Evaluation of LLM-Generated Distractors of Multiple-Choice Questions for the Japanese National Nursing Examination

Yûsei Kido[1], Hiroaki Yamada[1][a], Takenobu Tokunaga[1][b],
Rika Kimura[2][c], Yuriko Miura[2][d], Yumi Sakyo[2][e] and Naoko Hayashi[2][f]
[1]*School of Computing, Institute of Science Tokyo, Japan*
[2]*Graduate School of Nursing Science, St. Luke's International University, Japan*
{*kido.y.ad@m, yamada@c, take@c*}*.titech.ac.jp,* {*rikakimura, miura-yuriko, yumi-sakyo, naoko-hayashi*}*@slcn.ac.jp*

Keywords: Large Language Models, Japanese National Nursing Examination, Distractor Generation, Multiple-Choice Questions, Automatic Question Generation.

Abstract: This paper reports the evaluation results in the usefulness of distractors generated by large language models (LLMs) in creating multiple-choice questions for the Japanese National Nursing Examination. Our research questions are: "(RQ1) Do question writers adopt LLM-generated distractor candidates in question writing?" and "(RQ2) Does providing LLM-generated distractor candidates reduce the time for writing questions?". We selected ten questions from the proprietary mockup examinations of the National Nursing Examination administered by a prep school, considering the analysis of the last ten-year questions of the National Nursing Examination. Distractors are generated by seven different LLMs, given a stem and a key for each question of the above ten, and they are compiled into the distractor candidate sets. Given a stem and a key for each question, 15 domain experts completed questions by filling in three distractors. Eight experts are provided with the LLM-generated distractor candidates; the other seven are not. The results of comparing the two groups provided us with affirmative answers to both RQs. The current evaluation remains subjective from the viewpoint of the question writers; it is necessary to evaluate whether questions generated with the assistance of LLM work in a real examination setting. Our future plan includes administering a large-scale mockup examination using both human-made and LLM-assisted questions and analysing the differences in the responses to both types of questions.

1 INTRODUCTION

Automatic question generation (AQG) is one of the active research areas in Artificial Intelligence (AI) and is expected to reduce the burden on question writers in various education domains. There have been a series of comprehensive surveys on the AQG studies (Alsubait et al. 2015, Kurdi et al. 2020, Faraby et al. 2023). Alsubait et al.Alsubait et al. (2015) covered 81 papers published up to 2014 and reported language learning is the dominant as the target domain.

Kurdi et al. Kurdi et al. (2020) followed Alsubait et al. Alsubait et al. (2015) to collect and analyse 93 papers on AQG published from 2015 to 2019. The domain distribution is similar to that reported by Alsubait et al. Alsubait et al. (2015), i.e. the language learning domain remains dominant.

The studies covered by these two surveys before 2019 adopt traditional approaches: template-based, rule-based (Liu et al. 2010) or statistical-based (Kumar et al. 2015, Gao et al. 2019). The significant development of neural networks in the 2010s led Faraby et al. Faraby et al. (2023) to collect 224 neural network-based AQG papers published between 2016 and early 2022. Many of these studies utilise large datasets originally developed for Question-Answering (QA) systems, such as SQuAD (Rajpurkar et al. 2016), NewsQA[1] for training neural AQG systems.

After the appearance of ChatGPT[2] at the end of 2022, numerous large language models (LLMs) fol-

[a] https://orcid.org/0000-0002-1963-958X
[b] https://orcid.org/0000-0002-1399-9517
[c] https://orcid.org/0000-0001-9660-4471
[d] https://orcid.org/0009-0003-5270-603X
[e] https://orcid.org/0000-0001-9519-5792
[f] https://orcid.org/0000-0002-7058-692X

[1] https://www.microsoft.com/en-us/research/project/newsqa-dataset/

[2] https://chat.openai.com

Kido, Y., Yamada, H., Tokunaga, T., Kimura, R., Miura, Y., Sakyo, Y. and Hayashi, N.
Evaluation of LLM-Generated Distractors of Multiple-Choice Questions for the Japanese National Nursing Examination.
DOI: 10.5220/0013460300003932

In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 1*, pages 754-764
ISBN: 978-989-758-746-7; ISSN: 2184-5026

lowed. Their versatile and high performance for various tasks without fine-tuning greatly impacted academia and industry (Liu et al. 2023). AQG is also a potential application of LLMs. For instance, Perkoff et al. Perkoff et al. (2023) compared three types of LLM architectures, T5 (Raffel et al. 2020), BART (Lewis et al. 2020) and GPT-2 (Radford et al. 2019), in generating reading comprehension questions and concluded that T5 was the most promising. Yuan et al. Yuan et al. (2023) used GPT-3 to generate questions and chose better questions among automatically generated candidates. Oh et al. Oh et al. (2023) utilised LLM for paraphrasing references to improve the evaluation metrics for AQG. Shin and Lee Shin and Lee (2023) conducted a human evaluation of ChatGPT-generated multiple-choice questions (MCQs) for language learners, in which 50 language teachers evaluated a mixed set of human-made and ChatGPT-made MCQs without knowing their origins. They reported that both types of MCQs were of comparable quality.

We follow this line of research by utilising LLMs to generate MCQs. This research is a part of the project funded by the Ministry of Health, Labour and Welfare (MHLW) of Japan, which aims to automate administering the National Nursing Examination. Thus, our target domain is nursing; specifically, we aim to generate questions for the Japanese National Nursing Examination. Unlike the language learning domain, the challenge of this domain is the difficulty of human evaluation by domain experts. The quality of questions in this domain must be guaranteed at the national examination level, which is difficult by automatic evaluation. In addition, experts with experience in writing questions of National Nursing Examination are far fewer in number than language teachers; we have difficulty in recruiting expert evaluators. Kido. et al. Kido. et al. (2024) reported the feasibility study of using the LLM-generated questions for the Japanese National Nursing Examination. Our present study extends their preliminary evaluation by utilising LLM-generated questions in real-world question-writing settings involving domain experts. Following Kido. et al. Kido. et al. (2024), we focus on generating distractors of MCQs since it is a most burdensome task in question writing. Instead of fully automatic distractor generation, we take an approach to generate distractor candidates by LLMs and propose them to human question writers. The objective of this study is to evaluate the effectiveness and efficiency of distractor generation by LLMs for human question writing. To this end, we set up two research questions.

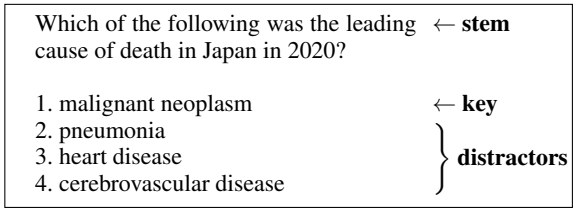

Figure 1: Example of the essential questions (Translation by authors).

RQ1. Do question writers adopt LLM-generated distractor candidates in question writing? (Effectiveness)

RQ2. Does providing LLM-generated distractor candidates reduce the time for writing questions? (Efficiency)

In the following, we briefly describe the Japanese National Nursing Examination and the analysis of the questions in the last ten-year examinations (section 2), then explain the experimental design (section 3). The experimental results and discussion (section 4) follow before the conclusion.

2 JAPANESE NATIONAL NURSING EXAMINATION

Registered nurses must pass the National Nursing Examination in Japan. Graduating from a college or university with a nursing curriculum is a prerequisite to the examination. The examination covers a wide range of subjects to confirm the knowledge about nursing from various perspectives.

The examination questions are in the form of MCQ and classified into three: essential, general, and situational. This study focuses on the essential questions as the first step of the project. Since they consist of simple recall-type questions asking important fundamental knowledge, generating their distractors would be a plausible task for LLMs. The subjects of the essential questions are organised into a three-level hierarchical structure consisting of 16 major subjects, 49 intermediate subjects and 252 minor subjects. The first column of Table 1 lists the 16 major subjects. The essential part consists of 50 questions that assess necessary basic knowledge. A score of 80% or higher on the essential questions is necessary to pass the examination. The questions are intended to check the examinees' knowledge of nursing and not to select examinees for a certain quota. Figure 1 shows an essential question example. A question consists of a stem (question sentences), a key (correct choice) and three or rarely four distractors (incorrect choices).

Table 1: Evaluation of the last ten-year essential questions of Japanese National Nursing Examination (Number of questions and percentages in parentheses).

Major subject \ Correct response rate (CRR) / Discrimination index (DI)	Question class I: [.90, .99) ≥ .2	II: ≥ .99 –	III: < .90 ≥ .2	IV: < .90 < .2	V: [.90, .99) < .2	Total
1. Health Indicators/ Definition and understanding of health	13 (34.2)	4 (10.5)	9 (23.7)	2 (5.3)	10 (26.3)	38
2. Health and life/ Factors affecting health and wellbeing	5 (18.5)	2 (7.4)	12 (44.4)	3 (11.1)	5 (18.5)	27
3. Basics of the health care/Social security system	6 (31.6)	1 (5.3)	5 (26.3)	0 (0.0)	7 (36.8)	19
4. Nursing ethics	2 (18.2)	2 (18.2)	3 (27.3)	1 (9.1)	3 (27.3)	11
5. Basic laws and regulations related to nursing	0 (0.0)	1 (9.1)	1 (9.1)	1 (9.1)	8 (72.7)	11
6. Characteristics of human beings	1 (10.0)	2 (20.0)	2 (20.0)	1 (10.0)	4 (40.0)	10
7. Human growth and development/ Characteristics of each period of the life cycle	11 (22.9)	6 (12.5)	19 (39.6)	2 (4.2)	10 (20.8)	48
8. Patients and families as nursing subjects	0 (0.0)	1 (33.3)	1 (33.3)	0 (0.0)	1 (33.3)	3
9. Major field of nursing and its functions	2 (8.3)	3 (12.5)	7 (29.2)	2 (8.3)	10 (41.7)	24
10. Structure and function of the human body	17 (32.1)	5 (9.4)	17 (32.1)	5 (9.4)	9 (17.0)	53
11. Pathology and nursing care/Diseases and signs	15 (19.7)	12 (15.8)	30 (30.3)	7 (9.2)	19 (25.0)	76
12. Pharmacokinetics, pharmacodynamics and therapeutics management	11 (36.7)	2 (6.7)	8 (26.7)	1 (3.3)	8 (26.7)	30
13. Basic nursing skills	2 (9.5)	6 (28.6)	5 (23.8)	0 (0.0)	8 (38.1)	21
14. Daily living assistance skills	6 (16.2)	11 (29.7)	2 (5.4)	1 (2.7)	17 (45.9)	37
15. Nursing skills to ensure patient safety and comfort	1 (4.0)	12 (48.0)	4 (16.0)	1 (4.0)	7 (28.0)	25
16. Nursing skills associated with medical treatment	8 (11.9)	19 (28.4)	14 (20.9)	9 (13.4)	17 (25.4)	67
Total	100 (20.0)	89 (17.8)	132 (26.4)	36 (7.2)	143 (28.6)	500

Table 2: Choice type distribution in the past ten-year essential questions.

Choice type	#Questions
Noun phrase	309
Sentence	57
Numerics	96
Figure & table	22
Exceptional questions	16
Total	500

The choices can be words or phrases like in Figure 1, longer descriptions in clauses and sentences, numerical values, graphs and tables. Table 2 shows the distribution of the choice types in the essential questions of the examinations over the past ten years, provided by MHLW, the body responsible for the National Nursing Examination. This study considers only questions with choices of words, phrases and sentences[3]. Although multi-modal LLMs have been actively studied recently, questions with figures and tables are not popular in past examinations. Questions with numerical choices are better suited to rule-based approaches, e.g. setting appropriate error offsets against the correct value will make reasonable distractors.

As we obtained the test takers' examination results of the past questions, we evaluate the 500 questions from the last ten years regarding the correct response rate (CRR) and discrimination index (DI). A high CRR value means that the question is easy, and a high DI value means that it can distinguish high- and low-ability test takers well. Based on our past experiences, we set the threshold for CRR and DI at 0.9 and 0.2, respectively. We consider questions with more than 0.9 (and less than 0.99) of CRR and more than 0.2 of DI "good questions" (Class I in Table 1) and others "questions to improve". The questions to improve are further classified into Class II, III, IV and V based on the CRR and DI values as shown in Table 1. For example, Class II is a too-easy question class. Table 1 indicates that major subjects 1, 3, 10 and 12 have relatively many good questions. In contrast, subjects 2, 9, 13, 14, 15 and 16 have a few.

[3]The first two types in Table 2

3 METHOD

In the experiments, given a stem and a key, human question writers are instructed to complete a question by providing three distractors. In the following, we may call completing a question by filling in the distractors "question writing". The question writers may refer to the distractor candidates generated by LLMs and adopt them, or they might create their original distractors. A set of distractor candidates is made from distractors generated by multiple different LLMs.

3.1 Question Writers

We ask five questions for each question writer. Considering the workload and time constraints of the question writers, five questions per person was the limit to recruit a sufficient number of question writers. On this condition, we recruited 16 question writers who have experience writing questions for the past Japanese National Nursing Examination. Half of them are novices with experience of less than five years, and the other half are veterans with more than or equal to five years of experience. One veteran question writer quit during the experiment; the actual number of participating question writers is 15. No author of this paper is included in these 15 participants. They are divided into two groups: Group A(ssisted) of eight writers that are provided with the LLM-generated distractor candidates, and Group C(ontrol) of seven writers without candidates. The writers in Group C must complete the questions by creating their original distractors. The participants are paid 3,000 JPY for the completion of five questions.

3.2 Materials

Questions

We selected ten questions from 250 essential questions in the mockup examinations of the National Nursing Examination administered by a prep school. These 250 questions are not open to the public. We obtained them under contract, together with the examination results. According to our analysis of the past National Nursing Examination in section 2, we first choose ten major subjects: 1, 3, 10 and 12, which include many good questions, and 2, 9, 13, 14, 15 and 16, which include a few good questions. We would like to see if providing LLM-generated distractor candidates contributes to further improvement for the former subject group and necessary improvement for the latter. We classify the prep school questions into five classes (I, II, III, IV and V) based on the same criteria adopted for the past National Nursing Examination analysis. A question from each major subject, 1, 3, 10 and 12, is randomly selected; a question from each major subject, 2, 9, 13, 14, 15 and 16, is randomly chosen to obtain ten questions in total. As a result, we have two questions from class I, seven from class III and one from class V.

Employed LLMs

We utilise the seven LLMs listed in Table 3 to generate distractor candidates. The first three are open-source models. Swallow-2 and Swallow-3 are made through continuously pre-training the Llama model with large Japanese corpora (Fujii et al. 2024, Okazaki et al. 2024). Their difference comes from the base Llama model, i.e. Llama2 and Llama3. MedSwallow-3 is a model in which Swallow-3 is further continuously pre-trained with Japanese medical texts (Iwasawa et al. 2024). These three models are fine-tuned using the questions from the last ten-year National Nursing Examination provided by MHLW and the mock examinations provided by the prep school mentioned above. The prep school questions do not overlap with our target questions. The total number of questions is 576, divided into 518 (90%) for training and 58 (10%) for development. The development set is used to decide training termination during fine-tuning. To save computational resources for fine-tuning, we adopt the QLoRA technique that introduces a low-rank matrix for parameter tuning and 4-bit quantisation of parameters (Dettmers et al. 2023). The available hyper-parameters for fine-tuning are set as follows: LoRA rank= 8, batch size= 1, learning rate= 10^{-4} and the number of epoches= 10. These values were empirically decided without an exhaustive hyperparameter search. We adopt the model with the minimum loss on the development dataset. The open models employ greedy decoding for inference.

Table 3: LLMs used to generate distractor candidates.

Model	Short name
1. Llama2-Swallow-70b-instruct-v0.1	Swallow-2
2. Llama3-Swallow-70B-Instruct-v0.1	Swallow-3
3. Llama3-Preferred-MedSwallow-70B	MedSwallow-3
4. GPT-3.5-turbo (0613)	GPT-3.5
5. GPT-3.5-turbo (0613) finetuned	GPT-3.5-FT
6. GPT-4 (0613)	GPT-4
7. GPT-4o (240513)	GPT-4o

The last four models are utilised through Microsoft Azure API. At the time of the experiment, fine-tuning was available only for GPT-3.5-turbo. We prepared two models for GPT-3.5-turbo, i.e. with and without fine-tuning. The available hyper-parameters of fine-tuning GPT-3.5-turbo are batch size and the number of epochs; they are set to 1 (default value) and 5 (maximum value), respectively. The training data for fine-tuning is the same as for the Swallow family. For inference, the temperature parameter is set to 0, and the top_p parameter is 0.95.

3.3 Procedure

Generating Distractor Candidate Sets

For each question, the above seven LLMs generate four distractors (28 in total) using the zero-shot

Zero-shot prompt

USER: Give us four distractors for the four-choice question "⟨stem⟩" with the correct answer "⟨answer⟩".

Five-shot prompt

USER: Give us four distractors for the four-choice question "⟨stem⟩" with the correct answer "⟨answer⟩".

ASSISTANT: Distractors:

- ⟨distractor$_1$⟩
- ⟨distractor$_2$⟩
- ⟨distractor$_3$⟩

— four more exemplars here —

USER: Give us four distractors for the four-choice question "⟨stem⟩" with the correct answer "⟨answer⟩".

Figure 2: Prompts to LLMs (Translation).

Table 4: Number of generated distractors by LLMs.

Q	Distractor	Duplicated	Candidate	Gold
1	13	4	6	3
2	28	0	14	0
3	25	2	11	1
4	20	8	10	0
5	24	4	10	1
6	27	1	13	0
7	18	6	9	0
8	25	3	11	0
9	8	5	5	1
10	27	1	13	0
Total	215	34	102	6
Ave.	21.5	3.4	10.2	0.6

prompt for the four fine-tuned models and the five-shot prompt for the three API models except for GPT-3.5-FT. As too many candidates would increase the cognitive load on the question writers, we decided to present about ten candidates to them. Assuming that the LLM-generated distractors would be further narrowed down, we decided to let the seven LLMs generate two to three times as many as the candidates to present, i.e. four distractors per LLM.

Figure 2 shows the translation of the prompts[4]. USER and ASSISTANT denote LLM user and LLM roles, respectively. The angle-bracketed word such as ⟨stem⟩ denotes a placeholder to fill with appropriate strings for the question before submitting it to the models. The exemplars for the five-shot prompts are randomly chosen from the training data.

The second and third columns in Table 4 show the type number of distractors ("Distractor") and those that are generated from multiple LLMs ("Duplicated") for each question. The average number of distractors and duplicated distractors are 21.5 and 3.4, respectively, suggesting that the models generate diverse distractors across all models.

To reduce the number of suggesting distractors to the question writers, starting from the LLM-generated distractor set, we follow the steps below to create distractor candidate sets.

1. We discard the distractors that are the same as the key for the question. There was one such distractor for the seventh question (Q7).
2. We collect the distractors generated by multiple models (The "Duplicated" column in Table 4).
3. We add distractors to the above collections so that every collection includes at least two distractors from each model. The insufficient distractors for a model are supplemented by randomly selecting the distractors generated by that model.

The resultant collections are the distractor candidate sets to provide the question writers. The column "Candidates" in Table 4 shows the number of distractors in the distractor candidate sets. The "Gold" column in Table 4 indicates the number of distractors that are the same as the original distractors of the question. The candidate sets include less than one gold distractor on average.

Assigning Questions to Question Writers

Group A and C of question writers work on the same ten questions. Each writer group is divided into two subgroups, each containing half novices and the other half veterans. The ten questions are divided into two, QS1 and QS2, and each subset is assigned to each subgroup. All five QS2 questions belong to class III, while the QS1 questions consist of two class I, two class III and one class V. Therefore, four question writers, two novices and two veterans in each subgroup (three for the five questions in Group C due to the participant withdrawal) work on the same five questions, QS1 or QS2.

Instruction to the Question Writers

We instruct the question writers to provide the appropriate three distractors for the given stem and key of five questions. The responses are collected through the Google Form platform because the participants are in distant locations. Group A is provided with a list of LLM-generated distractor candidates without details about the candidate generation process. They are just told that the distractor candidates are gener-

[4]The original prompts are in Japanese.

Table 5: Number of distractors per question (upper half; type) and per question writer (QW) (bottom half; token). The numbers in parentheses indicate those generated by LLMs.

QS	Q	Distractors	Adoption rate	Gold	Multi	Control	Valid (q6)
1	1	4 (4)	1.00	3 (3)	4 (4)	4 (4)	2
	2	10 (6)	0.60	0 (0)	2 (2)	0 (0)	4
	3	10 (5)	0.50	1 (1)	2 (2)	2 (2)	3
	4	11 (8)	0.73	0 (0)	1 (1)	2 (2)	4
	5	9 (7)	0.78	1 (1)	2 (2)	4 (2)	4
	Ave.	8.8 (6.0)	0.72				
2	6	10 (5)	0.50	0 (0)	1 (1)	0 (0)	3
	7	8 (1)	0.13	1 (0)	3 (0)	4 (0)	0
	8	10 (3)	0.30	0 (0)	2 (2)	0 (0)	0
	9	8 (5)	0.63	2 (1)	1 (1)	2 (1)	0
	10	12 (4)	0.33	0 (0)	0 (0)	0 (0)	2
	Ave.	9.6 (3.6)	0.38				
	Total	92 (48)	0.52	8 (6)	18 (15)	18 (11)	22
	QW	Distractors	Adoption rate	Gold	Multi	Control	Valid (q6)
	A1N1	15 (15)	1.00	2 (2)	10 (10)	3 (3)	6
	A1N2	15 (15)	1.00	3 (3)	5 (5)	5 (5)	1
	A2N1	15 (2)	0.13	2 (1)	4 (1)	4 (1)	1
	A2N2	15 (10)	0.67	1 (0)	6 (3)	3 (0)	1
	Ave.	15 (10.5)	0.70				
	A1V1	15 (4)	0.27	4 (4)	3 (3)	6 (4)	3
	A1V2	15 (12)	0.80	2 (2)	9 (9)	6 (6)	8
	A2V1	12 (8)	0.67	2 (1)	5 (4)	2 (1)	1
	A2V2	15 (3)	0.20	2 (1)	1 (1)	3 (1)	3
	Ave.	14.3 (6.8)	0.48				
	Total	117 (69)	0.59	18 (14)	43 (36)	32 (21)	24

ated by LLMs. Instead of adopting the candidates, they may provide their original distractors.

After providing three distractors, they are asked to answer the following questionnaire. The questions q2 to q5 should be answered by points on the five-point Likert scale, with one being "disagree" and five being "agree"; q6 is answered by ticking the checkbox for valid distractors in the list. The response formats are shown in square brackets.

q1 : How long did you need to create three distractors? [minutes]

q2 : The workload for creating distractors is heavier than that for the National Nursing Examination. [1–5]

q3 : The LLM-generated distractor candidates help create distractors. [1–5]

q4 : The LLM-generated distractor candidates are inspiring for brainstorming for question writing. [1–5]

q5 : The LLM-generated distractor candidates disturb your free thinking. [1–5]

q6 : Which LLM-generated distractor candidates were valid or were adopted? Choose all that apply. [List of the distractor candidates with checkbox]

Group C works on the same ten questions without the distractor candidates; they must create their original distractors. After providing three distractors, they are asked to answer q1 and q2 of the above questionnaire.

4 RESULTS AND DISCUSSION

4.1 RQ1: Effectiveness

Table 5 shows the number of distractors provided by Group A ("Distractor"), those that are the same as the original distractors of the question ("Gold"), those that are from multiple question writers ("Multi") and those that are the same as the distractors from Group C ("Control"). The numbers in parentheses correspond to the LLM-generated distractors. The last column ("Valid (q6)") indicates the number of LLM-

Table 6: Number of distractors in the candidate sets per LLM (The adopted numbers in parentheses).

Model	Distractors	Adoption rate	Gold
Swallow-2	23 (18)	0.78	4
Swallow-3	25 (17)	0.68	5
MedSwallow-3	22 (13)	0.59	4
GPT-3.5	23 (15)	0.65	3
GPT-3.5-FT	22 (11)	0.50	1
GPT-4	25 (11)	0.44	4
GPT-4o	26 (12)	0.46	4

generated distractors that were not adopted by the question writers but judged valid in the questionnaire (q6). The upper half of the table shows the question-wise type numbers, and the bottom half shows the question writer (QW)-wise token numbers. The naming convention of the question writers is as follows. The first letter indicates Group A or C, the second number indicates the question subgroup, QS1 or QS2, five questions each, the third letter indicates Novice or Veteran, and the last number is the identifier used to distinguish question writers with the same above attributes.

The bottom rows ("Total") indicate that 48 out of 92 (0.52; type in the upper half) and 69 out of 117 (0.59; token in the bottom half) distractors[5] in the completed questions come from LLMs. We generated 102 distractor candidates in total by LLMs as shown in Table 4, 48 of which (0.48) are adopted by the question writers. When we add the 22 "Valid" distractors to these adopted, the number goes up to 70 (0.69).

Among the adopted 48 LLM-generated distractors, 11 (0.22) overlap with the distractors created by the Group C writers ("Control") who do not refer to the LLM-generated candidates. These 11 LLM-generated distractors can be considered as high quality as those created by human experts. The rest, on the other hand, which are not thought of by the Group C writers, suggests that LLMs can generate novel distractors. The overlap between the 48 LLM-generated and gold distractors is also small, 6 out of 48 (0.13). These novel distractors have also been qualified by the Group A writers. These facts suggest that the LLM-generated distractors can assist question writing in terms of their quality and novelty.

Difference in LLMs

We generated the distractor candidate set by merging the outputs from seven different LLMs, which make 102 distractors in total (Table 4). Table 6 shows the number of distractors in the candidate sets presented to the Group A writers. The number of adopted is shown in the parentheses. Contrary to our expectations, Swallow-2, a rather older model, has the highest adoption rate. MedSwallow-3 trained with medical texts has a lower adoption rate than its base model, Swallow-3. The terminology between the medical and nursing domains might have some gaps. The GPT-4 family provides many distractors in the candidate set, but their adoption rates are worse than those of the GPT-3.5 family. This is another counter-expectation result.

[5]A2V1 could not complete a question; therefore, their total distractor number is less than 15.

Table 7: Relation of numbers between generating models and adopting question writers.

#QW							
4	0	0	0	0	1	0	0
3	0	1	0	1	1	0	1
2	5	3	1	0	1	0	0
1	21	10	0	0	0	0	2
0	42	11	0	0	1	0	0
#models	1	2	3	4	5	6	7

Among the candidate set, there are six "Gold" distractors that are the same as the original question's distractors as shown in Table 4. All six distractors were adopted by the question writers (the upper half of Table 5). Table 6 shows the number of generated "Gold" distractors for each LLM. Unlike the adoption rate, the models, except for the GPT-3.5 family, replicate the original distractors well. A high replication rate does not always lead to a high adoption rate, suggesting that only intrinsic evaluation using gold distractors is not sufficient for evaluating generated distractors.

There are duplicated distractors both in LLMs and in question writers. We investigate the relationship between these duplications. Table 7 shows the number of distractors that are generated by multiple models and adopted by multiple question writers. The x-axis indicates the number of models that generated a distractor, while the y-axis indicates the number of question writers who adopted the distractor. For instance, "10" in the cell (2, 1) means that there are ten distractors that were generated by two models and adopted by one question writer. The row "0" corresponds to the distractors that any question writers did not adopt. We can not see a strong correlation between the numbers of generating models and adopting question writers; a Pearson correlation coefficient is 0.47.

Difference in Questions

The adoption rate in the upper half of Table 5 varies depending on the questions, ranging from 0.13 to 1.00. We investigate the characteristics of the question for which LLM-generated distractors are likely to

Table 8: Average response values of questionnaire per question (upper half) and per question writer (QW) (bottom half) (SDs in parentheses).

QS	Q	Group	q1 Time	q2 Workload	q3 Helpful	q4 Inspiring	q5 Disturb	q6 Valid
1	1	A	4.0 (4.1)	1.0 (0.0)	4.8 (0.5)	4.5 (0.6)	1.8 (1.0)	2
		C	2.8 (1.7)	2.0 (1.2)				
	2	A	9.5 (6.4)	2.5 (1.3)	3.8 (1.9)	3.8 (0.5)	2.3 (1.3)	4
		C	8.3 (3.5)	3.3 (1.3)				
	3	A	5.0 (3.6)	2.3 (1.0)	4.5 (0.6)	4.3 (0.5)	1.5 (0.6)	3
		C	3.5 (1.9)	2.8 (1.0)				
	4	A	5.5 (6.4)	2.0 (1.2)	4.5 (0.6)	4.3 (0.5)	1.3 (0.5)	4
		C	3.5 (1.0)	2.5 (0.6)				
	5	A	5.5 (3.1)	2.3 (1.5)	4.8 (0.5)	4.5 (0.6)	1.8 (1.0)	4
		C	5.8 (3.0)	2.0 (1.2)				
	Ave.	A	5.9 (2.1)	2.0 (0.6)	4.5 (0.4)	4.3 (0.3)	1.7 (0.4)	3.4
		C	4.8 (2.3)	2.5 (0.6)				
2	6	A	8.3 (4.7)	1.8 (0.5)	3.8 (1.3)	2.8 (1.0)	2.8 (1.7)	3
		C	28.3 (12.6)	2.0 (1.0)				
	7	A	6.0 (4.1)	2.5 (1.0)	1.8 (0.5)	1.5 (0.6)	2.8 (1.3)	0
		C	33.3 (25.2)	2.7 (1.5)				
	8	A	8.5 (7.9)	2.5 (1.0)	2.5 (1.9)	1.5 (1.0)	2.3 (1.5)	0
		C	40.0 (17.3)	2.7 (1.5)				
	9	A	5.3 (4.2)	3.0 (1.6)	2.3 (1.9)	2.3 (1.9)	1.5 (1.0)	0
		C	38.3 (44.8)	3.0 (2.0)				
	10	A	12 (6.7)	2.5 (1.3)	3.5 (1.3)	2.0 (0.8)	2.3 (1.5)	2
		C	46.7 (37.9)	2.7 (1.5)				
	Ave.	A	8.0 (2.6)	2.5 (0.4)	2.8 (0.8)	2.0 (0.6)	2.3 (0.5)	
		C	37.3 (7.0)	2.6 (0.4)				
Ave.		A	7.0 (2.5)	2.2 (0.6)	3.6 (1.1)	3.1 (1.3)	2.0 (0.5)	2.2
		C	21.0 (17.9)	2.6 (0.4)				

QW	Group	q1 Time	q2 Workload	q3 Helpful	q4 Inspiring	q5 Disturb	q6 Valid
A1N1	A	3.4 (1.1)	1.2 (0.5)	4.6 (0.6)	4.6 (0.6)	2.2 (1.3)	6
A1N2	A	2.2 (0.8)	3.0 (1.2)	4.6 (0.6)	4.2 (0.5)	1.4 (0.9)	1
A2N1	A	14.0 (6.5)	2.6 (0.6)	2.2 (1.3)	1.8 (0.8)	1.4 (0.9)	1
A2N2	A	4.0 (2.5)	1.0 (0.0)	3.6 (2.0)	2.0 (1.7)	1.2 (0.5)	1
	Ave.	5.9 (5.5)	2.0 (1.0)	3.8 (1.1)	3.2 (1.5)	1.6 (0.4)	2.3
C1N1	C	5.0 (3.1)	1.8 (0.8)				
C1N2	C	3.6 (3.7)	3.0 (0.0)				
C2N1	C	68.0 (21.7)	3.8 (1.1)				
C2N2	C	27.0 (6.7)	3.0 (0.0)				
	Ave.	25.9 (30.0)	2.9 (0.8)				
Ave.		15.9 (22.7)	2.4 (1.0)				
A1V1	A	6.0 (5.2)	2.6 (0.9)	3.8 (1.6)	3.8 (0.5)	1.8 (0.5)	3
A1V2	A	12.0 (2.7)	1.2 (0.5)	4.8 (0.5)	4.4 (0.6)	1.4 (0.6)	8
A2V1	A	5.8 (3.0)	3.4 (1.1)	2.6 (1.5)	2.4 (1.1)	3.0 (1.2)	1
A2V2	A	8.4 (4.4)	2.8 (0.5)	2.6 (1.3)	1.8 (0.8)	3.6 (0.9)	3
	Ave.	8.1 (2.9)	2.5 (0.9)	3.5 (1.1)	3.1 (1.2)	2.5 (1.0)	3.8
C1V1	C	4.4 (3.1)	3.6 (0.9)				
C1V2	C	6.0 (2.2)	1.6 (0.6)				
C2V1	C	17.0 (8.4)	1.0 (0.0)				
	Ave.	9.1 (6.9)	2.1 (1.4)				
Ave.		8.5 (4.5)	2.3 (1.0)				

be adopted. We have two types of questions in terms of the choice type: noun phrases and sentences (Table 2); Q8 and Q10 in our question set have sentence choices, and the others have noun phrase choices. We calculate the micro-averaged adoption rate for these groups, obtaining 0.61 for noun phrase choices and 0.32 for sentence choices. As our data size is small, we can not draw a decisive conclusion; the difference suggests that noun-phrase candidates tend to be more adopted than sentence candidates.

Concerning the difference in question sets, QS1 and QS2, the average adoption rates are 0.72 for QS1 and 0.38 for QS2. Removing the outliers Q1 and Q7 reduces the difference, but they are still 0.64 (QS1) and 0.44 (QS2). We suspect the high adoption rate for QS1 comes from the mixture of question classes; QS1 consists of two class I, two class III and one class V. We need to collect more data regarding different characteristics to draw a decisive conclusion. A practical approach would be developing a usable tool in real settings that provides question writers with distractor candidates and then collecting question instances through its actual operation.

Difference in Question Writers

There is also a large variation in the adoption rate among the question writers, from 0.13 to 1.00 (the bottom half of Table 5). As with the questions, we investigate the characteristics of the question writers who likely adopt the LLM-generated distractors. An obvious feature is their degree of experience, i.e. novices vs veterans. We calculate the micro-averaged adoption rate for four novices and four veterans to obtain 0.70 for novices and 0.48 for veterans. The veterans tend to adopt less LLM-generated distractors. Less experienced question writers may be less confident in their own decisions and, therefore, more susceptible to the LLM suggestion. We then investigate the effect of the question sets in each group. In the novice group, the adoption rate is 1.0 for QS1 and 0.40 for QS2. In contrast, they are 0.53 and 0.43 in the veteran group. Again, the novices are more affected by the difference in question sets.

4.2 RQ2: Efficiency

Table 8 shows the macro-averaged response values of the questionnaire per question (upper half) and per question writer (bottom half). The "Time" column indicates completion time in minutes, and the columns "Workload" to "Disturb" are points on the five-point Likert scale, with one being "disagree" and five being "agree". The "Valid" column shows the number of distractors that were not adopted but considered valid. The numbers in parentheses denote the standard deviation. Responses to q3 to q6 are available only for Group A, as Group C was not provided the LLM-generated distractors.

Comparing the bottom two lines in the upper half of the table, providing distractor candidates reduces the average time to complete a question by a third, i.e. 21 to 7 minutes. Although the differences are slight, Group A's average workload values are smaller than Group C's, i.e. 2.2 vs 2.6. The average values of q3 and q4 exceed 3.0, meaning that the writers consider the LLM-generated distractors helpful for question writing and thinking of distractors. The average disturbance value of 2.0 suggests that the LLM-generated distractors do not disturb the writer's free thinking.

Difference in Question Sets

The upper half of Table 8 shows that, on average, Group A takes only a third of the time of Group C to complete a question. However, when we look at the differences for individual questions in the upper half of the table, Group A takes a longer or comparable time to complete questions in QS1 than Group C. The average time for QS1 is 5.9 for Group A and 4.8 for Group C, while that for QS2 is 8.0 and 37.3, respectively. The time reduction mainly comes from QS2. As in the analysis of effectiveness, the differences in the composition of question classes can be a reason. This tendency reversed for the workload score (q2). The average workload score for QS1 is 2.0 for Group A and 2.5 for Group C, while that for QS2 is 2.5 and 2.6, respectively. The difference between Groups A and C is more significant for QS1.

Differences in Experience

We calculated the average time for novices and veterans from the bottom half of Table 8, regardless of the group, to find that the novices took almost twice as much time (15.9) as the veterans (8.5). Furthermore, in the novice group, the Group C writers took 4.4 times as much time (25.9) as the Group A writers (5.9), whereas there is little difference between Group A (8.1) and C (9.1) in the veteran group. These differences suggest that concerning the question completion time, the LLM-generated distractor candidates impact more on less experienced question writers.

We did the same analysis on the workload scores (q2). The average workload scores for the novice and veteran groups are 2.4 and 2.3, respectively. The difference is smaller than that of the completion time. However, we have a different view on the availability of the LLM-generated candidates. The average scores

Table 9: Correlation between questionnaire responses and adoption rate.

	q1	q2	q3	q4	q5
ρ	−0.47	−0.41	0.67	0.59	−0.25

in the novice group are 2.0 for Group A and 2.9 for Group C, whereas in the veteran group, they are 2.5 and 2.1, showing less difference. The LLM-generated distractor candidates again impact more on less experienced question writers.

Questionnaire Response and Adoption Rate

We investigate the relationship between the adoption rate and the responses to the questionnaire. Using each question writer and each question as a single data point, the correlation between the responses to the questionnaire question and the adoption rate is calculated. Table 9 shows the Pearson correlation coefficients between the adoption rate and responses to q1 to q5 of the questionnaire. We observe weak or mild correlations between the adoption rate and question writers' subjective responses except for q5.

5 CONCLUSION

This paper evaluated the effectiveness and efficiency of LLM-generated distractors for question writing of the National Nursing Examination. To this end, we set two research questions: "RQ1: Do question writers adopt LLM-generated distractor candidates in question writing? (effectiveness)" and "RQ2: Does providing LLM-generated distractor candidates reduce the time for writing questions? (efficiency)". We conducted the experiment where 15 experts completed questions by filling three distractors, given a stem and a key for each question. Half of the experts were provided LLM-generated distractor candidates, and the other half were not. Half of them have more than or equal to five years of experience in writing the National Nursing Examination questions, and the rest have experience of less than five years. The results provided us with affirmative answers to both RQs, which aligns with the past research in a different domain, e.g. (Shin and Lee 2023). We also found that less experienced question writers are more susceptible to LLM-generated distractors. This experience bias raises a new research issue of the need for strict guidelines in the usage of LLM-generated distractors.

The present experiment has several limitations. First, the number of questions and question writers is limited. In the National Nursing Examination, we need 50 questions for the essential part. More than that number of questions must be created in the preparatin phase. Ten questions in our experiments are far fewer than those in the real examination. They do not cover all subjects introduced in section 2 either. In addition, we would like to have more question writers participating in the experiment. The present number of participants is not enough to draw a decisive conclusion in some aspects. However, as we noted in the introduction section, it is difficult to recruit many experts in our domain only for research purposes. One direction would be realising a LLM-based question writing support system and collecting data from the real question writing process.

Secondly, we focused on the essential questions in the National Nursing Examination in this work. However, the National Nursing Examination consists of three types of questions: essential, general and situational. The choices of the latter two types of questions could be more complicated. Therefore, further refinement of prompts for LLMs would be necessary. We plan to extend our target to the latter two in the succeeding project.

Finally, the present evaluation remains subjective from the viewpoint of question writers. Considering that the objective of the examination is assessing the test takers' knowledge, it is necessary to evaluate whether questions generated with the assistance of LLM work in knowledge assessment. We plan to administer a large-scale mockup examination that includes both LLM-assisted and human-made questions and conduct comparable analyses.

ACKNOWLEDGEMENT

This work is supported by Grant-in-Aid for Scientific Research and Health, Labour and Welfare Sciences, Grand Number 22AC1003.

REFERENCES

T. Alsubait, B. Parsia, and U. Sattler. Ontology-based multiple choice question generation. *KI - Künstliche Intelligenz*, 30, 11 2015. doi: 10.1007/s13218-015-0405-9.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *CoRR*, abs/2305.14314, 2023.

S. A. Faraby, A. Adiwijaya, and A. Romadhony. Review on neural question generation for education purposes. *International Journal of Artificial Intelligence in Education*, pages 1–38, 2023.

K. Fujii, T. Nakamura, M. Loem, H. Iida, M. Ohi, K. Hattori, H. Shota, S. Mizuki, R. Yokota, and N. Okazaki.

Continual pre-training for cross-lingual llm adaptation: Enhancing Japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, pages 1–25, University of Pennsylvania, USA, Oct. 2024.

Y. Gao, L. Bing, W. Chen, M. Lyu, and I. King. Difficulty controllable generation of reading comprehension questions. pages 4968–4974, 08 2019. doi: 10.24963/ijcai.2019/690.

J. Iwasawa, K. Suzuki, and W. Kawakami. Llama3 preferred medswallow 70b, 2024. URL https://huggingface.co/pfnet/Llama3-Preferred-MedSwallow-70B.

Y. Kido., H. Yamada., T. Tokunaga., R. Kimura., Y. Miura., Y. Sakyo., and N. Hayashi. Automatic question generation for the Japanese National Nursing Examination using large language models. In *Proceedings of the 16th International Conference on Computer Supported Education - Volume 1*, pages 821–829. INSTICC, SciTePress, 2024. ISBN 978-989-758-697-2. doi: 10.5220/0012729200003693.

G. Kumar, R. Banchs, and L. D'Haro. Automatic fill-the-blank question generator for student self-assessment. pages 1–3, 10 2015. doi: 10.1109/FIE.2015.7344291.

G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30:121 – 204, 2020.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

M. Liu, R. A. Calvo, and V. Rus. Automatic question generation for literature review writing support. In *International Conference on Intelligent Tutoring Systems*, 2010. URL https://api.semanticscholar.org/CorpusID:13917826.

Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, H. He, A. Li, M. He, Z. Liu, Z. Wu, L. Zhao, D. Zhu, X. Li, N. Qiang, D. Shen, T. Liu, and B. Ge. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017, 2023. ISSN 2950-1628. doi: https://doi.org/10.1016/j.metrad.2023.100017. URL https://www.sciencedirect.com/science/article/pii/S2950162823000176.

S. Oh, H. Go, H. Moon, Y. Lee, M. Jeong, H. S. Lee, and S. Choi. Evaluation of question generation needs more references. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6358–6367, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.396.

N. Okazaki, K. Hattori, H. Shota, H. Iida, M. Ohi, K. Fujii, T. Nakamura, M. Loem, R. Yokota, and S. Mizuki. Building a large Japanese Web corpus for large language models. In *Proceedings of the First Conference on Language Modeling*, COLM, pages 1–18, University of Pennsylvania, USA, Oct. 2024.

E. M. Perkoff, A. Bhattacharyya, J. Z. Cai, and J. Cao. Comparing neural question generation architectures for reading comprehension. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 556–566, 2023.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed: 2024-11-15.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), Jan. 2020. ISSN 1532-4435.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

D. Shin and J. H. Lee. Can ChatGPT make reading comprehension testing items on par with human experts? *Language Learning & Technology*, 27(3):27–40, 2023.

X. Yuan, T. Wang, Y.-H. Wang, E. Fine, R. Abdelghani, H. Sauzéon, and P.-Y. Oudeyer. Selecting better samples from pre-trained LLMs: A case study on question generation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12952–12965, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.820.