CTU Journal of Innovation and Sustainable Development
ISSN 2588-1418 | e-ISSN 2815-6412

DOI:10.22144/ctujoisd.2024.325

VQABG: Vietnamese question/answers benchmark generator for field-specific chatbot ground-truth dataset using EMINI (Exact Match wIth Numeric Information) indicator

Ngo-Ho Anh-Khoa[1], Vo Khuong-Duy[2], and Ngo-Ho Anh-Khoi[3*]

[1]*Quantmetry, France*

[2]*Adhightech Ltd., Viet Nam*

[3]*Faculty of Information Technology, Nam Can Tho University, Viet Nam*

**Corresponding author (nhakhoi@nctu.edu.vn)*

Article info.

Received 1 Jul 2024
Revised 4 Sep 2024
Accepted 5 Oct 2024

Keywords

Automatic question answering generator, Chatbot, Generative Artificial Intelligence, Vietnamese language

ABSTRACT

Currently, the application of generative Artificial Intelligence for developing specialized chatbots in Vietnamese is an inevitable trend. However, one of the most challenging aspects of assessing the quality of Vietnamese chatbot products is creating a specialized benchmark in a question-and-answer format. Typically, this benchmark is manually crafted by industry experts, which can be extremely costly. In contrast, for English, we can use bag-of-words model toolkits and grammatical structure architectures to generate appropriate questions automatically based on pre-existing answers from the original data. However, there is almost no complete model available for this task in Vietnamese. Regarding quality assessment, this is usually performed manually by experts using Human Evaluation (HE) indicators, which is also costly. Therefore, the aim of this study is to propose an algorithmic architecture specifically designed for the Vietnamese language. This architecture will automatically generate a set of question-and-answer queries to create a benchmark, as well as facilitate the development of a mechanism for automatic, straightforward, cost-effective, and accurate quality assessment for Vietnamese chatbots. We refer to this system as the Vietnamese Question/Answers Benchmark Generator (VQABG) and propose an innovative evaluation indicator called the Exact Match with Numeric Information (EMINI).

1. INTRODUCTION

In the field of Artificial Intelligence (AI), a benchmark is a method used to evaluate, compare the performance and capability of various models, algorithms, or AI systems. A benchmark usually incorporates a standard dataset and a series of predetermined tests to measure accuracy, speed, reliability, and other attributes of an AI model or system. These benchmarks aid researchers and developers in enhancing and optimizing AI models and assessing their performance based on specific measurement standards.

Current research concentrates on defining a benchmark for chatbot systems in Vietnamese, emphasizing the creation of a question/answer set for evaluating chatbot reliability. One popular technique used for this evaluation involves a validation set of corresponding questions and answers and compares chatbot answers to correct responses. However, this validation set is often manually created, requiring a team of industry experts to read the text and create questions as well as suitable responses. While other methods exist, they're less common. No matter which method is used, most current evaluation systems require human intervention, whether direct or indirect. This process can be expensive and heavily reliant on expert knowledge, potentially leading to biases or errors.

Particularly within the Vietnamese research context, with scarce experts, creating a quality benchmark is tough. After the benchmark creation, evaluating chatbot responses based on the benchmark's standard responses also poses a challenge. Typically, determining the reliability of a chatbot's responses requires painstaking human effort to compare chatbot answers manually to the reference set. As aforementioned, this task is not only heavily reliant on expert knowledge in the field, but it can also be inherently subjective or error-prone during the evaluation process.

This research focuses on solving this issue with two objectives: first, to auto-generate a set of questions/responses for Chatbot from input sources, and second, to enable the generated set to autonomously assess the quality of Chatbot responses. The second objective is crucial, given that chatbot responses are often complex and human-like. relying on human assessment not only incurs costs but also poses accuracy related risks.

In most languages, auto-generating appropriate questions from responses is a substantial research area requiring much effort. This problem is mostly due to the complex grammatical structures in most languages, specifically issues related to verb tense agreement. As a result, in most studies and real production processes, manually created benchmarks are the only option, even for relatively simple languages such as English. According to our research, no automatic model exists for generating a question-answer set from text for Vietnamese. As such, creating fully automatic Vietnamese chatbot question/answer sets remains an open issue.

Given the simple grammatical structure of Vietnamese, our research team discovered that some questions could easily be created from simple affirmations or denials. By substituting a component in the sentence with words like "what", "when" or "how much" (in Vietnamese) without compromising grammatical accuracy, these questions can be made. Leveraging this advantage, we propose creating automatic question-and-answer sets in Vietnamese-based on a definite input source. Not only does this allow for the creation of a wholly automatic and accurate benchmark, but it also provides an opportunity to set up an automatic assessment system for chatbot responses.

Evaluating chatbots is usually based on three following factors: efficiency, perfection, and satisfaction (Casas et al., 2020). Efficiency emphasizes whether the chatbot can accomplish the proposed tasks. Perfection reflects the similarity between the answer generated by the chatbot and the human answer. Satisfaction measures the level of user satisfaction when using chatbots. Perfection and satisfaction are usually evaluated through user feedback. In this study, we focus on the effectiveness factor: assessing the chatbot's ability to provide accurate answers. Typically, this ability is evaluated through a benchmark set of question-and-answer lists. To date, most benchmarks are manually created: a group of people read a text and produce questions related to this text. As for the answers, benchmarks can use multiple-choice, short answers (single entities, meaning the answer is a word or a group of words, not a complete sentence), and free-form answers (meaning the answer must be a complete sentence but does not follow a certain structure). However, regardless of the type, until now most evaluation systems use human elements to perform the evaluation task, an approach is relatively expensive, causing difficulty for the overall evaluation problem, especially in the development of small products with low costs.

Current benchmarks can be divided into three groups: Reading Comprehension Benchmarks, Question Answering Benchmarks, and Cloze Benchmarks (Rajpurkar et al., 2016). The Reading Comprehension Benchmark evaluates the ability to synthesize multiple sentences in a text to answer a question. This benchmark is complex and requires a chatbot's reasoning ability. Thus, this set of questions and answers is usually manually created. Some benchmarks in this group include those of Hirschman et al. (1999) and Richardson et al. (2013). The Cloze Benchmark evaluates the chatbot's ability to fill in the blanks. The question is a sentence or a text passage with a few words or phrases omitted. The omitted words would be answers to these questions. This benchmark can be automatically generated. However, it is difficult to evaluate a chatbot's reasoning ability with this benchmark. One example of this benchmark is The

Children's Book Test by Hill et al. (2015). The Question-Answering Benchmark often relies on extracting the answer directly from the text. That means a part of a sentence is extracted to be the answer. The rest of the sentence is transformed into a question. Some noticeable studies following this approach are Voorhees et al. (2000), Ferrucci et al., (2010), Yang et al., (2015), Le-Hong et al. (2018), Le et al. (2022) and Islam et al. (2023). These studies are partially manually made. In the case of Yang et al. (2015), a search query from user's history is taken to form the question set and answers are extracted from Wikipedia pages the user clicked on. In the search history, queries starting with question words (for example, what, who, when, why, and how) and ending with a question mark are considered as questions. Wikipedia pages, text passages, or sentences that might contain the answer are manually picked. In the work of Le-Hong et al. (2018) and Le et al. (2022), they manually create question answer benchmarks for the Vietnamese language. Several recent works use a large language model as an annotator to generate question-answer pairs (Lyu et al., 2024; Kenneweg et al., 2024). Our study also focuses on creating a Question-Answering Benchmark, trying to automate the process of extracting information from text and transforming text into questions. Except for a few special cases (for example, Q&A section on websites), there's little chance to find readily available question/answer datasets, especially in specialty fields, in most cases only either the question or worse, it is nearly impossible to find an available question set for a specialty problem. Conversely, in most languages, generating accurate questions automatically from answers is a huge field of study, however, the achievements are relatively small, due to complex grammatical structures in most languages, especially the problem of verb conjugation or linguistic agreement (Ngo Ho, 2021; Kenneweg et al., 2024; Lyu et al., 2024), it is a complex issue in languages. The RAGAS technique (Retrieval Augmented Generation ASsessment) could be leveraged to create a few solutions to eliminate manual benchmarks, but with benchmarks based on LLM to evaluate LLM, it almost just wants to focus on evaluating “satisfaction” or more precisely, “human-likeness of the answer" instead of focusing on evaluating the reliability of the answer (Es et al., 2023). Therefore, in most research, or actual product implementation process, using manual benchmarks is almost the only choice.

To our knowledge, there are no automated models to generate a set of questions-answers from text for the Vietnamese language. However, in Vietnamese, due to the monosyllabic grammar characteristics, some forms of questions could easily be formed from affirmative/negative sentences by replacing a sentence component with question words like “what”, “when”, “how much” without reducing the grammatical accuracy. Capitalizing on this feature, the research proposes to create question/answers in Vietnamese-based on given sources, allowing to produce a completely automatic standard set, as well as allowing the establishment of automated evaluation testing on chatbot answers, thus reducing the cost of manually performing the standard tasks, which are usually costly. In this model, the most challenging part is identifying the words or word combinations (word groups) that need to be taken to form the answer, as Vietnamese is an isolating language. In addition, the model must be able to identify accurate question words to replace these words.

This study considers using statistical data in questions to increase the accuracy of answers. Since data is often easier to verify than pure text, which often gets disrupted by the intelligent response mechanism of chatbots, causing information elements to become chaotic and complicating the reliability verification by industry experts, consequently, querying data allows automatic calibration evaluations to become feasible without human intervention. Through setting up algorithms that automatically generate data query questions in Vietnamese-based on a certain input source, this study creates a completely automated approach to calibrate the reliability of responses from chatbots. This not only reduces costs associated with manual verification but also enhances the quality of the evaluation. The goal is to simply reduce costs related to manually conducting verification tasks, which are among the most expensive tasks.

2. MATERIALS AND METHOD

2.1. Vietnamese Question/Answers Benchmark Generator (VQABG)

The aim of this research is twofold. First, we would like to generate a set of query-based question/answers automatically from the input source to create a dedicated standard set in a simple and cost-effective way. Secondly, this standard set eases the quality control of chatbot responses, which is quite challenging due to the human-like, complex nature of such responses. This makes human manual checks not only costly but also risky, for a wrong answer in a complex, multiple-component format could confuse the checker into rating it as correct. As chatbot responses adapt to human conversation style, rating the quality of answering general questions is a big challenge.

The research idea is based on a simple fact that "numbers don't lie" meaning that even if chatbot responses may be complex, multi-component and difficult to understand, this will only happen with descriptive responses, but for questions asking numerical information, e.g. amount, number, year, the accuracy of responses could easily be determined. Another feature of this question type is its ability to check answers automatically without manual human work, satisfying both search criteria for this research. Therefore, the research proposes to build an automated process to create a set of question/answers based on queries, with the goal of creating the standard set as well as developing a method to evaluate chatbot responses automatically. This is designed for specialty chatbots in Vietnamese, with the aim of reducing costs and improving the accuracy of model evaluation.

Due to the straightforward grammatical structure of Vietnamese, our research team found that it is possible to generate questions from basic affirmations or negations quite easily. By replacing certain elements in a sentence with interrogative words such as "what" ("cái gì" in Vietnamese), "when" ("khi nào" in Vietnamese), or "how much" ("bao nhiêu" in Vietnamese), we can form questions while maintaining grammatical correctness. Capitalizing on this feature, we suggest developing automated question-and-answer pairs in Vietnamese derived from a specific input source. This approach not only facilitates the creation of a fully automated and precise benchmark but also enables the establishment of an automatic evaluation system for chatbot interactions. It must be stated clearly that this method applies exclusively to Vietnamese, as the structural characteristics of the Vietnamese language allow for this extremely simple change to be implemented. The corresponding tasks in other languages, such as English, are exceedingly complex; therefore, this method is rarely used to perform such equivalent tasks, and as a result, there is almost no similar research in English or other European languages in general. Consequently, this technique is considered solely in the context of Vietnamese chatbot research. One of the contributions of our study is the discovery of this simple truth about the Vietnamese language.

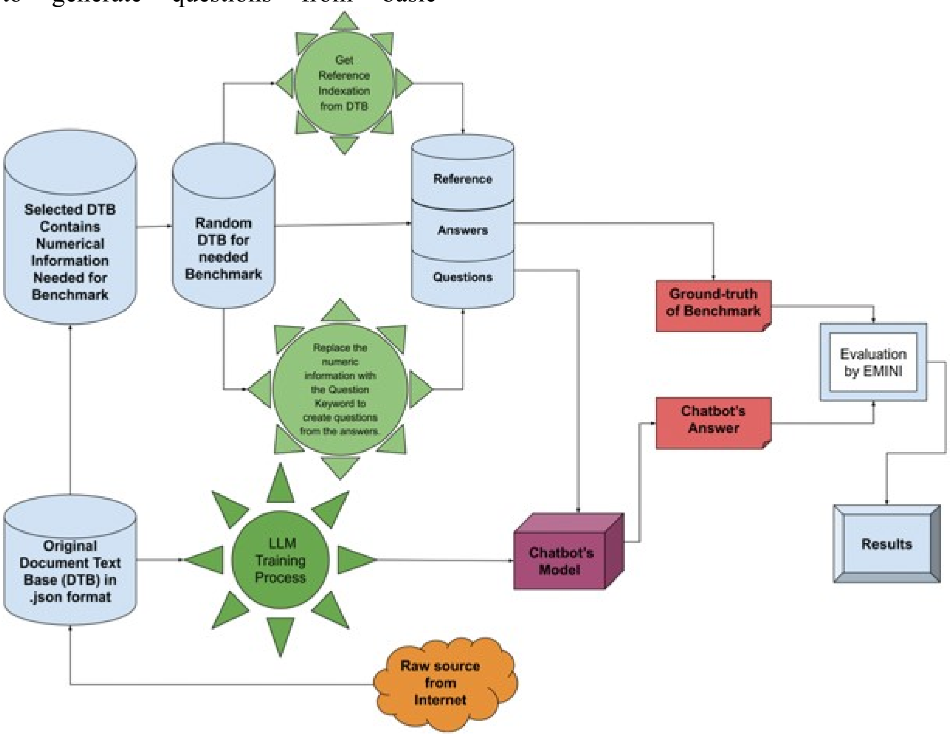

Figure 1. Processing diagram of the VQABG technique

The VQABG technique can be summarized as follows: (1) Data is mined from raw internet sources, filtered to remove noise (e.g., spam, ads), and broken down into individual sentences to form the source data. (2) From this source data, a portion or the entirety is selected to create a standard set. (3) The standard set is further filtered to retain only sentences containing numerical information, specifically those with ASCII values representing digits (0-9). To generate questions, numerical content in the sentences is replaced with the keyword "bao nhiêu" (meaning "how much" in Vietnamese). When a sentence contains more than one number, all numerical content is replaced with appropriate question keywords. These modified sentences form the questions of the standard set, while the original numerical content forms the corresponding answers. The position of each sentence within its paragraph and news group is recorded as a reference for the standard set.

During evaluation, each question from the standard set is input into the chatbot to generate a response. The chatbot's answer is then compared to the standard set's answer using an Exact Match (EM) index, which measures the accuracy of the chatbot model. Below is a flowchart illustrating the VQABG technique (Figure 1).

2.2. Evaluation with EMINI indicator

We propose an improved Exact Match metric (EM) (Chen et al., 2024; Lyu et al., 2024) by combining EM with the evaluation of numeric information: EMINI (Exact Match wIth Numeric Information). This helps to measure reliability in answers used for the VQABG technique. This is a relatively straightforward measurement. For a pair of question/answer, if the keywords in the predicted answer exactly match the keywords in the reference answer, then EMINI = 1; otherwise, EMINI = 0. It is important to note that the ground truth of the benchmark is an answer with accurate numerical data or accurate keywords (e.g., datetime). Thus, a score of 1 is given to an answer if the chatbot provides accurate data or correct keywords in the answer, and conversely. Simultaneously, the study also employs the ASS (Answer Semantic Similarity) metric from the RAGAS (Retrieval Augmented Generation Assessment) technique, which is widely used to measure chatbot reliability (Es et al., 2023), for comparison with the results measured by EMINI. As mentioned in the earlier section, ASS, like most of the measures currently in use, focuses on evaluating “satisfaction” or, more precisely, “human-likeness of the answer” rather than concentrating on assessing the accuracy of the answer.

By analyzing these two different metrics in specific examples, the advantages of using the VQABG technique to evaluate chatbot reliability can be observed. Regarding the reliability evaluation with the calibration set, the system simply searches for the numerical information data portion in the chatbot's answer, it is considered a correct answer if found, and vice versa, if the portion of numerical information data is not present in the answer it is considered incorrect. Therefore, the chatbot must provide an accurate quantitative response and an accurate reference to the information source for the response to be deemed correct. Since the numerical information data portion can easily be separated from the answer, therefore, the evaluation process can be fully automated.

3. RESULTS AND DISCUSSION

To test the accuracy of the chatbot's responses, it is necessary to implement a specialized database. The study chooses the VEID database to illustrate the VQABG technique. This research designed an automated system to generate questions and answers from source data from the VnEconomy community, the largest economic information source in Viet Nam, to date. The research utilized a Python script to retrieve all economic information data from the website of this major economic information source, VnEconomy. The extracted data is saved as text files (*.txt*, *.pdf*, *.html*) to create the original database. This original database is then structured into hierarchical files to document the source reference. Next, the data is processed by sentence segmentation. Finally, the data is stored in a hierarchical structure in “.json” format, creating the Vietnamese Economy Information Database (VEID). The VEID database includes 27,682 documents, totaling 30,186,806 words (as of May 1, 2024). VEID consists of nine main categories:

– Enterprises: including 3,547 documents with 3,774,021 words.

– Digital economy: including 2,562 documents with 2,730,966 words.

– Finance: including 2,900 documents with 3,278,218 words.

– World [Economy]: including 2,773 documents with 2,879,443 words.

– Market: including 2,767 documents with 3,217,327 words.

– Real estate: including 3,268 documents with 3,279,093 words.

– Social [Economy]: including 3,952 documents with 4,067,845 words.

– Investment: including 4,256 documents with 4,822,709 words.

– Focus [Economy]: including 1,657 documents with 2,137,184 words.

The research has created the benchmark sets using the VQABG technique, consisting of 100 random questions from the database (Vietnamese Economy Keyword Information Database, VENID-R100) with keyword information through the VQABG technique. For comparison, the study utilizes two LLM models: Gpt-3.5-turbo with the following parameters: chunk_size = 1000; n_ctx = 4096; temperature = 0.01; top_p = 1; top_k = 10.

The results from running the chatbot model and the evaluation yielded the following outcomes:

Table 1. Comparison of overall results between ASS and EMINI in the VNEIQAD-R100 benchmark set

VQABG	ASS ± std	EMINI
Gpt-3.5-turbo	92,14% ± 7,69%	56%

Looking at the results table, it is evident that there is a significant discrepancy of nearly 50% between the ASS and EMINI measurements. This means that more than half of the responses that EMINI considers incorrect have been assessed by ASS as correct. The choice of ASS in RAGAS only makes this difference more pronounced; using other measurement metrics from RAGAS yields similar results, although the disparity may not be as large as with ASS. This is because ASS is very characteristic of assessing 'correctness' based on satisfaction rather than reliability, which is a weakness. If traditional benchmarking methods are used, they can yield very satisfactory results but completely lack the accuracy of the answers. Clearly, this evidence shows the mistake of relying predominantly on current measurement indicators in most studies, while it is quite challenging or modest to develop benchmarks using the EM measurement index. This also indicates that EMINI, along with the VQABG technique, is a benchmarking method worth considering. In the following section, the study will cite some typical cases of question results from the VNEIQAD-R100 benchmark set to illustrate the issues faced in the currently developed benchmark sets alongside the ASS measurement index or similar ones, as well as to highlight the advantages of the EMINI index along with the VQABG technique.

Table 2. Example 1-th of the results for question 100-th in the VNEIQAD-R100 standard set

<table>
<tr><td rowspan="2">Benchmark VNEIQAD-R100
[No of Question: 100]</td><td>Question</td><td>Ông Lê Hồng Việt, Tổng Giám đốc FPT Smart Cloud cho rằng AI mới bùng nổ bao nhiêu năm gần đây? (EN: Mr. Le Hong Viet, General Director of FPT Smart Cloud, thinks that AI has exploded in how many years?)</td><td colspan="2">Model Evaluation</td></tr>
<tr><td>Correct Answer</td><td>Ông Lê Hồng Việt, Tổng Giám đốc FPT Smart Cloud cho rằng AI mới bùng nổ 2 năm gần đây. (EN: Mr. Le Hong Viet, General Director of FPT Smart Cloud, said that AI has only exploded in the last 2 years.)</td><td>ASS</td><td>EMINI</td></tr>
<tr><td></td><td>Chatbot’ s Answer</td><td>FPT Smart Cloud cho rằng AI mới bùng nổ trong vòng 10 năm gần đây. (EN: FPT Smart Cloud believes that AI has only exploded in the last 10 years.)</td><td>94,88%</td><td>0</td></tr>
</table>

Note: The chatbot answered incorrectly, but ASS rated it as correct, while EMINI accurately assessed it as incorrect.

Table 3. Example 2-th of the results for question 83-th in the VNEIQAD-R100 standard set

<table>
<tr><td rowspan="2">Benchmark VNEIQAD-R100 [No of Question: 83]</td><td>Question</td><td>Đơn vị cho biết, năm thành phố là Bắc Kinh, Thượng Hải, Singapore, Sydney, Tokyo hiện chiếm bao nhiêu% công suất hoạt động của tất cả trung tâm dữ liệu trên toàn châu Á-Thái Bình Dương? (EN: The unit said, the five cities of Beijing, Shanghai, Singapore, Sydney, Tokyo currently account for what percentage of operating capacity of all data centers across Asia-Pacific?)</td><td colspan="2">Model Evaluation</td></tr>
<tr><td>Correct Answer</td><td>Đơn vị cho biết, năm thành phố là Bắc Kinh, Thượng Hải, Singapore, Sydney, Tokyo hiện chiếm 62% công suất hoạt động của tất cả trung tâm dữ liệu trên toàn châu Á-Thái Bình Dương. (EN: The five cities of Beijing, Shanghai, Singapore, Sydney and Tokyo currently account for 62% of all data center capacity across Asia-Pacific, the unit said.)</td><td>ASS</td><td>EMINI</td></tr>
<tr><td></td><td>Chatbot’s Answer</td><td>62%</td><td>76,24%</td><td>1</td></tr>
</table>

Note: The chatbot answered correctly, but ASS rated it as somewhat incorrect, while EMINI accurately assessed it as correct.

Table 4. Example 3-th of the results for question 58-th in the VNEIQAD-R100 standard set

<table>
<tr><td rowspan="2">Benchmark VNEIQAD-R100 [No of Question: 58]</td><td>Question</td><td>Trong bối cảnh phát triển ứng dụng di động năng động, sự phát triển của Thực tế tăng cường (AR) và Thực tế ảo (VR) trong các lĩnh vực đa dạng có thể sẽ là một xu hướng quan trọng vào năm bao nhiêu? (EN: In the dynamic mobile app development landscape, the growth of Augmented Reality (AR) and Virtual Reality (VR) in diverse domains is likely to be an important trend in which year?)</td><td colspan="2">Model Evaluation</td></tr>
<tr><td>Correct Answer</td><td>Trong bối cảnh phát triển ứng dụng di động năng động, sự phát triển của Thực tế tăng cường (AR) và Thực tế ảo (VR) trong các lĩnh vực đa dạng có thể sẽ là một xu hướng quan trọng vào năm 2024. (EN: In the dynamic mobile app development landscape, the growth of Augmented Reality (AR) and Virtual Reality (VR) in diverse domains is likely to be a key trend in 2024.)</td><td>ASS</td><td>EMINI</td></tr>
<tr><td></td><td>Chatbot’s Answer</td><td>Năm 2024. (EN: Year 2024)</td><td>83,28%</td><td>1</td></tr>
</table>

Note: The chatbot answered correctly, but ASS rated it as somewhat incorrect, while EMINI accurately assessed it as correct.

Table 5. Example 4-th of the results for question 81-th in the VNEIQAD-R100 standard set

Benchmark *VNEIQAD-R100 [No of Question: 81]*	**Question**	Đó là lời khuyên của ông Nguyễn Hải Nam, Giám đốc Công ty trách nhiệm hữu hạn Credit bao nhiêu.AI tại tọa đàm "Phòng ngừa rủi ro trong thanh toán không dùng tiền mặt" do Tạp chí Kinh tế Việt Nam - VnEconomy tổ chức sáng ngày bao nhiêu/10? (EN: That is the advice of Mr. Nguyen Hai Nam, Director of Credit bao nhieu.AI Limited Liability Company at the seminar "Risk prevention in non-cash payments" organized by Vietnam Economic Magazine - VnEconomy on the morning of which day in October?)	**Model Evaluation**	
	Correct Answer	Đó là lời khuyên của ông Nguyễn Hải Nam, Giám đốc Công ty trách nhiệm hữu hạn Credit 360.AI tại tọa đàm "Phòng ngừa rủi ro trong thanh toán không dùng tiền mặt" do Tạp chí Kinh tế Việt Nam - VnEconomy tổ chức sáng ngày 2/10. (EN: That is the advice of Mr. Nguyen Hai Nam, Director of Credit 360.AI Limited Liability Company at the seminar "Risk prevention in non-cash payments" organized by Vietnam Economic Magazine - VnEconomy on the morning of October 2.)	**ASS**	**EMINI**
	Chatbot' s Answer	Lời khuyên của ông Nguyễn Hải Nam, Giám đốc Công ty trách nhiệm hữu hạn Credit 360.AI tại tọa đàm "Phòng ngừa rủi ro trong thanh toán không dùng tiền mặt" do Tạp chí Kinh tế Việt Nam - VnEconomy tổ chức sáng ngày 2/10. (EN: Advice from Mr. Nguyen Hai Nam, Director of Credit 360.AI Limited Liability Company at the seminar "Risk prevention in non-cash payments" organized by Vietnam Economic Magazine - VnEconomy on the morning of October 2.)	99,4%	1

Note: The chatbot answered correctly, and both ASS and EMINI rated it as correct.

In the first three examples, it seems that the EMINI index yields more accurate results compared to ASS. In both first two examples, ASS gives high scores even when the chatbot's response is incorrect (example 1) or correct but lacking sufficient information (example 2). Meanwhile, EMINI correctly assesses the situation: it scores 0 when the answer is wrong and scores 1 when the answer is correct. In the third example, both ASS and EMINI assign the correct score when the chatbot answers correctly, but ASS gives a lower score, likely because the formulation of the chatbot's response does not match the correct answer in terms of semantic similarity. Finally, in the fourth example, both ASS and EMINI evaluate accurately when the chatbot answers correctly. Overall, if we only consider whether the chatbot's answers are correct or incorrect, EMINI seems to provide more accurate results. However, it is important to note that ASS also evaluates the similarity between the chatbot's response and the correct answer in terms of semantics, which is crucial for assessing the phrasing of the response. In the case of evaluating the reliability of the answers, EMINI is a reliable metric to measure with the VQABG technique.

Table 6. Example 5-th of the results for question 54-th in the VNEIQAD-R100 standard set

Benchmark *VNEIQAD-R100* *[No of Question: 54]*			**Model Evaluation**	
	Question	Theo đó, bao nhiêu thủ tục hành chính thuộc lĩnh vực đầu tư thực hiện trực tuyến gồm: …? (EN: Accordingly, how many administrative procedures in the investment sector are performed online, including: …?)		
	Correct Answer	Theo đó, 04 thủ tục hành chính thuộc lĩnh vực đầu tư thực hiện trực tuyến gồm: … (EN: Accordingly, 04 administrative procedures in the investment sector are implemented online, including: …)	**ASS**	**EMINI**
	Chatbot' s Answer	Có tổng cộng 4 thủ tục hành chính thuộc lĩnh vực đầu tư được thực hiện trực tuyến. (EN: There are a total of 4 administrative procedures in the investment sector that are performed online.)	88,73%	0

Note: The chatbot answered correctly, ASS rated it as somewhat incorrect, and EMINI mistakenly rated it as incorrect.

Table 7. Example 6-th of the results for question 67-th in the VNEIQAD-R100 standard set

Benchmark *VNEIQAD-R100* *[No of Question: 67]*			**Model Evaluation**	
	Question	Cục Phát thanh, truyền hình và thông tin điện tử đã rà quét và yêu cầu ngăn chặn bao nhiêu nhóm (groups) trên Facebook có nội dung tiêu cực […] và yêu cầu Facebook ngăn chặn bao nhiêu nhóm về nội dung này?		
	Correct Answer	Cục Phát thanh, truyền hình và thông tin điện tử đã rà quét và yêu cầu ngăn chặn 43 nhóm (groups) trên Facebook có nội dung tiêu cực […] và yêu cầu Facebook ngăn chặn 47 nhóm về nội dung này. (EN: The Department of Radio, Television and Electronic Information has scanned and requested to block how many groups on Facebook with negative content […] and requested Facebook to block how many groups with this content? The Department of Radio, Television and Electronic Information has scanned and requested to block 43 groups on Facebook with negative content […] and requested Facebook to block 47 groups with this content.)	**ASS**	**EMINI**
	Chatbot' s Answer	Cục Phát thanh, truyền hình và thông tin điện tử đã rà quét và yêu cầu ngăn chặn 43 nhóm (groups) trên Facebook có nội dung tiêu cực như hướng dẫn cách thức bùng nợ vay và 8 nhóm hướng dẫn cách tự tử. (EN: The Department of Radio, Television and Electronic Information has scanned and requested the blocking of 43 Facebook groups with negative content such as instructions on how to default on loans and 8 groups giving instructions on how to commit suicide.)	98,23%	0

Note: The chatbot answered correctly, ASS rated it correctly, and EMINI rated it incorrectly.

The research found two instances where EMINI incorrectly assessed the chatbot's response. The reason is that EMINI is currently configured to compare in the form of words containing numerical information, rather than converting them into numerical form. There are many complex numerical formats, and converting these can lead to inaccuracies. This error can be addressed by classifying content that contains numerical information into different groups that represent various types of numbers (natural numbers, complex numbers, fractions, etc.) and processing each type separately. Another cause is that when the source sentence contains more than one piece of numerical information, the resulting question will be in the form of a complex question (containing multiple query components) with several subordinate questions corresponding to the numerical data, leading to incomplete or insufficient answers from the chatbot regarding the required numerical data. This issue can be resolved by breaking down complex questions into smaller components, resulting in more generated questions, or by accepting complex questions as a complicated part of the benchmark.

4. CONCLUSION

The study will aim to improve these issues in future research. Currently, the application of generative AI to establish domain-specific chatbots in Vietnamese is an inevitable trend. However, one of the most challenging aspects of assessing the quality of Vietnamese chatbot products is the creation of a domain standard for Vietnamese chatbots in a question-answer format. Typically, this domain benchmark is manually created by industry experts, which is very costly. For English, we can use tools for modeling bag-of-words and grammatical structure architectures to automatically generate appropriate questions based on existing answers from the original data. For quality assessment tasks, this is usually done manually by experts (human evaluation), which is also quite expensive.

This study introduces the VQABG (Question/Answers Generating with Numerical Information) technique, which allows for the automatic generation of questions/answers that query numerical information in the source database in Vietnamese, also our proposed EMINI indicator of evaluation. Thus, it creates a domain-specific benchmark and evaluation indicator in Vietnamese while also enabling a mechanism for automatic quality assessment that is quick, cost-effective, and accurate for evaluating the quality of domain-specific Vietnamese chatbots.

REFERENCES

Casas, J., Tricot, M.-O., Abou Khaled, O., Mugellini, E., & Cudré-Mauroux, P. (2021). Trends & methods in chatbot evaluation. In *Companion publication of the 2020 International Conference on Multimodal Interaction (ICMI '20 Companion)* (pp. 280–286). Association for Computing Machinery. https://doi.org/10.1145/3395035.3425319

Chen, J., Lin, H., Han, X., & Sun, L. (2024, March). Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 16, pp. 17754–17762).

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). Ragas: Automated evaluation of retrieval augmented generation. arXiv preprint arXiv:2309.15217.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., ... & Welty, C. (2010). Building Watson: An overview of the DeepQA project. *AI magazine*, *31*(3), 59-79.

Hill, F., Bordes, A., Chopra, S., & Weston, J. (2015). The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Hirschman, L., Light, M., Breck, E., & Burger, J. D. (1999). Deep read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 325–332). Association for Computational Linguistics.

Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., & Vidgen, B. (2023). FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Kenneweg, T., Kenneweg, P., & Hammer, B. (2024). Retrieval augmented generation systems: Automatic dataset creation, evaluation and Boolean agent setup. *arXiv preprint arXiv:2403.00820.*

Le-Hong, P., & Bui, D. T. (2018). A factoid question answering system for Vietnamese. In *Companion Proceedings of the Web Conference 2018* (pp. 1049–1055).

Le, K., Nguyen, H., Le Thanh, T., & Nguyen, M. (2022, June). VIMQA: A Vietnamese dataset for advanced reasoning and explainable multi-hop question answering. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 6521–6529).

Lyu, Y., Li, Z., Niu, S., Xiong, F., Tang, B., Wang, W., Wu, H., Liu, H., Xu, T., & Chen, E. (2024). Crudrag: A comprehensive Chinese benchmark for retrieval-augmented generation of large language models. arXiv preprint arXiv:2401.17043.

Ngo Ho, A. K. (2021). *Generative probabilistic alignment models for words and sub-words: A systematic exploration of the limits and potentials of neural parametrizations* (Master's thesis, Université Paris-Saclay). Document and Text Processing. https://tel-03210116

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing.* (pp. ages 2383–2392). Austin, Texas. Association for Computational Linguistics.

Richardson, M., Burges, C. J. C., & Renshaw, E. (2013). MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 193–203). Association for Computational Linguistics.

Voorhees, E. M., & Tice, D. M. (2000). Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (SIGIR '00). (pp. 200–207). Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/345508.345577

Yang, Y., Yih, W.-t., & Meek, C. (2015). WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* (pp. 2013–2018). Lisbon, Portugal. Association for Computational Linguistics.