Article title: Ethical review of clinical research with generative AI: Evaluating ChatGPT's accuracy and reproducibility

Short title: Evaluating ChatGPT's accuracy and reproducibility

Authors:

Yasuko Fukataki[1, 2], Hayashi Wakako[1], Nishimoto Naoki[3], Yoichi M. Ito[1, 3]

Affiliations:

[1] Health Data Science, Department of Social Science, Graduate School of Medicine, Hokkaido University, Sapporo, Japan

[2] Biostatistics and Data Management, Sapporo Medical University, Sapporo, Japan

[3]Data Science Center, Promotion Unit, Institute of Health Science Innovation for Medical Care, Hokkaido University Hospital, Sapporo, Japan

Corresponding author:

Yasuko Fukataki

E-mail: yafukataki@sapmed.ac.jp

Abstract

This study evaluated the accuracy and reproducibility of ChatGPT models, specifically GPT-4 and GPT-4o, by reviewing Japanese-language clinical research protocols and informed consent forms using Japanese prompts. The integration of generative AI technologies into clinical research ethics reviews has the potential to enhance consistency, reduce human error, and decrease the manual effort required to assess complex documents. This study primarily aimed to assess and compare the ability of these models to accurately extract and summarize key elements such as research objectives, study design, and ethical considerations, which are critical for ethical review processes. We developed and optimized custom prompts to improve the performance of the models, focusing on the essential aspects of the protocol and informed consent review. The results showed that GPT-4o achieved an 80% accuracy rate in identifying research objectives and a 100% accuracy rate for research design, indicating superior consistency compared with GPT-4, which, despite being slightly less accurate, still showed significant potential for application in ethics reviews. Furthermore, a comparison between customized GPTs and standard prompts revealed that customized GPTs provided significantly higher reproducibility and accuracy, underscoring the value of fine-tuning and Retrieval-Augmented Generation techniques for enhancing AI-assisted review processes. Additionally, challenges in parsing complex PDF documents were identified, highlighting the importance of standardized document formatting to ensure accurate AI analysis. These findings demonstrate the potential of AI-driven systems to improve the efficiency, accuracy, and standardization of research ethics evaluations, potentially setting new standards for AI integration in clinical research practice.

Author summary

This study examined the use of ChatGPT models, specifically GPT-4 and GPT-4o, to review clinical research protocols and informed consent forms written in Japanese, focusing on their ability to improve the review process in clinical research ethics. Our objective was to determine whether these AI models can reliably extract and summarize critical elements, such as research objectives, study design, and ethical considerations, from research documents. The GPT-4o model demonstrated high accuracy and reproducibility, particularly in evaluating research designs, with 80% and 100% accuracy rates for research objectives and research designs, respectively. GPT-4 also exhibited good performance, albeit with a slightly lower accuracy. The findings revealed that tailored prompts significantly enhanced model performance compared to standard prompts, highlighting the necessity for precise customization in AI-driven analyses. Furthermore, we discovered that well-structured documents are crucial for enabling accurate AI-assisted reviews. These results suggest that, with further refinement, ChatGPT could become a valuable tool for enhancing the consistency and efficiency of ethics committee evaluations, potentially reducing manual workload and setting new standards for AI integration in clinical research.

Introduction

An ethical review of clinical research is essential to ensure patient safety and reliability. Ethical review

committees evaluate the ethical appropriateness of research protocols and establish standards to protect patient rights and safety. However, the review processes of Institutional Review Boards (IRBs) globally and Certified Review Boards (CRBs) in Japan face several challenges [1-10]. For example, CRBs in Japan exhibit variability in the details and quality of review comments, necessitating the standardization and equalization of review quality [1-7]. Internationally, discrepancies in revision instructions for research protocols and consent documents, as well as inconsistencies in review times, have been reported [8-10]. This variability in review quality undermines the consistency and transparency of the review process. These issues often stem from differences in the interpretation of review standards and shortage of human resources among experts. Delays in the review process risk delaying the start of research, which, in turn, delays the introduction of new treatments and medical technologies. Moreover, inadequate or inappropriate revision comments based on ethical standards may compromise participant safety due to insufficient or improper modifications of protocols and informed consent forms (ICFs).

In recent years, advancements in Large Language Models (LLMs), machine learning, deep learning, and generative artificial intelligence (AI) have shown promise in medical fields, including medical image analysis, electronic health record analysis, and virtual health assistants [11-17]. LLMs, which are models based on deep learning, are a part of generative AI and are capable of performing natural language processing (NLP) tasks [18]. ChatGPT, a representative example of generative AI, has

77 demonstrated its ability to learn from large datasets and execute complex language tasks [19]. For
78 instance, ChatGPT's abilities have been widely recognized, achieving results comparable to passing
79 medical licensing exams in both Japan and the United States [20-22].
80
81 We believe that the advancement of these technologies presents potential improvements in the quality
82 and efficiency of IRB reviews of research protocols. However, to our knowledge, no prior research has
83 explored the application of generative AI to the IRB review process. One major reason is the high
84 confidentiality of research documents, such as protocols, , making it inappropriate for cloud-based
85 generative AI (e.g., ChatGPT, Gemini, Copilot) to analyze these documents. Furthermore, IRB reviews
86 typically require the perspectives of multiple experts, extensive knowledge and experience, and are
87 complex processes [1,23-25]. It remains unclear whether generative AI can fully replicate this process.
88 Additionally, evaluating AI's ability to ensure the reproducibility and accuracy of reviews is crucial,
89 but this has not been sufficiently validated.
90
91 By developing LLMs that operate in local environments, it becomes possible to safely process highly
92 confidential documents, thereby mitigating the risk of data leakage associated with cloud-based
93 generative AI. However, the development of LLMs in a local environment requires advanced technical
94 expertise, such as model training, optimization, and securing computational resources, which are
95 expected to demand significant time and effort.

Therefore, this study aimed to evaluate how current cloud-based generative AI technologies can contribute to the ethical review process by focusing on the accuracy and reproducibility of ChatGPT in analyzing and summarizing medical documents. Specifically, we developed four custom GPT models and assessed their ability to accurately read and summarize highly specialized, complex Japanese research protocols and informed consent documents. Additionally, we compared the latest GPT-4o model (as of May 13, 2024) with the conventional GPT-4 model to evaluate improvements in processing accuracy when handling Japanese documents. This preliminary study aimed to determine whether GPT models can produce accurate and reproducible outputs based on professional review standards for IRB reviews. This essential data will be used for more extensive testing and validation in upcoming trials. The initial evaluation results indicate that ChatGPT can accurately interpret specialized research protocols and ICFs, faithfully extracting and summarizing the specified content. These findings suggest that generative AI technology has potential utility in analyzing medical documents.

Future studies will explore the use of Retrieval-Augmented Generation (RAG) technology to assess whether AI can appropriately perform reviews from multiple expert perspectives. If cloud-based generative AI proves effective for IRB reviews, its implementation through secure, contracted cloud services may be considered, offering scalability and maintainability with faster deployment than local

LLM development. Additionally, based on this study's results, local LLM development can be considered, which could further enhance confidentiality and allow for customization to specific requirements.

In future, generative AI and LLM technologies are expected to rapidly extract and summarize information from research protocols and ICFs, generate comments consistent with review standards, and produce comprehensive review feedback. This would enable AI to play a complementary role in supporting the multidimensional reviews conducted by experts.

Furthermore, this study and future research may provide new solutions to address the current challenges in ethical reviews, leading to the development of AI-assisted review support tools that streamline the review process and enhance patient safety and research reliability. Importantly, the goal of this research is to clarify the supportive role that generative AI can play in the ethical review process, aiming to complement, rather than replace, expert reviews. The primary focus of this research was to explore the range and limitations of AI's role in supporting reviews and to identify efficient and appropriate methods for utilizing generative AI.

Results

In this study, we evaluated research protocols and informed consent documents using customized GPTs, focusing on research objectives and background, research design, and advantages and disadvantages.

We compared the accuracy and reproducibility of the GPT-4 and GPT-4o models, as well as the reproducibility of the outputs generated by the GPTs with those produced using standard prompts. Detailed results for each evaluation category are provided below.

Evaluation of accuracy and reproducibility

Accuracy evaluation of research objectives and background

The GPT-4o model produced outputs that completely matched the content of the research protocol in eight out of ten attempts, whereas the GPT-4 model achieved seven complete matches. In both models, discrepancies were found where “trastuzumab BS” was mistakenly shortened to “trastuzumab.” Specifically, the research protocol states, “In this context, we decided to provide trastuzumab BS, one of the matched treatments, to patients with solid tumors who tested positive for HER2 gene amplification through the above cancer genome profiling test system, and to explore its efficacy.” However, the output omitted “BS” from “trastuzumab BS” and only referred to “trastuzumab” (Table 1).

Table 1. Comparison of Accuracy Between GPT-4o and GPT-4 Models in Reviewing Research Protocols

Item	Consistency Rate (%)		Partial Consistency Rate （%）	
	GPT-4o	GPT-4	GPT-4o	GPT-4
Research objectives and background	80	70	20	30

Research design	100	100	0	0
Advantages and disadvantages (RP)	100	100	0	0
Advantages and disadvantages (ICF)	100	100	0	0

- **RP** Research protocol
- **ICF:** Informed consent form
- **Accuracy evaluation:** The consistency rate represents the percentage of outputs that perfectly matched the research protocol, while the partial consistency rate indicates cases where the main information was mostly consistent. (Based on the results of 10 trials)

Accuracy evaluation of research design

Regarding research design, both GPT-4o and GPT-4 produced outputs that were completely consistent with the content of the research protocol in all 10 attempts. Specifically, all outputs correctly included information such as “Open-label, single-arm, multicenter interventional study, target: patients with advanced solid tumors without standard treatment options, non-randomized, sample size: 41.” Additionally, some outputs accurately included supplementary information, such as study duration, interim analysis, primary endpoints, and administration methods (Table 1).

Accuracy evaluation of advantages and disadvantages

Regarding the advantages and disadvantages outlined in the research protocol and ICF, both GPT-4o

and GPT-4 produced outputs that were completely consistent with the content of the research protocol in all 10 attempts (Table 1).

Fig. 1 provides a visual representation of the differences in the consistency rates for each evaluation category, with the GPT-4o model showing higher overall consistency rates.

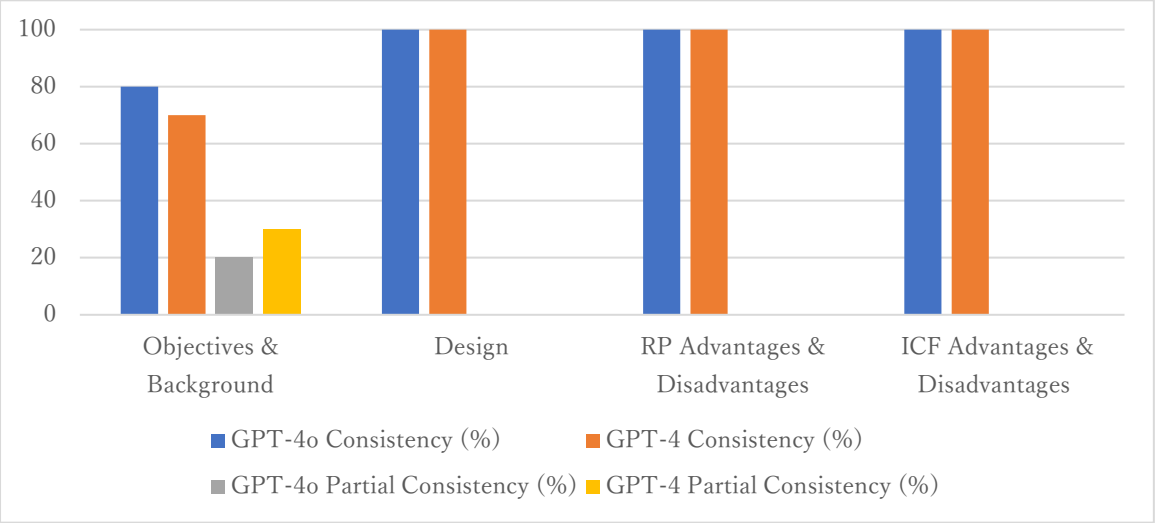

Fig. 1. Comparison of consistency and partial consistency rates for research protocol review

Reproducibility evaluation

For reproducibility, both the GPT-4o and GPT-4 models demonstrated high reproducibility in the categories of "Research objectives and background," "Research design," and "advantages and disadvantages." Specifically, reproducibility was considered “high” if 8 or more out of 10 trials were rated as "consistent" or "partially consistent" (Table 2).

Table 2. Comparison of Reproducibility in Reviewing Research Protocols Between GPT-4o and GPT-4 Models

Item	GPT-4o	GPT-4
Research objectives and background	High	High
Research Design	High	High
Advantages and disadvantages of research protocol	High	High
Advantages and disadvantages of ICF	High	High

・**Reproducibility evaluation:** Reproducibility was assessed based on the number of trials in which the output was judged to be consistent or partially consistent. Reproducibility was rated as “high” if 8 or more out of 10 attempts were consistent or partially consistent, “moderate” if 5-7 attempts were consistent or partially consistent, and “low” if 4 or fewer attempts were consistent or partially consistent.

Comparison of GPTs and standard ChatGPT prompt outputs

In addition to the outputs generated using the GPTs, we produced outputs using standard prompts to assess the advantages described in the research protocol. The results of GPT-4o showed that GPTs consistently generated two advantages across all 10 trials: (1) improvement in detailed information collection and follow-up care, and (2) an increase in treatment options.

In contrast, outputs generated using standard prompts produced a more diverse range of categories, with the number of advantages mentioned varying between 3 and 8 across the 10 trials (Tables 3 and

195 4). The categories of advantages produced by the GPTs and standard prompts are summarized in Table
196 4.

197

198 **Table 3. Number of benefits generated by GPTs and standard prompts (Model used: GPT-4o)**

Prompt Type	Number of Trials	Avg. Benefits	SD	Max	Min
GPTs	10 trials	2	0	2	2
Standard prompts	10 trials	5.3	1.8	8	3

199 • **Avg. Benefits:** Average number of benefits

200

201 **Table 4. Benefit categories generated by GPTs and standard prompts (Model used: GPT-4o)**

Category Name	GPTs	Standard Prompts
Detailed information collection and follow-up care	Yes	Yes
Increase in treatment options	Yes	Yes
Reduction of economic burden	N/A	Yes
Compensation for health risks	N/A	Yes
Provision of personalized treatment	N/A	Yes
Cost-effective treatment	N/A	Yes
Evaluation of treatment efficacy and safety	N/A	Yes
Collection and analysis of clinical data	N/A	Yes

Protection of personal information	N/A	Yes
Improvement in quality of life	N/A	Yes

Discussion

This study served as a preliminary evaluation of the accuracy and reproducibility of ChatGPT when reviewing Japanese research protocols using Japanese prompts. The results demonstrated different levels of accuracy and reproducibility between the GPT-4 and GPT-4o models in the review process. These findings provide essential baseline data for the future implementation of RAG techniques and fine-tuning.

First, the GPT-4o model achieved an 80% consistency rate for "Research Objectives and Background" and a 100% consistency rate for "Research Design." This suggests that the GPT-4o model could be valuable in ensuring consistency in the review process, indicating the potential for AI to efficiently and accurately analyze Japanese research protocols. Although GPT-4 showed slightly lower consistency rates, it maintained high accuracy, demonstrating performance levels that may be applicable in review tasks. These results underscore the importance of continuing comparative assessments between the GPT-4 and GPT-4o models. Evaluating the nuanced differences between the two models across different tasks and conditions will help deepen our understanding of their respective characteristics.

Previous studies have also highlighted the ongoing challenges of reproducibility and accuracy in AI applications within the medical field. For example, “The reproducibility issues that haunt health-care AI” points out that AI faces reproducibility issues in medical diagnoses and treatments attributed to inconsistent results across different datasets and conditions [26]. The findings of this study reaffirm the importance of reproducibility as a critical factor in the practical application of AI technologies, aligning with previous research. In particular, the evaluation of AI’s accuracy and reproducibility in analyzing medical documents remains a significant challenge for future technological development.

This will not only broaden the application scope of AI but also provide foundational data for optimal model selection. By assessing the flexibility and adaptability in handling complex medical documents, the groundwork for the practical application of AI tools can be established. Thus, the ongoing comparison between GPT-4 and GPT-4o is a critical step in maximizing AI performance and selecting the most effective tool for real-world applications.

Moreover, a comparison between the outputs generated using GPTs and those generated using standard prompts showed that the GPTs consistently exhibited high reproducibility, thereby improving the accuracy of the review process. In contrast, the outputs generated using standard prompts displayed variability in the extracted information, which was attributed to a lack of detailed prompt settings. These findings suggest that incorporating RAG techniques and fine-tuning can further enhance the

precision and consistency of AI-assisted reviews.

Technical challenges related to reading PDFs were also identified. In particular, when PDF files are highly structured or text is stored as images, ChatGPT cannot extract information accurately. It was confirmed that converting documents to Word format and standardizing the formatting of sections improved the reading and analysis accuracy. This indicates that document structure and format standardization are crucial factors for improving the accuracy of AI-assisted reviews, emphasizing the importance of proper document preparation for effective AI utilization. These findings provide practical guidelines for preventing incorrect information extraction.

Additionally, although ChatGPT demonstrated the potential to process specialized documents, such as research protocols, with a certain degree of accuracy and reproducibility in the medical field, its performance was found to be dependent on the document format and structure. For instance, distinguishing between similar technical terms, such as "trastuzumab BS" and "trastuzumab," proved challenging, indicating the need for further improvements. Maintaining consistency in terminology by using guidelines and templates when drafting documents is crucial. This will help further enhance the accuracy and reliability of AI-assisted reviews.

Future research should explore the optimal approach by combining RAG techniques and fine-tuning to improve the precision of handling technical terms. Additionally, we plan to investigate best practices

for document preparation to maximize AI performance. Ultimately, ChatGPT has the potential to contribute to the efficiency and accuracy of IRB review processes; however, further research and validation are required. This study lays a crucial foundation for improving review tasks using AI, and provides directions for future research and development.

Limitations

This study had two main limitations. First, because the appropriateness of the published review results was not disclosed, it was not possible to directly verify the accuracy of ChatGPT's evaluations against a reference set of correct answers. We recognize that this constraint imposes certain limitations on the accuracy and reproducibility assessments in this study. Therefore, the research team manually reviewed the research protocols and output results, evaluating accuracy and reproducibility based on predefined criteria (Tables 5–7). However, since the manual review process contains subjective elements, there is a potential for bias in the evaluation of the results.

Table 5. Four custom GPTs created for this study

Type	Overview of Prompts
GPTs1	Accurately extract the research objectives and background from the provided research protocol.
GPTs2	Accurately extract details such as target patients, trial phase, trial type, randomization, control group setup, blinding, and sample size.

GPTs3	Accurately extract the advantages and disadvantages for the subjects from sections 16.3.1 and 16.3.2 of the research protocol.
GPTs4	Accurately extract the advantages and disadvantages for the subjects from Section 6 of the ICF.

272

273 **Table 6. Criteria for consistency evaluation**

Item	Condition
Accuracy of content	・Specific numerical values match perfectly. ・All major information related to research objectives and background is included. ・All major information related to research design and methods is included.
Consistency of meaning	The meaning and intent of the original document are accurately conveyed.
Consistency of expression	Even if phrasing or word choices differ, the meaning and intent remain unchanged.

274

275 **Table 7. Criteria for accuracy evaluation**

Evaluation criteria	Description
Consistent	The output content matches the content of the research protocol completely and without error.
Partially consistent	The major information and intent of the output are mostly consistent with the research protocol, but minor discrepancies are observed. However, these discrepancies do not significantly impair the accuracy of the information, nor do they alter the meaning or

	intent.
Inconsistent	The output content does not match the major content of the research protocol, resulting in significant impairment of information accuracy or alteration of meaning or intent.

Second, since this study did not use an Application Programming Interface (API), the system did not automatically log which model (GPT-4 or GPT-4o) was used in each session. Consequently, it was not possible to systematically verify the model used from the downloaded data. However, because the researchers manually recorded the model used for each session, we believe this limitation did not significantly affect the overall reliability of the study. In future research, we plan to use the API to automate detailed logging to ensure greater objectivity and reproducibility. The use of the API is expected to further improve the efficiency and accuracy of the research.

Materials and Methods

Study overview

In this study, we evaluated the performance of GPT models (OpenAI, Inc., San Francisco, CA, USA) by reviewing research protocols. Specifically, we compared the latest GPT-4o model with the traditional GPT-4 model to assess the improvements in accuracy when processing Japanese documents with Japanese prompts. GPTs (Generative Pre-trained Transformers) are specialized versions of ChatGPTs that users can customize for specific tasks or topics designed to handle complex processes.

We evaluated the accuracy and reproducibility of both the GPT-4 and GPT-4o models using customized GPTs for the following evaluation categories: 1) research objectives and background, 2) research design, and 3) advantages and disadvantages of the research. Furthermore, after comparing the GPT-4 and GPT-4o models, we used the GPT-4o model to compare the outputs generated by the GPTs with those generated using standard prompts. For standard prompts, we used only the GPT-4o model to extract the advantages and compared the results with those of GPTs. This comparison allowed us to analyze the differences in reproducibility and accuracy between the models and methods.

Data source

The research protocols used in this study were mock review materials published on the official website of Japan's Ministry of Health, Labor, and Welfare [27-31]. These materials included 10 research projects from various disease areas: three in oncology, one in gastrointestinal diseases, one in rheumatoid arthritis, one in vaccines, one in medical devices, one in cardiovascular diseases, one in pediatric psychiatric disorders, and one in cognitive impairment. Each project included implementation plans, research protocols, explanatory clinical research documents, lists of collaborating investigators, procedure manuals for disease response, monitoring manuals, audit manuals, and expert evaluations.

For this study, we used the research protocol titled "Phase II Trial of Trastuzumab for Patients with

Advanced Solid Tumors without Standard Treatment Options Showing HER2 Gene Amplification" (【統合版】模擬審査①(HER2)_研究計画書.pdf). Initially, uploading the research protocol PDF directly to ChatGPT proved difficult for reading, analyzing, and extracting text due to inconsistencies in the PDF structure and quality. This issue was particularly pronounced when the text was stored as images or when the PDF creation process was not standardized. To address this, we converted the PDF into Word format using a general PDF-to-Word conversion tool and standardized the document formatting. This conversion significantly improved ChatGPT's ability to accurately analyze the research protocol. An expert evaluation document was used as reference to determine the four evaluation categories for the research protocol.

GPT model selection

In this study, we evaluated the performance of the GPT-4 and GPT-4o models. GPT-4 is a well-established high-performance model, rated highly in many text-generation tasks [32]. In contrast, GPT-4o, updated on May 13, 2024, is twice as fast as GPT-4 Turbo, while reducing costs by 50%. Thus, GPT-4o offers significant advantages in terms of cost efficiency and processing speed. Furthermore, GPT-4o sets new standards for speech recognition, multilingual support, and visual capabilities, achieving high accuracy in non-English languages due to improvements in multilingual capabilities [33]. Given these characteristics, we deemed it important to verify the potential accuracy improvements in the GPT-4o model.

330

331 # GPTs customization and use

332 In this study, we utilized the "GPTs" feature provided by OpenAI to create four customized GPTs [34]
333 aimed at evaluating research protocols and informed consent forms (ICFs) (Table 5). Each GPT was
334 designed to correspond to a specific evaluation category and serve the following purposes and roles:

335 ・**GPTs1:** A prompt designed to extract and summarize the "Research Objectives and Background"
336 from the research protocol. The prompt was instructed to accurately extract important terms and
337 abbreviations.

338 ・**GPTs2:** A prompt designed to extract and summarize the "Research Design and Methods" from the
339 research protocol, accurately capturing key information such as the trial phase and target patients.

340 ・**GPTs3:** A prompt designed to extract the advantages and disadvantages for participants from the
341 research protocol.

342 ・**GPTs4:** A prompt designed to extract the advantages and disadvantages for participants from the
343 ICF.

344

345 To evaluate the advantages and disadvantages, each GPT was designed to focus on specific sections
346 where these were described in the research protocol and ICF and to accurately extract the relevant
347 content. Using these GPTs, we compared the performance of GPT-4 and GPT-4o.

348

These models were developed using cutting-edge natural language processing (NLP) technology. GPTs are pretrained on vast amounts of text data and can adapt to tasks such as new text generation, question answering, and summarization. This enables GPTs to generate highly accurate human-like texts.

GPTs are superior to regular chat sessions in the following ways [34]:

1. **Pre-training on vast data:** They possess a wide range of knowledge and contextual understanding.
2. **Transformer architecture:** They generate high-precision, consistent responses by considering long-range dependencies.
3. **Adaptability to various tasks:** They are capable of text generation, summarization, translation, and answering questions.

We used pretrained GPT models without fine-tuning and set detailed conditions to achieve consistent processing. All prompts and review materials were created and processed in Japanese, considering the improved multilingual capabilities of the GPT-4o model and the fact that the research protocols were written in Japanese.

GPTs evaluation process

In this study, the following procedure was adopted to confirm the reproducibility of the ChatGPT

models:

1. Start a new chat session.
2. Select the model: Either GPT-4 or GPT-4o.
3. Select one of the four custom GPTs (GPTs1, GPTs2, GPTs3, GPTs4).
4. Upload the document for evaluation: Upload the research protocol or ICF to the selected custom GPT.
5. Output the results: Once the document is processed and results are output, initiate a new chat session. This ensures that each session is independent and free from the influence of previous sessions.
6. Repeat the process: Follow the steps as outlined below:
    - **GPTs1:** Conduct 10 runs with GPT-4, followed by 10 runs with GPT-4o.
    - **GPTs2:** Conduct 10 runs with GPT-4, followed by 10 runs with GPT-4o.
    - **GPTs3:** Conduct 10 runs with GPT-4, followed by 10 runs with GPT-4o.
    - **GPTs4:** Conduct 10 runs with GPT-4, followed by 10 runs with GPT-4o.

This process resulted in 80 evaluations across both the GPT-4 and GPT-4o models for each custom GPT. As a result of the above process, the values shown in Table 1 represent the agreement and partial agreement rates based on the results of 10 trials for each model. Specifically, using the four custom

GPTs listed in Table 2, we evaluated each model's performance in extracting and summarizing the research objectives and background, research design and methods, and advantages and disadvantages. Following this, we used only the GPT-4o model to compare the results of the GPTs with the output of the standard prompts, focusing on the extraction of advantages and analyzing the performance differences between the models and methods.

Standard prompt usage

Additionally, we conducted tasks to extract advantages from the research protocol using a regular ChatGPT interface without GPTs. In this approach, we instructed ChatGPT: "You are an IRB reviewer. Please extract the advantages for the participants from the uploaded research protocol and list them as bullet points." This mimicked a typical review process without specifying particular sections or information. The prompts were created entirely in Japanese and the review materials were also processed in Japanese.

Unlike GPTs, this method did not specify detailed settings or sections and was used to evaluate the diversity and variability of the output results. We used only the GPT-4o model for standard prompts, focusing on extracting the advantages for comparison with GPTs. This allowed us to assess the extent to which differences in accuracy and reproducibility arose between the use of the GPTs and standard prompts.

406

407 Standard prompt evaluation process

408 To confirm the reproducibility of the GPTs and standard prompts, the following procedure was adopted

409

410 1. Start a new chat session.

411 2. Select the GPT-4o model.

412 3. Enter the prompt into the input field.

413 4. Upload the document for evaluation: Upload the research protocol for evaluation.

414 5. Output Results: Once the document is processed and the results are output, start a new chat session.

415 This ensures that each session is independent and free from the influence of previous sessions.

416 6. Repeat the process: Repeat the above steps 10 times.

417

418 Supplementary file 1 includes key sections of the research protocol translated from Japanese

419 to English, as well as a list of the prompts created in Japanese and their corresponding output

420 results, also translated from Japanese to English.

421

422 Confirmation of accuracy and reproducibility

423 For the two categories, "Research Objectives and Background" and "Research Design and Methods,"

424 we confirmed whether the 10 outputs from each model were consistent with the content of the research

425 protocol, evaluating both accuracy and reproducibility. The evaluation criteria were as follows:

426

427 First, we established consistency criteria, as shown in Table 6, and used them to assess whether
428 ChatGPT's output matched the content of the research protocol. These criteria included content
429 accuracy, meaning consistency, and expression consistency. For content accuracy, we confirmed
430 whether the specific numbers and data were completely accurate, whether all key information related
431 to the research objectives and background was included, and whether all key information related to the
432 research design and methods was present. For meaning consistency, we ensured that the original
433 document’s intent was accurately conveyed. For expression consistency, we confirmed that even if the
434 phrasing or word choice differed, the meaning and intent remained the same. For example, phrases
435 like "the number of patients increased" and "patient numbers increased," or "the effectiveness of the
436 treatment was confirmed" and "treatment effectiveness was confirmed" would be considered
437 synonymous.

438

439 Based on the consistency criterion, we classified each trial output as "consistent," "partially
440 consistent," or "inconsistent" using the accuracy evaluation criteria in Table 7. "Consistent" was
441 defined as output content being completely in line with the main content of the research protocol.
442 "Partially consistent" referred to cases where there were minor inconsistencies, but the main
443 information or intent was mostly aligned. "Inconsistent" was when the major content did not match,

compromising the accuracy of the information.

Subsequently, reproducibility was evaluated based on the results of 10 trials. According to the reproducibility evaluation criteria in Table 8, if 8 or more of the 10 trials were deemed "consistent" or "partially consistent," reproducibility was rated as "high." If 5-7 trials were deemed "consistent" or "partially consistent," reproducibility was rated as "moderate." If 4 or fewer trials met the criteria, reproducibility was rated as "low."

Through this evaluation process, the accuracy and reproducibility of ChatGPT's output were thoroughly assessed.

Table 8. Criteria for reproducibility evaluation

Reproducibility level	Description
High	When 8 or more out of 10 trials are judged as "consistent" (including "consistent" and "partially consistent" in Table 7).
Moderate	When 5 to 7 out of 10 trials are judged as "consistent" (including "consistent" and "partially consistent" in Table 7).
Low	When 4 or fewer out of 10 trials are judged as "consistent" (including "consistent" and "partially consistent" in Table 7).

456

457 # Accuracy and reproducibility of advantages and disadvantages

458 Next, we used ChatGPT to extract the advantages and disadvantages from the "Benefits and Risks"
459 section of the research protocol and evaluated their accuracy and reproducibility. The evaluation
460 criteria were as follows

461

462 The advantages and disadvantages of the research protocol are listed in Table 9, whereas those of the
463 ICF are listed in Table 10. We evaluated whether the content extracted by ChatGPT matched these
464 entries based on the "Consistency" criteria described in Table 6.

465

466 **Table 9. Advantages and disadvantages listed in the research protocol**

Item	Description
Advantage 1	In addition to regular treatment, more detailed information collection and follow-up care can be conducted, leading to more appropriate and thorough treatment.
Advantage 2	The investigational drug, by recognizing HER2 gene amplification, may reveal its impact on the treatment of advanced solid tumors, potentially increasing treatment options.
Disadvantage 1	A larger amount of blood collection than usual is expected.
Disadvantage 2	Administration of the investigational drug may cause side effects.

467

Table 10. Advantages and disadvantages listed in the ICF

Item	Description
Advantage 1	If the investigational drug Trastuzumab Biosimilar X proves effective, there is a possibility that the progression of cancer may temporarily halt, or symptoms may improve.
Advantage 2	In the future, if Trastuzumab Biosimilar X proves effective in recognizing HER2 gene amplification in solid tumors, it may increase treatment options.
Disadvantage 1	A larger amount of blood collection than usual is expected.
Disadvantage 2	There is a possibility of side effects from the administration of Trastuzumab Biosimilar X.
Disadvantage 3	Compared to regular treatment, there may be an increase in hospital visits, the duration of hospital stays, and the number of tests conducted.

Additionally, we performed the same task of extracting the advantages and disadvantages from the research protocol using a standard ChatGPT session without using GPTs. In the standard ChatGPT session, the user simply instructed, "As an IRB reviewer, please extract the advantages for participants from the attached research protocol and output them as bullet points," without setting any detailed task-specific configurations.

This comparison allowed us to assess the differences in the information extracted when using GPTs versus when not using them, thereby evaluating the utility of GPTs in complementing the IRB review

process.

Acknowledgments

We would like to express our gratitude to Associate Professor Kenji Hirata from the Department of Diagnostic Imaging, Graduate School of Medicine, Hokkaido University, for providing valuable insights related to this project. We also thank Professor Shiro Hinotsu from the Division of Biostatistics and Data Management, Sapporo Medical University, for dedicating his time to discuss this topic with us. Finally, we would like to thank Editage (www.editage.jp) for English language editing.

References

1. 厚生労働省，第 2 回　厚生科学審議会　臨床研究部会　議事録
[cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/shingi2/0000179030.html

2. 厚生労働省，第 10 回厚生科学審議会臨床研究部会　議事録
[cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/newpage_03963.html

3. 厚生労働省，第 10 回厚生科学審議会臨床研究部会資料 1 ：臨床研究・治験の推進に係る論点整理
[cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/newpage_03565.html

4. 厚生労働省，第 34 回厚生科学審議会臨床研究部会　議事録[cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/newpage_38195.html

5. 厚生労働省，第 34 回厚生科学審議会臨床研究部会　資料 1 ：認定臨床研究審査委員会について [cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/newpage_37286.html

6. 厚生労働省，第 35 回厚生科学審議会臨床研究部会　議事録 [cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/newpage_43555.html

7. 厚生労働省，第 35 回厚生科学審議会臨床研究部会　資料 1 － 6 ：認定臨床研究審査委員会について [cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/newpage_42147.html

8. Larson E, Bratts T, Zwanziger J, Stone P. A Survey of IRB process in 68 U.S. hospitals. J Nurs Scholarsh. 2004;36:260-264. doi: 10.1111/j.1547-5069.2004.04047.x.

9. Dyrbye LN, Thomas MR, Mechaber AJ, Eacker A, Harper W, Massie FS Jr, et al. Medical education research and IRB review: An analysis and comparison of the IRB review process at six institutions. Acad Med. 2007;82:654-660. doi: 10.1097/ACM.0b013e31806747b8.

10. Abbott L, Grady C. A Systematic Review of the Empirical Literature Evaluating IRBs: What we know and what we still need to learn. J Empir Res Hum Res Ethics. 2011;6:3-19. doi: https://doi.org/10.1525/jer.2011.6.1.3

11. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med. 2023;388:1201–1208. doi: 10.1056/NEJMp2301212.

12. Nazario-Johnson L, Zaki HA, Tung GA. Use of large language models to predict neuroimaging. J Am Coll Radiol. 2023;20:1004-1009. doi: 10.1016/j.jacr.2023.06.008.

13. Lee KH, Lee RW, Kwon YE. Validation of a deep learning chest X-ray interpretation model: integrating large-scale AI and large language models for comparative analysis with ChatGPT. Diagnostics (Basel). 2023;14:90. doi: 10.3390/diagnostics14010090.

14. Denecke K, May R, Rivera Romero O, LLM Health Group. Potential of large language models in health care: Delphi Study. J Med Internet Res. 2024;26:e52399. doi: 10.2196/52399.

15. Hirosawa T, Harada Y, Tokumasu K, Ito T, Suzuki T, Shimizu T. Evaluating ChatGPT-4's diagnostic accuracy: Impact of visual data integration. JMIR Med Inform. 2024;12:e55627. doi: 10.2196/55627.

16. Zhu Q, Chen X, Jin Q, Hou B, Mathai TS, Mukherjee P, et al. Leveraging professional radiologists' expertise to enhance LLMs' evaluation for radiology reports. ArXiv [Preprint]. 2024:arXiv:2401.16578v3. doi: 10.48550/arXiv.2401.16578.

17. Al Nazi Z, Peng W. Large language models in healthcare and medical domain: A review. Comput Lang. 2024. doi: 10.48550/arXiv.2401.06775.

18. Naveed H, Khan AU, Mian A, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. arXiv. 2023. doi: 10.48550/arXiv.2307.06435.

19. OpenAI. Introducing ChatGPT [cited 2024 Sep 13]. Available from: https://openai.com/blog/chatgpt.

20. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. doi: 10.2196/45312. Erratum in: JMIR Med Educ. 2024;10:e57594. doi: 10.2196/57594.

21. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: Evaluation Study. JMIR Form Res. 2023;7:e48023. doi: 10.2196/48023.

554 22. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, et al. Performance of generative pretrained
555 transformer on the national medical licensing examination in Japan. PLOS Digit Health. 2024;3:e0000433. doi:
556 10.1371/journal.pdig.0000433.
557
558 23. 厚生労働省，第 2 回　厚生科学審議会　臨床研究部会　資料 3　認定臨床研究審査委員会について
559 [cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/shingi2/0000175731.html
560
561 24. 厚生労働省，第 5 回　厚生科学審議会　臨床研究部会　議事録
562 [cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/shingi2/0000188562.html
563
564 25. 厚生労働省，第 5 回　厚生科学審議会　臨床研究部会　資料 2　認定臨床研究審査委員会について
565 [cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/shingi2/0000184433.html
566
567 26. Sohn E. The reproducibility issues that haunt health-care AI. Nature. 2023;613:402-403. doi:
568 10.1038/d41586-023-00023-2.
569 27. 厚生労働省，第 8 回厚生科学審議会臨床研究部会　議事録
570 [cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/newpage_02822.html
571
572 28. 厚生労働省，第 8 回厚生科学審議会臨床研究部会　参考資料 3 ：臨床研究法の施行状況等について

[cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/newpage_02524.html

29. 厚生労働省，第 9 回厚生科学審議会臨床研究部会　議事録

[cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/newpage_03737.html

30. 厚生労働省，第 9 回厚生科学審議会臨床研究部会　資料 3 ：臨床研究・治験活性化等に関する取組等について

[cited 2024 Sep 13]. Available from: https://www.mhlw.go.jp/stf/newpage_03163.html

31. 模擬審査資料　模擬審査における認定委員会の審査資料一式について（平成 31 年 3 月 28 日厚生労働省威勢曲研究開発振興課事務連絡）令和元年模擬審査資料（課題 1～5），令和元年模擬審査資料（課題 6～10）[cited 2024 Apr 9]. Available from: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000163417.html

32. OpenAI. GPT-4 is OpenAI’s most advanced system, producing safer and more useful responses [cited 2024 Sep 13]. Available from: https://openai.com/product/gpt-4.

33. OpenAI. Hello GPT-4o　[cited 2024 Sep 13]. Available from: https://openai.com/index/hello-gpt-4o/

34. OpenAI. Introducing GPTs　[cited 2024 Sep 13]. Available from: https://openai.com/index/introducing-gpts/