

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

GRADUATION THESIS

**Thesis title: Improving Vietnamese Question
Generation Using Reinforcement Learning**

Nguyen Hoang Dang
dang.nh204873@sis.hust.edu.vn

Major: Computer Science
Specialization: Data Science

Supervisor: PhD. Tran The Hung

Signature

Department: Computer Science

School: School of Information and Communications Technology

HANOI, 06/2024

ACKNOWLEDGMENT

This is my graduation thesis, marking the end of 4 years of study at Hanoi University of Science and Technology. During the past 4 years, I feel that I have had the opportunity to learn and grow through the chance to be exposed to problems that are not only new and applicable but also difficult. With the enthusiasm of the teachers and their dedicated guidance, I am even more passionate about the field of study I have chosen. These are not only useful knowledge but also a solid foundation for me to continue building and pursuing my educational path. I express my thanks to all the teachers who instructed me to complete my subjects here. These lessons always help me in the process of building my future career. Last but not least, I want to express my gratitude to PhD. Tran The Hung, who inspired me to select, complete the thesis, and achieve the expected results.

ABSTRACT

In the field of natural language processing, automatic question generation is reckoned to be a challenge. This is due to the ambiguity where many questions can lead to the same answer. We want our model, after being fine-tuned by on reinforcement learning task, to be able to generate questions that not only resemble the style in the data set but also appear natural and contextual. We introduce a selection method for the preference dataset which is then used for a preference-based reinforcement learning method to align the policy language model. Additionally, we enlarge the training dataset using synthetic data generated from a large language model (LLM) fine-tuned with instruction. Our training method enhances the model stability on different sampling methods, generating better-rated questions in terms of AI evaluation. Quantitatively, our method improves the model performance, showing an increase in Rouge-L (+3.44), BLEU-4 (+3.18), and Win rate (+12.73). The advancement made by our method on the model is also explained in detail.

Student

(Signature and full name)

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION.....	1
1.1 Problem Statement.....	1
1.2 Background and Problems of Research	3
1.3 Research Objective and Conceptual Framework	4
1.4 Contributions	4
1.5 Organization of Thesis	4
CHAPTER 2. LITERATURE REVIEW	6
2.1 Scope of Research	6
2.2 Related Works	6
2.2.1 Question-answering dataset.....	6
2.2.2 Automatic Question Generation.....	6
2.2.3 LLM as evaluator	7
2.3 Reinforcement Learning from Human & AI Feedback	7
2.3.1 Reinforcement Learning in Natural Language Generation	7
2.3.2 Reward from Human & AI Feedback	8
2.4 Proximal Policy Optimization	10
2.5 Direct Preference Optimization (DPO)	13
CHAPTER 3. METHODOLOGY	15
3.1 Overview	15
3.2 Models	16
3.2.1 ViT5	16
3.2.2 Vistral-7b-chat.....	16
3.3 Dataset	17
3.3.1 UIT-ViQuAD.....	17

3.3.2 Data Processing	18
3.4 Augmentation Using Synthetic Data	19
3.5 Efficient Fine-tuning	22
3.5.1 Quantization	22
3.5.2 Low-rank Adaptation	23
3.6 Supervised Fine-tuning	24
3.7 Model Aligning	25
3.7.1 Self-play Fine-tuning with DPO (SDPO)	25
3.7.2 Self-preference Selection with AI Feedback	26
CHAPTER 4. NUMERICAL RESULTS	29
4.1 Evaluation Parameters	29
4.1.1 RougeL Score	29
4.1.2 BLEU-4 Score	30
4.1.3 BERTScore	30
4.1.4 Win rate	31
4.2 Simulation Methods	32
4.2.1 Generation settings	32
4.2.2 Experiments	33
4.2.3 Results and Discussion	34
CHAPTER 5. CONCLUSIONS	36
5.1 Summary	36
5.2 Suggestion for Future Works	36
REFERENCE	40
APPENDIX	42

A. PROMPTS AND TRAINING SETTINGS.....	42
A.1 Examples of LLM prompt	42
A.1.1 Prompt for data augmentation.....	42
A.1.2 Prompt for preference selection	43
A.2 Detailed experiment settings	44
B. ADDITIONAL EMPIRICAL RESULTS	47

LIST OF FIGURES

Figure 2.1	Description of how reinforcement learning works.	7
Figure 2.2	Language model as a policy $\pi_{\theta}(a s)$	8
Figure 2.3	The process by which a set of priorities is constructed based on humans' or AIs' feedback.	9
Figure 3.1	Data pipeline.	15
Figure 3.2	Self-play preference set selection.	25
Figure 3.3	The preference set is constructed based on the LLM responses.	27
Figure 4.1	Win rate of $Vistral_{Our}$ versus $Vistral_{SFT}$	35

LIST OF TABLES

Table 4.1	Evaluation scores	34
Table B.1	Additional scores measured on different generation parameters.	47

LIST OF ABBREVIATIONS

Abbreviation	Definition
API	Application Programming Interface
AQG	Automatic Question Generation
DPO	Direct Preference Optimization
LLM	Large Language Model
PPO	Proximal Policy Optimization
RL	Reinforcement Learning
RLAIF	Reinforcement Learning from Artificial Intelligence Feedback
RLHF	Reinforcement Learning from Human Feedback
SDPO	Self-play Fine-tuning with Direct Preference Optimization
SFT	Supervised Fine-tuning