ViGPTQA - State-of-the-Art LLMs for Vietnamese Question Answering: System Overview, Core Models Training, and Evaluations

Minh-Thuan Nguyen[1♡], **Khanh-Tung Tran**[1♡]*, **Vincent Nguyen**[2], **Xuan-Son Vu**[3†]
[1]DopikAI JSC, Hanoi, Vietnam
[2]University of Orleans, INSA CVL, LIFO EA, France
[3]Umeå University, Sweden
{minhthuan.nguyen, kt.tran}@dopikai.com
{vincent.nguyen}@univ-orleans.fr
{sonvx}@cs.umu.se

Abstract

Large language models (LLMs) and their applications in low-resource languages (such as in Vietnamese) are limited due to lack of training data and benchmarking datasets. This paper introduces a practical real-world implementation of a question answering system for Vietnamese, called ViGPTQA, leveraging the power of LLM. Since there is no effective LLM in Vietnamese to date, we also propose, evaluate, and open-source an instruction-tuned LLM for Vietnamese, named ViGPT. ViGPT demonstrates exceptional performances, especially on real-world scenarios. We curate a new set of benchmark datasets that encompass both AI- and human-generated data, providing a comprehensive evaluation framework for Vietnamese LLMs. By achieving state-of-the-art results and approaching other multilingual LLMs, our instruction-tuned LLM underscores the need for dedicated Vietnamese-specific LLMs. Our open-source model supports customized and privacy-fulfilled Vietnamese language processing systems.

1 Introduction

Large language models (LLMs), especially instruction-following models have achieved remarkable success in a wide range of natural language processing (NLP) tasks, demonstrating their ability to understand and generate human-like text. These models, including proprietary models such as ChatGPT, BingAI, and Bard, and open-source models such as LLaMA (Touvron et al., 2023), Alpaca (Taori et al., 2023), and Vicuna (Zheng et al., 2023), have been trained on vast amounts of text data, enabling them to learn intricate language patterns and capture semantic nuances.

While LLMs have shown impressive performance on various languages, there has been a noticeable gap in efforts dedicated to developing LLMs for Vietnamese, a low-resource language. Vietnamese possesses its own linguistic characteristics and contextual nuances, making it imperative to explore and optimize language models tailored to this unique language. Additionally, initial efforts for evaluating performances of multilingual LLMs (Lin et al., 2021; Zheng et al., 2023) have been carried out only for dominant languages, such as English and Chinese. As a result, it is important for thorough evaluation of Vietnamese LLMs. Comprehensive benchmarking will offer insights into the capabilities and potential limitations of LLMs when used with Vietnamese, enabling researchers and developers to fine-tune and optimize these models for optimal performance.

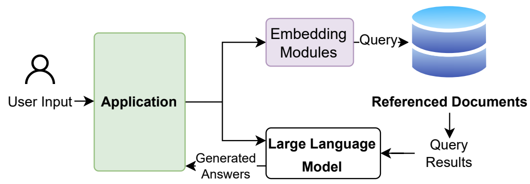

Figure 1: Our ViGPTQA system powered by LLM, combined with an embedding module to query and extract input for factual and referenced responses.

LLMs have significantly empowered various applications across multiple domains (Li et al., 2023; Wu et al., 2023). For example, they can be used to create question answering systems that provide more accurate and informative responses than traditional systems. Figure 1 illustrates our real-world question answering system called *ViGPTQA*, in which LLMs are combined with an embedding module to query and extract input from users for factual and referenced responses. A crucial point to highlight is that these applications require a highly proficient LLM to be feasible.

In this work, we propose, implement, evaluate, and open-source a Vietnamese monolingual instruction-tuned LLM, named *ViGPT*. By fine-tuning a pre-trained language model with specific instructions, ViGPT aims to enhance its perfor-

* Work conducted during research at DopikAI Labs;
† Corresponding author; ♡ Equal contributions

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 754–764
December 6-10, 2023 ©2023 Association for Computational Linguistics

mance and adaptability to the Vietnamese language. In addition to the general-purpose ViGPT model, we also introduce a law domain-specific variant, named *ViGPT-Law*, to power our ViGPTQA system. ViGPT-Law is specifically trained on a legal text corpus, allowing it to generate more accurate and informative responses to law-related queries. Moreover, we curate a comprehensive set of benchmark datasets specifically designed for evaluating Vietnamese LLMs. These datasets include both AI-generated and human-generated data, covering a wide range of emergent capabilities evaluations and task-specific challenges. This diverse range of benchmark datasets offer a standardized framework for assessing and comparing the performances of Vietnamese LLMs.

Our main contributions are listed as follows:

- We present ViGPTQA system, a practical real-world implementation of a question answering system for Vietnamese, harnessing the capabilites of LLM.
- We contribute an instruction-tuned LLM for Vietnamese, named ViGPT, with multiple variants, including domain-specific models.
- We curate a new set of benchmark datasets that encompass both AI-generated and human-generated data, providing a comprehensive evaluation framework for Vietnamese LLMs.
- We benchmark our proposed model on established datasets for Vietnamese on various tasks (question answering, named entiy recognition) and practical use cases with exceptional performances compared to previous methods.

Source code, benchmark datasets, and model weights are made publicly available at https://github.com/DopikAI-Labs/ViGPT for further advancement of customized and privacy-fulfilled systems for Vietnamese language processing.

2 Related Work

The development of large language models (LLMs) has gained significant attention in the natural language processing (NLP) community, leading to a plethora of research efforts on various aspects of LLMs for different languages (Zeng et al., 2021; Touvron et al., 2023; Taori et al., 2023; Zheng et al., 2023; Peng et al., 2023). Instruction-following language models have emerged as a promising direction to enable LLMs to generate targeted and controlled outputs based on user instructions. Recent studies have explored various methods for fine-tuning LLMs with instruction data, enhancing their performance on specific tasks and domains (Koleva et al., 2022; Qiao et al., 2022; Li et al., 2023; Wu et al., 2023; Chen et al., 2023). Prior research on Vietnamese language processing has been carried out to pre-train Vietnamese monolingual language models (Duong et al., 2021), with downstream application to tasks such as question answering (Phan et al., 2022; Tran et al., 2023), named entity recognition (Vu et al., 2019; Tran et al., 2023), and text summarization (Phan et al., 2022), exploring challenges specific to Vietnamese language. Existing models show promise in traditional NLP tasks but lack dedicated efforts for Vietnamese-specific LLMs and real-world applications. To the best of our knowledge, as of the time of writing this work, this is one of the first studies to introduce a billion-parameter Vietnamese instruction-tuned LLM with thoroughly benchmarked results, emphasizing real-world applicability.

3 Methodology

In this section, we will outline our fine-tuning approach and data curation process for training the generic ViGPT model and adapting it to the specific domain of Vietnamese laws, referred to as ViGPT-Law. Our primary objective is to expand the boundaries of LLM for Vietnamese, thereby empowering our ViGPTQA system.

3.1 ViGPT Finetuning Approach

Figure 2 demonstrates our finetuning strategy for ViGPT and its variants. As indicated in previous works (Touvron et al., 2023), there are two main crucial factors in training a high-quality instruction-following language model: a strong pre-trained model and high-quality instruction-following data. In our literature review, we assessed the current state of pre-trained large language models for Vietnamese. Regarding the second challenge, we leveraged a dataset comprising 52K instruction-following samples released by Alpaca (Taori et al., 2023). As the dataset was in English, we utilized OpenAI's gpt-3.5-turbo model (OpenAI, 2022) to translate the data into Vietnamese.

However, it is important to acknowledge that the 52K Alpaca dataset, as well as the transla-

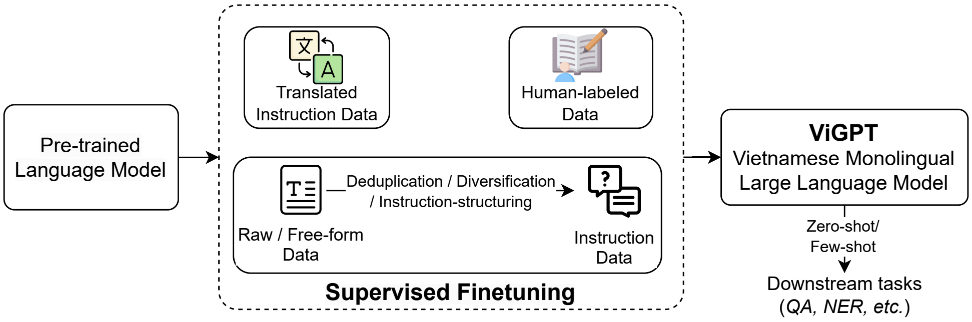

Figure 2: Overview of our finetuning strategies for ViGPT.

tion process conducted by another LLM. It might overlook or hallucinate distinct features unique to the Vietnamese language (e.g., characteristics of characters in Vietnamese novels or specific regulations). To address these limitations, we collected and processed an additional set of 1107 question-answering samples created by native Vietnamese users. It covers topics such as Vietnam's history, geography, and literature. We also utilized 3000 extractive question-answering samples from the VinewsQA dataset (Nguyen et al., 2020b) and generated 5000 synthetic abstractive question-answering samples within the Vietnamese law domain. Combining these 9107 native samples with the 52K examples, we performed supervised fine-tuning of our LLMs, resulting in *ViGPT-v2*. This variant differs from *ViGPT-v1*, which was only fine-tuned with the translated Alpaca dataset. The fine-tuning process employed Hugging Face's training framework (Huggingface), incorporating techniques like Fully Sharded Data Parallel, mixed precision training, and Low-Rank Adaptation (Hu et al., 2022).

3.2 Specific Domain Adaptation with ViGPT

Here, we discuss our finetuning strategy, focusing on utilizing the capabilities of ViGPT for a real-world application (ViGPTQA system) within the specific domain of Vietnamese laws. This approach can be extended to diverse applications in various other domains.

To adapt the LLMs for law domain, we initiated the process by gathering Vietnamese law documents (Vu, 2021). We meticulously curated this dataset by eliminating duplicate entries and documents containing fewer than 100 tokens. As a result, we collected 252425 Vietnamese law-related documents. This monolingual dataset served as the foundation for pre-training the LLMs to adapt them to this specific domain. For the generation of synthetic abstractive question-answering samples, we adopted the test set generation approach outlined in (Lance et al., 2023a). Leveraging the capabilities of the gpt-3.5-turbo model, this process automatically generated question-answer pairs based on text chunks. The key to harnessing the full potential of the gpt-3.5-turbo model lay in the provision of relevant context and suitable prompts, as mentioned in (Lance et al., 2023b). To apply this process, we randomly selected 1000 Vietnamese law documents, segmented them into 5000 chunks, each consisting of 4000 characters, and inputted them into the gpt-3.5-turbo model to produce question-answering pairs. Finally, we obtained 5000 synthetic question-answering samples. This synthetic data was leveraged as instruction-following training examples, as mentioned in Subsection 3.1.

As our base pre-trained LLM appears to lack substantial knowledge regarding Vietnamese law, to address the real-world abstractive question answering task, we first continue to pre-train the VietAI/gpt-j-6B model using our collected monolingual law dataset with next-word prediction task. This allows us to obtain a pre-trained LLM with extensive knowledge of Vietnamese law, referred to as *ViGPT-Law*. Then, we further finetune this *ViGPT-Law* model using the translated 52K Alpaca and the expanded version. This expanded version includes the initial 52K translated Alpaca and the 5000 synthetic Vietnamese law question-answering pairs that were collected as mentioned above. The outcome of this fine-tuning process was the creation of two distinct models: *ViGPT-Law-v1* and *ViGPT-Law-v2*, respectively. This process helps enhance the model's understanding of the legal domain.

In a real-world scenario, particularly in a domain like law that demands high accuracy, it is crucial for the LLM chatbot to provide precise and contextually relevant answers, referencing specific laws from official documents. To this end, we deploy a

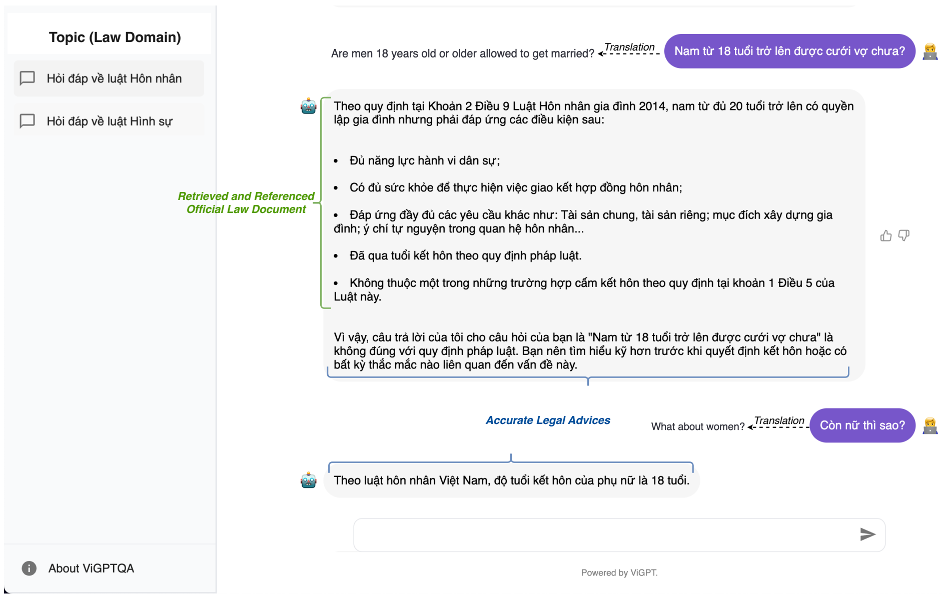

Figure 3: A demonstration of our deployed ViGPTQA system's user interface, featuring interactions in the Vietnamese legal field, where pairs of user's questions and their corresponding ViGPTQA's answers are presented.

vector database plugin for our chatbot, as shown in Figure 1. The process involves an embedding module that utilizes a similarity function to embed input questions and query the database of law documents. The module then combines the closest document and the question, providing input to ViGPT for accurate answer generation. This approach ensures that the chatbot can deliver precise and reliable responses in the law domain, meeting the demands of users seeking accurate legal information. Results are shown and analyzed in Section 4.1.

4 Experiments

In this section, we conduct experiments on traditional benchmarking tasks for Vietnamese language models to demonstrate the effectiveness of our proposed large language model, ViGPT. We first evaluate ViGPT and its variants on a newly curated Vietnamese law question answering dataset to assess the ability of the LLMs in powering ViGPTQA system. Then, we thoroughly benchmark various characteristics of ViGPT, including truthfulness and reasoning capability, and compare it against multilingual LLMs. The results show the strong capability of our model and highlight areas for future improvement.

4.1 Abstractive Question Answering

We evaluate our ViGPT models' performance on the abstractive question answering (AQA) task in Vietnamese. AQA requires models to comprehend input question and context, generating human-like answers that may not be exact replicas of specific text spans. As there is no benchmark available for this task in Vietnamese, we introduce ViLawsQA, a curated dataset from official law documents of Vietnam. Questions, answers, and legal citations are collected from the official Vietnamese law website (Vu, 2021), and questions suitable for the AQA task are selected, resulting in 1020 context-question-answer samples spanning across 27 law categories.

An example interaction with pairs of question and corresponding answer within the Vietnamese law domain is provided in Figure 3, showcasing the front-end interface of our deployed ViGPTQA system.

To assess models' performances, we utilize automated metrics (ROUGE-1, BLEU-1, and BLEU-4) and human evaluation. Three Vietnamese annotators are asked to score 200 random samples using a 0-4 Likert scale, where a score of 4 indicates a perfect answer and 0 signifies a totally false answer. Scores of 3, 2, and 1 represent mostly true, half true, and partly true answers, respectively. Average scores are used for model evaluation. Note that we report both ROUGE-1 scores from 2 implementations: one from our implementation that used the correct tokenizer for Vietnamese, and another labeled as ROUGE-1-Non-Unicode, which is calculated using the Python library *rouge-score* (Google, 2022) and the widely-used wrapper library *evaluate* (HuggingFace, 2022). The latter implementation employs an unchangeable text tokenizer that filters out all Unicode characters, including all characters

Index	Model	Context	ROUGE-1 (Unicode)	ROUGE-1-Non-Unicode*	BLEU-1	BLEU-4	Human
1	vi_mrc$_{large}$	No	0	0	0	0	
2	gpt-3.5-turbo	No	33.99	46.87	23.57	11.95	2.11
3	ViGPT-v1	No	30.83	53.17	20.08	8.02	0.48
4	ViGPT-v2	No	30.97	53.12	20.25	8.57	0.51
5	ViGPT-Law-v1	No	31.25	48.52	23.41	10.21	0.64
6	ViGPT-Law-v2	No	31.42	52.64	25.53	14.98	1.23
7	vi_mcr$_{large}$	ground truth	14.85	17.43	2.26	1.62	0.62
8	ViGPT-Law-v2	vietnamese-sbert	42.10	58.22	30.53	17.82	2.17
9	ViGPT-Law-v2	embedding-ada-002	43.22	59.21	32.34	19.76	2.24
10	ViGPT-Law-v2	ground truth	**45.33**	**59.62**	**33.82**	**21.11**	**2.52**

Table 1: Abstractive Vietnamese Question Answering Task - ViLawsQA task. *Here we note that ROUGE-1-Non-Unicode scores are calculated using the python library *rouge-score* (Google, 2022) and the popular wrapper library *evaluate* (HuggingFace, 2022), which uses an unchangeable text tokenizer that removes all Unicode characters, including all Vietnamese punctuation. This suboptimal approach for comparing Vietnamese texts may lead to incorrect benchmarking results that do not fully capture the richness of the language.

with Vietnamese punctuation. This suboptimal approach for comparing Vietnamese texts may lead to inaccurate benchmarking results. Nevertheless, we have included the results from this less suitable implementation to raise awareness and encourage further research into more accurate evaluation methods for the Vietnamese language.

Table 1 displays our experimental results on the ViLawsQA task. The *Context* column indicates whether models utilize the given context to answer the question. Experiments 1 to 6 assess the model's ability to answer questions based solely on the input question. The vi_mrc$_{large}$ (Binh, 2021) model fails without the given context as it is an extractio model, while the gpt-3.5-turbo model achieves the highest human evaluation score of 2.11 out of 4. Among our four models, the *ViGPT-Law-v2* model obtains the highest score at 1.23, showcasing the effectiveness of our domain adaptation and synthetic data generation process (Section 3.2). However, these results suggest that the majority of answers provided by all models are not useful for humans. Therefore, when given a question, it is vital to retrieve relevant documents to support the model in answering based on that knowledge.

Experiments 7 to 10 demonstrate models' performance when context is provided. We use two top Vietnamese text semantic retrieval models, vietnamese-sbert (Hieu, 2022) and embedding-ada-002 (Greene et al., 2022), to retrieve the context for given questions. Ground truth context is human-crafted and contains necessary information to answer the question. Experiment 7 reveals vi_mrc$_{large}$ model performs poorly compared to our models on all evaluation metrics, even with ground truth context. This is due to the task requires comprehension and synthesis of the answer from the given context, which is challenging for an extraction model. Experiment 10 shows that our model generates useful answers for humans when given the ground truth context, scoring 2.52 out of 4 on human evaluation. However, this scenario is not always practical as obtaining correct context for each question is difficult. Experiments 8 and 9 demonstrate our solutions for retrieving suitable context perform well on all four evaluation metrics, scoring 2.17 and 2.24 on human evaluation, respectively. Although vietnamese-sbert scores slightly lower than embedding-ada-002, its open-source nature and ease of deployment make it advantageous for real-world applications compared to the paid embedding-ada-002 model.

4.2 Extractive Question Answering

We benchmark our ViGPT models on Vietnamese extraction-based machine reading comprehension (MRC) datasets, including ViCoQA (Luu et al., 2021), ViNewsQA (Nguyen et al., 2020b), ViWikiQA (Do et al., 2021), and ViQuAD 2.0 (Nguyen et al., 2022). We compare the performances of our models, ViGPT-v1 and ViGPT-v2, with the state-of-the-art models vi_mrc$_{large}$ (Binh, 2021). F1 and Similarity scores are leveraged as automatic metrics. F1 measures token overlap between predicted and human-annotated answers, while Similarity score assesses semantic similarity between two answers using vietnamese-sbert model (Hieu, 2022). Human evaluation is also performed by scoring 200 randomly selected question-answer pairs on a 0-4 Likert scale similar to the abstractive question answering task above.

The experiment results in Table 2 demonstrate that our *ViGPT-v1*$\star$ model performs poorly in terms of both F1 and human scores across the four tasks,

Model	ViWikiQA			ViCoQA			ViNewsQA			ViQuAD 2.0		
	F1	Sim.	Human	F1	Sim.	Human	F1	Sim.	Human	F1	Sim.	Human
vi_mrc$_{large}$	**54.15**	61.52	**2.73**	**63.54**	66.26	**2.57**	23.58	40.78	1.72	**72.42**	**70.72**	**2.82**
ViGPT-v2	51.55	**63.25**	2.67	61.93	**73.01**	2.52	**52.98**	**70.97**	**2.69**	52.31	65.57	2.41
ViGPT-v1⋆	10.78	50.92	0.85	20.71	58.22	1.26	10.19	58.74	0.97	25.9	61.12	0.91
ViGPT-v2⋆	45.23	58.92	2.22	48.01	63.02	2.27	39.01	52.96	1.75	47.76	61.41	2.32

Table 2: Extractive Vietnamese Question Answering Tasks. ⋆ denotes few-shot fine-tuning.

indicating that the 52K instruction-following Alpaca dataset is not effective for fine-tuning the LLM on Vietnamese MRC tasks. However, the *ViGPT-v2⋆* model, which incorporates a subset of 3K samples from the ViNewsQA dataset into the 52K Alpaca data, performs well, achieving approximately 50.0 F1 scores and receiving human ratings of over 2.0 for most tasks. This shows the strong ability of our large and general model to solve this task. Additionally, the *ViGPT-v2* model, trained on the entire training dataset for these tasks, performs almost as well as the vi_mrc$_{large}$ model on three tasks (ViQuAD 2.0, ViCoQA, and ViNewsQA) and outperforms the vi_mrc$_{large}$ model on the ViWikiQA task in terms of all F1, Similarity, and human scores.

Model	F1
ETNLP$_{MULTI}$ (Vu et al., 2019)	91.09
XLM-R$_{large}$ (Nguyen et al., 2020a)	93.8
PhoBERT$_{base}$ (Nguyen and Nguyen, 2020)	94.2
ViT5$_{base\ 1024-length}$ (Phan et al., 2022)	94.5
ViT5$_{large\ 1024-length}$ (Phan et al., 2022)	93.8
ViDeBERTa$_{large}$ (Tran et al., 2023)	**95.3**
VietAI/gpt-j-6B⋆	68.65
ViGPT-v1⋆	**69.31**
ViGPT-v2⋆	<u>68.92</u>

Table 3: Evaluation results (%) for NER task on PhoNER dataset. ⋆ denotes few-shot fine-tuning.

4.3 Named Entity Recognition

Here, we explore the performance of LLMs on the NER task in a few-shot scenario, where only a small number of samples are available for fine-tuning. We randomly select 100 samples from the training set of PhoNER (Truong et al., 2021) to train three models: *ViGPT-v1*, *ViGPT-v2*, and VietAI/gpt-j-6B. We use supervised fine-tuning and freeze the model's weights, only finetune the classifier head on top. The evaluation results on PhoNER test set are presented in Table 3.

Remarkably, even with just 100 training samples, *ViGPT-v1* achieves a commendable level of performance in terms of F1-score (69.31) and accuracy (91.85%), followed by *ViGPT-v2*, with an F1-score of 68.92. These results are noteworthy when compared to previous approaches that relied on fine-tuning with the entire training set, consisting of 5000 samples. Additionally, when compared to the pre-trained only model, the efficacy of instruction-based fine-tuning for few-shot learning on downstream tasks has not been previously studied. Despite this, both *ViGPT-v1* and *ViGPT-v2* still demonstrate better performance compared to VietAI/gpt-j-6B, making them a more preferable choice for NER tasks. These results emphasize the potential of ViGPTs in real-world tasks where minimal training data is available, making them a highly practical and effective solution.

4.4 ViTruthfulQA

We present ViTruthfulQA, a dataset for evaluating truthfulness of a LLM in generating answers to questions, similar to (Lin et al., 2021). Our dataset consists primarily of samples focused on various aspects of Vietnam's information, including history, geography, and literature. We curate the dataset to be adversarial by inputting the samples through gpt-3.5-turbo and filtering out questions that can be easily answered by the model. As mentioned in Subsection 4.1, we also report the less suitable ROUGE-1 score, ROUGE-1-Non-Unicode, for inclusion.

We compare *ViGPT-v1* and *ViGPT-v2* with five methods: vilm/vietcuna-3b (vilm ai, 2023), which is also a LLM trained with SFT objective, VietAI/gpt-neo-1.3B (VietAI, 2021) and VietAI/gpt-j-6B (pre-trained only methods), gpt-3.5-turbo, a multilingual LLM, and BingAI - based on gpt-3.5-turbo with Internet access plugin. In this work, we did not include results of popular open-source models such as Llama and Llama2 or closed-source such as Google Bard, since they cannot stably generate Vietnamese answers for our questions; their responses are mostly in English. The benchmarking results, as shown in Table 4, highlight the capabilities of ViGPT models compared to other Vietnamese LLMs, where our model outperforms previous approaches, demonstrating a significant gap in terms of truthfulness (human

Model	Human	ROUGE-1 (Unicode)	ROUGE-1-Non-Unicode	BLEU-1	BLEU-4
vilm/vietcuna-3b	6.02	27.10	37.76	43.29	9.8
VietAI/gpt-neo-1.3B	7.23	12.92	22.52	8.87	1.01
VietAI/gpt-j-6B	7.93	14.91	23.81	10.35	2.00
gpt-3.5-turbo	**29.91**	31.84	51.02	32.84	7.27
BingAI †	**74.08**	38.85	53.78	51.62	**18.61**
ViGPT-v1	18.50	27.73	46.67	40.26	5.73
ViGPT-v2	25.45	**43.26**	**56.56**	**57.53**	14.25

Table 4: Evaluation results on VitruthfulQA. † denotes method has access to the Internet.

evaluation) score. Despite having a smaller number of parameters compared to gpt-3.5-turbo, our model still performs admirably (25.45 in human evaluation score of *ViGPT-v2* compared to 29.91 of gpt-3.5-turbo). It is worth noting that BingAI, which has addtional plugins that allow for internet access, theoretically should be able to answer all questions. However, its actual truthfulness performance is 73.88%, indicating room for future improvement. Additionally, there is a strong correlation between human-based metrics and automated evaluation metrics, specifically ROUGE-1, BLEU-1, and BLEU-4. *ViGPT-v2* achieves the best scores in ROUGE-1 (56.67) and BLEU-1 (57.53), and comes in second place in BLEU-4, closely following BingAI.

Baseline	Baseline Score	ViGPT-v2 Score
vilm/vietcuna-3b	109.0	**369.0**
VietAI/gpt-neo-1.3B	167.0	**322.0**
VietAI/gpt-j-6B	150.0	**369.0**
ViGPT-v1	268.0	**313.0**
gpt-3.5-turbo	**658.5**	319.0

Table 5: Average score judged by gpt-4 on 80 translated samples of Vicuna-Instructions-80.

4.5 Automatic Evaluation with LLM

Vicuna-Instructions-80 is a dataset synthesized by gpt-4 with 80 challenging questions across 8 categories, including knowledge, math, Fermi, counterfactual, roleplay, generic, coding, writing, common-sense. The dataset is translated to Vietnamese using gpt-3.5-turbo with human corrections. Moreover, following the original approach, we perform automatic evaluation of models on this dataset using gpt-4 as the evaluator. Relevant works have found a strong agreement of over 80% between human evaluators and strong LLMs that act as evaluator, such as gpt-4 (Zheng et al., 2023). We benchmark ViGPT-v2 against five baselines: ViGPT-v1, vilm/vietcuna-3b, VietAI/gpt-neo-1.3B, VietAI/gpt-j-6B, and gpt-3.5-turbo. Results in Table 5 demonstrate the effectiveness and usefulness of our model, as we surpass all other Vietnamese-specific language models. Moreover, our proposed model, ViGPT-v2, outperforms more than half of the questions with respect to all other monolingual LLMs, highlighting its superiority. However, it is important to acknowledge that our model still faces a significant performance gap when compared to gpt-3.5-turbo. The performance difference is due to gpt-3.5-turbo's task-specific fine-tuning and the significant scale gap between our method (6B parameters) and gpt-3.5-turbo (175B parameters). Hence, one potential future direction is to scale our ViGPTs to the size of current multilingual models.

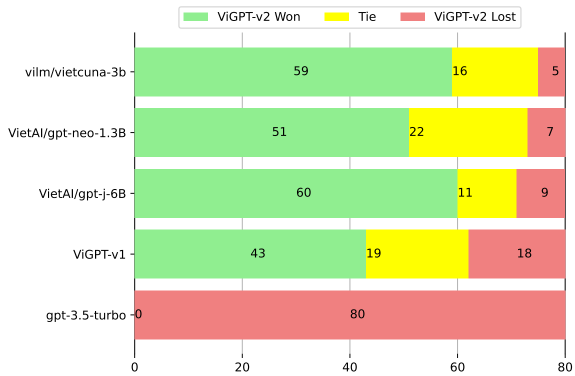

Figure 4: Response comparison assessed by gpt-4.

5 Conclusion

This work focuses on advancing large language models for the Vietnamese language and developing real-world applications, namely ViGPTQA. We introduce ViGPT, an instruction-following LLM for Vietnamese. We propose novel datasets for fine-tuning language models (instruction data), adaptation to a specific domain (ViLaws dataset), and benchmarking datasets for Vietnamese LLMs. Initial evaluations showcase the usefulness of ViGPT and variants on downstream tasks, and its emergent capability, compared to other multilingual LLMs. We provide public access to our datasets, model codes, and weights, fostering collaboration and enabling reproducibility. Our research contributes to the development of LLMs for Vietnamese, paving the way for specialized and efficient LLMs.

6 Limitations

Although our ViGPT models demonstrate promising results in various critical Vietnamese NLP tasks, such as machine reading comprehension and named entity recognition, they still exhibit certain limitations in achieving high performance. Firstly, the quality of the instruction-following data for Vietnamese is relatively low, and it is insufficient to help the LLMs handle multiple tasks within a single model. This limitation became evident during our benchmarking and analysis, particularly when incorporating our proposed law-domain specific knowledge. However, it is crucial to develop a general-purpose LLM for the Vietnamese language, regardless of specific domains, to address this limitation effectively. Secondly, regarding fairness and bias, while ViGPTs have demonstrated sufficient truthfulness in its generated answers, there is still a large gap compared to absolute truthfulness, as with other models. Furthermore, additional experiments are warranted to further evaluate the fairness of the model, ensuring that biases are adequately addressed and mitigated.

7 Acknowledgment

The authors would like to thank the DopikAI Technology Company (http://DoPik.AI), Quang-Anh Bui, Xuan-Vu Dinh, Tien-Tung Bui, Thanh-Tu Nguyen, and many annotators for their hard work to support the evaluation task. Without their support, the work would not have been possible.

References

Nguyen Vu Le Binh. 2021. Machine reading comprehension special for the vietnamese language. https://huggingface.co/nguyenvulebinh/vi-mrc-large. Accessed: 2023-07-23.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Phong Nguyen-Thuan Do, Nhat Duy Nguyen, Tin Van Huynh, Kiet Van Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Sentence extraction-based machine reading comprehension for vietnamese. *CoRR*, abs/2105.09043.

Duong, Thanh, and Binh. 2021. GPT-J 6B on vietnamese news. https://huggingface.co/VietAI/gpt-j-6B-vietnamese-news. Accessed: 2023-07-23.

Google. 2022. Python rouge implementation. Accessed: 2023-07-23.

R. Greene, T. Sanders, L. Weng, and A. Neelakantan. 2022. New and improved embedding model. https://openai.com/blog/new-and-improved-embedding-model. Accessed: 2023-07-23.

Le Ngoc Hieu. 2022. Sentence transformer model for the vietnamese language. https://huggingface.co/keepitreal/vietnamese-sbert. Accessed: 2023-07-23.

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*.

Huggingface. Huggingface framework. https://huggingface.co/. Accessed: 2023-07-23.

HuggingFace. 2022. Evaluate library. Accessed: 2023-07-23.

Aneta Koleva, Martin Ringsquandl, Mark Buckley, Rakeb Hasan, and Volker Tresp. 2022. Named entity recognition in industrial tables using tabular language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 348–356, Abu Dhabi, UAE. Association for Computational Linguistics.

Lance, Danil, and Ben. 2023a. Langchain evaluator. https://github.com/langchain-ai/auto-evaluator. Accessed: 2023-07-23.

Lance, Danil, and Ben. 2023b. Langchain evaluator prompt. https://github.com/hwchase17/langchain/blob/master/langchain/chains/qa_generation/prompt.py. Accessed: 2023-07-23.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge.

Stephanie C. Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*.

Son T. Luu, Mao Nguyen Bui, Loi Duc Nguyen, Khiem Vinh Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Conversational machine reading comprehension for vietnamese healthcare texts. *CoRR*, abs/2105.01542.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020a. A Vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kiet Nguyen, Son Quoc Tran, Luan Thanh Nguyen, Tin Van Huynh, Son Thanh Luu, and Ngan Luu-Thuy Nguyen. 2022. VLSP 2021-ViMRC challenge: Vietnamese machine reading comprehension. *VNU Journal of Science: Computer Science and Communication Engineering*, 38(2).

Kiet Van Nguyen, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020b. New vietnamese corpus for machine readingcomprehension of health news articles. *CoRR*, abs/2006.11138.

OpenAI. 2022. Gpt-3.5 docs. https://platform.openai.com/docs/models/gpt-3-5. Accessed: 2023-07-23.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4.

Long Phan, Hieu Trung Tran, Hieu Chi Nguyen, and Trieu H. Trinh. 2022. Vit5: Pretrained text-to-text transformer for vietnamese language generation. In *North American Chapter of the Association for Computational Linguistics*.

Lingfeng Qiao, Chen Wu, Ye Liu, Haoyuan Peng, Di Yin, and Bo Ren. 2022. Grafting pre-trained models for multimodal headline generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 244–253, Abu Dhabi, UAE. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Cong Dao Tran, Nhut Huy Pham, Anh Tuan Nguyen, Truong Son Hy, and Tu Vu. 2023. ViDeBERTa: A powerful pre-trained language model for Vietnamese. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1071–1078, Dubrovnik, Croatia. Association for Computational Linguistics.

Thinh Hung Truong, Mai Hoang Dao, and Dat Quoc Nguyen. 2021. COVID-19 Named Entity Recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

VietAI. 2021. GPT-Neo 1.3b on vietnamese news. https://huggingface.co/VietAI/gpt-neo-1.3B-vietnamese-news. Accessed: 2023-07-23.

vilm ai. 2023. Vietnamese large language model. https://huggingface.co/vilm/vietcuna-3b. Accessed: 2023-07-23.

Bui Tuong Vu. 2021. Official vietnamese law website. https://thuvienphapluat.vn/hoi-dap-phap-luat. Accessed: 2023-07-23.

Xuan-Son Vu, Thanh Vu, Son Tran, and Lili Jiang. 2019. ETNLP: A visual-aided systematic approach to select pre-trained embeddings for a downstream task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1285–1294, Varna, Bulgaria.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, Chen Li, Ziyan Gong, Yifan Yao, Xinjing Huang, Jun Wang, Jianfeng Yu, Qi Guo, Yue Yu, Yan Zhang, Jin Wang, Hengtao Tao, Dasen Yan, Zexuan Yi, Fang Peng, Fangqing Jiang, Han Zhang, Lingfeng Deng, Yehong Zhang, Zhe Lin, Chao Zhang, Shaojie Zhang, Mingyue Guo, Shanzhi Gu, Gaojun Fan, Yaowei Wang, Xuefeng Jin, Qun Liu, and Yonghong Tian. 2021. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Reproducibility

The finetuning process employed Hugging Face's training framework (Huggingface), utilizing techniques such as Fully Sharded Data Parallel, mixed precision training, and Low-Rank Adaptation (Hu et al., 2022). Based on these techniques, our training pipeline can be completed with low cost, in around 12 hours on a single RTX 3090 24Gb GPU. Warmup steps, batch size, learning rate, and cutoff length are set to 100, 64, 0.004, and 1024, respectively. For Low-Rank Adaptation, we set lora rank, lora alpha, lora dropout, and lora target modules to

16, 16, 0.05, and [q_proj, v_proj], respectively. In generation stage, we adopt top-p sampling as the default decoding method with a temperature = 0.5, top-p = 0.7, and repetition penalty = 1.2.

B Automatic Evaluation with LLM for ViGPT-v1

Baseline	Baseline Score	ViGPT-v1 Score
vilm/vietcuna-3b	116.0	**395.0**
VietAI/gpt-neo-1.3B	189.0	**360.0**
VietAI/gpt-j-6B	167.0	**375.0**
gpt-3.5-turbo	**654.5**	348.0

Table 6: Comparisons of ViGPT-v1 with baselines, judged by gpt-4 on 80 translated samples of Vicuna-Instructions-80.

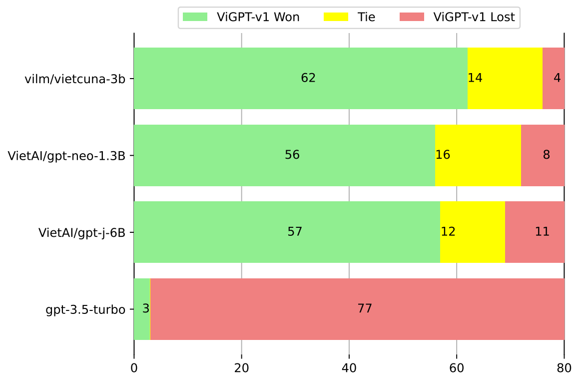

Figure 5: Response comparison of ViGPT-v1 with baselines, assessed by gpt-4.

We perform comparison for ViGPT-v1 with other baselines on Vicuna-Instructions-80 as in section 4.5, with gpt-4 as the automated judge. The results illustrated in Table 6 and Figure 5 confirm our findings: both ViGPT-v1 and ViGPT-v2 outperform other monolingual language models for Vietnamese, demonstrating the effectiveness of our finetuning strategy.

C Datasets Description

In this section, we provide a detailed description about the novel proposed training and evaluation datasets for ViGPT models.

C.1 Training Datasets

- **Vietnamese Alpaca Instruction-Following Data.** We utilized the gpt-3.5-turbo model to translate 52K samples of Alpaca instruction-following (Taori et al., 2023) into Vietnamese. This dataset enables us to establish an initial Vietnamese instruction-following model and explore the cross-language generalization capability of instruction-tuning.

- **Vietnamese Question Answering Data.** As the knowledge within the 52K Alpaca dataset is general and not specific to Vietnamese, we have compiled and curated an additional dataset comprising 1107 question-answering samples generated by native Vietnamese users. These samples cover topics such as Vietnam's history, geography, and literature.

- **Vietnamese Extractive Question Answering Data.** To enhance our model's ability to comprehend the provided context for answering questions, we incorporate a limited subset of 3000 samples from the VinewsQA dataset (Nguyen et al., 2020b) into our instruction-tuning dataset. Each sample in this span-extraction dataset comprises an *input*: the question, an *instruction*: the passage containing the answer span text, and an *output*: the answer to the question extracted from the provided passage.

- **ViLawsQA Training Set.** Owing to the absence of a dataset for abstractive question answering tasks in Vietnamese, we present a collection of 5000 synthetic samples in the field of Vietnamese law. These samples are generated through the test set generation process within langchain-ai (Lance et al., 2023a). This process employs the gpt-3.5-turbo model or ChatGPT to automatically formulate question-answering samples based on segments of text and appropriate prompts (Lance et al., 2023b). In this study, we initially randomly selected 1000 Vietnamese legal documents, dividing them into 5000 text segments each comprising 4000 characters. Coupled with fitting prompts, these segments were inputted into the gpt-3.5-turbo model to derive question-answering pairs. This synthetic data, consisting of (question, text segment, answer) combinations, was incorporated into our instruction-following training dataset.

- **Vietnamese Law Documents Data.** In order to adapt LLMs for the Vietnamese law domain, we gathered Vietnamese law documents from the official Vietnamese law website (Vu, 2021). These documents underwent processing involving the elimination of duplicates and documents containing fewer than 100 tokens. Ultimately, we acquired a dataset comprising 252425 documents related to Vietnamese law. This monolingual dataset was used for pre-training LLMs to facilitate specific domain adaptation. In our work, lengthy documents were segmented into chunks, each with a maximum length of 1024 tokens.

C.2 Evaluation Datasets

- **ViLawsQA Test Set.** We have reserved a set of 1020 samples for evaluating Vietnamese question-answering performance in the law domain. Each sample in the test set comprises an *input*: the question, *instruction*: contextual information required to answer the question, and *output*: the ground truth answer. During evaluation, models receive the input and instruction and must effectively extract the relevant information from the instruction to generate accurate answers.

- **VitruthfulQA.** We propose ViTruthfulQA, a dataset that is comparable to (Lin et al., 2021) for assessing how truthful an LLM is while generating responses to questions. The majority of the samples in our dataset are devoted to different aspects of Vietnam's knowledge, such as its history, geography, and literature. By running the samples through gpt-3.5-turbo and filtering queries that the model can easily answer, we design the dataset to be adversarial. The final number of samples are 213. Each sample in the dataset includes an *input*: question about a known fact related to Vietnam, *Correct answers*: various different ways to response to the question correctly, and *Incorrect answers*: answers that are incorrect or mimic common falsehood about the question. The desired usage of this dataset is to evaluate truthfulness ability of language models by comparing their generated responses with the set Correct answers and Incorrect answers.

- **Vietnamese Vicuna-Instructions-80.** Vicuna-Instructions-80 (Chiang et al., 2023) is a dataset with 80 questions that baseline models find challenging, across 8 categories, including knowledge, math, Fermi, counterfactual, roleplay, generic, coding, writing, common-sense, created by gpt-4. The dataset is translated to Vietnamese using gpt-3.5-turbo with human corrections. The dataset is carried out to evaluate models' capabilities such as reasoning, hallucination, etc. Following the original approach, we perform automatic evaluation of models on this dataset using gpt-4 as the evaluator. Relevant works have found a strong agreement of over 80% between human evaluators and strong LLMs that act as evaluator, such as gpt-4 (Zheng et al., 2023).