

# Large Language Model-Based Generation and Assessment of Radiological Technologist Licensure Exam Items in Japan

Toshimune Ito, Toru Ishibashi, Tatsuya Hayashi, Shinya Kojima, Kazumi Sogabe

Submitted to: JMIR Medical Education  
on: August 08, 2025

**Disclaimer:** © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

## *Table of Contents*

---

<b>Original Manuscript</b> .....	<b>5</b>
<b>Supplementary Files</b> .....	<b>29</b>
Figures .....	<b>30</b>
Figure 1.....	<b>31</b>
Figure 2.....	<b>32</b>
CONSORT (or other) checklists.....	<b>33</b>
CONSORT (or other) checklist 0.....	<b>34</b>

Preprint  
JMIR Publications

# Large Language Model-Based Generation and Assessment of Radiological Technologist Licensure Exam Items in Japan

Toshimune Ito<sup>1,2</sup> PhD; Toru Ishibashi<sup>1</sup> PhD; Tatsuya Hayashi<sup>1,2</sup> PhD; Shinya Kojima<sup>1,2</sup> PhD; Kazumi Sogabe<sup>1,2</sup> PhD

<sup>1</sup>Department of Medical Radiological Faculty of Medical Technology Teikyo University Tokyo JP

<sup>2</sup>Department of Medical Radiological Graduate School of Medical Technology Teikyo University Itabashi-ku JP

## Corresponding Author:

Toshimune Ito PhD  
Department of Medical Radiological  
Faculty of Medical Technology  
Teikyo University  
2-11-1 Kaga, Itabashi-ku,  
Tokyo  
JP

## Abstract

**Background:** For health professionals, mock examinations are an essential tool for assessing learning outcomes and reinforcing preparation for national licensure. However, traditionally written items are often inconsistent in difficulty, coverage, and clarity. Meanwhile, large language models (LLMs) have demonstrated high accuracy in medical exams and could potentially support the development of item banks, but the educational quality of artificial intelligence (AI)-generated questions remains underexplored.

**Objective:** To identify the most accurate LLM for the Japanese National Examination for Radiological Technologists and, using that model, to generate and evaluate blueprint-aligned multiple-choice questions.

**Methods:** Four LLMs, OpenAI o3, o4-mini, o4-mini-high (OpenAI), and Gemini 2.5 Flash (Google), were evaluated for their accuracy across all 200 items of the 77th Japanese National Examination for Radiological Technologists. The model with the highest accuracy (OpenAI o3) was then used to generate 192 multiple-choice items, adhering to the official blueprint. Expert reviewers rated these AI-generated items across five educational criteria: item difficulty, factual accuracy, content coverage, appropriateness of wording, and instructional usefulness. Statistical analyses were applied to compare model performance and item quality.

**Results:** OpenAI o3 achieved the highest accuracy overall (90.0%) and on non-image items (92.5%), significantly outperforming o4-mini on the full set ( $P = 0.0234$ ). Based on the expert reviewers' scores, the AI-generated items strongly performed in terms of item difficulty (4.29), factual accuracy (4.18), and content coverage (4.73), whereas significantly lower scores were seen for appropriateness of wording (3.92) and instructional usefulness (3.60) ( $P < 0.05$ ).

**Conclusions:** OpenAI o3 can generate radiological licensure questions aligned with national standards in terms of difficulty and content accuracy. There are limitations in wording clarity and pedagogical feedback, although these can be addressed through editorial review. This approach can facilitate efficient collaboration between AI and faculty in developing mock exams and also holds promise for supporting scalable, syllabus-aligned assessment in health professional education.

(JMIR Preprints 08/08/2025:81807)

DOI: <https://doi.org/10.2196/preprints.81807>

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

**Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [JMIR Publications](#), I will be able to make my accepted manuscript PDF available to anyone at any time.

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in [JMIR Publications](#), I will be able to make my accepted manuscript PDF available to anyone at any time.

Preprint  
JMIR Publications

**Original Manuscript**



# Large Language Model-Based Generation and Assessment of Radiological Technologist Licensure Exam Items in Japan

**Short Running Title:** LLM-generated Licensure Examination Items

Ito, T<sup>1\*,2</sup>., Ishibashi, T<sup>1</sup>., Hayashi, T<sup>1,2</sup>., Kojima, S<sup>1,2</sup>., Sogabe, K.<sup>1,2</sup>

<sup>1</sup>Department of Medical Radiology, Faculty of Medical Technology, Teikyo University, 2-11-1 Kaga, Itabashi-ku, Tokyo, 173-8605, Japan

<sup>2</sup>Department of Medical Radiology, Graduate School of Medical Technology, Teikyo University, 2-11-1 Kaga, Itabashi-ku, Tokyo, 173-8605, Japan

**\*Corresponding Author:** Toshimune Ito

Department of Medical Radiology, Faculty of Medical Technology, Teikyo University, 2-11-1 Kaga, Itabashi-ku, Tokyo, 173-8605, Japan

Tel: 03-3964-7053; Fax: 03-3964-7052; Email: [toito@med.teikyo-u.ac.jp](mailto:toito@med.teikyo-u.ac.jp)

## Abstract

**Background:** For health professionals, mock examinations are an essential tool for assessing learning outcomes and reinforcing preparation for national licensure. However, traditionally written items are often inconsistent in difficulty, coverage, and clarity. Meanwhile, large language models (LLMs) have demonstrated high accuracy in medical exams and could potentially support the development of item banks, but the educational quality of artificial intelligence (AI)-generated questions remains underexplored.

**Objective:** To identify the most accurate LLM for the Japanese National Examination for Radiological Technologists and, using that model, to generate and evaluate blueprint-aligned multiple-choice questions.

**Methods:** Four LLMs, OpenAI o3, o4-mini, o4-mini-high (OpenAI), and Gemini 2.5 Flash (Google), were evaluated for their accuracy across all 200 items of the 77th Japanese National Examination for Radiological Technologists. The model with the highest accuracy (OpenAI o3) was then used to generate 192 multiple-choice items, adhering to the official blueprint. Expert reviewers rated these AI-generated items across five educational criteria: item difficulty, factual accuracy, content coverage, appropriateness of wording, and instructional usefulness. Statistical analyses were applied to compare model performance and item quality.

**Results:** OpenAI o3 achieved the highest accuracy overall (90.0%) and on non-image items (92.5%), significantly outperforming o4-mini on the full set ( $P = 0.0234$ ). Based on the expert reviewers' scores, the AI-generated items strongly performed in terms of item difficulty (4.29), factual accuracy (4.18), and content coverage (4.73), whereas significantly lower scores were seen for appropriateness of wording (3.92) and instructional usefulness (3.60) ( $P < 0.05$ ).

**Conclusions:** OpenAI o3 can generate radiological licensure questions aligned with national standards in terms of difficulty and content accuracy. There are limitations in wording clarity and pedagogical feedback, although these can be addressed through editorial review. This approach can

facilitate efficient collaboration between AI and faculty in developing mock exams and also holds promise for supporting scalable, syllabus-aligned assessment in health professional education.

### Keywords

large language models; mock examination; licensure exam; radiological technologists; educational evaluation; OpenAI; medical education; item generation; AI-assisted learning; visual reasoning



## Introduction

Mock examinations are a key pedagogical tool in training programs for health professionals. These are designed to consolidate the knowledge required for national licensure and to gauge students' achievement [1-3]. In particular, multiple-choice formats are valuable because they enable the systematic, efficient appraisal of the broad foundational knowledge expected in clinical practice, making them integral to the quality of the curriculum. However, most items are written by individual instructors that draw on past examinations or personal clinical experience, and their difficulty and content validity are rarely subjected to systematic review [4,5]. These can result in biases in content coverage, inconsistencies in wording, and variable educational usefulness, which undermine the stability of learning outcome assessments.

Several studies have reported the high accuracy of large language models (LLMs) in health professional licensure examinations, owing to their rapid advancements [6-9]. In text-based multiple-choice questions, models have begun to match or surpass human test-takers while generating rationales and keyword-level explanations that can serve as formative feedback [10-13]. These suggest the potential utility of LLM-assisted item writing during the construction of high-quality question banks. However, most research has centered on the accuracy of LLMs in answering existing licensure items [14-16], while empirical evidence regarding the educational quality of questions authored by LLMs remains scarce [13,17]. A comprehensive appraisal that includes (i) appropriate difficulty, (ii) completeness and accuracy of content coverage, (iii) clarity of option wording, and (iv) usefulness of accompanying explanations is necessary to address this knowledge gap and clarify the practical value of AI-supported mock examinations, as well as its limitations.

This study evaluated the quality of AI-generated multiple-choice questions based on the Japanese National Examination for Radiological Technologists. Several LLMs were used to answer the exam, then the highest-performing model was used to generate a set of mock items. These AI-generated questions were then evaluated across several aspects (i.e., item-level difficulty, item-level factual accuracy, accuracy of content coverage, appropriateness of wording, and instructional usefulness) through blinded expert review and statistical analysis. By doing so, this study aims to provide empirical data on the educational soundness of AI-generated items, as well as highlight any emerging challenges.

## Methods

### 1. Models and Study Period

Four LLMs released in February 2025 were evaluated: OpenAI o3, OpenAI o4-mini, OpenAI o4-mini-high (all OpenAI), and Gemini 2.5 Flash (Google). The evaluations were done from March 14 to May 8, 2025 using the publicly accessible browser versions, and the desired engine was explicitly selected via the menu of each platform. The item generation study was conducted from May 15 to June 28, 2025 using OpenAI o3, the model with the best answer accuracy. All prompts for accuracy testing and item generation followed an identical template; these were supplied to every model in the same wording in Japanese to ensure consistency. Browsing functions were disabled, such that the outputs relied solely on the pretrained parameters and provided materials.

### 2. Answer Accuracy

Answer accuracy was assessed based on all 200 items of the 77th National Examination

for Radiological Technologists, administered on 20 February 2025. All items were multiple-choice, and question stems containing images were presented unchanged. Each model was given the question stem and options in Japanese, then instructed to select the correct answer(s) in single-best or multiple-select format. Table 1 lists the subjects and the number of items per subject. Due to the differences in each model, the input procedures were adapted accordingly. For the OpenAI models, the 200 items were split into four text files containing 50 items each, and image-based stems were accompanied by PNG files labeled with the item number. However, since Gemini permits only one file upload, the stems and options were pasted directly into the prompt while attaching an image file as needed. All inputs were entered manually. A representative interaction is shown in Figure 1.

Table 1 Breakdown of the 2025 Japanese National Exam Questions by Subject

Subject	Questions	Composition	Items w/ Images
	(n)	(%)	(n)
Diagnostic Imaging	20	10	12
Techniques Nuclear Medicine	20	10	1
Technology Radiation Therapy	20	10	1
Technology Medical Imaging Informatics	10	5	1
Healthcare Safety	8	4	0
Management Basic Medical Sciences	30	15	0
Radiation Science & Engineering	36	18	2
X-ray Imaging Equipment	20	10	0

X-ray Imaging Techniques	20	10	9
Image Engineering	6	3	1
Radiation Safety			
Management	10	5	0
Total	200	100	27

The outputs of the model were compared to the official answer key issued by the Ministry of Health, Labour and Welfare ([https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou\\_iryuu/iryuu/topics/tp250428-06.html](https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryuu/iryuu/topics/tp250428-06.html)). The correct and incorrect responses were counted overall for 200 items and separately for the 173 items that did not require image interpretation (i.e., non-image items). Statistical significance was tested across models.

### 3. Item Generation

#### 3.1 Generation Procedure

The mock items were generated using OpenAI o3, since it had the highest accuracy among all four models. Image-based stems were excluded, since all models performed poorly on these. Using the same examination as a blueprint, OpenAI o3 was used to produce 192 questions across 14 subjects (Table 2), matching the same distribution of items. The model was supplied with text files containing the past 5 y of examination items and the official test specifications, ensuring its alignment with test objectives. Browsing remained disabled. Since Healthcare Safety Management is a new domain introduced in 2025, thereby lacking any historical reference items, it was excluded from the mock item generation. Items were generated separately for each subject in Japanese, and each output included the stem, five options, the key, and a brief rationale.

Table 2 Distribution of Artificial Intelligence (AI)-Generated Mock Items

<b>Subject</b>	<b>Blueprint Target (n)</b>	<b>Generated (n)</b>
Diagnostic Imaging Techniques	20	20
Nuclear Medicine Technology	20	20
Radiation Therapy Technology	20	20
Medical Imaging Informatics	10	10
Basic Medical Sciences	30	30
Radiation Science & Engineering	36	36
X-ray Imaging Equipment	20	20
X-ray Imaging Techniques	20	20
Image Engineering	6	6
Radiation Safety Management	10	10
Healthcare Safety Management <sup>†</sup>	8	0
<b>Total</b>	<b>200</b>	<b>192</b>

<sup>†</sup>Since Healthcare Safety Management was only recently introduced as a new subject in the 2025 blueprint, it was excluded from the mock item generation

### 3.2 Evaluation of Generated Items

All 192 generated questions were reviewed by experts of the subject matter; these were faculty members with at least 5 y of experience as subject coordinators in radiological technology programs and who routinely author mock examinations. Items were assigned to reviewers by discipline, and each question was evaluated by one expert. The reviewers used a five-point scale (1 = very inappropriate, 5 = highly appropriate) to rate each item based on item difficulty, factual accuracy, accuracy of content coverage, appropriateness of wording, and instructional usefulness. For each criterion, we calculated the median score and tested the statistical significance of the proportion of high ratings ( $\geq 4$ ). The evaluation framework, which is based on faculty experience with national examination item writing, is presented in Table 3.

Table 3 Evaluation of generated items

<b>Question No.</b>	<b>Evaluation Criterion</b>	<b>Rating Scale*</b>
1	Item difficulty	1–5

2	Factual accuracy	1–5
3	Accuracy of content coverage	1–5
4	Appropriateness of wording	1–5
5	Instructional usefulness	1–5

\*Rating scale definition: 1 = unacceptable; 2 = Major revision needed; 3 = Revisable; 4 = Minor revision; 5 = Adoptable.

#### 4. Statistical Analysis

Statistical analysis was performed using JMP® 18 (JMP Statistical Discovery LLC, Cary, NC, USA). Cochran's Q test was initially used to examine overall differences in answer accuracy; when significant, pairwise differences were probed with McNemar test using Bonferroni correction. The item generation study used a one-sided Wilcoxon signed-rank test ( $H_0$ : median  $\leq 4$ ). Statistical significance was set at  $P < 0.05$  for all analyses.

### Results

#### 1. Answer Accuracy

The accuracy of the LLMs on the full 200-item set and the non-image 173-item set is shown in Table 4. All models consistently scored lower in the full set versus the non-image set, with OpenAI o3 achieving the best results at 90% and 92.5%, respectively. A significant difference was seen between OpenAI o3 and OpenAI o4-mini on the full set, whereas no significant differences were seen among models on the non-image set.

Table 4 Model Accuracies and Statistical Comparisons on 200 Benchmark Questions and 173 Non-Image Questions

	200 questions	173 non-image
--	---------------	---------------

	<b>questions</b>	
Model accuracy (%)		
OpenAI-o4-mini-high	86.0 (80.5, 90.1)	88.4 (82.8, 92.4)
OpenAI-o4-mini	82.5 (76.6, 87.1)	86.7 (80.8, 91.0)
OpenAI-o3	90.0 (85.1, 93.4)	92.5 (87.6, 95.6)
Gemini 2.5 Flash	83.0 (77.2, 87.6)	89.6 (84.1, 93.3)
Cochran's Q test ( <i>P</i> -value)	0.0138*	0.1007
Pairwise McNemar test		
(Bonferroni-adjusted p-value)		
OpenAI-o4-mini-high vs OpenAI-		
o4-mini	1.00	NA
OpenAI-o4-mini-high vs OpenAI-		
o3	0.442	NA
OpenAI-o4-mini-high vs Gemini		
2.5 Flash	1.00	NA
OpenAI-o4-mini vs OpenAI-o3	0.0234*	NA
OpenAI-o4-mini vs Gemini 2.5		
Flash	1.00	NA
OpenAI-o3 vs Gemini 2.5 Flash	0.064	NA

Values in parentheses represent 95% confidence intervals. An asterisk denotes statistical significance ( $P < 0.05$ ). NA, not applicable.

## 2. Item Generation

Table 5 presents the scores and statistics for all 192 questions, while Figure 2 illustrates the prompt template and sample outputs. Among item difficulty, factual accuracy, and accuracy of content coverage, the medians and the proportions of scores  $\geq 4$  did not differ significantly, although accuracy of content coverage had the highest score. Meanwhile, instructional usefulness had a significantly lower score than appropriateness of wording.

Table 5 Evaluation Scores for Survey Questions

Questio n	Mean Score (95 % CI)	P-Value†
1	4.29 (4.11, 4.46)	<0.0001 *
2	4.18 (3.98, 4.38)	0.0010 *
3	4.73 (4.60, 4.86)	<0.0001 *
4	3.92 (3.73, 4.11)	0.4453
5	3.60 (3.41, 3.80)	0.9989

Values in parentheses represent 95% confidence intervals. An asterisk denotes statistical significance ( $P < 0.05$ ).

† One-sided Wilcoxon signed-rank test against the null hypothesis that the median score is  $\leq 4$ .

## Discussion

This study compared four LLMs in terms of answer accuracy on the Japanese National Examination for Radiological Technologists. The top performer, OpenAI o3, was used to generate the mock test, which was then evaluated by experts in terms of educational quality. As shown in Table 4, OpenAI o3 achieved the highest overall accuracy of 90.0% (95%CI 85.1%–93.4%) across all 200 questions, which differed significantly only from OpenAI-o4-mini after Bonferroni correction ( $P = 0.0234$ ). However, after excluding 27 image-based items, the differences in accuracy among models disappeared (Cochran Q,  $P = 0.1007$ ), which could be attributed to variations in multimodal reasoning capability.

Built on a GPT-4 lineage, OpenAI o3 integrates a CLIP-style visual encoder with high-resolution patches and is pretrained on a diverse visual corpus that includes medical images and employs unified self-attention across linguistic and visual tokens [18,19]. These features seem to aid in capturing low-contrast regions and anatomical patterns characteristic of computed tomography and radiography, limiting the accuracy drop

attributable to image items to just 2.5%. In contrast, OpenAI o4-mini is a lightweight variant designed for speed and lower cost. Its visual module uses simplified patch embeddings at reduced resolution, and its pre-training presumably involves fewer images [20,21]. This results in coarse visual representation, which can overlook subtle diagnostic cues, thereby resulting in a 4.2% decline on image items and the only statistically significant gap relative to OpenAI o3. Meanwhile, OpenAI o4-mini-high supplements the mini architecture with targeted fine-tuning for medical images and partial recovery of high-resolution inputs [22,23]. This additional training appears to improve visual mapping, and as a result, the image-item penalty is only 2.4%, with no significant difference versus OpenAI o3 ( $P = 0.442$ ). Lastly, Gemini 2.5 Flash employs a two-tower design, wherein an external vision encoder converts images to textual tags before language processing [24,25]. The sparse tag vocabularies and missing anatomical labels in medical images can hinder the extraction of image-specific cues. Although the overall accuracy of Gemini did not differ significantly from OpenAI o3, it still scored 6.6% lower on image items.

The pronounced performance spread on image-based questions could be mainly attributed to the aggressive parameter reduction in OpenAI o4-mini and the information loss inherent in the image-to-tag pipeline in Gemini, both of which weaken visual feature representation. However, OpenAI o3 and o4-mini-high have higher resolution encoders and benefit from medical-specific fine-tuning. However, due to the limited sample sizes and proprietary nature of the detailed model architectures, these explanations remain partly hypothetical. Nevertheless, these findings highlight the importance of the visual module scale and the presence of medical-domain training when selecting an LLM for the development of AI-generated questions in this field.

Building on these findings, the 192 items generated by OpenAI o3 were rated by subject matter experts across five educational criteria (Table 5). Item difficulty had an average score of 4.29, indicating good alignment with the national examination standard. Factual accuracy averaged 4.18, mirroring the high answer accuracy of the model. The accuracy of content coverage was scored 4.73, indicating that the large context window of the model close adhered to the official syllabus. In line with this, AI-generated questions for the National Medical Licensing Examination have demonstrated similar strengths [26]. In contrast, relatively lower scores were seen for appropriateness of wording (3.92) and instructional usefulness (3.60), highlighting the weaknesses of unedited LLM output in terms of ambiguous phrasing and overly generic feedback. These findings are also consistent with earlier studies [26-28]. In summary, OpenAI o3 appears capable of rapidly producing syllabus-aligned item frameworks with an appropriate level of difficulty, although editorial refinement by faculty is needed to further polish these drafts into a high-quality mock-exam bank.

This study has several limitations. First, the image-based items were excluded from expert review, thus precluding the assessment of visual tasks. Second, each question was evaluated by a single expert, and thus inter-rater reliability could not be assessed. Third, the LLMs used are cloud-based services that can change in terms of internal weights and inference algorithms without notice. Even with the same version label and prompt, future runs could result in wording differences, thereby threatening reproducibility [29-31].

In the future, visual encoders are expected to operate at a higher resolution and undergo additional tuning for medical domains. This could enable LLMs to automatically generate image-based items across modalities (e.g., computed tomography, magnetic

resonance imaging, and ultrasound), thus bringing mock exams closer to clinical reality. Further improvements in the feedback system could also be seen. By delivering adaptive feedback that varies in depth according to each learner's proficiency, students can be provided with on-demand, targeted remediation material. LLMs could also be used to map items to the national blueprint in real time, enabling the detection and correction of domain imbalances while reducing faculty workload. Lastly, aligning these models with overseas licensure frameworks could expand their use to ultimately support a multilingual, multi-profession, international mock-exam bank.

## Conclusions

This study demonstrated that an LLM (OpenAI o3) can attain high accuracy on national radiological technology examination, as well as generate new multiple-choice items with appropriate difficulty, factual correctness, and syllabus coverage, as evaluated by experts. Although the AI-generated questions fell short in terms of wording clarity and pedagogical feedback, these can be mitigated through targeted editorial review. Practically speaking, LLMs can be used to draft content that is eventually refined by the faculty. This workflow could enable the more efficient development of mock examinations and reinforce curriculum alignment without imposing additional burden on instructors. However, performance gaps on image-based items, the absence of inter-rater reliability data, and the inherent volatility of cloud-hosted models underscore the need for cautious implementation and transparent reporting of model metadata. Nevertheless, future advancements in high-resolution visual encoders and medical-specific tuning can close this multimodal gap, while adaptive feedback functions and automated blueprint mapping can further extend the educational value of AI-generated

assessments. After overcoming these barriers in terms of technical improvements and reproducibility safeguards, LLMs can be a strong asset in radiological technology education, which can even extend to the licensure preparations of other allied health professionals worldwide.



### Acknowledgments

The authors thank Hiroki Ohtani, Hiroki Saito, Tatsuru Ota, Kiyoshi Hishiki, and Masao Fujihara of Teikyo University for their careful evaluation of the problem statements and for the constructive feedback that strengthened this study. During the preparation of this work, the authors used OpenAI o3 to enhance readability and proof-read the English text. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication. Final English proofreading was performed by Enago Co., Ltd.

### Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

### Conflicts of Interest

None declared.

### Data Availability

Data sharing is not applicable to this article as no data sets were generated or analyzed during this study.

### Ethical Approval and Informed Consent

The Ethics Committee of Teikyo University, and the committee concluded that ethical approval was not required as this research is intended for evaluating test question quality and does not constitute human medical research. Written informed consent for

publication was obtained from the individual (or their legal guardian) included in the study

#### Authors' Contributions

Conceptualization: T.I., K.S.

Data Curation: T. Ishibashi

Software / Analysis Tools: S.K.

Formal Analysis: T.H.

Writing – Original Draft: T.I.

#### Abbreviations

AI: Artificial intelligence

LLMs: Large language models

## References

1. Al-Sheikh MH, Albaker W, Ayub MZ. Do mock medical licensure exams improve performance of graduates? Experience from a Saudi medical college. *Saudi J Med Sci.* 2022;10(2):157-161. doi:10.4103/sjmms.sjmms\_173\_21
2. Scott NP, Martin TW, Schmidt AM, Shanks AL. Impact of an online question bank on resident in-training exam performance. *J Med Educ Curric Dev.* 2023;10:23821205231206221. doi:10.1177/23821205231206221
3. Siab F, Key M, Morrissey H, Ball P. Pharmacy students' opinions of using mock questions to prepare for summative examinations. *Int J Curr Pharm Res.* 2020 Jul;12(4):58-65. doi:10.22159/ijcpr.2020v12i4.39079
4. Alawgali GA. An evaluation of a final year multiple choice questions examination at Faculty of Medicine – University of Benghazi. *Open Access Maced J Med Sci.* 2024;12:132-137. doi:10.3889/oamjms.2024.11798
5. Khan R, Payne M, Chinnery H, McLean M. Barriers and facilitators to writing quality items for medical school assessments - a scoping review. *Med Educ.* 2019 May 2;19(1):123. doi:10.1186/s12909-019-1544-8
6. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health.* 2023 Feb 9;2(2):e0000198. doi:10.1371/journal.pdig.0000198
7. Tanaka Y, Nakata T, Aiga K, et al. Performance of generative pretrained transformer on the National Medical Licensing Examination in Japan. *PLOS Digit Health.* 2024 Jan 23;3(1):e0000433. doi:10.1371/journal.pdig.0000433

8. Yang J, Siriwat L, Saowaprut P, Wabina RS. Performance of large language models on Thailand's national medical licensing examination: a cross-sectional study. *J Educ Eval Health Prof.* 2025 Jan 1;22:16. doi:10.3352/jeehp.2025.22.16
9. Zhu S, Hu W, Yang Z, Yan J, Zhang F. Qwen-2.5 outperforms other large language models in the Chinese National Nursing Licensing Examination: retrospective cross-sectional comparative study. *JMIR Med Inform.* 2025 Jan 10;13:e63731. doi:10.2196/63731
10. Tomova M, Roselló Atanet I, Sehy V, Sieg M, März M, Mäder P. Leveraging large language models to construct feedback from medical multiple-choice questions. *Sci Rep.* 2024 Nov 13;14(1):27910. doi:10.1038/s41598-024-79245-x
11. Kondo T, Okamoto M, Kondo Y. Pilot study on using large language models for educational resource development in Japanese radiological technologist exams. *Med Sci Educ.* 2025 Apr;35(2):919-927. doi:10.1007/s40670-024-02251-1
12. Sabaner M, Karatepe Hashas AS, Mutibayraktaroglu KM, Yozgat Z, Klefter O, Subhi Y. The performance of artificial intelligence-based large language models on ophthalmology-related questions in Swedish proficiency test for medicine: ChatGPT-4 omni vs Gemini 1.5 Pro. *AJO Int.* 2024 Dec 11;1(4):100070. doi:10.1016/j.ajoint.2024.100070
13. Mistry NP, Saeed H, Rafique S, Le T, Obaid H, Adams SJ. Large language models as tools to generate radiology board-style multiple-choice

- questions. *Acad Radiol.* 2024 Sep 1;31(9):3872-3878. doi:10.1016/j.acra.2024.06.046
14. Brin D, Sorin V, Konen E, Nadkarni G, Glicksberg BS, Klang E. How large language models perform on the United States Medical Licensing Examination: a systematic review. *medRxiv.* 2023 Sep 3. doi:10.1101/2023.09.03.23294842
15. Zong H, Wu R, Cha J, et al. Large language models in worldwide medical exams: platform development and comprehensive analysis. *J Med Internet Res.* 2024 Dec 27;26:e66114. doi:10.2196/66114
16. Rosoł M, Gąsior J, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Medical Final Examination. *Sci Rep.* 2023 Nov 22;13(1):20512.
17. Kim J, Chua M, Lorenzo A, et al. Use of AI (GPT-4)-generated multiple-choice questions for the examination of surgical subspecialty residents: report of feasibility and psychometric analysis. *Can Urol Assoc J.* 2025 Feb 24;19(6):9020. doi:10.5489/cuaj.9020
18. Zhang Y, Pan Y, Zhong T, et al. Potential of multimodal large language models for data mining of medical images and free-text reports. *Meta-Radiology.* 2024 Dec 1;2(4):100103. doi:10.1016/j.metrad.2024.100103
19. Soni N, Ora M, Agarwal A, Yang T, Bathla G. A review of the opportunities and challenges with large language models in radiology: the road ahead. *AJNR Am J Neuroradiol.* 2025 Jul 1;46(7):1292-1299. doi:10.3174/ajnr.A8589

20. Alsabbagh AR, Mansour T, Al-Kharabsheh M, et al. MiniMedGPT: efficient large vision-language model for medical visual question answering. *Pattern Recognit Lett.* 2025 Mar 1;189:8-16. doi:10.1016/j.patrec.2025.01.001
21. Zhu D, Chen J, Shen X, Li X, Elhoseiny M. MiniGPT-4: enhancing vision-language understanding with advanced large language models. *arXiv. Preprint posted online April 2023.* doi:10.48550/arXiv.2304.10592
22. Dong X, Zhang P, Zang Y, et al. InternLM-XComposer2-4KHD: a pioneering large vision-language model handling resolutions from 336 pixels to 4K HD. *Advances in Neural Information Processing Systems*, 37, 42566-42592 *arXiv. Preprint posted online April 2024.* doi:10.48550/arXiv.2404.06512
23. Wang Z, Huang Y, Wu Y, et al. Fusion side tuning: a parameter and memory efficient fine-tuning method for high-resolution medical image classification. In: *Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Lisbon: IEEE; 2024.
24. Reid M, Globerson A, Lacey G, et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. Google DeepMind. *arXiv. Preprint posted online March 2024.* doi:10.48550/arXiv.2403.05530
25. Boostani M, Bánvölgyi A, Goldust M, et al. Diagnostic performance of GPT-4o and Gemini Flash 2.0 in acne and rosacea. *Int J Dermatol.* 2025. doi:10.1111/ijd.17729
26. Fleming S, Morse K, Kumar A, et al Assessing the potential of USMLE-like exam questions generated by GPT-4. *medRxiv. Preprint posted online April 2023.* doi:10.1101/2023.04.25.23288588

27. Zhou Z, Rizwan AP, Rogoza N, Chung AD, Kwan BY. Differentiating between GPT-generated and human-written feedback for radiology residents. *Curr Probl Diagn Radiol.* 2025 Oct;54(5):574-578. doi:10.1067/j.cpradiol.2025.02.002
28. Kuusemets L, Parve K, Ain K, Kraav T. Assessing AI-generated (GPT-4) versus human-created MCQs in mathematics education: a comparative inquiry into vector topics. *Int J Educ Math Sci Technol.* 2024;12(6):1538-1558. doi:10.46328/ijemst.4440
29. Ma W, Yang C, Kästner C. (Why) is my prompt getting worse? Rethinking regression testing for evolving LLM APIs. In: *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI 2024* Apr 14, pp. 166-171. doi:10.1145/3644815.3644950
30. Schroeder K, Wood-Doughty Z. Can you trust LLM judgments? Reliability of LLM-as-a-judge. arXiv. Preprint posted online December 2024. doi:10.48550/arXiv.2412.12509
31. Renze M, Guven E. The effect of sampling temperature on problem solving in large language models. arXiv. Preprint posted online February 2024. doi:10.18653/v1/2024.findings-emnlp.432

## Figure Legends

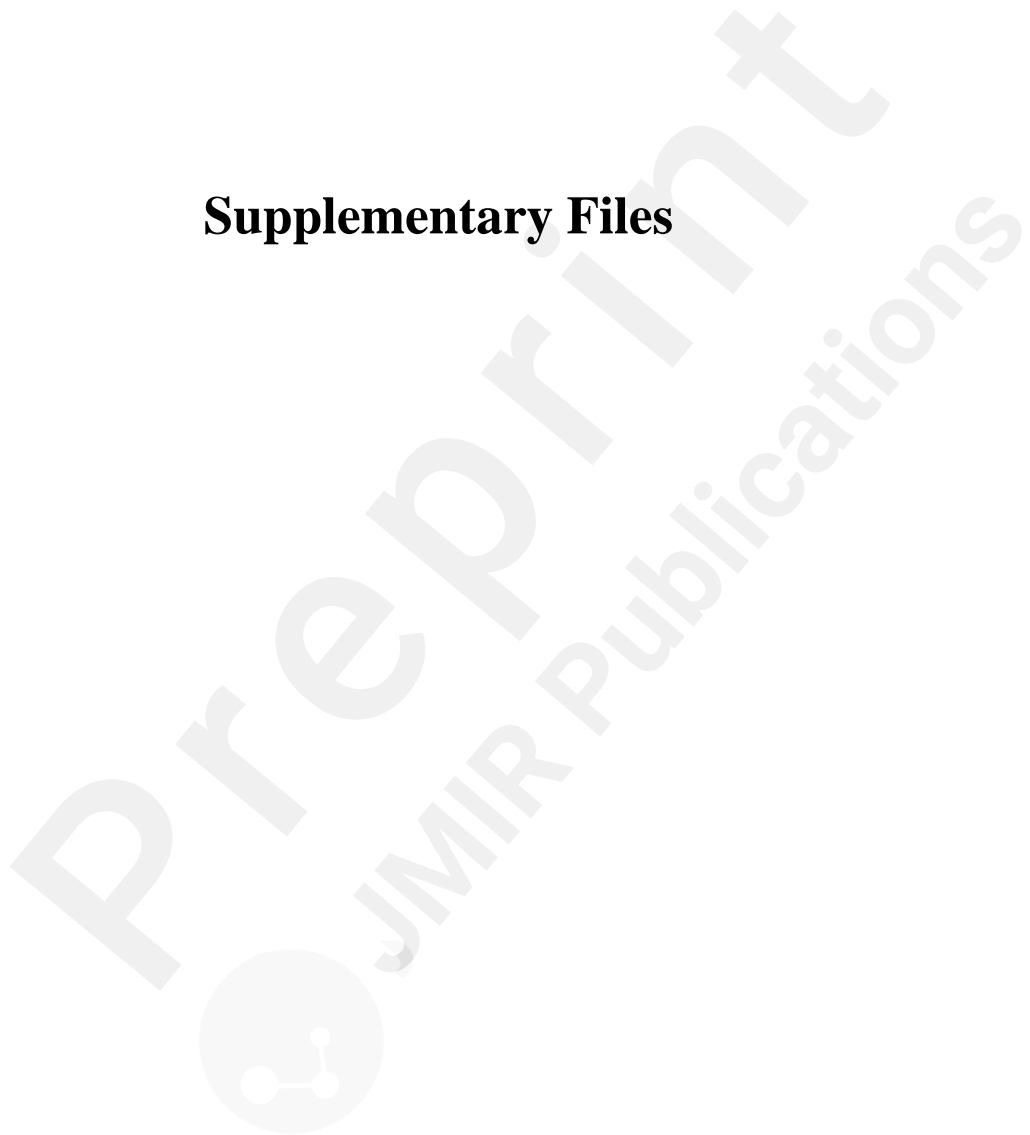
Figure 1. Representative interaction with a large language model.

This diagram illustrates the workflow used to evaluate the answer accuracy of large language models (LLMs). The LLMs were given prompts to answer each question (including text and images when applicable) in Japanese, with specific instructions for answer selection and formatting. The output included the selected answer, a confidence score, and a brief explanation. Note: All actual prompts and inputs were entered in Japanese, but this example is shown in English for illustration purposes.

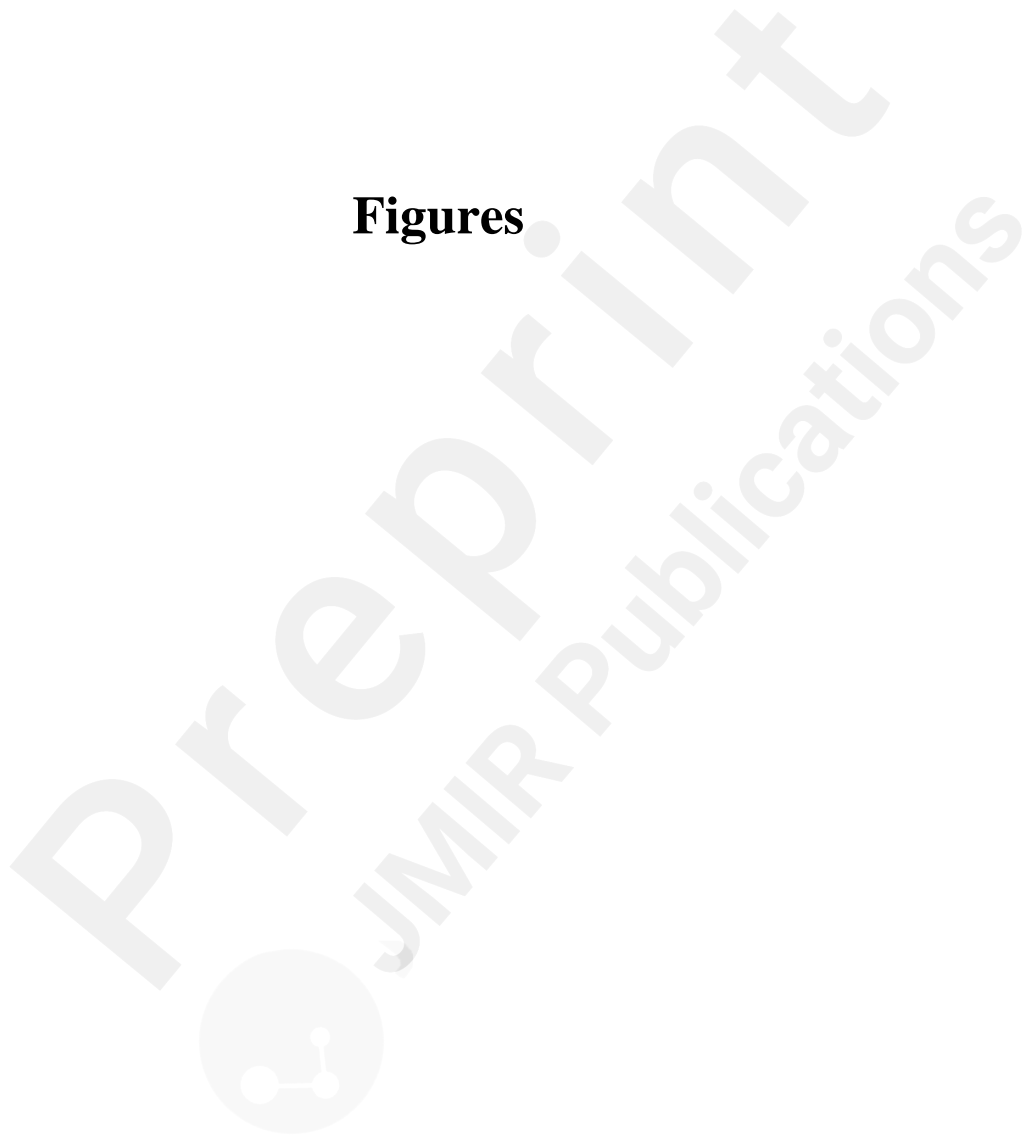
Figure 2. Prompt summary and representative example of item generation.

(A) Summary of the prompts used to instruct the language model to generate original mock questions aligned with the National Examination for Radiological Technologists. The summary outlines the role of the model, input references, specifications of generation, item-creation rules, and output format. (B) The actual prompt and representative response generated by the model. The prompt included specific formatting and content-generation instructions written in Japanese. The response shows the generated item, correct answers, and explanation in Japanese.

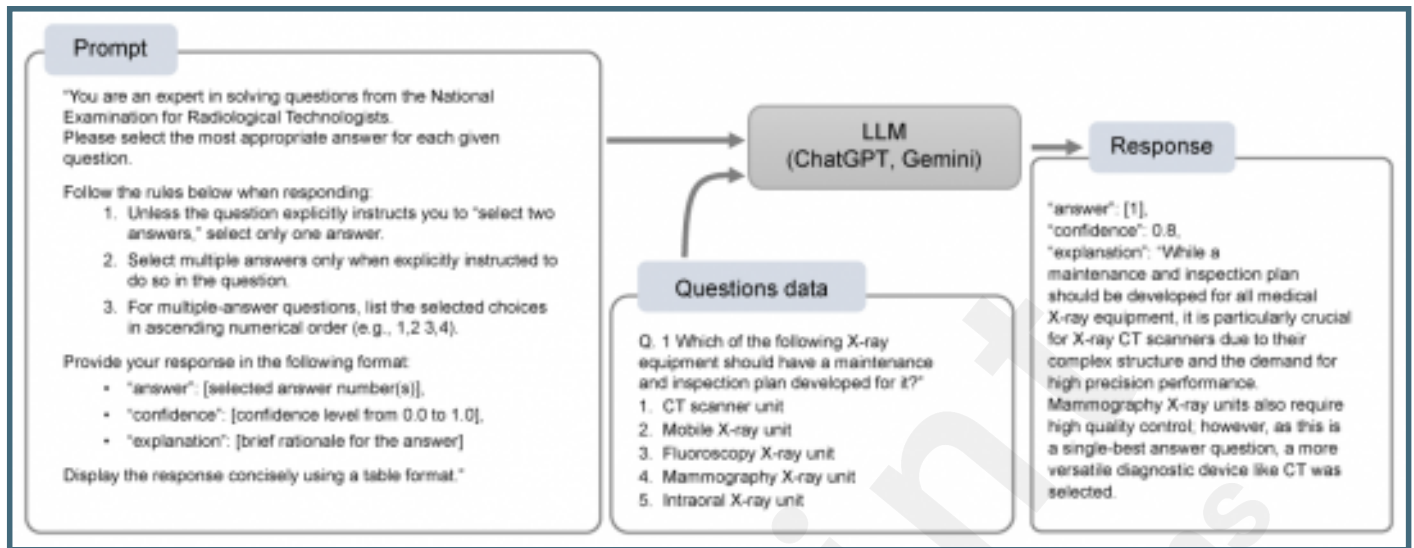
## Supplementary Files



## Figures



Representative interaction with a large language model.



Prompt summary and representative example of item generation.

**A**

**Role & General Instruction:**  
Generate practice questions for the National Examination for Radiological Technologists that reflect the official exam's tone, difficulty, and content focus.

**Reference materials:**  
Use items 1-144 from the past five years of the National Examination (72nd to 76th, 2020-2024) as reference material to guide question generation in terms of content and format.

**Specifications for Item Generation:**  
Create the specified number of questions for the designated subject, ensuring coverage of the subtopics and keywords provided in the reference materials.

**Instructions for Item Creation:**  
Generate original five-choice questions that follow the style and level of past exams. Use both single-answer and two-answer formats, marking correct answers and including brief explanations for each.

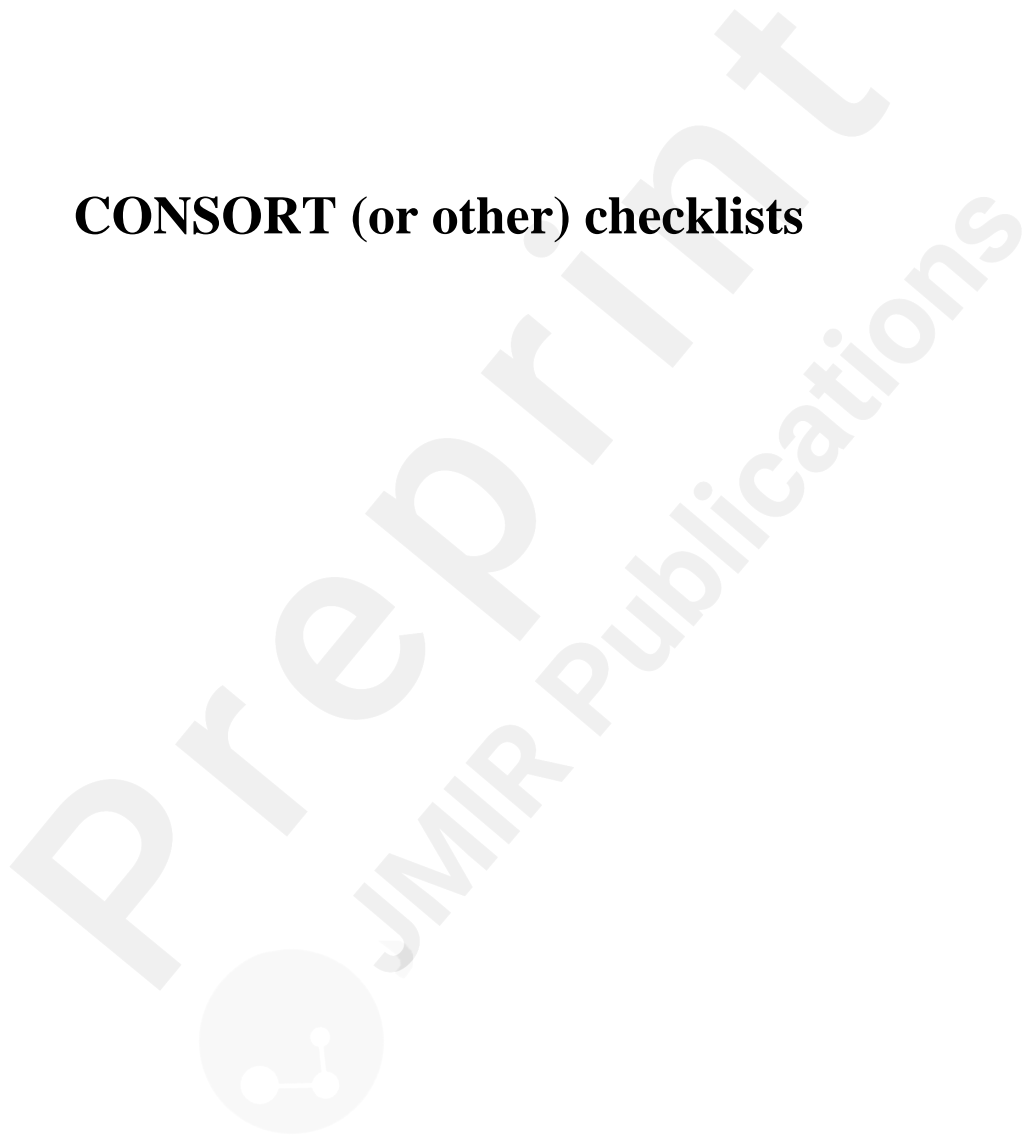
**Output Format:**  
Present each item in a standardized format with a numbered question, five answer choices, clearly indicated correct answer(s), and a brief explanation for all options.

---

**B**

Prompt (Original Japanese)	Response (Generated in Japanese)
<p>「あなたは診療放射線技師国家試験の問題作成の専門家です。以下の参考資料を基に、国家試験本番と同等の文体・難易度・出題傾向を持つ練習問題を作成してください。</p> <p>参考資料：過去5年分（第72回（2020年）～第76回（2024）の診療放射線技師国家試験問題集（問1から問144）</p> <p>問1 性ホルモンが腫瘍の増殖に関わるのはどれか。2つ選べ。</p> <ol style="list-style-type: none"> <li>1. 卵巣</li> <li>2. 陰茎癌</li> <li>3. 尿道癌</li> <li>4. 子宮体癌</li> <li>5. 前立腺癌</li> </ol> <p>：</p> <p>：</p> <p>問144 変性が起こるのはどこか。</p> <ol style="list-style-type: none"> <li>1. 肺</li> <li>2. 卵管</li> <li>3. 卵巣</li> <li>4. 子宮頸部</li> <li>5. 子宮内膜</li> </ol> <p>※出題内容の指定※ - 科目：※[放射線物理学]※ - 問題数：※[10問]※ - 範囲・キーワード：※[出題範囲(DF_2_放射線物理学)]※</p> <p>※科目※の小項目のみを抽出し表示する。</p> <p>※出題にあたっての条件※ - 国試本番と同等の口調・専門性・難易度を再現すること。過去問の典型的な表現（「～はどれか。」など）や形式にならう。 - 問題は※完全新規※に作成すること（参考資料と※同一の問題文にならない※ように留意すること）。過去問をヒントにテーマや表現を変えて出題してよいが解答そのものは厳しく考案する。 - 出題形式：※6択択一形式※（正答1つ）と※5択複数形式※（正答2つ）を組み合わせる。全体の約※30%※を正答1つの問題、約※30%※を正答2つの問題とする。 - ※選択肢※は問題ごとに※1～5の5つ※を用いること（内容は出題範囲に必ず沿ったもの）。 - 正答が2つある問題では、問題文の末尾に「※2つ選べ。※」と明記し、受験者が2つ選ぶ形式であることを示す。 - ※解答表示※：各問題の選択肢の直後に「正答：○」と正解選択肢を記号で示す（正解が複数ある場合は「正答：2と4」のように列記）。 - ※解説※：各問題に対して、正解の理由および各選択肢の解説を添えること。他の選択肢が誤りである理由や関連知識も含め、受験者の理解が深まる説明を書く。</p> <p>※出力フォーマット※：以下の形式で問題文、選択肢、正答、解説を記載すること（各問題を通し番号で番号付けする）。</p> <p>問題1. ～（問題文）～はどれか。 ※複数正解の場合、文末に「2つ選べ。」</p> <ol style="list-style-type: none"> <li>1. 選択肢1</li> <li>2. 選択肢2</li> <li>3. 選択肢3</li> <li>4. 選択肢4</li> <li>5. 選択肢5</li> </ol> <p>正答：2 ※複数正解の場合「正答：2と4」</p> <p>解説：問題文や選択肢に関する解説。正解が正しい理由と他の選択肢が誤りである理由をそれぞれ説明する。」</p>	<p>練習問題(全10問)</p> <p>問題1 水中において1.25 MeVの光子が物質と相互作用する際、最も卓越する現象はどれか。 光電効果 コンプトン散乱 電子対生成 レイリー散乱 光核反応 正答：2 解説：光電吸収は<math>E(\gamma) &lt; 2m_0c^2</math>、電子対生成は<math>E(\gamma) &gt; 2m_0c^2</math>に支配されるが、低原子番号物質でMeV付近ではコンプトン散乱が最大となる。レイリー散乱は無電離散乱で電子が小さく、光核反応は高エネルギー域(数10 MeV以上)で顕在化する。</p> <p>問題2 線エネルギー付与(LET)が最大となるのはどれか。 5 MeV <math>\alpha</math>粒子 25 MeV 陽子 6 MeV 電子 1 MeV <math>\gamma</math>線 120 keV X線 正答：1 解説：LETは電荷の2乗に比例し速度の2乗に反比例する。重くて遅い<math>\alpha</math>粒子(<math>He^{2+}</math>)は密度エネルギー付与が大きく、他の放射線より高LETとなる。</p> <p>問題3 光電吸収係数を増大させる要因はどれか。2つ選べ。 低原子番号物質 低エネルギー光子 材料密度の低下 高原子番号物質 高エネルギー光子 正答：2と4 解説：光電吸収係数は概ね<math>\rho Z^3 E^{-3}</math>に比例する。よって低エネルギー(<math>\gamma</math>)・高Z(4)が係数を大きくする。密度は質量減弱係数を変えず(1/30%)はむしろ係数を小さくする。</p> <p>問題4 X線管で発生する制動X線の最短波長を1/2に短縮するには管電圧(V)をどのように変化させればよいか。 2倍にする 1/2倍にする 1/√2倍にする 変化させない 正答：1 解説：最短波長<math>\lambda_{min} = hc/eV</math>であり<math>1/V</math>。半減させるには電圧を2倍にする必要がある。</p> <p>問題5 熱中性子の速度に影響する物理量はどれか。2つ選べ。 温度 中性子の質量 原子核の原子番号 中性子のスピン 電離密度 正答：1と2 解説：Maxwell-Boltzmann分布より<math>v_{rms} = \sqrt{3kT/m}</math>、速度は温度(1)と質量(2)に依存し、核種のZやスピン、媒質の電離密度とは直接関係しない。</p>

## **CONSORT (or other) checklists**



STROBE checklist.

URL: <http://asset.jmir.pub/assets/145674167fda3a94629f8995901c4fbf.pdf>

