ViGLUE: A Vietnamese General Language Understanding Benchmark and Analysis of Vietnamese Language Models

Minh-Nam Tran[1,2*], **Phu-Vinh Nguyen**[1,2*], **Long Nguyen**[1,2†], **Dien Dinh**[1,2]
[1]Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{tmnam20,npvinh20}@apcs.fitus.edu.vn, {nhblong,ddien}@fit.hcmus.edu.vn

Abstract

As the number of language models has increased, various benchmarks have been suggested to assess the proficiency of the models in natural language understanding. However, there is a lack of such a benchmark in Vietnamese due to the difficulty in accessing natural language processing datasets or the scarcity of task-specific datasets. **ViGLUE** [1], the proposed dataset collection, is a **Vi**etnamese **G**eneral **L**anguage **U**nderstanding **E**valuation benchmark developed using three methods: translating an existing benchmark, generating new corpora, and collecting available datasets. ViGLUE contains twelve tasks and encompasses over ten areas and subjects, enabling it to evaluate models comprehensively over a broad spectrum of aspects. Baseline models utilizing multilingual language models are also provided for all tasks in the proposed benchmarks. In addition, the study of the available Vietnamese large language models is conducted to explore the language models' ability in the few-shot learning framework, leading to the exploration of the relationship between specific tasks and the number of shots.

1 Introduction

Since the introduction of the Transformer architecture (Vaswani et al., 2017) and its variations, there has been significant progress in many natural language processing tasks. The expansion of pre-trained language models leveraging that design, such as BERT (Devlin et al., 2018), GPT (Radford et al., 2018), and T5 (Raffel et al., 2020), is primarily responsible for this progress. Besides that, the need to evaluate such models for natural language understanding has been raised. GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) are introduced as well-designed benchmarks to evaluate English models in NLU, lacking the ability to analyze models in other languages. Consequently, particular benchmarks for other languages have been proposed, such as the CLUE benchmark (Xu et al., 2020) for Chinese, the FLUE benchmark (Le et al., 2020) for French, the KLUE benchmark (Park et al.) for Korean, the RussianSuperGLUE (Shavrina et al., 2020) for Russian, or the IndicGLUE (Kakwani et al., 2020) for Indian. They are developed for assessing the performance of language-specific pre-trained language models.

While numerous benchmarks exist for examining the NLU capabilities of language models, there is a lack of comparable benchmarks in Vietnamese, particularly those that are openly accessible for instant usage. Therefore, this study aims to address this gap by developing and introducing a comprehensive benchmark designed to evaluate the NLU capabilities of language models in Vietnamese. Creating such benchmarks is crucial for assessing the performance of language models in Vietnamese and fostering advancements in natural language processing specific to this language. To establish such a benchmark, it is necessary to carefully choose and organize a wide range of tasks that encompass different linguistic features, contextual complexities, topic diversity, and domain variety.

This study introduces ViGLUE as an evaluation framework over twelve NLU tasks, detailed in Table 1. The creation of ViGLUE involves utilizing three methods: the translation of existing benchmarks, the collection of publically available datasets, and the building of new corpora. Eight of the nine tasks of GLUE (Wang et al., 2018) are initially translated into Vietnamese and thereafter gone under back-translation to verify the translation quality. Furthermore, two publicly available Vietnamese datasets, namely VSFC (Nguyen et al., 2018) and VSMEC (Ho et al., 2020a), have

[*]Both authors contributed equally to this research.
[†] Corresponding author.
[1]Source code is available at: https://github.com/trminhnam/ViGLUE and the dataset is published at: https://huggingface.co/datasets/tmnam20/ViGLUE.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 4174–4189
June 16-21, 2024

Corpus	Train	Validation	Test	Method	Metric	Domain
Natural Language Inference Tasks						
MNLI	392,702	9,815	9,796	Semi-translating	Acc.	Miscellaneous
QNLI	104,743	5,463	5,463	Semi-translating	Acc.	Wikipedia
RTE	2,490	277	3,000	Semi-translating	Acc.	Miscellaneous
VNRTE	12,526	3,137	-	Constructing	Acc.	News
WNLI	635	71	146	Semi-translating	Acc.	Fiction books
Sentiment Analysis Tasks						
SST2	67,349	872	1,821	Semi-translating	Acc.	Movie reviews
VSFC	11,426	1,538	3,166	Collecting	Acc.	Student feedback
VSMEC	5,548	686	693	Collecting	Acc.	Social media
Similarity and Paraphrase Tasks						
MRPC	3,668	408	1,725	Semi-translating	Acc./F1	News
QQP	363,846	40,430	390,965	Semi-translating	Acc./F1	Quora QA
Single-Sentence Tasks						
CoLA	8,551	1,043	1,063	Semi-translating	MCC	Miscellaneous
VToC	7,293	1,831	-	Constructing	Acc.	News

Table 1: Task statistics of ViGLUE. NLI stands for natural language inference. Acc is for accuracy, and MCC stands for Matthews correlation coefficient. Column "Method" points out the method to obtain the corresponding corpus.

been collected to broaden the evaluation scope of ViGLUE to include students' feedback and multimedia comments, respectively. Finally, two new tasks are created based on the Vietnamese news documents with the help of the GPT model. The objective of creating additional corpora is to broaden the topic coverage of the ViGLUE benchmark.

Besides constructing the benchmark, several large Vietnamese language models have been studied on the proposed benchmark with zero-, one-, and few-shot learning to explore their ability to understand Vietnamese. The larger the model's capacity is, the more beneficial the model receives under the few-shot evaluation. Meanwhile, the performance of small pre-trained language models decreased when the number of shots increased.

In conclusion, the contributions are:

- First, a public Vietnamese general language understanding evaluation benchmark is proposed with twelve tasks clustered into four groups covering several domains.
- Second, the baseline models leveraging multilingual language models are proposed to provide a comparison of the pre-trained models in Vietnamese language understanding tasks.
- Finally, an empirical study is demonstrated with few-shot learning on the Vietnamese large language models and multilingual large language models incorporating Vietnamese knowledge on the proposed benchmark.

2 Related Work

In natural language processing, evaluating the ability of the language models on natural language understanding is necessary. With the birth of pre-trained Transformer-architecture modes such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2018), GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) appear as standard benchmarks to test the methods on different aspects. The GLUE (Wang et al., 2018) dataset focuses on text genres and degrees of difficulty. At the same time, SuperGLUE (Wang et al., 2019) is an extended version of GLUE (Wang et al., 2018) with improvement in difficulty and novel tasks. However, while GLUE and SuperGLUE focus on English, using them to evaluate large language models that are created for other languages would be a noticeable problem.

In response to this need, many research groups spent effort on creating comparable datasets in different languages, utilizing the building pipeline of GLUE and SuperGLUE, while these evaluation frameworks just assessed models in English. Consequently, plenty of benchmarks for language-specific natural language understanding were developed. The CLUE benchmark (Xu et al., 2020) was designed with nine various Chinese NLU tasks, constructed from multiple resources (Chinese news, app description, etc.). The FLUE benchmark (Le et al., 2020) was developed for the French language. The specialty of this dataset is that it includes three over six tasks from cross-lingual datasets. Moreover, the KLUE benchmark (Park et al.) is created with two goals: covering diverse aspects of NLU in Korean and minimizing redundancy among tasks.

To construct those benchmarks, techniques such as data gathering (KLUE (Park et al.), CLUE (Xu et al., 2020)), re-using existing datasets (FLUE (Le et al., 2020)) are employed. An alternative approach is to create new multilingual variances of the original dataset with human and machine translation. XNLI (Conneau et al., 2018), XCOPA (Ponti et al., 2020), XTREME (Hu et al., 2020), and the RussianSuperGLUE (Shavrina et al., 2020) are constructed in this way. However, this method depends on the quality of the translation tools (Conneau et al., 2018) and may not maintain the context of specific tasks. Despite this problem, the translation method offers the advantage of utilizing the reliable pool of workers employed in the original datasets (Conneau et al., 2018). Furthermore, it has been observed that the hypotheses and semantic aspect of the samples remain consistent when applied to different languages (Conneau et al., 2018).

Having such datasets like GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) without any restrictions, free to access, and convenient to use for Vietnamese would be crucial to evaluate large language models such as PhoGPT (Nguyen et al., 2023), Vietcuna. Even though there are some Vietnamese datasets, such as the Vietnamese Students' Feedback Corpus (VSFC) (Nguyen et al., 2018), the Vietnamese Social Media Emotion Corpus (VSMEC) (Ho et al., 2020b), and the COVID-19 named entity recognition dataset for Vietnamese (Truong et al., 2021). However, the availability of additional resources is restricted from direct open access due to the necessity of obtaining user approval, which is a significant inconvenience for researchers working with Vietnamese NLP datasets.

Hence, a new Vietnamese NLU benchmark is introduced. This dataset is constructed by utilizing the translating method, building from news sources, and collecting the existing datasets. Nevertheless, the back-translation process is proposed to assure the translation quality of the translated texts in terms of both lexical and semantic elements.

3 ViGLUE Overview

ViGLUE contains twelve tasks obtained by three methods: collecting, constructing, and translating. The task statistics are shown in Table 1. The licenses are discussed in the Appendix A.

3.1 Tasks

In ViGLUE, tasks are clustered into four groups: natural language inference tasks, sentiment analysis tasks, similarity and paraphrase tasks, and single-sentence tasks. Each part below lists the tasks in each cluster and describes them.

3.1.1 Natural Language Inference Tasks

MNLI, originated from MultiNLI (Williams et al., 2017), is a corpus of multi-genre texts. It is built to assess natural language models' performance in sentence comprehension. The objective is to determine the relationship between a given premise and hypothesis, namely whether the hypothesis logically implies the premise, contradicts the premise, or has no logical connection between them.

QNLI is constructed from the Stanford Question Answering Dataset (Rajpurkar et al., 2016). It contains multiple questions and paragraphs, where a single in each section serves as a solution to the corresponding question. The GLUE benchmark modified the original dataset to introduce a new task of determining whether the context sentence is a proper answer to the provided question.

RTE (recognizing textual entailment) is similar to a natural language inference task since it evaluates the ability of machine models to comprehend the semantic meaning of sentences (Dagan et al., 2010). Following GLUE, this corpus comprises three datasets, including RTE1 (Dagan et al., 2005), RTE2 (Bar-Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009). Furthermore, it is worth noting that there are two distinct labels, entailment and not entailment, because the GLUE authors merged neutral and contradiction labels into "not entailment" class.

VNRTE, short for Vietnamese News Recognizing Textual Entailment, is built on online news doc-

uments crawled from VnExpress, an official Vietnamese online news webpage. This includes about 16,000 Vietnamese sentences being separated to entail and not entail labels, roughly 8,000 sentences for each label. This task is specially designed by leveraging two tools, semantic similarity search and the GPT model to create the samples.

WNLI comes from the Winograd Schema Challenge (Levesque et al., 2012). The original dataset evaluates the machine learning models for understanding the meaning of the ambiguous pronoun inside the sentence and choosing the correct reference from a list of options. The task WNLI in the GLUE benchmark is converted into a classification problem, where the sentence containing the proper pronoun is classified as entail or not entail.

3.1.2 Sentiment Analysis

SST2, also known as the Stanford Sentiment Treebank (Socher et al., 2013), contains individual sentences extracted from movie reviews and annotated by humans. In this task, the models must recognize the sentence's sentiment as positive or negative, a binary classification problem.

VSFC, known as Vietnamese Students' Feedback Corpus (Van Nguyen et al., 2018), contains more than 16,000 student responses about lectures, curriculum, facilities, etc. For each text piece in this task, the model has to predict whether that response is positive, negative, or neutral.

VSMEC, which is the Vietnamese Social Media Emotion Corpus (Ho et al., 2020a). This dataset includes nearly 7,000 sentences labeled with six fundamental emotions (enjoyment, disgust, sadness, anger, fear, surprise) or "other" for the sentence with an emotion not belonging to the six above.

3.1.3 Similarity and Paraphrase Tasks

MRPC, short for Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005), is a collection of sentence pairs extracted from internet newswire articles. The task requires the model to recognize the semantic equivalence between two sentences.

QQP is a paraphrase-based problem that involves question pairs sourced from the Quora website (Iyer et al., 2017). The task is to decide whether two questions are semantically equivalent or have the same answer even if the questions are different.

3.1.4 Single-sentence Tasks

CoLA, introduced by Warstadt et al. (2019), provides annotated samples for the language acceptability task (or grammar error detection task). The corpus is constructed using literary publications and scholarly articles within linguistic theory.

VToC, stands for Vietnamese Topic Classification. This dataset is just another variance of the VNRTE dataset, where each sentence is labeled with the article's topic. VToC covers 15 topics, including Automobile, Business, Digital, Education, Entertainment, Health, Law, Life, News, Perspective, Relax, Science, Sports, Travel, and World.

3.2 Dataset Construction

The target is to create a high-quality benchmark that is easily accessible and freely available to all individuals. To build the ViGLUE benchmark, all three mentioned methods to construct the NLU benchmarks are utilized, including translating source benchmarks to Vietnamese, gathering available datasets, and making new corpora. The following information outlines the process of creating ViGLUE by employing these techniques.

3.2.1 Semi-Translating

The translation approach, as utilized in previous studies (Conneau et al., 2018; Ponti et al., 2020), is employed to transform eight subsets of the GLUE benchmark from English to Vietnamese. These subsets include CoLA, MNLI, MRPC, QNLI, QQP, RTE, SST2, and WNLI, as outlined in Table 1. To mitigate the potential for interdependence across features while translating, we handle them individually. By utilizing the GLUE dataset and employing the Google Translate API, the output dataset effectively inherits the contributions of the worker pool, as well as the varying levels of task complexity and diversity (Williams et al., 2017). Furthermore, the meanings of the premises and hypotheses remain unchanged, which is useful for inference tasks.

Instead of using human verification similar to XCOPA (Ponti et al., 2020), XNLI (Conneau et al., 2018) and XGLUE (Liang et al., 2020), the back-translation (Edunov et al., 2018) method is used. For each translated task, we sampled one hundred examples and translated all of them back into the original language (English in particular). Then, the original and back-translated sentences are compared by calculating BLEU, METEOR, and semantic similarity scores to judge the translation quality.

Semantic similarity is used to ensure the consistency of the meaning through the translating process. It is computed by calculating the cosine similarity score between the semantic representations of original and back-translated sen-

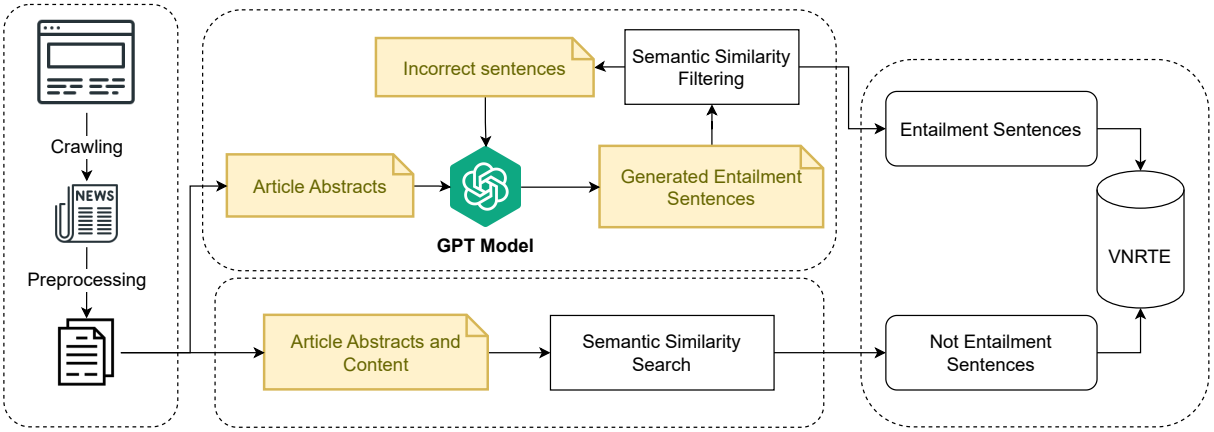

Figure 1: The VNRTE creation pipeline has three stages, including (1) article crawling and preprocessing, (2) generating entailment sentences, and (3) semantic searching inside articles for non-entailment sentences.

tences. First, the representation is obtained by using a pre-trained Siamese BERT (Reimers and Gurevych, 2019) network $f(x;\theta)$ with pre-trained parameters θ. In this case, the network (or embedding model) transforms input text x into a k-dimensional dense embedding vector $e = f(x;\theta)$. Then, the semantic similarity between sentences x_1 and x_2 is determined by cosine similarity, calculated as $score(x_1, x_2) = \frac{e_1 \cdot e_2}{||e_1|| \cdot ||e_2||}$, where e_1, e_2 are embedding vectors obtained by feeding x_1, x_2 through the embedding model, respectively.

The lexical aspect is also guaranteed by computing BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) metrics between the original and back-translated sentences. BLEU is a common evaluation metric for comparing translation results, while METEOR extends the ability of BLEU by matching the candidates based on the surface form, stemmed form, and meaning.

The quality checking results by back translation are provided in Table 2. Except for SST2 and CoLA, The BLEU metric of all subsets over 50 scores, and the METEOR value of all tasks encompass 70. The semantic similarity score also gets high results, all larger than 90 except SST2. On average, the translated corpora maintain the meaning of the text, which is the most important aspect when the task of GLUE focuses on inference or semantic meaning between two sentences.

3.2.2 Constructing

Besides the translated tasks from the GLUE benchmark, two corpora are constructed based on the Vietnamese online articles. The creation pipeline of VNRTE is visualized in Figure 1.

Subset	BLEU	METEOR	Semantic Similarity
CoLA	46.53	76.38	93.40
MNLI	50.51	77.28	93.22
MRPC	60.03	83.40	95.46
QNLI	57.97	81.37	95.44
QQP	58.46	83.14	95.90
RTE	60.23	83.38	96.74
SST2	39.74	68.72	88.52
WNLI	50.06	77.07	94.04
Avg	53.50	79.41	94.30

Table 2: Results of backtranslation evaluation using BLEU, METEOR, and Semantic similarity metrics. The Semantic Similarity is the average cosine similarity score between original GLUE sentence embeddings and back-translated GLUE sentence embeddings obtained from a pre-trained Siamese BERT network (Reimers and Gurevych, 2019). All the metrics are scaled to 100.

First, the raw documents were obtained by crawling through the VnExpress Online Newspaper, an official Vietnamese online news platform. After performing the preprocessing step, including removing unnecessary components like URLs and HTML tags, a clean document set is created.

For the VNRTE corpus, there are two stages to design the task, which are described as follows:

- The first flow (or the upper flow in Figure 1) uses the abstract sentences of the articles, followed by rewriting by GPT model to obtain sentences with similar meanings. To avoid the error of the generative model, such as generated output having a different meaning to the original sentence, semantic similarity filtering is applied. The method is to compute the cosine similarity score between the embedding

vector of the abstract statement and the generated text, then filter out which pairs have a score less than 0.9 to feed the corresponding abstract into the model for rewriting the entailment sentence. This stage loops four times and removes all the sentence pairs having a score less than 0.9 at the end.

- For the second flow, the abstract statement of the article is used with its corresponding news content. In this stage, the objective is to utilize the entire document content to identify the combination of sentences that do not entail each other. Using the abstract statement as the anchor text, similarity search scores are computed between the anchor and the content sentences. Subsequently, a sentence is randomly selected, with a similarity score falling within the range of 0.3 to 0.4, to designate it as the sentence that still has the same topic with the anchor text but does not entail it.

Concerning the VToC corpus, it utilizes the abstract sentence from each article as the input data for the Vietnamese topic classification task. The abstract sentence is selected from the clean document along with the document's topic, which serves as the label for this problem. Because VnExpress is an official Vietnamese online news platform, the quality of the abstract statement, document topic, and article content is asserted before publication.

3.2.3 Collecting

The collecting approach utilizes the pre-existing Vietnamese datasets to assess tasks. With this method, two Vietnamese datasets, including VSFC (Van Nguyen et al., 2018) and VSMEC (Ho et al., 2020a), are collected and preprocessed by removing emoticons and emojis. Ultimately, the two datasets are regarded as the tasks for assessing large language models. ViGLUE has expanded its range of tasks and benefits from the previously well-constructed datasets by employing this method.

3.3 Dataset Analysis

The statistics of each task in the ViGLUE dataset, including the number of sentences, the number of tokens, vocabulary size, and the average number of tokens per sentence, are provided in Table 3. To measure these statistics, `Underthesea`[2] library is used for sentence and word tokenization.

[2] https://github.com/undertheseanlp/underthesea

According to the data in Table 3, it is evident that QQP has the largest vocabulary size, with 69,796 unique tokens, surpassing the second largest, QNLI, which has 50,759 unique tokens. Despite MNLI having the highest token count, its vocabulary size ranks third, with 41,701 unique tokens. The average sentence length, measured in tokens per sentence (tps), is highest for VNRTE, with an average of 30.24 tps. VToC, MRPC, RTE, and QNLI follow with average sentence lengths of 29.72, 28.80, 27.26, and 26.70 tps respectively. The average sentence length of other assignments is less than 25.

Corpus	#Sents	#Tokens	#Vocab	Avg#TpS
CoLA	9,621	102,290	3,819	10.63
MNLI	848,739	17,989,715	41,701	21.20
MRPC	11,970	344,761	10,714	28.80
QNLI	227,513	6,073,566	50,759	26.70
QQP	903,686	13,287,371	69,796	14.70
RTE	7,982	217,599	9,980	27.26
SST2	68,569	959,762	7,319	14.00
VNRTE	6,436	194,595	8,366	30.24
VSFC	1,583	21,647	1,157	13.67
VSMEC	878	9,573	1,785	10.90
VToC	1,916	56,940	5,471	29.72
WNLI	1,767	26,190	1,426	14.82
Avg	174,221	3,273,667	17,691	20.22

Table 3: ViGLUE task statistics. #Sents denotes the number of sentences; #Tokens denotes the number of tokens; #Vocab denotes the vocabulary size, and Avg#TpS denotes the average number of tokens per sentence.

4 Experiments

This section describes the baseline models used in the experiments conducted on the ViGLUE benchmark and the analysis of evaluation results in the model's language understanding capability.

4.1 Baselines

The ViGLUE benchmark utilizes two types of baseline models: the non-trained technique and the fine-tuning of pre-trained language models.

4.1.1 Majority Baseline

For each task, the class label with the highest proportion is taken as a prediction over the entire test set. This algorithm is also known as the ZeroR classifier (Aher and Lobo, 2012). Even lacking predictive capabilities, it offers a strong baseline for the classification tasks (Nasa and Suman, 2012).

	Natural Language Inference					Sentiment Analysis			Similarity/Paraphrase		Single-sentence Tasks	
Model	**MNLI**	**QNLI**	**RTE**	**VNRTE**	**WNLI**	**SST2**	**VCSFC**	**VSMEC**	**MRPC**	**QQP**	**CoLA**	**VToC**
Metric	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc./F1	Acc./F1	MCC	Acc.
ZeroR	35.45	50.54	52.71	53.11	**56.34**	50.92	50.85	31.20	68.38/81.22	63.38/0.00	0.00	6.77
mBERT	79.66	89.11	**70.76**	99.97	**56.34**	88.42	93.62	53.64	85.29/88.85	89.12/85.16	14.13	81.43
XLM-R_{base}	81.61	88.17	62.45	**100.00**	**56.34**	89.45	94.95	55.25	83.82/88.26	89.46/85.87	3.64	83.07
XLM-R_{large}	35.45	**91.23**	67.51	**100.00**	54.93	**90.14**	**95.39**	37.9	**88.24/91.64**	**90.48/87.22**	0.0	**87.82**
mDeBERTaV3	**83.34**	89.99	69.31	99.97	**56.34**	89.79	95.07	**55.39**	86.52/90.05	89.98/86.69	**19.62**	80.88

Table 4: Evaluation results of the baseline models on the validation subset. MNLI uses the validation_matched. All tasks use the accuracy metric, except that CoLA uses the Matthews correlation coefficient (MCC). All the results are multiplied by 100 and rounded to two decimal places. The best result of each task is shown in **bold**.

4.1.2 Pre-trained Models

To get the baseline models, the following pre-trained encoder-only Transformer models incorporating Vietnamese knowledge are used:

- mBERT is the multilingual variant of BERT (Devlin et al., 2018), which is one of the first pre-trained language models trained with two self-supervised learning tasks, namely masked language modeling and next sentence-prediction, on the multilingual unlabeled data.
- XLM-RoBERTa (or XLM-R) (Conneau et al., 2020) is a masked language model trained on 2.5TB of data in 100 languages to boost performance on multilingual downstream tasks.
- mDeBERTa V3 (He et al., 2021a) refers to a collection of DeBERTa V3 models that have been trained using CC100 data. In addition, it employs ELECTRA-Style pre-training with Gradient Disentangled Embedding Sharing (He et al., 2021b) to effectively perform unsupervised learning on unlabeled corpora.

The mentioned pre-trained models are selected due to their multilingual ability, providing good weight initialization for the baseline models and comprehensive coverage of the tokenizer's vocabulary to prevent out-of-vocabulary phenomena. In addition, mBERT and mDeBERTaV3 are utilized in their basic variants, whilst XLM-RoBERTa is employed in its base and large configurations.

4.1.3 Fine-tuning

Although all the tasks in ViGLUE are different regarding the specific task (sentiment analysis, paraphrase, natural language inference), they are sequence classification problems. Consequently, the output of the Transformer at the first index in the sequence is fed into a classifier and optimized via cross-entropy loss. The training configuration consists of 3 epochs each session, utilizing a learning rate 2e-5 and employing the Adam optimizer. For every task, each model undergoes three times fine-tuning, and the best performance checkpoint on the validation set is selected to be reported.

4.2 Benchmark Results

Evaluation results of baseline models for each task in ViGLUE are reported in Table 4. The `evaluate`[3] library is used to load and compute the metric.

XLM-R_{large} surpasses all models on seven over twelve tasks: QNLI, VNRTE, VSFC, MRPC, QQP, SST2, and VToC. It is noticeable that XLM-R_{large} has superior performance in similarity and paraphrase tasks, surpassing all other models in this group. In addition, XLM-R_{large} achieves the greatest performance compared to the different models in two out of three single-sentence tasks. However, its accuracy drops significantly in task VSMEC, at 37.9% compared to around 53% to 55% of other fine-tuned models. When comparing XLM-R_{large} to XLM-R_{base}, XLM-R_{base} performance does not exceed its large version on any task. It indicates that increasing model capacity enhances the model's performance on most tasks.

The mDeBERTaV3 model ranks second on the ViGLUE benchmark, outperforming other models in MNLI, VSMEC, and CoLA tasks. Nevertheless, in tasks where mDeBERTaV3 does not outperform other models, the gap in performance between them is negligible, with a margin of less than 1%.

On the task WNLI, all models achieve the same accuracy, 56.34%, equal to the ZeroR classifier (majority model), except XLM-R_{large}. It's because the task is too challenging for the model to recognize the entailment relationship between the premise and the hypothesis of the WNLI samples. The GLUE benchmark explained a difference be-

[3] https://github.com/huggingface/evaluate

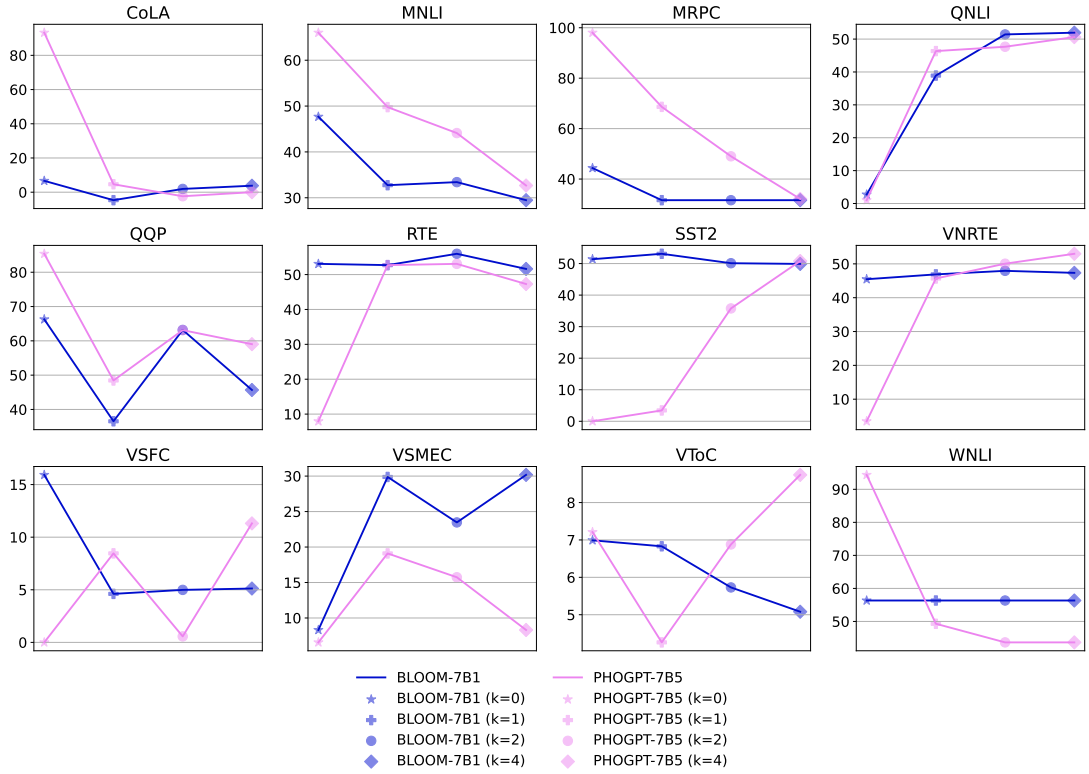

Figure 2: Evaluation results of BLOOM-7B1 and PhoGPT on twelve ViGLUE tasks (multiplied by 100). All tasks are reported with accuracy, except for CoLA using Matthews correlation coefficient.

tween the distribution of the train set and the test set, indicating that the baseline models overfit the train subset of WNLI in this finetuning schema.

In short, XLM-R_{large} is a multilingual language model that outperforms mBERT and mDeBERTaV3 through the evaluation results in seven out of twelve tasks. In addition, it has been shown that expanding the model size enhances the model's ability to understand general language.

5 Few-shot Learning with Vietnamese Large Language Models

Besides finetuning the pre-trained encoder-only Transformer models, several Vietnamese large language models, also known as autoregressive language models, are evaluated on ViGLUE to compare their language understanding ability.

This experiment applies the few-shot learning framework (Brown et al., 2020), also referred to as in-context learning, to the BLOOM model family (Workshop et al., 2022) and PhoGPT model (Nguyen et al., 2023). The advantage of this approach is eliminating the need for model finetuning, thereby reducing the hardware requirements associated with LLMs. Additionally, the notation k denotes the number of samples in the context.

From Figure 2 in Appendix C.1, PhoGPT outperforms BLOOM-7B1 on a small number of tasks. When using zero-shot learning ($k = 0$), PhoGPT demonstrates superior performance compared to BLOOM-7B1 on six out of twelve tasks. Similarly, BLOOM-7B1 gets better performance compared to PhoGPT on five out of twelve tasks for $k = 1$, seven out of twelve tasks for $k = 2$, and five out of twelve tasks for $k = 4$. Furthermore, it is evident that increasing the number of examples in the prompts adversely affects the model's efficacy, namely on CoLA, MNLI, MRPC, and WNLI. It is noticeable that zero-shot learning achieves the highest scores in CoLA, MNLI, MRPC, QQP, and WNLI, to other values of k even though it does not offer any task-specific instruction. Nevertheless, the concept of increasing the number of examples resulting in a performance improvement remains valid for QNLI and VNRTE. The statistics for QQP, SST2, VSFC, and VSMEC exhibit uneven fluctuations when the number of shots changes.

Figure 3, in Appendix C.2, visualizes the few-shot learning performance of the BLOOM model family. It is obvious that increasing k benefits the models on QNLI for any model size. In contrast, it also shows that CoLA, MNLI, MRPC, and QQP

do not benefit from few-shot learning since the performance at $k = 0$ always gets the highest values among all k values. Regarding the model size, the largest model, BLOOM-176B, achieves roughly identical performance to other configurations on most workloads (CoLA, MNLI, MRPC, and VToC). It outperforms models with lower capacities only on QNLI, RTE, and SST2, while for other tasks, it cannot beat them. When the model size increases from 560M to larger, the performance also increases on several tasks, such as QNLI, QQP, RTE, SST2, VSFC, and VSMEC. Nevertheless, the BLOOM-560M model, which is the smallest in size, outperforms the other models in the BLOOM family in both VNRTE and WNLI tasks. On inference tasks, raising the number of parameters in multilingual language models sometimes fails to increase the performance.

6 Conclusion

By providing ViGLUE, the issue of missing a Vietnamese natural language understanding benchmark is tackled. ViGLUE is built using three methods, including translating available benchmarks, constructing new corpora, and gathering available Vietnamese datasets. The baseline models finetuned on multilingual language models are provided for all ViGLUE tasks and XLM-RoBERTa$_{\text{large}}$ achieve the best performance on seven over twelve tasks compared to mBERT and mDeBERTaV3. Besides exploring the encoder architecture, large language models are also examined using few-shot learning. We observe that on CoLA, MNLI, MRPC, QQP, and WNLI, the models perform better without any task instruction. In contrast, models achieve greater performance on QNLI and VNRTE when increasing the number of samples in the context.

The number of shots employed in the experiments is restricted to only four values (0, 1, 2, and 4). For further investigation, experimenting with different and large values of the number of examples in the context is considered. In addition, tasks in ViGLUE focus on short sentences or sentence pairs. Future work should focus on long sentences or text at a higher level. Finally, ViGLUE covers news, Wikipedia, textbooks, and publication domains. Therefore, expanding the scope of ViGLUE to medical, law, and education is a good direction.

Limitation

Despite the benchmark spread over twelve corpora, ViGLUE still has some limitations as follows.

In the GLUE benchmark, CoLA requires the models to classify if the input sentence is grammatically correct or not. However, the machine translator sometimes corrects the translated text, leading to the wrong label in the translated CoLA tasks. The way to fix this is to sophisticate the translated sentence where the label is `unacceptable`. Furthermore, using a human translator instead of a machine translator is an alternative approach.

The number of shots used in the few-shot learning contexts is limited due to the high hardware requirements when increasing the number of shots. Therefore, the observation above is only true for cases where $k \in \{0, 1, 2, 4\}$. Investigating more values of k is a direction for future research.

Potential Risks

The ViGLUE benchmark contains datasets for sentiment analysis tasks, which may include negative-feeling sentences. Hence, users of ViGLUE should be aware of this negative aspect and avoid developing such an unethical model based on sentences expressing a negative emotion. We guarantee that any biases present in the benchmark are inadvertently introduced, and our objective is not to cause harm to any individual or entity. We strongly advocate for the proper utilization of the datasets to drive advancements in natural language processing.

Acknowledgements

This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

References

Sunita B Aher and LMRJ Lobo. 2012. Comparative study of classification algorithms. *International Journal of Information Technology*, 5(2):239–243.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4. Venice.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. 2010. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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. 2020a. Emotion recognition for vietnamese social media text. In *Computational Linguistics: 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019, Hanoi, Vietnam, October 11–13, 2019, Revised Selected Papers 16*, pages 319–333. Springer.

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. 2020b. Emotion recognition for vietnamese social media text. In *Computational Linguistics: 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019, Hanoi, Vietnam, October 11–13, 2019, Revised Selected Papers 16*, pages 319–333. Springer.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. 2017. First quora dataset release: Question pairs.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Chitra Nasa and Suman Suman. 2012. Evaluation of different classification techniques for web data. *International journal of computer applications*, 52(9):34–40.

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.

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

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 *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

S Park, J Moon, S Kim, WI Cho, J Han, J Park, C Song, J Kim, Y Song, T Oh, et al. Klue: Korean language understanding evaluation. arxiv 2021. *arXiv preprint arXiv:2105.09680*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

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*.

Kiet Van Nguyen, Vu Duc Nguyen, Phu XV Nguyen, Tham TH Truong, and Ngan Luu-Thuy Nguyen. 2018. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th international conference on knowledge and systems engineering (KSE)*, pages 19–24. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Appendix

A Licenses and Terms of Use

Since ViGLUE uses tasks translated from the GLUE benchmark, it inherits all of the licenses available in the GLUE. Furthermore, information about licenses of two collected datasets, VSFC and VNRTE, is also provided in Table 5.

Task	License
CoLA	CC BY 4.0
MLNI	CC BY 4.0
MRPC	CC BY 4.0
QNLI	CC BY-SA 4.0
QQP	CC BY 4.0
RTE	CC BY-SA 3.0
SST2	CC BY 4.0
VNRTE	CC BY-NC-ND 4.0
VSFC	-
VSMEC	-
VToC	CC BY-NC-ND 4.0
WNLI	CC BY 4.0

Table 5: Licenses of tasks in ViGLUE. Notation "-" denotes that there is no information about the license.

For CoLA, MRPC, QQP, SST2, and WNLI, they do not provide licenses for these tasks so the license of the GLUE benchmark, which is CC BY 4.0, is used instead. The two task VNRTE and VToC, which are created from the content of VNExpress website, are published under CC BY-NC-ND 4.0 and do not serve for commercial use. Finally, ViGLUE is published under a CC BY 4.0 license, and we highly recommend that ViGLUE should be used only for academic purposes only.

B Training Setup

B.1 Hardware Requirements

All the training sessions are run on a single NVIDIA A100-PCIE-40GB, with 64GB of RAM and 12 CPU cores. For inference and few-shot learning runs, the same hardware setting is applied.

B.2 Training Hyperparameters

For finetuning the baseline models, the AdamW optimizer is used with $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1\text{e}{-8}$. For a single training session, the model is trained for three epochs with a global batch size of 32, a learning rate of $2\text{e}{-5}$, and fp32 precision. The model undergoes finetuning for each task in three sessions, each using a different seed value: 1, 10, and 100 accordingly. Moreover, long model inputs are truncated to a maximum of 256 tokens.

B.3 Model Sizes and Training Time

The model capacity is shown in Table 6 and the training period of the experiments for each model across all tasks is provided in Table 7.

Model	#Params
mBERT	178M
XLM-R_{base}	278M
XLM-R_{large}	560M
mDeBERTaV3	279M

Table 6: Number of trainable parameters for the multilingual pre-trained models finetuned as the baselines.

Task	mBERT	XLM-R_{base}	XLM-R_{large}	mDeBERTaV3
CoLA	00:01:08	00:02:44	00:02:27	00:02:37
MNLI	02:34:51	02:50:50	05:38:01	04:03:56
MRPC	00:00:52	00:01:07	00:02:45	00:02:04
QNLI	00:39:39	00:46:07	01:38:38	01:22:58
QQP	01:20:18	03:16:38	03:59:45	02:56:57
RTE	00:01:14	00:01:14	00:03:54	00:01:51
SST2	00:13:30	00:17:15	00:29:52	00:27:18
VNRTE	00:04:06	00:04:24	00:12:38	00:08:31
VSFC	00:02:27	00:02:48	00:07:08	00:04:39
VSMEC	00:01:10	00:01:24	00:03:37	00:02:14
VToC	00:01:24	00:01:36	00:04:14	00:02:41
WNLI	00:00:09	00:00:09	00:00:38	00:00:20

Table 7: Training time of the baseline models on twelve ViGLUE tasks. The time follows HH:MM:SS format.

B.4 BERT-like Model Benchmark Results

The models are loaded and trained with the script from the `transformers` framework (Wolf et al., 2020). Besides the multilingual language models as the baselines, additional results of the Vietnamese models are provided, including PhoBERT (Nguyen and Nguyen, 2020) and ViDeBERTa (Tran et al., 2023).

The benchmark results of all training sessions are reported in Table 8. Matthews correlation coefficient is used for CoLA, accuracy/F1 score are used for MRPC and QQP, and accuracy is used for the rest of the tasks of ViGLUE. Each of the chosen pre-trained models is finetuned on each task in three sessions, using seed values of 1, 10, and 100, respectively. The model's highest performance on each challenge, across three different seeds, is indicated by underlining. The highest performance on each task among all models is displayed in **bold**.

C Few-shot Learning Evaluation Results

C.1 BLOOM-7B1 vs PhoGPT

For the highly fair comparison between a multilingual language model containing Vietnamese knowledge and a large language model mainly pre-trained on Vietnamese text, BLOOM-7B1 and PhoGPT are chosen in this experiment, with the number of parameters at 7.1 billion and 7.5 billion, respectively. The evaluation results are shown in Figure 2.

C.2 Benchmark Results of BLOOM Family

There are six configurations in the BLOOM model family with the number of parameters 560M, 1.1B, 1.7B, 3B, 7.1B, and 176B. The evaluation results on twelve tasks are visualized in Figure 3.

C.3 Community Model Evaluation

In addition to evaluating models from publications, many Vietnamese large language models published by the community are utilized to assess the ViGLUE benchmark. The objective is to offer a concise measurement of the language understanding capabilities of the models to the community, helping them in selecting a suitable Vietnamese LLM. The results are shown in Table 9.

Besides BLOOM and PhoGPT, two Vietnamese large language model families provided by the community are evaluated on the ViGLUE benchmark. The models are listed as follows,

- The Vietcuna family with three models: `vilm/vietcuna-3b`, `vilm/vietcuna-3bv2`, and `vilm/vietcuna-7b-v3`.
- The Hoa group with two models: `vlsp-2023-vllm/hoa-1b4` and `vlsp-2023-vllm/hoa-7b`.

Model	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST2	VNRTE	VSFC	VSMEC	VToC	WNLI
Multilingual Models												
mBERT	10.93	79.63	85.29/88.85	88.52	89.12/85.16	65.70	88.42	99.97	93.62	48.10	80.83	49.30
	10.09	79.35	83.09/87.43	89.11	88.86/84.74	64.98	88.42	99.97	93.05	51.02	81.43	**56.34**
	14.13	79.66	83.58/87.75	88.85	89.06/85.13	**70.76**	88.19	99.90	93.62	53.64	81.32	53.52
XLM-R_{base}	3.64	80.79	82.60/87.34	88.17	89.45/85.78	62.45	88.19	99.97	94.50	51.31	83.07	56.34
	0.00	81.61	83.82/88.26	87.90	89.44/85.80	60.29	88.30	99.90	94.50	53.06	82.96	46.48
	0.00	80.88	83.82/88.17	87.83	89.46/85.87	51.26	89.45	**100.00**	94.95	55.25	82.85	**56.34**
XLM-R_{large}	0.00	35.45	86.76/90.29	91.09	**90.48/87.22**	67.51	50.92	**100.00**	**95.39**	32.51	86.46	45.07
	0.00	35.45	68.38/81.22	**91.23**	90.10/86.82	47.29	89.11	**100.00**	95.14	37.90	**87.82**	43.66
	0.00	31.82	**88.24/91.64**	49.46	63.18/0.00	61.73	**90.14**	99.94	95.01	37.61	87.49	54.93
mDeBERTaV3	15.29	83.33	84.31/87.92	89.99	89.97/86.58	63.54	89.22	99.97	95.07	53.35	80.56	43.66
	19.62	**83.34**	85.05/89.32	89.84	89.98/86.69	69.31	89.79	99.81	95.01	53.64	80.88	56.34
	17.95	83.21	86.52/90.05	89.75	89.88/86.55	69.31	89.45	99.87	94.57	**55.39**	80.72	**56.34**
Vietnamese Models												
$PhoBERT_{base}$	16.28	82.73	81.86/87.15	89.49	89.87/86.34	62.82	**90.94**	99.97	95.26	56.56	86.02	56.34
	14.59	82.90	82.60/87.16	89.73	89.81/86.30	65.70	90.37	99.97	**95.51**	56.41	86.24	**54.93**
	17.03	82.81	81.37/86.52	89.71	89.84/86.27	65.70	90.71	99.97	95.20	56.27	86.07	56.34
$PhoBERT_{large}$	14.99	32.95	82.35/87.19	89.44	89.95/86.48	61.01	89.33	99.94	95.33	**60.79**	**88.86**	56.34
	0.00	**84.19**	**85.54/89.41**	89.97	**90.30/86.84**	66.79	89.79	**100.00**	95.01	60.35	88.04	**56.34**
	0.00	31.82	85.29/89.17	**90.92**	90.00/86.46	**59.93**	90.14	99.97	95.45	59.04	88.69	56.34
$PhoBERT_{base}V2$	**22.28**	83.40	83.09/87.70	90.13	90.15/86.82	70.40	90.60	99.97	**95.51**	59.18	84.43	53.52
	22.17	83.23	82.60/86.92	90.39	90.23/86.95	**71.12**	90.02	99.97	95.33	58.02	85.20	53.52
	13.11	83.39	83.33/87.86	90.08	90.01/86.62	66.79	89.68	**100.00**	95.33	57.00	85.53	53.52
$ViDeBERTa_{xsmall}$	0.00	68.79	71.81/81.72	78.46	83.26/77.18	51.26	76.49	99.55	87.11	30.32	19.39	**57.75**
	0.00	69.18	70.83/81.44	78.29	83.19/76.23	54.15	77.06	99.71	86.67	33.09	17.04	43.66
	0.00	69.06	73.53/82.97	78.60	83.06/76.59	54.87	76.83	99.49	86.10	31.63	16.49	46.48
$ViDeBERTa_{base}$	0.00	32.95	68.38/81.22	78.42	83.96/77.76	50.90	50.80	99.27	79.85	31.34	20.43	56.34
	0.00	44.72	68.38/81.22	70.11	83.86/77.58	51.62	67.89	99.55	81.17	32.07	15.73	56.34
	0.00	40.97	71.32/81.75	78.56	84.04/78.08	49.82	55.39	99.52	84.14	30.32	12.45	43.66

Table 8: Evaluation results of the multilingual baseline models on twelve tasks of ViGLUE. Additional results for the Vietnamese models are also measured. For all tasks, accuracy is reported except MRPC and QQP using accuracy/F1 score and CoLA using the Matthews correlation coefficient. The model is finetuned for each task in three sessions with three corresponding seed values, 1, 10, and 100. The highest performance of the model on each task across three seeds is underlined while the highest performance on each task across all models is shown in **bold**.

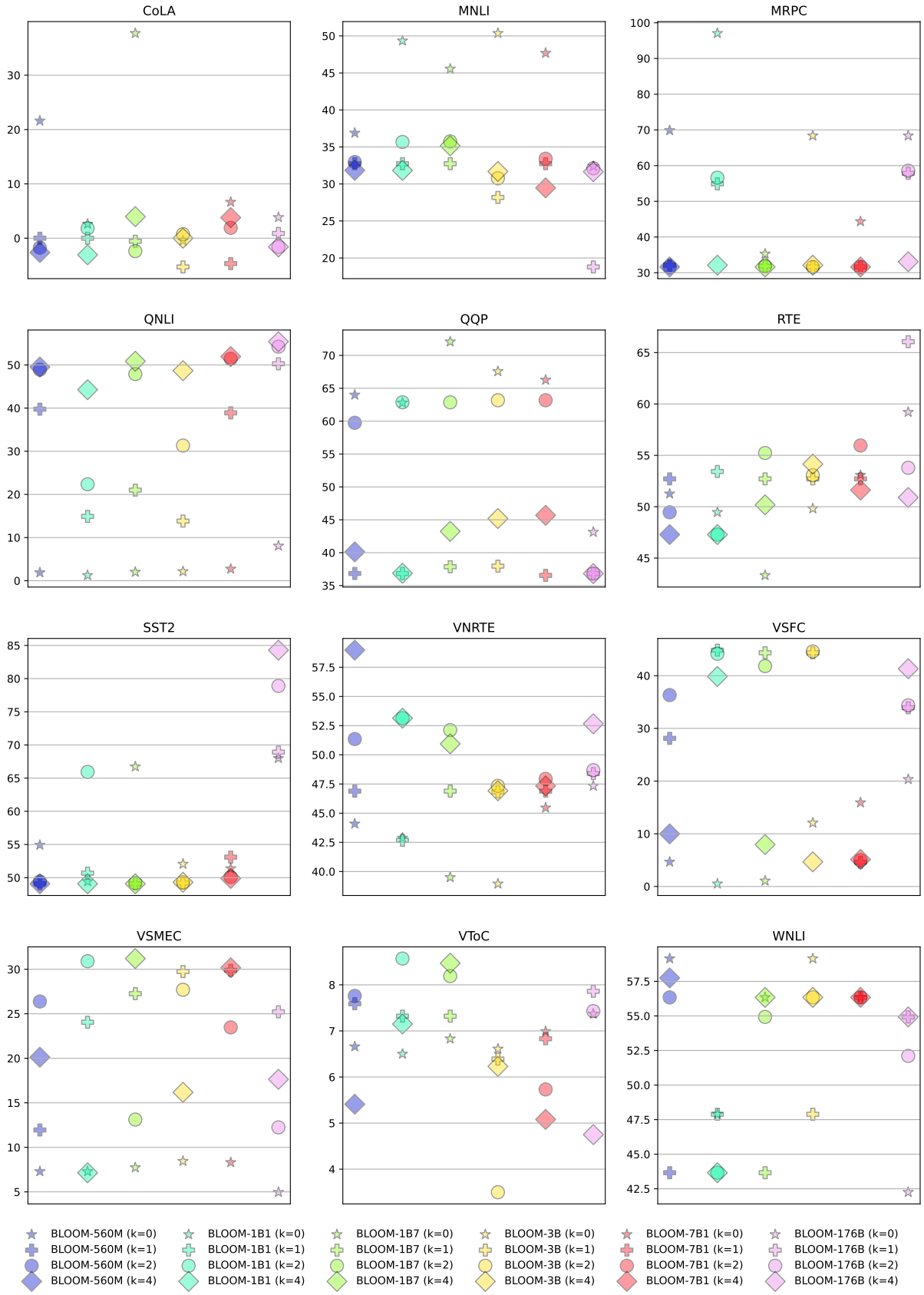

Figure 3: Few-shot learning performance of BLOOM model family with $k = 0, 1, 2, 4$. The star, the plus, the circle, and the diamond symbols represent few-shot learning results with $k = 0$, $k = 1$, $k = 2$, and $k = 4$ respectively. The metric used in the visualization is accuracy, except for CoLa, which reported the Matthew correlation coefficient.

Model		BLOOM 560M	BLOOM 1B1	BLOOM 1B7	BLOOM 3B	BLOOM 7B1	BLOOM 176B	Vietcuna-3B	Vietcuna-3B-v2	Vietcuna-7B-v3	Phogpt-7B5	Hoa 1B4	Hoa 7B
CoLa MCC	0	21.59	2.62	37.72	-0.44	6.65	3.84	0.29	2.71	1.39	**93.25**	31.02	0.05
	1	0.00	0.00	-0.56	-5.28	-4.67	0.88	-02.07	0.00	3.93	**04.63**	2.10	1.81
	2	-1.74	1.81	-2.38	0.74	1.90	-1.59	0.00	0.00	1.37	-2.37	-1.76	**3.71**
	4	-2.62	-3.05	**3.96**	0.00	3.78	-1.60	0.00	0.00	-3.80	0.00	-0.64	3.26
MNLI Acc.	0	36.89	49.35	45.55	50.36	47.68	32.38	23.82	22.89	32.74	**66.01**	37.72	46.96
	1	32.73	32.72	32.73	28.18	32.73	18.78	25.63	28.32	18.45	**49.77**	32.72	26.83
	2	32.96	35.66	35.75	30.76	33.40	32.12	25.18	27.36	17.91	**44.10**	33.02	34.88
	4	31.83	31.82	35.16	31.68	29.46	31.65	25.65	29.54	27.25	32.67	**38.93**	33.52
MRPC Acc./F1	0	69.85/71.72	97.05/97.86	35.29/10.20	68.38/70.61	44.36/32.23	68.38/81.11	24.01/6.62	24.01/1.27	28.92/0.00	98.03/98.57	48.77/40.45	**99.50/99.64**
	1	32.1/1.42	54.90/67.60	31.86/3.47	31.61/0.00	31.61/0.00	57.84/69.28	50.49/66.10	59.31/74.14	29.16/0.00	**68.62/81.34**	68.38/81.22	63.97/77.55
	2	31.61/0.00	56.61/65.22	31.86/1.41	31.61/0.00	31.61/0.00	58.57/70.50	45.58/60.63	50.98/65.63	27.45/01.33	49.01/54.18	**69.85/81.83**	68.38/81.22
	4	31.61/0.00	32.10/2.12	31.61/0.00	32.10/1.42	31.61/0.00	33.08/13.88	48.03/62.27	56.61/70.93	25.73/0.65	32.10/2.80	34.31/11.25	**66.66/79.64**
QNLI Acc.	0	1.88	1.24	2.01	2.14	2.72	8.07	45.30	**72.43**	62.73	1.00	1.59	1.73
	1	39.75	14.90	20.99	13.82	38.89	**50.26**	40.76	44.31	19.14	46.34	48.47	17.92
	2	48.91	22.35	47.90	31.31	51.41	**54.23**	46.98	49.64	50.26	47.66	48.36	18.35
	4	49.53	44.26	50.86	48.67	51.94	**55.39**	47.59	50.46	50.53	50.61	48.21	30.58
QQP Acc./F1	0	63.99/4.56	62.79/30.45	72.08/40.92	67.57/25.28	66.27/16.83	43.12/54.92	15.50/12.87	16.81/19.94	17.87/24.61	85.38/76.93	61.65/7.74	**86.52/79.88**
	1	36.84/53.82	36.82/53.82	37.86/53.96	37.94/53.56	36.54/52.14	36.82/53.75	32.09/47.97	35.80/52.67	29.78/45.16	**48.45/45.82**	44.08/43.16	37.04/53.75
	2	59.74/10.23	62.86/2.48	62.86/2.65	**63.18/0.79**	**63.18/0.00**	36.81/53.75	32.95/48.93	36.35/53.30	20.16/20.57	63.12/0.10	62.61/3.05	43.15/42.92
	4	40.12/51.52	36.87/53.81	43.23/46.26	45.19/54.32	45.68/37.79	36.81/53.82	34.92/51.35	36.67/53.65	21.11/27.75	59.02/4.14	**60.60/15.23**	36.87/53.72
RTE Acc.	0	51.26	49.45	43.32	49.81	53.06	59.20	**77.25**	72.20	75.09	7.94	51.98	43.68
	1	52.70	53.42	52.70	52.70	52.70	**66.06**	62.81	53.79	61.73	52.70	52.70	52.70
	2	49.45	47.29	55.23	53.06	55.95	53.79	**62.81**	58.84	**62.81**	53.06	54.87	55.23
	4	47.29	47.29	50.18	54.15	51.62	50.90	**69.31**	59.56	63.17	47.29	50.18	47.29
SST2 Acc.	0	54.93	49.42	66.74	52.06	51.37	68.00	**87.15**	86.35	69.15	0.00	48.96	50.00
	1	49.08	50.68	49.08	49.19	53.09	68.92	65.48	50.57	**77.75**	3.44	49.08	49.19
	2	49.42	65.94	49.19	49.31	50.11	**78.89**	72.82	72.82	77.86	35.77	50.34	50.91
	4	49.08	49.08	49.08	49.31	49.88	**84.28**	63.07	70.18	81.53	50.80	48.27	58.37
WNLI Acc.	0	59.15	47.88	56.33	59.15	56.33	42.25	42.25	43.66	46.47	**94.36**	57.74	60.56
	1	43.66	47.88	43.66	47.88	**56.33**	54.92	47.88	43.66	50.70	49.29	43.66	43.66
	2	**56.33**	43.66	54.92	**56.33**	**56.33**	52.11	45.07	45.07	49.29	43.66	52.11	45.07
	4	**57.74**	43.66	56.33	56.33	56.33	54.92	42.25	46.47	49.29	43.66	56.33	45.07
VNRTE Acc.	0	44.08	42.87	39.49	38.95	45.45	47.33	89.76	**97.44**	90.14	3.47	46.82	40.51
	1	46.89	42.68	46.89	46.82	46.89	48.39	**92.92**	87.56	90.85	45.68	46.89	47.24
	2	51.35	53.13	52.11	47.33	47.94	48.67	91.83	83.39	**92.12**	50.04	47.05	74.78
	4	58.97	53.13	50.94	46.92	47.33	52.66	89.06	83.51	**92.92**	52.98	47.21	56.64
VToC Acc.	0	6.66	6.49	6.82	6.60	6.99	**7.37**	6.22	6.17	4.58	7.20	6.99	6.49
	1	7.59	7.31	7.31	6.38	6.82	7.86	5.78	5.35	7.04	4.25	6.82	**10.15**
	2	7.75	8.57	8.19	3.49	5.73	7.42	6.17	5.95	6.55	6.88	7.48	**10.37**
	4	5.40	7.15	8.46	6.22	5.07	4.75	6.60	4.75	3.71	8.73	8.57	**9.61**
VSFC Acc.	0	4.67	0.50	1.07	12.06	15.91	20.34	40.99	37.71	**41.88**	0.00	2.33	4.67
	1	28.11	**44.85**	44.34	44.34	4.61	33.92	42.32	40.36	34.30	8.46	44.47	44.53
	2	36.32	44.15	41.81	**44.59**	4.99	34.36	43.20	39.67	33.10	0.56	44.53	44.53
	4	9.98	39.86	7.95	4.67	5.11	41.31	44.47	34.23	39.48	11.30	23.81	**48.26**
VSMEC Acc.	0	7.28	7.28	7.72	8.45	8.30	4.95	**11.07**	8.30	6.70	6.55	7.43	8.89
	1	11.95	24.05	27.25	29.73	**29.88**	25.21	19.53	20.11	12.09	19.09	18.07	22.44
	2	26.38	**30.90**	13.11	27.69	23.46	12.24	21.28	16.03	7.14	15.74	12.97	4.95
	4	20.11	7.14	**31.19**	16.18	30.17	17.63	5.68	11.51	10.34	8.30	12.39	6.12

Table 9: Large language models evaluation on all tasks. The highest F1 score is denoted with the underline while the highest accuracy and MCC, which is short for Matthews correlation coefficient, are marked in **bold**.