InformGen: An AI Copilot for Accurate and Compliant Clinical Research Consent Document Generation

Zifeng Wang[1], Junyi Gao[2], Benjamin Danek[1], Brandon Theodorou[1], Ruba Shaik[3], Shivashankar Thati[1], Seunghyun Won[4], Jimeng Sun[1,3#]

[1] School of Computing and Data Science, University of Illinois Urbana-Champaign, Urbana, IL, USA

[2] Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, UK

[3] Carle Illinois College of Medicine, University of Illinois Urbana-Champaign, Champaign, IL, USA

[4] Seoul National University Bundang Hospital, Gyeonggi, Republic of Korea

[#]Corresponding author. Emails: jimeng@illinois.edu

Abstract

Leveraging large language models (LLMs) to generate high-stakes documents, such as informed consent forms (ICFs), remains a significant challenge due to the extreme need for regulatory compliance and factual accuracy. Here, we present InformGen, an LLM-driven copilot for accurate and compliant ICF drafting by optimized knowledge document parsing and content generation, with humans in the loop. We further construct a benchmark dataset comprising protocols and ICFs from 900 clinical trials. Experimental results demonstrate that InformGen achieves near 100% compliance with 18 core regulatory rules derived from FDA guidelines, outperforming a vanilla GPT-4o model by up to 30%. Additionally, a user study with five annotators shows that InformGen, when integrated with manual intervention, attains over 90% factual accuracy, significantly surpassing the vanilla GPT-4o model's 57%-82%. Crucially, InformGen ensures traceability by providing inline citations to source protocols, enabling easy verification and maintaining the highest standards of factual integrity.

Introduction

Informed consent forms (ICFs) constitute a fundamental pillar of ethical clinical research, functioning as the primary medium through which participants comprehend study procedures, risks, benefits, and their rights.[1] Clinical trials necessitate multiple ICF iterations to accommodate site-specific requirements, institutional review board stipulations, and linguistic-policy adaptations for multinational studies.[2] The ICF development process requires a comprehensive analysis of clinical trial protocols, which often span tens to hundreds of pages. This resource-intensive endeavor significantly impacts clinical research efficiency, potentially delaying participant enrollment and extending study timelines,[3] while simultaneously compromising trial inclusivity through suboptimal ICF quality and accessibility.[4]

Large language models (LLMs), trained on vast textual corpora, initially emerged as tools for narrative text generation applications.[5] Their integration into medicine has advanced substantially, facilitating clinical text generation and summarization[6] while enabling patient-oriented conversational interfaces.[7] Within clinical trials research, investigators have adapted LLMs for protocol development tasks, including the formulation of eligibility criteria and generation of trial synopses.[8,9] Recent investigations adopted LLMs to simplify ICFs to facilitate their accessibility.[10,11] Besides, a particularly promising method is retrieval-augmented generation (RAG),[12] wherein LLMs dynamically incorporate reference documents during text generation. RAG enables LLMs to process extensive source documentation while producing outputs with explicit citations grounded in sources, thereby facilitating human review and verification.[13]

Despite the advancements, the application of LLMs for ICF drafting remains largely unexplored, presenting several critical challenges. First, regulatory compliance poses a major hurdle, as ICFs must adhere to FDA guidelines on structure and language,[14] as well as institution-specific policies that vary across research sites.[15,16] Second, clinical trial documents demand exceptional information fidelity, requiring strict adherence to trial protocols while ensuring content remains verifiable by human reviewers, an aspect insufficiently examined in prior research. Third, there is a lack of dedicated evaluation frameworks for AI-generated regulatory documents. Traditional natural language processing (NLP) metrics, such as BLEU scores,[17] primarily assess text overlap and semantic similarity but fail to capture critical ICF-specific quality dimensions. Besides, existing AI-driven approaches for ICFs predominantly focus on isolated aspects, such as readability,[18–20] rather than addressing the full spectrum of requirements for regulatory compliance and content reliability. In particular, a recent study found that none of the LLMs tested, including GPT-4 and Gemini, were capable of generating a legally compliant ICF and only covered fewer than one-third of the required risks, procedural details, and preparatory instructions.[21]

In this work, we make three key contributions: (1) InformBench, a benchmark dataset comprising over 900 paired clinical trial protocols and their corresponding ICFs, representing the largest AI-ready ICF drafting dataset to date; (2) an evaluation framework aligned with FDA guidelines[14] to systematically assess compliance and factual accuracy, leveraging both automated and human evaluations for benchmarking LLM performance; and (3) InformGen, a specialized LLM pipeline that enhances human-AI collaboration in ICF generation. Our evaluation demonstrates that this approach achieves near-perfect regulatory compliance (in most cases 100%), maintains high factual accuracy (80-95%), and offers variable and editable evidence, within a human-in-the-loop framework.

Results

Overview of InformGen

Figure 1 presents an overview of the development and evaluation of the InformGen framework for informed consent form (ICF) drafting. The name InformGen reflects three aspects of "informed" principles in this study. First, it is knowledge- and policy-informed, as it

integrates input from clinical trial protocols, FDA regulations, and site-specific policies to guide document generation. Second, it is human-informed, enabling seamless verification and correction of both inputs and outputs within the pipeline. Finally, it is designed specifically as an AI copilot for informed consent form generation. Specifically, InformGen consists of three key stages: (1) knowledge document parsing, (2) document drafting, and (3) content evaluation (See "Methods").

Knowledge document parsing processes two main types of documents prior to ICF generation as illustrated in Figure 1a. (see "Methods"). The first type includes the *guideline document* and *consent template*, which provide regulatory guidelines and site-specific policies, respectively. The guideline document applies universally to all ICFs and requires parsing only once, while the consent template is site-specific and remains applicable to all future ICFs from the same site. These documents are parsed and transformed into structured rules that are attached to relevant ICF sections. The second type of document is the *clinical trial protocol*, which is essential for each individual clinical trial and contains critical information required for drafting the ICF. Specifically, two procedure-specific components are extracted to inform trial participants: the trial schedule and the risks associated with experimental procedures. To ensure accuracy, we take an additional step to extract and standardize the *Schedule of Assessments* (SOA) table and the *procedure-specific risk* table for manual inspection before incorporating them into the ICF generation process. This step is necessary because a standard retrieval-augmented generation (RAG) approach often struggles to capture nuanced details when processing lengthy protocol documents spanning hundreds of pages.

Figure 1b illustrates the document drafting pipeline, which follows a RAG paradigm to ensure that all generated content is grounded in authoritative knowledge documents. Each generated section includes citation links indicating the exact source paragraph and section, enabling human verification. Aligned with site-specific policies, the pipeline incorporates section-level guidelines as directives and retrieves relevant reference text chunks to generate the corresponding ICF sections. For key sections such as *"Procedures"*, the pipeline integrates the *Schedule of Assessments* (SOA) table as an additional input, while for *"Risks and Discomforts"*, it incorporates the procedure-specific risk table to enhance accuracy and compliance.

Overview of InformBench

Figure 1c illustrates the content evaluation framework, which assesses the generated content along two key dimensions. First, *compliance* is evaluated by measuring the proportion of regulatory and site-specific rules that are correctly followed. We provide the parsed rules in Supplementary Figure 1. Second, *factual accuracy* is assessed by calculating the ratio of key facts correctly covered in the generated content, based on reference human-written ICFs and the corresponding trial protocols.

To comprehensively evaluate both generic LLMs and the optimized InformGen framework in terms of compliance and factual accuracy, we constructed the InformBench benchmark. The dataset was built by first retrieving all available clinical trial protocols and ICFs from ClinicalTrials.gov. Among these, oncology-related ICFs exhibited higher quality in terms of formatting and adherence to regulatory standards. Consequently, we selected a subset primarily focused on oncology-related conditions, yielding a total of 900 trials with at least one available ICF and corresponding protocol document. The dataset statistics are presented in Figure 2. The included trials span all common cancer types[22] and all four clinical trial phases. Most protocol documents range from 50 to 100 pages, while ICFs typically contain 10 to 20 pages. A small subset of lower-quality protocols, often associated with trials labeled as having an "Unknown" phase, contain fewer than 10 pages. In subsequent sections, we analyze how the quality and complexity of input protocols impact the quality of ICF generation.

Guided by an ICF template from Stanford,[15] we focused on four critical and challenging sections: *"Purpose of Research"*, *"Procedures"*, *"Duration of Study Involvement"*, and *"Risks and Discomforts"*. Content quality was assessed specifically for these sections.

Given the diverse template formats across ICFs, we applied a standardization approach to categorize each parsed section into one of the four sections or mark it as not applicable.

Evaluation of compliance

Figure 3 presents the compliance evaluation results of InformGen and baseline models. As previously discussed, we focused on four key ICF sections, each associated with multiple regulatory rules (Supplementary Figure 1). To assess compliance, we conducted an automatic evaluation using GPT-4o on all ICFs, complemented by a manual evaluation on a subset of 100 trials. For each section in each trial, we evaluated whether the generated content adhered to each applicable rule, categorizing it as either met or violated. The compliance score was then computed as the proportion of rules successfully met. In the human evaluation, we introduced an intermediate category, "Partial Yes", allowing annotators to indicate cases where a rule was only partially satisfied. For comparison, we implemented a baseline method that follows a standard RAG approach, supplemented with general guidance on drafting the target section (referred to as "prompt"). We denote this approach as *RAG+Prompting*. Notably, both InformGen and RAG+Prompting utilize the same underlying LLM, either GPT-4o or GPT-4o-mini, to ensure a fair comparison.

Figure 3a presents the automatic evaluation results comparing InformGen and the baseline. We tested two variants of each method using GPT-4o and GPT-4o-mini, where the former represents a state-of-the-art LLM and the latter a lightweight version. The results indicate that InformGen, with explicit regulatory guidance and site-specific policies, achieved near 100% compliance in most cases, regardless of the underlying LLM. In contrast, the RAG+Prompting baseline met compliance requirements only for the *"Purpose of Research"* and *"Duration of Study Involvement"* sections. The rules governing these sections are relatively straightforward, such as clearly stating the purpose of the clinical study and specifying its expected duration (Supplementary Figure 1). However, RAG+Prompting exhibited lower compliance rates for the *"Procedures"* section (82%-90%) and the *"Risks and Discomforts"* section (73%-75%). For *"Procedures"*, RAG+Prompting often failed to explicitly distinguish experimental procedures unique to the research, a key regulatory requirement. In the *"Risks and Discomforts"* section, failure modes were more varied, reflecting the complexity of accurately identifying risks such as potential privacy loss, reasonably foreseeable discomforts, and the need to balance comprehensiveness without overwhelming the reader (Supplementary Figure 1). These challenges highlight the advantages of an optimized InformGen framework in mitigating such issues and ensuring higher compliance.

Figure 3b provides a trial-level comparison of InformGen and *RAG+Prompting*, further broken down by clinical trial phases. Across all phases, InformGen consistently outperforms the baseline, achieving a winning rate of over 90% in most cases. Additionally, InformGen maintains a compliance rate exceeding 90% across the majority of trials, reinforcing the overall trend observed in the average performance analysis.

Figure 3c presents the confusion matrix comparing the alignment between automatic evaluation using GPT-4o and human evaluation. This analysis validates the reliability of using GPT-4o as an automated compliance judge. To ensure consistency with human annotators, GPT-4o was tasked with assessing rules using the same categories: "Yes", "No", or "Partial Yes". Human assessments were treated as the ground truth. The results indicate that GPT-4o tended to be more stringent than human annotators, particularly for the *"Purpose of Research"* and *"Duration of Study Involvement"* sections, where the true negative rate (TNR) was noticeably lower than the true positive rate (TPR). For *"Procedures"* and *"Risks and Discomforts"*, GPT-4o also demonstrated strong agreement with human evaluations, achieving a TPR above 80% and a TNR exceeding 85%. These findings suggest that GPT-4o serves as a reliable compliance judge. Importantly, since this evaluation method does not require a ground truth ICF, it can be applied as a post-checking mechanism for any newly drafted ICF to ensure maximum compliance.

Figure 3d compares the compliance performance of InformGen and RAG+Prompting across four key ICF sections, evaluated by both human annotators and GPT-4o on 100 clinical trials. InformGen consistently outperforms the baseline, achieving near-perfect

compliance in *"Purpose of Research"* and *"Duration of Study Involvement"* according to both evaluations. In contrast, RAG+Prompting shows noticeably lower compliance, particularly in the *"Duration of Study Involvement"* section. The gap between the methods is even more pronounced in the *"Procedures"* and *"Risks and Discomforts"* sections, where InformGen maintains high compliance, while RAG+Prompting struggles significantly, with compliance rates nearly halved. These results underscore InformGen 's effectiveness in adhering to regulatory requirements, particularly in sections requiring precise procedural and risk-related details.

Evaluation of factual accuracy

Research consent forms impose stringent requirements on the accuracy of the information conveyed to clinical trial participants, with critical legal and ethical implications.[23] However, LLMs have been observed to occasionally fabricate or omit essential trial-specific details, such as procedural risks and study schedules, because of the complexity and volume of the input knowledge documents like protocols. As such, we establish an evaluation framework for factual accuracy. For each ICF section, we first extract up to five key facts from human-written reference ICFs. The generated content is then assessed against these reference facts to determine whether each key fact is correctly represented (Figure 1c). The extracted key facts are tailored to the specifics of each section. For instance, in *"Duration of Study Involvement"*, key facts include the total study duration in months and the expected number of participant visits. This approach allows us to systematically measure the factual accuracy of the generated content.

Figure 4a presents the factual accuracy results of InformGen and RAG+Prompting across three key ICF sections: *"Risks and Discomforts"*, *"Procedures"*, and *"Duration of Study Involvement"*. The results represent the average accuracy per key fact across 900 clinical trials. InformGen achieves over 85% factual accuracy in the *"Risks and Discomforts"* and *"Procedures"* sections, demonstrating a significant improvement over RAG+Prompting by 10% and 30% in absolute accuracy, respectively. This highlights the effectiveness of incorporating the SOA table and procedure-specific risk table extraction. In contrast, RAG+Prompting often omits crucial details in risk and procedure descriptions due to retrieval limitations. For the *"Duration of Study Involvement"* section, the factual accuracy difference between the two methods is less pronounced. Analysis reveals that errors primarily stem from ambiguities in defining the expected duration of active participation, even when using the SOA table. Additionally, inaccuracies introduced during protocol document parsing further contribute to errors. These findings underscore the importance of incorporating a human-in-the-loop approach to enhance factual accuracy.

A similar pattern emerges in the trial-level factual accuracy comparison (Figure 4b), where InformGen achieves a winning rate of 70%-80% for Phase 1 to Phase 3 trials. An exception is observed in *"Phase 4 or Unknown"* trials, where InformGen attains a lower winning rate of 54.4% over RAG+Prompting. Further analysis indicates that protocol quality negatively impacts the results, as 17.2% of *"Unknown"* phase trial protocols contain fewer than 10 pages, and 68.7% have fewer than 50 pages (Figure 2). This trend is further validated by the factual accuracy analysis based on protocol length (Figure 4d), which reveals a significant drop in accuracy for the *"Duration of Study Involvement"* section when protocol documents contain fewer than 10 pages. Additionally, Figure 4d highlights a general decline in performance as protocol length increases, though RAG+Prompting exhibits a steeper drop, while InformGen remains more stable. This discrepancy arises because vanilla retrieval approaches struggle to extract complete information from longer, more complex knowledge documents, whereas InformGen is more resilient to document length variations.

Figure 4c illustrates the agreement between human annotators and the GPT-4o judge (AI evaluation) in the fact-checking process. Treating human evaluation as the ground truth, we calculate the true positive rate (TPR) and true negative rate (TNR) for AI evaluations. The results show that AI evaluation achieves a TPR of 0.84–0.96 and a TNR of 0.79–0.95, demonstrating strong alignment with human assessments. Additionally, we compare the absolute factual accuracy between human and AI evaluations for 100 clinical

trials (Figure 4e). The overall accuracy is similar between the two evaluation methods, though AI evaluation tends to be more stringent. InformGen consistently outperforms RAG+Prompting across all three sections, with the largest performance gap observed in the *"Procedures"* section, where InformGen achieves 75% accuracy compared to only 30% for the baseline. These results highlight the effectiveness of the optimized InformGen pipeline in enhancing factual accuracy for generated content.

Human intervention improves compliance and factual accuracy

Previous experiments have demonstrated the superiority of InformGen over the standard RAG-plus-prompting approach in ensuring regulatory compliance and improving factual accuracy in ICF generation. While compliance issues have been effectively addressed, a gap remains in achieving perfect factual accuracy, particularly in the *"Duration of Study Involvement"* section, representing the "last mile" for making this tool fully practical. Notably, all prior experiments were conducted under a fully automated setup, without human intervention. In real-world applications, human experts can be incorporated into the process to review and correct errors, further enhancing accuracy. To explore this potential, we have developed a demo platform that integrates human oversight, as illustrated in Figure 5, and evaluated to what extent the lift it can bring.

Figure 5a presents example parsing results for two critical knowledge documents: the procedure-specific risk table and the schedule of assessment (SOA) table. Before InformGen proceeds with content generation, human users can review these intermediate outputs to ensure accuracy. Each procedure-risk pair, along with its frequency description, is linked to the corresponding source pages in the trial protocols, allowing users to verify the original context and make necessary edits. Similarly, the SOA table is standardized into a machine-readable format and linked to the extracted table from the protocol for detailed inspection. Once approved, these documents can be consistently applied across all ICFs for the specific trial. As illustrated in Figure 5b, the generated ICF content includes inline citations that link directly to the source content, with key references highlighted to indicate their role in LLM-based generation. This feature enables users to efficiently verify content against source documents and make necessary refinements. Additionally, users can leverage the AI copilot to assist in rewriting or refining the initial draft, ensuring both accuracy and clarity.

Figure 5c presents the manually evaluated factual accuracy results for 20 clinical trials, comparing two methods: RAG+Prompting and InformGen +Human. The former represents the baseline approach, while the latter incorporates human intervention to refine the SOA and procedure-risk tables before generating content for the three key ICF sections. The results demonstrate the effectiveness of human-in-the-loop refinement, with InformGen achieving up to 90% factual accuracy when using GPT-4o as the underlying LLM. This represents a substantial absolute accuracy improvement of 15% to 30%, with high statistical significance. Moreover, errors in the generated content can be directly traced back to the extracted tables, allowing for efficient post-processing corrections. In contrast, the standard RAG approach struggles to identify missing content, often requiring a full proofread of the entire protocol. Additionally, we observed that GPT-4o-mini has difficulty accurately calculating duration-related values, suggesting that a more powerful base LLM is preferable for optimal performance.

Discussion

Informed consent forms (ICFs) are essential documents for participant enrollment and onboarding in clinical research, subject to strict compliance and factual accuracy requirements to uphold ethical standards. Large language models (LLMs) have been utilized for drafting clinical trial documents[8, 9, 24] and for simplifying ICFs to enhance accessibility.[10, 11] However, the application of LLMs for ICF drafting remains largely unexplored, and no standardized framework exists for rigorously evaluating compliance and factual accuracy in LLM-generated clinical trial documents, particularly in ICFs. To address this

gap, we constructed a dataset comprising protocols and ICFs from 900 clinical trials. We systematically parsed these documents to develop a benchmark dataset, InformBench, along with a comprehensive evaluation framework to assess both compliance with regulatory standards and factual accuracy in LLM-generated ICF content. This evaluation framework can be executed by human experts but also by LLM as a judge, which, through our experiments, is highly consistent with how human experts' preferences.

On InformBench, we implemented LLM-based ICF drafting using GPT-4o and GPT-4o-mini, following the standard approach of retrieval-augmented generation (RAG)[12] combined with prompt engineering.[25] Our experiments indicate that this RAG+Prompt method, even when leveraging state-of-the-art LLMs, falls short in meeting compliance and factual accuracy requirements, highlighting significant challenges in adopting LLMs for high-stakes document drafting in clinical trials. To address these limitations, we developed InformGen, an optimized ICF drafting pipeline designed to enhance both compliance and factual accuracy while incorporating human oversight. We compared InformGen against RAG+Prompting using the same underlying LLMs on InformBench. The results demonstrate the superiority of InformGen: it achieved near 100% compliance, ensuring adherence to regulatory requirements for language and content completeness. Additionally, InformGen outperformed RAG+Prompting in factual accuracy, with an absolute improvement of 10%-20% in the *"Risks and Discomforts"* and *"Procedures"* sections.

In a fully automated generation setup, we observed that InformGen did not show a significant improvement in the *"Duration of Study Involvement"* section. To address this, we developed a demo platform that enables human interaction with InformGen, allowing users to review and refine intermediate knowledge documents, including the procedure-specific risk table and the schedule of assessment (SOA) table extracted from clinical trial protocols. Using this platform, we conducted a user study on 20 clinical trials, where participants could verify and modify these extracted documents before ICF generation. The results demonstrated a substantial benefit from human intervention, with InformGen achieving an absolute accuracy improvement of 20%-30% and reaching an overall accuracy of 80%-95%, including significant gains in the study duration section. This has highlighted the benefit and necessity in enhancing human-AI interaction when drafting high-stakes clinical trial documents such as ICFs.

This study has several limitations. (1) While InformGen is designed for clinical researchers, enhancing the experience for research participants remains an important consideration, particularly in addressing linguistic and regional diversity.[4] (2) Although InformGen focuses on consent documents in clinical research, its application could be expanded to other clinical research documents and consent forms beyond clinical trials, such as surgical consent forms,[11] which remain largely unexplored. (3) The use of generic LLMs raises privacy concerns, as it requires data sharing with third-party model providers. Recent advancements in developing specialized LLMs based on open-source models[26] present a promising direction for improving security while leveraging clinical data to enhance model performance. (4) A larger-scale user study is needed to assess not only compliance and factual accuracy but also the practical benefits, including real-time efficiency gains in clinical workflows.

LLMs have made significant advancements in medical AI applications. InformGen highlights the critical need for adapting generic LLMs to clinical research while ensuring regulatory compliance and factual accuracy, while also incorporating human expertise in the loop. We anticipate that this paradigm will drive further research and expand LLM applications in clinical research.

Methods

Collection and Building of InformBench

To construct InformBench, we retrieved the full ClinicalTrials.gov database from the AACT database,[27] which contains over 500,000 clinical trials. We first applied filters to identify oncology-related trials, starting with common cancer types[22] and using GPT-4o

to generate an expanded set of keywords for each type. These keywords were then used to match target conditions in the "conditions" table. Next, we extracted trial-provided documents, including protocols and informed consent forms, from the "provided_documents" table. Since not all trials included these documents, we retained 900 oncology trials with available materials. All retrieved documents were in PDF format and were converted into machine-readable markdown format using AWS Textract,[28] an optical character recognition (OCR) service. Textract also facilitated section-wise segmentation of ICF and protocol content, which we leveraged to structure the dataset into samples with paired input protocol content and the output sections of ICFs.

The collected ICFs exhibit diverse structures, as they follow different templates and vary in quality. To enable a fair comparison across clinical trials, we standardized section classification. Specifically, for each parsed raw section, we utilized GPT-4o-mini to classify it into one of the predefined target sections. The target sections were determined based on the most critical components of the Stanford ICF template,[15] including *"Purpose of Research"*, *"Procedures"*, *"Duration of Study Involvement"*, and *"Risks and Discomforts"*. Sections classified under the same target category were then merged to form a standardized reference ICF section, which serves as the basis for evaluation.

FDA establishes mandatory regulatory requirements for all ICFs. To systematically evaluate compliance, we developed a set of rules based on the FDA guidelines for informed consent,[14] specifically focusing on the *"Basic Elements of Informed Consent"* section. This section outlines the essential components that an ICF must include, along with their intended purpose and explanations. We extracted and structured these components into a rule table, where each rule is mapped to a corresponding standard ICF section, enabling its use as a reference for compliance evaluation. The complete rule table is presented in Supplementary Figure 1. To assess content compliance, we provide the rule name and description as input to the GPT-4o judge, prompting it to determine whether each rule is followed or violated, producing a binary compliance assessment.

To generate labels for factual accuracy evaluation, we leveraged the ICF template, which outlines the key elements that must be included in each section. For example, in the *"Purpose of Research"* section, the template specifies language such as: *"This research study is looking for [state number of people] with [disease or condition]. [Clarify if enrollment will occur throughout the United States or internationally]. Stanford University expects to enroll [state number] research study participants."* Based on these guidelines, we developed section-specific criteria for extracting key information. In this case, relevant elements include the total number of participants, target conditions, and primary trial locations. We then provided these structured extraction guidelines to GPT-4o, enabling large-scale automated extraction across all sections of the 900 clinical trials in the dataset.

Clinical trial protocol parsing

Before drafting a site-specific ICF for a trial, the first step is to comprehend the clinical trial protocol, as it serves as the foundation for study design, treatment schedules, and potential risks. Clinical trial protocols typically span tens to hundreds of pages, making it time-consuming even for experienced experts to extract relevant information for ICF drafting, as key details are often dispersed throughout the document. To address this challenge, we adopt a retrieval-augmented generation (RAG) approach, beginning by embedding the protocol into a vector database. This enables efficient retrieval of relevant information, ensuring that ICF drafting is grounded in the most pertinent trial details.

To process a protocol PDF, we first utilized Textract,[28] an OCR service from AWS, to convert the raw file into Markdown, a machine-readable format that includes basic annotations of the page and section structure. Next, we segmented the Markdown file into text chunks using LangChain's `MarkdownHeaderTextSplitter`,[29] associating each chunk with its corresponding first- and second-level headers and page numbers as metadata. We then applied OpenAI's embedding model, `text-embedding-3-small`,[30] to convert each text chunk into a 128-dimensional embedding vector. These embeddings were stored in a vector database, with Amazon MemoryDB as the primary implementation, though the approach is compatible with other open-source alternatives. This setup enables real-time

retrieval, allowing any user query to be transformed into embeddings and matched against the most relevant protocol chunks efficiently.

Under the InformGen framework, two specialized knowledge documents require parsing: the schedule of assessment (SOA) table and the procedure-risk table. For SOA extraction, we developed an LLM-driven pipeline that begins by analyzing the raw protocol PDF to identify pages containing the SOA table. The selected pages, along with their parsed Markdown content, are then processed using GPT-4o with vision capabilities, following a structured extraction process: (1) detecting key procedures and time points, (2) extracting procedure-timepoint pairs along with corresponding activities, and (3) reconstructing the SOA table into a machine-readable format, such as an Excel spreadsheet, for user review. For the procedure-risk table, a similar pipeline is applied: (1) identifying pages mentioning experimental procedures, (2) extracting associated risks and discomforts for each procedure, including expected likelihood and relevant notes, and (3) reconstructing the procedure-risk table for user verification. This structured approach ensures accurate extraction and facilitates human oversight before integrating these documents into ICF drafting. The example output SOA table and procedure-risk table can be found in Figure 5a.

Document drafting

InformGen is designed to generate ICFs section by section while adhering to site-specific templates. To achieve this, we first parse the site templates into a structured format to provide clear guidance for LLMs. For each section in the template, we extract the content guidelines and, if necessary, break them down into multiple steps for better clarity. Each section is then associated with multiple content guidelines, specifying what should be included in the section. Additionally, we extract relevant keywords from these guidelines to generate search queries for retrieving pertinent information from the vector database (Figure 1a). This structured approach ensures that InformGen produces ICF content that aligns with site-specific requirements.

As illustrated in Figure 1b, InformGen utilizes the parsed consent template as input for site-specific policies and follows a structured pipeline for document generation: (1) generating a search query with metadata filters, (2) retrieving the most relevant chunks, and (3) drafting the section based on content guidance. In step (1), we leverage the table of contents from the target protocol, along with content guidance and relevant keywords, to prompt the LLM to synthesize a search query. This query includes metadata-based filtering to determine which sections of the protocol should be retrieved. This approach is particularly beneficial for lengthy protocols segmented into thousands of chunks, where a purely embedding-based ranking may overlook critical information. By incorporating metadata filtering, we significantly reduce the candidate set, improving retrieval precision and ensuring that essential content is effectively captured. In step (3), alongside generating content, we also instruct the model to produce inline citations linking to the retrieved input chunks. These citations enable users to trace the generated information back to the original protocol document for verification and review (Figure 5b).

InformGen incorporates specialized inputs for sections requiring precise details on schedules and risks associated with experimental procedures, specifically *"Procedures"*, *"Duration of Study Involvement"*, and *"Risks and Discomforts"*. For the first two sections, the SOA table is used as an additional input. To facilitate processing, we segment the SOA table by time points, converting it into structured paragraphs that describe the activities occurring at each stage. These paragraphs are then provided to the model, which expands them into readable ICF content. Similarly, for the *"Risks and Discomforts"* section, risks are categorized by procedure and transformed into paragraph form, allowing the LLM to generate well-structured and accurate content.

References

[1] Grady, C. Enduring and emerging challenges of informed consent. *New England Journal of Medicine* **372**, 855–862 (2015).

[2] Nishimura, A. *et al.* Improving understanding in the research informed consent process: a systematic review of 54 interventions tested in randomized control trials. *BMC medical ethics* **14**, 1–15 (2013).

[3] Koonrungsesomboon, N., Laothavorn, J. & Karbwang, J. Understanding of essential elements required in informed consent form among researchers and institutional review board members. *Tropical Medicine and Health* **43**, 117–122 (2015).

[4] Velez, M. A. *et al.* Consent document translation expense hinders inclusive clinical trial enrolment. *Nature* **620**, 855–862 (2023).

[5] Yuan, A., Coenen, A., Reif, E. & Ippolito, D. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, 841–852 (2022).

[6] Van Veen, D. *et al.* Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine* **30**, 1134–1142 (2024).

[7] Wan, P. *et al.* Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial. *Nature Medicine* **30**, 2878–2885 (2024).

[8] Wang, Z., Xiao, C. & Sun, J. Autotrial: Prompting language models for clinical trial design. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12461–12472 (2023).

[9] Lin, J., Xu, H., Wang, Z., Wang, S. & Sun, J. Panacea: A foundation model for clinical trial search, summarization, design, and recruitment. *medRxiv* 2024–06 (2024).

[10] Mirza, F. N. *et al.* Using chatgpt to facilitate truly informed medical consent. *NEJM AI* **1**, AIcs2300145 (2024).

[11] Ali, R. *et al.* Bridging the literacy gap for surgical consents: an ai-human expert collaborative approach. *NPJ Digital Medicine* **7**, 63 (2024).

[12] Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **33**, 9459–9474 (2020).

[13] Gao, T., Yen, H., Yu, J. & Chen, D. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2023).

[14] Food, Administration, D. *et al.* Informed consent guidance for irbs, clinical investigators, and sponsors. *Silver Spring, MD: Food and Drug Administration* (2023).

[15] Forms & consent templates. https://researchcompliance.stanford.edu/panels/hs/forms-templates. Accessed: 2025-03-14.

[16] Consent and assent form templates. https://irb.ucsf.edu/consent-and-assent-form-templates. Accessed: 2025-03-14.

[17] Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318 (2002).

[18] Decker, H. *et al.* Large language model- based chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA network open* **6**, e2336997–e2336997 (2023).

[19] Vaira, L. A. *et al.* Evaluating ai-generated informed consent documents in oral surgery: A comparative study of chatgpt-4, bard gemini advanced, and human-written consents. *Journal of Cranio-Maxillofacial Surgery* **53**, 18–23 (2025).

[20] Shi, Q. *et al.* Transforming informed consent generation using large language models: Mixed methods study. *JMIR Medical Informatics* **13**, e68139 (2025).

[21] Raimann, F. J., Neef, V., Hennighausen, M. C., Zacharowski, K. & Flinspach, A. N. Evaluation of ai chatbots for the creation of patient-informed consent sheets. *Machine learning and knowledge extraction* **6**, 1145–1153 (2024).

[22] Common cancer types. https://www.cancer.gov/types/common-cancers. Accessed: 2025-1-14.

[23] Allen, J. W., Schaefer, O., Porsdam Mann, S., Earp, B. D. & Wilkinson, D. Augmenting research consent: should large language models (llms) be used for informed consent to clinical research? *Research Ethics* 17470161241298726 (2024).

[24] Markey, N., El-Mansouri, I., Rensonnet, G., van Langen, C. & Meier, C. From rags to riches: Utilizing large language models to write documents for clinical trials. *Clinical Trials* (2025).

[25] Brown, T. *et al.* Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds.) *Advances in Neural Information Processing Systems*, vol. 33, 1877–1901 (Curran Associates, Inc., 2020). URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[26] Wang, Z. *et al.* A foundation model for human-ai collaboration in medical literature mining. *arXiv preprint arXiv:2501.16255* (2025).

[27] Improving public access to aggregate content of clinicaltrials.gov. https://aact.ctti-clinicaltrials.org/. Accessed: 2025-03-14.

[28] Amazon textract. https://aws.amazon.com/cn/textract/. Accesed: 2025-03-19.

[29] How to split markdown by headers. https://python.langchain.com/docs/how_to/markdown_header_metadata_splitter. Accessed: 2025-03-20.

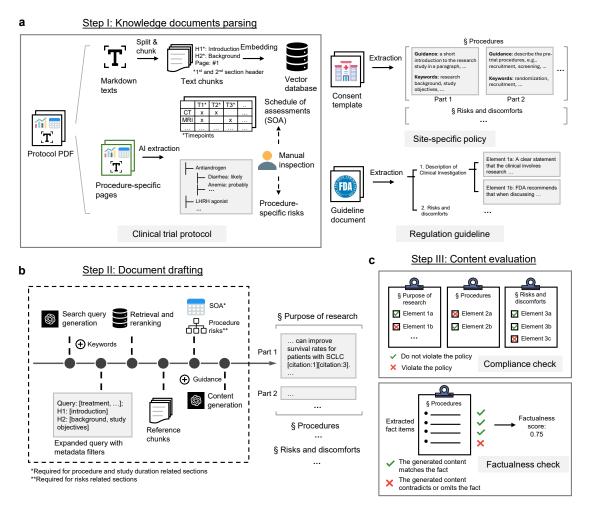[30] Vector embeddings. https://platform.openai.com/docs/guides/embeddings. Accessed: 2025-03-20.

Figure 1: **Overview of InformGen Workflow**. **a**, The clinical trial protocol is parsed as the primary knowledge document. The raw PDF is converted to Markdown text, segmented into chunks with metadata on section titles and page numbers, and encoded into dense embeddings for storage in a vector database. Additionally, two specialized knowledge documents, the schedule of assessment (SOA) table and the procedure-risk table, are extracted from protocols and subjected to human review before use in content generation to ensure high factual accuracy. On the right, the FDA guidelines for ICF drafting are processed to create regulatory policies for compliance evaluation, while the site-specific consent template is parsed to provide structured instructions for content generation. **b**, InformGen generates ICF sections using a retrieval-augmented generation pipeline. It first formulates search queries with metadata filters, retrieves the most relevant chunks, and applies parsed instructions to generate the content. For sections requiring trial schedules or risk-related details, the SOA or procedure-risk table is incorporated as an additional input. **c**, The generated content is evaluated for regulatory compliance and factual accuracy. Section-specific rule sets, derived from FDA guidelines, are used to assess whether the content violates any compliance requirements. Additionally, key facts are extracted from human-written reference ICFs as ground truth and compared against the generated content to compute an accuracy score.

Characteristics		Phase I	Phase II	Phase III	Phase IV	Unknown
n		200	515	56	15	326
Year, n (%)	1999-2010	15 (7.5)	45 (8.7)	7 (12.5)	0 (0.0)	9 (2.8)
	2010-2015	53 (26.5)	136 (26.4)	16 (28.6)	0 (0.0)	36 (11.0)
	2015-2020	95 (47.5)	261 (50.7)	22 (39.3)	10 (66.7)	164 (50.3)
	2020-2024	36 (18.0)	65 (12.6)	7 (12.5)	5 (33.3)	107 (32.8)
Region, n (%)	United States	187 (94.4)	481 (94.7)	46 (83.6)	12 (85.7)	215 (67.6)
	China	7 (3.5)	13 (2.6)	8 (14.5)	1 (7.1)	20 (6.3)
	Other	4 (2.0)	15 (3.0)	4 (7.3)	1 (7.1)	83 (26.1)
Enrollment, n (%)	<10	48 (24.0)	115 (22.3)	1 (1.8)	4 (26.7)	27 (8.3)
	10-100	147 (73.5)	339 (65.8)	13 (23.2)	10 (66.7)	161 (49.4)
	100-1000	5 (2.5)	59 (11.5)	35 (62.5)	1 (6.7)	112 (34.4)
	>1000	0 (0.0)	2 (0.4)	0 (0.0)	0 (0.0)	26 (8.0)
Protocol pages, n (%)	<10	6 (3.0)	13 (2.5)	3 (5.4)	4 (26.7)	56 (17.2)
	10-50	45 (22.5)	136 (26.4)	20 (35.7)	10 (66.7)	224 (68.7)
	50-100	107 (53.5)	295 (57.3)	20 (35.7)	1 (6.7)	43 (13.2)
	>100	42 (21.0)	71 (13.8)	13 (23.2)	0 (0.0)	3 (0.9)
ICF pages, n (%)	<10	19 (9.5)	56 (10.9)	12 (21.4)	5 (33.3)	209 (64.1)
	10-20	121 (60.5)	314 (61.0)	37 (66.1)	9 (60.0)	103 (31.6)
	20-30	43 (21.5)	112 (21.7)	4 (7.1)	1 (6.7)	12 (3.7)
	>30	17 (8.5)	33 (6.4)	3 (5.4)	0 (0.0)	2 (0.6)
Conditions, n (%)	Bladder Cancer	4 (2.0)	13 (2.5)	0 (0.0)	2 (13.3)	6 (1.8)
	Breast Cancer	26 (13.0)	59 (11.5)	6 (10.7)	1 (6.7)	56 (17.2)
	Colorectal Cancer	16 (8.0)	23 (4.5)	1 (1.8)	1 (6.7)	31 (9.5)
	Endometrial Cancer	0 (0.0)	6 (1.2)	0 (0.0)	0 (0.0)	2 (0.6)
	Kidney Cancer	5 (2.5)	19 (3.7)	1 (1.8)	1 (6.7)	5 (1.5)
	Leukemia	25 (12.5)	49 (9.5)	5 (8.9)	2 (13.3)	8 (2.5)
	Liver Cancer	3 (1.5)	12 (2.3)	2 (3.6)	0 (0.0)	18 (5.5)
	Lung Cancer	16 (8.0)	61 (11.8)	13 (23.2)	0 (0.0)	27 (8.3)
	Melanoma	10 (5.0)	23 (4.5)	0 (0.0)	0 (0.0)	5 (1.5)
	Non-Hodgkin Lymphoma	27 (13.5)	47 (9.1)	3 (5.4)	0 (0.0)	8 (2.5)
	Pancreatic Cancer	12 (6.0)	24 (4.7)	1 (1.8)	0 (0.0)	13 (4.0)
	Prostate Cancer	15 (7.5)	58 (11.3)	13 (23.2)	2 (13.3)	29 (8.9)
	Thyroid	2 (1.0)	6 (1.2)	0 (0.0)	1 (6.7)	5 (1.5)
	Other Cancer	67 (33.5)	158 (30.7)	16 (28.6)	4 (26.7)	132 (40.5)
	Others	3 (1.5)	13 (2.5)	0 (0.0)	2 (13.3)	15 (4.6)

ICF: Informed Consent Form. The percentages for conditions do not sum to 100% because a single clinical trial may target multiple conditions. Similarly, the percentages for trial phases may exceed 100% as a single trial can span multiple phases.

Figure 2: **Dataset statistics.** Characteristics of clinical trials and the associated protocols and informed consent forms used in the experiments. Year: the start year of the clinical trial; Region: which region this trial was primarily initiated and conducted; Enrollment: the number of actual enrollment subjects in the trial; Protocol pages: the number of pages of the clinical trial protocol; ICF pages: the number of pages of the human written consent form document; Conditions: the primary targeted conditions of the trial.
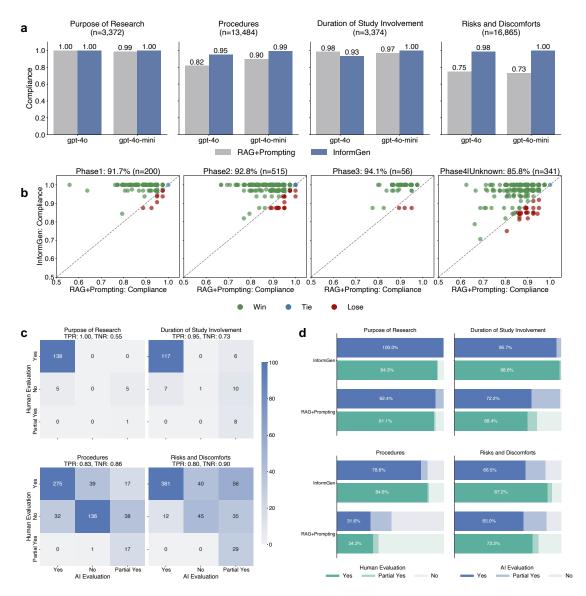
Figure 3: **Evaluating compliance of InformGen and baselines**. **a**, Automatic evaluation of the compliance rate of the generated content by InformGen and the baseline RAG+Prompting, using GPT-4o as a judge. Both methods are evaluated in two variants, utilizing either GPT-4o or GPT-4o-mini as the underlying LLM. **b**, In-depth trial-level comparison of InformGen and RAG+Prompting, analyzing compliance rates across all four clinical trial phases. **c**, Confusion matrix comparing compliance judgments made by GPT-4o and human annotators for the generated ICF content. TPR: True Positive Rate; TNR: True Negative Rate. **d**, Manual evaluation and GPT-4o-based assessment of compliance for the generated content across 100 trials.
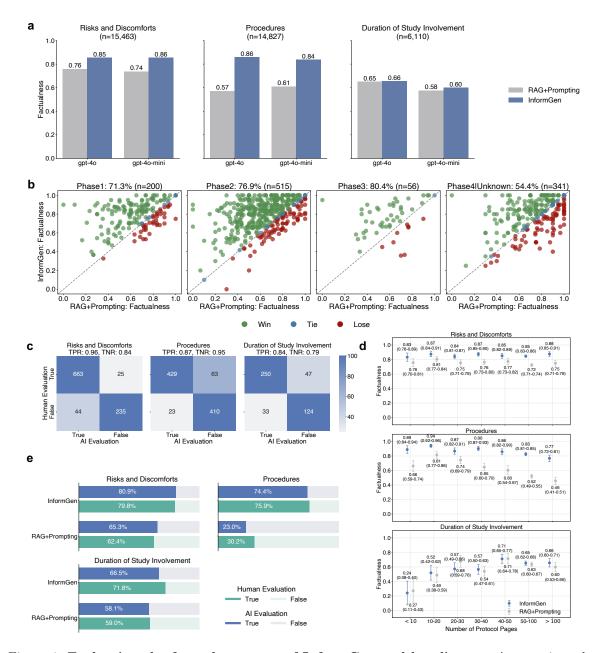
Figure 4: **Evaluating the factual accuracy of InformGen and baselines**. **a**, Automatic evaluation of the factual accuracy of content generated by InformGen and the baseline RAG+Prompting. Each ICF section is assessed against 3∼5 key facts, with n indicating the total number of evaluated facts. **b**, Trial-level breakdown of results by clinical trial phase. A side-by-side comparison of InformGen and RAG+Prompting for each trial shows that InformGen achieves a winning rate of 55%-80% in factual accuracy improvement. **c**, Confusion matrix comparing factual accuracy evaluations from GPT-4o (automatic evaluation) with human annotators. TPR: True Positive Rate; TNR: True Negative Rate. **d**, Analysis of factual accuracy across trials with varying protocol complexity. The x-axis represents the number of protocol pages, showing that InformGen exhibits greater robustness compared to RAG+Prompting as protocol complexity increases. **e**, Comparison of manual evaluation and GPT-4o-based assessment of factual accuracy across 100 trials.

15

Figure 5: **Development of the InformGen platform for consent document generation with human-in-the-loop**. **a**, Illustration of the demo platform for AI-assisted consent document drafting. The interface allows users to review and refine the processed procedure-risk table and schedule of assessment (SOA) table extracted from the input protocol. Users can also trace back information to the original protocol for verification. **b**, Visualization of the generated ICF content, displayed section by section. The sidebar lists the generated sections, while the main panel presents the content with inline citations, which are linked to the source documents for user review. **c**, Manual evaluation of factual accuracy comparing RAG+Prompting with InformGen incorporating human oversight (InformGen +Human). The results show a significant accuracy improvement of 20%-30%, with InformGen +Human achieving over 90% factual accuracy.

Section Title	Rule Name	Rule Description
Purpose of Research	1. Description of Clinical Investigation: 1a	A clear statement that the clinical investigation involves research is important to make prospective subjects aware that, although preliminary data (bench, animal, pilot studies, literature) may exist, the purpose of the subject's participation is primarily to contribute to research
Purpose of Research	Additional 6. Number of Subjects: 6a	When appropriate, the informed consent process must state the approximate number of subjects involved in the clinical investigation (21 CFR 50.25(b)(6)).
Duration of Study Involvement	Clinical Investigation Duration: 1h	1h. The informed consent process must clearly describe the expected duration of the subject's participation in the clinical investigation (see 21 CFR 50.25(a)(1)), which includes their active participation as well as long-term follow-up, if appropriate.
Duration of Study Involvement	Additional 4. Consequences of Subject's Decision to Withdraw: 4c	Prospective subjects should be made aware of the number of visits and approximate time required for participation in the study to help them determine if they are able to make the commitment to complete the study.
Procedures	1. Description of Clinical Investigation: 1c	The description of the clinical investigation should identify tests or procedures required by the protocol that would not be part of their care outside of the research; for example, drawing blood samples for a pharmacokinetic study. Note that all experimental procedures must be identified as such (21 CFR 50.25(a)(1)).
Procedures	1. Description of Clinical Investigation: 1d	1d. Procedures related solely to research must be explained (for example, protocol-driven versus individualized dosing, randomized assignment to treatment, blinding of subject and investigator, and receipt of placebo if the study is placebo-controlled) (21 CFR 50.25(a)(1)).
Procedures	1. Description of Clinical Investigation: 1e	The description should include relevant information on what is known about both the test article (e.g., the investigational product under study) and the control. For example, the description should indicate whether the test article is approved or cleared for marketing and describe the use(s) for which it has been approved or cleared. The description should also provide relevant information about any control used in the study: for example, whether the control is FDA-approved or cleared for marketing, considered a medically recognized standard of care, or is a placebo (including an explanation of what a placebo is).
Procedures	1. Description of Clinical Investigation: 1f	The consent process should outline what the subject's participation will involve in order to comply with the protocol, for example, the number of clinic visits, maintenance of diaries, and medical or dietary restrictions (including the need to avoid specific medications or activities, such as participation in other clinical investigations).
Procedures	1. Description of Clinical Investigation: 1g	1g. If there is follow-up period, prospective subjects must be informed of the procedures that will occur during such follow-up (21 CFR 50.25(a)(1)).
Voluntary Participation	8. Voluntary Participation: 8a	8a. Subjects must be informed that they may decline to take part in the clinical investigation or may stop participation at any time without penalty or loss of benefits to which subjects are entitled (21 CFR 50.25(a)(8)). Language that limits the subject's right to decline to participate or withdraw from the clinical investigation must not be used.
Voluntary Participation	8. Voluntary Participation: 8b	Written withdrawal from the study by the subject is not a requirement.
Participant Responsibilities	8. Voluntary Participation: 8a	Subjects must be informed that they may decline to take part in the clinical investigation or may stop participation at any time without penalty or loss of benefits to which subjects are entitled (21 CFR 50.25(a)(8)). Language that limits the subject's right to decline to participate or withdraw from the clinical investigation must not be used.
Withdrawal from Study	Additional 2. Involuntary Termination of Subject's Participation: 2a	When appropriate, the consent process must inform the subject of anticipated circumstances under which the investigator may end the subject's participation without the subject's consent (21 CFR 50.25(b)(2)).
Withdrawal from Study	Additional 4. Consequences of Subject's Decision to Withdraw: 4a	4a. When withdrawal from a clinical investigation may adversely affect the subject, the informed consent process must explain the withdrawal procedures that are recommended in order to ensure the subject's safety and should specifically state why the withdrawal procedures are important to the subject's welfare. The consent process must explicitly inform the subject of the potential adverse effects of premature termination of the investigational intervention.
Withdrawal from Study	Additional 4. Consequences of Subject's Decision to Withdraw: 4b	4b. If applicable, the consent process must explain whether a subject who withdraws early will receive future study payments. The subject should also be informed as to how the data that have already been collected will be handled.
Possible Risks, Discomforts, and Inconveniences	2. Risks and Discomforts: 2a	The informed consent process must describe the reasonably foreseeable risks or discomforts to the subject. This includes risks or discomforts of tests, interventions and procedures required by the protocol (including protocol-specified standard medical procedures, exams, and tests), with a particular focus on those that carry significant risk of morbidity or mortality.
Possible Risks, Discomforts, and Inconveniences	2. Risks and Discomforts: 2b	Possible risks or discomforts due to changes to a subject's medical care (e.g., by changing the subject's stable medication regimen or by stopping the subject's current treatment and randomizing them to either the investigational drug or placebo) should be addressed.
Possible Risks, Discomforts, and Inconveniences	2. Risks and Discomforts: 2c	2c. Where relevant, participants should also be made aware of the possibility of unintended disclosures of private information and be provided with an explanation of measures to protect a subject's privacy and data, and limitations to those measures.
Possible Risks, Discomforts, and Inconveniences	2. Risks and Discomforts: 2d	2d. Reasonably foreseeable discomforts to the subject must also be described (21 CFR 50.25(a)(2)). For example, the consent form should disclose that the subject may be uncomfortable having to stay in one position or experience claustrophobia-like symptoms during an MRI.
Possible Risks, Discomforts, and Inconveniences	2. Risks and Discomforts: 2e	It is not necessary to describe all possible risks, especially if doing so could make the form overwhelming for subjects to read. Information on risks that are more likely to occur and those that are serious should be described so that prospective subjects can understand the nature of the risk. The discussion may include information on whether a risk is reversible, and the probability of the risk based on existing data. Information on what may be done to mitigate serious risks, and risks and discomforts more likely to occur, should also be considered for inclusion.
Possible Risks, Discomforts, and Inconveniences	2. Risks and Discomforts: 2f	If applicable, the consent document should include a description of the reasonably foreseeable risks to the subject, but also the potential for risk to "others" (for example, radiation therapy where close proximity to subjects post-procedure may create some risk to others). In situations where there may be a risk to others, efforts to mitigate the potential risk (e.g., using separate bathrooms) may be included in the consent document.
Possible Risks, Discomforts, and Inconveniences	2. Risks and Discomforts: 2g	When appropriate, a statement must be included that a particular treatment or procedure may involve currently unforeseeable risks to the subject (or to the subject's embryo or fetus, if the subject is or may become pregnant) (21 CFR 50.25(b)(1).
Possible Risks, Discomforts, and Inconveniences	Additional 1. Unforeseeable Risks: 1a	If long-term safety studies (such as bench and animal testing) are not completed, the informed consent process should explain that researchers have not completed studies that may identify potential risks (e.g., carcinogenicity or teratogenicity studies), and that there may be potential risks to subjects (or to the embryo or fetus, if the subject is or may become pregnant) that are currently unforeseeable (21 CFR 50.25(b)(1)).
Possible Risks, Discomforts, and Inconveniences	Additional 1. Unforeseeable Risks: 1b	Sponsors may want to consider whether appropriate birth control measures and notifying the investigator of pregnancy should be included in the protocol and addressed in the informed consent document.
Alternatives	1. Description of Clinical Investigation: 1b	FDA recommends that when discussing the required elements of informed consent with prospective subjects, there should first be a discussion of the care a patient would likely receive if not part of the research, if relevant, and then the potential subject should be provided with information about the research.
Alternatives	4. Alternative Procedures or Treatments: 4a	Prospective subjects must be informed of the appropriate alternatives available to them, including a description of the care they would be likely to receive if they choose not to participate in the research. This includes alternatives such as approved therapies for the patient's disease or condition, other forms of therapy (e.g., surgical) or diagnosis, and when appropriate, supportive care with no disease-directed therapy. This disclosure must include a description of the current medically recognized standard of care, particularly in studies of medical products intended to treat or diagnose serious diseases or conditions.
Alternatives	4. Alternative Procedures or Treatments: 4b	4b. A full risk/benefit explanation of alternatives may not be appropriate to include in the written document.
Participant's Rights	8. Voluntary Participation: 8a	Subjects must be informed that they may decline to take part in the clinical investigation or may stop participation at any time without penalty or loss of benefits to which subjects are entitled (21 CFR 50.25(a)(8)). Language that limits the subject's right to decline to participate or withdraw from the clinical investigation must not be used.
Participant's Rights	8. Voluntary Participation: 8b	Written withdrawal from the study by the subject is not a requirement.
Participant's Rights	Additional 5. Providing Significant New Findings to Subjects: 5a	The consent process must, when appropriate, include a statement that significant new findings that may relate to the subject's willingness to continue participation, such as new risk information, will be provided to the subject (21 CFR 50.25(b)(5)).
Confidentiality	5. Confidentiality: 5a	The consent process must describe the extent to which confidentiality of records identifying subjects will be maintained (21 CFR 50.25(a)(5)) and should identify all entities, for example, the study sponsor, the research team, regulatory agencies, and/or ethics committee members, who may gain access to the records relating to the clinical investigation.
Confidentiality	5. Confidentiality: 5b	The consent process must also note the possibility that FDA may inspect records (21 CFR 50.25(a)(5)) and should not state or imply that FDA needs permission from the subject for access to the records. Under the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, FDA does not need permission to inspect records containing protected health information (45 CFR 164.512).

Supplementary Figure 1: **Rule set for evaluating compliance of generated content.** *"Section Title"*: Manually assigned section to which the rule applies and is used for assessment. *"Rule Name"*: The designated name of the rule. *"Rule Description"*: A detailed explanation of the rule, derived from the FDA guideline document.