A Large-Scale Benchmark for Vietnamese Sentence Paraphrases

Sang Quang Nguyen[1, 2], **Kiet Van Nguyen**[1, 2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
sangnq.19@grad.uit.edu.vn
kietnv@uit.edu.vn

Abstract

This paper presents ViSP, a high-quality Vietnamese dataset for sentence paraphrasing, consisting of 1.2M original–paraphrase pairs collected from various domains. The dataset was constructed using a hybrid approach that combines automatic paraphrase generation with manual evaluation to ensure high quality. We conducted experiments using methods such as back-translation, EDA, and baseline models like BART and T5, as well as large language models (LLMs), including GPT-4o, Gemini-1.5, Aya, Qwen-2.5, and Meta-Llama-3.1 variants. To the best of our knowledge, this is the first large-scale study on Vietnamese paraphrasing. We hope that our dataset and findings will serve as a valuable foundation for future research and applications in Vietnamese paraphrase tasks. The dataset is available for research purposes at https://github.com/ngwgsang/ViSP.

1 Introduction

Sentences or phrases that express the same idea but use different words are called paraphrases (Bhagat and Hovy, 2013). Paraphrase helps create a richer amount of data, but still retains the main meaning of the sentence used.

Paraphrases generation is crucial for various tasks such as: In question answering (Bernhard and Gurevych, 2008; Dong et al., 2017; Gan and Ng, 2019), by generating paraphrases of the retrieved answers, QA systems can provide more comprehensive and nuanced responses; In information retrieval (Wallis, 1993; Zukerman et al., 2002), paraphrasing can help search engines find relevant documents even if the user's query doesn't match the exact wording of the documents; Machine translation (Callison-Burch et al., 2006; Russo-Lassner et al., 2005) , paraphrasing techniques can enhance translation accuracy by generating more natural and semantically equivalent translations and chat bot (Marceau et al., 2022), paraphrasing enables chat bot to respond more flexibly and naturally to user queries, adapting to variations in phrasing.

Although Vietnamese is widely spoken languages, Vietnamese is referred to as a low-resource language in NLP. Most previous work in paraphrase generation has focused mainly on English, such as MS-COCO (Lin et al., 2014), PAWS (Alzantot et al., 2018), QQP [1], ParaSCI (Dong et al., 2021). Although there are multilingual datasets such as TaPaCo (Scherrer, 2020), the number of Vietnamese sentence pairs is only 962, the number of sentence pairs is too small and because they are translated from English, the meaning will not be fluent. Some other works related to paraphrasing, such as ViQP (Nguyen et al., 2023b), have the limitation that their scope is only in questions, and questions in Vietnamese have a completely different structure than normal sentences.

In this paper, two our main contributions are described as follows:

1. **The creation of ViSP, the first large-scale dataset for Vietnamese sentence paraphrasing.** We developed a dataset containing over 1.2 million pairs of Vietnamese sentences across diverse topics. Each original sentence is accompanied by multiple paraphrases, all manually verified by a team of annotators to ensure high quality and accuracy.

2. **Comprehensive exploration of Vietnamese sentence paraphrasing.** We evaluated baseline models and compared their performance with traditional methods, such as rule-based approaches and back translation, as well as human performance. This analysis highlights the relative strengths and limitations of automated paraphrase generation for Vietnamese.

We hope ViSP together with our empirical study

[1]https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

Findings of the Association for Computational Linguistics: NAACL 2025, pages 1045–1060

can serve as a starting point for future Vietnamese paraphrase research and applications.

2 Dataset Creation

In this section, we introduce the process of constructing the ViSP dataset (see Figure 1), which includes Collecting, Preprocessing, Exampling, Generating and Validating.

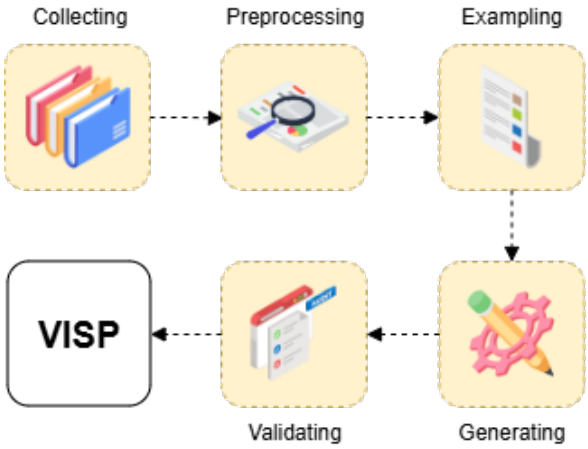


Figure 1: The overview process of creating our dataset ViSP.

2.1 Data Collection

We collect sentences from publicly available resources that contain original Vietnamese documents, including the UIT-ViQuAD (Nguyen et al., 2020b), UIT-ViNewsQA (Van Nguyen et al., 2022), ALQAC (Nguyen et al., 2023a) and ViNLI (Huynh et al., 2022) datasets. These datasets provide a diverse range of data sourced from Vietnamese news articles and Wikipedia, offering valuable material for sentence paraphrasing task, respectively.

After collecting data from the available datasets, we proceed to extract sentences from context segments of the above data sources using underthesea, a Vietnamese NLP toolkit[2].

2.2 Preprocessing

First, we manually filtered the sentences to remove those that were incorrect, unsuitable for Vietnamese language norms, or contained offensive language.

Next, we classified the sentences based on their topic using the Gemini (Team et al., 2023). The model categorized sentences into various labels, including health, society, lifestyle, science, culture, computer, law, sports, business and other. This step allowed us to organize the sentences by their subject matter, offering a comprehensive overview of different domains within the Vietnamese language context.

2.3 Exampling

To evaluate the generative performance of the Gemini model, we divided the team into two groups: the generation group $\{H_1, H_2\}$ and the evaluation group $\{H_3, H_4, H_5, H_6, H_7\}$. We randomly selected 350 sentences, consisting of 300 for testing and 50 for generate the Few-shot prompt, referred to as the few-shot corpus. The annotators in the evaluation group were tasked with manually generating paraphrases for the selected sentences, followed by cross-validation of the paraphrases among the evaluators. The generation group individually crafted paraphrases manually, providing a direct comparison against the AI group $\{G_1, G_2, G_3\}$. We split the dataset into 6 rounds $\{R1, R2, R3, R4, R5, R6\}$, each consisting of 50 sentences. The Few-shot prompts were randomly selected from 10 out of the 50 samples in the Few-shot corpus, which had been created by the evaluation group.

	R1	R2	R3	R4	R5	R6
G_1	**73.10**	68.30	**70.78**	69.56	**70.87**	68.18
G_2	69.51	68.74	68.98	69.45	70.44	65.55
G_3	70.28	66.93	68.45	69.50	69.16	65.09
H_1	72.11	66.79	70.64	**70.25**	70.56	67.13
H_2	71.08	**69.13**	70.75	68.30	69.55	**69.48**

Table 1: Compare Gemini with Few-shot examples performance and human performance across six rounds on the BLEU-4.

Table 1 demonstrate that Gemini significantly outperforms human efforts in paraphrase generation across multiple rounds. Specifically, the model achieved a win rate of 83.33% against H_1 and 66.67% against H_2. These results underscore the effectiveness of AI in replacing manual paraphrase generation, offering both cost savings and greater coverage.

2.4 Data Generation

We used the highest-performing prompt from section 2.3 to generate paraphrases from the cleaned and labeled dataset of original sentences from section 2.2. The paraphrase generation task can be

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

Input

s: Berlin trở thành địa điểm thành phố được viếng thăm nhiều thứ ba tại châu Âu. (*English: Berlin becomes the third most visited city in Europe.*)

k: 2

Output

p_1**:** Berlin là thành phố được du khách viếng thăm nhiều thứ ba tại châu Âu. (*English: Berlin is the third most visited city in Europe by tourists.*)
p_2**:** Xếp thứ ba về số lượng du khách viếng thăm tại châu Âu là thành phố Berlin. (*English:Ranked third in terms of number of visitors in Europe is the city of Berlin.*)

Table 2: Example of input and output of sentence paraphrase task.

formalized as follows. For each input, consisting of an original sentence s, the number k of paraphrases to be generated, and the chosen Few-shot prompt f, the model M generates a set of paraphrases using the formula 1:

$$M_f(s, k) = \{p_1, p_2, \ldots, p_k\} \quad (1)$$

In this setup, the task is to generate k paraphrases $\{p_1, p_2, \ldots, p_k\}$ that convey the original meaning while varying the structure and wording of the sentence s.

2.5 Data Validation

Automatic evaluation of the generation results from large language models (LLMs) can be easily achieved when ground truths from existing datasets are available (Zhu et al., 2023). However, open-ended data like paraphrasing or translation, human validation is necessary (Long et al., 2024). A straightforward idea is to provide some generated samples to human experts, who will then determine whether they are correct.

We established a review process involving seven annotators to ensure the quality of the paraphrased sentences generated by Gemini (Team et al., 2023). Each original-paraphrase sentence pair was evaluated by three annotators, corresponding to three votes. Annotators assessed each pair as True or False. A pair was considered valid if it received at least two True votes out of three. Sentence pairs were marked as False if their meaning was not preserved after paraphrasing or if they contained grammatical or spelling errors, based on a checklist (See Appendix C). A pair that received two or more False votes were removed from the dataset.

Original

SpaceX đang thử nghiệm các nguyên mẫu tàu tại cơ sở của họ ở nam Texas, tuy nhiên cả 4 phiên bản bay thử gần đây đều kết thúc bằng vụ nổ.
(*English: SpaceX is testing the prototypes of the spacecraft at their facility in southern Texas; however, all four recent test flights have ended in explosions.*)

Paraphrase

Các mẫu tàu đang được thử nghiệm tại cơ sở SpaceX ở phía nam Texas, nhưng cả 4 phiên bản thử nghiệm bay gần đây đã không thành công.
(*English: The spacecraft prototypes are being tested at SpaceX's facility in southern Texas, but all four recent test flights have failed.*)

Table 3: Example of an incorrect paraphrase pair violating the SEMANTIC EQUIVALENCE constraint.

Table 3 presents an example of SEMANTIC EQUIVALENCE error for the generated paraphrases. Sentences with errors were removed from the dataset to ensure high-quality standards. Across the entire dataset, the average error rate was 4.49% (See Appendix C).

2.6 Dataset Analysis

2.6.1 Overall Statistics

The statistics of the training, validation and test sets of the ViSP dataset are described in Table 4. In the table, we present number of original, the average number of paraphrases per original, the average lengths of original and paraphrased sentences, as well as the vocabulary sizes for both original and paraphrased sentences across all sets.

Statistics	Train	Val	Test
Sentence Pair†	406,308	391,044	380,590
Original	33,030	6,929	6,963
Avg. paraphrase per original	2.97	6.91	6.80
Avg. original length	21.90	21.47	21.53
Avg. paraphrase length	22.95	23.36	23.35
Original vocab	42,135	15,826	15,952
Paraphrase vocab	45,460	20,277	20,248

Table 4: Statistics of the training, validation, and test sets of the ViSP dataset. † denotes that total number of paraphrase pairs generated from all possible combinations.

2.6.2 Data Faithfulness and Diversity

We evaluate the dataset using BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to measure semantic similarity between generated paraphrases and original sentences by comparing n-grams. As

shown in Table 5, BLEU-4 scores for the Train, Val, and Test sets are 63.66, 67.24, and 66.83, while ROUGE-2 scores are 72.4, 73.32 and 72.99, indicating strong semantic alignment across all subsets.

To assess paraphrase diversity, we use DIST-1 and DIST-2(Li et al., 2016), which measure unique unigrams and bigrams, as well as Entropy-based metrics ENT-4, which capture the distributional richness of generated paraphrases, and Jaccard, which gauges lexical overlap. The DIST-1 scores are 94.94, 95.56, and 95.42, and the DIST-2 scores are 94.74, 95.01, and 94.96 for the Train, Validation, and Test sets, respectively. Additionally, the ENT-4 scores are 5.71, 6.52, and 6.51, while the Jaccard scores are 53.85, 51.61, and 51.39 across the same sets. The consistently high Distinct and ENT-4 values, accompanied by the lower Jaccard on the validation and test sets, suggest that the paraphrases exhibit a diverse lexical distribution, minimizing redundancy while maintaining coherence. The slightly higher diversity metrics in these sets also indicate that the paraphrases are more varied, improving evaluation robustness by ensuring broader linguistic diversity.

Type	Metric	Train	Val	Test
Semantic	BLEU-4	63.66	67.24	66.83
	ROUGE-2	72.40	73.32	72.99
Diversity	DIST-1	94.94	95.56	95.42
	DIST-2	94.74	95.01	94.96
	ENT-4	5.71	6.52	6.51
	Jaccard	53.85	51.61	51.39
Human Eval	INF	4.74	4.73	4.78
	REL	4.64	4.50	4.71
	FLU	4.86	4.83	4.80
	COH	4.86	4.90	4.89

Table 5: Evaluation of semantic faithfulness, diversity, and human evaluation metrics on the Train, Validation, and Test sets.

Additionally, we conduct a manual evaluation by human experts on 200 randomly selected samples from each of the train, validation, and test sets. Human evaluators assess the paraphrases using a 5-point scale across four key dimensions, based on (Grusky et al., 2018): INF, REL, FLU and COH (see Appendix B.2). Before the evaluation, we measured inter-annotator agreement using Fleiss' Kappa (Fleiss, 1971) for the task of rating paraphrase sentence pairs with five labels corresponding to the 5-point scale (ratings from 1 to 5). The Fleiss' Kappa values for four human metrics was 0.7252, 0.7144, 0.7634, and 0.7481, respectively. According to the interpretation guidelines by (Landis, 1977), these Kappa values indicate substantial agreement among the annotators. As shown in Table 5, the scores were rated quite well, ranging from 4.71 to 4.89.

2.6.3 Topic Based Analysis

In Table 6, *Health* and *Society* are the most common topics, making up about 33% and 19% of the total dataset, respectively. This disparity occurs because the dataset originates from UIT-ViNewsQA and UIT-ViQuAD (see Appendix A), which primarily focus on these two topics. The other topics are more evenly spread, each covering around 3% to 6% of the data.

Topic	Train	Val	Test	All
Health	11,381	2,443	2,367	16,193
Society	7,088	1,183	1,222	9,495
Culture	2,189	425	407	3,023
Computer	1,669	494	475	2,640
World	3,192	399	415	4,008
Sports	1,401	387	405	2,195
Science	2,001	593	616	3,212
Lifestyle	1,947	522	507	2,978
Law	1,804	327	377	2,510
Business	1,045	342	330	1,719

Table 6: Distribution of topics across the Train, Validation, Test, and All sets in the dataset, statistics based on the number of original sentences. For examples of each topic of sentence, see the Appendix Table 13.

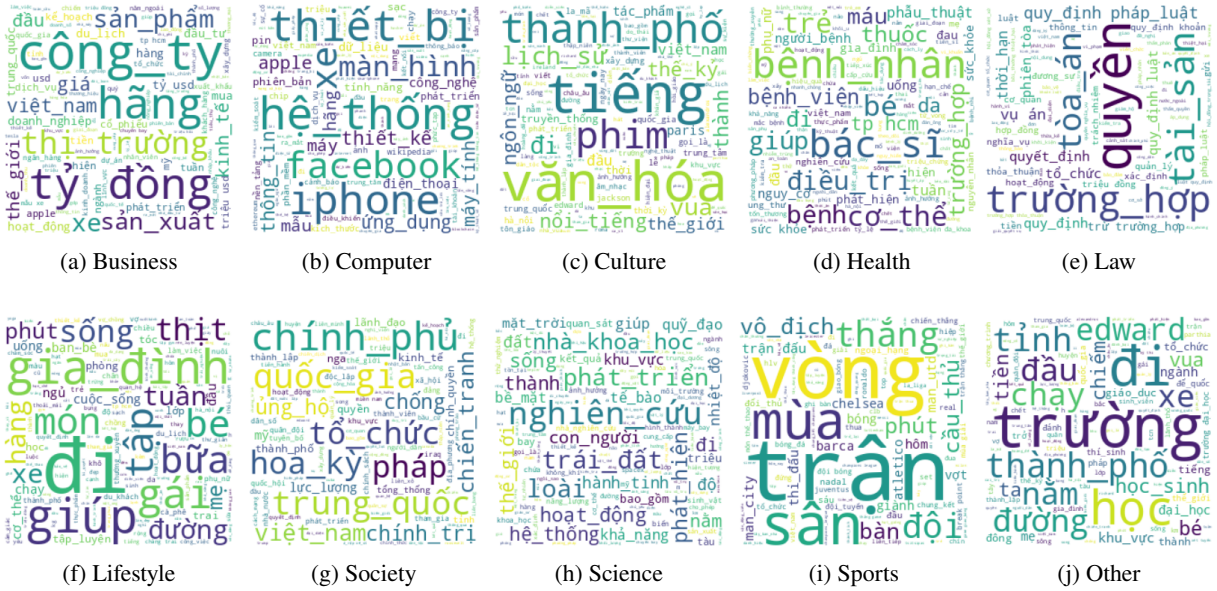
2.6.4 Length Based Analysis

Table 7 shows the combined distribution of sentence lengths across both the original and paraphrased sentences. The majority of sentences are between 11 and 20 words, accounting for approximately 43.05% of the dataset, which is the highest percentage among all length ranges. In contrast, sentences with more than 51 words represent the lowest percentage, comprising only about 0.65% of the dataset.

3 Experiments and Results

3.1 Human Performance

Following human performance concept of other study like (Nguyen et al., 2020b; Huynh et al.,



(a) Business (b) Computer (c) Culture (d) Health (e) Law

(f) Lifestyle (g) Society (h) Science (i) Sports (j) Other

Figure 2: Word clouds illustrating the most frequent words for each topic in the ViSP dataset.

Words	Train	Validation	Test	All
1-10	10,111	4,009	3,983	18,103
11-20	56,043	23,800	23,657	103,500
21-30	42,945	18,737	18,419	80,101
31-40	16,291	6,414	6,255	28,960
41-50	4,975	1,514	1,651	8,140
51+	826	355	372	1,553

Table 7: Combined distribution of sentence lengths across the Train, Validation, Test, and All sets, including both original and paraphrase sentences.

2022), we recruited five native Vietnamese speakers to perform the paraphrasing task. These individuals had no prior experience with paraphrasing tasks. Each annotator was asked to generate three paraphrases for a given set of sentences. Before starting, they were trained on the concept of paraphrasing and provided with guidelines to ensure that the paraphrases retained the original meaning while introducing lexical and structural variations.

Next, we randomly selected a subset of 300 samples, with 150 drawn from the test set and 150 from the validation set. This subset was designated as Test_{300} for further evaluation.

3.2 Re-Implemented Methods and Baselines

In this section, we re-implemented the following method and models on our dataset as described in Section 3.

EDA (Wei and Zou, 2019) applies simple transformations such as random deletion (RD), random swap (RS), random insertion (RI), and synonym replacement (SR). For RI and SR, we replace WordNet with the PhoW2V model (Nguyen et al., 2020a) to generate Vietnamese synonyms.

Back Translation leverages translation between languages to produce semantically similar sentences. We use the en2vi and vi2en models from (Nguyen et al., 2022) for this process.

We experiment with several pre-trained sequence-to-sequence models for paraphrase generation, including **mBART** (Tang et al., 2020), **BARTpho** (Tran et al., 2021), **mT5** (Xue, 2020), and **ViT5** (Phan et al., 2022). These models were chosen for their strengths in both multilingual and Vietnamese-specific tasks. **mBART** and **mT5** provide robust multilingual capabilities, while **BARTpho** and **ViT5** are optimized for Vietnamese, offering language-specific nuances.

3.3 Evaluation Metrics

We use BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019) to evaluate paraphrase quality, and Distinct-N (Li et al., 2016), Entropy-N (Shannon, 1948), and Jaccard (Jaccard, 1901) to measure diversity. For a detailed breakdown of evaluation metrics, see Appendix B.

3.4 Experimental Settings

We use a single NVIDIA Tesla A100 GPU via Google Colaboratory[3] to fine-tune all models on

[3]https://colab.research.google.com/

our dataset. When fine-tuning, we set the length of max length of sentence is 96 tokens, learning rate is 1e-5, batch size is 16 and training with five epochs.

3.5 Experimental Results

3.5.1 Single Paraphrase Evaluation

In the realm of single paraphrase generation, Table 8 shows that BARTpho-word$_{large}$ leads in BLEU-4 and ROUGE-2, with values of 72.06 and 76.06 on the Val set, respectively, indicating that the generated sentences are more similar to the reference paraphrases. It maintains strong performance on Test and Test$_{300}$, achieving BLEU-4 of 71.61 and 71.70, and ROUGE-2 of 75.78 and 76.22. While mBART$_{large}$ performs well, it achieves higher BERTScore across all sets, with 85.84 on Val and 86.17 on Test$_{300}$, suggesting that although the generated sentences differ more from the references, they retain better semantic similarity. ViT5-base also shows strong semantic preservation, with BERTScore reaching 85.37 on Test$_{300}$. Among monolingual models, BARTpho-syllable$_{large}$ performs well, with BERTScore of 85.62 on Test$_{300}$, reinforcing its effectiveness in generating faithful paraphrases. Human performance remains the upper bound, with BERTScore of 88.30, highlighting the gap between model-generated and human paraphrases. Among augmentation methods, Back Translation performs best, achieving a BERTScore of 79.99 on Test$_{300}$, while simpler methods like Random Deletion and Synonym Replacement show notably lower scores.

3.5.2 Multiple Paraphrases Evaluation

For multiple paraphrase generation, as shown in Table 9, BARTpho-word$_{large}$ leads the performance across all K values, with the highest BLEU-4 and ROUGE-2 scores, indicating that its generated paraphrases closely resemble the reference sentences. mBART$_{large}$, while slightly behind in BLEU-4 and ROUGE-2, achieves higher BERTScore for K = 3 and K = 5, suggesting strong semantic similarity. However, its performance declines at K = 10, showing reduced accuracy when generating more paraphrases. Among monolingual models with base architecture, ViT5$_{base}$ performs well at K = 10, achieving a BLEU-4 score of 66.92 and a BERTScore of 83.68. ViT5$_{base}$ also outperforms ViT5$_{large}$, showing greater stability and less degradation in accuracy with increasing paraphrase numbers.

3.5.3 Topic Based Evaluation

As shown in Figure 3, the results indicate that BARTpho-word$_{large}$ and mBART$_{large}$ consistently perform best across all topics. Within the T5 family, ViT5-base still outperforms other T5 models but also consistently surpasses all base BART models, including BARTpho-syllable$_{base}$ and BARTpho-word$_{base}$. This is evident across multiple categories, especially in *culture* and *sports*. However, the challenge appears more pronounced when evaluated by BLEU, whereas BERTScore remains relatively similar across topics. Additionally, *lifestyle*, *culture*, and *sports* emerge as the most challenging domains for all models, with the highest BLEU-4 scores in these categories hovering only around 70.

3.5.4 Length Based Evaluation

As shown in Figure 4, BARTpho-word$_{large}$ consistently outperforms other models in BLEU across all sentence lengths, achieving the highest scores in the 41–50 word range. Meanwhile, BARTpho-syllable$_{large}$ tends to yield stronger BERTScore values, highlighting the overall effectiveness of BARTpho-based models. mBART$_{large}$ follows closely in most cases; for instance, in the 31–40 word range, BARTpho-syllable$_{base}$ achieves 69.38 BLEU-4 and 85.32 BERTScore, slightly below mBART. In contrast, T5-based models (ViT5, mT5) show weaker performance, particularly in shorter sentences (1–20 words). Notably, mT5$_{large}$ scores as low as 18.56 BLEU-4 and 70.96 BERTScore in the 11–20 word range, significantly below the BART-based models, which consistently perform well across all sentence lengths.

3.5.5 Diversity Based Evaluation

Table 10 reports the diversity DIST-1, DIST-2 and entropy ENT-4 metrics, as well as the Jaccard scores, of the five beam-searched paraphrases generated by various models on the Test and Test$_{300}$ sets. Overall, BARTpho-word models exhibit strong performance, particularly BARTpho-word$_{large}$, which achieves competitive bigram diversity with a DIST-2 score of 94.70 on the Test set and 94.63 on Test$_{300}$, along with a high ENT-4 score of 5.28 on the Test set. Notably, BARTpho-syllable$_{large}$ attains the highest DIST-1 score of 95.75 on the Test set, while mBART$_{large}$ reaches the highest DIST-1 score of 95.72 on Test$_{300}$. Meanwhile, ViT5$_{large}$ leads in ENT-4 for the Test set with a score of 5.29. Additionally, although mT5$_{base}$

Model	Val			Test			$Test_{300}$		
	BLEU-4	ROUGE-2	BERTScore	BLEU-4	ROUGE-2	BERTScore	BLEU-4	ROUGE-2	BERTScore
RD	26.82	52.49	66.19	26.87	52.42	66.10	26.23	52.93	66.27
RS	14.79	48.89	63.75	14.76	48.70	63.76	14.69	48.86	64.12
RI + PhoW2V	29.50	57.79	69.79	29.43	57.63	69.69	29.86	58.49	70.05
SR + PhoW2V	22.81	46.48	63.30	22.67	46.46	63.28	22.86	46.14	63.10
BT + vinai-translate-v2	54.33	63.84	79.23	53.97	63.65	21.60	54.03	64.33	79.99
$mBART_{large}$	71.71	76.02	**85.84**	71.12	75.76	**85.74**	**72.23**	76.20	**86.17**
$mT5_{base}$	60.22	70.20	81.46	59.58	69.69	81.27	60.84	71.00	81.85
$mT5_{large}$	27.04	46.59	72.02	26.86	46.23	71.84	26.83	46.48	72.38
$BARTpho\text{-}syllable_{base}$	68.39	73.97	84.34	67.66	73.51	84.15	70.39	75.03	85.06
$BARTpho\text{-}syllable_{large}$	70.29	75.35	85.22	69.81	74.89	85.10	70.83	75.75	85.62
$BARTpho\text{-}word_{base}$	69.61	74.63	79.25	68.76	74.18	79.15	70.23	75.47	79.72
$BARTpho\text{-}word_{large}$	**72.06**	**76.06**	79.97	**71.61**	**75.78**	79.99	71.70	**76.22**	80.10
$ViT5_{base}$	70.20	74.91	85.08	69.75	74.58	85.00	71.24	75.69	85.37
$ViT5_{large}$	67.10	71.83	82.68	66.70	71.53	82.53	67.73	72.92	82.98
Human performance	-	-	-	-	-	-	94.97	88.29	88.30

Table 8: Evaluation of various models and methods on the Val, Test, and $Test_{300}$ sets of the ViSP dataset, assessing the best single paraphrased sentence generated by each model. The best overall results are highlighted in **bold**.

Model	K = 3			K = 5			K = 10		
	BLEU-4	ROUGE-2	BERTScore	BLEU-4	ROUGE-2	BERTScore	BLEU-4	ROUGE-2	BERTScore
$mBART_{large}$	70.35	75.23	**85.29**	69.76	74.84	**85.06**	66.16	71.50	81.57
$mT5_{base}$	58.68	68.85	80.32	57.78	68.27	79.88	54.68	66.36	78.36
$mT5_{large}$	26.98	46.39	71.51	27.07	46.48	71.46	27.29	46.70	71.09
$BARTpho\text{-}syllable_{base}$	66.81	72.99	83.63	66.20	72.56	83.37	62.26	68.94	79.54
$BARTpho\text{-}syllable_{large}$	69.02	74.43	84.65	68.48	74.07	84.44	67.06	73.20	**83.80**
$BARTpho\text{-}word_{base}$	68.03	73.70	78.50	67.42	73.31	78.34	64.09	70.21	75.54
$BARTpho\text{-}word_{large}$	**70.88**	**75.28**	79.39	**70.31**	**74.91**	79.23	**69.10**	**74.09**	78.88
$ViT5_{base}$	68.93	74.08	84.52	68.35	73.73	84.34	66.92	72.81	83.68
$ViT5_{large}$	66.19	71.24	81.93	65.63	70.89	81.68	63.98	69.87	80.85

Table 9: Evaluation of various models on the ViSP dataset's test set, focusing on K paraphrased sentences generated per input sentence (where K is the number of paraphrases). The BLEU-4, ROUGE-2 and BERTScore are averaged across all K paraphrases for each model. The best overall results are highlighted in **bold**. For a detailed breakdown of multiple paraphrase outputs, See Appendix D, Table 15.

achieves high Jaccard scores of 82.72 on the Test set and 87.98 on $Test_{300}$, $BARTpho\text{-}word_{large}$ obtains the lowest Jaccard scores of 57.99 (Test) and 58.26 ($Test_{300}$), signifying the greatest lexical diversity in its paraphrasing. These findings suggest that BARTpho-word excels at generating lexically varied and distributionally rich paraphrases, making it a robust option for applications requiring both diversity and consistency in paraphrasing.

4 Discussion

Multilingual vs. Monolingual. As (Conneau, 2019) highlight, while multilingual models offer flexibility across languages, monolingual models often excel in specialized tasks due to their focus on a single language's nuances. This distinction becomes apparent in our test results, where $BARTpho\text{-}word_{large}$ demonstrates a clear advantage in both single and multi-paraphrase generation. Initially, $BARTpho\text{-}word_{large}$ holds a clear edge in single paraphrase generation, achieving the highest BLEU-4 and ROUGE-2 scores. Although $mBART_{large}$ achieves higher BERTScore in single paraphrase generation, it experiences a notable decline in performance as the number of required paraphrases increases (K=5, K=10), with BLEU-4 dropping to 66.16. This performance drop aligns with findings from (Hu et al., 2020), which show that multilingual models struggle with generating numerous high-quality paraphrases. In contrast, monolingual models like BARTpho and ViT5 maintain strong performance in both single and multiple paraphrase tasks. Their focused training makes them better suited for tasks requiring high

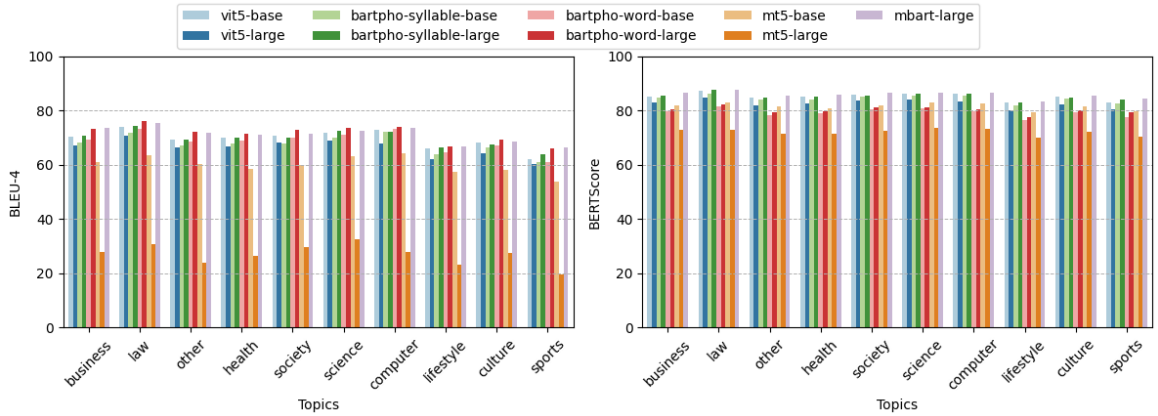


Figure 3: BLEU-4 and BERTScore across different topics.

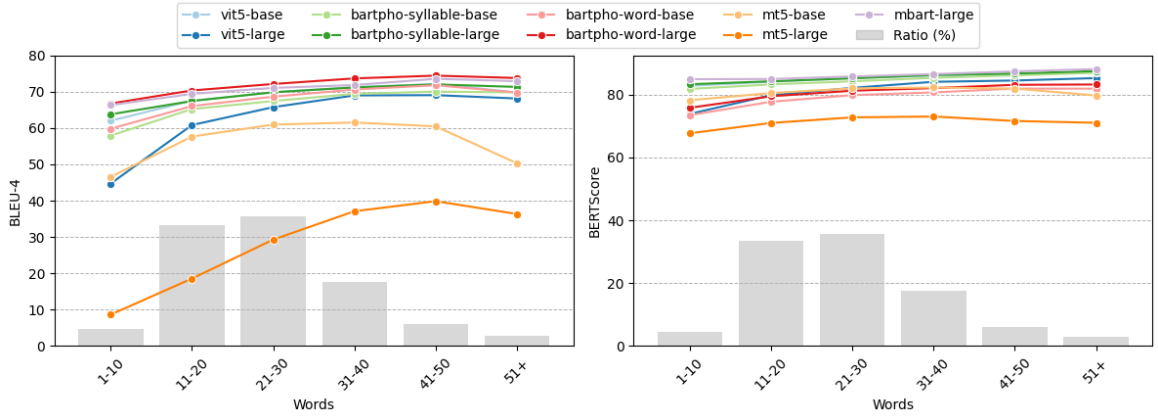


Figure 4: BLEU-4 and BERTScore across different lengths.

paraphrase diversity, consistently producing multiple outputs without losing quality.

Impact of Model Architecture. Research suggests that larger architectures do not always ensure better accuracy. For example, ViT5$_{\text{large}}$, mT5$_{\text{large}}$ does not surpass ViT5$_{\text{base}}$, mT5$_{\text{base}}$ accross all metrics, reflecting findings by (Kaplan et al., 2020) that size increases do not guarantee performance gains. Particularly when generating less common words, larger models may perform worse in paraphrasing due to vocabulary deviations (Brown, 2020). In contrast, BART-based models, such as BARTpho-word and BARTpho-syllable, consistently show improvements in both accuracy and diversity with increased model size, as shown by (Lewis, 2019), affirming the benefits of larger architectures in generating diverse paraphrases.

Impact of Sentence Structure. The structure of a sentence significantly influences the performance of paraphrase models. Different sentence types in Vietnamese—such as simple, compound, complex, and special—present varying levels of difficulty. As shown in Table 11, monolingual models like BARTpho-word$_{\text{large}}$ consistently achieve higher BLEU-4 scores for simple, compound, and complex sentences. This aligns with findings from (Isabelle et al., 2017), which suggest that models trained on a single language excel in capturing syntactic nuances. However, models often struggle with compound and complex sentences, which frequently include metaphorical and metonymic (See Appendix A, Table 13) expressions in Vietnamese, as noted by (Shutova et al., 2013), highlighting the challenge of paraphrasing non-standard and figurative structures. These results suggest the need for improved handling of complex syntactic and figurative forms in paraphrasing tasks.

LLM Performance. Table 12 shows that Vietnamese-specific LLMs, such as Vistral-7B-Chat (Van Nguyen et al., 2023), lag behind significantly, indicating weaker paraphrase generation capabilities compared to general-purpose models. This suggests that current Vietnamese-focused models may require further optimization

Model	Val				Test				Test_{300}			
	DIST-1	DIST-2	ENT-4	Jaccard	DIST-1	DIST-2	ENT-4	Jaccard	DIST-1	DIST-2	ENT-4	Jaccard
$\text{mBART}_{\text{large}}$	95.67	94.61	5.23	69.05	95.62	94.59	5.22	68.82	95.72	94.60	4.34	70.31
mT5_{base}	95.11	93.33	5.01	82.99	94.91	93.27	5.01	82.72	95.37	93.47	4.15	87.98
$\text{mT5}_{\text{large}}$	56.30	58.94	5.02	72.25	56.23	58.93	5.02	72.11	54.62	56.77	**4.51**	72.75
$\text{BARTpho-syllable}_{\text{base}}$	95.75	94.32	5.12	76.99	95.62	94.30	5.12	76.71	**96.17**	94.26	4.25	78.91
$\text{BARTpho-syllable}_{\text{large}}$	**95.90**	94.53	5.21	66.38	**95.75**	94.48	5.20	66.35	96.02	94.41	4.27	66.83
$\text{BARTpho-word}_{\text{base}}$	93.95	94.60	5.21	62.13	93.80	94.51	5.21	62.10	94.09	94.36	4.41	63.75
$\text{BARTpho-word}_{\text{large}}$	93.85	**94.77**	5.27	**58.00**	93.72	**94.70**	5.28	**57.99**	93.82	**94.63**	4.47	**58.26**
$\text{ViT5}_{\text{base}}$	95.33	94.15	5.14	75.70	95.14	94.05	5.14	75.49	95.48	94.01	4.27	77.36
$\text{ViT5}_{\text{large}}$	93.99	93.59	**5.30**	68.78	93.95	93.63	**5.29**	68.53	94.75	94.04	4.39	69.89
Human performance	-	-	-	-	-	-	-	-	95.54	94.96	6.48	52.33

Table 10: Evaluation of the 5 beam-searched paraphrases in terms of DIST-1, DIST-2, ENT-4, and Jaccard on the Val, Test, and Test_{300} sets. The **best** overall results are in bold, and the second best are underlined.

Model	simple	compound	complex
$\text{mBART}_{\text{large}}$	73.06	67.63	69.73
mT5_{base}	61.92	54.77	57.96
$\text{mT5}_{\text{large}}$	24.80	27.23	31.14
$\text{BARTpho-syllable}_{\text{base}}$	69.58	64.24	66.27
$\text{BARTpho-syllable}_{\text{large}}$	71.42	67.07	68.56
$\text{BARTpho-word}_{\text{base}}$	70.44	65.45	67.69
$\text{BARTpho-word}_{\text{large}}$	**73.40**	**68.66**	**70.27**
$\text{ViT5}_{\text{base}}$	71.60	66.56	68.39
$\text{ViT5}_{\text{large}}$	67.05	63.70	66.58

Table 11: BLEU-4 scores of various models on different sentence structures in the ViSP Test set. The best overall results are highlighted in **bold**. For a detailed breakdown of sentence structs, see Appendix A, Table 13.

Model	BLEU-4	ROUGE-2	BERTScore
GPT-4o Mini	52.73	65.55	81.82
Gemini 1.5 Flask	50.98	63.02	79.61
Vistral-7B-Chat	29.16	49.46	70.71
Aya-23-8B	42.15	59.52	75.21
Qwen2.5-7B	54.38	65.71	80.72
Meta-Llama-3.1-8B	60.32	69.34	82.40
Meta-Llama-3.1-70B	**65.51**	**73.21**	**84.27**
Human Performance	94.97	88.29	88.30

Table 12: Evaluation of various LLMs on the Test_{300} using BLEU-4, ROUGE-2, and BERTScore metrics. The best overall results are highlighted in **bold**.

or fine-tuning for paraphrase tasks. Among multilingual and general-purpose LLMs, Meta-Llama-3.1-70B (Dubey et al., 2024) achieves the best results, followed by Meta-Llama-3.1-8B (Dubey et al., 2024), GPT-4o (Hurst et al., 2024), and Qwen2.5-7B (Yang et al., 2024), which all demonstrate strong lexical and semantic alignment with human paraphrases. However, none of the models reach human performance, indicating room for improvement in semantic fidelity and lexical variation. Since all models are evaluated without fine-tuning, their performance is reasonable, especially for high-resource models like Meta-Llama-3.1 and GPT-4o. However, fine-tuning on Vietnamese-specific paraphrase datasets could further narrow the gap between AI-generated and human paraphrases.

5 Conclusion and Future Work

We introduced ViSP, a Vietnamese paraphrase dataset created using human annotations and LLM outputs, for evaluating and benchmarking paraphrase generation models. We tested models like mBART, BARTpho, ViT5, and mT5 across various sentence lengths and topics, highlighting the strengths and weaknesses of multilingual and monolingual approaches. Our evaluation covered accuracy (BLEU, ROUGE, BERTScore) and diversity (Distinct-N, Entropy-N, Jaccard), and we compared model-generated paraphrases with human performance to assess the gap between automated systems and human paraphrasing.

In future, we plan to extend ViSP to tasks like machine translation, question answering and retrieval augmented generation. Additionally, we aim to pretrain a Vietnamese paraphrasing model, addressing a key gap in domain-specific models. This model will target complex linguistic phenomena, including metaphor and metonymy in Vietnamese, which present significant challenges for natural language understanding and generation tasks. ViSP will also support developing robust sentence similarity models like SBERT (Reimers, 2019), advancing Vietnamese NLP research.

Limitations

While our models were fine-tuned on the ViSP dataset, they were trained under low-resource conditions, which means the overall performance may not be fully optimized. With more computational resources, further improvements could be achieved. During data creation, we employed the Few-shot method to guide the generation process. However, we have not yet compared this approach with other advanced techniques like Chain-of-Thought (Wei et al., 2022) or Tree-of-Thought (Yao et al., 2024), which could potentially yield better results in generating higher-quality paraphrases. Additionally, the current dataset lacks representation from certain specialized domains such as metaphor, mathematics and programing. This absence may affect the models ability to generalize to these specific areas.

Ethics Statement

The ViSP dataset was developed with adherence to ethical guidelines. Human annotators were informed and compensated fairly. All datasets used, including UIT-ViQuAD (Nguyen et al., 2020b), UIT-ViNewsQA (Van Nguyen et al., 2022), ALQAC (Nguyen et al., 2023a) and ViNLI (Huynh et al., 2022), were utilized in compliance with their respective licenses and terms of use. Additionally, in generating paraphrases with large language models (LLMs), we took steps to review and mitigate potential errors in the outputs, ensuring fairness and representativeness across different domains.

Acknowledgement

We sincerely appreciate the insightful comments and constructive feedback provided by the anonymous reviewers. This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under the grant number DS2025-26-01.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples.

Delphine Bernhard and Iryna Gurevych. 2008. Answering learners' questions by retrieving question paraphrases from social q&a sites.

Rahul Bhagat and Eduard Hovy. 2013. What Is a Paraphrase? *Computational Linguistics*, 39(3):463–472.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering.

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. Parasci: A large scientific paraphrase dataset for longer paraphrase generation.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, Florence, Italy. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022. ViNLI: A Vietnamese corpus for studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3858–3872, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. *arXiv preprint arXiv:1704.07431*.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

JR Landis. 1977. The measurement of observer agreement for categorical data. *Biometrics*.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey.

Louis Marceau, Raouf Belbahar, Marc Queudot, Nada Naji, Eric Charton, and Marie-Jean Meurs. 2022. Quick starting dialog systems with paraphrase generation.

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

Chau Nguyen, Son T Luu, Thanh Tran, An Trieu, Anh Dang, Dat Nguyen, Hiep Nguyen, Tin Pham, Trang Pham, Thien-Trung Vo, et al. 2023a. A summary of the alqac 2023 competition. In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–6. IEEE.

Kiet Van Nguyen, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020b. A vietnamese dataset for evaluating machine reading comprehension.

Sang Quang Nguyen, Thuc Dinh Vo, Duc PA Nguyen, Dang T Tran, and Kiet Van Nguyen. 2023b. Viqp: Dataset for vietnamese question paraphrasing. In *2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6. IEEE.

Thien Hai Nguyen, Tuan-Duy H. Nguyen, Duy Phung, Duy Tran-Cong Nguyen, Hieu Minh Tran, Manh Luong, Tin Duy Vo, Hung Hai Bui, Dinh Phung, and Dat Quoc Nguyen. 2022. A Vietnamese-English Neural Machine Translation System. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association: Show and Tell (INTERSPEECH)*.

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.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H Trinh. 2022. Vit5: Pretrained text-to-text transformer for vietnamese language generation. *arXiv preprint arXiv:2205.06457*.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Grazia Russo-Lassner, Jimmy Lin, and Philip Resnik. 2005. A paraphrase-based approach to machine translation evaluation. *LAMP-TR-125 CS-TR-4754 UMIACS-TR-2005-57, University of Maryland, College Park, MD*.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2021. Bartpho: pre-trained sequence-to-sequence models for vietnamese. *arXiv preprint arXiv:2109.09701*.

Chien Van Nguyen, Thuat Nguyen, Quan Nguyen, Huy Nguyen, Björn Plüster, Nam Pham, Huu Nguyen, Patrick Schramowski, and Thien Nguyen. 2023. Vistral-7b-chat-towards a state-of-the-art large language model for vietnamese.

Kiet Van Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2022. New vietnamese corpus for machine reading comprehension of health news articles. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

Peter Wallis. 1993. Information retrieval based on paraphrase. In *Proceedings of pacling conference*. Citeseer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks.

Ingrid Zukerman, Bhavani Raskutti, and Yingying Wen. 2002. Experiments in query paraphrasing for information retrieval. In *Australian Joint Conference on Artificial Intelligence*, pages 24–35. Springer.

Appendix

In this section, we provide supplementary information to support the main content of this paper. This includes additional details about the datasets, models, evaluation metrics, and methods used throughout our experiments.

A Dataset Details

The ViSP dataset is compiled from several publicly available sources, including:

1. **UIT-ViQuAD** (Nguyen et al., 2020b) This machine reading comprehension dataset includes over 23,000 human-generated question-answer pairs. These pairs are derived from 5,109 passages extracted from 174 Vietnamese Wikipedia articles, providing a rich source of information and ensuring a diverse range of topics and contexts.

2. **UIT-ViNewsQA** (Van Nguyen et al., 2022) This corpus consists of 22,057 question-answer pairs created by crowd-workers. These pairs are based on a collection of 4,416 Vietnamese healthcare news articles, with answers being textual spans directly taken from the corresponding articles.

3. **ALQAC** (Nguyen et al., 2023a): The ALQAC dataset contains thousands of multiple-choice question-answer pairs, sourced from Vietnamese legal documents. Each pair is carefully reviewed for clarity and accuracy, making it an essential resource for testing question answering models in the legal domain.

4. **ViNLI** (Huynh et al., 2022) The ViNLI corpus comprises over 30,000 human-annotated premise-hypothesis sentence pairs. These pairs are extracted from more than 800 online news articles, offering a substantial and varied dataset for natural language inference tasks.

Figure 5 shows the distribution of original sentence sources in the ViSP dataset, with UIT-ViNewsQA and ViNLI contributing the largest proportion. This suggests that ViSP is heavily influenced by news-related content, which may impact the linguistic patterns and domain coverage of the paraphrases.

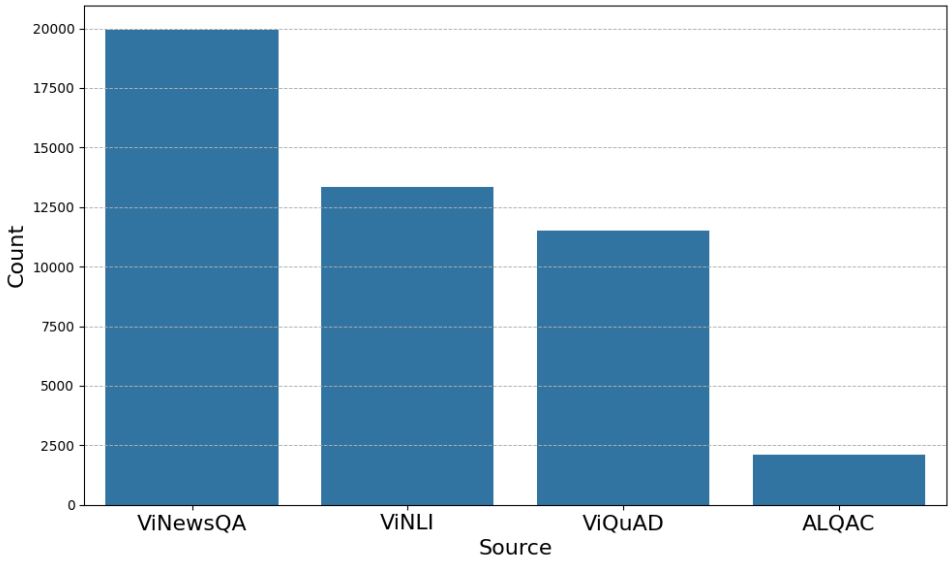


Figure 5: Distribution of sentence source in the ViSP dataset.

Sentence	Source	Topic	Structure
Vào ngày 23 tháng 9 năm 1846, nhà thiên văn Johann Galle đã phát hiện ra Sao Hải Vương ở vị trí lệch 1 độ so với tiên đoán của Urbain Le Verrier. (*English: On September 23, 1846, astronomer Johann Galle discovered Neptune 1 degree off from Urbain Le Verrier's prediction.*)	ViQuAD	Science	Complex
Theo giáo sư Long, mỗi nước có khuyến cáo khác nhau khi điều trị vi khuẩn HP. (*English: According to Professor Long, each country has different recommendations when treating HP bacteria.*)	ViNewsQA	Health	Simple
Hazard gia nhập Real hè 2019 từ Chelsea, theo bản hợp đồng trị giá 190 triệu USD - trong đó có 112 triệu USD trả trước. (*English: Hazard joined Real in the summer of 2019 from Chelsea, in a contract worth 190 million USD - including 112 million USD in advance.*) **#metonym**	ViNLI	Sports	Compound
Công ty đã hứng rất nhiều cuộc tấn công mạng, phải từ bỏ nhiều dịch vụ chủ chốt trước khi 'bán mình' cho đại gia viễn thông Mỹ Verizon. (*English: The company faced numerous cyberattacks and had to abandon several key services before 'selling itself' to the American telecom giant Verizon.*) **#metonym #metaphor**	ViNLI	Business	Compound
Trong khi đó, Điện Kremlin tuyên bố đang nghiên cứu khả năng tổ chức hội nghị này. (*English: Meanwhile, the Kremlin announced that it is studying the possibility of holding this conference.*) **#metonym**	ViQuAD	Society	Single
Những người có làn da ngăm đen tạo cảm giác khỏe mạnh, gợi cảm cho người đối diện. (*English: People with dark skin give a feeling of health and sexiness to the other person.*)	ViNewsQA	Lifestyle	Simple
Đơn khiếu nại phải kèm theo bản sao quyết định giải quyết khiếu nại lần đầu và các tài liệu kèm theo. (*English: The complaint must be accompanied by a copy of the initial complaint resolution decision and accompanying documents.*)	ALQAC	Law	Complex
Ellen DeGeneres, sinh năm 1958, là ngôi sao truyền hình hàng đầu tại Mỹ. (*English: Ellen DeGeneres, born in 1958, is a top television star in America.*)	ViNLI	Culture	Simple
Video đầu tiên hiện có hơn 2,1 triệu lượt xem chỉ sau một ngày đăng tải. (*English: The first video now has more than 2.1 million views after just one day of posting.*)	ViNLI	Other	Simple

Table 13: Examples of classifying sentences by topic in ViSP dataset.

B Metrics

B.1 Sematic and Diversity

1. **BLEU-4** (Papineni et al., 2002): This metric measures the precision of 4-grams between the generated paraphrase and the reference. A higher BLEU-4 value indicates greater syntactic and lexical alignment with the reference.

2. **ROUGE-2** (Lin, 2004): This metric calculates the recall of bigrams (2-grams) in the generated paraphrase compared to the reference. A higher ROUGE-2 value reflects better preservation of key content from the reference.

3. **BERTScore** (Zhang et al., 2019): This measure uses contextual embeddings to compare each token in the generated paraphrase with those in the reference. A higher BERTScore implies stronger semantic similarity and fidelity to the reference text.

4. **DIST-1 and DIST-2** (Li et al., 2016): These metrics capture the distinctiveness of unigrams and bigrams, respectively. Higher values indicate more diverse and less repetitive paraphrases.

5. **ENT-4** (Shannon, 1948): This is the entropy of 4-grams, reflecting the diversity and unpredictability of word combinations. A higher ENT-4 value suggests more varied and creative paraphrases.

6. **Jaccard** (Jaccard, 1901): This score measures the lexical overlap between the original sentence and its paraphrase. A lower Jaccard value indicates less overlap with the source, and hence greater

paraphrase diversity.

B.2 Human Eval

We conduct manual evaluations where human reviewers assess the quality of paraphrased sentences. Each paraphrase is evaluated based on four key criteria, with reviewers assigning a score from 1 (poor) to 5 (excellent) for each criterion:

1. **INF (Informativeness)**: How well does the paraphrase retain the original meaning?
2. **REL (Relevance)**: To what extent are the important facts and details preserved?
3. **FLU (Fluency)**: How fluent and natural does the sentence sound?
4. **COH (Coherence)**: How well do the sentence parts fit together to form a coherent whole?

C Paraphrase Verification Checklist

To ensure the quality and accuracy of the paraphrased sentences in our dataset, we implemented a verification process where annotators assessed whether each sentence pair constituted a valid paraphrase. Annotators were instructed to evaluate each sentence pair based on the above criteria in Table 14. If the paraphrased sentence met all the criteria, it was marked as a valid paraphrase. If it failed to meet any of the criteria, it was marked as invalid.

Rule	Question
SEMANTIC EQUIVALENCE	Does the paraphrased sentence convey the same meaning as the original sentence, preserving all key information without adding or omitting important details?
FLUENCY & GRAMMATICAL	Is the paraphrased sentence grammatically correct and fluent in Vietnamese?
STYLE & TONE CONSISTENCY	Does the paraphrase maintain the same style and tone as the original sentence?
NO CONTRADICTIONS	Does the paraphrase avoid contradicting any facts or statements in the original sentence?

Table 14: Checklist used by annotators to verify if a sentence pair is a valid paraphrase.

Figure 6 presents the distribution of errors across different paraphrase verification rules. The analysis reveals that SEMANTIC EQUIVALENCE is the most common source of errors, indicating that paraphrased sentences often fail to fully preserve the meaning of the original text. This suggests that maintaining semantic consistency remains a significant challenge in paraphrase generation. Additionally, while NO CONTRADICTIONS and STYLE & TONE CONSISTENCY exhibit lower error rates, FLUENCY & GRAMMATICAL still accounts for a noticeable portion of errors.

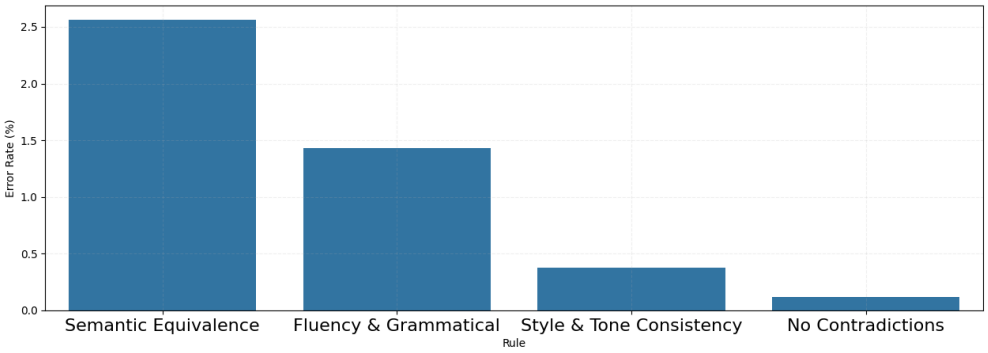


Figure 6: Error rate distribution across different paraphrase verification rule.

D Model Outputs

In Table 15, we present example outputs from the models across various experiments to further clarify and support the findings discussed in the main text. These supplementary results include detailed paraphrases generated by models like mBART, BARTpho, ViT5, and mT5.

Model	Paraphrases
Ground Truth	Đại lễ săn thỏ Phục sinh năm nay là lần đầu tiên được tổ chức kể từ năm 2017. (*English: This year's Easter Bunny Hunt is the first to be held since 2017.*) Sau 2017, năm nay là lần đầu tiên Đại lễ săn thỏ Phục sinh được tổ chức. (*English: After 2017, this year is the first time the Easter Bunny Hunt is held.*) Đại lễ săn thỏ Phục sinh đã được tổ chức trở lại sau 5 năm vắng bóng, kể từ năm 2017. (*English: The Easter Bunny Hunt has been held again after a 5-year absence, since 2017.*) Từ năm 2017 đến nay, Đại lễ săn thỏ Phục sinh đã không được tổ chức, cho đến năm nay. (*English: Since 2017, the Easter Bunny Hunt has not been held, until this year.*)
$\text{mBART}_{\text{large}}$	Đại lễ săn thỏ Phục sinh đã được tổ chức lần đầu tiên kể từ năm 2017. Đại lễ săn thỏ Phục sinh được tổ chức lần đầu tiên kể từ năm 2017. Đại lễ săn thỏ Phục sinh năm nay lần đầu tiên được tổ chức kể từ năm 2017.
mT5_{base}	Năm nay, Đại lễ săn thỏ Phục sinh được tổ chức kể từ 2017. Vào năm nay, Đại lễ săn thỏ Phục sinh được tổ chức kể từ 2017. Năm nay, Đại lễ săn thỏ Phục sinh đã được tổ chức kể từ 2017.
$\text{mT5}_{\text{large}}$	Săn thỏ Phục sinh được tổ chức kể từ 2017. Này là lần đầu tiên Đại lễ săn thỏ Phục sinh được tổ chức kể từ 2017. Săn thỏ Phục sinh được tổ chức kể từ 2017 năm nay.
$\text{BARTpho-syllable}_{\text{base}}$	Đây là lần đầu tiên Đại lễ săn thỏ Phục sinh được tổ chức kể từ năm 2017. Kể từ năm 2017, Đại lễ săn thỏ Phục sinh đã được tổ chức lần đầu tiên . Đại lễ săn thỏ Phục sinh năm nay là lần đầu tiên được tổ chức kể từ 2017.
$\text{BARTpho-syllable}_{\text{large}}$	Đại lễ săn thỏ Phục sinh năm nay đánh dấu lần đầu tiên Đại lễ được tổ chức kể từ năm 2017. Đại lễ săn thỏ Phục sinh năm nay đánh dấu lần đầu tiên được tổ chức kể từ năm 2017. Đại lễ săn thỏ Phục sinh năm nay đánh dấu lần đầu tiên Đại lễ này được tổ chức kể từ năm 2017.
$\text{BARTpho-word}_{\text{base}}$	Đại lễ săn thỏ Phục sinh đã được tổ chức lần đầu tiên kể từ năm 2017. Đại lễ săn thỏ Phục sinh năm nay là lần đầu tiên được tổ chức kể từ 2017. Đại lễ săn thỏ Phục sinh đã được tổ chức thường niên kể từ năm 2017.
$\text{BARTpho-word}_{\text{large}}$	Đại lễ săn thỏ Phục sinh năm nay là lần đầu tiên sự kiện này được tổ chức kể từ năm 2017. Đại lễ săn thỏ Phục sinh năm nay đánh dấu lần đầu tiên lễ hội này được tổ chức kể từ năm 2017. Đại lễ săn thỏ Phục sinh năm nay là sự kiện đầu tiên được tổ chức kể từ năm 2017.
$\text{ViT5}_{\text{base}}$	Đại lễ săn thỏ Phục sinh năm nay là lần đầu tiên được tổ chức kể từ năm 2017. Đại lễ săn thỏ Phục sinh năm nay lần đầu tiên được tổ chức kể từ năm 2017. Đại lễ săn thỏ Phục sinh năm nay là lần đầu tiên được tổ chức tại Việt Nam kể từ năm 2017.
$\text{ViT5}_{\text{large}}$	Việc tổ chức Đại lễ săn thỏ Phục sinh năm nay là lần đầu tiên được tổ chức kể từ năm 2017. Việc tổ chức Đại lễ săn thỏ Phục sinh là lần đầu tiên được tổ chức kể từ năm 2017. Người ta đã tổ chức Đại lễ săn thỏ Phục sinh lần đầu tiên kể từ năm 2017.

Table 15: Examples of paraphrases generated by various models for the original sentence "Năm nay là lần đầu tiên Đại lễ săn thỏ Phục sinh được tổ chức kể từ 2017." (*English: This year is the first time the Easter Bunny Hunt has been held since 2017.*). The highlighted show the lexical differences compared to the original sentence.