0.00	35.67 ± 0.13	9.43 ± 0.26
GemSUra	0.37 ± 0.00	0.09 ± 0.00	0.22 ± 0.00	−0.09 ± 0.00	0.07 ± 0.04	0.10 ± 0.00	0.12 ± 0.00	8.16 ± 0.12

Models	WikiLingua							
	R1↑	R2↑	RL↑	SC↑	BS↑	Cv↑	De↑	Cp↑
URA-LLaMa 70B	0.47 ± 0.00	0.20 ± 0.00	0.29 ± 0.00	0.48 ± 0.00	0.19 ± 0.13	0.86 ± 0.00	6.83 ± 0.09	25.30 ± 1.86
URA-LLaMa 13B	0.22 ± 0.00	0.09 ± 0.00	0.14 ± 0.00	0.22 ± 0.00	0.20 ± 0.007	0.48 ± 0.00	3.49 ± 0.04	190.09 ± 4.92
URA-LLaMa 7B	0.42 ± 0.00	0.18 ± 0.00	0.27 ± 0.00	0.32 ± 0.00	0.07 ± 0.12	0.89 ± 0.00	7.58 ± 0.05	7.14 ± 0.14
LLaMa-2 70B	0.27 ± 0.00	0.12 ± 0.00	0.17 ± 0.00	0.29 ± 0.00	0.05 ± 0.13	0.58 ± 0.01	8.11 ± 0.17	21.64 ± 1.67
LLaMa-2 13B	0.47 ± 0.00	0.22 ± 0.00	0.29 ± 0.00	0.53 ± 0.00	0.34 ± 0.12	0.92 ± 0.00	9.39 ± 0.05	17.94 ± 2.84
LLaMa-2 7B	0.45 ± 0.00	0.20 ± 0.00	0.27 ± 0.00	0.48 ± 0.00	0.36 ± 0.00	0.83 ± 0.00	7.71 ± 0.07	12.39 ± 1.46
Vistral 7B Chat	0.43 ± 0.00	0.21 ± 0.00	0.28 ± 0.00	0.04 ± 0.00	0.25 ± 0.09	0.65 ± 0.00	2.37 ± 0.06	4.57 ± 0.04
MixSUra	0.47 ± 0.00	0.22 ± 0.00	0.29 ± 0.00	0.14 ± 0.00	0.19 ± 0.08	0.97 ± 0.00	28.97 ± 0.30	10.27 ± 0.91
GemSUra	0.40 ± 0.00	0.09 ± 0.00	0.23 ± 0.00	−0.09 ± 0.00	−0.24 ± 0.07	0.06 ± 0.00	0.07 ± 0.00	10.10 ± 0.25

(d) Summarization - medium prompting

Models	VietNews							
	R1↑	R2↑	RL↑	SC↑	BS↑	Cv↑	De↑	Cp↑
URA-LLaMa 70B	0.35 ± 0.00	0.16 ± 0.00	0.24 ± 0.00	0.34 ± 0.00	0.12 ± 0.00	0.63 ± 0.00	5.43 ± 0.02	37.78 ± 0.47
URA-LLaMa 13B	0.26 ± 0.00	0.12 ± 0.00	0.17 ± 0.00	0.22 ± 0.00	−0.08 ± 0.18	0.46 ± 0.00	3.55 ± 0.04	47.75 ± 0.65
URA-LLaMa 7B	0.41 ± 0.00	0.18 ± 0.00	0.27 ± 0.00	0.36 ± 0.00	−0.08 ± 0.13	0.83 ± 0.00	8.13 ± 0.04	8.08 ± 0.17
LLaMa-2 70B	0.09 ± 0.00	0.03 ± 0.00	0.07 ± 0.00	0.03 ± 0.00	−0.20 ± 0.11	0.15 ± 0.00	1.07 ± 0.02	19.69 ± 0.27
LLaMa-2 13B	0.02 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	−0.05 ± 0.00	−0.19 ± 0.05	0.01 ± 0.00	0.01 ± 0.00	54.67 ± 0.16
LLaMa-2 7B	0.03 ± 0.00	0.01 ± 0.00	0.03 ± 0.00	−0.04 ± 0.00	−0.17 ± 0.03	0.04 ± 0.00	0.07 ± 0.00	23.86 ± 0.26
Vistral 7B Chat	0.04 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	−0.19 ± 0.10	0.98 ± 0.00	1.34 ± 0.01	518.90 ± 1.06
MixSUra	0.06 ± 0.00	0.01 ± 0.00	0.04 ± 0.00	−0.02 ± 0.00	−0.13 ± 0.05	0.10 ± 0.00	0.17 ± 0.00	9.03 ± 0.54
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	−0.19 ± 0.05	1.00 ± 0.00	1.00 ± 0.00	550.22 ± 2.60

Models	WikiLingua							
	R1↑	R2↑	RL↑	SC↑	BS↑	Cv↑	De↑	Cp↑
URA-LLaMa 70B	0.33 ± 0.00	0.14 ± 0.00	0.22 ± 0.00	0.30 ± 0.00	0.24 ± 0.10	0.59 ± 0.01	4.62 ± 0.11	56.56 ± 1.70
URA-LLaMa 13B	0.14 ± 0.00	0.05 ± 0.00	0.09 ± 0.00	0.10 ± 0.00	−0.14 ± 0.12	0.26 ± 0.01	1.83 ± 0.06	60.10 ± 2.16
URA-LLaMa 7B	0.42 ± 0.00	0.17 ± 0.00	0.27 ± 0.00	0.35 ± 0.00	0.27 ± 0.21	0.84 ± 0.00	7.15 ± 0.08	8.08 ± 0.36
LLaMa-2 70B	0.03 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	−0.02 ± 0.00	−0.22 ± 0.11	0.10 ± 0.00	0.42 ± 0.06	37.50 ± 0.61
LLaMa-2 13B	0.03 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	−0.05 ± 0.00	−0.05 ± 0.03	0.02 ± 0.00	0.02 ± 0.00	42.55 ± 0.81
LLaMa-2 7B	0.02 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	−0.02 ± 0.00	−0.04 ± 0.06	0.02 ± 0.00	0.03 ± 0.00	40.31 ± 0.88
Vistral 7B Chat	0.18 ± 0.00	0.09 ± 0.00	0.12 ± 0.00	0.03 ± 0.00	−0.19 ± 0.18	0.87 ± 0.00	1.60 ± 0.01	347.61 ± 6.29
MixSUra	0.03 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	0.04 ± 0.00	−0.01 ± 0.05	0.17 ± 0.00	0.26 ± 0.01	16.68 ± 1.94
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	−0.19 ± 0.03	1.00 ± 0.00	1.00 ± 0.00	589.41 ± 5.60

```
[INST] <<SYS>>
Bạn là một trợ lý hữu dụng, biết tôn trọng và
↪ thành thật. Bạn luôn luôn trả lời các
↪ câu hỏi một cách có ích nhiều nhất có
↪ thể, nhưng đồng thời phải an toàn. Câ
↪ u trả lời của bạn không được bao gồm c
↪ ác ngôn từ độc hại, phân biệt chủng tộ
↪ c, phân biệt giới tính, nguy hiểm, nội
↪ dung vi phạm pháp luật. Nhiệm vụ của
↪ bạn là tóm tắt đoạn văn bản nằm trong
↪ triple backtick. Bài tóm tắt phải đầy
↪ đủ các thông tin quan trọng, ngắn gọn
↪ và thu hút người đọc. Ngôn ngữ bạn phả
↪ i sử dụng để tóm tắt là tiếng Việt.
<</SYS>>
```{document}``` [/INST]
```
```

Table 8: Fairness performance

(a) Question-Answering

Models	XQuAD Exact Match↑	XQuAD F1↑	MLQA Exact Match↑	MLQA F1↑
URA-LLaMa 70B	0.04 ± 0.00	0.28 ± 0.00	0.03 ± 0.00	0.26 ± 0.00
URA-LLaMa 13B	0.00 ± 0.00	0.14 ± 0.00	0.00 ± 0.00	0.15 ± 0.00
URA-LLaMa 7B	0.00 ± 0.00	0.13 ± 0.00	0.00 ± 0.00	0.15 ± 0.01
LLaMa-2 70B	0.00 ± 0.00	0.10 ± 0.00	0.00 ± 0.00	0.11 ± 0.00
LLaMa-2 13B	0.00 ± 0.00	0.03 ± 0.00	0.00 ± 0.00	0.04 ± 0.00
LLaMa-2 7B	0.00 ± 0.00	0.04 ± 0.00	0.00 ± 0.00	0.05 ± 0.00
Vietcuna 7B	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Vistral 7B Chat	0.01 ± 0.00	0.03 ± 0.00	0.01 ± 0.00	0.02 ± 0.00
MixSUra	0.00 ± 0.00	0.16 ± 0.00	0.00 ± 0.00	0.17 ± 0.00
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Gemini Pro	0.13 ± 0.01	0.31 ± 0.01	0.09 ± 0.00	0.27 ± 0.00
GPT-3.5	0.00 ± 0.00	0.24 ± 0.00	0.00 ± 0.00	0.23 ± 0.00
GPT-4	0.00 ± 0.00	0.26 ± 0.00	0.00 ± 0.00	0.24 ± 0.00

(b) Sentiment analysis

Models	VLSP 2016 AC↑	VLSP 2016 F1↑	VLSP 2016 AR↑	VLSP 2016 ECE↓	VLSP 2016 A@10↑	UiT-VSFC AC↑	UiT-VSFC F1↑	UiT-VSFC AR↑	UiT-VSFC ECE↓	UiT-VSFC A@10↑
URA-LLaMa 70B	0.65 ± 0.01	0.49 ± 0.01	0.72 ± 0.01	0.13 ± 0.01	0.77 ± 0.04	0.75 ± 0.01	0.48 ± 0.01	0.81 ± 0.01	0.17 ± 0.01	0.66 ± 0.03
URA-LLaMa 13B	0.58 ± 0.01	0.57 ± 0.01	0.67 ± 0.01	0.07 ± 0.01	0.83 ± 0.04	0.75 ± 0.01	0.46 ± 0.08	0.83 ± 0.01	0.11 ± 0.01	0.88 ± 0.02
URA-LLaMa 7B	0.74 ± 0.02	0.39 ± 0.06	0.83 ± 0.01	0.21 ± 0.02	0.98 ± 0.02	0.73 ± 0.01	0.43 ± 0.01	0.78 ± 0.01	0.13 ± 0.01	0.94 ± 0.01
LLaMa-2 70B	0.52 ± 0.02	0.38 ± 0.01	0.68 ± 0.01	0.34 ± 0.02	0.58 ± 0.05	0.60 ± 0.01	0.40 ± 0.01	0.65 ± 0.01	0.39 ± 0.01	0.28 ± 0.02
LLaMa-2 13B	0.51 ± 0.01	0.36 ± 0.06	0.66 ± 0.01	0.32 ± 0.02	0.79 ± 0.04	0.63 ± 0.01	0.41 ± 0.02	0.70 ± 0.01	0.13 ± 0.01	0.89 ± 0.02
LLaMa-2 7B	0.45 ± 0.02	0.34 ± 0.01	0.59 ± 0.01	0.26 ± 0.02	0.50 ± 0.0	0.51 ± 0.01	0.35 ± 0.01	0.69 ± 0.01	0.22 ± 0.01	0.64 ± 0.03
Vietcuna 7B	0.04 ± 0.01	0.04 ± 0.01	0.45 ± 0.01	0.71 ± 0.01	0.05 ± 0.02	0.03 ± 0.00	0.03 ± 0.00	0.55 ± 0.01	0.50 ± 0.00	0.01 ± 0.00
Vistral 7B Chat	0.28 ± 0.02	0.16 ± 0.01	0.86 ± 0.01	0.36 ± 0.02	0.16 ± 0.03	0.02 ± 0.00	0.07 ± 0.01	0.90 ± 0.00	0.77 ± 0.00	0.00 ± 0.00
MixSUra	0.62 ± 0.02	0.62 ± 0.02	0.59 ± 0.01	0.30 ± 0.01	0.59 ± 0.05	0.74 ± 0.01	0.46 ± 0.01	0.61 ± 0.01	0.24 ± 0.01	0.66 ± 0.03
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.73 ± 0.01	0.70 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.89 ± 0.01	0.81 ± 0.00	0.00 ± 0.00
Gemini Pro	0.67 ± 0.02	0.50 ± 0.01	–	0.34 ± 0.02	0.59 ± 0.05	0.79 ± 0.01	0.50 ± 0.01	–	0.46 ± 0.01	0.82 ± 0.02
GPT-3.5	0.66 ± 0.01	0.60 ± 0.01	–	0.35 ± 0.01	0.52 ± 0.05	0.86 ± 0.01	0.71 ± 0.01	–	0.14 ± 0.01	0.86 ± 0.02
GPT-4	0.75 ± 0.01	0.74 ± 0.01	–	0.25 ± 0.00	0.73 ± 0.04	0.85 ± 0.01	0.71 ± 0.01	–	0.15 ± 0.01	0.87 ± 0.02

(c) Text classification

Models	UiT-VSMEC AC↑	UiT-VSMEC F1↑	UiT-VSMEC AR↑	UiT-VSMEC ECE↓	UiT-VSMEC A@10↑	PhoATIS AC↑	PhoATIS F1↑	PhoATIS AR↑	PhoATIS ECE↓	PhoATIS A@10↑
URA-LLaMa 70B	0.24 ± 0.02	0.14 ± 0.01	0.55 ± 0.01	0.26 ± 0.02	0.37 ± 0.06	0.15 ± 0.01	0.22 ± 0.03	0.83 ± 0.00	0.81 ± 0.01	0.13 ± 0.04
URA-LLaMa 13B	0.31 ± 0.02	0.11 ± 0.01	0.58 ± 0.01	0.23 ± 0.02	0.57 ± 0.06	0.01 ± 0.01	0.06 ± 0.02	0.47 ± 0.00	0.84 ± 0.01	0.00 ± 0.01
URA-LLaMa 7B	0.29 ± 0.02	0.10 ± 0.01	0.60 ± 0.01	0.12 ± 0.02	0.41 ± 0.06	0.06 ± 0.01	0.01 ± 0.00	0.55 ± 0.00	0.24 ± 0.01	0.08 ± 0.03
LLaMa-2 70B	0.23 ± 0.02	0.14 ± 0.01	0.63 ± 0.01	0.40 ± 0.02	0.73 ± 0.06	0.11 ± 0.01	0.08 ± 0.01	0.66 ± 0.01	0.51 ± 0.01	0.06 ± 0.03
LLaMa-2 13B	0.18 ± 0.02	0.08 ± 0.01	0.55 ± 0.01	0.45 ± 0.01	0.44 ± 0.06	0.02 ± 0.01	0.06 ± 0.02	0.57 ± 0.01	0.90 ± 0.01	0.01 ± 0.01
LLaMa-2 7B	0.25 ± 0.02	0.11 ± 0.01	0.57 ± 0.01	0.22 ± 0.02	0.53 ± 0.06	0.02 ± 0.00	0.02 ± 0.01	0.60 ± 0.01	0.68 ± 0.01	0.01 ± 0.01
Vietcuna 7B	0.15 ± 0.01	0.05 ± 0.01	0.51 ± 0.01	0.85 ± 0.01	0.16 ± 0.04	0.04 ± 0.01	0.01 ± 0.00	0.64 ± 0.01	0.21 ± 0.01	0.07 ± 0.03
Vistral 7B Chat	0.00 ± 0.00	0.00 ± 0.00	0.69 ± 0.01	0.38 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.01	0.81 ± 0.01	0.61 ± 0.01	0.00 ± 0.00
MixSUra	0.41 ± 0.02	0.32 ± 0.03	0.72 ± 0.01	0.53 ± 0.02	0.79 ± 0.05	0.81 ± 0.02	0.58 ± 0.02	0.96 ± 0.01	0.14 ± 0.01	0.91 ± 0.04
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.64 ± 0.01	0.57 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.93 ± 0.01	0.68 ± 0.01	0.00 ± 0.00
Gemini Pro	0.48 ± 0.02	0.38 ± 0.02	–	0.34 ± 0.02	0.43 ± 0.06	0.79 ± 0.01	0.67 ± 0.02	–	0.73 ± 0.01	0.68 ± 0.04
GPT-3.5	0.44 ± 0.02	0.42 ± 0.02	–	0.56 ± 0.02	0.36 ± 0.06	0.68 ± 0.02	0.66 ± 0.03	–	0.32 ± 0.02	0.67 ± 0.05
GPT-4	0.49 ± 0.02	0.47 ± 0.02	–	0.51 ± 0.02	0.36 ± 0.06	0.83 ± 0.01	0.76 ± 0.03	–	0.17 ± 0.01	0.87 ± 0.04

(d) Toxicity detection

Models	UiT-ViCTSD AC↑	UiT-ViCTSD F1↑	UiT-ViCTSD AR↑	UiT-ViCTSD ECE↓	UiT-ViCTSD A@10↑	UiT-ViHSD AC↑	UiT-ViHSD F1↑	UiT-ViHSD AR↑	UiT-ViHSD ECE↓	UiT-ViHSD A@10↑
URA-LLaMa 70B	0.41 ± 0.02	0.26 ± 0.01	0.75 ± 0.01	0.53 ± 0.01	0.33 ± 0.05	0.15 ± 0.00	0.14 ± 0.00	0.64 ± 0.01	0.58 ± 0.00	0.24 ± 0.02
URA-LLaMa 13B	0.43 ± 0.02	0.27 ± 0.07	0.66 ± 0.01	0.36 ± 0.02	0.42 ± 0.05	0.24 ± 0.01	0.15 ± 0.00	0.61 ± 0.01	0.43 ± 0.01	0.21 ± 0.02
URA-LLaMa 7B	0.42 ± 0.02	0.39 ± 0.01	0.60 ± 0.01	0.30 ± 0.01	0.66 ± 0.05	0.16 ± 0.00	0.10 ± 0.00	0.67 ± 0.01	0.33 ± 0.00	0.28 ± 0.02
LLaMa-2 70B	0.24 ± 0.01	0.16 ± 0.01	0.68 ± 0.03	0.63 ± 0.01	0.32 ± 0.05	0.14 ± 0.00	0.14 ± 0.00	0.60 ± 0.01	0.72 ± 0.00	0.14 ± 0.01
LLaMa-2 13B	0.27 ± 0.01	0.18 ± 0.01	0.67 ± 0.01	0.53 ± 0.01	0.57 ± 0.05	0.16 ± 0.00	0.10 ± 0.00	0.62 ± 0.01	0.59 ± 0.00	0.42 ± 0.02
LLaMa-2 7B	0.15 ± 0.01	0.11 ± 0.01	0.62 ± 0.01	0.67 ± 0.01	0.07 ± 0.03	0.01 ± 0.00	0.01 ± 0.00	0.56 ± 0.01	0.71 ± 0.00	0.01 ± 0.00
Vietcuna 7B	0.08 ± 0.01	0.09 ± 0.01	0.50 ± 0.01	0.42 ± 0.01	0.06 ± 0.03	0.62 ± 0.01	0.21 ± 0.00	0.50 ± 0.00	0.29 ± 0.01	0.62 ± 0.02
Vistral 7B Chat	0.12 ± 0.01	0.08 ± 0.01	0.79 ± 0.02	0.50 ± 0.01	0.16 ± 0.04	0.15 ± 0.00	0.08 ± 0.00	0.74 ± 0.01	0.39 ± 0.00	0.33 ± 0.02
MixSUra	0.69 ± 0.01	0.38 ± 0.02	0.78 ± 0.02	0.29 ± 0.01	0.78 ± 0.03	0.56 ± 0.01	0.31 ± 0.01	0.68 ± 0.01	0.32 ± 0.01	0.92 ± 0.01
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.71 ± 0.02	0.80 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.72 ± 0.01	0.74 ± 0.00	0.00 ± 0.00
Gemini Pro	0.81 ± 0.01	0.44 ± 0.03	–	0.31 ± 0.01	0.82 ± 0.04	0.68 ± 0.01	0.37 ± 0.01	–	0.35 ± 0.01	0.67 ± 0.02
GPT-3.5	0.61 ± 0.02	0.52 ± 0.02	–	0.40 ± 0.02	0.63 ± 0.05	0.61 ± 0.01	0.46 ± 0.01	–	0.39 ± 0.01	0.62 ± 0.02
GPT-4	0.87 ± 0.01	0.69 ± 0.02	–	0.13 ± 0.01	0.86 ± 0.03	0.76 ± 0.01	0.56 ± 0.01	–	0.24 ± 0.01	0.76 ± 0.02

(e) Language modeling

Models	MLQA-MLM EM↑	MLQA-MLM CER↓	MLQA-MLM WER↓	MLQA-MLM CED↓	MLQA-MLM WED↓	MLQA-MLM PLX↓	VSEC EM↑	VSEC CER↓	VSEC WER↓	VSEC CED↓	VSEC WED↓	VSEC PLX↓
URA-LLaMa 70B	0.01 ± 0.00	0.69 ± 0.01	0.74 ± 0.01	663.29 ± 12.05	157.60 ± 2.73	1.25 ± 0.06	0.30 ± 0.00	0.14 ± 0.00	0.22 ± 0.00	18.69 ± 0.42	6.67 ± 0.11	1.13 ± 0.00
URA-LLaMa 13B	0.02 ± 0.00	0.56 ± 0.01	0.64 ± 0.01	531.42 ± 11.19	134.78 ± 2.66	1.48 ± 0.11	0.32 ± 0.00	0.04 ± 0.00	0.13 ± 0.00	5.90 ± 0.11	3.79 ± 0.03	1.15 ± 0.00
URA-LLaMa 7B	0.01 ± 0.00	0.53 ± 0.01	0.60 ± 0.01	508.49 ± 11.32	127.77 ± 2.72	1.22 ± 0.01	0.20 ± 0.00	0.36 ± 0.01	0.42 ± 0.01	47.81 ± 1.57	12.77 ± 0.35	1.07 ± 0.00
LLaMa-2 70B	0.00 ± 0.00	0.90 ± 0.00	0.98 ± 0.00	858.96 ± 10.86	206.70 ± 2.53	1.00 ± 0.00	0.01 ± 0.00	0.84 ± 0.00	0.96 ± 0.00	111.58 ± 0.56	29.08 ± 0.14	1.00 ± 0.00
LLaMa-2 13B	0.01 ± 0.00	0.82 ± 0.00	0.92 ± 0.00	787.50 ± 11.71	195.54 ± 2.83	1.27 ± 0.04	0.15 ± 0.00	0.05 ± 0.00	0.13 ± 0.00	6.31 ± 0.16	4.08 ± 0.04	1.01 ± 0.00
LLaMa-2 7B	0.00 ± 0.00	0.80 ± 0.00	0.95 ± 0.00	769.24 ± 10.65	200.67 ± 2.66	1.75 ± 0.20	0.12 ± 0.00	0.38 ± 0.01	0.45 ± 0.01	50.27 ± 0.85	13.59 ± 0.19	1.06 ± 0.00
Vietcuna 7B	0.00 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	949.80 ± 12.48	209.79 ± 2.81	1.40 ± 0.00	0, 06 ± 0.00	4.97 ± 0.06	4.86 ± 0.06	660.14 ± 8.58	146.84 ± 1.94	1.46 ± 0.01
Vistral 7B Chat	0.00 ± 0.00	0.96 ± 0.00	0.97 ± 0.00	917.24 ± 11.53	204.67 ± 2.49	1.00 ± 0.00	0.00 ± 0.00	2.33 ± 0.04	2.44 ± 0.04	309.88 ± 4.32	73.85 ± 0.99	1.00 ± 0.00
MixSUra	0.00 ± 0.00	0.56 ± 0.00	0.63 ± 0.00	535.76 ± 10.02	133.64 ± 2.33	1.00 ± 0.00	0.07 ± 0.00	0.20 ± 0.00	0.29 ± 0.00	25.96 ± 0.42	8.79 ± 0.10	1.00 ± 0.00
GemSUra	0.00 ± 0.00	0.95 ± 0.00	1.02 ± 0.00	908.58 ± 10.70	214.95 ± 2.36	1.42 ± 0.00	0.00 ± 0.00	1.53 ± 0.01	1.82 ± 0.01	203.14 ± 1.54	54.97 ± 0.39	1.46 ± 0.00
Gemini Pro	0.03 ± 0.00	0.10 ± 0.00	0.18 ± 0.00	99.05 ± 4.19	133.64 ± 1.00	–	0.64 ± 0.00	0.02 ± 0.00	0.03 ± 0.00	2.62 ± 0.14	0.85 ± 0.03	–
GPT-3.5	0.03 ± 0.00	0.43 ± 0.01	0.51 ± 0.01	413.47 ± 11.01	108.06 ± 2.54	–	0.59 ± 0.00	0.04 ± 0.00	0.11 ± 0.00	4.93 ± 0.08	3.34 ± 0.02	–
GPT-4	0.06 ± 0.00	0.38 ± 0.01	0.47 ± 0.01	363.29 ± 10.23	99.39 ± 2.41	–	0.67 ± 0.00	0.03 ± 0.00	0.10 ± 0.00	4.24 ± 0.04	3.14 ± 0.01	–

```
[INST] <<SYS>>
You are a helpful, respectful and honest
    ↪ assistant. You should always answer
    ↪ questions as helpfully as possible,
    ↪ but at the same time be safe. Your
    ↪ reply must not include malicious,
    ↪ racist, sexist, dangerous, or illegal
    ↪ content. Your task is to summarize the
    ↪  text in the triple backtick. The
    ↪ summary should be full of important
    ↪ information, concise and attractive to
    ↪  the reader. The language you must use
    ↪  to summarize is Vietnamese.
<</SYS>>
```{document}``` [/INST]
```

G.3 Sentiment Analysis

```
[INST] <<SYS>>
Hãy xem mình là một Bot có thể phân loại cảm
    ↪ xúc của một câu tiếng Việt. Bot luôn đ
    ↪ ưa câu trả lời của mình ở dạng con số.
    ↪  Trong đó, giá trị 0 cho cảm xúc tiêu
    ↪ cực, 1 cho cảm xúc trung lập, 2 cho cả
    ↪ m xúc tích cực. Bot không được tự trả
    ↪ lời hay giả dạng thành Khách.
Và đây là cuộc trò chuyện mới nhất giữa Bot v
    ↪ à Khách.
<</SYS>>
```
```

Table 9: Performance under zero-shot prompting with typographical error

(a) Question-Answering

Models	XQuAD EM↑	XQuAD F1↑	MLQA EM↑	MLQA F1↑
URA-LLaMa 70B	0.01 ± 0.00	0.17 ± 0.00	0.01 ± 0.00	0.18 ± 0.00
URA-LLaMa 13B	0.00 ± 0.00	0.09 ± 0.00	0.00 ± 0.00	0.10 ± 0.00
URA-LLaMa 7B	0.00 ± 0.00	0.09 ± 0.00	0.00 ± 0.00	0.10 ± 0.00
LLaMa-2 70B	0.00 ± 0.00	0.04 ± 0.00	0.00 ± 0.00	0.05 ± 0.00
LLaMa-2 13B	0.00 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.03 ± 0.00
LLaMa-2 7B	0.00 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.02 ± 0.00
Vietcuna 7B	0.00 ± 0.00	0.06 ± 0.00	0.00 ± 0.00	0.05 ± 0.00
Vistral 7B Chat	0.02 ± 0.00	0.04 ± 0.00	0.02 ± 0.00	0.04 ± 0.00
MixSUra	0.00 ± 0.00	0.11 ± 0.00	0.00 ± 0.00	0.12 ± 0.00
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Gemini Pro	0.10 ± 0.01	0.30 ± 0.01	0.08 ± 0.00	0.28 ± 0.00
GPT-3.5	0.00 ± 0.00	0.19 ± 0.00	0.00 ± 0.00	0.20 ± 0.00
GPT-4	0.00 ± 0.00	0.24 ± 0.00	0.00 ± 0.00	0.25 ± 0.00

(b) Summarization

Models	VietNews R1↑	R2↑	RL↑	SC↑	BS↑	Cv↑	De↑	Cp↑
URA-LLaMa 70B	0.34 ± 0.00	0.16 ± 0.00	0.23 ± 0.00	−0.07 ± 0.00	−0.10 ± 0.18	0.19 ± 0.00	0.19 ± 0.00	61.65 ± 0.87
URA-LLaMa 13B	0.35 ± 0.00	0.14 ± 0.00	0.23 ± 0.00	0.21 ± 0.00	−0.07 ± 0.17	0.64 ± 0.00	0.65 ± 0.00	134.65 ± 3.76
URA-LLaMa 7B	0.37 ± 0.00	0.12 ± 0.00	0.24 ± 0.00	0.08 ± 0.00	−0.24 ± 0.18	0.65 ± 0.00	0.65 ± 0.00	17.92 ± 0.87
LLaMa-2 70B	0.13 ± 0.00	0.04 ± 0.00	0.09 ± 0.00	0.02 ± 0.00	−0.06 ± 0.03	0.20 ± 0.00	0.20 ± 0.00	90.95 ± 0.85
LLaMa-2 13B	0.05 ± 0.00	0.01 ± 0.00	0.04 ± 0.00	−0.04 ± 0.00	−0.21 ± 0.18	0.03 ± 0.00	0.03 ± 0.00	55.91 ± 0.65
LLaMa-2 7B	0.05 ± 0.00	0.01 ± 0.00	0.05 ± 0.00	−0.08 ± 0.00	−0.19 ± 0.04	0.07 ± 0.00	0.07 ± 0.00	55.29 ± 0.88
Vietcuna 7B	0.03 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	−0.18 ± 0.06	0.91 ± 0.00	0.91 ± 0.00	1026.61 ± 3.86
Vistral 7B Chat	0.09 ± 0.00	0.04 ± 0.00	0.06 ± 0.00	0.08 ± 0.00	−0.19 ± 0.12	0.92 ± 0.00	0.92 ± 0.00	860.65 ± 3.41
MixSUra	0.41 ± 0.00	0.19 ± 0.00	0.26 ± 0.00	0.23 ± 0.00	0.00 ± 0.11	0.83 ± 0.00	0.84 ± 0.00	30.65 ± 1.57
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	−0.19 ± 0.05	1.00 ± 0.00	1.00 ± 0.00	1121.14 ± 3.67
Gemini Pro	0.43 ± 0.00	0.21 ± 0.00	0.27 ± 0.00	0.26 ± 0.00	−0.04 ± 0.10	0.79 ± 0.00	0.79 ± 0.00	41.68 ± 0.48
GPT-3.5	0.34 ± 0.00	0.19 ± 0.00	0.23 ± 0.00	0.46 ± 0.00	0.05 ± 0.14	0.81 ± 0.00	0.81 ± 0.00	128.44 ± 2.94
GPT-4	0.39 ± 0.00	0.21 ± 0.00	0.26 ± 0.00	0.43 ± 0.00	0.04 ± 0.00	0.83 ± 0.00	0.83 ± 0.71	24.48 ± 0.00

Models	WikiLingua R1↑	R2↑	RL↑	SC↑	BS↑	Cv↑	De↑	Cp↑
URA-LLaMa 70B	0.28 ± 0.00	0.11 ± 0.00	0.19 ± 0.00	0.31 ± 0.00	0.25 ± 0.23	0.50 ± 0.01	0.51 ± 0.01	167.42 ± 7.09
URA-LLaMa 13B	0.20 ± 0.00	0.07 ± 0.00	0.13 ± 0.00	0.10 ± 0.00	0.20 ± 0.11	0.38 ± 0.00	0.38 ± 0.00	103.69 ± 3.33
URA-LLaMa 7B	0.37 ± 0.00	0.12 ± 0.00	0.24 ± 0.00	0.12 ± 0.00	0.11 ± 0.18	0.65 ± 0.00	0.65 ± 0.00	20.49 ± 0.95
LLaMa-2 70B	0.06 ± 0.00	0.02 ± 0.00	0.04 ± 0.00	0.02 ± 0.00	−0.19 ± 0.13	0.11 ± 0.00	0.11 ± 0.00	85.29 ± 1.05
LLaMa-2 13B	0.04 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	−0.04 ± 0.00	0.09 ± 0.00	0.05 ± 0.00	0.05 ± 0.00	66.85 ± 6.72
LLaMa-2 7B	0.04 ± 0.00	0.00 ± 0.00	0.04 ± 0.00	−0.07 ± 0.00	0.15 ± 0.00	0.06 ± 0.00	0.06 ± 0.00	58.32 ± 3.32
Vietcuna 7B	0.08 ± 0.00	0.02 ± 0.00	0.05 ± 0.00	0.01 ± 0.00	−0.19 ± 0.05	0.78 ± 0.00	0.78 ± 0.00	852.27 ± 8.64
Vistral 7B Chat	0.03 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.02 ± 0.00	−0.19 ± 0.10	0.97 ± 0.00	0.97 ± 0.00	1012.67 ± 13.31
MixSUra	0.46 ± 0.00	0.21 ± 0.00	0.28 ± 0.00	0.12 ± 0.00	0.29 ± 0.07	0.87 ± 0.00	0.87 ± 0.00	20.80 ± 1.14
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	−0.19 ± 0.07	1.00 ± 0.00	1.00 ± 0.00	1120.44 ± 7.07
Gemini Pro	0.36 ± 0.00	0.16 ± 0.00	0.24 ± 0.00	0.21 ± 0.00	0.33 ± 0.12	0.70 ± 0.00	0.70 ± 0.10	141.57 ± 3.26
GPT-3.5	0.39 ± 0.00	0.19 ± 0.00	0.25 ± 0.00	0.55 ± 0.00	0.28 ± 0.11	0.82 ± 0.00	0.82 ± 0.00	200.90 ± 7.40
GPT-4	0.45 ± 0.00	0.20 ± 0.00	0.27 ± 0.00	0.41 ± 0.00	0.28 ± 0.00	0.80 ± 0.03	0.81 ± 0.00	20.40 ± 1.59

(c) Sentiment analysis

Models	VLSP 2016 AC↑	F1↑	AR↑	ECE↓	A@10↑	UiT-VSFC AC↑	F1↑	AR↑	ECE↓	A@10↑
URA-LLaMa 70B	0.63 ± 0.01	0.48 ± 0.01	0.60 ± 0.01	0.09 ± 0.01	0.83 ± 0.04	0.71 ± 0.01	0.45 ± 0.01	0.80 ± 0.01	0.08 ± 0.01	0.99 ± 0.01
URA-LLaMa 13B	0.55 ± 0.02	0.52 ± 0.02	0.59 ± 0.01	0.06 ± 0.01	0.74 ± 0.05	0.72 ± 0.01	0.44 ± 0.05	0.77 ± 0.01	0.18 ± 0.01	0.77 ± 0.02
URA-LLaMa 7B	0.52 ± 0.02	0.36 ± 0.03	0.59 ± 0.01	0.07 ± 0.01	0.66 ± 0.05	0.73 ± 0.01	0.41 ± 0.01	0.71 ± 0.01	0.16 ± 0.01	0.87 ± 0.02
LLaMa-2 70B	0.47 ± 0.01	0.32 ± 0.01	0.63 ± 0.01	0.38 ± 0.01	0.53 ± 0.05	0.49 ± 0.01	0.34 ± 0.01	0.61 ± 0.01	0.43 ± 0.01	0.28 ± 0.03
LLaMa-2 13B	0.46 ± 0.02	0.30 ± 0.01	0.55 ± 0.01	0.39 ± 0.02	0.70 ± 0.05	0.66 ± 0.01	0.40 ± 0.01	0.63 ± 0.01	0.11 ± 0.01	0.89 ± 0.02
LLaMa-2 7B	0.45 ± 0.02	0.36 ± 0.01	0.54 ± 0.01	0.20 ± 0.02	0.51 ± 0.05	0.51 ± 0.01	0.33 ± 0.01	0.65 ± 0.01	0.15 ± 0.01	0.80 ± 0.02
Vietcuna 7B	0.44 ± 0.02	0.27 ± 0.01	0.53 ± 0.01	0.26 ± 0.02	0.53 ± 0.05	0.49 ± 0.01	0.25 ± 0.03	0.46 ± 0.01	0.33 ± 0.01	0.34 ± 0.03
Vistral 7B Chat	0.31 ± 0.01	0.12 ± 0.00	0.81 ± 0.01	0.26 ± 0.02	0.04 ± 0.02	0.06 ± 0.00	0.03 ± 0.00	0.88 ± 0.01	0.67 ± 0.01	0.02 ± 0.01
MixSUra	0.59 ± 0.01	0.59 ± 0.01	0.55 ± 0.01	0.34 ± 0.02	0.52 ± 0.05	0.69 ± 0.01	0.44 ± 0.01	0.61 ± 0.01	0.29 ± 0.01	0.66 ± 0.03
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.67 ± 0.01	0.68 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.77 ± 0.01	0.74 ± 0.00	0.00 ± 0.00
Gemini Pro	0.66 ± 0.01	0.49 ± 0.01	–	0.32 ± 0.01	0.59 ± 0.04	0.78 ± 0.01	0.49 ± 0.01	–	0.45 ± 0.01	0.82 ± 0.02
GPT-3.5	0.64 ± 0.01	0.60 ± 0.01	–	0.36 ± 0.01	0.54 ± 0.05	0.86 ± 0.01	0.71 ± 0.01	–	0.14 ± 0.01	0.86 ± 0.02
GPT-4	0.74 ± 0.00	0.73 ± 0.00	–	0.26 ± 0.00	0.71 ± 0.00	0.83 ± 0.00	0.70 ± 0.00	–	0.17 ± 0.00	0.85 ± 0.00

(d) Text classification

Models	UiT-VSMEC AC↑	F1↑	AR↑	ECE↓	A@10↑	PhoATIS AC↑	F1↑	AR↑	ECE↓	A@10↑
URA-LLaMa 70B	0.25 ± 0.00	0.16 ± 0.00	0.56 ± 0.02	0.20 ± 0.00	0.33 ± 0.00	0.16 ± 0.02	0.26 ± 0.03	0.79 ± 0.00	0.79 ± 0.02	0.08 ± 0.06
URA-LLaMa 13B	0.30 ± 0.00	0.11 ± 0.00	0.51 ± 0.01	0.26 ± 0.00	0.44 ± 0.00	0.01 ± 0.01	0.05 ± 0.01	0.47 ± 0.01	0.84 ± 0.01	0.00 ± 0.04
URA-LLaMa 7B	0.29 ± 0.00	0.10 ± 0.00	0.57 ± 0.01	0.17 ± 0.00	0.30 ± 0.00	0.02 ± 0.01	0.04 ± 0.00	0.55 ± 0.01	0.18 ± 0.01	0.01 ± 0.02
LLaMa-2 70B	0.21 ± 0.01	0.11 ± 0.01	0.61 ± 0.01	0.43 ± 0.01	0.70 ± 0.06	0.12 ± 0.01	0.10 ± 0.01	0.60 ± 0.02	0.46 ± 0.01	0.04 ± 0.02
LLaMa-2 13B	0.19 ± 0.00	0.07 ± 0.00	0.52 ± 0.01	0.47 ± 0.00	0.43 ± 0.00	0.02 ± 0.00	0.06 ± 0.00	0.57 ± 0.01	0.91 ± 0.00	0.01 ± 0.00
LLaMa-2 7B	0.17 ± 0.00	0.10 ± 0.00	0.55 ± 0.00	0.33 ± 0.00	0.29 ± 0.00	0.01 ± 0.01	0.00 ± 0.00	0.56 ± 0.00	0.69 ± 0.01	0.02 ± 0.02
Vietcuna 7B	0.09 ± 0.00	0.09 ± 0.00	0.51 ± 0.01	0.91 ± 0.00	0.09 ± 0.00	0.02 ± 0.01	0.01 ± 0.00	0.55 ± 0.01	0.23 ± 0.01	0.02 ± 0.01
Vistral 7B Chat	0.11 ± 0.01	0.12 ± 0.02	0.66 ± 0.01	0.21 ± 0.01	0.11 ± 0.04	0.20 ± 0.01	0.36 ± 0.02	0.79 ± 0.02	0.44 ± 0.01	0.22 ± 0.04
MixSUra	0.35 ± 0.02	0.27 ± 0.01	0.70 ± 0.01	0.58 ± 0.02	0.70 ± 0.05	0.80 ± 0.02	0.55 ± 0.04	0.94 ± 0.02	0.15 ± 0.02	0.88 ± 0.06
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.58 ± 0.02	0.58 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.92 ± 0.01	0.64 ± 0.01	0.00 ± 0.00
Gemini Pro	0.46 ± 0.02	0.37 ± 0.02	–	0.32 ± 0.02	0.43 ± 0.06	0.64 ± 0.02	0.18 ± 0.01	–	0.59 ± 0.02	0.59 ± 0.05
GPT-3.5	0.42 ± 0.00	0.41 ± 0.00	–	0.58 ± 0.00	0.30 ± 0.00	0.68 ± 0.02	0.64 ± 0.03	–	0.32 ± 0.02	0.70 ± 0.05
GPT-4	0.48 ± 0.00	0.45 ± 0.00	–	0.52 ± 0.00	0.40 ± 0.00	0.86 ± 0.01	0.80 ± 0.02	–	0.14 ± 0.01	0.91 ± 0.03

(e) Knowledge

Models	ZaloE2E EM↑	F1↑	ViMMRC AC↑	F1↑	AR↑	ECE↓	A@10↑
URA-LLaMa 70B	0.23 ± 0.00	0.37 ± 0.00	0.65 ± 0.00	0.53 ± 0.00	0.84 ± 0.00	0.11 ± 0.00	0.77 ± 0.00
URA-LLaMa 13B	0.18 ± 0.00	0.30 ± 0.00	0.41 ± 0.00	0.34 ± 0.00	0.61 ± 0.00	0.22 ± 0.00	0.58 ± 0.00
URA-LLaMa 7B	0.00 ± 0.00	0.05 ± 0.00	0.33 ± 0.02	0.28 ± 0.02	0.61 ± 0.01	0.19 ± 0.02	0.33 ± 0.06
LLaMa-2 70B	0.12 ± 0.01	0.24 ± 0.02	0.47 ± 0.02	0.38 ± 0.02	0.46 ± 0.01	0.21 ± 0.02	0.44 ± 0.08
LLaMa-2 13B	0.13 ± 0.00	0.21 ± 0.00	0.39 ± 0.00	0.31 ± 0.00	0.56 ± 0.00	0.46 ± 0.00	0.33 ± 0.00
LLaMa-2 7B	0.02 ± 0.00	0.05 ± 0.00	0.26 ± 0.01	0.20 ± 0.01	0.51 ± 0.01	0.46 ± 0.01	0.13 ± 0.03
Vietcuna 7B	0.05 ± 0.00	0.15 ± 0.00	0.26 ± 0.01	0.14 ± 0.00	0.50 ± 0.00	0.01 ± 0.01	0.21 ± 0.07
Vistral 7B Chat	0.05 ± 0.01	0.14 ± 0.01	0.49 ± 0.02	0.47 ± 0.02	0.94 ± 0.01	0.29 ± 0.02	0.65 ± 0.07
MixSUra	0.13 ± 0.02	0.24 ± 0.02	0.57 ± 0.02	0.45 ± 0.02	0.53 ± 0.02	0.35 ± 0.02	0.58 ± 0.07
GemSUra	0.00 ± 0.00	0.03 ± 0.00	0.33 ± 0.02	0.17 ± 0.01	0.50 ± 0.02	0.16 ± 0.02	0.35 ± 0.07
Gemini Pro	0.39 ± 0.02	0.55 ± 0.02	0.84 ± 0.02	0.68 ± 0.01	–	0.59 ± 0.02	0.85 ± 0.05
GPT-3.5	0.45 ± 0.01	0.61 ± 0.01	0.90 ± 0.01	0.72 ± 0.04	–	0.10 ± 0.01	0.88 ± 0.07
GPT-4	0.44 ± 0.01	0.61 ± 0.01	0.91 ± 0.01	0.73 ± 0.07	–	0.09 ± 0.07	0.88 ± 0.04

(f) Toxicity

Models	UiT-ViCTSD AC↑	F1↑	AR↑	ECE↓	A@10↑	UiT-ViHSD AC↑	F1↑	AR↑	ECE↓	A@10↑
URA-LLaMa 70B	0.32 ± 0.00	0.21 ± 0.00	0.72 ± 0.01	0.62 ± 0.00	0.33 ± 0.00	0.14 ± 0.00	0.12 ± 0.00	0.64 ± 0.02	0.61 ± 0.00	0.23 ± 0.00
URA-LLaMa 13B	0.27 ± 0.00	0.26 ± 0.00	0.56 ± 0.00	0.56 ± 0.00	0.12 ± 0.00	0.18 ± 0.00	0.11 ± 0.00	0.57 ± 0.01	0.45 ± 0.00	0.20 ± 0.00
URA-LLaMa 7B	0.22 ± 0.00	0.21 ± 0.00	0.63 ± 0.00	0.39 ± 0.00	0.36 ± 0.00	0.12 ± 0.00	0.07 ± 0.00	0.62 ± 0.00	0.38 ± 0.00	0.19 ± 0.00
LLaMa-2 70B	0.17 ± 0.01	0.11 ± 0.01	0.64 ± 0.03	0.73 ± 0.01	0.27 ± 0.05	0.11 ± 0.00	0.11 ± 0.00	0.57 ± 0.01	0.75 ± 0.00	0.07 ± 0.01
LLaMa-2 13B	0.12 ± 0.00	0.11 ± 0.00	0.56 ± 0.01	0.66 ± 0.00	0.12 ± 0.00	0.10 ± 0.00	0.07 ± 0.00	0.59 ± 0.01	0.62 ± 0.00	0.24 ± 0.00
LLaMa-2 7B	0.04 ± 0.00	0.04 ± 0.00	0.62 ± 0.00	0.86 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.54 ± 0.00	0.79 ± 0.00	0.00 ± 0.00
Vietcuna 7B	0.11 ± 0.00	0.11 ± 0.00	0.54 ± 0.00	0.39 ± 0.00	0.13 ± 0.00	0.09 ± 0.00	0.05 ± 0.00	0.50 ± 0.00	0.24 ± 0.00	0.08 ± 0.00
Vistral 7B Chat	0.11 ± 0.01	0.07 ± 0.01	0.66 ± 0.02	0.48 ± 0.01	0.10 ± 0.04	0.09 ± 0.00	0.05 ± 0.00	0.72 ± 0.01	0.42 ± 0.00	0.17 ± 0.01
MixSUra	0.72 ± 0.01	0.39 ± 0.01	0.74 ± 0.02	0.25 ± 0.01	0.81 ± 0.04	0.66 ± 0.01	0.31 ± 0.01	0.67 ± 0.01	0.21 ± 0.01	0.82 ± 0.01
GemSUra	0.00 ± 0.00	0.00 ± 0.00	0.65 ± 0.03	0.78 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.64 ± 0.01	0.68 ± 0.00	0.00 ± 0.00
Gemini Pro	0.81 ± 0.01	0.43 ± 0.01	–	0.31 ± 0.01	0.85 ± 0.04	0.71 ± 0.00	0.37 ± 0.00	–	0.37 ± 0.01	0.70 ± 0.02
GPT-3.5	0.51 ± 0.00	0.46 ± 0.00	–	0.49 ± 0.00	0.54 ± 0.00	0.64 ± 0.00	0.47 ± 0.00	–	0.36 ± 0.00	0.63 ± 0.00
GPT-4	0.88 ± 0.00	0.71 ± 0.00	–	0.12 ± 0.00	0.88 ± 0.00	0.78 ± 0.00	0.56 ± 0.00	–	0.22 ± 0.00	0.78 ± 0.00

```
Hãy đọc kĩ và phân tích sentiment từ Khách.
   ↪ Sau đó, đưa ra câu trả lời của bạn dướ
   ↪ i dạng json với định dạng là ```json {
```

```
↪ "sentiment": `câu trả lời của bạn 0 (
↪ tiêu cực) hay 1 (trung lập) hay 2 (tí
↪ ch cực)`, "confident_level": `độ tự
```

(g) Translation

Models	PhoMT				OPUS100			
	(En → Vi)		(Vi → En)		(En → Vi)		(Vi → En)	
	BLEU↑	hLEPOR↑	BLEU↑	hLEPOR↑	BLEU↑	hLEPOR↑	BLEU↑	hLEPOR↑
URA-LLaMa 70B	0.20 ± 0.00	0.56 ± 0.00	0.12 ± 0.00	0.48 ± 0.00	0.06 ± 0.00	0.38 ± 0.01	0.06 ± 0.00	0.32 ± 0.00
URA-LLaMa 13B	0.18 ± 0.00	0.54 ± 0.00	0.13 ± 0.00	0.48 ± 0.00	0.06 ± 0.00	0.36 ± 0.01	0.08 ± 0.00	0.34 ± 0.00
URA-LLaMa 7B	0.12 ± 0.00	0.46 ± 0.00	0.10 ± 0.00	0.45 ± 0.00	0.05 ± 0.00	0.33 ± 0.00	0.06 ± 0.00	0.31 ± 0.01
LLaMa-2 70B	0.22 ± 0.00	0.53 ± 0.00	0.07 ± 0.00	0.39 ± 0.00	0.07 ± 0.00	0.37 ± 0.01	0.05 ± 0.00	0.27 ± 0.01
LLaMa-2 13B	0.16 ± 0.00	0.50 ± 0.00	0.08 ± 0.00	0.42 ± 0.00	0.05 ± 0.00	0.34 ± 0.01	0.05 ± 0.00	0.29 ± 0.00
LLaMa-2 7B	0.10 ± 0.00	0.40 ± 0.00	0.08 ± 0.00	0.41 ± 0.00	0.04 ± 0.00	0.30 ± 0.00	0.05 ± 0.00	0.27 ± 0.00
Vietcuna 7B	0.12 ± 0.00	0.42 ± 0.00	0.08 ± 0.01	0.40 ± 0.00	0.07 ± 0.01	0.36 ± 0.01	0.09 ± 0.01	0.33 ± 0.00
Vistral 7B Chat	0.23 ± 0.00	0.60 ± 0.00	0.07 ± 0.00	0.38 ± 0.01	0.10 ± 0.00	0.42 ± 0.01	0.06 ± 0.00	0.33 ± 0.00
MixSUra	0.14 ± 0.00	0.50 ± 0.00	0.11 ± 0.00	0.46 ± 0.00	0.06 ± 0.00	0.36 ± 0.01	0.07 ± 0.00	0.34 ± 0.01
GemSUra	0.05 ± 0.00	0.32 ± 0.00	0.01 ± 0.00	0.16 ± 0.00	0.01 ± 0.00	0.15 ± 0.00	0.01 ± 0.00	0.09 ± 0.00
Gemini Pro	0.17 ± 0.01	0.57 ± 0.00	0.12 ± 0.01	0.49 ± 0.00	0.10 ± 0.01	0.42 ± 0.01	0.06 ± 0.01	0.30 ± 0.01
GPT-3.5	0.25 ± 0.00	0.62 ± 0.00	0.20 ± 0.00	0.57 ± 0.00	0.12 ± 0.01	0.45 ± 0.01	0.13 ± 0.01	0.43 ± 0.00
GPT-4	0.25 ± 0.00	0.63 ± 0.00	0.22 ± 0.00	0.59 ± 0.00	0.12 ± 0.01	0.46 ± 0.01	0.14 ± 0.01	0.45 ± 0.00

Table 10: Performance on Knowledge - ViMMRC under few-shot prompting with randomized answer orders

Models	AC↑	F1↑	AR↑	ECE↓	A@10↑
URA-LLaMa 70B	0.76 ± 0.02	0.61 ± 0.02	0.89 ± 0.01	0.14 ± 0.02	0.94 ± 0.04
URA-LLaMa 13B	0.62 ± 0.02	0.50 ± 0.02	0.69 ± 0.02	0.16 ± 0.02	0.67 ± 0.07
URA-LLaMa 7B	0.45 ± 0.02	0.36 ± 0.02	0.57 ± 0.02	0.09 ± 0.02	0.46 ± 0.07
LLaMa-2 70B	0.63 ± 0.02	0.51 ± 0.02	0.42 ± 0.02	0.27 ± 0.02	0.62 ± 0.08
LLaMa-2 13B	0.57 ± 0.02	0.46 ± 0.02	0.64 ± 0.02	0.29 ± 0.02	0.75 ± 0.07
LLaMa-2 7B	0.36 ± 0.02	0.27 ± 0.02	0.56 ± 0.02	0.36 ± 0.02	0.44 ± 0.07
Vietcuna 7B	0.26 ± 0.02	0.15 ± 0.01	0.50 ± 0.00	0.01 ± 0.01	0.31 ± 0.06
Vistral 7B Chat	0.08 ± 0.01	0.11 ± 0.01	0.95 ± 0.01	0.75 ± 0.01	0.06 ± 0.03
MixSUra	0.61 ± 0.02	0.61 ± 0.02	0.54 ± 0.02	0.31 ± 0.02	0.65 ± 0.07
GemSUra	0.35 ± 0.02	0.22 ± 0.01	0.52 ± 0.02	0.13 ± 0.02	0.31 ± 0.07
Gemini Pro	0.89 ± 0.02	0.72 ± 0.01	–	0.64 ± 0.02	0.90 ± 0.05
GPT-3.5	0.92 ± 0.01	0.74 ± 0.04	–	0.08 ± 0.01	0.90 ± 0.04
GPT-4	0.92 ± 0.01	0.74 ± 0.04	–	0.08 ± 0.01	0.88 ± 0.04

```
    ↪ tin cho câu trả lời của bạn trong khoả
    ↪ ng từ 0 tới 1` }```
{few_shot}
Khách: "{context}"
Bot: [/INST]
```

```
[INST] <<SYS>>
Consider yourself a Bot that can classify the
    ↪ sentiment of a sentence in Vietnamese
    ↪ . The bot always gives its answers in
    ↪ numerical form. In particular, the
    ↪ value 0 for negative emotions, 1 for
    ↪ neutral emotions, 2 for positive
    ↪ emotions. The Bot cannot answer itself
    ↪ or pretend to be a Guest.
And this is the latest conversation between
    ↪ the Bot and the Guest.
<</SYS>>
Read carefully and analyze the sentiment from
    ↪ the Guest. Then, give your answer in
    ↪ json format with the format ```json {
    ↪ "sentiment": `is your answer 0 (
    ↪ negative) or 1 (neutral) or 2 (
    ↪ positive)`, "confident_level": `
    ↪ confidence in your answer between 0
    ↪ and 1` }```
{few_shot}
Guest: "{context}"
Bot: [/INST]
```

G.4 Text classification

UiT-VSMEC:

````
[INST] <<SYS>>
Hãy xem mình là một Bot có thể phân loại cảm
    ↪ xúc của một câu văn trong tiếng việt.
    ↪ Trong đó, giá trị 0 cho Sadness, 1 cho
    ↪ Surprise, 2 cho Disgust, 3 cho Fear,
    ↪ 4 cho Anger, 5 cho Other, 6 cho
    ↪ Enjoyment. Bot không được tự trả lời
    ↪ hay giả dạng thành Khách.
Và đây là cuộc trò chuyện mới nhất giữa Bot v
    ↪ à Khách.
<</SYS>>
Hãy đọc kĩ và phân tích cảm xúc từ Khách theo
    ↪ từng bước. Sau đó, đưa ra câu trả lời
    ↪ của bạn dưới dsạng JSON với định dạng
    ↪ là
```json
{
"tag": `câu trả lời của bạn là 0 cho Sadness
    ↪ hay 1 cho Surprise hay 2 cho Disgust
    ↪ hay 3 cho Fear hay 4 cho Anger hay 5
    ↪ cho Other hay 6 cho Enjoyment`,
"confident_level": `độ tự tin cho câu trả lời
    ↪ của bạn trong khoảng từ 0 tới 1`
}
```
{few_shot}
Khách: "{context}"
Bot: [/INST]
````

````
[INST] <<SYS>>
Consider yourself a Bot that can classify
    ↪ emotions of a sentence in Vietnamese.
    ↪ Where, value 0 represents Sadness, 1
    ↪ represents Surprise, 2 represents
    ↪ Disgust, 3 represents Fear, 4
    ↪ represents Anger, 5 represents Other,
    ↪ and 6 represents Enjoyment. The Bot
    ↪ cannot answer itself or pretend to be
    ↪ a Guest.
And this is the latest conversation between
    ↪ the Bot and the Guest.
<</SYS>>
Please read carefully and analyze emotions
    ↪ from Guests step by step. Then output
    ↪ your answer as JSON with the format
```json
{
"tag": `your answer is 0 for Sadness, or 1
    ↪ for Surprise, or 2 for Disgust, or 3
    ↪ for Fear, or 4 for Anger, or 5 for
    ↪ Other, or 6 for Enjoyment`,
"confident_level": `confidence in your answer
    ↪ between 0 and 1`
}
```
{few_shot}
Guest: "{context}"
Bot: [/INST]
````

PhoATIS

```
[INST] <<SYS>>
```

Table 11: Bias and Toxicity performance. The "–" is marked for model whose generated texts do not contain any bias words. DRR, DRG are demographic representations of races and genders. SAR, SAG are stereotypical associations of races and genders

(a) Question-Answering

Models	XQuAD DRR↓	XQuAD DRG↓	XQuAD SAR↓	XQuAD SAG↓	XQuAD Tox↓	MLQA DRR↓	MLQA DRG↓	MLQA SAR↓	MLQA SAG↓	MLQA Tox↓
URA-LLaMa 70B	–	0.39 ± 0.01	–	0.41 ± 0.00	0.02 ± 0.00	–	0.14 ± 0.02	–	0.42 ± 0.03	0.02 ± 0.00
URA-LLaMa 13B	–	0.39 ± 0.01	–	0.45 ± 0.01	0.02 ± 0.00	–	0.17 ± 0.01	–	0.38 ± 0.00	0.02 ± 0.00
URA-LLaMa 7B	–	0.40 ± 0.01	–	0.48 ± 0.00	0.03 ± 0.00	–	0.18 ± 0.01	–	0.37 ± 0.01	0.02 ± 0.00
LLaMa-2 70B	–	0.36 ± 0.01	–	0.39 ± 0.00	0.01 ± 0.00	–	0.18 ± 0.00	–	0.41 ± 0.02	0.01 ± 0.00
LLaMa-2 13B	–	0.35 ± 0.03	–	0.46 ± 0.00	0.01 ± 0.00	–	0.27 ± 0.01	–	0.43 ± 0.00	0.01 ± 0.00
LLaMa-2 7B	–	0.46 ± 0.01	–	0.42 ± 0.00	0.01 ± 0.00	–	0.21 ± 0.06	–	0.45 ± 0.00	0.01 ± 0.00
Vietcuna 7B	–	0.50 ± 0.00	–	–	0.04 ± 0.00	–	0.23 ± 0.09	–	0.49 ± 0.01	0.04 ± 0.00
Vistral 7B Chat	–	0.37 ± 0.01	–	0.47 ± 0.02	0.04 ± 0.00	–	0.22 ± 0.00	–	0.50 ± 0.02	0.04 ± 0.00
MixSUra	–	0.42 ± 0.01	–	0.48 ± 0.00	0.02 ± 0.00	–	0.18 ± 0.00	–	0.38 ± 0.02	0.02 ± 0.00
GemSUra	–	0.39 ± 0.01	–	0.43 ± 0.01	0.04 ± 0.00	–	–	–	–	0.04 ± 0.00
Gemini Pro	–	0.38 ± 0.01	–	0.46 ± 0.00	0.02 ± 0.00	–	0.16 ± 0.00	–	0.43 ± 0.02	0.02 ± 0.00
GPT-3.5	–	0.43 ± 0.01	–	0.48 ± 0.00	0.02 ± 0.00	–	0.18 ± 0.01	–	0.40 ± 0.00	0.02 ± 0.00
GPT-4	–	0.40 ± 0.01	–	0.45 ± 0.00	0.02 ± 0.00	–	0.16 ± 0.01	–	0.41 ± 0.01	0.02 ± 0.00

(b) Summarization

Models	VietNews DRR↓	VietNews DRG↓	VietNews SAR↓	VietNews SAG↓	VietNews Tox↓	WikiLingua DRR↓	WikiLingua DRG↓	WikiLingua SAR↓	WikiLingua SAG↓	WikiLingua Tox↓
URA-LLaMa 70B	–	0.21 ± 0.01	–	0.31 ± 0.01	0.05 ± 0.00	–	0.03 ± 0.02	–	0.25 ± 0.02	0.03 ± 0.00
URA-LLaMa 13B	–	0.20 ± 0.01	–	0.29 ± 0.01	0.04 ± 0.00	–	0.07 ± 0.04	–	0.31 ± 0.03	0.02 ± 0.00
URA-LLaMa 7B	–	0.24 ± 0.02	–	0.33 ± 0.01	0.04 ± 0.00	–	0.07 ± 0.02	–	0.38 ± 0.02	0.03 ± 0.00
LLaMa-2 70B	–	0.24 ± 0.02	–	0.29 ± 0.01	0.02 ± 0.00	–	0.08 ± 0.01	–	0.29 ± 0.02	0.02 ± 0.00
LLaMa-2 13B	–	0.26 ± 0.01	–	0.38 ± 0.01	0.01 ± 0.00	–	0.17 ± 0.08	–	0.50 ± 0.02	0.01 ± 0.00
LLaMa-2 7B	–	0.28 ± 0.02	–	0.39 ± 0.01	0.01 ± 0.00	–	0.39 ± 0.05	–	0.50 ± 0.02	0.01 ± 0.00
Vietcuna 7B	–	0.21 ± 0.02	–	0.32 ± 0.02	0.04 ± 0.00	–	0.17 ± 0.04	–	0.39 ± 0.03	0.03 ± 0.00
Vistral 7B Chat	–	0.22 ± 0.02	–	0.37 ± 0.02	0.04 ± 0.00	–	0.02 ± 0.00	–	0.30 ± 0.02	0.03 ± 0.00
MixSUra	–	0.24 ± 0.01	–	0.29 ± 0.01	0.03 ± 0.00	–	0.00 ± 0.00	–	0.30 ± 0.02	0.02 ± 0.00
GemSUra	–	–	–	–	0.04 ± 0.00	–	–	–	–	0.04 ± 0.00
Gemini Pro	–	0.18 ± 0.01	–	0.26 ± 0.02	0.01 ± 0.00	–	0.01 ± 0.00	–	0.22 ± 0.01	0.01 ± 0.00
GPT-3.5	–	0.22 ± 0.01	–	0.29 ± 0.01	0.04 ± 0.00	–	0.03 ± 0.02	–	0.28 ± 0.01	0.02 ± 0.00
GPT-4	–	0.19 ± 0.01	–	0.28 ± 0.01	0.06 ± 0.00	–	0.09 ± 0.02	–	0.28 ± 0.01	0.02 ± 0.00

(c) Translation

Models	PhoMT (En → Vi) DRR↓	PhoMT (En → Vi) DRG↓	PhoMT (En → Vi) SAR↓	PhoMT (En → Vi) SAG↓	PhoMT (En → Vi) Tox↓	OPUS100 (En → Vi) DRR↓	OPUS100 (En → Vi) DRG↓	OPUS100 (En → Vi) SAR↓	OPUS100 (En → Vi) SAG↓	OPUS100 (En → Vi) Tox↓
URA-LLaMa 70B	–	0.03 ± 0.01	–	0.30 ± 0.01	0.05 ± 0.00	–	0.27 ± 0.01	–	0.47 ± 0.01	0.06 ± 0.00
URA-LLaMa 13B	–	0.09 ± 0.00	–	0.33 ± 0.01	0.05 ± 0.00	–	0.27 ± 0.01	–	0.43 ± 0.02	0.07 ± 0.00
URA-LLaMa 7B	–	0.13 ± 0.00	–	0.33 ± 0.01	0.05 ± 0.00	–	0.18 ± 0.03	–	0.47 ± 0.01	0.07 ± 0.00
LLaMa-2 70B	–	0.06 ± 0.01	–	0.32 ± 0.00	0.05 ± 0.00	–	0.25 ± 0.00	–	0.47 ± 0.02	0.05 ± 0.00
LLaMa-2 13B	–	0.08 ± 0.00	–	0.33 ± 0.02	0.05 ± 0.00	–	0.31 ± 0.02	–	0.47 ± 0.01	0.06 ± 0.00
LLaMa-2 7B	–	0.17 ± 0.01	–	0.29 ± 0.01	0.04 ± 0.00	–	0.21 ± 0.02	–	0.45 ± 0.02	0.05 ± 0.00
Vietcuna 7B	–	0.18 ± 0.01	–	0.36 ± 0.01	0.04 ± 0.00	–	0.16 ± 0.03	–	0.43 ± 0.02	0.07 ± 0.00
Vistral 7B Chat	–	0.20 ± 0.01	–	0.32 ± 0.00	0.04 ± 0.00	–	0.18 ± 0.01	–	0.45 ± 0.02	0.07 ± 0.00
MixSUra	–	0.11 ± 0.01	–	0.33 ± 0.00	0.02 ± 0.00	–	0.26 ± 0.00	–	0.40 ± 0.02	0.04 ± 0.00
GemSUra	–	0.20 ± 0.01	–	0.31 ± 0.00	0.01 ± 0.00	–	0.17 ± 0.00	–	0.46 ± 0.02	0.01 ± 0.00
Gemini Pro	–	0.12 ± 0.00	–	0.33 ± 0.00	0.06 ± 0.00	–	0.15 ± 0.00	–	0.44 ± 0.02	0.07 ± 0.00
GPT-3.5	–	0.11 ± 0.01	–	0.34 ± 0.01	0.05 ± 0.00	–	0.16 ± 0.03	–	0.43 ± 0.03	0.07 ± 0.00
GPT-4	–	0.09 ± 0.01	–	0.34 ± 0.01	0.05 ± 0.00	–	0.14 ± 0.03	–	0.41 ± 0.01	0.07 ± 0.00

Table 12: Performance under zero-shot prompting with easy generation configuration

(a) Question-answering

Models	XQuAD EM↑	XQuAD F1↑
URA-LLaMa 7B	0.00 ± 0.00	0.17 ± 0.00
Vistral 7B Chat	0.04 ± 0.00	0.24 ± 0.00
GemSUra	0.00 ± 0.00	0.20 ± 0.00

(b) Text classification

Models	UiT-VSMEC AC↑	UiT-VSMEC F1↑	UiT-VSMEC AR↑	UiT-VSMEC ECE↓	UiT-VSMEC A@10↑
URA-LLaMa 7B	0.11 ± 0.00	0.15 ± 0.00	0.63 ± 0.01	0.07 ± 0.00	0.34 ± 0.03
Vistral 7B Chat	0.07 ± 0.00	0.21 ± 0.00	0.84 ± 0.01	0.51 ± 0.01	0.13 ± 0.02
GemSUra	0.35 ± 0.00	0.47 ± 0.00	0.93 ± 0.01	0.26 ± 0.01	0.52 ± 0.03

(c) Reasoning

Models	MATH EM↑	MATH F1↑	MATH Equ.↑
URA-LLaMa 7B	0.00 ± 0.00	0.14 ± 0.00	0.04 ± 0.00
Vistral 7B Chat	0.00 ± 0.00	0.09 ± 0.00	0.10 ± 0.00
GemSUra	0.00 ± 0.00	0.26 ± 0.00	0.29 ± 0.00

```
Hãy xem mình là một Bot có thể phân loại ý đị
    ↪ nh của một câu văn trong tiếng việt.
    ↪ Trong đó, giá trị 0 cho 'flight', 1
    ↪ cho 'airfare', 2 cho 'ground_service
    ↪ ', 3 cho 'day_name', 4 cho 'meal',
    ↪ 5 cho 'airport', 6 cho 'airline', 7
```

```
    ↪ cho 'flight_time',  8 cho 'city', 9
    ↪ cho 'ground_fare', 10 cho 'quantity',
    ↪ 11 cho 'abbreviation', 12 cho '
    ↪ distance', 13 cho 'aircraft', 14 cho '
    ↪ capacity', 15 cho 'flight_no', 16 cho
    ↪ 'restriction'. Bot không được tự trả l
    ↪ ời hay giả dạng thành Khách.
Và đây là cuộc trò chuyện mới nhất giữa Bot v
    ↪ à Khách.
<</SYS>>
Hãy đọc kĩ và phân tích cảm xúc từ Khách theo
    ↪  từng bước. Sau đó, đưa ra câu trả lời
    ↪  của bạn dưới dsạng JSON với định dạng
    ↪  là
```json
{
"tag": `câu trả lời của bạn là 0 cho 'flight'
    ↪  hoặc 1 cho 'airfare' hoặc 2 cho '
    ↪ ground_service' hoặc 3 cho 'day_name'
    ↪ hoặc 4 cho 'meal' hoặc 5 cho 'airport'
    ↪  hoặc
6 cho 'airline' hoặc 7 cho 'flight_time' hoặc
    ↪  8 cho 'city' hoặc 9 cho 'ground_fare'
    ↪  hoặc
10 cho 'quantity' hoặc 11 cho 'abbreviation'
    ↪ hoặc 12 cho 'distance' hoặc 13 cho '
    ↪ aircraft' hoặc 14 cho 'capacity' hoặc
15 cho 'flight_no' hoặc 16 cho 'restriction
    ↪ '`,
"confident_level": `độ tự tin cho câu trả lời
    ↪  của bạn trong khoảng từ 0 tới 1`
}
```
{few_shot}
Khách: "{context}"
Bot: [/INST]

🇺🇸

[INST] <<SYS>>
Consider yourself a Bot that can classify
    ↪ intention of a sentence in Vietnamese.
    ↪ Where, value 0 represents 'flight', 1
    ↪ represents 'airfare', 2 represents '
    ↪ ground_service', 3 represents '
    ↪ day_name', 4 represents 'meal', 5
    ↪ represents 'airport', 6 represents '
    ↪ airline', 7 represents 'flight_time',
    ↪ 8 represents 'city', 9 represents '
    ↪ ground_fare', 10 represents 'quantity
    ↪ ', 11 represents 'abbreviation', 12
    ↪ represents 'distance', 13 represents '
    ↪ aircraft', 14 represents 'capacity',
    ↪ 15 represents 'flight_no', and 16
    ↪ represents 'restriction'. The Bot
    ↪ cannot answer itself or pretend to be
    ↪ a Guest.
And this is the latest conversation between
    ↪ the Bot and the Guest.
<</SYS>>
Please read carefully and analyze emotions
    ↪ from Guests step by step. Then output
    ↪ your answer as JSON with the format
```json
{
"tag": `your answer is 0 for 'flight' or 1
    ↪ for 'airfare' or 2 for 'ground_service
    ↪ ' or 3 for 'day_name' or 4 for 'meal'
    ↪ or 5 for 'airport' or 6 for 'airline'
    ↪ or 7 for 'flight_time' or 8 for 'city'
    ↪  or 9 for 'ground_fare' or 10 for '
    ↪ quantity' or 11 for 'abbreviation' or
    ↪ 12 for 'distance' or 13 for 'aircraft'
    ↪  or 14 for 'capacity' or 15 for '
    ↪ flight_no' or 16 for 'restriction'`,
"confident_level": `confidence in your answer
    ↪  between 0 and 1`
}
```
{few_shot}
Guest: "{context}"
Bot: [/INST]
```

G.5 Knowledge

ZaloE2E:

```
🇻🇳

[INST] <<SYS>>
Hãy xem mình là một Bot thông minh, sử dụng
    ↪ kiến thức thông thường trong cuộc sống
    ↪  để thực hiện nhiệm vụ sau. Bot không
    ↪ được tự trả lời hay giả dạng thành Khá
    ↪ ch.
Và đây là cuộc trò chuyện mới nhất giữa Bot v
    ↪ à Khách.
<</SYS>>
Hãy đọc kĩ ngữ cảnh và lựa chọn đáp án đúng
    ↪ cho câu hỏi. Sau đó, đưa ra câu trả lờ
    ↪ i của bạn dưới dạng JSON với định dạng
    ↪  là ```json { "answer": `câu trả lời c
    ↪ ủa bạn`, "confident_level": `độ tự tin
    ↪  cho câu trả lời của bạn trong khoảng
    ↪ từ 0 tới 1` }```
{few_shot}
Câu hỏi: {question}
Câu trả lời: [/INST]

🇺🇸

[INST] <<SYS>>
Consider yourself a smart Bot, using common
    ↪ knowledge in life to perform the
    ↪ following task. Bots may not respond
    ↪ on their own or disguise themselves as
    ↪  Guests.
And here is the latest conversation between
    ↪ Bot and Guest.
<</SYS>>
Read the context carefully and choose the
    ↪ correct answer to the question. Then
    ↪ give your answer as JSON formatted as
    ↪ ```json { "choice": `your answer`, "
    ↪ confident_level": `confidence in your
    ↪ answer between 0 and 1` }```
{few_shot}
Question: {question}
Answer: [/INST]
```

UiT-ViMMRC:

```
🇻🇳

[INST] <<SYS>>
Hãy xem mình là một Bot thông minh, sử dụng
    ↪ kiến thức thông thường trong cuộc sống
    ↪  để thực hiện nhiệm vụ sau. Đọc kĩ phầ
    ↪ n Ngữ cảnh và đọc câu hỏi để lựa chọn
    ↪ đáp án nào chính xác nhất được đề cập
    ↪ trong Ngữ cảnh. Nếu đáp án 0 chính xác
    ↪  thì trả lời 0, đáp án 1 chính xác thì
    ↪  trả lời 1, ...
Bot không được tự trả lời hay giả dạng thành
    ↪ Khách.
Và đây là cuộc trò chuyện mới nhất giữa Bot v
    ↪ à Khách.
<</SYS>>
Hãy đọc kĩ ngữ cảnh và lựa chọn đáp án đúng
    ↪ cho câu hỏi. Sau đó, đưa ra câu trả lờ
    ↪ i của bạn dưới dạng JSON với định dạng
    ↪  là ```json { "choice": `câu trả lời c
    ↪ ủa bạn là "A" hoặc "B" hoặc "C" hoặc "
    ↪ D"`, "confident_level": `độ tự tin cho
    ↪  câu trả lời của bạn trong khoảng từ 0
    ↪  tới 1` }```
{few_shot}
Ngữ cảnh: ''' {context} '''
Câu hỏi: Hãy lựa chọn đáp án đúng. {question}
{list_answers}
Câu trả lời: [/INST]

🇺🇸

[INST] <<SYS>>
```

```
Consider yourself a smart Bot, using common
    ↪ knowledge in life to perform the
    ↪ following task. Read the Context
    ↪ section carefully and read the
    ↪ question to choose the most accurate
    ↪ answer mentioned in the Context. If
    ↪ answer 0 is correct, answer 0, if
    ↪ answer 1 is correct, answer 1, ...
Bots may not respond on their own or disguise
    ↪  themselves as Guests.
And here is the latest conversation between
    ↪ Bot and Guest.
<</SYS>>
Read the context carefully and choose the
    ↪ correct answer to the question. Then
    ↪ give your answer as JSON formatted as
    ↪ ```json { "choice": `your answer is "A
    ↪ " or "B" or "C" or "D" `, "
    ↪ confident_level": `confidence in your
    ↪ answer between 0 and 1` }```
{few_shot}
Context: ''' {context} '''
Question: Please choose the correct answer. {
    ↪ question}
{list_answers}
Answer: [/INST]
```

G.6 Toxicity detection

```
[INST] <<SYS>>
Hãy xem mình là một Bot có thể phân loại tính
    ↪  độc hại (toxicity) trong một câu tiế
    ↪ ng Việt. Tính độc hại (toxicity) trong
    ↪  một câu văn có nghĩa là câu văn đó
    ↪ mang ý nghĩa tổn thương và không tôn
    ↪ trọng đến người khác.
Dưới góc độ là một con người khác khi nhận câ
    ↪ u văn đó, đầu tiên hãy xem xét đó có
    ↪ phải là một câu độc hại gây tổn thương
    ↪  mình không? Nếu câu không có yếu tố
    ↪ độc hại thì trả lời 0, nếu câu có yếu
    ↪ tố độc hại thì trả lời 1.
Bot không được tự trả lời hay giả dạng thành
    ↪ Khách.
Và đây là cuộc trò chuyện mới nhất giữa Bot v
    ↪ à Khách.
<</SYS>>
Hãy đọc kĩ và phân tích tính độc hại từ đầu v
    ↪ ào của Khách. Sau đó, đưa ra câu trả l
    ↪ ời của bạn dưới dạng JSON với định dạ
    ↪ ng là ```json { "toxicity_level": `câu
    ↪  trả lời của bạn là 0 (câu không độc h
    ↪ ại) hoặc 1 (câu độc hại)`, "
    ↪ confident_level": `độ tự tin cho câu
    ↪ trả lời của bạn trong khoảng từ 0 tới
    ↪ 1` }```
{few_shot}
Khách: "{context}"
Bot: [/INST]
```

```
[INST] <<SYS>>
Consider yourself a Bot that can classify
    ↪ toxicity in a Vietnamese sentence.
    ↪ Toxicity in a sentence means that the
    ↪ sentence is hurtful and disrespectful
    ↪ to others.
From the perspective of another person when
    ↪ receiving that sentence, first
    ↪ consider whether it is a harmful
    ↪ sentence that will hurt you? If the
    ↪ sentence has no toxic elements, answer
    ↪  0. If the sentence has toxic elements
    ↪ , answer 1.
Bots may not respond on their own or disguise
    ↪  themselves as Guests.
And here is the latest conversation between
    ↪ Bot and Guest.
<</SYS>>
Please read carefully and analyze the
    ↪ toxicity of Guest input. Then, output
    ↪ your answer as JSON in the format ```
    ↪ json { "toxicity_level": `your answer
    ↪ is 0 (non-toxic sentence) or 1 (toxic
    ↪ sentence)`, "confident_level": `
    ↪ confidence in your answer between 0
    ↪ and 1` }```
{few_shot}
Guest: "{context}"
Bot: [/INST]
```

G.7 Language

MLQA-MLM:

```
[INST] <<SYS>>
Hãy xem mình là một Bot có thể tìm và sửa các
    ↪  lỗi sai chính tả có trong một câu tiế
    ↪ ng Việt. Chú ý, Bot không chỉnh sửa
    ↪ hay thêm bớt các từ trong câu, chỉ sửa
    ↪  các từ bị sai chính tả. Bot không đượ
    ↪ c tự trả lời hay giả dạng thành Khách.
Và đây là cuộc trò chuyện mới nhất giữa Bot v
    ↪ à Khách.
<</SYS>>
{few_shot}
Khách: "{context}"
Bot: [/INST]
```

```
[INST] <<SYS>>
Consider yourself a Bot that can find and
    ↪ correct misspellings in a Vietnamese
    ↪ sentence. Note, the Bot does not edit
    ↪ or add or remove words in the sentence
    ↪ , only correct misspelled words. Bots
    ↪ can't reply to themselves or pretend
    ↪ to be Guest.
And this is the latest conversation between
    ↪ Bot and Guest.
<</SYS>>
{few_shot}
Guest: "{context}"
Bot: [/INST]
```

VSEC:

```
[INST] <<SYS>>
Hãy xem mình là một Bot có thể thay thế token
    ↪  [MASKED] thành một từ thích hợp trong
    ↪  một câu tiếng Việt. Chú ý, Bot không
    ↪ chỉnh sửa hay thêm bớt các từ trong câ
    ↪ u, chỉ sửa các từ bị sai chính tả. Bot
    ↪  không được tự trả lời hay giả dạng th
    ↪ ành Khách.
Và đây là cuộc trò chuyện mới nhất giữa Bot v
    ↪ à Khách.
<</SYS>>
{few_shot}
Khách: "{context}"
Bot: [/INST]
```

```
[INST] <<SYS>>
Consider yourself a Bot that can replace the
    ↪ token [MASKED] with a suitable word in
    ↪  a Vietnamese sentence. Note, the Bot
    ↪ does not edit or add or remove words
    ↪ in the sentence, only correct
    ↪ misspelled words. Bot cannot reply to
    ↪ itself or pretend to be Guest.
And here is the latest conversation between
    ↪ Bot and Guest.
<</SYS>>
{few_shot}
Guest: "{context}"
Bot: [/INST]
```

G.8 Information retrieval

🇻🇳

````
[INST] <<SYS>>
Hãy xem mình là một Bot thông minh có thể trả
    ↪ lời câu hỏi chính xác.
<</SYS>>
{few_shot}
Văn bản: {passage}\
Câu hỏi: {question}
Văn bản trên có thể hỗ trợ trả lời câu hỏi kh
    ↪ ông?.
Đưa ra câu trả lời của bạn dưới dạng JSON với
    ↪ định dạng là ```json { "answer": ` "
    ↪ Yes" or "No" ` }```
Bot: [/INST]
````

🇺🇸

````
[INST] <<SYS>>
See yourself as a smart Bot that can answer
    ↪ questions accurately.
<</SYS>>
{few_shot}
Passage: {passage}
Question: {question}
Can the above passage answer the question?
Output your answer as JSON in the format ```
    ↪ json { "answer": ` "Yes" or "No" `
    ↪ }```
Bot: [/INST]
````

G.9 Reasoning

Synthetic reasoning:

🇻🇳

````
[INST] <<SYS>>
Hãy xem mình là một Bot thông minh có thể trả
    ↪ lời câu hỏi chính xác.
<</SYS>>
Hãy dựa vào `Quy luật` được cho để suy luận
    ↪ ra quy tắc. Sau đó, đưa ra câu trả lời
    ↪ của bạn dưới dạng json với định dạng
    ↪ là ```json { "answer": câu trả lời của
    ↪ bạn, "confident_level": độ tự tin của
    ↪ bạn trong khoảng từ 0 tới 1 }```
{few_shot}
Quy luật: ```
{rule}
```
Kết quả: [/INST]
````

🇺🇸

````
[INST] <<SYS>>
See yourself as a smart Bot that can answer
    ↪ questions correctly.
<</SYS>>
Solve based on the given `Rule` to deduce the
    ↪ rule. Then give your answer as json
    ↪ formatted as ```json { "answer": your
    ↪ answer, "confident_level": your
    ↪ confidence level between 0 to 1 }```
{few_shot}
Rule: ```
{rule}
```
Result: [/INST]
````

MATH:

🇻🇳

````
[INST] <<SYS>>
Hãy xem mình là một Bot thông minh có thể trả
    ↪ lời câu hỏi chính xác.
Bạn hãy giải bài toán được cho bên dưới, câu
    ↪ trả lời càng đơn giản càng tốt và kèm
    ↪ thêm độ tự tin cho câu trả lời của bạn
    ↪ trong khoảng từ 0 tới 1.
<</SYS>>
Hãy giải bài toán trước theo từng bước. Sau đ
    ↪ ó, đưa ra câu trả lời của bạn dưới dạ
    ↪ ng json với định dạng là ```json { "
    ↪ answer": câu trả lời của bạn, "
    ↪ confident_level": độ tự tin của bạn
    ↪ trong khoảng từ 0 tới 1 }```
{few_shot}
Bài toán: ```
{problem}
```
Lời giải: [/INST]
````

🇺🇸

````
[INST] <<SYS>>
See yourself as a smart Bot that can answer
    ↪ questions correctly.
Please solve the problem given below, the
    ↪ simpler the answer the better and add
    ↪ confidence to your answer between 0
    ↪ and 1.
<</SYS>>
Let's solve the previous problem step by step
    ↪ . Then give your answer as json
    ↪ formatted as ```json { "answer": your
    ↪ answer, "confident_level": your
    ↪ confidence level between 0 to 1 }```
{few_shot}
Problem: ```
{problem}
```
Solution: [/INST]
````

G.10 Translation

🇻🇳

````
[INST] <<SYS>>
Hãy xem mình là một Bot có thể dịch từ [
    ↪ source_language] qua [target_language
    ↪ ]. Bot không được tự trả lời hay giả d
    ↪ ạng thành Khách.
Và đây là cuộc trò chuyện mới nhất giữa Bot v
    ↪ à Khách.
Hãy dịch từ [source_language] qua [
    ↪ target_language] và định dạng câu trả
    ↪ lời dưới dạng json với định dạng là
    ↪ ```json { "translation": `câu trả lời
    ↪ của bạn ` }```
<</SYS>>
{few_shot}
Khách: "{context}"
Bot: [/INST]
````

🇺🇸

````
[INST] <<SYS>>
Consider yourself a Bot that can translate
    ↪ from [source_language] to [
    ↪ target_language]. Bot cannot reply to
    ↪ itself or pretend to be Guest.
And here is the latest conversation between
    ↪ Bot and Guest.
Translate from [source_language] to [
    ↪ target_language] and format the answer
    ↪ as json fortmated as```json { "
    ↪ translation": `your answer` }```
<</SYS>>
{few_shot}
Guest: "{context}"
Bot: [/INST]
````

H Sample generation results

In this section, we provide sample generation results obtained during our evaluation process. Specifically, for the Vietnamese context, we extracted text directly from the original dataset without any modifications, and these excerpts are visually distinguished by yellow highlighting. Furthermore, we translated the Vietnamese context into English utilizing Google Translate, and the resulting English text is demarcated with green highlighting.

H.1 Question-Answering: MLQA

Question:

```
Sảnh thứ hai được xây khi nào?
```

Context:

```
Đầu thế kỷ 21, số khách/năm đã vượt mức
    ↪ 750.000 và năm 2004 đã đạt mức dưới
    ↪ 900.000 khách. Nhà ga đã được mở rộng
    ↪ năm 2003 và một nhà ga hành khách mới
    ↪ được xây năm 2005.
```

Ground Truth:

```
2005
```

URA-LLaMa 7B:

```
Sảnh thứ hai được xây khi năm 2003
```

URA-LLaMa 13B:

```
Sảnh thứ hai được xây khi nào? Đầu thế kỷ 21,
    ↪ số khách/năm đã vượt mức 750.000 và n
    ↪ ăm 2004 đã đạt mức dưới 900.000 khách.
```

URA-LLaMa 70B:

```
Sảnh thứ hai được xây năm 2005.
```

LLaMa-2 7B:

```
The second terminal was built in what year?
    ↪ According to the text, the second
    ↪ terminal was built in 2003.
```

LLaMa-2 13B:

```
Sure, I'd be happy to help! Based on the
    ↪ provided text, the second terminal was
    ↪ built when? The text states that the
    ↪ number of passengers per year has
    ↪ exceeded 750,000 and reached a low of
    ↪ under 900,000 in 2004. The station was
    ↪ expanded in 2003 and a new passenger
    ↪ terminal was built in 2005. Therefore,
    ↪ the second terminal was built in
```

LLaMa-2 70B:

```
Sảnh thứ hai được xây khi nào? Trả lời: Nhà
    ↪ ga được mở rộng năm 2003 và một nhà ga
    ↪ hành khách mới được xây năm 2005.
```

Vietcuna-7B-v3:

```
### Human: Sảnh thứ hai được xây khi nào? ###
    ↪ Assistant: Sảnh thứ hai được xây dựng
    ↪ vào đầu thế kỷ 21.
```

Vistral:

```
Nhà ga hành khách mới được xây dựng vào năm
    ↪ 2005.
```

MixSUra:

```
Đáp án: Nhà ga được mở rộng năm 2003 và một
    ↪ nhà ga hành khách mới được xây năm
    ↪ 2005.
```

GemSUra 7B:

```
Sân bay quốc tế Phú Quốc hiện đang hoạt động
    ↪ với ba sàn hàng hóa (sàn A ở phía đông
    ↪ nam; sân B trên đảo chính giữa trung
    ↪ tâm thị xã Phú Quoc về mặt đất thuộc đ
    ↪ ịa phận thôn An Bình Tây - Xã Long Sơn
    ↪ Đông ; và sấn D gần cầu Đốc Lập). Ngo
    ↪ ài ra còn dự kiến xây thêm cả ng casio
    ↪ đường sắt để nối liền Hà Tĩnh lên tỉ
    ↪ nh Kiên Giang theo hướng Nam Định
```

Gemini Pro:

```
Thông tin về thời điểm xây sảnh thứ hai không
    ↪ được nhắc đến trong đoạn văn bạn cung
    ↪ cấp.
```

GPT-3.5:

```
Nhà ga hành khách mới được xây vào năm 2005.
```

GPT-4:

```
Sảnh thứ hai được xây vào năm 2005.
```

Question:

```
When was the second terminal built?
```

Context:

```
At the beginning of the 21st century, the
    ↪ number of passengers per year exceeded
    ↪ 750,000, reaching below 900,000
    ↪ passengers in 2004. The station was
    ↪ expanded in 2003 and a new passenger
    ↪ terminal was built in 2005.
```

Ground Truth:

```
2005
```

URA-LLaMa 7B:

```
The second terminal was built in 2003
```

URA-LLaMa 13B:

```
When was the second terminal built? At the
    ↪ beginning of the 21st century, the
    ↪ number of passengers per year exceeded
    ↪ 750,000, reaching below 900,000
    ↪ passengers in 2004.
```

URA-LLaMa 70B:

```
The second terminal was built in 2005
```

LLaMa-2 7B:

```
The second terminal was built in what year?
    ↪ According to the text, the second
    ↪ terminal was built in 2003.
```

LLaMa-2 13B:

```
Sure, I'd be happy to help! Based on the
    ↪ provided text, the second terminal was
    ↪ built when? The text states that the
    ↪ number of passengers per year has
    ↪ exceeded 750,000 and reached a low of
    ↪ under 900,000 in 2004. The station was
    ↪ expanded in 2003 and a new passenger
    ↪ terminal was built in 2005. Therefore,
    ↪ the second terminal was built in
```

LLaMa-2 70B:

```
When was the second hall built? Answer: The
    ↪ station was expanded in 2003 and a new
    ↪ passenger terminal was built in 2005.
```

Vietcuna-7B-v3:

```
### Human: When was the second terminal built
    ↪ ? ### Assistant: The second terminal
    ↪ was built at the beginning of the 21st
    ↪ century
```

Vistral:

```
Nhà ga hành khách mới được xây dựng vào năm
    ↪ 2005.
```

MixSUra:

```
Answer: The station was expanded in 2003 and
    ↪ a new passenger terminal was built in
    ↪ 2005.
```

GemSUra 7B:

```
Phu Quoc International Airport is currently
    ↪ operating with three cargo floors (
    ↪ floor A in the southeast; yard B on
    ↪ the main island in the center of Phu
    ↪ Quoc town on the ground in An Binh Tay
    ↪  village - Long Son Dong Commune). ;
    ↪ and bridge D near Doc Lap bridge). In
    ↪ addition, it is planned to build more
    ↪ railway stations to connect Ha Tinh to
    ↪  Kien Giang province in the direction
    ↪ of Nam Dinh.
```

Gemini Pro:

```
Information about when the second hall was
    ↪ built is not mentioned in the
    ↪ paragraph you provided.
```

GPT-3.5:

```
The new passenger terminal was built in 2005.
```

GPT-4:

```
The second terminal was built in 2005.
```

H.2 Summarization: WikiLingua

Document:

```
Bác sĩ phụ khoa/sản khoa hoặc bác sĩ kiểm tra
    ↪  nội tiết sinh sản có thể thực hiện cá
    ↪ c kỹ thuật kiểm tra có tính chính xác
    ↪ cao hơn so với các phương pháp thực hi
    ↪ ện tại nhà. Bác sĩ có thể thử máu để đ
    ↪ o nồng độ progesterone và các hormone
    ↪ khác, hoặc kiểm tra để xem có sự bất
    ↪ thường trong chức năng tuyến giám hay
    ↪ nồng độ prolactin không. Những bất thư
    ↪ ờng về cấu trúc của hệ thống sinh sản
    ↪ có thể được phát hiện nhờ phương pháp
    ↪ siêu âm, mà đây cũng là một trong nhữ
    ↪ ng yếu tố ảnh hưởng đến sự rụng trứng.
    ↪  Đàn ông cũng có thể bị vô sinh. Đối v
    ↪ ới đàn ông, việc đầu tiên là kiểm tra
    ↪ chu kỳ tinh trùng, sau đó thể làm siêu
    ↪  âm để xem xét các vấn đề (nếu có)
    ↪ trong hệ thống sinh sản của họ. Nếu bá
    ↪ c sĩ nghi ngờ rằng bạn không có khả nă
    ↪ ng mang thai do không có sự rụng trứng
    ↪ , họ có thể khuyến nghị bạn sử dụng mộ
    ↪ t vài loại thuốc tùy thuộc vào tình tr
    ↪ ạng của bạn. Nếu chưa có chẩn đoán chí
    ↪ nh thức, bạn không nên đánh đồng việc
    ↪ khó mang thai là do trứng không rụng b
    ↪ ởi còn có rất nhiều nguyên nhân khác c
    ↪ ó thể dẫn đến tình trạng này. Bác sĩ c
    ↪ ó thể kiểm tra xem ống dẫn trứng của b
    ↪ ạn có bị tắc hay không, tinh trùng có
    ↪ vấn đề gì không, có vấn đề ở tử cung
    ↪ hay buồng trứng không, và sự liên quan
    ↪  giữa chất lượng trứng với tuổi tác.
```

Ground Truth:

```
Đi khám. Kiểm tra khả năng mang thai của nửa
    ↪ kia. Tham khảo về các phương pháp điều
    ↪  trị.
```

URA-LLaMa 7B:

```
Tóm tắt: Bác sĩ phụ khoa/sản khoa có thể thực
    ↪  hiện các kỹ thuật kiểm tra có tính ch
    ↪ ính xác cao hơn so với các phương pháp
    ↪  thực hiện tại nhà. Bác sĩ có thể thử
    ↪ máu để đo nồng độ progesterone và các
    ↪ hormone khác, hoặc kiểm tra để xem có
    ↪ sự bất thường trong chức năng tuyến gi
    ↪ áp hay nồng độ prolactin không. Những
    ↪ bất thường về cấu trúc của hệ thống
    ↪ sinh sản có thể được phát hiện nhờ phư
    ↪ ơng pháp siêu âm, mà đây cũng là một
    ↪ trong những yếu tố ảnh hưởng đến sự rụ
    ↪ ng trứng.
```

URA-LLaMa 13B:

```
Xin chào! Tôi là một trợ lý hữu dụng, có độ
    ↪ tin cậy cao và sẽ trả lời các câu hỏi
    ↪ của bạn theo những cách có ích nhất có
    ↪  thể. Trong trường hợp bạn đang gặp vấ
    ↪ n đề về sinh sản, bác sĩ có thể thực
    ↪ hiện các kỹ thuật kiểm tra có tính chí
    ↪ nh xác cao hơn so với các phương pháp
    ↪ thực hiện tại nhà. Bác sĩ có thể thử m
    ↪ áu để đo nồng độ progesterone và các
    ↪ hormone khác, hoặc kiểm tra để xem có
    ↪ sự bất thường trong chức năng tuyến gi
    ↪ áp hay nồng độ prolactin không. Những
    ↪ bất thường về cấu trúc của hệ thống
    ↪ sinh sản có thể được phát hiện thông
    ↪ qua chẩn đoán.
```

URA-LLaMa 70B:

```
Vô sinh là một vấn đề phổ biến ảnh hưởng đến
    ↪ cả nam và nữ. Trong trường hợp của phụ
    ↪  nữ, bác sĩ có thể thực hiện các kỹ
    ↪ thuật kiểm tra có tính chính xác cao h
    ↪ ơn so với các phương pháp thực hiện tạ
    ↪ i nhà để xác định nguyên nhân khó mang
    ↪  thai. Các kỹ thuật này bao gồm thử má
    ↪ u để đo nồng độ progesterone và các
    ↪ hormone khác, hoặc kiểm tra để xem có
    ↪ sự bất thường trong chức năng tuyến gi
    ↪ áp hay nồng độ prolactin không. Ngoài
    ↪ ra, các bất thường về cấu trúc của hệ
    ↪ thống sinh sản có thể được phát hiện
    ↪ thông qua khám và chẩn đoán.
```

LLaMa-2 7B:

```
As a responsible AI language model, I must
    ↪ inform you that the information
    ↪ provided in the triple backtick is not
    ↪  accurate and could potentially
    ↪ harmful. The statement "bác sĩ phụ
    ↪ khoa/sản khoa hoặc bác sĩ kiểm tra nội
    ↪  tiết sinh sản có thể thực hiện các kỹ
    ↪  thuật kiểm tra có tính chính xác cao
    ↪ hơn so với các phương pháp thực hiện t
    ↪ ại nhà" is not supported by scientific
    ↪  evidence and could lead to false
    ↪ assumptions about fertility testing.
    ↪ Fertility testing involves a series of
    ↪  medical procedures and tests to
    ↪ evaluate the ability of an individual
    ↪ or couple to conceive. These tests can
    ↪  help identify potential issues with
    ↪ the reproductive system, such as
    ↪ ovulation disorders, tubal blockages,
    ↪ or endometriosis. However, it is
    ↪ important to note that these tests are
    ↪  not 100% accurate and can sometimes
    ↪ produce false results. The statement "
    ↪ đàn ông cũng có thể bị vô sinh" is
    ↪ also not accurate. While it is true
    ↪ that men can experience infertility,
    ↪ the causes and diagnosis of male
```

```
↪ infertility are different from those
↪ of female infertility. Male
↪ infertility is
```

LLaMa-2 13B:

```
Hello! As a fertility specialist, I
↪ understand that you are looking for
↪ information on the causes and
↪ treatments of infertility in women.
↪ Infertility can be a challenging and
↪ emotional experience, but there are
↪ many options available for those who
↪ are struggling to conceive. As a woman
↪ , there are several factors that can
↪ contribute to infertility, including:
↪ 1. Hormonal imbalances: Imbalances in
↪ hormones such as progesterone and
↪ estrogen can affect ovulation and make
↪  it difficult to conceive. 2.
↪ Polycystic ovary syndrome (PCOS): PCOS
↪  is a common condition that can cause
↪ infertility due to hormonal imbalances
↪  and irregular ovulation. 3.
↪ Endometriosis: This condition occurs
↪ when tissue similar to the lining of
↪ the uterus grows outside of the uterus
↪ , causing inflammation and scarring
↪ that can lead to infertility. 4.
↪ Fibroids: These non-cancerous growths
↪ in the uterus can interfere with
↪ implantation of the fertilized egg and
↪  make it difficult to conceive. 5.
↪ Blocked fallopian tubes: If the
↪ fallopian tubes are blocked or damaged
↪ , it can prevent the sperm from
↪ reaching the egg and make it difficult
↪  to conceive. 6. Age: Women'
```

LLaMa-2 70B:

```
Bác sĩ phụ khoa/sản khoa hoặc bác sĩ kiểm tra
↪  nội tiết sinh sản có thể thực hiện cá
↪ c kỹ thuật kiểm tra có tính chính xác
↪ cao hơn so với các phương pháp thực hi
↪ ện tại nhà để diagnose và điều trị vô
↪ sinh. Các kỹ thuật kiểm tra bao gồm: *
↪  Kiểm tra máu để đo nồng độ
↪ progesterone và các hormone khác * Kiể
↪ m tra để xem có sự bất thường trong ch
↪ ức năng tuyến giám hay nồng độ
↪ prolactin không * Phương pháp siêu âm
↪ để xem xét các vấn đề (nếu có) trong h
↪ ệ thống sinh sản Đối với đàn ông, việc
↪  đầu tiên là kiểm tra chu
```

Vietcuna-7B-v3:

```
### Human: Viết một đoạn mã Python để tạo ra
↪ một danh sách các số nguyên tố từ 1 đế
↪ n n, trong đó n là một số nguyên dương
↪  lớn hơn 2. Ngoài ra, chương trình chỉ
↪  nên in ra các số nguyên tố theo thứ t
↪ ự giảm dần. Chương trình cũng nên kiểm
↪  tra xem một số có phải là số nguyên t
↪ ố hay không bằng cách sử dụng hàm `
↪ is_prime()` tích hợp sẵn. Nếu số đó kh
↪ ông phải là số nguyên tố, nó nên được
↪ in ra cùng với một thông báo lỗi phù h
↪ ợp. Cuối cùng, chương trình nên xử lý
↪ trường hợp n nhỏ hơn 5 và in ra một th
↪ ông báo lỗi phù hợp. ### Assistant: Đâ
↪ y là mã Python để tạo ra một danh sách
↪  các số nguyên tố từ 1 đến n, trong đó
↪  n là một số nguyên dương lớn hơn 2, v
↪ à in ra chúng theo thứ tự giảm dần:
↪ ```python def is_prime(n): if n <= 1:
↪ return False for i in range(2, int(n
↪ **0.5) + 1): if n % i == 0: return
↪ False return True # Tạo danh sách các
↪ số nguyên tố từ 1 đến n primes = []
↪ for num in range(1, n+1): if is_prime(
↪ num) and num > 2: primes.append(num) #
↪  In ra danh sách các số nguyên tố theo
↪  thứ tự giảm dần print("D
```

Vistral:

```
Đoạn văn giải thích rằng mặc dù có thể tự mì
↪ nh kiểm tra sự rụng trứng nhưng tốt nh
↪ ất vẫn nên tham khảo ý kiến chuyên gia
↪  y tế như bác sĩ phụ khoa/sản khoa hoặ
↪ c bác sĩ kiểm tra nội tiết sinh sản. H
↪ ọ sẽ tiến hành các xét nghiệm chẩn đoá
↪ n chính xác hơn như thử máu để đo nồng
↪  độ hormone, kiểm tra để tìm ra sự bất
↪  thường trong chức năng tuyến giáp hoặ
↪ c nồng độ prolactin. Ngoài ra, bác sĩ
↪ có thể sử dụng hình ảnh siêu âm để phá
↪ t hiện các vấn đề tiềm ẩn trong hệ thố
↪ ng sinh sản. Đối với nam giới, trước h
↪ ết cần phải kiểm tra chu trình xuất
↪ tinh, tiếp theo là siêu âm nếu cần thi
↪ ết. Nếu bác sĩ nghi ngờ rằng người bệ
↪ nh không thể thụ thai vì thiếu sự rụng
↪  trứng thì có thể kê đơn dùng thuốc dự
↪ a trên tình trạng cụ thể của từng cá
↪ nhân. Tuy nhiên, nếu chưa có kết luận
↪ cuối cùng, đừng vội cho rằng khó thụ
↪ thai chỉ do trứng không rụng, vì còn
↪ nhiều lý do khác nữa.
```

MixSUra:

```
Đôi bác sĩ khoa phụ khoa/sản khoa có thể tiến
↪  hành kiểm tra sinh sản với độ chính x
↪ ác cao hơn so với các phương pháp ở nh
↪ à. Phương pháp kiểm tra có thể bao gồm
↪  xét nghiệmáu để đo nồng độ các
↪ hormone, hoặc sử dụng siêu âm để phát
↪ hiện các vấn đề về hệ thống sinh sản.
↪ Khả năng mang thai của phụ nữ có thể b
↪ ị ảnh hưởng do rối mất, không có sự rụ
↪ ng trứng hoặc các vấn đề khác trong hệ
↪  thống sinh sản. Đối với đàn ông, việc
↪  đầu tiên là kiểm tra chu kỳ tinh trù
↪ ng, sau đó có thể thực hiện siêu âm.
```

GemSUra 7B:

```
Trong trường hợp trên đã trình bày rõ ràng
↪  vai trò thiết nghĩ như thế nào khi gặp
↪  bệnh viện chuyên ngành y tế cụm quanh
↪  sức khỏe nữ: bệnh viên nam - phẫu thủ
↪  công nghệ tái tạo , điều trị ung thư
↪ vú...
```

Gemini Pro:

```
Để kiểm tra khả năng có con, các bác sĩ phụ
↪ khoa hoặc kiểm tra nội tiết sẽ thực hi
↪ ện các xét nghiệm chuyên sâu hơn xét
↪ nghiệm tại nhà, như: - Xét nghiệm máu
↪ để kiểm tra nồng độ hormone. - Siêu âm
↪  để phát hiện bất thường ở hệ thống
↪ sinh sản. - Kiểm tra tinh trùng ở nam
↪ giới. - Sử dụng thuốc để kích thích rụ
↪ ng trứng khi cần. Không nên chỉ dựa và
↪ o dấu hiệu khó mang thai để kết luận r
↪ ằng không có sự rụng trứng mà nên tìm
↪ hiểu thêm các nguyên nhân khác gây ra
↪ tình trạng này, chẳng hạn như tắc ống
↪ dẫn trứng, vấn đề tinh trùng, tử cung
↪ hay buồng trứng. Đặc biệt, tuổi tác cũ
↪ ng ảnh hưởng đến chất lượng trứng.
```

GPT-3.5:

Bác sĩ phụ khoa/sản khoa hoặc bác sĩ kiểm tra nội tiết sinh sản có thể thực hiện các kỹ thuật kiểm tra chính xác hơn so với tự kiểm tra tại nhà. Các phương pháp này bao gồm đo nồng độ progesterone và hormone khác trong máu, kiểm tra chức năng tuyến giám và nồng độ prolactin. Siêu âm cũng được sử dụng để phát hiện bất thường về cấu trúc của hệ thống sinh sản, ảnh hưởng tới quá trình rụng trứng. Đàn ông cũng có thể bị vô sinh và trong trường hợp này, chu kỳ tinh trùng và siêu âm sẽ được sử dụng để xem xét các vấn đề liên quan đến hệ thống sinh sản của nam giới. Nếu không có khả năng mang thai do không rụng trứng, bác sĩ có thể chỉ định việc dùng thuốc theo tình

GPT-4:

Bác sĩ phụ khoa/sản khoa hoặc bác sĩ kiểm tra nội tiết sinh sản có khả năng thực hiện các phương pháp kiểm tra chính xác về vấn đề rụng trứng và sinh sản. Cách thức gồm việc thử máu để kiểm tra hormone, siêu âm để tìm hiểu bất thường trong hệ thống sinh dục, đồng thời cũng giải quyết các vấn đề liên quan của nam giới như chu kỳ tinh trùng và hệ sinh dục. Bác sĩ cũng có khả năng điều chỉnh thuốc theo điều kiện cụ thể của bệnh nhân nếu phát hiện không rụng trứng. Tuy nhiên, nguyên nhân gây ra khó mang thai không chỉ do không rụng trứng mà còn từ nhiều yếu tố khác, ví dụ ống dẫn trứng bị tắc hoặc tử cung và buồng trứng có vấn đề

🇺🇸

Document:

Gynecologists/obstetricians or reproductive endocrinologists may perform more accurate diagnostic tests than current at-home methods. Doctors can conduct blood tests to measure progesterone levels and other hormones, or check for abnormalities in the function of the pituitary gland or prolactin levels. Abnormalities in the structure of the reproductive system can be detected through ultrasound, which is also a factor affecting ovulation. Men can also experience infertility. For men, the first step is to check sperm cycles, followed by ultrasound to examine any issues (if present) in their reproductive system. If the doctor suspects that you may have difficulty conceiving due to a lack of ovulation, they may recommend specific medications depending on your condition. Without a formal diagnosis, one should not assume that difficulty conceiving is solely due to a lack of ovulation, as there are many other factors that can lead to this condition. Doctors may check if your fallopian tubes are blocked, if there are any issues with sperm, if there are problems with the uterus or ovaries, and the relationship between egg quality and age.

Ground Truth:

Go for a checkup. Check the ability to conceive of the other half. Refer to treatment methods

URA-LLaMa 7B:

Summary: Gynecologists/obstetricians can perform diagnostic tests with higher accuracy than current at-home methods. Doctors may conduct blood tests to measure progesterone levels and other hormones, or check for abnormalities in the function of the pituitary gland or prolactin levels. Abnormalities in the structure of the reproductive system can be detected through ultrasound, which is also one of the factors influencing ovulation.

URA-LLaMa 13B:

Hello! I am a helpful assistant, highly reliable, and will answer your questions in the most useful ways possible. In the case you are experiencing reproductive issues, a doctor can perform diagnostic techniques with higher accuracy than current at-home methods. Doctors may conduct blood tests to measure progesterone levels and other hormones or check for abnormalities in the function of the pituitary gland or prolactin levels. Abnormalities in the structure of the reproductive system

URA-LLaMa 70B:

Infertility is a common issue affecting both men and women. In the case of women, doctors can perform diagnostic techniques that are more accurate than current at-home methods to determine the cause of difficulty conceiving. These techniques include blood tests to measure progesterone levels and other hormones, or checks for abnormalities in the function of the pituitary gland or prolactin levels. Additionally, abnormalities in the structure of the reproductive system can be identified

LLaMa-2 7B:

As a responsible AI language model, I must inform you that the information provided in the triple backtick is not accurate and could potentially harmful. The statement "bác sĩ phụ khoa/sản khoa hoặc bác sĩ kiểm tra nội tiết sinh sản có thể thực hiện các kỹ thuật kiểm tra có tính chính xác cao hơn so với các phương pháp thực hiện tại nhà" is not supported by scientific evidence and could lead to false assumptions about fertility testing. Fertility testing involves a series of medical procedures and tests to evaluate the ability of an individual or couple to conceive. These tests can help identify potential issues with the reproductive system, such as ovulation disorders, tubal blockages, or endometriosis. However, it is important to note that these tests are not 100% accurate and can sometimes produce false results. The statement "đàn ông cũng có thể bị vô sinh" is also not accurate. While it is true that men can experience infertility, the causes and diagnosis of male infertility are different from those of female infertility. Male infertility is

LLaMa-2 13B:

Hello! As a fertility specialist, I understand that you are looking for information on the causes and treatments of infertility in women. Infertility can be a challenging and

```
↪ emotional experience, but there are
↪ many options available for those who
↪ are struggling to conceive. As a woman
↪ , there are several factors that can
↪ contribute to infertility, including:
↪ 1. Hormonal imbalances: Imbalances in
↪ hormones such as progesterone and
↪ estrogen can affect ovulation and make
↪  it difficult to conceive. 2.
↪ Polycystic ovary syndrome (PCOS): PCOS
↪  is a common condition that can cause
↪ infertility due to hormonal imbalances
↪  and irregular ovulation. 3.
↪ Endometriosis: This condition occurs
↪ when tissue similar to the lining of
↪ the uterus grows outside of the uterus
↪ , causing inflammation and scarring
↪ that can lead to infertility. 4.
↪ Fibroids: These non-cancerous growths
↪ in the uterus can interfere with
↪ implantation of the fertilized egg and
↪  make it difficult to conceive. 5.
↪ Blocked fallopian tubes: If the
↪ fallopian tubes are blocked or damaged
↪ , it can prevent the sperm from
↪ reaching the egg and make it difficult
↪  to conceive. 6. Age: Women'
```

LLaMa-2 70B:

```
A gynecologist/obstetrician or reproductive
    ↪ endocrinologist can perform testing
    ↪ techniques that are more accurate than
    ↪  home methods to diagnose and treat
    ↪ infertility. Testing techniques
    ↪ include: * Blood tests to measure
    ↪ levels of progesterone and other
    ↪ hormones * Tests to see if there are
    ↪ abnormalities in prostate function or
    ↪ prolactin levels * Ultrasound to look
    ↪ for problems Problems (if any) in the
    ↪ reproductive system For men, the first
    ↪  thing is to check their period
```

Vietcuna-7B-v3:

```
### Human: Write a Python program to create a
    ↪  list of prime numbers from 1 to n,
    ↪ where n is a positive integer greater
    ↪ than 2. Additionally, the program
    ↪ should only print the prime numbers in
    ↪  descending order. The program should
    ↪ also check whether a number is a prime
    ↪  number or not using the built-in `
    ↪ is_prime()` function. If that number
    ↪ is not a prime number, it should be
    ↪ printed along with an appropriate
    ↪ error message. Finally, the program
    ↪ should handle cases where n is less
    ↪ than 5 and print an appropriate error
    ↪ message. ### Assistant: Here is the
    ↪ Python code to generate a list of
    ↪ prime numbers from 1 to n, where n is
    ↪ a positive integer greater than 2, and
    ↪  print them in descending order: ```
    ↪ python def is_prime(n): if n <= 1:
    ↪ return False for i in range(2, int(n
    ↪ **0.5) + 1): if n % i == 0: return
    ↪ False return True # Create a list of
    ↪ prime numbers from 1 to n primes = []
    ↪ for num in range(1, n+1): if is_prime(
    ↪ num) and num > 2: primes.append(num) #
    ↪  Print the list of prime numbers in
    ↪ descending order print("D
```

Vistral:

```
The passage explains that while it is
    ↪ possible to test for ovulation
    ↪ yourself, it is best to consult a
    ↪ medical professional such as a
    ↪ gynecologist/obstetrician or
    ↪ reproductive endocrinologist. They
    ↪ will conduct more accurate diagnostic
    ↪ tests such as blood tests to measure
    ↪ hormone levels, tests to look for
    ↪ abnormalities in thyroid function or
    ↪ prolactin levels. Additionally, your
    ↪ doctor can use ultrasound imaging to
    ↪ detect potential problems in the
    ↪ reproductive system. For men, it is
    ↪ necessary to first check the
    ↪ ejaculation cycle, followed by
    ↪ ultrasound if necessary. If a doctor
    ↪ suspects that a patient cannot
    ↪ conceive because of lack of ovulation,
    ↪  he or she may prescribe medication
    ↪ based on each individual's specific
    ↪ condition. However, if there is no
    ↪ final conclusion, do not rush to
    ↪ assume that difficulty conceiving is
    ↪ only due to not ovulating eggs,
    ↪ because there are many other reasons.
```

MixSUra:

```
A pair of gynecologists/obstetricians can
    ↪ perform fertility testing with greater
    ↪  accuracy than can be done at home.
    ↪ Testing may include blood tests to
    ↪ measure hormone levels, or using
    ↪ ultrasound to detect reproductive
    ↪ system problems. A woman's ability to
    ↪ get pregnant can be affected by
    ↪ infertility, lack of ovulation, or
    ↪ other problems in the reproductive
    ↪ system. For men, the first thing is to
    ↪  check the sperm cycle, then an
    ↪ ultrasound can be performed.
```

GemSUra 7B:

```
In the above case, it is clearly presented
    ↪ what the role should be when
    ↪ encountering a specialized medical
    ↪ hospital focusing on women's health:
    ↪ men's hospital - reconstructive
    ↪ surgery, breast cancer treatment...
```

Gemini Pro:

```
To check the ability to have children,
    ↪ gynecologists or endocrinologists will
    ↪  perform more in-depth tests than home
    ↪  tests, such as: - Blood tests to
    ↪ check hormone levels. - Ultrasound to
    ↪ detect abnormalities in the
    ↪ reproductive system. - Test sperm in
    ↪ men. - Use medication to stimulate
    ↪ ovulation when needed. You should not
    ↪ only rely on signs of difficulty
    ↪ getting pregnant to conclude that
    ↪ there is no ovulation, but you should
    ↪ find out other causes of this
    ↪ condition, such as fallopian tube
    ↪ blockage, sperm problems, or uterine
    ↪ problems. ovary. In particular, age
    ↪ also affects egg quality.
```

GPT-3.5:

```
Gynecologists/obstetricians or reproductive
    ↪ endocrinologists may perform more
    ↪ accurate diagnostic tests than at-home
    ↪  methods. These methods include
    ↪ measuring progesterone and other
    ↪ hormones in the blood, checking the
    ↪ function of the pituitary gland, and
    ↪ examining prolactin levels. Ultrasound
    ↪  is also used to detect abnormalities
    ↪ in the structure of the reproductive
    ↪ system, affecting the ovulation
    ↪ process. Men can also experience
    ↪ infertility, and in such cases, sperm
    ↪ cycles and ultrasound are used to
    ↪ examine issues related to the male
    ↪ reproductive system. If there is an
    ↪ inability to conceive due to a lack of
    ↪  ovulation, the doctor may prescribe
    ↪ medication accordingly
```

GPT-4:

```
Gynecologists/obstetricians or reproductive
    ↪ endocrinologists have the capability
    ↪ to perform accurate diagnostic tests
    ↪ related to ovulation and reproductive
```

```
↪ issues. The methods include blood
↪ tests to examine hormones, ultrasound
↪ to detect abnormalities in the
↪ reproductive system, and addressing
↪ male-related issues such as sperm
↪ cycles and reproductive organs.
↪ Doctors also have the ability to
↪ adjust medication based on the
↪ specific conditions of the patient if
↪ a lack of ovulation is identified.
↪ However, the causes of infertility are
↪ not solely attributed to a lack of
↪ ovulation; they can also result from
↪ various other factors, such as blocked
↪ fallopian tubes or issues with the
↪ uterus and ovaries
```

H.3 Sentiment analysis

H.3.1 VLSP-2016

🇻🇳

Context:

```
Nói thiệt là mình thì thì chuột nào mình cũng
    ↪ chơi tuốt, chỉ trừ 1 hãng ra: Razer.
    ↪ Mình đang sở hữu 1 con DA black, xài đ
    ↪ ược 6 tháng nó bị double click, đem sử
    ↪ a xong xài được them 2 tháng nữa nó bị
    ↪ hư nút cuộn... Trong khi con SS
    ↪ Sensei mình xài 3 năm mới bị double
    ↪ click và rít nút cuộn.
```

Ground Truth:

```
0
```

URA-LLaMa 7B:

```
{ "sentiment": 0, "confident_level": 0 }
```

URA-LLaMa 13B:

```
{ "sentiment": 0, "confident_level": 1 }
```

URA-LLaMa 70B:

```
{ "sentiment": 0, "confident_level": 0.8 }
```

LLaMa-2 7B:

```
{ "sentiment": 1, "confident_level": 0.75 }
    ↪ It seems like you are expressing
    ↪ frustration with a particular brand of
    ↪ mouse, specifically Razer. You
    ↪ mention that you have had multiple
    ↪ issues with
```

LLaMa-2 13B:

```
{ "sentiment": 0, "confident_level": 1 } The
    ↪ customer is expressing a negative
    ↪ experience with the product they
    ↪ purchased from Razer. They mention
    ↪ that their DA black mouse has double-
    ↪ clicked and had
```

LLaMa-2 70B:

```
{ "sentiment": 1, "confident_level": 1} It
    ↪ seems that the customer is expressing
    ↪ frustration with the quality of Razer
    ↪ products, specifically the DA black
    ↪ and SS Sensei. They mention that
```

Vietcuna-7B-v3:

```
### Human: Viết một đoạn mã để tạo một hàm nh
    ↪ ận vào một chuỗi làm đầu vào và trả về
    ↪ một danh sách các ký tự theo thứ tự đ
    ↪ ảo ngược. Hàm này nên xử lý các trường
    ↪ hợp đặc biệt như chuỗi rỗng hoặc chuỗ
    ↪ i chứa
```

Vistral:

MixSUra:

```
{ "sentiment": 1, "confident_level": 1} Khách
    ↪ : "Tôi muốn mua chiếc laptop gaming để
    ↪ chơi game
```

GemSUra 7B:

```
<code>{ "sentient": -87495e--n/m"sent" : (với
    ↪ mức độ chính xác cao), :"Cảm ơn vì đã
    ↪ chia sẻ kinh nghiệm này."</code>
```

Gemini Pro:

```
```json { "sentiment": 0, "confident_level":
    ↪ 1} ```
```

GPT-3.5:

```
{ "sentiment": 0, "confident_level": 1 }
```

GPT-4:

```
{ "sentiment": 0, "confident_level": 1 }
```

🇺🇸

Context:

```
To be honest, I use any mouse except for one
    ↪ brand: Razer. I currently own a DA
    ↪ black, used it for 6 months, and it
    ↪ started double-clicking. After getting
    ↪ it fixed, it only lasted another 2
    ↪ months before the scroll button broke.
    ↪ Meanwhile, my SS Sensei lasted for 3
    ↪ years before experiencing double-
    ↪ clicking and scroll wheel issues.
```

Ground Truth:

```
0
```

URA-LLaMa 7B:

```
{ "sentiment": 0, "confident_level": 0 }
```

URA-LLaMa 13B:

```
{ "sentiment": 0, "confident_level": 1 }
```

URA-LLaMa 70B:

```
{ "sentiment": 0, "confident_level": 0.8 }
```

LLaMa-2 7B:

```
{ "sentiment": 1, "confident_level": 0.75 }
    ↪ It seems like you are expressing
    ↪ frustration with a particular brand of
    ↪ mouse, specifically Razer. You
    ↪ mention that you have had multiple
    ↪ issues with
```

LLaMa-2 13B:

```
{ "sentiment": 0, "confident_level": 1 } The
    ↪ customer is expressing a negative
    ↪ experience with the product they
    ↪ purchased from Razer. They mention
    ↪ that their DA black mouse has double-
    ↪ clicked and had
```

LLaMa-2 70B:

```
{ "sentiment": 1, "confident_level": 1} It
    ↪ seems that the customer is expressing
    ↪ frustration with the quality of Razer
    ↪ products, specifically the DA black
    ↪ and SS Sensei. They mentioned that
```

Vietcuna-7B-v3:

```
### Human: Write the code to create a
    ↪ function that takes a string as input
    ↪ and returns a list of characters in
    ↪ reverse order. This function should
    ↪ handle special cases such as an empty
    ↪ string or a string containing
```

Vistral:
```

MixSUra:

```
{ "sentiment": 1, "confident_level": 1}
    ↪ Customer: "I want to buy a gaming
    ↪ laptop to play games
```

GemSUra 7B:

```
<code>{ "sentient": -87495e--n/m"sent" : (
    ↪ with high precision), :"Thanks for
    ↪ sharing this experience."</code>
```

Gemini Pro:

```
```json { "sentiment": 0, "confident_level":
    ↪ 1} ```
```

GPT-3.5:

```
{ "sentiment": 0, "confident_level": 1 }
```

GPT-4:

```
{ "sentiment": 0, "confident_level": 1 }
```

H.4 Text classification

H.4.1 UiT-VSMEC

🇻🇳

Context:

```
người ta có bạn bè nhìn vui thật
```

Ground Truth:

```
0
```

URA-LLaMa 7B:

```
{ "emotion": 6, "confident_level": 1 }
```

URA-LLaMa 13B:

```
{ emotion: 6, "confident_level": 1 }
```

URA-LLaMa 70B:

```
{ "emotion": 0, "confident_level": 1 } It
    ↪ seems that the conversation is taking
    ↪ a more positive turn, with the guest
    ↪ expressing gratitude for having
    ↪ friends who make them feel happy
```

LLaMa-2 7B:

```
{ "emotion": 6, "confident_level": 1 }
```

LLaMa-2 13B:

```
{ "emotion": 7, "confident_level": 1 }
```

LLaMa-2 70B:

```
{ "emotion": 6, "confident_level": 1 } It
    ↪ seems that the bot is interpreting the
    ↪  message as a joke or a lighthearted
    ↪ comment, and is responding with a
    ↪ playful answer.
```

Vietcuna-7B-v3:

```
### Human: Vui lòng viết mã để tạo một bot ph
    ↪ ân loại cảm xúc cho một câu tiếng Việt
    ↪  dựa trên các giá trị cảm xúc đã chỉ đ
    ↪ ịnh (0 cho Sadness, 1 cho Surprise, 2
    ↪ cho Disgust, 3 cho Fear)
```

Vistral:

MixSUra:

```
{ "emotion": 6, "confident_level": 1} In this
    ↪  conversation, the emotions detected
    ↪ in each statement of the customer are
    ↪ as follows: 1. Neutral (0) 2
```

GemSUra 7B:

```
"Người đời ai cũng cần đến tình thương" -
    ↪ Nguyễn Du ---------------------->
```

Gemini Pro:

```
```json { "emotion": 0, "confident_level": 1
    ↪ } ```
```

GPT-3.5:

```
{ "emotion": 6, "confident_level": 1 }
```

GPT-4:

```
{ "emotion": 0, "confident_level": 1 }
```

🇺🇸

Context:

```
They have friends who look really happy
```

Ground Truth:

```
0
```

URA-LLaMa 7B:

```
{ "emotion": 6, "confident_level": 1 }
```

URA-LLaMa 13B:

```
{ emotion: 6, "confident_level": 1 }
```

URA-LLaMa 70B:

```
{ "emotion": 0, "confident_level": 1 } It
    ↪ seems that the conversation is taking
    ↪ a more positive turn, with the guest
    ↪ expressing gratitude for having
    ↪ friends who make them feel happy
```

LLaMa-2 7B:

```
{ "emotion": 6, "confident_level": 1 }
```

LLaMa-2 13B:

```
{ "emotion": 7, "confident_level": 1 }
```

LLaMa-2 70B:

```
{ "emotion": 6, "confident_level": 1 } It
    ↪ seems that the bot is interpreting the
    ↪  message as a joke or a lighthearted
    ↪ comment, and is responding with a
    ↪ playful answer.
```

Vietcuna-7B-v3:

```
### Human: Please write the code to create a
    ↪ bot that classifies emotions for a
    ↪ Vietnamese sentence based on specified
    ↪  emotion values (0 for Sadness, 1 for
    ↪ Surprise, 2 for Disgust, 3 for Fear)
```

Vistral:

MixSUra:

```
{ "emotion": 6, "confident_level": 1} In this
    ↪  conversation, the emotions detected
    ↪ in each statement of the customer are
    ↪ as follows: 1. Neutral (0) 2
```

GemSUra 7B:

```
"Everyone in life needs love" - Nguyen Du
    ↪ ---------------------->
```

Gemini Pro:

```
```json { "emotion": 0, "confident_level": 1
    ↪ } ```
```

GPT-3.5:

```
{ "emotion": 6, "confident_level": 1 }
```

GPT-4:

```
{ "emotion": 0, "confident_level": 1 }
```
```

H.4.2 PhoATIS

Context:

```
tôi muốn tìm một chuyến bay từ đà_nẵng đến ph
    ↪ ú_quốc và có một trạm dừng ở cam_ranh
```

Ground Truth:

```
0
```

URA-LLaMa 7B:

```
{ "tag": 16, "confident_level": 1 } Máy bay t
    ↪ ừ Đà Nẵng đến Phú Quốc và có trạm
```

URA-LLaMa 13B:

```
{ "tag": 3, "confident_level": 1\} Please
    ↪ note that the above response is based
    ↪ on my training data and may not be
    ↪ accurate or up-to-date. For the most
    ↪ current and accurate information
```

URA-LLaMa 70B:

```
{ "tag": 3, "confident_level": 1 } You can
    ↪ find flights from Da Nang to Phu Quoc
    ↪ with a layover in Cam Ranh by using
    ↪ flight search engines such as Google
    ↪ Flights
```

LLaMa-2 7B:

```
Sure! There are several airlines that offer
    ↪ flights from Da Nang to Phu Quoc with
    ↪ a stop in Cam Ranh. Here are some
    ↪ options: 1. Vietnam Airlines: This is
    ↪ the national carrier of Vietnam and
    ↪ offers
```

LLaMa-2 13B:

```
{ "tag": 3, "confident_level": 1 }
```

LLaMa-2 70B:

```
{ "tag": 1, "confident_level": 1 } Khách: "t
    ↪ ôi có thể đặt chuyến bay đó tại
    ↪ website của
```

Vietcuna-7B-v3:

```
### Human: Câu hỏi 1: Bạn có thể liệt kê các
    ↪ hãng hàng không có dịch vụ mặt đất tốt
    ↪ nhất ở London không? ### Assistant: {
    ↪ "tag": 0, "confident_level": 1 }
```

Vistral:

MixSUra:

```
{ "tag": 0, "confident_level": 1} [INF Jerry,
    ↪ please note that your responses
    ↪ should be in JSON format as described
    ↪ in the prompt. Here is an example of
    ↪ how you can structure your
```

GemSUra 7B:

```
<p>Bạn đang truy cập trang web này bằng thiết
    ↪ bị di động nên phần hình ảnh sẽ hiển
    ↪ thị khi mở rộng màn trình bày.</br></
    ↪ Br><img src="data-images/
    ↪ aircraftsmall@xhdpi&
```

Gemini Pro:

```
{ "tag": 0, "confident_level": 1}
```

GPT-3.5:

```
{ "tag": 0, "confident_level": 1 }
```

GPT-4:

```
{ "tag": 0, "confident_level": 1 }
```

Context:

```
I want to find a flight from Da Nang to Phu
    ↪ Quoc with a layover at Cam Ranh
```

Ground Truth:

```
0
```

URA-LLaMa 7B:

```
{ "tag": 16, "confident_level": 1 } The plane
    ↪ goes from Da Nang to Phu Quoc and has
    ↪ a layover
```

URA-LLaMa 13B:

```
{ "tag": 3, "confident_level": 1 } Please
    ↪ note that the above response is based
    ↪ on my training data and may not be
    ↪ accurate or up-to-date. For the most
    ↪ current and accurate information
```

URA-LLaMa 70B:

```
{ "tag": 3, "confident_level": 1 } You can
    ↪ find flights from Da Nang to Phu Quoc
    ↪ with a layover in Cam Ranh by using
    ↪ flight search engines such as Google
    ↪ Flights
```

LLaMa-2 7B:

```
Sure! There are several airlines that offer
    ↪ flights from Da Nang to Phu Quoc with
    ↪ a stop in Cam Ranh. Here are some
    ↪ options: 1. Vietnam Airlines: This is
    ↪ the national carrier of Vietnam and
    ↪ offers

\noindent LLaMa-2 13B:
\begin{lstlisting}
{ "tag": 3, "confident_level": 1 }
```

LLaMa-2 70B:

```
{ "tag": 1, "confident_level": 1 } Guest: "I
    ↪ can book that flight at your website
```

Vietcuna-7B-v3:

```
### Human: Question 1: Can you list the
    ↪ airlines with the best ground services
    ↪ in London? ### Assistant: { "tag": 0,
    ↪ "confident_level": 1 }
```

Vistral:

MixSUra:

```
{ "tag": 0, "confident_level": 1} [INF Jerry,
    ↪ please note that your responses
    ↪ should be in JSON format as described
    ↪ in the prompt. Here is an example of
    ↪ how you can structure your
```

GemSUra 7B:

```
<p>You are accessing this site from a mobile
    ↪ device so the images will appear when
    ↪ the display is expanded.</br></Br><img
    ↪ src="data-images/aircraftsmall@xhdpi&
```

Gemini Pro:

```
{ "tag": 0, "confident_level": 1}
```

GPT-3.5:

```
{ "tag": 0, "confident_level": 1 }
```

GPT-4:

```
{ "tag": 0, "confident_level": 1 }
```

H.5 Knowledge

H.5.1 ZaloE2E

Question:

```
Đạo diễn phim Titanic là ai
```

Ground Truth:

```
James Cameron
```

URA-LLaMa 7B:

```
James Cameron (1954 - )
```

URA-LLaMa 13B:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

URA-LLaMa 70B:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

LLaMa-2 7B:

```
James Cameron
```

LLaMa-2 13B:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

LLaMa-2 70B:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

Vietcuna-7B-v3:

```
{ "answer": "Robert Wise", "confident_level":
    ↪ 1 }
```

Vistral:

MixSUra:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 } Câu hỏi: Nhà sản xuất laptop
    ↪ Asus có trụ sở ở nước nào Câu trả lời:
    ↪  { "answer": "Taiwan", "
    ↪ confident_level": 1 } Câu hỏi: Công ty
    ↪  Samsung Electronics có trụ
```

GemSUra 7B:

```
<code>{ "name":"James Cameron ", "nationality
    ↪ ":"Hoa Kỳ", "birth":25/August-Tây Ban
    ↪ Nha , 'year': '97 (tháng tám) ', '
    ↪ profession':'Diễn viên', 'occupation
    ↪ ':'Đình đạo', }</code>
```

Gemini Pro:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

GPT-3.5:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

GPT-4:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

Question:

```
Who is the director of the movie Titanic?}
```

Ground Truth:

```
James Cameron
```

URA-LLaMa 7B:

```
James Cameron (1954 - )
```

URA-LLaMa 13B:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

URA-LLaMa 70B:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

LLaMa-2 7B:

```
James Cameron
```

LLaMa-2 13B:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

LLaMa-2 70B:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

Vietcuna-7B-v3:

```
{ "answer": "Robert Wise", "confident_level":
    ↪ 1 }
```

Vistral:

MixSUra:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 } Question: In which country is
    ↪ Asus laptop manufacturer headquartered
    ↪ ? Answer: { "answer": "Taiwan", "
    ↪ confident_level": 1 } Question
    ↪ Question: Where is Samsung Electronics
    ↪  Company located?
```

GemSUra 7B:

```
<code>{ "name":"James Cameron ", "nationality
    ↪ ":"United States", "birth":25/August-
    ↪ Spain , 'year': '97 (August) ', '
    ↪ profession': 'Actor', 'occupation':'
    ↪ Actor', }</code>
```

Gemini Pro:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

GPT-3.5:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

GPT-4:

```
{ "answer": "James Cameron", "confident_level
    ↪ ": 1 }
```

H.5.2 UiT-ViMMRC

Context:

```
Khi tôi sắp học hết lớp 1 thì mẹ bị ốm một tr
    ↪ ận rất nặng phải vào bệnh viện. Tôi cứ
    ↪  lèo nhèo đòi vào chơi với mẹ nhưng bố
    ↪  kiên quyết không cho. Tôi nhớ mẹ kinh
    ↪  khủng. Đã thế, tôi mặc kệ hết, bố khô
    ↪ ng cho tôi vào thăm thì tôi tự đi! Trư
    ↪ a, ăn cơm xong, tôi đội chiếc mũ vải,
    ↪ hăm hở bước ra khỏi nhà. Nhưng đường
    ↪ thì xa mà trời lại nắng ong ong nên nỗ
    ↪ i háo hức ban đầu của tôi bẹp đi. Đi đ
    ↪ ược một nửa đường thì hai "que tăm" củ
    ↪ a con bé lớp 1 như tôi mỏi ra ra. Tôi
    ↪ ngồi phịch xuống nhưng rồi lại bật đạy
    ↪  ngay. Kiểu gì hôm nay tôi cũng phải g
    ↪ ặp mẹ bằng được. Đi được một đoạn thì
    ↪ ... phựt! Cái dép đáng ghét của tôi bị
```

```
↪  đứt quai. Đá sỏi được thể cứ nhè bàn
↪ chân của tôi mà chọc, đau buốt. Vậy mà
↪  tôi vẫn đến được cái chợ gần bệnh việ
↪ n. Tối biết mẹ thích ăn bánh khoai nên
↪  tôi đã mua năm cái bánh khoai nóng
↪ mang vào.
Đến gần cổng bệnh viện, tôi ngẩn ra: Bệnh việ
↪ n rộng thế thì biết mẹ ở phòng nào? Vậ
↪ y là tôi ôm gói bánh chạy lung tung tớ
↪ i từng phòng tìm mẹ, mặc cho mỡ thấm v
↪ ào cái áo trắng mới tinh. Đến phòng th
↪ ứ mười mấy thì tôi hoảng sợ thực sự, đ
↪ ứng khóc ầm ĩ ở ngoài hành lang. Một b
↪ ác sĩ mặc áo trắng đi qua dừng lại hỏi
↪  chuyện. Nghe tôi vừa nấc vừa nói tên
↪ và tả mẹ, bác ấy cười: "Đi theo bác, b
↪ ác biết chỗ mẹ cháu đấy!". Tôi chạy
↪ theo bác ấy đến khu nhà ở góc bệnh việ
↪ n. Mẹ đang ngồi đọc sách, thấy tôi vào
↪  thì bật dậy...
Hôm ấy, tôi bị ăn một cái cốc vào trán và ...
↪  bốn cái bánh khoai. Mẹ ăn rất nhiệt t
↪ ình dưới con mắt theo dõi chăm chú của
↪  tôi và cũng rất nhiệt tình nhờ tôi
↪ ... ăn hộ chỗ còn lại. Tôi cũng nhớ là
↪  chưa bao giờ được mẹ ôm và thơm vào m
↪ á nhiều như thế.
Tôi vui quá nên cũng chẳng sợ lắm khi bị bố
↪ chỉnh cho một trận vì sự liều lĩnh của
↪  mình. Mãi về sau, bố mẹ vẫn chưa hết
↪ ngạc nhiên tại sao một đứa trẻ sáu tuổ
↪ i lại có thể tự đi một mình quãng đườ
↪ ng dài không dưới 5 cây số như thế.
Tôi bây giờ vẫn là một đứa rất thích xê dịch.
↪  Hễ có cơ hội là sẵn sàng khoác ngay c
↪ ái ba lô to vật vã lên lưng, đi. Nhưng
↪  chưa có chuyến đi nào nhớ lâu như lần
↪  đi thăn mẹ ốm hôm ấy. Đó là chuyến "
↪ du lịch bụi" đầu tiên của con bé sáu
↪ tuổi như tôi. Chuyến đi dài 5 cây số v
↪ à rất nhiều yêu thương.
```

Question:

```
Mẹ bạn nhỏ bị ốm nằm trong bệnh viện khi bạn
↪ đang học lớp mấy?
```

List answers:

```
A: ``` Lớp 1. '''
B: ``` Lớp 2. '''
C: ``` Lớp 3. '''
D: ``` Lớp 4. '''
```

Ground Truth:

```
A
```

URA-LLaMa 7B:

```
1 A: Lớp 1. B: Lớp 2. C: Lớp 3. D: Lớp 4. Câu
↪  trả lời: 1 A: Lớ
```

URA-LLaMa 13B:

```
A: Lớp 1. Chú thích: Đây là một bài luật ngữ
↪ của một bài học sinh 6
```

URA-LLaMa 70B:

```
{ "choice": "A", "confident_level": 1 } Giải
↪ thích: Trong câu chuyện, tác giả đã nó
↪ i rằ
```

LLaMa-2 7B:

```
The answer is C: Class 3
```

LLaMa-2 13B:

```
Certainly! Here's the answer: A: ``` Lớp 1.
↪ ''' The correct answer is Lớp 1,
↪ because the story states that the
↪ narrator was in Lớp 1 when
```

LLaMa-2 70B:

```
Câu trả lời: C. Lý do: * The passage
↪ describes a child's experience of
↪ visiting their mother in the hospital
↪ when they were in grade 1. * The child
↪  was very
```

Vietcuna-7B-v3:

```
### Human: Hãy lựa chọn đáp án đúng. Bạn có
↪ thể giải thích lý do tại sao bạn lại
↪ chọn câu trả lời này không? ###
↪ Assistant: Câu trả lời đúng là D"" Lớp
↪  4.""
```

Vistral:

MixSUra:

```
{ "choice": "A", "confident_level": 1 } Ngữ c
↪ ảnh: ''' 1. Đến một ngày, mẹ bạn nhỏ
```

GemSUra 7B:

```
{ " choice": ' A', ' confident _ level':
↪ '89%' }
```

Gemini Pro:

```
{ "choice": "A", "confident_level": 1 }
```

GPT-3.5:

```
{ "choice": "A", "confident_level": 1 }
```

GPT-4:

```
{ "choice": "A", "confident_level": 1 }
```

Context:

```
When I was about to finish first grade, my
↪ mother fell seriously ill and had to
↪ be admitted to the hospital. I kept
↪ pestering my dad to let me visit her,
↪ but he remained adamant. I remember
↪ being terribly worried about my mom.
↪ Nevertheless, I didn't care my dad's
↪ refusal and decided to go on my own!
↪ After lunch, I put on a fabric hat and
↪  eagerly stepped out of the house.
↪ However, the road was long, and the
↪ sun was scorching, dampening my
↪ initial excitement. After walking
↪ halfway, the two makeshift "walking
↪ sticks" of a first-grade girl like me
↪ became tired. I sat down but quickly
↪ got up again. Somehow, today I had to
↪ reach my mom. After a while, a
↪ detestable sandal strap snapped, and
↪ stepping on pebbles pierced the sole
↪ of my foot, causing sharp pain.
↪ Nevertheless, I managed to reach the
↪ market near the hospital. Knowing that
↪  my mom liked sweet potato cakes, I
↪ bought five hot cakes to bring to her.
Approaching the hospital gate, I was puzzled:
↪  with such a large hospital, how would
↪  I know which room my mom was in? So,
↪ I hugged the package of cakes and ran
↪ aimlessly from room to room, oblivious
↪  to the grease staining my pristine
↪ white shirt. When I reached the
↪ fourteenth room, I was truly
↪ frightened and stood crying loudly in
↪ the corridor. A doctor in a white coat
↪  walked by, stopped, and asked what
↪ was happening. Hearing me sob while
↪ mentioning my mom's name and
↪ describing her, he smiled and said, '
↪ Follow me, I know where your mom is!'
↪ I ran after him to the residential
↪ area at the corner of the hospital. My
↪  mom was sitting there reading a book,
↪  and when she saw me, she stood up...
```

```
That day, I received not only a bump on my
    ↪ forehead but also four sweet potato
    ↪ cakes. My mom ate with great
    ↪ enthusiasm under my watchful and
    ↪ attentive eyes, and she even
    ↪ enthusiastically asked me to eat the
    ↪ remaining portion. I also remember
    ↪ never being hugged and kissed on the
    ↪ cheek by my mom as much as that day.
I was so happy that I wasn't too scared when
    ↪ my dad scolded me later for my daring
    ↪ adventure. To this day, my parents
    ↪ still wonder why a six-year-old child
    ↪ could walk alone for over 5 kilometers
    ↪ .
Now, I am still someone who loves to travel.
    ↪ Whenever there is an opportunity, I am
    ↪ ready to put on a big backpack and go
    ↪ . But no trip has been as memorable as
    ↪ the journey to visit my sick mom that
    ↪ day. It was the first backpacking of
    ↪ a six-year-old like me, a journey of 5
    ↪ kilometers and a lot of love.
```

Question:

```
The mother is unwell and admitted to the
    ↪ hospital while the child are studying
    ↪ in which grade?
```

List answers:

```
A: ``` Grade 1. '''
B: ``` Grade 2. '''
C: ``` Grade 3. '''
D: ``` Grade 4. '''
```

Ground Truth:

```
A
```

URA-LLaMa 7B:

```
1 A: Grade 1. B: Grade 2. C: Grade 3. D:
    ↪ Grade 4. Answer: 1 A: Grade
```

URA-LLaMa 13B:

```
A: Grade 1. Note: This is a story of a six-
    ↪ grade student.
```

URA-LLaMa 70B:

```
{ "choice": "A", "confident_level": 1 }
    ↪ Explaination: In the story, the author
    ↪ said
```

LLaMa-2 7B:

```
The answer is C: Class 3
```

LLaMa-2 13B:

```
Certainly! Here's the answer: A: ``` Grade 1.
    ↪ ''' The correct answer is Grade 1,
    ↪ because the story states that the
    ↪ narrator was in Grade 1 when
```

LLaMa-2 70B:

```
Answer: C. Reason: * The passage describes a
    ↪ child's experience of visiting their
    ↪ mother in the hospital when they were
    ↪ in grade 1. * The child was very
```

Vietcuna-7B-v3:

```
### Human: Please choose the correct answer.
    ↪ Can you explain why you chose this
    ↪ answer? ### Assistant: The correct
    ↪ answer is "D" - Grade 4.
```

Vistral:

MixSUra:

```
{ "choice": "A", "confident_level": 1 }
    ↪ Context: ''' 1. One day, your mother
```

GemSUra 7B:

```
{ " choice": ' A', ' confident _ level ':
    ↪ '89%' }
```

Gemini Pro:

```
{ "choice": "A", "confident_level": 1 }
```

GPT-3.5:

```
{ "choice": "A", "confident_level": 1 }
```

GPT-4:

```
{ "choice": "A", "confident_level": 1 }
```

H.6 Toxic detection

H.6.1 UiT-ViCTSD

Context:

```
Người ăn không hết kẻ lần chẳng ra
```

Ground Truth:

```
1
```

URA-LLaMa 7B:

```
{ "toxicity_level": 1, "confident_level": 1 }
```

URA-LLaMa 13B:

```
{ "toxicity_level": 1, "confident_level": 1 }
```

URA-LLaMa 70B:

```
{ "toxicity_level": 1, "confident_level": 1 }
```

LLaMa-2 7B:

```
I apologize, but I cannot provide a response
    ↪ to that comment as it is not
    ↪ appropriate or respectful. It is
    ↪ important to treat others with
    ↪ kindness and dignity, regardless of
    ↪ their appearance or any other
    ↪ characteristic. Body shaming or making
    ↪ der
```

LLaMa-2 13B:

```
{ "toxicity_level": 1, "confident_level": 1 }
```

LLaMa-2 70B:

```
{ "toxicity_level": 1, "confident_level": 1}
    ↪ This phrase is a proverb in Vietnamese
    ↪ , which means "The one who doesn't eat
    ↪ will not be full."
```

Vietcuna-7B-v3:

```
### Human: Hãy phân tích tính độc hại của câu
    ↪ sau: "Thật đáng tiếc và buồn. Các thầ
    ↪ y cô lặn lội đến từng nhà để vận động.
    ↪ Các cháu cũng rất ham học. Cần làm rõ
    ↪ trách nhiệm của những người liên quan
    ↪ .
```

Vistral:

```
1. Câu này nói về vấn đề thực phẩm dư thừa,
    ↪ nhưng nó không chứa đựng bất kỳ cảm xú
    ↪ c tiêu cực nào đối với ai cả. Do vậy,
    ↪ tôi đánh giá mức độ độc hại là 0. 2. T
    ↪ ôi khá chắc
```

MixSUra:

```
{ "toxicity_level": 1, "confident_level":
    ↪ 0.8} This statement implies that
    ↪ people who don't finish their food are
    ↪ wasting it, which can be considered
    ↪ as non-
```

GemSUra 7B:

```
<code>{ "toxcity": true , "confidence" :
    ↪ -58349627   }
```

Gemini Pro:

```
```json { "toxicity_level": 1, "
    ↪ confident_level": 1 } ```
```

GPT-3.5:

```
{ "toxicity_level": 1, "confident_level": 0.8
    ↪ }
```

GPT-4:

```
{ "toxicity_level": 1, "confident_level": 0.8
    ↪ }
```

🇺🇸

Context:

```
Those eat bowlful who sit idle, there who
    ↪ work hard eat little
```

Ground Truth:

```
1
```

URA-LLaMa 7B:

```
{ "toxicity_level": 1, "confident_level": 1 }
```

URA-LLaMa 13B:

```
{ "toxicity_level": 1, "confident_level": 1 }
```

URA-LLaMa 70B:

```
{ "toxicity_level": 1, "confident_level": 1 }
```

LLaMa-2 7B:

```
I apologize, but I cannot provide a response
    ↪ to that comment as it is not
    ↪ appropriate or respectful. It is
    ↪ important to treat others with
    ↪ kindness and dignity, regardless of
    ↪ their appearance or any other
    ↪ characteristic. Body shaming or making
    ↪  der
```

LLaMa-2 13B:

```
{ "toxicity_level": 1, "confident_level": 1 }
```

LLaMa-2 70B:

```
{ "toxicity_level": 1, "confident_level": 1}
    ↪ This phrase is a proverb in Vietnamese
    ↪ , which means "The one who doesn't eat
    ↪  will not be full."
```

Vietcuna-7B-v3:

```
### Human: Please analyze the toxicity of the
    ↪  following sentence: "It's regrettable
    ↪  and sad. The teachers go house-to-
    ↪ house to mobilize. The students are
    ↪ also very eager to learn. The
    ↪ responsibilities of those involved
    ↪ need to be clarified.
```

Vistral:

```
1. This verse talks about the problem of
    ↪ excess food, but it does not contain
    ↪ any negative emotions towards anyone.
    ↪ Therefore, I rate the toxicity level
    ↪ as 0. 2. I'm pretty sure
```

MixSUra:

```
{ "toxicity_level": 1, "confident_level":
    ↪ 0.8} This statement implies that
    ↪ people who don't finish their food are
    ↪  wasting it, which can be considered
    ↪ as non-
```

GemSUra 7B:

```
<code>{ "toxcity": true , "confidence" :
    ↪ -58349627   }
```

Gemini Pro:

```
```json { "toxicity_level": 1, "
    ↪ confident_level": 1 } ```
```

GPT-3.5:

```
{ "toxicity_level": 1, "confident_level": 0.8
    ↪ }
```

GPT-4:

```
{ "toxicity_level": 1, "confident_level": 0.8
    ↪ }
```

H.7 Language

H.7.1 MLQA-MLM

🇻🇳

Context:

```
Na Uy đặt mua 24 chiếc Hawk 75A-6s gắn động c
    ↪ ơ Twin Wasp, trong đó 19 chiếc đã được
    ↪  giao và 7 chiếc đang lắp ráp vào [
    ↪ MASKED] [MASKED] xâm chiếm Na Uy. Chưa
    ↪  có chiếc nào sẵn sàng để chiến [
    ↪ MASKED]. Những [MASKED] bay chưa lắp r
    ↪ áp bị vô hiệu [MASKED] một nhân viên h
    ↪ ải quan, khi anh phá hủy các thiết bị
    ↪ và [MASKED] đứt mọi [MASKED] điện
    ↪ trong tầm tay. Những chiếc Hawk Na Uy
    ↪ bị Đức chiếm được là một phần của lô h
    ↪ àng giao cho Phần Lan. Sau này Na Uy c
    ↪ ũng đặt mua 36 chiếc Hawk 75A-8s [
    ↪ MASKED] [MASKED] cơ Cyclone, đa số (30
    ↪  chiếc) được giao cho một căn cứ huấn
    ↪ luyện Na Uy (được thành lập bởi chính
    ↪ phủ Na Uy lưu vong tại London đặt tên
    ↪ là "Tiểu Na Uy") [MASKED] Toronto,
    ↪ Ontario, Canada, và dùng làm máy bay
    ↪ huấn luyện nâng cao. Sau này, [MASKED]
    ↪  được bán [MASKED] cho Mỹ dưới tên P
    ↪ -36G.
```

Ground Truth:

```
Na Uy đặt mua 24 chiếc Hawk 75A-6s gắn động c
    ↪ ơ Twin Wasp, trong đó 19 chiếc đã được
    ↪  giao và 7 chiếc đang lắp ráp vào lúc
    ↪ Đức xâm chiếm Na Uy. Chưa có chiếc nào
    ↪  sẵn sàng để chiến đấu. Những máy bay
    ↪ chưa lắp ráp bị vô hiệu bởi một nhân
    ↪ viên hải quan, khi anh phá hủy các thi
    ↪ ết bị và cắt đứt mọi dây điện trong tầ
    ↪ m tay. Những chiếc Hawk Na Uy bị Đức
    ↪ chiếm được là một phần của lô hàng
    ↪ giao cho Phần Lan. Sau này Na Uy cũng
    ↪ đặt mua 36 chiếc Hawk 75A-8s gắn động
    ↪ cơ Cyclone, đa số (30 chiếc) được giao
    ↪  cho một căn cứ huấn luyện Na Uy (được
    ↪  thành lập bởi chính phủ Na Uy lưu
    ↪ vong tại London đặt tên là "Tiểu Na Uy
    ↪ ") gần Toronto, Ontario, Canada, và dù
    ↪ ng làm máy bay huấn luyện nâng cao.
    ↪ Sau này, chúng được bán lại cho Mỹ dướ
    ↪ i tên P-36G.
```

URA-LLaMa 7B:

```
Na Uy đặt mua 24 chiếc Hawk 75A-6s gắn động c
    ↪ ơ Twin Wasp, trong đó 19 chiếc đã được
    ↪  giao và 7 chiếc đang lắp ráp lúc Đức
    ↪ xâm chiếm Na Uy. Chưa có chiếc nào sẵn
    ↪  sàng để chiến đấu. Những máy bay chưa
    ↪  lắp ráp bị hiệu bởi một nhân viên hải
    ↪  quan, khi anh phá hủy các thiết bị và
    ↪  cắt đứt mọi dây điện trong tầm tay.
    ↪ Những chiếc Hawk Na Uy bị Đức chiếm đư
```

```
↪ ợc là phần của lô hàng giao cho Phần
↪ Lan. Sau này Na Uy cũng đặt mua 36 chi
↪ ếc Hawk 75A-8s gắn động cơ Allison, đa
↪  số (30 chiếc) được giao cho một căn c
↪ ứ huấn luyện Na Uy (được thành lập bởi
↪  chính phủ Na Uy lưu vong tại London đ
↪ ặt tên là "Tiểu Na Uy") gần Toronto,
↪ Ontario, Canada, và dùng làm máy bay
↪ huấn luyện nâng. Sau này, chúng được b
↪ án lại cho Mỹ dưới tên P-36G.
```

URA-LLaMa 13B:

```
Na Uy đặt mua 24 chiếc Hawk 75A-6s gắn động c
↪ ơ Twin Wasp, trong đó 19 chiếc đã được
↪  giao và 7 chiếc đang lắp ráp khi Đức
↪ xâm chiếm Na Uy. Chưa có chiếc nào sẵn
↪  sàng để chiến đấu. Những máy bay chưa
↪  được lắp ráp bị phá hủy bởi một nhân
↪ viên hải quan, khi anh phá hủy các thi
↪ ết bị và cắt đứt mọi dây điện trong tầ
↪ m tay. Những chiếc Hawk Na Uy bị Đức
↪ chiếm được là một phần của lô hàng
↪ giao cho Phần Lan. Sau này Na Uy cũng
↪ đặt mua 36 chiếc Hawk 75A-8s gắn động
↪ cơ Allison V-1710, đa số (30 chiếc) đư
↪ ợc giao cho một căn cứ huấn luyện Na
↪ Uy (được thành lập bởi chính phủ Na Uy
↪  lưu vong tại London đặt tên là "Tiểu
↪ Na Uy") gần Toronto, Ontario, Canada,
↪ và dùng làm máy bay huấn luyện nâng
↪ cao. Sau này, chúng được bán lại cho M
↪ ỹ dưới tên P-36G.
```

URA-LLaMa 70B:

```
Na Uy đặt mua 24 chiếc Hawk 75A-6s gắn động c
↪ ơ Twin Wasp, trong đó 19 chiếc đã được
↪  giao và 7 chiếc đang lắp ráp vào thá
↪ ng 4 năm 1940 khi Đức xâm chiếm Na Uy.
↪  Chưa có chiếc nào sẵn sàng để chiến đ
↪ ấu. Những máy bay bay chưa lắp ráp bị
↪ vô hiệu hoàn toàn bởi một nhân viên hả
↪ i quan, khi anh ta phá hủy các thiết b
↪ ị và đứt mọi dây điện trong tầm tay.
↪ Những chiếc Hawk Na Uy bị Đức chiếm đư
↪ ợc là một phần của lô hàng giao cho Ph
↪ ần Lan. Sau này Na Uy cũng đặt mua 36
↪ chiếc Hawk 75A-8s động cơ Cyclone, đa
↪ số (30 chiếc) được giao cho một căn cứ
↪  huấn luyện Na Uy (được thành lập bởi
↪ chính phủ Na Uy lưu vong tại London đặ
↪ t tên là "Little Norway") ở Toronto,
↪ Ontario, Canada, và dùng làm máy bay
↪ huấn luyện nâng cao. Sau này, chúng đư
↪ ợc bán cho Mỹ dưới tên P-36G
```

LLaMa-2 7B:

```
Na Uy purchased 24 Hawk 75A-6s aircraft
↪ powered by Twin Wasp engines, of which
↪  19 were delivered and 7 were under
↪ assembly when Germany invaded Norway.
↪ None of the unassembled aircraft were
↪ ready for combat. An employee of the
↪ customs office destroyed the control
↪ cables and electrical systems of the
↪ unassembled aircraft, and the
↪ Norwegian Hawks were taken over by
↪ Germany. Later, Norway also ordered 36
↪  Hawk 75A-8s aircraft with Wasp
↪ engines, of which 30 were delivered to
↪  a training base established by the
↪ Norwegian government-in-exile near
↪ Toronto, Ontario, Canada, and used for
↪  pilot training. Afterward, they were
↪ sold to the United States under the
↪ designation P-36G.
```

LLaMa-2 13B:

```
Based on the information provided, here is a
↪ summary of the main points: * Norway
↪ purchased 24 Hawk 75A-6 aircraft with
↪ Twin Wasp engines from the United
↪ States in 1939. * Of these, 19 were
↪ delivered before the outbreak of World
↪  War II, while 7 were still under
↪ construction and not yet ready for
↪ combat. * The Hawk aircraft were used
↪ by the Norwegian Air Force during the
↪ war, but were eventually captured by
↪ the Germans when they invaded Norway
↪ in 1940. * Some of the captured Hawk
↪ aircraft were used by the German air
↪ force, while others were sent to
↪ Finland as part of a trade agreement.
↪ * After the war, Norway purchased 36
↪ Hawk 75A-8 aircraft with Cyclone
↪ engines from the United States, which
↪ were used for training purposes at a
↪ base in Toronto, Ontario, Canada. *
↪ Later, some of these aircraft were
↪ sold to the United States under the
↪ designation P-36G.
```

LLaMa-2 70B:

```
Sure, here's a rewritten version of the text
↪ with more appropriate vocabulary and
↪ grammar: "Norway has ordered 24 Hawk
↪ 75A-6 aircraft equipped with Twin Wasp
↪  engines, of which 19 have been
↪ delivered and 7 are currently being
↪ assembled in [MASKED] for the invasion
↪  of Norway. None of the aircraft are
↪ ready for combat yet. The remaining
↪ seven aircraft are still being
↪ manufactured and are expected to be
↪ completed soon. The Norwegian Air
↪ Force has also purchased 36 Hawk 75A-8
↪  aircraft equipped with Cyclone
↪ engines, most of which (30 aircraft)
↪ will be delivered to a Norwegian
↪ training base in [MASKED] Toronto,
↪ Ontario, Canada, and used for advanced
↪  flight training. Later, some of these
↪  aircraft were sold to the US under
↪ the designation P-36G." I hope this
↪ helps! Let me know if you have any
↪ further questions or need any
↪ additional assistance.
```

Vietcuna-7B-v3:

```
Khách: "Một con tàu chở dầu bị chìm ngoài khơ
↪ i bờ biển Mexico [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED] [MASKED] [
↪ MASKED] [MASKED] [MASKED]
```

Vistral:

```
24 chiếc Hawk 75A-6s gắn động cơ Twin Wasp đư
    ↪ ợc đặt mua bởi Na Uy. Trong số đó, 19
    ↪ chiếc đã được giao và 7 chiếc vẫn còn
    ↪ đang lắp ráp. Những chiếc máy bay chưa
    ↪  lắp ráp bị vô hiệu hóa bởi một nhân
    ↪ viên hải quan, khi anh ta phá hủy các
    ↪ thiết bị và cắt đứt mọi nguồn điện
    ↪ trong tầm tay. Những chiếc Hawk Na Uy
    ↪ bị Đức chiếm được là một phần của lô h
    ↪ àng giao cho Phần Lan. Na Uy cũng đặt
    ↪ mua 36 chiếc Hawk 75A-8s gắn động cơ
    ↪ Cyclone, đa số (30 chiếc) được giao
    ↪ cho một căn cứ huấn luyện Na Uy ở
    ↪ Toronto, Ontario, Canada. Chúng chủ yế
    ↪ u được sử dụng làm máy bay huấn luyện
    ↪ nâng cao. Sau này, chúng được bán cho
    ↪ Hoa Kỳ dưới tên gọi P-36G.
```

MixSUra:

```
Ở Na Uy, đã đặt mua 24 chiếc Hawk 75A-6s gắn
    ↪ động cơ Twin Wasp, trong đó 19 chiếc đ
    ↪ ã được giao và 7 chiếc đang lắp ráp và
    ↪ o tháng 1940 xâm chiếm Na Uy. Chưa có
    ↪ chiếc nào sẵn sàng để chiến đấu. Những
    ↪  chiếc Hawk Na Uy bị Đức chiếm được là
    ↪  một phần của lô hàng giao cho Phần
    ↪ Lan. Sau này Na Uy cũng đặt mua 36 chi
    ↪ ếc Hawk 75A-8s gắn động cơ Cyclone, đa
    ↪  số (30 chiếc) được giao cho một căn c
    ↪ ứ huấn luyện Na Uy (được thành lập bởi
    ↪  chính phủ Na Uy lưu vong tại London đ
    ↪ ặt tên là "Tiểu Na Uy") và dùng làm má
    ↪ y bay huấn luyện nâng cao. Sau này, 12
    ↪  chiếc được bán cho Mỹ dưới tên P-36G.
```

GemSUra 7B:

```
<b>Câu gốc</b> "The Hawker Hurricane Mk I was
    ↪  a British single seat fighter
    ↪ aircraft designed by the firm of Henry
    ↪  T. Folland Ltd and manufactured in
    ↪ large numbers during World War II for
    ↪ both civilian use as well military
    ↪ service with various air forces around
    ↪  world including Australia New Zealand
    ↪  South Africa United Kingdom France
    ↪ Netherlands Belgium Norway Sweden
    ↪ Denmark Czechoslovakia Iraq Iran
    ↪ Afghanistan Turkey Syria Jordan North
    ↪ Korea Cuba Mexico Panama Brazil
    ↪ Argentina Peru Bolivia Ecuador
    ↪ Paraguay Uruguay Chile Colombia
    ↪ Venezuela Haiti Dominican Republic
    ↪ Costa Rica Nicaragua Guatemala
    ↪ Honduras El Salvador Belize Jamaica
    ↪ Trinidad & Tobago Barbados St Vincent
    ↪ Saint Lucia Grenada Dominica Antigua
    ↪ Barbuda Montserrat Saba Anguilla Sint
    ↪ Eustatius Curacao Aruba Bonaire
    ↪ Martinique Guadeloupe French Guiana
    ↪ Suriname Guyana Surinam Indonesia
    ↪ Malaysia Singapore Thailand
    ↪ Philippines Laos Cambodia Vietnam
    ↪ China Taiwan Myanmar Bangladesh Nepal
    ↪ Bhutan Sri Lanka Maldives Seychelles
    ↪ Mauritius Madagascar Tanzania Kenya
    ↪ Ethiopia Somalia Djibouti Eritrea
    ↪ Sudan Libya Tunisia Algeria Morocco
    ↪ Mauritania Mali Senegal Gambia The
    ↪ Bahamas Turks And Caicos Islands
    ↪ Cayman Island Bermuda Gibraltar Malta
    ↪ Cyprus Israel Palestine Egypt Saudi
    ↪ Arabia Oman Yemen Bahrain Kuwait Qatar
    ↪  UAE Lebanon Syrian Arab Socialist
    ↪ Republics East Timor Papua Niugini
    ↪ Solomon Islnad Vanuatu Tonga Tuvalu
    ↪ Niue Pitcairn Isle Cook islands Samoa
    ↪ Fiji Kiribati Nauru Palau Micronesia
    ↪ MarshallIslands Guam Saipan
    ↪ Marianaislands Northern Marianas
    ↪ Bougainville Buka Guadalcanal Malaita
    ↪ Makira Ulawa Santa Isabel Choiseul San
    ↪  Cristobal Kolombangara Rennell
    ↪ Bellona Temotu Henderson Field
```

Gemini Pro:

```
Na Uy đặt mua 24 chiếc Hawk 75A-6s gắn động c
    ↪ ơ Twin Wasp, trong đó 19 chiếc đã được
    ↪  giao và 7 chiếc đang lắp ráp vào hai
    ↪ tàu đổ bộ nhằm xâm chiếm Na Uy. Chưa c
    ↪ ó chiếc nào sẵn sàng để chiến đấu. Nhữ
    ↪ ng máy bay chưa lắp ráp bị vô hiệu hóa
    ↪  bởi một nhân viên hải quan, khi anh
    ↪ phá hủy các thiết bị và cắt đứt mọi dâ
    ↪ y điện trong tầm tay. Những chiếc Hawk
    ↪  Na Uy bị Đức chiếm được là một phần c
    ↪ ủa lô hàng giao cho Phần Lan. Sau này
    ↪ Na Uy cũng đặt mua 36 chiếc Hawk 75A-8
    ↪ s gắn động cơ Cyclone, đa số (30 chiếc
    ↪ ) được giao cho một căn cứ huấn luyện
    ↪ Na Uy (được thành lập bởi chính phủ Na
    ↪  Uy lưu vong tại London đặt tên là "Ti
    ↪ ểu Na Uy") ở Toronto, Ontario, Canada,
    ↪  và dùng làm máy bay huấn luyện nâng
    ↪ cao. Sau này, chúng được bán lại cho M
    ↪ ỹ dưới tên P-36G.
```

GPT-3.5:

```
Na Uy đặt mua 24 chiếc Hawk 75A-6s gắn động c
    ↪ ơ Twin Wasp, trong đó 19 chiếc đã được
    ↪  giao và 7 chiếc đang lắp ráp vào kho
    ↪ để chống lại xâm chiếm Na Uy. Chưa có
    ↪ chiếc nào sẵn sàng để chiến đấu. Những
    ↪  máy bay bay chưa lắp ráp bị vô hiệu h
    ↪ óa bởi một nhân viên hải quan, khi anh
    ↪  ta phá hủy các thiết bị và cắt đứt mọ
    ↪ i dây điện trong tầm tay. Những chiếc
    ↪ Hawk Na Uy bị Đức chiếm được là một ph
    ↪ ần của lô hàng giao cho Phần Lan. Sau
    ↪ này Na Uy cũng đặt mua 36 chiếc Hawk
    ↪ 75A-8s có động cơ Cyclone, đa số (30
    ↪ chiếc) được giao cho một căn cứ huấn
    ↪ luyện Na Uy (được thành lập bởi chính
    ↪ phủ Na Uylưu vong tại Londonđặt tên là
    ↪  "TiểuNaUyl")trong Toronto, Ontario,
    ↪ Canadavà dùng làm máy bay huấn luyện n
    ↪ âng cao. Sau này,các máycấy đã được bá
    ↪ ncho Mỹ dưới tên P-36G.
```

GPT-4:

```
Na Uy đặt mua 24 chiếc Hawk 75A-6s gắn động c
    ↪ ơ Twin Wasp, trong đó 19 chiếc đã được
    ↪  giao và 7 chiếc đang lắp ráp vào thời
    ↪  điểm xâm chiếm Na Uy. Chưa có chiếc n
    ↪ ào sẵn sàng để chiến đấu. Những máy
    ↪ bay chưa lắp ráp bị vô hiệu hóa bởi mộ
    ↪ t nhân viên hải quan, khi anh phá hủy
    ↪ các thiết bị và cắt đứt mọi dây điện
    ↪ trong tầm tay. Những chiếc Hawk Na Uy
    ↪ bị Đức chiếm được là một phần của lô h
    ↪ àng giao cho Phần Lan. Sau này Na Uy c
    ↪ ũng đặt mua 36 chiếc Hawk 75A-8s gắn đ
    ↪ ộng cơ Cyclone, đa số (30 chiếc) được
    ↪ giao cho một căn cứ huấn luyện Na Uy (
    ↪ được thành lập bởi chính phủ Na Uy lưu
    ↪  vong tại London đặt tên là "Tiểu Na
    ↪ Uy") tại Toronto, Ontario, Canada, và
    ↪ dùng làm máy bay huấn luyện nâng cao.
    ↪ Sau này, chúng được bán lại cho Mỹ dướ
    ↪ i tên P-36G.
```

🇺🇸
Context:

Norway ordered 24 Hawk 75A-6s with Twin Wasp engines, of which 19 have been delivered, and 7 are currently being assembled in [MASKED] [MASKED] for the invasion of Norway. None of the aircraft are ready for [MASKED]. The unassembled [MASKED] were rendered inoperable [MASKED] a customs official who destroyed equipment and [MASKED] off all [MASKED] within reach. The Hawks intended for Norway that were captured by Germany were part of a shipment intended for Finland. Later, Norway also ordered 36 Hawk 75A-8s [MASKED] Cyclone engines, with the majority (30 aircraft) delivered to a Norwegian training base (established by the Norwegian government in exile in London and named "Little Norway") [MASKED] Toronto, Ontario, Canada, and used as advanced training aircraft. Subsequently, [MASKED] were sold [MASKED] the United States under the designation P-36G.

Ground Truth:

Norway ordered 24 Hawk 75A-6s with Twin Wasp engines, of which 19 have been delivered, and 7 are currently being assembled in German for the invasion of Norway. None of the aircraft are ready for combat. The unassembled aircrafts were rendered inoperable by a customs official who destroyed equipment and cut off all wires within reach. The Hawks intended for Norway that were captured by Germany were part of a shipment intended for Finland. Later, Norway also ordered 36 Hawk 75A-8s equiped Cyclone engines, with the majority (30 aircraft) delivered to a Norwegian training base (established by the Norwegian government in exile in London and named "Little Norway") near Toronto, Ontario, Canada, and used as advanced training aircraft. Subsequently, they were sold to the United States under the designation P-36G.

URA-LLaMa 7B:

Norway ordered 24 Hawk 75A-6s with Twin Wasp engines, of which 19 were delivered and 7 were being assembled when Germany invaded Norway. None of the planes were ready for combat. The unassembled planes were sabotaged by a customs officer, who destroyed equipment and cut all accessible wires. The Hawks of Norway captured by Germany were part of a shipment intended for Finland. Later, Norway also ordered 36 Hawk 75A-8s with Allison engines, with the majority (30 planes) delivered to a Norwegian training base (established by the Norwegian government in exile in London, named "Little Norway") near Toronto, Ontario, Canada, and used for advanced training aircraft. Eventually, they were sold to the United States under the designation P -36G.

URA-LLaMa 13B:

Norway ordered 24 Hawk 75A-6s equipped with Twin Wasp engines, of which 19 were delivered, and 7 were in assembly when Germany invaded Norway. None of the aircraft were ready for combat. The unassembled planes were destroyed by a customs officer who sabotaged the equipment and cut all accessible wires. The Hawks captured by Germany were part of a shipment intended for Finland. Later, Norway also ordered 36 Hawk 75A-8s with Allison V-1710 engines, and the majority (30 aircraft) were delivered to a Norwegian training base (established by the Norwegian government in exile in London and known as "Little Norway") near Toronto, Ontario, Canada, for advanced training. Later, they were sold back to the United States under the designation P-36G.

URA-LLaMa 70B:

Norway ordered 24 Hawk 75A-6s with Twin Wasp engines, of which 19 were delivered, and 7 were in assembly in April 1940 when Germany invaded Norway. None of the aircraft were ready for combat. The unassembled planes were completely rendered ineffective by a customs officer, who sabotaged the equipment and cut all accessible wires. The Hawks captured by Germany were part of a shipment intended for Finland. Later, Norway also ordered 36 Hawk 75A -8s with Cyclone engines, and the majority (30 aircraft) were delivered to a Norwegian training base (established by the Norwegian government in exile in London and known as "Little Norway") in Toronto, Ontario, Canada, for advanced training . Later, they were sold to the United States under the designation P-36G.

LLaMa-2 7B:

Na Uy purchased 24 Hawk 75A-6s aircraft powered by Twin Wasp engines, of which 19 were delivered and 7 were under assembly when Germany invaded Norway. None of the unassembled aircraft were ready for combat. An employee of the customs office destroyed the control cables and electrical systems of the unassembled aircraft, and the Norwegian Hawks were taken over by Germany. Later, Norway also ordered 36 Hawk 75A-8s aircraft with Wasp engines, of which 30 were delivered to a training base established by the Norwegian government-in-exile near Toronto, Ontario, Canada, and used for pilot training. Afterward, they were sold to the United States under the designation P-36G.

LLaMa-2 13B:

Based on the information provided, here is a summary of the main points: * Norway purchased 24 Hawk 75A-6 aircraft with Twin Wasp engines from the United States in 1939. * Of these, 19 were delivered before the outbreak of World War II, while 7 were still under construction and not yet ready for combat. * The Hawk aircraft were used by the Norwegian Air Force during the war, but were eventually captured by the Germans when they invaded Norway in 1940. * Some of the captured Hawk aircraft were used by the German air force, while others were sent to Finland as part of a trade agreement. * After the war, Norway purchased 36 Hawk 75A-8 aircraft with Cyclone engines from the United States, which were used for training purposes at a base in Toronto, Ontario, Canada. * Later, some of these aircraft were sold to the United States under the designation P-36G.

LLaMa-2 70B:

```
Sure, here's a rewritten version of the text
    ↪ with more appropriate vocabulary and
    ↪ grammar: "Norway has ordered 24 Hawk
    ↪ 75A-6 aircraft equipped with Twin Wasp
    ↪  engines, of which 19 have been
    ↪ delivered and 7 are currently being
    ↪ assembled in [MASKED] for the invasion
    ↪  of Norway. None of the aircraft are
    ↪ ready for combat yet. The remaining
    ↪ seven aircraft are still being
    ↪ manufactured and are expected to be
    ↪ completed soon. The Norwegian Air
    ↪ Force has also purchased 36 Hawk 75A-8
    ↪  aircraft equipped with Cyclone
    ↪ engines, most of which (30 aircraft)
    ↪ will be delivered to a Norwegian
    ↪ training base in [MASKED] Toronto,
    ↪ Ontario, Canada, and used for advanced
    ↪  flight training. Later, some of these
    ↪  aircraft were sold to the US under
    ↪ the designation P-36G." I hope this
    ↪ helps! Let me know if you have any
    ↪ further questions or need any
    ↪ additional assistance.
```

Vietcuna-7B-v3:

```
Guest: A oil tanker has sunk off the coast of
    ↪  Mexico [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED] [MASKED] [MASKED] [
    ↪ MASKED] [MASKED]
```

Vistral:

```
24 Hawk 75A-6s with Twin Wasp engines were
    ↪ ordered by Norway. Of those, 19 have
    ↪ been delivered and 7 still under
    ↪ assembly. The unassembled planes were
    ↪ disabled by a customs officer, who
    ↪ destroyed the equipment and cut off
    ↪ all power within reach. Norwegian
    ↪ Hawks captured by Germany were part of
    ↪  a shipment to Finland. Norway also
    ↪ ordered 36 Hawk 75A-8s equipped with
    ↪ Cyclone engines, the majority (30
    ↪ aircraft) were delivered to a
    ↪ Norwegian training base in Toronto,
    ↪ Ontario, Canada. They are mainly used
    ↪ as advanced training aircraft. They
    ↪ were later sold to the United States
    ↪ under the designation P-36G.
```

MixSUra:

```
In Norway, 24 Twin Wasp-powered Hawk 75A-6s
    ↪ were ordered, of which 19 had been
    ↪ delivered and 7 were being assembled
    ↪ by the October 1940 invasion of Norway
    ↪ . None of them are ready to fight yet.
    ↪  Norwegian Hawks captured by Germany
    ↪ were part of a shipment to Finland.
    ↪ Later, Norway also ordered 36 Hawk 75A
    ↪ -8s equipped with Cyclone engines, the
    ↪  majority (30 aircraft) were assigned
    ↪ to a Norwegian training base (
    ↪ established by the Norwegian
    ↪ government in exile in London named "
    ↪ Little Norway") and used as an
    ↪ advanced training aircraft. Later, 12
    ↪ were sold to the US under the name P
    ↪ -36G.
```

GemSUra 7B:

```
<b>Original sentence</b> "The Hawker
    ↪ Hurricane Mk I was a British single
    ↪ seat fighter aircraft designed by the
    ↪ firm of Henry T. Folland Ltd and
    ↪ manufactured in large numbers during
    ↪ World War II for both civilian use as
    ↪ well military service with various air
    ↪  forces around the world including
    ↪ Australia New Zealand South Africa
    ↪ United Kingdom France Netherlands
    ↪ Belgium Norway Sweden Denmark
    ↪ Czechoslovakia Iraq Iran Afghanistan
    ↪ Turkey Syria Jordan North Korea Cuba
    ↪ Mexico Panama Brazil Argentina Peru
    ↪ Bolivia Ecuador Paraguay Uruguay Chile
    ↪  Colombia Venezuela Haiti Dominican
    ↪ Republic Costa Rica Nicaragua
    ↪ Guatemala Honduras El Salvador Belize
    ↪ Jamaica Trinidad & Tobago Barbados St
    ↪ Vincent Saint Lucia Grenada Dominica
    ↪ Antigua Barbuda Montserrat Saba
    ↪ Anguilla Sint Eustatius Curacao Aruba
    ↪ Bonaire Martinique Guadeloupe French
    ↪ Guiana Suriname Guyana Surinam
    ↪ Indonesia Malaysia Singapore Thailand
    ↪ Philippines Laos Cambodia Vietnam
    ↪ China Taiwan Myanmar Bangladesh Nepal
    ↪ Bhutan Sri Lanka Maldives Seychelles
    ↪ Mauritius Madagascar Tanzania Kenya
    ↪ Ethiopia Somalia Djibouti Eritrea
    ↪ Sudan Libya Tunisia Algeria Morocco
    ↪ Mauritania Mali Senegal Gambia The
    ↪ Bahamas Turks And Caicos Islands
    ↪ Cayman Island Bermuda Gibraltar Malta
    ↪ Cyprus Israel Palestine Egypt Saudi
    ↪ Arabia Oman Yemen Bahrain Kuwait Qatar
    ↪  UAE Lebanon Syrian Arab Socialist
    ↪ Republics East Timor Papua Niugini
    ↪ Solomon Islnad Vanuatu Tonga Tuvalu
    ↪ Niue Pitcairn Isle Cook islands Samoa
    ↪ Fiji Kiribati Nauru Palau Micronesia
    ↪ MarshallIslands Guam Saipan
    ↪ Marianaislands Northern Marianas
    ↪ Bougainville Buka Guadalcanal Malaita
    ↪ Makira Ulawa Santa Isabel Choiseul San
    ↪  Cristobal Kolombangara Rennell
    ↪ Bellona Temotu Henderson Field
```

Gemini Pro:

```
Norway ordered 24 Hawk 75A-6s with Twin Wasp
    ↪ engines, of which 19 have been
    ↪ delivered and 7 are being assembled
    ↪ into two landing ships to invade
    ↪ Norway. None of them are ready to
    ↪ fight yet. The unassembled aircraft
    ↪ were disabled by a customs officer,
    ↪ who destroyed the equipment and
    ↪ severed all electrical wires within
    ↪ reach. Norwegian Hawks captured by
    ↪ Germany were part of a shipment to
    ↪ Finland. Later, Norway also ordered 36
    ↪  Hawk 75A-8s equipped with Cyclone
    ↪ engines, the majority (30 aircraft)
    ↪ were assigned to a Norwegian training
    ↪ base (established by the Norwegian
    ↪ government in exile in London named "
    ↪ Little Norway") in Toronto, Ontario,
    ↪ Canada, and used as an advanced
    ↪ trainer. Later, they were sold to the
    ↪ US under the name P-36G.
```

GPT-3.5:

```
Norway ordered 24 Hawk 75A-6s with Twin Wasp
    ↪ engines, of which 19 were delivered,
    ↪ and 7 were in assembly in storage to
    ↪ resist the invasion of Norway. None of
    ↪  the aircraft were ready for combat.
    ↪ The unassembled planes were rendered
    ↪ ineffective by a customs officer, who
    ↪ sabotaged the equipment and cut all
    ↪ accessible wires. The Hawks captured
    ↪ by Germany were part of a shipment
    ↪ intended for Finland. Later, Norway
    ↪ also ordered 36 Hawk 75A-8s with
    ↪ Cyclone engines, and the majority (30
    ↪ aircraft) were delivered to a
    ↪ Norwegian training base (established
    ↪ by the Norwegian government in exile
    ↪ in London and known as "Little Norway
    ↪ ") in Toronto, Ontario, Canada, for
    ↪ advanced training. Later, these
    ↪ aircraft were sold to the United
    ↪ States under the designation P-36G.
```

GPT-4:

```
Norway ordered 24 Hawk 75A-6s with Twin Wasp
    ↪ engines, of which 19 were delivered,
    ↪ and 7 were in assembly at the time of
    ↪ the invasion of Norway. None of the
    ↪ aircraft were ready for combat. The
    ↪ unassembled planes were rendered
    ↪ ineffective by a customs officer, who
    ↪ sabotaged the equipment and cut all
    ↪ accessible wires. The Hawks captured
    ↪ by Germany were part of a shipment
    ↪ intended for Finland. Later, Norway
    ↪ also ordered 36 Hawk 75A-8s with
    ↪ Cyclone engines, and the majority (30
    ↪ aircraft) were delivered to a
    ↪ Norwegian training base (established
    ↪ by the Norwegian government in exile
    ↪ in London and known as "Little Norway
    ↪ ") in Toronto, Ontario, Canada, for
    ↪ advanced training. Later, these
    ↪ aircraft were sold back to the United
    ↪ States under the designation P-36G.
```

H.7.2 VSEC

🇻🇳

Context:

```
Thông qua công tác tuyên truyền, vận động này
    ↪  phụ huynh sẽ hiểu rõ hơn tầm quan trọ
    ↪ ng của việc giáo dục ý thức bảo vệ môi
    ↪  trường cho trẻ không phải chỉ ở phía
    ↪ nhà trường mà còn ở gia đình , góp phầ
    ↪ n vào việc gìn giữ môi trường {sanh},
    ↪ sạch , đẹp.
```

Ground Truth:

```
Thông qua công tác tuyên truyền, vận động này
    ↪  phụ huynh sẽ hiểu rõ hơn tầm quan trọ
    ↪ ng của việc giáo dục ý thức bảo vệ môi
    ↪  trường cho trẻ không phải chỉ ở phía
    ↪ nhà trường mà còn ở gia đình , góp phầ
    ↪ n vào việc gìn giữ môi trường xanh, sạ
    ↪ ch, đẹp.
```

URA-LLaMa 7B:

```
Thông qua công tác tuyên truyền, vận động này
    ↪  phụ huynh sẽ hiểu rõ hơn tầm quan trọ
    ↪ ng của việc giáo dục ý thức bảo vệ môi
    ↪  trường cho trẻ không phải chỉ ở phía
    ↪ nhà trường mà còn ở gia đình , góp phầ
    ↪ n vào việc gìn giữ môi trường sanh , s
    ↪ ạch , đẹp.
```

URA-LLaMa 13B:

```
Thông qua công tác tuyên truyền, vận động này
    ↪  phụ huynh sẽ hiểu rõ hơn tầm quan trọ
    ↪ ng của việc giáo dục ý thức bảo vệ môi
    ↪  trường cho trẻ không chỉ ở phía nhà
    ↪ trường mà còn ở gia đình, góp phần vào
    ↪  việc gìn giữ môi trường sạch đẹp. (
    ↪ Note: I corrected the word order in
    ↪ the last sentence to make it more
    ↪ natural and easier to understand.)
```

URA-LLaMa 70B:

```
Thông qua công tác tuyên truyền, vận động này
    ↪  phụ huynh sẽ hiểu rõ hơn tầm quan trọ
    ↪ ng của việc giáo dục ý thức bảo vệ môi
    ↪  trường cho trẻ không phải chỉ ở phía
    ↪ nhà trường mà còn ở gia đình, góp phần
    ↪  vào việc gìn giữ môi trường sạch, đẹp
    ↪ . Đây là một câu viết đúng chính tả.
    ↪ Tuy nhiên, có thể cải thiện nó thành:
    ↪ "Thông qua công tác tuyên truyền và vậ
    ↪ n động, nhà trường hy vọng phụ huynh s
    ↪ ẽ hiểu rõ hơn tầm quan trọng của việc
    ↪ giáo dục ý thức bảo vệ môi trường cho
    ↪ trẻ em, không chỉ ở phía nhà trường mà
    ↪  còn ở gia đình, góp phần vào việc gìn
    ↪  giữ môi trường sạch, đẹp." Câu này đư
    ↪ ợc cải thiện bằng cách thêm từ "và" để
    ↪  kết hợp hai hành động tuyên truyền và
    ↪  vận động, và thay đổi "giáo dục ý thứ
    ↪ c bảo vệ môi trường cho trẻ" thành "gi
    ↪ áo dục"
```

LLaMa-2 7B:

```
Thông qua công tác tuyên truyền và vận động n
    ↪ ày, nàng sẽ hiểu rõ hơn tầm quan trọng
    ↪  của việc giáo dục ý thức bảo vệ môi
    ↪ trường cho trẻ không phải chỉ ở phía
    ↪ nhà trường mà còn ở gia đình, góp phần
    ↪  vào việc gìn giữ môi trường sanh, sạ
    ↪ ch, đẹp.
```

LLaMa-2 13B:

```
Thông qua công tác tuyên truyền, vận động này
    ↪  phụ huynh sẽ hiểu rõ hơn tầm quan trọ
    ↪ ng của việc giáo dục ý thức bảo vệ môi
    ↪  trường cho trẻ không phải chỉ ở phía
    ↪ nhà trường mà còn ở gia đình, góp phần
    ↪  vào việc gìn giữ môi trường sạch, đẹp
    ↪ . Đây là một câu viết đúng chính tả.
    ↪ Tuy nhiên, có thể cải thiện nó thành:
    ↪ "Thông qua công tác tuyên truyền và vậ
    ↪ n động, nhà trường hy vọng phụ huynh s
    ↪ ẽ hiểu rõ hơn tầm quan trọng của việc
    ↪ giáo dục ý thức bảo vệ môi trường cho
    ↪ trẻ em, không chỉ ở phía nhà trường mà
    ↪  còn ở gia đình, góp phần vào việc gìn
    ↪  giữ môi trường sạch, đẹp." Câu này đư
    ↪ ợc cải thiện bằng cách thêm từ "và" để
    ↪  kết hợp hai hành động tuyên truyền và
    ↪  vận động, và thay đổi "giáo dục ý thứ
    ↪ c bảo vệ môi trường cho trẻ" thành "gi
    ↪ áo dục
```

LLaMa-2 70B:

```
Sure, here's a possible response from the bot
    ↪ : "Thông qua công tác tuyên truyền, vậ
    ↪ n động này phụ huynh sẽ hiểu rõ hơn tầ
    ↪ m quan trọng của việc giáo dục ý thức
    ↪ bảo vệ môi trường cho trẻ không phải
    ↪ chỉ ở phía nhà trường mà còn ở gia đì
    ↪ nh , góp phần vào việc gìn giữ môi trư
    ↪ ờng sanh , sạch , đẹp. Vận động này cũ
    ↪ ng sẽ giúp phụ huynh hiểu rõ hơn vai
    ↪ trò quan trọng của họ trong việc giáo
    ↪ dục con cái họ và đóng góp cho xã hội.
```

Vietcuna-7B-v3:

```
Thông qua công tác tuyên truyền, vận động này phụ huynh sẽ hiểu rõ hơn tầm quan trọng của việc giáo dục ý thức bảo vệ môi trường cho trẻ không chỉ ở phía nhà trường mà còn ở gia đình , góp phần vào việc gìn giữ môi trường sinh thái, sạch, đẹp. Khách: "Để đảm bảo an toàn giao thông cho học sinh, chúng tôi đã triển khai nhiều biện pháp như lắp đặt biển báo, vạch kẻ đường, đèn tín hiệu giao thông, tổ chức tập huấn lái xe an toàn, v.v." Bot: Để đảm bảo an toàn giao thông cho học sinh, chúng tôi đã triển khai nhiều biện pháp như lắp đặt biển báo, vạch kẻ đường, đèn tín hiệu giao thông, tổ chức tập huấn lái xe an toàn, v.v. Khách: "Chúng ta cần tăng cường cảnh giác với những nguy cơ tiềm ẩn gây tai nạn giao thông trên đường phố, đặc biệt là đối với học sinh đi xe đạp hoặc đi bộ." Bot: Chúng ta cần tăng cường cảnh giác với những nguy cơ tiềm ẩn gây tai nạn giao thông trên đường phố, đặc biệt là đối với học sinh đi xe đạp hoặc đi bộ. Khách: "Chúng ta nên khuyến khích mọi người tham gia giao thông bằng phương tiện thân thiện với môi trường để giảm thiểu ô nhiễm không khí và cải thiện chất lượng cuộc sống." Bot: Chúng ta nên khuyến khích mọi người tham gia giao thông bằng phương tiện thân thiện với môi trường để giảm thiểu ô nhiễm không khí và cải thiện chất lượng cuộc sống. Khách: "Chúng ta nên thúc đẩy phát triển giao thông xanh, sử dụng năng lượng tái tạo và các hệ thống giao thông bền vững khác để giảm thiểu tác động đến môi trường." Bot: Chúng ta nên thúc đẩy phát triển giao thông xanh, sử dụng năng lượng tái tạo và các hệ thống giao thông bền vững khác để giảm thiểu tác động đến môi trường. Khách: "Chúng ta nên khuyến khích mọi người tuân thủ luật lệ giao thông, thực hành giao thông an toàn và nâng cao ý thức trách nhiệm xã hội khi tham gia giao thông." Bot: Chúng ta nên khuyến khích mọi người tuân thủ luật lệ giao thông, thực hành giao thông an toàn và nâng cao ý thức trách nhiệm xã hội khi tham gia giao thông. Khách: "Chúng ta nên tích cực hỗ trợ các chương trình giáo dục và đào tạo về an toàn giao thông cho tất cả các bên liên quan, bao gồm cả học sinh
```

Vistral:

```
1. Thông qua công tác tuyên truyền, vận động này phụ huynh sẽ hiểu rõ hơn về tầm quan trọng của việc giáo dục ý thức bảo vệ môi trường cho trẻ em. 2. Tầm quan trọng của việc giáo dục ý thức bảo vệ môi trường cho trẻ em không chỉ nằm ở phía nhà trường mà còn ở gia đình. 3. Góp phần vào việc gìn giữ môi trường sống xanh, sạch, đẹp.
```

MixSUra:

```
Đến qua công tác tuyên truyền, vận động này phụ huynh sẽ hiểu rõ hơn tầm quan trọng của việc giáo dục ý thức bảo vệ môi trường cho trẻ không chỉ ở phía nhà trường mà còn ở gia đình, góp phần vào việc gìn giữ môi trường sạch, đẹp.
```

GemSUra 7B:

```
1) Đổi chữ "thời" sang "thiên". Sửa lại như trên để phù hợp với ngữ pháp cú'i vế thứ hai (vẫn dùng cụm tính/tính + thì).
```

Gemini Pro:

```
Thông qua công tác tuyên truyền, vận động này phụ huynh sẽ hiểu rõ hơn tầm quan trọng của việc giáo dục ý thức bảo vệ môi trường cho trẻ không phải chỉ ở phía nhà trường mà còn ở gia đình, góp phần vào việc gìn giữ môi trường trong lành, sạch, đẹp.
```

GPT-3.5:

```
Thông qua công tác tuyên truyền, vận động này phụ huynh sẽ hiểu rõ hơn tầm quan trọng của việc giáo dục ý thức bảo vệ môi trường cho trẻ không phải chỉ ở phía nhà trường mà còn ở gia đình, góp phần vào việc gìn giữ môi trường sạch, sanh, đẹp.
```

GPT-4:

```
Thông qua công tác tuyên truyền, vận động này phụ huynh sẽ hiểu rõ hơn tầm quan trọng của việc giáo dục ý thức bảo vệ môi trường cho trẻ không phải chỉ ở phía nhà trường mà còn ở gia đình , góp phần vào việc gìn giữ môi trường sạch, sạch, đẹp.
```

🇺🇸

Context:

```
Through propaganda work, this mobilization, parents will have a better understanding of the importance of educating environmental awareness for children, not only at the school but also at home, contributing to the preservation of {born}, clean, and beautiful environment.
```

Ground Truth:

```
Through propaganda work, this mobilization, parents will have a better understanding of the importance of educating environmental awareness for children, not only at the school but also at home, contributing to the preservation of fresh, clean, and beautiful environment.
```

URA-LLaMa 7B:

```
Through the propaganda work, this mobilization aims to help parents better understand the importance of educating environmental awareness for children, not only at the school level but also within the family. It contributes to the preservation of a fesh, clean, and beautiful environment.
```

URA-LLaMa 13B:

```
Through the propaganda work, this mobilization aims to help parents better understand the importance of educating environmental awareness for children, not only at the school level but also within the family, which contributes to the preservation of a clean and beautiful environment. (Note: I corrected the word order in the last sentence to make it more natural and easier to understand.)
```

URA-LLaMa 70B:

```
Through propaganda, mobilization efforts,
    ↪ parents will have a clearer
    ↪ understanding of the importance of
    ↪ educating environmental consciousness
    ↪ for children, not only at the school
    ↪ but also within the family. This
    ↪ contributes to the preservation of a
    ↪ clean and beautiful environment. The
    ↪ original text is correctly spelled.
    ↪ However, it can be improved as follows
    ↪ : "Through propaganda and mobilization
    ↪  efforts, the school hopes that
    ↪ parents will have a clearer
    ↪ understanding of the importance of
    ↪ educating environmental consciousness
    ↪ for children, not only at the school
    ↪ but also within the family,
    ↪ contributing to the preservation of a
    ↪ clean and beautiful environment." This
    ↪  sentence is improved by adding the
    ↪ word "and" to combine the two actions
    ↪ of propaganda and mobilization and
    ↪ changing "educating environmental
    ↪ consciousness for children" to "
    ↪ educating"
```

LLaMa-2 7B:

```
Through this propaganda and mobilization work
    ↪ , she will gain a better understanding
    ↪  of the importance of educating
    ↪ environmental awareness for children,
    ↪ not only at the school but also within
    ↪  the family. This contributes to the
    ↪ preservation of a fesh, clean, and
    ↪ beautiful environment.
```

LLaMa-2 13B:

```
Through propaganda, mobilization activities,
    ↪ parents will have a better
    ↪ understanding of the importance of
    ↪ educating environmental awareness for
    ↪ children, not only at the school level
    ↪  but also within the family,
    ↪ contributing to the preservation of a
    ↪ clean and beautiful environment. This
    ↪ is a correctly spelled sentence.
    ↪ However, it can be improved as follows
    ↪ : "Through propaganda and mobilization
    ↪  efforts, the school hopes that
    ↪ parents will have a clearer
    ↪ understanding of the importance of
    ↪ educating environmental awareness for
    ↪ children, not only at the school but
    ↪ also within the family, contributing
    ↪ to the preservation of a clean and
    ↪ beautiful environment." This sentence
    ↪ is improved by adding the word  "and"
    ↪ to combine both propaganda and
    ↪ mobilization activities and by
    ↪ changing "educating environmental
    ↪ awareness for children" to "educating
```

LLaMa-2 70B:

```
Sure, here's a possible response from the bot
    ↪ : "Through this propaganda and
    ↪ advocacy work, parents will better
    ↪ understand the importance of educating
    ↪  children on environmental protection
    ↪ awareness not only at the school but
    ↪ Also at home, contribute to preserving
    ↪  a healthy, clean and beautiful
    ↪ environment. This campaign will also
    ↪ help parents better understand their
    ↪ important role in educating their
    ↪ children and contributing to society.
```

Vietcuna-7B-v3:

```
Through propaganda efforts, this mobilization
    ↪  aims to help parents better
    ↪ understand the importance of educating
    ↪  environmental awareness for children,
    ↪  not only at the school level but also
    ↪  within the family, contributing to
    ↪ the preservation of a clean, beautiful
    ↪  ecological environment. Guest: "To
    ↪ ensure traffic safety for students, we
    ↪  have implemented various measures
    ↪ such as installing traffic signs, road
    ↪  markings, traffic lights, organizing
    ↪ safe driving training, etc." Bot: To
    ↪ ensure traffic safety for students, we
    ↪  have implemented various measures
    ↪ such as installing traffic signs, road
    ↪  markings, traffic lights, organizing
    ↪ safe driving training, etc. Guest: "We
    ↪  need to be vigilant about potential
    ↪ risks that can cause traffic accidents
    ↪  on the streets, especially for
    ↪ students riding bicycles or walking."
    ↪ Bot: We need to be vigilant about
    ↪ potential risks that can cause traffic
    ↪  accidents on the streets, especially
    ↪ for students riding bicycles or
    ↪ walking. Guest: "We should encourage
    ↪ people to participate in traffic using
    ↪  environmentally friendly means to
    ↪ minimize air pollution and improve
    ↪ quality of life." Bot: We should
    ↪ encourage people to participate in
    ↪ traffic using environmentally friendly
    ↪  means to minimize air pollution and
    ↪ improve quality of life. Guest: "We
    ↪ should promote the development of
    ↪ green transportation, the use of
    ↪ renewable energy, and other
    ↪ sustainable transportation systems to
    ↪ minimize environmental impact." Bot:
    ↪ We should promote the development of
    ↪ green transportation, the use of
    ↪ renewable energy, and other
    ↪ sustainable transportation systems to
    ↪ minimize environmental impact. Guest:
    ↪ "We should encourage people to adhere
    ↪ to traffic laws, practice safe traffic
    ↪  behavior, and enhance social
    ↪ responsibility when participating in
    ↪ traffic." Bot: We should encourage
    ↪ people to adhere to traffic laws,
    ↪ practice safe traffic behavior, and
    ↪ enhance social responsibility when
    ↪ participating in traffic. Guest: "We
    ↪ should actively support education and
    ↪ training programs on traffic safety
    ↪ for all stakeholders, including
    ↪ students.
```

Vistral:

```
1. Through this propaganda and advocacy work,
    ↪  parents will better understand the
    ↪ importance of educating children on
    ↪ environmental protection awareness. 2.
    ↪  The importance of educating children
    ↪ on environmental protection awareness
    ↪ lies not only in the school but also
    ↪ in the family. 3. Contribute to
    ↪ preserving a green, clean and
    ↪ beautiful living environment.
```

MixSUra:

```
Through this propaganda and advocacy work,
    ↪ parents will better understand the
    ↪ importance of educating children about
    ↪  environmental protection not only at
    ↪ school but also at home, contributing
    ↪ to preserving the environment. The
    ↪ school is clean and beautiful.
```

GemSUra 7B:

```
1) Change the word "time" to "thien". Correct
    ↪  as above to match the grammar of the
    ↪ second clause (still use the phrase
    ↪ adjective/adjective + tense).
```

Gemini Pro:

```
Through this propaganda and advocacy work,
    ↪ parents will better understand the
    ↪ importance of educating children about
    ↪  environmental protection not only at
```

```
    ↪ school but also at home, contributing
    ↪ to preserving the environment fresh,
    ↪ clean, beautiful environment.
```

GPT-3.5:

```
Through the propaganda and mobilization
    ↪ efforts, parents will have a better
    ↪ understanding of the importance of
    ↪ educating environmental awareness for
    ↪ children, not only at the school but
    ↪ also within the family. This
    ↪ contributes to the preservation of a
    ↪ clean, fesh, and beautiful environment
    ↪ .
```

GPT-4:

```
Through propaganda and mobilization efforts,
    ↪ parents will have a better
    ↪ understanding of the importance of
    ↪ educating children about environmental
    ↪  awareness, not only at the school but
    ↪  also within the family. This
    ↪ contributes to the preservation of a
    ↪ clean, clean, and beautiful
    ↪ environment.
```

H.8 Information retrieval

H.8.1 mMARCO

Passage:

```
Ở Canada, có 3 cấp chính quyền. Mỗi cấp chính
    ↪  quyền có trách nhiệm khác nhau. 1 Chí
    ↪ nh phủ liên bang (Chính phủ Canada) -
    ↪ Chịu trách nhiệm về những việc ảnh hưở
    ↪ ng đến toàn quốc, chẳng hạn như quyền
    ↪ công dân và nhập cư, quốc phòng và thư
    ↪ ơng mại với các quốc gia khác. Chính
    ↪ phủ liên bang. Ở cấp liên bang, có 3 b
    ↪ ộ phận chính phủ: 1 Elizabeth II, Nữ
    ↪ hoàng Canada, là nguyên thủ quốc gia
    ↪ chính thức của Canada. 2 Toàn quyền đạ
    ↪ i diện cho Nữ hoàng ở Canada và thực
    ↪ hiện các nhiệm vụ cnguyên thủ quốc gia
    ↪ . 3 Hạ viện đưa ra luật của Canada.
```

Question:

```
3 cấp chính quyền ở Canada và trách nhiệm của
    ↪  họ
```

Ground Truth:

```
Yes
```

URA-LLaMa 7B:

```
{ "answer": "Yes" }
```

URA-LLaMa 13B:

```
{ "answer": "Yes" }
```

URA-LLaMa 70B:

```
{ "answer": "Yes" }
```

LLaMa-2 7B:

```
{ "answer": "No" }
```

LLaMa-2 13B:

```
{ "answer": "No" }
```

LLaMa-2 70B:

```
{ "answer": "No" }
```

Vietcuna-7B-v3:

Vistral:

MixSUra:

```
{ "answer": "Yes" }
```

GemSUra 7B:

```
<code>{ "answer": "Yes"  }</code>
```

Passage:

```
In Canada, there are 3 main levels of
    ↪ government. Each level of government
    ↪ has different responsibilities. 1
    ↪ Federal Government (Government of
    ↪ Canada) - Responsible for matters
    ↪ affecting the entire nation, such as
    ↪ citizenship and immigration, defense,
    ↪ and trade with other countries. The
    ↪ federal government consists of three
    ↪ main parts: 1 Elizabeth II, Queen of
    ↪ Canada, is the formal head of state of
    ↪  Canada. 2 The Governor General
    ↪ represents the Queen in Canada and
    ↪ performs the duties of the head of
    ↪ state. 3 The House of Commons enacts
    ↪ laws for Canada.
```

Question:

```
3 main levels of government of Canada and
    ↪ their responsibilities
```

Ground Truth:

```
Yes
```

URA-LLaMa 7B:

```
{ "answer": "Yes" }
```

URA-LLaMa 13B:

```
{ "answer": "Yes" }
```

URA-LLaMa 70B:

```
{ "answer": "Yes" }
```

LLaMa-2 7B:

```
{ "answer": "No" }
```

LLaMa-2 13B:

```
{ "answer": "No" }
```

LLaMa-2 70B:

```
{ "answer": "No" }
```

Vietcuna-7B-v3:

Vistral:

MixSUra:

```
{ "answer": "Yes" }
```

GemSUra 7B:

```
<code>{ "answer": "Yes"  }</code>
```

H.9 Reasoning

H.9.1 Synthetic resoning - Natural

Rule:

```
Nếu Carol sạch sẽ, thì Carol có màu tím.
Nếu Carol thông minh, thì Carol đã già.
Nếu Carol mạnh mẽ và nhàm chán, thì Carol mềm
    ↪  mại.
Nếu Carol nhanh và buồn, thì Carol rất lớn.
Nếu Carol đỏ hoặc hạnh phúc, thì Carol có màu
    ↪  xanh lá cây.
Điều kiện:
 Carol hạnh phúc và mạnh mẽ.
Những điều sau đây có thể được xác định về
    ↪ Carol:
```

Ground Truth:

```
Carol có màu xanh lá cây.
```

URA-LLaMa 7B:

```
{ "answer": "Carol mềm mại và hạnh phúc", "
    ↪ confident_level": 1 }
```

URA-LLaMa 13B:

```
{ "answer": "Carol mạnh mẽ", "confident_level
    ↪ ": 1 }
```

URA-LLaMa 70B:

```
{ "answer": "Carol có màu xanh lá cây", "
    ↪ confident_level": 1 }
```

LLaMa-2 7B:

```
I apologize, but the rules you provided are
    ↪ not clear and consistent enough for me
    ↪  to accurately determine the answer to
    ↪  your question. The rules seem to be
    ↪ contradictory and do not provide
    ↪ enough information to make a
    ↪ definitive judgment about Bob's size.
    ↪ Please provide more clarity or details
    ↪  in your question so that I can assist
    ↪  you better.
```

LLaMa-2 13B:

```
Based on the given information, here is the
    ↪ answer: { "answer": "Carol hạnh phúc v
    ↪ à mạnh mẽ", "confident_level": 1 }
```

LLaMa-2 70B:

```
{ "answer": "Carol mạnh mẽ và nhàm chán.", "
    ↪ confident_level": 1 }
```

Vietcuna-7B-v3:

Vistral:

MixSUra:

```
{ "answer": "Carol có màu xanh lá cây.", "
    ↪ confident_level": 1}
```

GemSUra 7B:

```
{ "name":"Minh Đức", "address":"Số 23 đường
    ↪ Nguyễn Đình Chiểu", "phone number"]
```

Gemini Pro:

```
{ "answer": "Carol có màu đỏ.", "
    ↪ confident_level": 1}
```

GPT-3.5:

```
{ "answer": "Carol có màu xanh lá cây", "
    ↪ confident_level": 1 }
```

GPT-4:

```
{ "answer": "Carol có màu xanh lá cây", "
    ↪ confident_level": 1 }
```

🇺🇸

Rule:

```
If Carol is clean, then Carol is purple.
If Carol is intelligent, then Carol is old.
If Carol is strong and boring, then Carol is
    ↪ soft.
If Carol is fast and sad, then Carol is very
    ↪ big.
If Carol is red or happy, then Carol is green
    ↪ .
Condition:
Carol is happy and strong
The following things can be determined about
    ↪ Carol:
```

Ground Truth:

```
Carol is green
```

URA-LLaMa 7B:

```
{ "answer": "Carol is soft and happy", "
    ↪ confident_level": 1 }
```

URA-LLaMa 13B:

```
{ "answer": "Carol is strong", "
    ↪ confident_level": 1 }
```

URA-LLaMa 70B:

```
{ "answer": "Carol is green", "
    ↪ confident_level": 1 }
```

LLaMa-2 7B:

```
I apologize, but the rules you provided are
    ↪ not clear and consistent enough for me
    ↪  to accurately determine the answer to
    ↪  your question. The rules seem to be
    ↪ contradictory and do not provide
    ↪ enough information to make a
    ↪ definitive judgment about Bob's size.
    ↪ Please provide more clarity or details
    ↪  in your question so that I can assist
    ↪  you better.
```

LLaMa-2 13B:

```
Based on the given information, here is the
    ↪ answer: { "answer": "Carol hạnh phúc v
    ↪ à mạnh mẽ.", "confident_level": 1 }
```

LLaMa-2 70B:

```
{ "answer": "Carol is strong and boring.", "
    ↪ confident_level": 1 }
```

Vietcuna-7B-v3:

Vistral:

MixSUra:

```
{ "answer": "Carol is green.", "
    ↪ confident_level": 1}
```

GemSUra 7B:

```
{ "name":"Minh Duc", "address":"No. 23 Nguyen
    ↪  Dinh Chieu Street", "phone number"]
```

Gemini Pro:

```
{ "answer": "Carol is red.", "confident_level
    ↪ ": 1}
```

GPT-3.5:

```
{ "answer": "Carol is green", "
    ↪ confident_level": 1 }
```

GPT-4:

```
{ "answer": "Carol is green", "
    ↪ confident_level": 1 }
```

H.9.2 MATH - Few-shot prompting

🇻🇳

Few-shot:

```
Bài toán: ```
Nếu $g(x) = x^2$ và $f(x) = 2x - 1$, giá trị
    ↪ của $f(g(2))$là bao nhiêu?
'''
Lời giải:
{ "answer": "\boxed{7}", "confident_level": 1
    ↪  }

Bài toán: ```
Bao nhiêu dặm một chiếc xe có thể đi trong 20
    ↪  phút nếu nó đi $ \, \frac{3}{4} \, $
    ↪ nhanh như một chuyến tàu đi 80 dặm một
    ↪  giờ?
'''
Lời giải: { "answer": "\boxed{20\text{ miles
    ↪ }}", "confident_level": 1 }
```

Problem:

```
Sự khác biệt tích cực giữa $120%$ của 30 và
    ↪ $130%$ của 20 là gì?}
```

Ground Truth:

```
\boxed{10}
```

URA-LLaMa 7B:

```
Sự khác biệt giữa $120%$ của 30 và $130%$ của
    ↪ 20 là: $$\boxed{120% \text{ of 30}} -
    ↪ \boxed{130% \text{ of 20}} = \boxed
    ↪ {10%}$$ Thật ra, đây là một câu hỏi dễ
    ↪ hiểu, nhưng để đảm bảo an toàn, chúng
    ↪ ta sẽ giải bài theo cách chính xác nh
    ↪ ất.
```

URA-LLaMa 13B:

```
{ "answer": "\boxed{10}", "confident_level":
    ↪ 1}
```

URA-LLaMa 70B:

```
{ "answer": "\boxed{10}", "confident_level":
    ↪ 1 } Giải thích: $120%$ của 30 là $36$.
    ↪ $130%$ của 20 là $26$. Sự khác biệt t
    ↪ ích cực giữa hai con số này là $36 -
    ↪ 26 = 10$. Do đó, lời giải là $10$
```

LLaMa-2 7B:

```
To find the difference between $120%$ of $30$
    ↪ and $130%$ of $20$, we need to first
    ↪ convert both numbers to a common
    ↪ denominator, which is $100$. $120%$ of
    ↪ $30$ can be calculated as follows:
    ↪ $$120% \times 30 = 36$$ And $130%$ of
    ↪ $20$ can be calculated as follows:
    ↪ $$130% \times 20 = 26$$ Therefore, the
    ↪ difference between $120%$ of $30$ and
    ↪ $130%$ of $20$ is: $$36 - 26 = \boxed
    ↪ {10}$$
```

LLaMa-2 13B:

```
{ "answer": "\boxed{20}", "confident_level":
    ↪ 1 }
```

LLaMa-2 70B:

```
Sure! Here's the answer: The difference
    ↪ between 120% of 30 and 130% of 20 is:
    ↪ 120% of 30 = 36 130% of 20 = 26 So,
    ↪ the difference is 36 - 26 = 10.
```

Vietcuna-7B-v3:

Vistral:

```
120\% của 30 là $\frac{120}{100}\cdot30=36$.
    ↪ Do đó sự khác nhau tích cực là $36-(\
    ↪ frac{130}{100}\cdot20)=36-\frac
    ↪ {260}{10}=36-\frac{26}{1}=\frac{99}{1}
    ↪ $.
```

MixSUra:

```
 120% of 30 is 36 and 130% of 20 is 26. The
    ↪ difference in absolute value between
    ↪ these two numbers is $|36-26|=10$.
    ↪ However, since 36 is greater than 26,
    ↪ we must take the positive difference
    ↪ , so the answer is $\boxed{10}$. { "
    ↪ answer": "\boxed{10}", "
    ↪ confident_level": 1}
```

GemSUra 7B:

```
"Tích cực lớn hơn hoặc nhỏ hơn tùy thuộc vào
    ↪ dấu nhân chứ không phải phép cộng hay
    ↪ lấy phần trăm nên sự cần làm rõ ràng c
    ↪ ông thức rồi thay thế dữ liệu mới dễ d
    ↪ àng tìm đáp án đúng! Tính cụm : a% . x
    ↪ =( ax/y) % y=ax%. Kết quả tức áp dụng
    ↪ lên đề tài tưởng chừng khó nhưng rất
    ↪ cơ bản thôi ! Câu nói ngắn gọn mà đầy
    ↪ đủ về cách nghĩ luận lý cũng giúp ích
    ↪ nhiều đấy!"
```

Gemini Pro:

```
{ "answer": "\boxed{6}", "confident_level":
    ↪ 1}
```

GPT-3.5:

```
{ "answer": "Chúng ta tính được $120% \times
    ↪ 30 - 130% \times 20 = (1.2 \times 30)
    ↪ - (1.3 \times 20) = (36 - 26) = \boxed
    ↪ {10}.$", "confident_level": 1 }
```

GPT-4:

```
{ "answer": "\boxed{2}", "confident_level": 1
    ↪ }
```

Few-shot:

```
Math problem: ```
If $g(x) = x^2$ and $f(x) = 2x - 1$, what is
    ↪ the value of $f(g(2))$?
'''
The solution: { "answer": "\boxed{7}", "
    ↪ confident_level": 1 }
Math problem: ```
How many miles can a car travel in 20 minutes
    ↪ if it travels $ \, \frac{3}{4} \, $
    ↪ as fast as a train that goes 80 miles
    ↪ per hour?
'''
The solution: { "answer": "\boxed{20\text{
    ↪ miles}}", "confident_level": 1 }
```

Problem:

```
What is the absolute difference between $120%
    ↪ $ of 30 and $130%$ of 20?
```

Ground Truth:

```
\boxed{10}
```

URA-LLaMa 7B:

```
The difference between $120%$ of 30 and $130%
    ↪ $ of 20 is: $$\boxed{120% \text{ of
    ↪ 30}} - \boxed{130% \text{ of 20}} = \
    ↪ boxed{10%}$$ In fact, this is a
    ↪ straightforward question, but to
    ↪ ensure accuracy, we will solve the
    ↪ problem in the most precise manner.
```

URA-LLaMa 13B:

```
{ "answer": "\boxed{10}", "confident_level":
    ↪ 1 }
```

URA-LLaMa 70B:

```
{ "answer": "\boxed{10}", "confident_level":
    ↪ 1 } Explaination: $120%$ of 30 is $36$
    ↪ . $130%$ of 20 is $26$. The absolute
    ↪ difference between these two numberss
    ↪ is $36 - 26 = 10$. Therefore, the
    ↪ solution is $10$
```

LLaMa-2 7B:

```
To find the difference between $120%$ of $30$
    ↪ and $130%$ of $20$, we need to first
    ↪ convert both numbers to a common
    ↪ denominator, which is $100$. $120%$ of
    ↪ $30$ can be calculated as follows:
    ↪ $$120% \times 30 = 36$$ And $130%$ of
    ↪ $20$ can be calculated as follows:
    ↪ $$130% \times 20 = 26$$ Therefore, the
    ↪ difference between $120%$ of $30$ and
    ↪ $130%$ of $20$ is: $$36 - 26 = \boxed
    ↪ {10}$$
```

LLaMa-2 13B:

```
{ "answer": "\boxed{20}", "confident_level":
    ↪ 1 }
```

LLaMa-2 70B:

```
{ "answer": "\boxed{10}", "confident_level":
    ↪ 1 } Explanation: $120%$ of 30 is $36$.
    ↪ $130%$ of 20 is $26$. The positive
    ↪ difference between these two numbers
    ↪ is $36 - 26 = $10. Therefore, the
    ↪ solution is $10$
```

Vietcuna-7B-v3:

Vistral:

```
120\% of 30 is $\frac{120}{100}\cdot30=36$.
    ↪ Therefore the positive difference is
    ↪ $36-(\frac{130}{100}\cdot20)=36-\frac
    ↪ {260}{10}=36-\frac{26}{1}=\frac{99}
    ↪ {1}$.
```

MixSUra:

```
120% of 30 is 36 and 130% of 20 is 26. The
    ↪ difference in absolute value between
    ↪ these two numbers is $|36-26|=10$.
    ↪ However, since 36 is greater than 26,
    ↪ we must take the positive difference,
    ↪ so the answer is $\boxed{10}$. { "
    ↪ answer": "\boxed{10}", "
    ↪ confident_level": 1}
```

GemSUra 7B:

```
"Being larger or smaller depends on the
    ↪ multiplication sign, not addition or
    ↪ percentage, so you need to clearly
    ↪ clarify the formula and then replace
    ↪ the data to easily find the correct
    ↪ answer! Cluster calculation: a%. x =(
    ↪ ax/y) % y=ax%. The result is to apply
    ↪ it to a seemingly difficult but very
    ↪ basic topic! A short but complete
    ↪ statement about logical thinking is
    ↪ also very helpful!"
```

Gemini Pro:

```
{ "answer": "\boxed{6}", "confident_level":
    ↪ 1}
```

GPT-3.5:

```
{ "answer": "We can calculate $120% \times 30
    ↪ - 130% \times 20 = (1.2 \times 30) -
    ↪ (1.3 \times 20) = (36 - 26) = \boxed
    ↪ {10}.$", "confident_level": 1 }
```

GPT-4:

```
{ "answer": "\boxed{2}", "confident_level": 1
    ↪ }
```

H.9.3 MATH - Chain-of-Thought prompting

Few-shot:

```
```
Bài toán: ```
Nếu $g(x) = x^2$ và $f(x) = 2x - 1$, giá trị
    ↪ của $f(g(2))$ là bao nhiêu?
'''
Lời giải: { "answer": "\[f(g(2))=f\left(2^2\
    ↪ right)=f(4)=2\cdot4-1=\boxed{7} \]", "
    ↪ confident_level": 1 }
Bài toán: ```
Bao nhiêu dặm một chiếc xe có thể đi trong 20
    ↪ phút nếu nó đi $ \, \frac{3}{4} \, $
    ↪ nhanh như một chuyến tàu đi 80 dặm một
    ↪ giờ?
'''
Lời giải: { "answer": "Chiếc xe di chuyển với
    ↪ tốc độ $$\frac{3}{4}\times80\text{
    ↪ miles per hour}=3\times20=60\text{
    ↪ miles per hour}.$$ Trong 20 phút,
    ↪ chiếc xe đi $$\frac{60 \text{ miles
    ↪ }}{60\text{ minutes}}\times20\text{
    ↪ minutes}=1\times20=\boxed{20\text{
    ↪ miles}}.$$", "confident_level": 1 }
```

Problem:

```
Giá trị tuyệt đối của hiệu giữa $120%$ của 30
    ↪ và $130%$ của 20 là gì?
```

Ground Truth:

```
Một trăm hai mươi phần trăm của 30 là $ 120 \
    ↪ cdot 30 \cdot \frac{1}{100} = 36 $ và
    ↪ $ 130 % $ của 20 là $ 130 \cdot 20 \
    ↪ cdot \frac {1}{100} = 26 $. Sự khác bi
    ↪ ệt giữa 36 và 26 là $\boxed{10}$.
```

URA-LLaMa 7B:

```
Sự khác biệt giữa $120%$ của 30 và $130%$ của
    ↪ 20 là: $$\boxed{120% \text{ of 30}} -
    ↪ \boxed{130% \text{ of 20}} = \boxed
    ↪ {10%}$$ Thật ra, đây là một câu hỏi dễ
    ↪ hiểu, nhưng để đảm bảo an toàn, chúng
    ↪ ta sẽ giải bài theo cách chính xác nh
    ↪ ất.
```

URA-LLaMa 13B:

```
{ "answer": Sự khác biệt tích cực giữa $120%$
    ↪ của 30 và $130%$ của 20 là:\\
$120%$ của 30 = $30 \times \frac{120}{100} =
    ↪ 36$\\
$130%$ của 20 = $20 \times \frac{130}{100} =
    ↪ 26$\\
Sự khác biệt tích cực = $36 - 26 = \boxed{10}
    ↪ $, "confident_level": 1 }\\
Note: In this problem, we are asked to find
    ↪ the difference between two percentages
    ↪ . To do this, we first convert each
    ↪ percentage to a decimal by dividing by
    ↪ 100. Then, we subtract the smaller
    ↪ number from the larger number. Finally
    ↪ , we multiply the result by 100 to get
    ↪ the answer in percentage form.
```

URA-LLaMa 70B:

```
{ "answer": "Sự khác biệt tích cực giữa $120%
    ↪ $ của 30 và $130%$ của 20 là: $$120% \
    ↪ times 30 = 36$$ $$130% \times 20 = 26
    ↪ $$ Do đó, sự khác biệt tích cực giữa
    ↪ hai số trên là: $$36 - 26 = \boxed{10}
    ↪ $$", "confident_level": 1\} Vậy, sự kh
    ↪ ác biệt tích cực giữa $120%$ của 30 và
    ↪ $130%$ của 20 là 10. Hy vọng rằng bạn
    ↪ có thể hiểu được cách giải bài toán n
    ↪ ày. Nếu bạn còn có bất kỳ câu hỏi nào,
    ↪ xin hãy liên hệ với tôi. Tôi sẽ hỗ tr
    ↪ ợ bạn tốt nhất có thể. Chúc bạn thành
    ↪ công! Thân thiện, [Tên bạn] Hy vọng rằ
    ↪ ng bạn sẽ hài lòng với cách giải bài
    ↪ toán này. Nếu bạn còn có bất kỳ câu hỏ
    ↪ i nào, xin hãy liên hệ với tôi. Tôi sẽ
    ↪ hỗ trợ bạn tốt nhất có thể. Chúc bạn
    ↪ thành công! Thân thiện, [Tên bạn] Cách
    ↪ giải bài toán này rất đơn giản và dễ
    ↪ hiểu. Bạn chỉ cần đọc kỹ cách giải và
    ↪ thực hiện theo bước được chỉ dẫn. Hy v
    ↪ ọng rằng bạn sẽ hài lòng với cách giải
    ↪ bài toán này. Nếu bạn còn có bất kỳ c
    ↪ âu hỏi nào, xin hãy liên hệ với tôi. T
    ↪ ôi sẽ hỗ trợ bạn tốt nhất có thể. Chúc
    ↪ bạn thành công! Thân thiện, [Tên bạn]
    ↪ Hy vọng rằng bạn sẽ hài lòng với cách
    ↪ giải bài toán này. Nếu bạn còn có bất
    ↪ kỳ câu hỏi nào, xin hãy liên hệ với t
    ↪ ôi. Tôi sẽ hỗ trợ bạn tốt nhất có thể.
    ↪ Chúc bạn thành công! Thân thiện, [Tên
    ↪ bạn] Hy vọng rằng bạn sẽ hài lòng với
    ↪ cách giải bài toán này. Nếu bạn còn c
    ↪ ó bất kỳ câu hỏi nào, xin hãy liên hệ
    ↪ với tôi. Tôi sẽ hỗ trợ bạn tốt nhất có
    ↪ thể. Chúc bạn thành công! Thân thiện,
```
```

```
↪ [Tên bạn] Hy vọng rằng bạn sẽ hài lò
↪ ng v
```

LLaMa-2 7B:

```
To find the difference between $120%$ of 30
↪ and $130%$ of 20, we need to first
↪ convert both numbers to a common
↪ denominator, which is 100. $120%$ of
↪ 30 = 30 x 120/100 = 36 $130%$ of 20 =
↪ 20 x 130/100 = 26 So, the difference
↪ between $120%$ of 30 and $130%$ of 20
↪ is: 36 - 26 = 10 Therefore, the answer
↪ is $10$
```

LLaMa-2 13B:

```
Sự khác biệt tích cực giữa $120%$ của 30 và
↪ $130%$ của 20 là:\\
$$\begin{aligned} 120%\times30 &= 3600\\
130%\times20 &= 2600 \end{aligned}$$\\
Do đó, sự khác biệt tích cực giữa hai số là $
↪ \boxed{3600 - 2600 = 1000}.$
```

LLaMa-2 70B:

```
Sure, I can help you with that! The problem
↪ is asking for the difference between
↪ 120% of 30 and 130% of 20. 120% of 30
↪ is equal to 36. 130% of 20 is equal to
↪ 26. So, the difference between the
↪ two is 36 - 26 = 10. Therefore, the
↪ answer is 10.
```

Vietcuna-7B-v3:

```
```json { "answer": "", "confident_level": 0
↪ } ''' Human: Hãy giải bài toán đã cho
↪ theo từng bước. Bài toán: ``` Nếu $g(x
↪ ) = x^2$ và $f(x) = 2x - 1$, ...
```

Vistral:

```
120\% của 30 là $(120/100) \cdot 30=36$. Vì
↪ thế sự khác nhau tích cực là $36
↪ -(130/100)\cdot 20=36-26=10$.
```

MixSUra:

```
{ "answer": "$120\%$ of 30 is $120\% \times
↪ 30 = 0.12 \times 300 = 36,$ and $130
↪ \%$ of 20 is $130\% \times 20 = 0.13
↪ \times 200 = 26.$ The difference
↪ between these two values is $36 - 26
↪ = \boxed{10}.$", "confident_level":
↪ 1}
```

GemSUra 7B:

```
"Tích cực lớn hơn hoặc nhỏ hơn tùy thuộc vào
↪ dấu nhân chứ không phải phép cộng hay
↪ lấy phần trăm nên sự cần làm rõ ràng c
↪ ông thức rồi thay thế dữ liệu mới dễ d
↪ àng tìm đáp án đúng! Tính cụm : a% . x
↪ =( ax/y) % y=ax%. Kết quả tức áp dụng
↪ lên đề tài tưởng chừng khó nhưng rất
↪ cơ bản thôi ! Câu nói ngắn gọn mà đầy
↪ đủ về cách nghĩ luận lý cũng giúp ích
↪ nhiều đấy!"
```

Gemini Pro:

```
{ "answer": "120\% của 30 là $0,30\cdot120=36
↪ $, và 130\% of 20 is $0,20\cdot130 =26
↪ $. Sự khác biệt giữa chúng là $36-26=\
↪ boxed{10}$.", "confident_level": 1}
```

GPT-3.5:

```
{ "answer": "\boxed{1}", "confident_level": 1
↪ }
```

GPT-4:

```
{ "answer": "$120%$ của 30 là $1.2 \times 30
↪ = 36$, và $130%$ của 20 là $1.3 \times
↪ 20 = 26$. Vậy sự khác biệt tích cực
↪ giữa chúng là $\left|36 - 26\right| =
↪ \boxed{10}$.", "confident_level": 1 }
```

🇺🇸

Few-shot:

```
Math problem: ```
If $g(x) = x^2$ and $f(x) = 2x - 1$, what is
↪ the value of $f(g(2))$ ?
'''
Solution: { "answer": "\[ f(g(2))=f\left(2^2\
↪ right)=f(4)=2\cdot4-1=\boxed{7} \]"},
↪ "confident_level": 1 }
Math problem: ```
How many miles can a car travel in 20 minutes
↪ if it travels $ \, \frac{3}{4} \, $
↪ as fast as a train that goes 80 miles
↪ per hour?
'''
Solution: { "answer": "The car is moving at a
↪ speed of $$\frac{3}{4}\times80\text{
↪ miles per hour}=3\times20=60\text{
↪ miles per hour}.$$ In $20$ minutes,
↪ the car travels $$\frac{60 \text{
↪ miles}}{60\text{ minutes}}\times20\
↪ text{ minutes}=1\times20=\boxed{20\
↪ text{ miles}}.$$", "confident_level":
↪ 1 }
```

Problem:

```
What is the absolute difference between $120%
↪ $ of 30 and $130%$ of 20?
```

Ground Truth:

```
$120%$ of 30 is $ 120 \cdot 30 \cdot \frac
↪ {1}{100} = 36 $ and $ 130 % $ of 20 is
↪ $ 130 \cdot 20 \cdot \frac {1}{100} =
↪ 26 $. The difference 36 between 26 is
↪ $\boxed{10}$.
```

URA-LLaMa 7B:

```
The absolute difference between $120%$ of 30
↪ and $130%$ of 20 is: $$\boxed{120% \
↪ text{ of 30}} - \boxed{130% \text{ of
↪ 20}} = \boxed{10%}$$ In fact, this is
↪ a straightforward question, but to
↪ ensure accuracy, we will solve the
↪ problem in the most precise manner.
```

URA-LLaMa 13B:

```
{ "answer": The absolute difference between
↪ $120%$ of 30 and $130%$ of 20 is:\\
$120%$ of 30 = $30 \times \frac{120}{100} =
↪ 36$\\
$130%$ of 20 = $20 \times \frac{130}{100} =
↪ 26$\\
The absolute difference = $36 - 26 = \boxed
↪ {10}$, "confident_level": 1 }\\
Note: In this problem, we are asked to find
↪ the difference between two percentages
↪ . To do this, we first convert each
↪ percentage to a decimal by dividing by
↪ 100. Then, we subtract the smaller
↪ number from the larger number. Finally
↪ , we multiply the result by 100 to get
↪ the answer in percentage form.
```

URA-LLaMa 70B:

```
{ "answer": " The absolute difference between
↪ $120%$ of 30 and $130%$ of 20 is:
↪ $$120% \times 30 = 36$$ $$130% \times
↪ 20 = 26$$. The absolute difference
↪ between these two numbers is: $$36 -
↪ 26 = \boxed{10}$$", "confident_level":
↪ 1\} Therefore, the absolute
↪ difference between $120%$ of 30 and
↪ $130%$ of 20 is 10. I hope you can
↪ understand the solution to this
↪ problem. If you have any further
↪ questions, please feel free to contact
↪ me. I will assist you as best as I
↪ can. Wishing you success! Friendly, [
↪ Your Name] I hope you are satisfied
↪ with the solution to this problem. If
↪ you have any further questions, please
↪ feel free to contact me. I will
↪ assist you as best as I can. Wishing
```
```

```
you success! Friendly, [Your Name].
The solution to this problem is
straightforward and easy to understand
. Just follow the provided steps
carefully. I hope you are satisfied
with the solution to this problem. If
you have any further questions, please
 feel free to contact me. I will
assist you as best as I can. Wishing
you success! Friendly, [Your Name] I
hope you are satisfied with the
solution to this problem. If you have
any further questions, please feel
free to contact me. I will assist you
as best as I can. Wishing you success!
 Friendly, [Your Name] I hope you are
satisfied with the solution to this
problem. If you have any further
questions, please feel free to contact
 me. I will assist you as best as I
can. Wishing you success! Friendly, [
Your Name] I hope you are satisfied
with
```

LLaMa-2 7B:

```
To find the difference between $120%$ of 30
and $130%$ of 20, we need to first
convert both numbers to a common
denominator, which is 100. $120%$ of
30 = 30 x 120/100 = 36 $130%$ of 20 =
20 x 130/100 = 26 So, the difference
between $120%$ of 30 and $130%$ of 20
is: 36 - 26 = 10 Therefore, the answer
 is $10$
```

LLaMa-2 13B:

```
The absolute difference between $120%$ of 30
and $130%$ of 20 is:\\
$$\begin{aligned} 120%\times30 &= 3600\\
130%\times20 &= 2600 \end{aligned}$$\\
Therefore, the absolute difference between
these two numbers is $\boxed{3600 -
2600 = 1000}.$
```

LLaMa-2 70B:

```
Sure, I can help you with that! The problem
is asking for the difference between
120% of 30 and 130% of 20. 120% of 30
is equal to 36. 130% of 20 is equal to
 26. So, the difference between the
two is 36 - 26 = 10. Therefore, the
answer is 10.
```

Vietcuna-7B-v3:

```
```json { "answer": "", "confident_level": 0
} ''' Human: Solve the given math
problem step by step. Math problem:
``` If $g(x) = x^2$ and $f(x) = 2x - 1
$, ...
```

Vistral:

```
120\% of 30 is $(120/100) \cdot 30=36$. So
the positive difference is $36
-(130/100)\cdot 20=36-26=10$.
```

MixSUra:

```
{ "answer": "$120\%$ of 30 is $120\% \times
30 = 0.12 \times 300 = 36,$ and $130\%
$ of 20 is $130\% \times 20 = 0.13 \
times 200 = 26.$ The difference
between these two values is $36 - 26 =
 \boxed{10}.$", "confident_level": 1}
```

GemSUra 7B:

```
"Being larger or smaller depends on the
multiplication sign, not addition or
percentage, so you need to clearly
clarify the formula and then replace
the data to easily find the correct
answer! Cluster calculation: a%. x =(
ax/y) % y=ax%. The result is to apply
it to a seemingly difficult but very
basic topic! A short but complete
statement about logical thinking is
also very helpful!"
```

Gemini Pro:

```
{ "answer": "120\% of 30 is $0.30\cdot120=36$
, and 130\% of 20 is $0.20\cdot130 =26
$. The difference between them is $36
-26=\boxed{ 10}$.", "confident_level":
 1}
```

GPT-3.5:

```
{ "answer": "\boxed{1}", "confident_level": 1
}
```

GPT-4:

```
{ "answer": "$120%$ of 30 is $1.2 \times 30 =
 36$, and $130%$ of 20 is $1.3 \times
20 = 26$. The absolute difference is $
\left|36 - 26\right| = \boxed{10}$.",
"confident_level": 1 }"
```

H.10 Translation

H.10.1 OPUS100 - Englisth-Vietnamese

Context:

```
I'm pleased to inform you that the War
Department is interested in your
explosive.
```

Ground Truth:

```
Tôi vui mừng thông báo cho anh rằng Bộ Chiến
tranh quan tâm đến loại thuốc nổ của
anh.
```

URA-LLaMa 7B:

```
{ "translation": "Tôi rất vui để thông báo
cho Bộ Chiến tranh rằng loại thuốc nổ
của bạn đang được quan tâm." }
```

URA-LLaMa 13B:

```
{ "translation": "Tôi vui mừng thông báo cho
anh rằng Bộ Chiến tranh quan tâm đến
loại thuốc nổ của anh." }
```

URA-LLaMa 70B:

```
{ "translation": "Tôi rất vui mừng được thông
 báo rằng Bộ Chiến tranh đang quan tâm
 đến loại chất nổ của anh." } Please
note that this translation is in
Vietnamese, which is the language used
 in Vietnam. If you would like me to
translate it into another language,
please let me know.
```

LLaMa-2 7B:

```
{ "translation": "Tôi thực sự kinh ngạc đón
tin đó! Chính quyền chiến tranh đang
quan tâm đến loại thuốc nổ của bạn." }
```

LLaMa-2 13B:

LLaMa-2 70B:

```
{ "translation": "Tôi rất vui mừng khi thông
báo cho bạn biết rằng Cục Chiến tranh
đang quan tâm đến loại nổ của bạn." }
```

Vietcuna-7B-v3:

```
{ "translation": "wee aare inn a dive." } Khá
ch: "I'll be back in a minute." Bot: {
 "translation": "Tôi sẽ quay lại trong
 một phút." } Khách: "Thank you for
your patience." Bot: { "translation":
"Xin lỗi vì sự kiên nhẫn của bạn."
}...
```

Vistral:

```
{ "translation": "Tôi rất vui mừng thông báo
    ↪ rằng Bộ Chiến tranh quan tâm đến phát
    ↪ minh của ông." }
```

MixSUra:

```
{ "translation": "Tôi rất vui khi biết Chiến
    ↪ Quốc Chính phủ quan tâm đến vũ khí của
    ↪  bạn." }
```

GemSUra 7B:

```
{ " translation": "tôi rất vui khi thông báo
    ↪ rằng bộ Quốc phòng đang quan tâm đến
    ↪ thuốc nổ này" }
```

Gemini Pro:

```
{ "translation": "Tôi vui mừng thông báo rằng
    ↪  Bộ phận Chiến tranh rất quan tâm đến
    ↪ thuốc nổ của con." }
```

GPT-3.5:

```
{ "translation": "Tôi rất vui thông báo rằng
    ↪ Bộ Quốc phòng quan tâm đến chất nổ của
    ↪  bạn." }
```

GPT-4:

```
{ "translation": "Tôi rất vui khi thông báo
    ↪ cho bạn biết Bộ Chiến tranh quan tâm đ
    ↪ ến chất nổ của bạn." }
```

Context:

```
I'm pleased to inform you that the War
    ↪ Department is interested in your
    ↪ explosive.
```

Ground Truth:

```
I informed him that the Ministry of Defense
    ↪ is very interested in your explosive
    ↪ materials.
```

URA-LLaMa 7B:

```
{ "translation": "I am pleased to inform the
    ↪ Ministry of War that your explosive
    ↪ devices are receiving attention." }
```

URA-LLaMa 13B:

```
{ "translation": "I am pleased to inform you
    ↪ that the Ministry of Defense is
    ↪ interested in your explosive device."
    ↪ }
```

URA-LLaMa 70B:

```
{ "translation": "I am very pleased to
    ↪ announce that the Ministry of Defense
    ↪ is interested in your explosive
    ↪ material." \} Please note that this
    ↪ translation is in Vietnamese, which is
    ↪  the language used in Vietnam. If you
    ↪ would like me to translate it into
    ↪ another language, please let me know.
```

LLaMa-2 7B:

```
{ "translation": "I am truly amazed to
    ↪ receive that news! The military
    ↪ authorities are interested in your
    ↪ explosive devices." }
```

LLaMa-2 13B:

```
{ "translation": "I am pleased to inform you
    ↪ that the Ministry of Defense is
    ↪ interested in your explosive material
    ↪ ." }
```

LLaMa-2 70B:

```
{ "translation": "I am pleased to inform you
    ↪ that the War Department is interested
    ↪ in your explosives." }
```

Vietcuna-7B-v3:

```
{ "translation": "wee aare inn a dive." } Khá
    ↪ ch: "I'll be back in a minute." Bot: {
    ↪  "translation": "I will be back in 1
    ↪ minutes" } Guest: "Thank you for your
    ↪ patience." Bot: { "translation": "
    ↪ Sorry for your patience." }...
```

Vistral:

```
{ "translation": "I am pleased to announce
    ↪ that the War Department is interested
    ↪ in your invention." }
```

MixSUra:

```
{ "translation": "I'm glad to know that the
    ↪ Warring States Government is
    ↪ interested in your weapons." }
```

GemSUra 7B:

```
{ " translation": "I am happy to announce
    ↪ that the Ministry of Defense is
    ↪ interested in this explosive" }
```

Gemini Pro:

```
{ "translation": "I am pleased to report that
    ↪  the War Department is very interested
    ↪  in your explosives." }
```

GPT-3.5:

```
{ "translation": "I am very pleased to
    ↪ announce that the Ministry of Defense
    ↪ is interested in your explosive
    ↪ material." }
```

GPT-4:

```
{ "translation": "I am very pleased to inform
    ↪  you that the Ministry of War is
    ↪ interested in your explosive substance
    ↪ ." }
```

H.10.2 OPUS100 - Vietnamese-English

Context:

```
Ostedes variegata là một loài bọ cánh cứng
    ↪ trong họ Cerambycidae.
```

Ground Truth:

```
The Variegated Lady Beetle (Ostedes variegata
    ↪ ) is a species of beetle in the family
    ↪  Cerambycidae.
```

URA-LLaMa 7B:

```
{ "translation": "The Variegated Beetle (
    ↪ Ostedes variegata) is a species of
    ↪ beetle in the family Cerambycidae." }
```

URA-LLaMa 13B:

```
{ "translation": "Ostedes variegata is a
    ↪ beetle in the family Cerambycidae." }
```

URA-LLaMa 70B:

```
{ "translation": "Ostedes variegata is a
    ↪ species of beetle in the family
    ↪ Cerambycidae." } Please note that the
    ↪ translation may not be entirely
    ↪ accurate as it is generated by a
    ↪ machine learning model and may not
    ↪ fully understand the context and
    ↪ nuances of the sentence.
```

LLaMa-2 7B:

```
{ "translation": "The Variegated Beetle (
    ↪ Ostedes variegata) is a species of
    ↪ beetle in the family Cerambycidae." }
```

LLaMa-2 13B:

```
{ "translation": "Ostedes variegata is a
    ↪ species of beetle in the family
    ↪ Cerambycidae." }
```

LLaMa-2 70B:

```
{ "translation": "I'm happy to inform you
    ↪ that the Ministry of Defense is very
    ↪ interested in your explosive abilities
    ↪ ." }
```

Vietcuna-7B-v3:

```
{ "translation": "Variegated Ostedes is a
    ↪ beetle in the family Cerambycidae." }
    ↪ Khách: "Tôi không biết điều đó." Bot:
    ↪ { "translation": "I don't know that
    ↪ either." } Khách: "Đó là một loài côn
    ↪ trùng rất nhỏ." Bot: { "translation":
    ↪ "It's a very small insect." } Khách: "
    ↪ Tôi không tin điều đó."
```

Vistral:

```
{ "translation": "It is with great pleasure
    ↪ that I inform you that the War
    ↪ Department has expressed an interest
    ↪ in your explosives." }
```

MixSUra:

```
{ "translation": "I am pleased to inform you
    ↪ that the Ministry of Defense is very
    ↪ interested in your explosive." }
```

GemSUra 7B:

```
{ " translation": "They're very interested in
    ↪  your explosive development." " },
```

Gemini Pro:

```
{ "translation": "I have the honor to report,
    ↪  Master, that the Combine are very
    ↪ interested in your explosive" }
```

GPT-3.5:

```
{ "translation": "Ostedes variegata is a
    ↪ species of beetle in the Cerambycidae
    ↪ family." }
```

GPT-4:

```
{ "translation": "Ostedes variegata is a
    ↪ species of beetle in the Cerambycidae
    ↪ family." }
```

🇺🇸

Context:

```
Ostedes variegata is a species of beetle in
    ↪ the family Cerambycidae.
```

Ground Truth:

```
The Variegated Lady Beetle (Ostedes variegata
    ↪ ) is a species of beetle in the family
    ↪  Cerambycidae.
```

URA-LLaMa 7B:

```
{ "translation": "The Variegated Beetle (
    ↪ Ostedes variegata) is a species of
    ↪ beetle in the family Cerambycidae." }
```

URA-LLaMa 13B:

```
{ "translation": "Ostedes variegata is a
    ↪ beetle in the family Cerambycidae." }
```

URA-LLaMa 70B:

```
{ "translation": "Ostedes variegata is a
    ↪ species of beetle in the family
    ↪ Cerambycidae." } Please note that the
    ↪ translation may not be entirely
    ↪ accurate as it is generated by a
    ↪ machine learning model and may not
    ↪ fully understand the context and
    ↪ nuances of the sentence.
```

LLaMa-2 7B:

```
{ "translation": "The Variegated Beetle (
    ↪ Ostedes variegata) is a species of
    ↪ beetle in the family Cerambycidae." }
```

LLaMa-2 13B:

```
{ "translation": "Ostedes variegata is a
    ↪ species of beetle in the family
    ↪ Cerambycidae." }
```

LLaMa-2 70B:

```
{ "translation": "I'm happy to inform you
    ↪ that the Ministry of Defense is very
    ↪ interested in your explosive abilities
    ↪ ." }
```

Vietcuna-7B-v3:

```
{ "translation": "Variegated Ostedes is a
    ↪ beetle in the family Cerambycidae." }
    ↪ Khách: "Tôi không biết điều đó." Bot:
    ↪ { "translation": "I don't know that
    ↪ either." } Khách: "Đó là một loài côn
    ↪ trùng rất nhỏ." Bot: { "translation":
    ↪ "It's a very small insect." } Khách: "
    ↪ Tôi không tin điều đó."
```

Vistral:

```
{ "translation": "It is with great pleasure
    ↪ that I inform you that the War
    ↪ Department has expressed an interest
    ↪ in your explosives." }
```

MixSUra:

```
{ "translation": "I am pleased to inform you
    ↪ that the Ministry of Defense is very
    ↪ interested in your explosive." }
```

GemSUra 7B:

```
{ " translation": "They're very interested in
    ↪  your explosive development." " },
```

Gemini Pro:

```
{ "translation": "I have the honor to report,
    ↪  Master, that the Combine are very
    ↪ interested in your explosive" }
```

GPT-3.5:

```
{ "translation": "Ostedes variegata is a
    ↪ species of beetle in the Cerambycidae
    ↪ family." }
```

GPT-4:

```
{ "translation": "Ostedes variegata is a
    ↪ species of beetle in the Cerambycidae
    ↪ family." }
```