IIAI Open Conference Publication Series
IIAI Letters on Business and Decision Science
Vol. 005, LBDS346
DOI: https://doi.org/10.52731/lbds.v005.346

Proposal of a Haiku Evaluation Method Using Large Language Model and Prompt Engineering

Shunki Tomizawa [*], Soichiro Yokoyama [*],
Tomohisa Yamashita [*], Hidenori Kawamura [*]

Abstract

In this paper we describe the development of a haiku evaluation system using Large Language Model (LLM). We propose several prompting methods for haiku evaluation and selection, and verify the performance of the proposed methods using an automatically evaluable haiku dataset. We also performed haiku evaluation and selection on a large haiku database containing over 100 million verses using the proposed methods and validated their effectiveness through a questionnaire survey of haiku poets. The main contributions of this paper are as follows. First, we investigated the effectiveness of the procedures for demonstrating the validity of a number of haiku rating systems, including the creation of rating datasets and the results of subjective ratings through questionnaires. Second, we investigated methods for conducting haiku evaluation using LLM and prompt engineering.

Keywords: haiku evaluation, human evaluation, Large Language Model (LLM), prompt engineering

1 Introduction

The rapid development of artificial intelligence technology has opened up new possibilities for creative activities and artistic works. In traditional art fields such as painting, music and literature, works generated by deep learning models have attracted much attention. Discussions have intensified on how these works relate to human sensibility and creativity, and what new value they bring. In the generation of artistic works using deep learning models, the evaluation of these works is essential for improving their quality and is an integral part of the creative process. There are various types of work evaluation, such as classifying the genre to which the work belongs [1], generating explanatory or critical texts about the work [2], and selecting from multiple candidates [3].

[*] Hokkaido University, Sapporo, Japan

In this research, we focus on haiku in Japanese — a literary art form that has been cherished in Japan for centuries - and perform evaluation by selecting from multiple candidates using Large Language Model (LLM). Haiku is the world's shortest fixed-form poem, consisting of 17 syllables in a 5-7-5 pattern and including a seasonal word (*kigo*) that expresses seasonal scenes or emotions. Its purpose is to express and share with the reader the scenes or emotions the poet wants to convey in just 17 syllables. However, since a reader's impression of a haiku depends greatly on his or her own experiences, knowledge, and sensibilities, different readers often appreciate different haiku for different reasons, which may be different from what the author intended to convey. Because of these characteristics, judging haiku is difficult.

Here we focus on *kukai*, which is one aspect of haiku culture. A *kukai* is a gathering where haiku poets bring the haiku they have written and evaluate each other's work. In the evaluation of haiku, the ratings given by poets in *kukai* serve as an important benchmark for artificial intelligence-based evaluation. There are various types of *kukai*, but in this study we focus on one of the most common forms, where haiku are composed on a given theme. Participants submit haiku written on the given theme. It is important to determine whether artificial intelligence can appropriately interpret and evaluate the thematic content of haiku. Furthermore, the results of this research are expected to be applicable to the evaluation of various artistic works, which require an understanding of the author's creative intent embedded in the work.

Against this background, this research aims to select high quality haiku composed on a given theme from a large number of haiku candidates, and investigates methods to achieve this using LLM and prompt engineering. In addition, we validate the performance of the proposed methods using a haiku dataset evaluated by haiku poets and a questionnaire survey of haiku poets.

2 Related Works

Early poetry generation systems used rule-based methods. For example, a system was developed to automatically generate haiku from user-selected phrases, using seasonal words and dictionaries, and optimizing composition by scoring [4].

The advent of neural network-based language models has advanced automatic poetry generation in various languages, including English [5, 6], French [7], and Chinese [8], enabling high quality poems with controllable themes and rhythms. Zugarini et al. proposed a syllable-based neural model that mimics Dante's *Divine Comedy* [9]. Belouadi et al. introduced an end-to-end token-free model for poetry style generation [10], which learns poetic styles such as rhythm and rhyme directly from data.

In Japanese haiku generation, convolutional neural networks have been used to estimate haiku quality [11], and evaluation criteria have been established through surveys [12]. We evaluate haiku generated by LLM and those composed by humans, with the aim of verifying whether LLM can effectively evaluate and select haiku related to a given theme.

Advances in LLM such as GPT-2 [5], GPT-3 [13], and T5 [14] have shown high performance in NLP tasks. conversational LLM like ChatGPT [15] enable interactive communication with users and expand the application range of generative models. We are using GPT-4 [16] to tackle haiku evaluation.

The effective use of LLM has been enhanced by prompting methods such as zero-shot and few-shot learning [13], chain-of-thought prompting [17, 18], and techniques to im-

prove response accuracy [19, 20, 21, 22]. We propose a method for evaluating haiku by incorporating model-generated sentences into prompts using chain-of-thought prompting.

3 Types of Haiku and Mutual Evaluation by Humans

3.1 About Haiku

Haiku is considered to be the smallest form of poetry in the world and has been a popular form of poetry in Japan for over 600 years. This study focuses on the evaluation of the most common type of haiku, namely yuuki-teikei haiku in Japanese. According to the Japan Traditional Haiku Association, the requirements for a seasonal fixed-form haiku are twofold: it must be "composed of 17 syllables in a 5-7-5 pattern" and "contain a seasonal word (*kigo*)". Only a few works have deviated from these rules; however, these works are not appropriate as a first step in haiku evaluation and will not be considered in this study. In addition, the use of techniques such as *kireji* (cutting words) such as "ya" and "kana" or personification to convey the scenes observed or phenomena felt by the author within the limited 17 syllables is both the difficulty and the charm of haiku.

3.2 Evaluation of Haiku in *Kukai*

In haiku culture, poets bring the haiku they have written to meetings called *kukai* for mutual evaluation. At *kukai*, participants can critique each other's haiku and discuss with the authors, or vote for the haiku they think are good and share the reasons for their votes. Through these activities, they share each other's sensibilities and improve their own knowledge and skills. Submitting haiku to a *kukai* is called *toku* (submitting haiku), and selecting from the submitted haiku is called *senku* (selecting haiku). Since selection is generally done anonymously, the number of votes each haiku receives becomes a quantitative value that purely reflects the quality of the haiku in that *kukai*, without the influence of information about the author such as his or her haiku experience. Therefore, by using the number of votes as an indicator, it is possible to evaluate the performance of haiku evaluation and selection methods using LLM. Voting methods vary depending on the *kukai*. In this study, we make a qualitative evaluation of the haiku selected by the LLM based on *tokusen* (special selection), which is awarded to the haiku considered the best, and *namisen* (regular selection), in which several haiku are selected following the special selection.

Kukai can be divided into two types based on the restrictions placed on the haiku to be submitted. One is the type where a specific seasonal word, called *kendai*, is specified, and haiku containing that *kendai* are submitted. The other is the type where a theme, such as a particular situation or image, is specified, and haiku that fit that theme are submitted. In this study, as mentioned in Section 1, we assume a *kukai* in which haiku that fit the latter type of theme are submitted. We investigate methods for evaluating and selecting haiku that fit arbitrary themes from among several haiku using LLM. Because a *kukai* is typically attended by haiku poets belonging to a school of haiku, not all haiku poets will share the same assessment. However, the number of votes obtained here still provides important information for the haiku to be evaluated by haiku poets. It is important to confirm that the LLM can evaluate and select the haiku voted for by haiku poets.

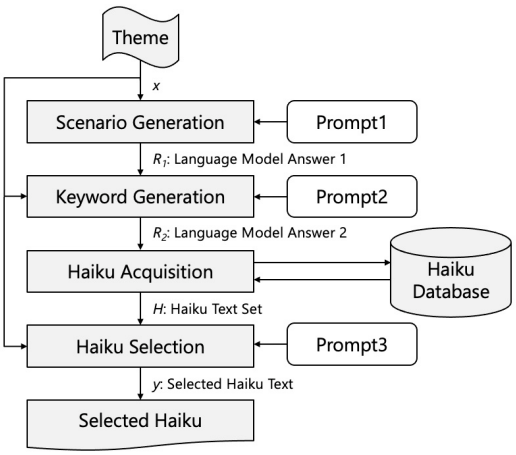

Figure 1: Overview of the proposed method. The haiku to be selected are in Japanese. The actual prompts are in Japanese, but are presented here in English translation.

4 Haiku Evaluation Method and Selection Algorithm

4.1 Overview of the Proposed Method

This section describes the proposed method for judging haiku. An overview of the method is shown in Figure 1. Given a theme text x, the method outputs a haiku y composed on the theme. The method consists of four modules: scenario generation, keyword generation, haiku acquisition, and step-by-step haiku selection. Each module is explained in detail below.

4.2 Scenario Generation Module

The scenario generation module uses an LLM to generate ten scenarios and their descriptions based on the theme. First, the given theme text x is inserted into the text prompt template T_1 to create the text prompt P_1. The prompt template T_1 is shown in Table 1. It consists of instructions for scenario generation, guidelines for step-by-step reasoning [13], and specifications for the output format. The text prompt P_1 is input to the LLM, and the pairs of generated scenarios and their descriptions $R_1 = \{s_{ij} \mid i = 1,2,\ldots,10; j = 1,2\}$ are output. Here, s_{i1} represents the i-th generated scenario text, and s_{i2} represents the description text of the i-th generated scenario.

Table 1: Prompt for the Scenario Generation Module. {request} is replaced with the theme text x.

Please generate 10 scenarios or situations that you think are suitable for composing haiku on the theme "{request}".
Please approach this task step-by-step. However, please follow the conditions below.
###Conditions
- After first showing your thoughts, output the following:
Scenario or Situation 1 and its description
Scenario or Situation 2 and its description
... (and so on up to 10)

4.3 Keyword Generation Module

The keyword generation module is a necessary component for acquiring haiku from the database described in the next section. This module generates 10 keywords related to the theme using the LLM. First, the output R_1 of the scenario generation module is inserted into the text prompt template T_2 to create the text prompt P_2. The prompt template T_2 is shown in Table 2. It consists of instructions for word generation, specifications for the output format, and guidance on the characteristics of the words to be generated. The text prompt P_2 is input to the LLM, and the generated keywords $R_2 = \{k_i \mid i = 1, 2, \ldots, 10\}$ are output.

Table 2: Prompt for the Keyword Generation Module. {request} is replaced with the theme text x, and {scenario} is replaced with the output R_1 of the scenario generation module.

From the theme "{request}", the following scenarios or situations can be imagined. {scenario}
Based on the above, please generate 10 words that you think are suitable for composing haiku on the theme "{request}".
However, please follow the conditions below.
###Conditions
- The output should be only "Word 1, Word 2, ..." (and so on up to 10)
- The words should be directly related to the theme

4.4 Haiku Acquisition Module

The haiku acquisition module acquires the haiku to be evaluated. This module supports acquisition from a pre-constructed haiku dataset and from a haiku database. These methods are explained below.

- **Acquisition from Haiku Dataset**

  In this method, the haiku to be selected are manually collected in advance to create text data. This module reads the text data and outputs it as is. Note that the keyword generation module is not used when acquiring haiku with this method.

- **Acquisition from Haiku Database**

  First, we explain the haiku database. The haiku included in the database are extracted from sentences generated by a language model trained on haiku data using Long Short-Term Memory (LSTM), a deep learning model. The haiku are selected based on conditions of seasonal fixed-form haiku, such as syllable counts (5-7-5) and inclusion of seasonal words (*kigo*) [23]. More than 100 million haiku are accumulated, and all haiku are annotated with evaluation scores estimated by the deep learning model. The quality of the haiku included in the database is explained in Section 5.2. haiku from the database are acquired using the words k_i $(i = 1, 2, \ldots, 10)$ output by the keyword generation module. Using each word k_i as input, haiku that include that word are acquired, and the resulting text data are output.

4.5 Step-by-Step Haiku Selection Module

The step-by-step haiku selection module performs sequential selection of haiku obtained by the haiku acquisition module. Depending on the number of haiku to be selected, they are divided into groups, and each group is input into the LLM for evaluation and selection. This method is explained in detail below.

Table 3: Prompt for the Evaluation and Selection Module when using Scenarios and Evaluation Scores. {(haiku, scores)} contains pairs of haiku to be evaluated and their evaluation scores. {selectionnum} contains the number of haiku to be selected by the LLM, determined by the function *h*.

The following scenarios or situations are considered suitable for composing haiku on the theme "{request}": {scenario} ### Request Please carefully evaluate each of the following haiku and select {selectionnum} haiku that are clearly related to the theme "{request}" and are likely to be rated highest by many people. Each haiku has scores for the following three evaluation items. The scores are on a four-point scale: 1 (Highly applicable), 2 (Applicable), 3 (Neither), 4 (Not applicable). When selecting haiku, please consider these scores. ### Evaluation Items Item 1: Is the meaning clear? Explanation 1: Is what is being said understandable or content that could realistically occur? Item 2: Is there empathy? Explanation 2: Can one empathize based on one's past experiences or thoughts? Item 3: Can the author's feelings be read? Explanation 3: Are the author's subjective emotions apparent? {(haiku, scores)} However, please follow the conditions below. ### Conditions - Write out the selected haiku one per line, and output only {selectionnum} lines. - Do not output any text other than the input haiku, such as reasons.

4.5.1 Division of Haiku to be Selected

In this section, we describe the method for dividing the haiku to be selected. In order to properly evaluate high quality haiku composed on a given theme using an LLM, it is necessary to consider the number of haiku input into the LLM at one time. In this study, based on the number of tokens given to the LLM at once and empirical rules from the authors, we set the maximum number of haiku input into the LLM at one time to 50. We then determined the number of haiku to divide and input into the LLM. Let N be the number of haiku to be selected; we define a function g that determines the number of haiku per group as

$$g(N) = \begin{cases} 50 & \text{if } N > 100, \\ 20 & \text{if } 100 \geq N > 50, \\ 10 & \text{if } 50 \geq N > 20, \\ 5 & \text{if } N \leq 20. \end{cases}$$

Based on function g, we define a function h that determines the number of haiku to be selected by the LLM from each divided group as

$$h(N) = \begin{cases} 10 & \text{if } N > 100, \\ 5 & \text{if } 100 \geq N > 50, \\ 2 & \text{if } 50 \geq N > 20, \\ 1 & \text{if } N \leq 20. \end{cases}$$

4.5.2 Evaluation and Selection of Haiku Using LLM

In this section, we describe the methods for evaluating and selecting haiku using an LLM. The evaluation methods are divided into four patterns based on differences in the prompts given to the LLM. These four evaluation methods are explained in detail below.

- **Evaluation Using a Simple Prompt**

  This method evaluates and selects haiku using a simple prompt that includes only the instructions for evaluating and selecting haiku and specifications for the output format.

- **Evaluation Using Evaluation Scores**

  This method uses a prompt that includes for evaluating and selecting haiku, specifications for the output format specifications, and the evaluation scores of the haiku to be selected. The evaluation scores are four-level ratings on three main items observed in generally well-regarded haiku, determined through interviews with haiku poets. The three evaluation items and their descriptions are as follows:

  - **meaningful**: Is the content understandable as a haiku?
  - **empathy**: Can you empathize based on one's past experiences or thoughts?
  - **author's feelings**: Are the author's subjective emotions apparent?

  The haiku are scored on a four-point scale for each evaluation item: 1 (Highly applicable), 2 (Applicable), 3 (Neither), 4 (Not applicable).

  The evaluation scores are generated for the haiku to be selected using the LLM. The input to the LLM is the haiku and a text prompt. This prompt is a few-shot CoT prompting consisting of the following three elements [24]:

  - Evaluation items and their meanings.
  - Three few-shot examples that include instructions to generate evaluation scores, reasoning processes, and output format specifications.
  - Instructions to generate evaluation scores.

  The few-shot examples are pairs of questions for evaluating haiku and sample outputs; the LLM outputs evaluation scores following the format of the sample outputs. Elements included in the sample outputs are seasonal words, morphological analysis, cutting words, and explanations of the haiku determined through interviews with haiku poets.

  By including the generated evaluation scores with the haiku to be selected and evaluating them using the LLM, we expect that incorporating the evaluation scores into the prompts will improve the accuracy of the LLM in highly rating haiku that poets consider to be good.

- **Evaluation Using Chain-of-Thought Prompting**

  This method uses a prompt that includes instructions for evaluating and selecting haiku, output format specifications, and the output text of the scenario generation module. By providing the output of the scenario generation module to the LLM, it is

expected that the LLM will be able to better interpret the meaning of the haiku to be evaluated and whether the haiku is composed on the given theme, and reflect this in the evaluation.

- **Evaluation Using CoT Prompting and Evaluation Scores**

  This method uses a prompt that includes instructions for for evaluating and selecting haiku, output format specifications, the output R_1 of the scenario generation module, and the evaluation scores. Table 3 shows an example of the actual prompt.

4.5.3 Haiku Selection Algorithm

The flow of the haiku selection module's processing, using the haiku division and selection functions explained earlier, is shown in Algorithm 1.

Algorithm 1 Haiku Selection Algorithm

Require: N, haiku, prompt
Ensure: Selected_haiku
  Selected_haiku $\leftarrow \emptyset$
  for $i = 1$ to $\left\lfloor \frac{N}{g(N)} \right\rfloor$ **do**
    group$_i \leftarrow \{\text{haiku}[g(N)\cdot(i-1)+1], \ldots, \text{haiku}[g(N)\cdot i]\}$
    Selected_haiku $\leftarrow$ Selected_haiku $\cup$ LLM(group$_i$, prompt, $h(N)$)
  end for
  newN $\leftarrow h(N) \times \left\lfloor \frac{N}{g(N)} \right\rfloor$
  if newN > 20 **then**
    Selected_haiku $\leftarrow f$(newN, Selected_haiku, prompt)
  end if
    return Selected_haiku

The function f in the algorithm is a recursive function defined as follows:

$$f(N, \text{haiku}, \text{prompt}) = \begin{cases} \bigcup_{i=1}^{\left\lfloor \frac{N}{g(N)} \right\rfloor} \text{LLM}(\text{group}_i, \text{prompt}, h(N)) & \text{if } N \leq 20 \\ f\left(h(N) \times \left\lfloor \frac{N}{g(N)} \right\rfloor, \bigcup_{i=1}^{\left\lfloor \frac{N}{g(N)} \right\rfloor} \text{LLM}(\text{group}_i, \text{prompt}, h(N)), \text{prompt} \right) & \text{otherwise} \end{cases}$$

Here, the function LLM represents the processing of the LLM that selects $h(N)$ haiku based on the given haiku and prompt. The algorithm 1 is explained as follows.

- **Step 1:** Initialize `Selected_haiku` as an empty set. This variable stores the haiku selected throughout the algorithm.
- **Step 2:** Divide the N haiku into $\left\lfloor \frac{N}{g(N)} \right\rfloor$ groups based on $g(N)$ and store them in `group`$_i$. Then, the LLM selects $h(N)$ haiku from `group`$_i$ and adds them to `Selected_haiku`. This operation is performed for all groups.
- **Step 3:** Calculate `newN`, the number of haiku selected by the LLM. If `newN` exceeds 20, recursively call the algorithm to divide into groups and select haiku using the LLM.

- **Step 4:** When the number of haiku `newN` becomes 20 or less, perform the grouping and haiku selection using the LLM one last time, return the results, and terminate the algorithm.

This algorithm will eventually select up to a maximum of four haiku.

5 Experiments

In this study, we conducted two experiments using the four haiku evaluation methods presented in Section 4.5.2 to verify the performance of the proposed method. In the first experiment, we created a dataset that includes haiku submitted to a contest that solicited haiku on a specific theme, and by using this dataset as the haiku to be evaluated, we verified whether the proposed method could select haiku composed on the theme. In addition, by examining the proportion of haiku considered excellent works by haiku poets among those selected by each method, we discussed trends in haiku quality between the methods. In the second experiment, we used haiku acquired from a database containing over 100 million entries as evaluation targets and verified whether the proposed method could select high quality haiku composed on the given theme. Furthermore, we conducted a questionnaire survey of haiku poets regarding the haiku selected by each method and the excellent works composed by humans to validate the performance of the proposed method.

5.1 Performance Validation of the Proposed Method using Haiku Dataset

5.1.1 Objective

The objective is to verify whether the proposed method can select high quality haiku that match the given theme. In order to conduct the verification, we need a dataset that includes haiku submitted in response to a specific theme and where the quality of the submitted haiku has been evaluated. Therefore, we collected haiku submitted to the Ehime Toyota haiku contest, which were planned by Marcobo.com Co., Ltd. and haiku unrelated to the theme submitted to the ”Fukushi Kukai” which also were planned by Marcobo.com Co., Ltd. to create a dataset of haiku to be evaluated. Ehime Toyota is a contest that invites haiku from beginners to haiku poets, regardless of their haiku experience, and the submitted haiku are evaluated by haiku poets who are also the chief editors of haiku magazines. Fukushi Kukai is held once a month and is attended by about 15 haiku poets with haiku experience ranging from 3 to 30 years. This dataset can be used to verify both whether the selected haiku are composed on the theme and whether they are high quality haiku. We validate the performance of the proposed method by analyzing the results of haiku selection using each proposed method on the created dataset.

5.1.2 Setup

We collected a total of 500 haiku submitted to the first to eighth editions of Ehime Toyota haiku contest. The 500 haiku included 8 Grand Prizes, 24 Excellence Awards, and 48 Honorable Mentions. In addition, we collected a total of 500 haiku not composed on the theme, submitted to the Fukushi Kukai. The haiku from the Fukushi Kukai were randomly sampled from haiku submitted in 14 sessions held between June 2023 and July 2024, in which the author participated, excluding haiku that appeared to be composed on the theme. Using the 1,000 haiku dataset, we collected 30 haiku selected at the final stage

of the selection algorithm for each evaluation method. The model used in the experiment was gpt-4o-2024-08-06 via the OpenAI API.

5.1.3 Results

For each method, we calculated the proportion of the 30 selected haiku that were composed on the theme, as well as the proportion of haiku that were prize-winning or higher, excellence award or higher, and grand prize, respectively. The results are shown in Table 4.

Table 4: Proportion of selected haiku for each method that meet each item regarding theme relevance and haiku quality.

	Haiku Composed on the theme	Prize-Winning or higher	Excellence Award or higher	Grand Prize
Simple	100%	30%	17%	0%
Evaluation Score	100%	47%	23%	50%
CoT	100%	20%	7%	0%
Evaluation Score + CoT	100%	20%	7%	13%

First, all four methods selected haiku composed on the theme at 100%. Next, looking at the items related to haiku quality, we can see that the method using evaluation scores achieved the best results in all three items. Among the 30 haiku selected by the method using evaluation scores, 14 were prize-winning or higher, and it was able to select 4 grand prize haiku, of which only 8 existed among the 1,000 haiku evaluated. Table 5 shows the average values of the three types of evaluation scores assigned to the haiku selected and not selected by the method using evaluation scores.

Table 5: Average evaluation scores assigned to the haiku selected and not selected by the evaluation method using evaluation scores.

	Is the Meaning Understandable?	Does It Have Empathy?	Can the Author's Feelings Be Read?
Selected Haiku	1.00	1.50	1.60
Non-Selected Haiku	1.80	2.03	2.23

Looking at Table 5, we can see that the selected haiku received better evaluation scores than the non-selected haiku in all items. Moreover, all evaluation scores of the selected haiku were either 1 (Highly applicable) or 2 (Applicable), whereas 25% of the evaluation scores of the non-selected haiku were 3 (Neither) or 4 (Not Applicable). By assigning evaluation scores to the haiku under evaluation and reflecting these scores during selection, it was found that the proportion of selecting high quality haiku that would receive votes in *kukai* increases. On the other hand, it was found that the two methods using CoT had a lower proportion of selecting high quality haiku compared to the other two methods that did not use CoT.

5.1.4 Discussion

Looking at the process of the selection algorithm, all four methods selected only haiku composed on the theme in the initial stage of the algorithm. Also, it was confirmed that in the methods using CoT, many of the selected haiku were about the scenarios generated by the scenario generation module or scenarios combining them. According to the author's subjective evaluation, among the 80 prize-winning or higher haiku used in this experiment, 26 were composed about the scenarios generated by the scenario generation module. Since

the methods using CoT tend to select haiku composed about the generated scenarios, they are more likely to prioritize selecting those haiku over excellent works, which is considered to be the reason why the proportions in the three items related to haiku quality in Table 4 were lower. By using the scenario module only in the initial stage of the selection algorithm and assigning evaluation scores at the stage where only haiku composed on the theme become the haiku to be evaluated, it is considered that high quality haiku composed on the theme can be selected more efficiently.

5.2 Performance Validation of the Proposed Method using Haiku Database

5.2.1 Objective

The objective is to verify whether the proposed method can select high quality haiku that match the given theme from haiku database described in Section4.4. The generated haiku in the database are annotated with scores representing haiku-like qualities estimated by the deep learning model, but a high score does not necessarily mean that the haiku is good from the haiku poet's point of view. It has been confirmed that among the top several tens to hundreds of haiku in the database, a few high quality haiku are included. However, as the number of haiku that can be manually verified is limited, it is important to evaluate and select the haiku in the database by the proposed method and to check the quality of the haiku in the database. From the above, in order to verify the performance of the proposed method and to check the quality of the haiku in the database, we conducted a questionnaire survey of haiku poets on the haiku selected by the proposed method. In addition, the haiku used in the questionnaire survey included the excellent works from Ehime Toyota explained in Experiment 1. This allows us to compare the quality of the haiku selected by the proposed method with excellent works composed by humans, thereby confirming both the performance of the proposed method and the quality of the haiku in the database.

5.2.2 Setup

We set the theme as "Life with Car" and selected 30 haiku from the database using each evaluation method. Ideally, we would evaluate all haiku in the database; however, evaluating approximately 120 million haiku is not realistic due to token limits and time constraints. Therefore, in this study, we used the 10 words generated by the keyword generation module. Among the haiku that include each of those words, we selected the top 200 haiku based on the evaluation values annotated to them and used them as the haiku to be evaluated. Thus, the total number of haiku to be evaluated was 2,000. For each evaluation method, we repeated the selection until 30 different haiku were collected. The model used in the experiment was gpt-4o-2024-08-06 via the OpenAI API. We also collected a total of 30 haiku consisting of 8 Grand Prize works and 22 Excellence Award works submitted to Ehime Toyota from the first to the eighth editions. The Excellent works were randomly sampled from 24 haiku.

In the questionnaire survey, for each of the 150 haiku, we asked questions regarding "Relevance to the theme" and "Quality of the Haiku." Tables 6 and 7 show the question items and response options for each.

In addition, since some haiku in the database may be grammatically incorrect or incomprehensible from the perspective of haiku poets, we asked them to answer the following three items as well.

Table 6: Question and response options regarding relevance to the theme.

Do you think it is related to the theme?
Strongly agree
Somewhat agree
Somewhat disagree
Strongly disagree

Table 7: Question and response options regarding haiku quality based on *kukai* selection.

Would you select it in a *kukai*?
Give a special selection (*tokusen*)
Give a regular selection (*namisen*)
Considered but not selected for regular selection
Not selected

- Is it a grammatically correct haiku?

  Two options: ”Correct,” ”Incorrect”

- Is it a meaningful haiku?

  Two options: ”Meaningful,” ”Meaningless”

- Are the phrases connected?

  Two options: ”Connected,” ”Not connected”

The respondents were eight haiku poets with haiku experience ranging from 3 to 15 years.
The aggregation method is as follows:

- **Relevance to the theme**: Proportion of haiku where the responses were ”Strongly agree” or ”Somewhat agree”
- **Regular selection candidate**: Proportion of haiku selected as candidates for regular selection
- **Regular or special selection**: Proportion of haiku selected as ”Regular selection” or ”Special selection”
- **Grammar**: Average score of all haiku when ”Grammatically correct” responses are given 1 point, and ”Incorrect” responses are given 0 points
- **Meaningful**: Average score of all haiku when ”Meaningful haiku” responses are given 1 point, and ”Meaningless haiku” responses are given 0 points
- **Phrase connection**: Average score of all haiku when ”Connected” responses are given 1 point, and ”Not connected” responses are given 0 points

5.2.3 Results

Table 8 shows the results of the questionnaire survey.

Table 8: Questionnaire results on theme relevance and haiku quality.

	Relevance to the theme	Candidate for Regular Selection	Regular or Special Selection	Grammar	Meaningful	Phrase Connection
Simple	75%	38%	6%	75%	65%	65%
Evaluation Score	75%	42%	10%	80%	72%	68%
CoT	62%	40%	9%	84%	75%	75%
Evaluation Score + CoT	66%	41%	10%	81%	70%	70%
Human	**95%**	**89%**	**69%**	**94%**	**93%**	**88%**

As shown in Table 8, the excellent works submitted to Ehime Toyota showed the best results in all items regarding theme relevance and haiku quality. Comparing the four evaluation methods, we found that the methods using simple prompts without the scenario generation module and the method using only evaluation scores selected haiku composed on the theme at a high rate. Since the scenarios generated by the scenario generation module are reflected in the keyword generation module, the keywords generated by methods using CoT often included words not directly related to the theme "Life with Car," such as "family" and "sunset." Upon examining the haiku acquired using such words, we found zero or only a few haiku composed on "Life with Car." Therefore, when using the scenario generation module, these results occurred because there were few haiku composed on the theme among the haiku to be evaluated. On the other hand, in items such as grammar, meaningfulness, and phrase connection, the method using CoT prompting was the best. This result suggests that by incorporating the content generated by the scenario generation module into the evaluation prompts, it becomes possible to select meaningful haiku that appropriately express these scenarios at a high rate.

5.2.4 Discussion

The human-created haiku targeted in this questionnaire are 30 excellent works selected from over 1,000 haiku submitted to Ehime Toyota, so they are considered to be among the top less than 1% in quality among the haiku included in the current database. Therefore, we found that in the current experimental setup, the total of 30 haiku selected did not reach the level of the excellent works of Ehime Toyota. However, we received comments from haiku poets who evaluate the haiku submitted to Ehime Toyota, stating that the haiku selected from the database using the proposed method are of comparable quality to the human-created haiku that narrowly missed winning prizes, although few haiku reach the level of the excellent works. Furthermore, the fact that there were extremely few haiku in the database that contained the keywords generated by the keyword generation module for the theme "Life with Car" is considered a reason for the inferior results compared to the human-created haiku. There are hundreds of thousands of haiku in the database that contain adjectives commonly used in haiku, such as "cool" and "lonely," and among the top 200 of these, there are several haiku of quality that could receive votes in a *kukai*, with some even receiving the highest number of votes. However, haiku containing words such as "steering wheel," "engine," and "speed"—generated in this experiment—numbered only a few thousand. According to the author's subjective evaluation, over 90% of the top 200 haiku were nonsensical, and there were hardly any haiku that would receive votes in a *kukai*. From these results, it became clear that there is a need to set multiple different themes and re-examine the quality of the haiku in the database and the performance of the proposed method.

On the other hand, among the selected haiku, there were several that received high evaluations from haiku poets, with comments that they were of comparable quality to excellent works. We present two of these haiku in Figure 2.

The haiku selected using the method with evaluation scores received a score of 1 (Highly applicable) in all three evaluation criteria and poetically depicted the scene of petals fluttering down in the silence after the engine's sound stops. The haiku selected using CoT prompting and evaluation scores not only had all evaluation scores as 1 (Highly applicable) but also described scenarios related to "drive" and "winter morning" generated by the scenario generation module. It splendidly captured the beauty of the scenery and a sense of

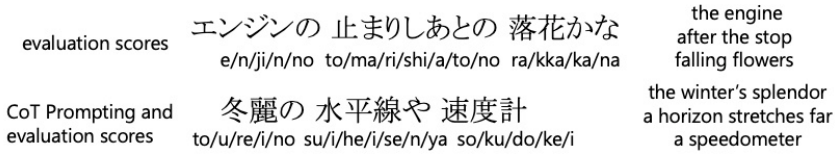

Figure 2: Examples of haiku used in the experiment. Each haiku was selected using the method with evaluation scores and the method using CoT prompting and evaluation scores.

speed.

6 Conclusion

In this study, we proposed a haiku evaluation method using LLM and prompt engineering and validated its performance through a created evaluation dataset and a questionnaire survey of haiku poets. In the haiku evaluation process, we confirmed that generating scenarios associated with the given theme and evaluation scores for the haiku, and reflecting them in the evaluation, increases the proportion of selecting high quality haiku composed on the theme. In addition, the results of the questionnaire survey indicated that by using haiku generated by deep learning model as evaluation targets, it is possible to select haiku that are, at a certain rate, close in quality to human-created haiku to a certain extent. In the future, by refining the evaluation method based on the results of this study, we aim to build an efficient and highly accurate haiku evaluation system.

References

[1] Chen Jiyang et al. *A Hybrid Parallel Computing Architecture Based on CNN and Transformer for Music Genre Classification*. Electronics 2024, 2024.

[2] Panos Achlioptas et al. *ArtEmis: Affective Language for Visual Art*. CVPR2021, 2021.

[3] Carlos Hernandez-Olivan and Jose R. Beltran. *Music Composition with Deep Learning: A Review*. arXiv, 2021.

[4] Naoko Tosa, Hideto Obara, and Michihiko Minoh. *Hitch Haiku: An Interactive Supporting System for Composing Haiku Poem*. ICEC2008, 2008.

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. Technical report, 2019.

[6] Brendan Bena and Jugal Kalita. *Introducing Aspects of Creativity in Automatic Poetry Generation*. ICON-2019, 2019.

[7] Mika Hamalainen, Khalid Alnajjar, and Thierry Poibeau. Modern french poetry generation with roberta and gpt-2. 2022.

[8] Xingxing Zhang and Mirella Lapata. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680, Doha, Qatar, oct 2014. Association for Computational Linguistics.

[9] Andrea Zugarini, Stefano Melacci, and Marco Maggini. *Neural Poetry: Learning to Generate Poems using Syllables*. ICANN2019, 2019.

[10] Jonas Belouadi and Steffen Eger. *ByGPT5: End-to-End Style-conditioned Poetry Generation with Token-free Language Models*. ACL2023, 2023.

[11] Shinji Kikuchi, Keizo Kato, Junya Saito, Seiji Okura, Kentaro Murase, Takaya Yamamoto, and Akira Nakagawa. *Quality Estimation for Japanese Haiku Poems Using Neural Network*. IEEE2016, 2016.

[12] Jimpei Hitsuwari, Yoshiyuki Ueda, Woojin Yun, and Michio Nomura. *Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry*. Elsevier Ltd., 2023.

[13] Tom B. Brown et al. *Language Models are Few-Shot Learners*. OpenAI, 2020.

[14] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Journal of machine learning research, 2019.

[15] *Introducing ChatGPT*. OpenAI, 2022.

[16] *GPT-4 Technical Report*. OpenAI, 2023.

[17] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. Advances in Neural Information Processing Systems, 2022.

[18] Mirac Suzgun, Nathan Scales, Nathanael Scharli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. *Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them*. ACL2023, 2023.

[19] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*.

[20] Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. *Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4*. arXiv, 2023.

[21] Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. *Reframing Instructional Prompts to GPTk's Language*. ACL2022, 2022.

[22] Xuezhi Wang, Jason Wei, Quoc Le Dale Schuurmans, Sharan Narang Ed Chi, Aakanksha Chowdhery, and Denny Zhou. *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. ICLR2023, 2023.

[23] Kodai Hirata, Soichiro Yokoyama, Tomohisa Yamashita, and Hidenori Kawamura. *Implementation of Autoregressive Language Models for Generation of Seasonal Fixed-form Haiku in Japanese*. KICSS2022, 2022.

[24] Shunki Tomizawa, Soichiro Yokoyama, Tomohisa Yamashita, and Hidenori Kawamura. *Proposal for a haiku evaluation mechanism using CoT Prompting with a large-scale language model*. JSAI2024, 2024.