

Article

Not peer-reviewed version

Retrieval-Augmented Generation Enhanced GPT-4.1 to Support Clinical Trial Informed Consent Review for Data Reuse

[Lameck Mbangula Amugongo](#)*, Lena Schaller, R. Maarten van Dijk, Helene Wendt, Enrica Zanuttigh, Claudia Neumann, Andreas Freisinger, Jaroslaw Deska

Posted Date: 4 March 2026

doi: 10.20944/preprints202603.0380.v1

Keywords: generative pre-trained transformer 4 (GPT-4); retrieval augmented generation (RAG); informed consent forms (ICFs)



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Retrieval-Augmented Generation Enhanced GPT-4.1 to Support Clinical Trial Informed Consent Review for Data Reuse

Lameck Mbangula Amugongo ^{1,*} , Lena Schaller ², R. Maarten van Dijk ², Helene Wendt ², Enrica Zanuttigh ², Claudia Neumann ¹, Andreas Freisinger ¹ and Jaroslaw Deska ¹

¹ Biostatistics and Data Sciences Department, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riß, Germany

² Staburo GmbH, 81549 Munich, Germany

* Correspondence: mbangula_lameck.amugongo@boehringer-ingelheim.com

Abstract

Background: Regulatory frameworks such as the Belmont Report, the Common Rule, and the Declaration of Helsinki require informed consent to ensure participants understand a study's purpose and can make voluntary decisions about their involvement. Regulations including the General Data Protection Regulation (Regulation (EU) 2016/679) further emphasise that consent must be freely given and revocable without disadvantage. Although informed consent forms (ICFs) are intended to be clear and accessible, they have become increasingly lengthy and complex. Large language models (LLMs) offer potential to navigate and interpret this complexity and have shown promise in biomedical information extraction tasks. However, their susceptibility to hallucinations limits reliability in high stakes settings. Retrieval augmented generation (RAG) can mitigate such errors. This study evaluates the integration of LLMs with RAG for reviewing data reuse language in ICFs and their ability to interpret complex textual structures. **Methods:** Firstly, we processed 438 ICFs from different trials, including multi-countries, languages and versions of ICFs. Using expertly curated prompts, we extracted information about data reuse using GPT-4.1. Comparing the LLM-generated data reuse outputs with human expert ground truth, we evaluated accuracy and the time required to extract information for each ICF. To further validate the workflow, we evaluated an independent set of 488 ICFs spanning additional trials, languages, and regions. For this cohort, we assessed the correctness of LLM outputs along with the quality of supporting evidence provided by the model. **Results:** Across 438 ICFs, the system achieved 81.6% accuracy, which increased to 90% in a subsequent evaluation of additional 488 ICFs after prompt optimisation. Using a RAG-based approach, the system accurately extracted data reuse information across multiple languages and identified nuanced international regulatory requirements. **Conclusion:** This approach has the potential to significantly alleviate administrative burdens by automating labour-intensive processes, while also generating insights that could inform the standardisation of ICF creation. Ultimately, these advancements may contribute to reduce the complexity of ICFs, thereby improving their readability and comprehensibility for participants.

Keywords: generative pre-trained transformer 4 (GPT-4); retrieval augmented generation (RAG); informed consent forms (ICFs)

1. Introduction

Informed consent forms (ICFs) are crucial for studies that involve human subjects, safeguarding the well-being, rights and safety of clinical study participants [1]. ICFs serve as essential documentation that clinical researchers often consult to verify what participants have agreed to regarding the use of their data collected during a study. This includes samples pertaining to biomarkers, imaging, deoxyribonucleic acid (DNA) and other collected data that can be used for translational science or

other scientific improvements. Reviewing information about data reuse in ICFs is a manual process, involving a dedicated team assessing ICFs one-by-one to verify data reuse permissions. This manual process is tedious and time-consuming, requiring reviewers to navigate complex language and nuanced variations across countries and languages to extract relevant information about data reuse, a challenge that becomes even more pronounced in multinational clinical trials where multiple country-specific versions of the same ICF further increase review complexity. Without relevant information from ICFs about data reuse, participants' data cannot be reused beyond the purpose of the study for which it was collected, hindering advancements in translational medicine and other scientific research. The recent development of large language models (LLMs) presents opportunities to navigate the complex language in informed consent using artificial intelligence. LLMs have shown significant potential in similar biomedical tasks such as information extraction, authoring clinical documents and patient trial matching [2]. However, LLMs are prone to hallucination, generating misleading, false or non-existent information [3]. To solve this problem, retrieval augmented generation (RAG) has been proposed [4]. A recent survey has shown that RAG has the potential to reduce hallucinations of LLMs, especially in critical sectors such as healthcare [5]. The ability of LLMs to navigate the intricate and specialised language within ICFs remains largely unexplored. Shi et al.[6] have shown that LLMs can generate ICF documents with improved readability, understandability, and actionability, without sacrificing accuracy or completeness. In another study, a chat-based application was created to enable users to interact with ICF documents, providing precise and prompt answers about data reusability [7]. The proposed chatbot is limited to only one ICF document at a time and requires users to type their questions. This can lead to inconsistencies in LLM responses, as each user may write their prompt differently. In this study, we investigated the integration of LLMs into the clinical workflow for ICF review, focusing on their capacity to comprehend and process the complex linguistic structures inherent to these documents. By leveraging RAG, GPT-4.1 and expert-defined prompts, our findings demonstrate that LLMs are capable of extracting relevant information related to data reuse. Our RAG-based application can process multiple ICF documents at a time, improving the efficiency of ICF document analysis.

2. State of the art

A recent systematic review highlights an increasing number of studies exploring the use of RAG-based LLMs across diverse aspects of clinical workflows [5]. However, the use of RAG-based LLMs in the informed consent process is relatively new. To date, only a few studies have explored the potential for LLMs to process ICF documents. Most of these studies focus on LLM-generated ICF content, analysing readability, accuracy, and ethical implications. For example, early research by Decker et al.[8], demonstrated that LLM-based chatbots can outperform traditional surgeon-authored consent materials in terms of clarity and completeness, particularly when describing surgical benefits and alternatives. These findings suggest that LLMs, when appropriately integrated into the consent workflow, may enhance participant understanding and reduce the risk of misinformation. In a recent study, Vaira et al.[9] evaluated the quality and readability of ICFs generated by various chat models such as Generative Pre-Trained Transformer-4 (GPT-4)[10] and Bard Gemini [11]. Their results show that chatGPT4 generated the highest-rated consent documents, demonstrating AI's promise in improving patient communication. In another study, Shi et al.[6] reported that LLM-generated ICFs are comparable to human-authored documents in terms of accuracy and completeness, with significantly improved readability and actionability.

Raimann et al.[12] explored the performance of three chat models, namely GPT-3.5, GPT-4 and Gemini, in creating information sheets for six common anesthesiologic procedures, guided by patient questions. Multiple drafts of anesthesia forms were created and refined, with results evaluated against established standard checklists. Illustrating the potential for the application of AI in informed consent sheets but underscoring a critical limitation, none of the tested models were able to produce legally compliant IC sheets without human oversight. This highlights the current boundary of LLM

capabilities and the necessity of hybrid approaches that combine machine efficiency with human judgment.

Beyond technical performance, emerging literature has begun to address the relational and ethical dimensions of LLM use in informed consent. Rudra et al.[13] advocate for a supportive role of LLMs, emphasising their potential to detect patient distress and alert healthcare providers, thereby facilitating empathetic communication. These perspectives reinforce the notion that LLMs should augment and not replace human interactions, preserving the trust and emotional nuance essential to ethical consent processes. Finally, Allen, Levy, and Wilkinsonc [14] introduce a conceptual framework for informed autonomy, arguing that, while LLMs can enhance information processing and value clarification, they also risk undermining patient agency if not carefully deployed. Their proposed guidelines for implementation reflect a growing consensus that ethical integration of LLMs in healthcare requires ongoing evaluation, transparency, and patient-centred design [5].

3. Methods

The advent of LLMs has transformed natural language processing (NLP), shifting NLP from task-specific representation to task-agnostic pre-training [15]. LLMs have demonstrated performance comparable to humans in many different tasks [16,17]. However, LLMs are static snapshots with knowledge cut-off dates, e.g., Google Gemini 2.5 Pro incorporates data up to January 2025 and the cut-off for GPT-4.1 is June 2024 [18]. Moreover, LLMs lack access to proprietary information, which includes ICF documents. To address these limitations, we proposed a RAG workflow to augment the LLM with ICF documents containing information about data reuse. The information from the ICF documents minimises hallucinations, enabling the LLM to generate reliable responses supported by extracted text, which act as verifiable references of what the LLM used to generate the outcome.

3.1. RAG Workflow

Figure 1 shows the RAG workflow. The process begins with the user uploading an ICF document (or multiple documents). The uploaded ICF is then ingested in an in-memory vector store. Afterwards, we perform a context-based hybrid (keyword and semantic) search to retrieve relevant information from the ICF document. The relevant content is then passed to the LLM with expert-defined prompts. Finally, the LLM-generated outcome for data reuse is presented as a response to the user to aid the decision-making.

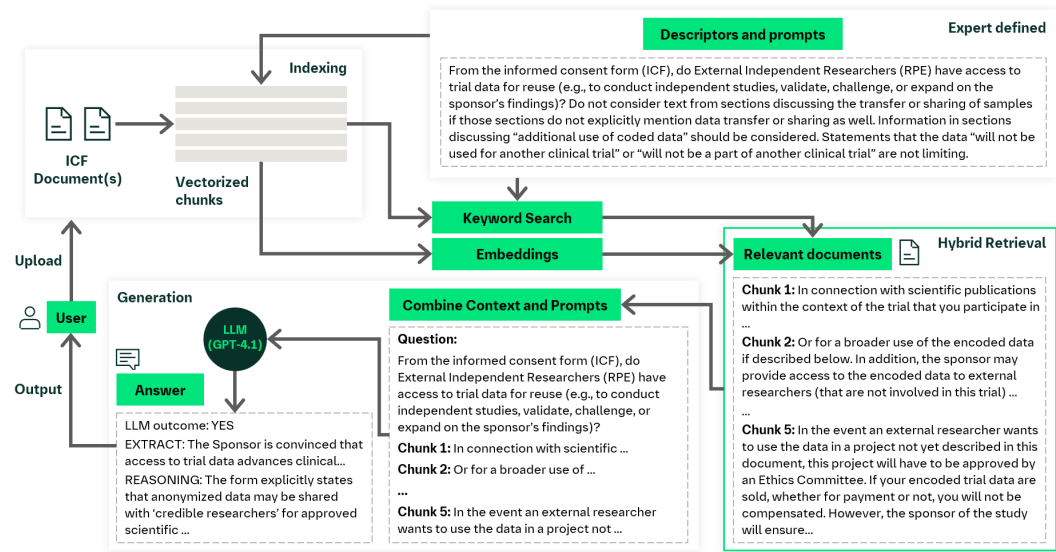


Figure 1. Workflow illustrating the schematic RAG pipeline.

3.2. Vector Database

To securely and efficiently process uploaded ICF documents, we implemented an in-memory vector store. First, text was extracted from the ICF document and segmented through a process called “chunking”. Chunking is necessary because a single ICF document typically contains at least 15 pages, and embedding models are limited by the number of tokens they can process within a single sequence. To determine the optimal chunking strategy, we tested different chunk sizes of 500, 1000 and 2000 tokens. After chunking, the text segments were transformed into high-dimensional embeddings using an OpenAI text-embedding-3-small model[19] to capture semantic meaning in the ICF document. The text-embedding-3-small model was chosen because of its high efficiency and better performance over predecessor models, such as text-embedding-ada-002 [19]. The text embeddings, together with associated metadata, were then instantiated within the LangChain MemoryVectorStore, allowing for rapid similarity search and contextual retrieval of relevant consent clauses. The in-memory architecture was deliberately chosen to ensure compliance with data minimisation principles, averting persistent storage of ICF information, thereby aligning with ethical standards in research data governance and regulations [20] while maintaining technical efficiency in semantic retrieval. Figure 2 shows an example representing three chunk embeddings.

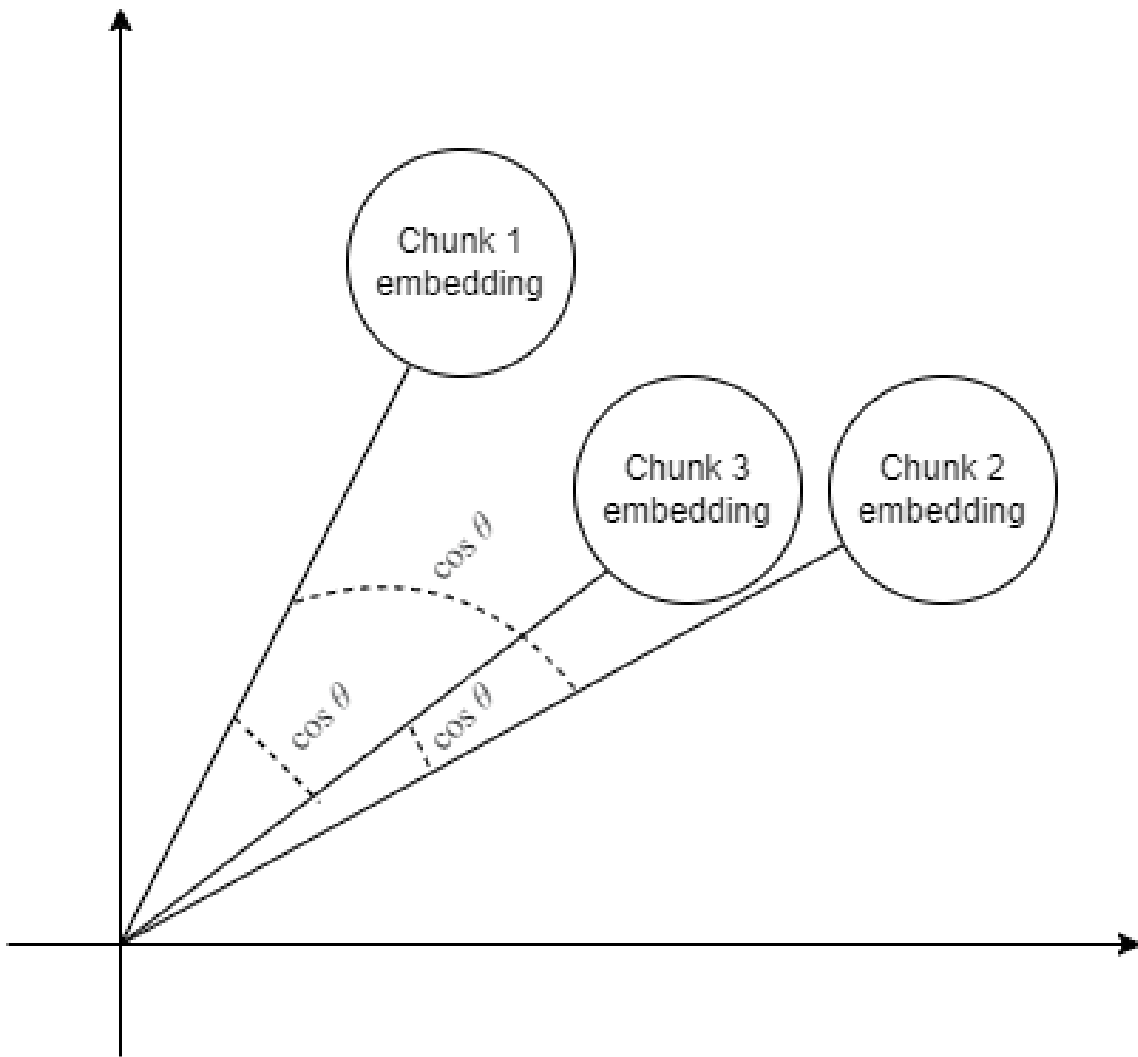


Figure 2. An example of embedding representing three chunks with their respective cosine similarity.

3.3. Prompting Strategy

Prompting is fundamental to guide LLMs to generate outputs. In a nutshell, a prompt is a set of instructions that help the model comprehend the type of response expected by the user. There are different prompting strategies aiming to improve prompt clarity, contextual alignment, and structural flow to provide the most accurate response to the user query [21]. In this study, we explored different prompting strategies, i.e., zero-shot and few-shot prompting, with few-shot prompting chosen as the final strategy due to its ability to ground LLM responses by providing one or more examples. This helps the model comprehend what text to look for in the ICF document and the format in which the model should provide the responses. An example of few-shot prompting is provided in table 1. Table 1. An example of a prompt to extract relevant information if the patient has consented to allow their data to be used to improve the trial medication and related medications.

Table 1. An example of a prompt to extract relevant information if the patient has consented to allow their data to be used to improve the trial medication and related medications.

Example
From the information provided in the informed consent form (ICF). Can trial data be used to improve medication quality and related substances? This info is in the confidentiality and data privacy section. Look for phrases such as: "understand how the trial drug and similar drugs work in the body and the study drug mode of action" If it is mentioned that data can be reused to understand how similar trial drug works or better understand related diseases, then reply YES. If it is explicitly mentioned that data "cannot" be used to improve related medication or similar trial drugs, reply NO. If it is not explicitly mentioned that data can be used to understand similar trial drugs or related medication or no information is available, reply NA. Extract information from document to support the answer. Descriptor: MED_TR_REL

The full list of prompts are provided in Table A1 of Appendix. The descriptors COUNTRY, TRIAL_NO, VERSION and VERSION_DATE were not assessed as they do not affect data sharing.

3.4. Retrieval

To ensure accurate and contextually relevant extraction from ICFs, we employed both pre-retrieval and post-retrieval techniques. Pre-retrieval optimisation focused on indexing strategies, including segmentation of the uploaded ICF into granular text chunks, alignment of text with associated metadata, and embedding of prompts into vector space representations. For retrieval, we employed a hybrid approach, combining similarity search and keyword search. First, we performed a similarity search on the vector store, returning the top-k = 5 most relevant segments. Next, a BM25 (best matching 25) retriever was applied to the same indexed text and metadata to perform a keyword-based search, likewise retrieving the top five results. Post-retrieval, the outputs of both methods were integrated into an ensemble retriever, which combined the BM25 keyword search and vector similarity search with weighted contributions (0.3 and 0.7, respectively). This hybrid retrieval strategy allowed us to balance lexical precision with semantic nuance in the ICF language, thereby improving the quality of the retrieved information and ensuring that both explicit terminology and contextual meaning within the ICF were captured.

3.5. LLM Model

Initially, we explored several models, such as GPT-3.5 and GPT-4o. However, after the release of GPT-4.1, which introduced major improvements such as long context and instruction following, we opted to use GPT-4.1 in our RAG pipeline. This decision was driven by its superior use of context and enhanced long-context [22]. To improve LLM generation, we developed a structured workflow using LangChain to create a RAG pipeline that integrates information retrieval with LLM outcome generation for data reuse. For each prompt, the system retrieved the top-5 most relevant sections from the ICF, which served as the external knowledge from which the LLM derived its data reuse assessment.

For structured and reliable outputs, we defined a structured data model using Pydantic's Base Model class, which provides automatic validation and serialisation of fields through type annotations. The extracted data schema specifies four components: descriptor, outcome, supporting extract, and LLM reasoning, thereby enforcing consistency in outputs and reducing ambiguity in downstream analysis. Such schema-driven approaches are critical in biomedical research, where methodological rigor and reproducibility are important.

3.6. Application

To enable users to access and apply the workflow without the need to write or execute code, we developed a web-based application using Streamlit [23]. Streamlit was selected due to its simplicity and rapid development capabilities, which allow for efficient creation and deployment of interactive data applications. The final application integrates the RAG-based workflow for analysing ICFs directly within the Streamlit interface, providing an intuitive and user-friendly environment for performing all analyses.

3.7. Deployment

Due to the confidential nature of ICFs, data security is essential. In our study, all LLM calls were performed through application programming interface calls in Boehringer Ingelheim secure environment deployed on Amazon web services (AWS) private cloud, upholding data integrity and privacy. The application was deployed on Posit Connect, a platform that enables users to securely share insights, automate tasks and deploy various data science applications (posit). Within Boehringer Ingelheim, Posit Connect is deployed on a private AWS cloud, allowing users to securely access the system and analyse ICFs.

3.8. Evaluation

To evaluate the performance of our RAG pipeline, we conducted an experiment using 438 ICF documents and tested different chunk sizes of 500, 1000 and 2000. We applied various metrics, such as accuracy (correctness of LLM outcome), completeness (correctness of LLM outcome and correct supporting extraction) and the evaluation of LLM reasoning. To compute accuracy, we compared the LLM outcome with the ground truth provided by subject matter experts. Because we did not have ground truth for supporting text from the ICF document, human experts manually evaluated the LLM outcome and reasoning. Evaluation was only performed on data sharing descriptors (see Table 2). In total, 21 data sharing descriptors were analysed, identifying mentions of intended recipient (i.e. LIC, RPC, RPE, SPAFI), reuse purpose (MED_TR, MED_TR_REL, MED_ALL, DIS_ALL, DIS_TR, DIS_TR_REL, DIS_TR_TA, DPROD_ALL, DPROD_TR_REL, TRPROD_ALL, TRPROD_TR_REL, QUAL_ALL), data retention time (DLINK30, RR50, RR80) and content (RESTRICTIVE_WORDING, GENOMIC_INFO).

Table 2. Data sharing descriptors investigated and their definitions.

Descriptor	Definition	Outcome
Recipient		
<i>Data can be shared to:</i>		
LIC	Owners of a licensed product.	YES NO NA
RPC	Research partners and collaborators of the Sponsor.	YES NO NA
RPE	External independent researchers.	YES NO NA
SPAFI	Sponsors, affiliates, or third-party processors.	YES NO NA
Purpose		
<i>Data can be shared to:</i>		
MED_TR	Learn about the trial medication.	YES NO NA
MED_TR_REL	Learn about medications related to the trial medication.	YES NO NA
MED_ALL	Learn about any medications.	YES NO NA
DIS_ALL	Learn about any disease.	YES NO NA
DIS_TR	Learn about the disease studied in the trial.	YES NO NA
DIS_TR_REL	Learn about diseases related to the trial disease.	YES NO NA
DIS_TR_TA	Learn about diseases in a specific therapeutic area.	YES NO NA
DPROD_ALL	Develop any diagnostic tools.	YES NO NA
DPROD_TR_REL	Develop diagnostic tools related to the trial disease.	YES NO NA
TRPROD_ALL	Develop any therapeutics.	YES NO NA
TRPROD_TR_REL	Develop therapeutics related to the trial disease.	YES NO NA
QUAL_ALL	Improve the quality of future clinical trials.	YES NO NA
Data Retention		
DLINK30	The subject re-identification link will be deleted after 30 years.	YES NO
RR50	Data will be retained for up to 50 years.	YES NO
RR80	Data will be retained for up to 80 years.	YES NO
Specific Text		
RESTRICTIVE_WORDING	Any text with other restrictions on data reuse.	YES NO
GENOMIC_INFO	Any text mentioning genomic data.	YES NO

3.9. Validation

To validate the performance of our RAG pipeline, experts selected 488 ICF documents from different trials, languages and regions. The LLM’s responses were assessed for accuracy, completeness, and relevance. We computed the number of ICFs which were correctly analysed for all data sharing descriptors. Finally, we computed the number of ICFs correctly analysed for data sharing, categorised by country of origin.

3.10. Statistical Analysis

For data sharing information, we were not only interested in the accuracy, completeness, and usefulness of the answers provided by the RAG based pipeline, but also in how the output impacts decision-making regarding data reuse. As such, we computed class specific and overall accuracies for each data sharing descriptor extracted from 488 ICFs. Each descriptor was evaluated across possible output categories as seen in Table 2. An output of “YES” indicates that the ICF contains text supporting data reuse for the specified descriptor, while an output of “NO” reflects that the ICF explicitly prohibits data reuse for that purpose. If no context is provided in the ICF to support or prohibit data reuse, then the output for the descriptor would be “NA”. However, for descriptors identifying clauses for data retention practices (DLINK30, RR50 and RR80), as well as the descriptors pinpointing specific language (GENOMIC_INFO and RESTRICTIVE_WORDING) were evaluated on a binary scale (YES/NO), where an output of “NO” applies both in the context of an explicit prohibition, and in the absence of any mention related to the descriptor. Reporting class-level assessment allowed us to thoroughly examine how well our workflow distinguished affirmative, ambiguous, or negative data sharing statements, while overall accuracy provided a high-level summary of system performance.

4. Results

The performance between GPT-4o and GPT-4.1 was comparable, see Figure A1 of the Appendix. GPT-4.1 outperformed GPT-4o on several descriptors, especially descriptors requiring reasoning and instruction following, such as those identifying further restrictive wording and genomic information.

From here onwards, all results were obtained using GPT-4.1 as the LLM. When testing the effect of different chunking sizes (see Table A2 of the Appendix), average accuracies of 79.5%, 80.5% and 81.6% were achieved across all data sharing descriptors for chunk sizes of 500, 1000 and 2000 tokens, respectively. To maximize accuracy, the chunk size of 2000 tokens was chosen for subsequent analyses. As shown in the screenshot in Figure 3, the application presents all relevant information to help the user efficiently make a decision about data-sharing. Emphasising user autonomy, the application also allows for user corrections where the LLM output is wrong. In terms of speed, our application processed ICF documents with an average time of 1.95 ± 0.35 minutes. The full LLM outcome results for an example ICF are provided in Figure A2 of the Appendix.

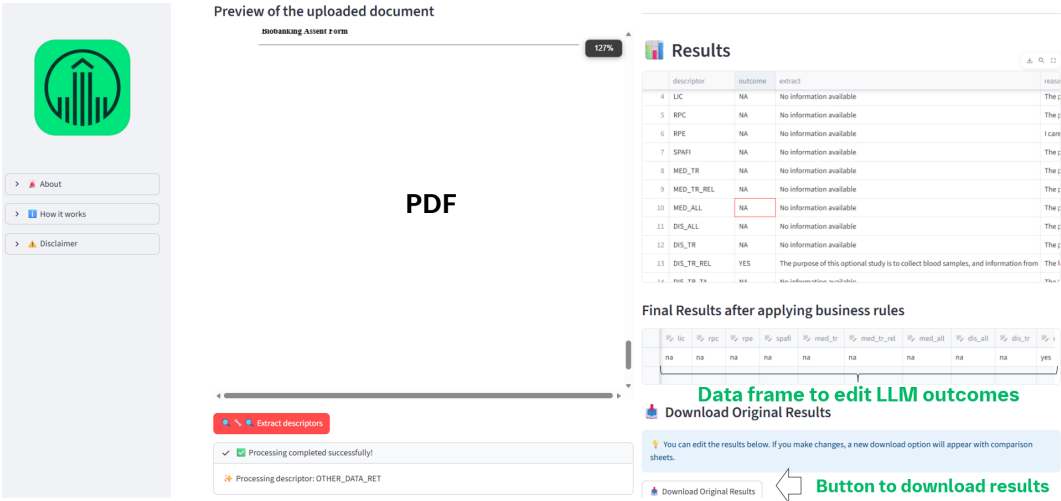


Figure 3. Streamlit application that enables users to access our automatic workflow to analyse ICFs. The application is intuitive, enabling the user to view the results and uploaded document side-by-side. The results can be edited to enable corrections. Once a user is satisfied with the results, the results can be downloaded.

4.1. Evaluation

Using chunksize 2000 tokens, the model achieved an overall accuracy of 81.6% across 438 ICF documents, indicating that 82% of all data sharing descriptor predictions matched the corresponding expert evaluations. However, some shortcomings were identified, such as an accuracy of less than 70% in several descriptors: MED_ALL, DIS_TR_TA, RESTRICTIVE_WORDING and GENOMIC_INFO. To address the shortcomings, we performed prompt optimisation and then validated the workflow.

4.2. Validation

For the validation of our workflow, we conducted further analysis of 488 ICF documents, achieving an average performance of 89.5% in terms of accuracy across all data-sharing descriptors, see Table 3. High accuracy (>80%) was observed across most descriptors, with only two descriptors, MED_ALL and RESTRICTIVE_WORDING, attaining lower accuracies of 74.6% and 71.1%, respectively.

In total, 186 ICFs analysed had completely correct outcomes for all data sharing descriptors. As shown in Table A3 of the Appendix, 441 ICFs from 26 different countries and 47 general template (no country affiliation) ICFs were analysed. ICFs from Turkey (2/2) and Norway (10/10) were analysed with 100% accuracy for data sharing. Similarly, 35/39 Polish ICFs were completely correctly analysed by our workflow.

4.3. Statistical analysis

Across 488 informed consent documents, our workflow demonstrated strong performance for most data sharing descriptors. The highest accuracy was achieved for GENOMIC_INFO, which reached 98.2% overall and showed near perfect detection of both affirmative (98%) and negative cases (98.9%). QUAL_ALL also performed robustly, with high and well-balanced accuracies across all classes and an overall accuracy of 95.1 percent. Several descriptors related to medicinal and disease specific

data reuse, including MED_TR (93.9%) and DIS_TR (92.4%), exhibited similarly strong performance, supported by perfect accuracy in the identification of NO cases.

Table 3. The performance of the system for all data-sharing descriptors on 488 ICF documents. The accuracy (ACC, as both percentage and fraction) for each output (YES, NA, NO), as well as accuracy per-descriptor are presented.

Descriptor	ACC - YES	ACC - NA	ACC - NO	Overall ACC
LIC	94.5 (260/275)	90.4 (189/209)	100 (4/4)	92.8 (453/488)
RPC	91.9 (351/382)	89.4 (93/104)	100 (2/2)	91.4 (446/488)
RPE	86.0 (270/314)	98.7 (147/149)	80.0 (20/25)	89.5 (437/488)
SPAFI	97.8 (409/418)	37.3 (25/67)	100 (3/3)	89.5 (437/488)
MED_TR	96.5 (384/398)	81.4 (70/86)	100 (4/4)	93.9 (458/488)
MED_TR_REL	96.1 (295/307)	58.2 (103/177)	100 (4/4)	82.4 (402/488)
MED_ALL	96.5 (272/282)	45.8 (88/192)	28.6 (4/14)	74.6 (364/488)
DIS_ALL	89.1 (197/221)	90.9 (210/231)	75.0 (27/36)	88.9 (434/488)
DIS_TR	96.3 (365/379)	77.5 (79/102)	100 (7/7)	92.4 (451/488)
DIS_TR_REL	98.2 (335/341)	66.2 (92/139)	87.5 (7/8)	88.9 (434/488)
DIS_TR_TA	96.4 (321/333)	71.5 (108/151)	100 (4/4)	88.7 (433/488)
DPROD_ALL	88.6 (171/193)	94.3 (265/281)	92.9 (13/14)	92.0 (449/448)
DPROD_TR_REL	92.7 (202/218)	95.5 (253/265)	80.0 (4/5)	94.1 (459/488)
TRPROD_ALL	90.1 (218/242)	79.1 (189/239)	85.7 (6/7)	84.6 (413/488)
TRPROD_TR_REL	92.6 (249/269)	81.8 (175/214)	80.0 (4/5)	87.7 (428/488)
QUAL_ALL	95.7 (265/277)	97.3 (144/148)	87.3 (55/63)	95.1 (464/488)
DLINK30	88.4 (114/129)	NA	96.1 (345/359)	94.1 (459/488)
RR50	100 (7/7)	NA	94.4 (454/481)	94.5 (461/488)
RR80	92.8 (77/83)	NA	96.8 (392/405)	96.1 (469/488)
RESTRICTIVE_WORDING	9.3 (10/108)	NA	88.7 (337/380)	71.1 (347/488)
GENOMIC_INFO	98.0 (387/395)	NA	98.9 (92/93)	98.2 (479/488)

Descriptors involving broader or more conditional statements showed greater variability. MED_TR_REL (82.4%) and DIS_TR_TA (88.7%) maintained high accuracy for YES classifications but demonstrated reduced performance for NA or NO outcomes, suggesting difficulty in interpreting nuanced or context dependent phrasing. MED_ALL showed the greatest imbalance, with high accuracy for YES cases (96.5%) but very low accuracy for NO cases (28.6%), resulting in a lower overall accuracy of 74.6%.

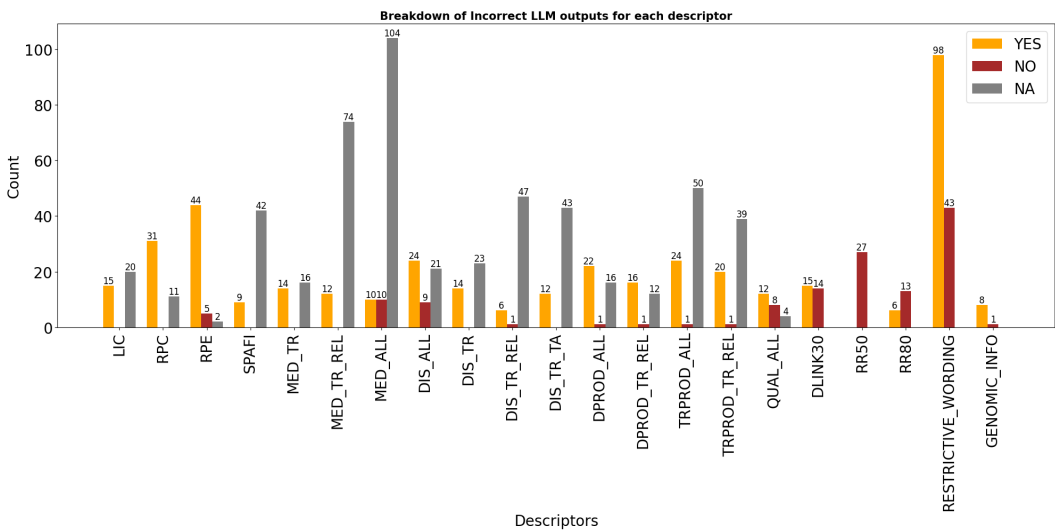


Figure 4. Distribution of LLM incorrect outputs. For each data sharing descriptor, the total number of cases where ground truths of “YES”, “NO”, and “NA” were misclassified is presented.

The lowest performance was observed for RESTRICTIVE_WORDING, which achieved an overall accuracy of 71.1%. This descriptor had particularly low accuracy in identifying YES cases (9.3%), indicating that restrictive or implicitly limiting statements remain challenging for automated extraction methods. As seen in Figure 4, the RESTRICTIVE_WORDING descriptor had the largest number of incorrect outcomes, followed by MED_ALL and MED_TR_REL. This suggests that, for these descriptors, subject matter experts should take additional care when reviewing LLM outcomes. The majority of incorrect outcomes (48.9%) were determined by subject matter experts to have a ground truth value of “NA”, which applies in cases where the document does not explicitly mention the conditions for data reuse for the descriptor in question.

Overall, the results underscore that LLMs have substantial potential to assist in navigating the complex linguistic structures of ICFs, with several descriptors demonstrating high accuracy and balanced precision–recall performance. These strong outcomes indicate that LLMs can reliably extract and interpret structured regulatory or biomedical information, supporting their integration into clinical workflows for ICF review.

5. Discussion

In this study we have developed a workflow integrating LLMs augmented with relevant context from ICF documents to automatically extract and analyse ICF documents to support the data reuse process. Our results show that, with expert-defined prompts, our RAG pipeline was able to extract information about data reuse with high accuracy.

To the best of our knowledge, we only found one study that explored the role of LLMs to analyse ICF documents for data reuse. This study proposed an adaptable framework using LLMs to aid the analysis of ICFs for data use agreements. They achieved model accuracy of 96.6% with their ICF Q&A chatbot [7]. However, their approach has several limitations: it is designed to process only one ICF at a time and its evaluation was performed using just two ICF documents. As a result, the generalisability and robustness of the chatbot across diverse ICFs, particularly across different languages and countries, remain unclear.

In contrast, we present a comprehensive analysis of more than 900 ICFs to evaluate and validate a novel integrated workflow for ICF review for data reuse. We responsibly leverage a GPT-4.1 LLM augmented with a RAG to automatically extract information related to data reuse. This workflow provides an intelligent, intuitive and efficient method for largescale ICF review. Importantly, our approach supports the simultaneous processing of multiple ICF documents and incorporates a human-in-the-loop design, enabling users to easily review and correct LLM outputs. Our workflow is intentionally human-centred; rather than replacing human reviewers, it aims to enhance the ICF review process by improving scalability and consistency while maintaining expert oversight. We observed a high accuracy rate for most of the data reuse descriptors targeted by our approach. However, poor performance was observed in descriptors such as MED_ALL and RESTRICTIVE_WORDING, highlighting persistent limitations related to ambiguity, edge case reasoning, and underrepresented concepts. The RESTRICTIVE_WORDING descriptor is particularly challenging to identify, even for human reviewers. In contrast to other descriptors, which are clearly defined by words such as medicine and disease, restrictions on data reuse do not rely on standard phrasings and are therefore intrinsically more difficult to identify. These weaknesses mirror broader concerns regarding LLM hallucinations and misinterpretations in high stakes settings, reinforcing the need for RAG and other grounding strategies to ensure reliability and safety. Our results also demonstrate that iterative prompt refinement can meaningfully improve performance. Following targeted optimisation, average accuracy across data sharing descriptors increased from 81.6% in the initial evaluation to 89.5% in the validation phase, indicating that contextual adjustments to prompt structure can yield substantial gains.

In addition, we found that the majority of misclassifications occurred when the expected output value is “NA”, see Figure 4. This pattern suggests that the LLM tends to infer either a confirmation (YES output) or denial (NO output) of data reuse in these cases, even if the document does not

contain context to support either decision. Importantly, this behaviour did not stem from unsupported hallucinations. For every decision, the model cited specific text from the ICF that influenced its output. After review by subject matter experts, the referenced passages and identified recurring phrases in the prompt that inadvertently triggered the model's reasoning. These observations suggest that a small number of linguistic patterns can bypass the intended prompt logic. As such continuous review of prompt logic versus model output can improve prompt optimisation and thus, reduce misclassifications.

Although several individual data sharing descriptors demonstrated high accuracy, this does not guarantee flawless performance at the document level. In our validation, only 186 of 488 ICFs achieved completely correct outcomes across all descriptors (38.1%), highlighting that even a single incorrect descriptor can lower the overall score. Given that, for each ICF, 21 descriptors were assessed, and the cumulative impact of one or two errors can substantially reduce the rate of fully correct documents despite strong per-descriptor performance. These weaknesses mirror broader concerns regarding LLM hallucinations and misinterpretations in high-stakes settings, reinforcing the need for RAG and other grounding strategies to ensure reliability and safety. Overall, our findings suggest that LLMs are well positioned to enhance clinical document understanding when appropriately constrained and supported by robust retrieval mechanisms, case reasoning, and underrepresented concepts.

Our study is not without limitations. First, after exploring the GPT-4o model, we opted to use GPT-4.1 without exploring other models. We chose GPT-4.1 because it has outperformed other models on several automation experimental tasks [24]. Additionally, GPT-4.1 offers exceptional performance including instruction following and reasoning at lower cost [22]. Second, in our study we used expert defined prompts to extract relevant information about data reuse. As such, we do not offer chatbot capabilities to allow users to further engage with the document. We chose this approach because we wanted reproducibility in terms of LLM outcomes. Additionally, allowing users to correct the LLM outputs is a more realistic application in clinical research where autonomy remains with the expert user. Third, even with RAG, misinterpretation may still occur, and we found that the LLM still gave wrong information in certain cases because of the inherent nature in how these systems work[25]. As such, we recommend the human-in-the-loop approach to verify LLM outputs regarding data sharing.

While our study focuses on analysing information from ICF documents, in the future we will leverage these insights to standardise ICF creation, accounting for different institutional review boards and country-specific ethical standards to optimise the process end-to-end. Subsequently, making ICF analysis more accurate. Future work will focus on further improvements of our workflow with minor prompt optimisation to capture nuances of language represented in different ICFs to avoid misclassifications. Additionally, we will quantify the time and cost saving impact of automating ICF review process.

Finally, in this study we have shown that LLMs can be successfully integrated into the clinical research for data reuse assessment. Using a few-shot prompting approach, our results show that it is possible to automatically extract relevant information about data reuse, with whom the data can be shared, and for which purpose the data can be reused. This will lead to increased efficiency, improved accuracy in the ICF review process and reduce manual efforts. Thus, further support translational medicine and expedite the discovery of therapeutic insights for the benefit of patients.

6. Conclusions

In this study we illustrate with few-shot prompting how an LLM incorporating RAG can be applied to solve downstream clinical research tasks, such as the review of ICFs for data reuse. During validation, our novel workflow incorporating GPT-4.1, expert defined prompts, and hybrid search achieved an accuracy of more than 85% in 17 out of 21 descriptors when compared to human ground truth. Furthermore, our automatic tool helps users understand the nuances of informed consent language by processing ICFs in various languages, including English, German, Dutch, French, Spanish, Korean, and Japanese. This illustrates the scalability of the solution and adaptability of LLMs to most

common languages. However, challenges such as hallucination or the LLM using the wrong reasoning may still occur. We have shown that this limitation can be mitigated by incorporating clear examples into prompts. Additionally, employing human-in-the-loop oversight allows subject matter experts to correct errors in the few instances where LLM outputs are inaccurate, thereby ensuring adherence to the responsible development and use of AI. Finally, our novel workflow illustrates how LLMs with RAG and few-shot prompting can be adapted for downstream tasks in clinical research, aiding experts to efficiently review, discern and determine appropriate use cases for data reuse.

Author Contributions: Conceptualization, LMA; methodology, LMA, LS, HW, RMvD; software, LMA; evaluation, HW; validation, LS; formal analysis, LS, HW; investigation, LMA, LS, HW, RMvD, EZ; resources, CN, AF and JD; data curation, LMA, LS; writing—original draft preparation, LMA; writing—review and editing, LS, HW, RMvD, EZ, CN, AF and JD; visualization, LMA.; supervision, LMA, RMvD, EZ, CN, AF, and JD.; project administration, CN, AF, and JD. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author(s).

Conflicts of Interest: LMA, CN, AF and JD are employed by the by Boehringer Ingelheim Pharma GmbH & Co. KG, and LS, HW, RMvD and EZ are employed by Staburo GmbH.

Abbreviations

The following abbreviations are used in this manuscript:

- GPT Generative pre-trained transformer
- ICFs Informed consent forms
- LLM Large language models
- RAG Retrieval augmented generation

Appendix A

Appendix A.1 Prompts

Table A1. The complete list of prompts used in our workflow.

Prompts
From the content of the informed consent form (ICF), identify the country associated with the document. The country may be explicitly named or implied through country codes or contextual clues (e.g., regulatory references, language, or location-specific contact details). Spanish language does not necessarily mean Spain is the country associated with the document and requires more thorough assessment. Use the following list of country names and codes as reference: For example: Algeria, AR/ARG (Argentina), AU/AUS (Australia), AT/AUT (Austria), BY (Belarus), BE/BEL (Belgium), Bosnia and Hercegovina, Botswana, BR (Brazil), BG (Bulgaria), CAN (Canada), CL (Chile), CN/CHN (China), CO (Colombia), HR (Croatia), CZ/CZE (Czech Republic), Ecuador, Egyptian (Egypt), EE (Estonia), FI (Finland), FR/FRA (France), GE (Georgia), DE (Germany), GRC (Greece), HU/HUN (Hungary), HKG (Hong Kong), IN/IND (India), ID (Indonesia), IE/IRL/UKIE (Ireland), IL/ISR (Israel), IT/ITA (Italy), JP/JPN (Japan), Jordan, KR/KOR (Korea), LV (Latvia),Lebanon, LT (Lithuania), Macedonia, MY/MYS (Malaysia), MX (Mexico), MD (Moldova), Monaco, Morocco, NLD (Netherlands), NZ/NZL (New Zealand), Norway, Panama, PE (Peru), Philippines,PL/POL (Poland), PT/PRT (Portugal), RO (Romania), RU (Russia), RS (Serbia), SG/SGP (Singapore), SK (Slovakia), SI (Slovenia), South Africa, ES/ESP (Spain), Sri Lanka, SE (Sweden), CH (Switzerland), TW/TWN (Taiwan), Thailand, Tunisia, Turkey, UAE, UA (Ukraine), UK/GBR (United Kingdom), US/USA (United States), Venezuela, Vietnam. Descriptor: COUNTRY
From the information provided in the informed consent form (ICF). What is the Trial No (Protocol No) mentioned in the ICF? The Trial No (Protocol No) is in this format: XXXX-XXXX, for example: Trial No: 1443-0004, 1368-0027. Extract where Trial number (Protocol number) is mentioned. Descriptor: TRIAL_NO
From the informed consent form (ICF), extract the version number of the document. Look for phrases like: "version-01", "version-b-01", "Main Consent Version # 2.0", etc. Do not confuse the version number with the version date. Extract the version as a digit only (e.g., 1, 2, 3). Do not include any other characters. Extract information from document to support the answer. Descriptor: VERSION
From the information provided in the informed consent form (ICF). What is the ICF version date? Extract where the date is mentioned. Extract the version date and return in the format DD-MMM-YYYY. For example: 05-Aug-2025. Do not include any other characters. Descriptor: VERSION_DATE
From the informed consent form (ICF), do Licensees (LIC) have access to trial data for reuse, including conducting clinical trials or further developing and commercializing the licensed product under agreed terms? Look for terms like: "transfer," "transaction," "merger," "acquisition," or references to another organization taking over development. Respond: "YES" if the ICF explicitly allows data access or transfer to licensees or successor entities. "NO" only if it is mentioned that licensees or successor entities explicitly prohibited from accessing or using trial data. "NA" if access is not mentioned or is unclear or no information is available. Extract information from document to support the answer. Descriptor: LIC
Continued on next page...

Prompts
From the informed consent form (ICF), do Research Partners or non-commercial Collaborators (RPC) have access to trial data for reuse? Reply "YES" if access is explicitly granted to research partners or collaborators (e.g., "research partners," "collaborating institutions," "non-commercial collaborators"). Reply "NO" only if data access by research partners or collaborators (e.g., "research partners," "collaborating institutions," "non-commercial collaborators") is explicitly prohibited. Reply "NA" if access is not mentioned or is unclear or no information is available. Extract information from document to support the answer. Descriptor: RPC
From the informed consent form (ICF), do External Independent Researchers (RPE) have access to trial data for reuse (e.g., to conduct independent studies, validate, challenge, or expand on the sponsor's findings)? Reply "YES" if the ICF explicitly allows data sharing with independent researchers (e.g., universities, scientific institutions). Reply "NO" only if access by independent researchers (e.g., universities, scientific institutions) is explicitly prohibited. Reply "NA" if access is not mentioned or is unclear or no information is available. Example indicator of "NA" reply: "Please be aware that no data or information can be used for any other research purposes in the future by any means without prior referral to the Research Ethics Committee of the Egyptian Ministry of Health and Population." Extract information from document to support the answer. Descriptor: RPE
From the informed consent form (ICF), do Sponsors, affiliates, or third-party processors (SPAFI) have access to trial data for reuse (e.g., analysis, storage, or management)? Reply "YES" if sponsors, affiliates, or third-party processors have access to trial data. Reply "NO" only if access is explicitly prohibited to sponsors, affiliates, or third-party processors to access trial data. Reply "NA" if it is not mentioned that sponsors, affiliates, or third-party processors are granted access to trial data or no information is available. Example indicators of YES reply: "Your coded data and bio samples are needed for the Sponsor, its research partners and service providers. . .". "The sponsor, designated personnel, and collaborating organizations may access your data. . ." Example indicator for NA reply: "The samples or parts of them may be transferred to the sponsor, its research partners and service providers (like clinical research organizations or laboratories) including companies belonging to the Boehringer Ingelheim Group of Companies." Extract information from document to support the answer. Descriptor: SPAFI
From the information provided in the informed consent form (ICF). Can trial data be used to improve medication quality and related substances? This includes understanding the medication's safety, efficacy, pharmacokinetics, pharmacodynamics, and interactions. Look for phrases such as: "understand how the trial drug work in the body and the study drug mode of action" If it is mentioned that data can be reused to understand how trial drug works, then reply YES. If it is explicitly mentioned that data "cannot" be used to understand how trial drug works, reply NO. If it is not explicitly mentioned that data can be used to understand how the trial drug works or no information is available, reply NA. Extract information from document to support the answer. Descriptor: MED_TR

Continued on next page...

Prompts
From the information provided in the informed consent form (ICF). Can trial data be used to improve medication quality and related substances? This info is in the confidentiality and data privacy section. Look for phrases such as: "understand how the trial drug and similar drugs work in the body and the study drug mode of action" If it is mentioned that data can be reused to understand how similar trial drug works or better understand related diseases, then reply YES. If it is explicitly mentioned that data "cannot" be used to improve related medication or similar trial drugs, reply NO. If it is not explicitly mentioned that data can be used to understand similar trial drugs or related medication or no information is available, reply NA. Extract information from document to support the answer. Descriptor: MED_TR_REL
From the informed consent form (ICF), can trial data be used to improve the quality or understanding of any medications or substances? This includes safety, efficacy, pharmacokinetics, pharmacodynamics, and interactions. Reply: "YES" if the ICF clearly allows data use for understanding or improving any medications or substances. "NO" only if it explicitly stated that use of data to improve or understand any medication or any trial drugs is not allowed. "NA" if this is not mentioned or is unclear or no information is available. Example indicator of "NA" reply: "Please be aware that no data or information can be used for any other research purposes in the future by any means without prior referral to the Research Ethics Committee of the Egyptian Ministry of Health and Population." Extract information from document to support the answer. Descriptor: MED_ALL
From the informed consent form (ICF), can trial data be reused broadly to study the trial disease and any other diseases (not just related ones)? Reply "YES" if the ICF explicitly allows reuse to understand other diseases, associated illnesses, risk factors, or affected populations (e.g., "to better understand yours and other diseases"). Reply "NO" only if the broad use of data to understand other diseases is explicitly prohibited. Reply "NA" if it is not mentioned that data can be used to understand any other diseases or no information is available. Extract information from document to support the answer. Descriptor: DIS_ALL
From the information provided in the informed consent form (ICF). Can trial data be reused to learn about the trial disease? This includes detailed analysis of the disease's nature, progression, and impact on the trial's patient population. Example: "better understand yours, related diseases and associated health problems" If it is mentioned that data can be use to allow for detailed analysis to provide more understanding of the disease's nature, progression, and impact on the trial's patient population reply YES. If it is explicitly mentioned that data "cannot" be used to understand the disease's nature, progression and impact on trial's patient population, reply NO. If it is not explicitly mentioned that data can be used to understand the disease's nature, progression, and impact on the trial's patient population or no information is available, reply NA. Extract information from document to support the answer. Descriptor: DIS_TR
Based on the informed consent form (ICF), can trial data be reused to understand the trial disease and related conditions? Reply "YES" if the ICF explicitly allows data use to study the trial disease and related or similar diseases (e.g., "to better understand your disease and related health problems"). Reply "NO" if it explicitly prohibits such use. Reply "NA" if this is not mentioned or is unclear. Extract information from document to support the answer. Descriptor: DIS_TR_REL

Continued on next page...

Prompts
From the informed consent form (ICF), can trial data be reused to study the trial disease or other diseases within specific therapeutic areas (e.g., cardiovascular, oncology, immunotherapy, respiratory)? Respond based on the following criteria: Reply "YES" if the ICF explicitly allows reuse to understand diseases within these or other defined therapeutic areas. Reply "NO" only if it is mentioned that data "cannot" be used to understand other diseases within the defined therapeutic area/areas. Reply "NA" if the use of data to understand diseases within these or other defined therapeutic areas is not mentioned or is unclear. Extract information from document to support the answer. Descriptor: DIS_TR_TA
From the informed consent form (ICF), determine whether trial data can be broadly reused to support the development of any diagnostic products — including unrelated ones — such as new technologies, improved tools, or diagnostic strategies. Respond based on the following criteria: Reply "YES" if the ICF explicitly allows data reuse for any diagnostic product development. Reply "NO" if only the ICF explicitly prohibits data reuse specifically for any diagnostic product development. Reply "NA" if the ICF does not mention use of data for any diagnostic product development or is unclear or no information available. Example indicator of NA reply: "Please be aware that no data or information can be used for any other research purposes in the future by any means without prior referral to the Research Ethics Committee of the Egyptian Ministry of Health and Population." Extract information from document to support the answer. Descriptor: DPROD_ALL
From the information provided in the informed consent form (ICF). Can trial data be reused to develop diagnostic products related to the trial disease? This includes creating, testing, and refining diagnostic tools, devices, or procedures specifically for the disease under investigation. Example: "develop diagnostic tests for, or drugs to treat yours and related diseases." If it is mentioned that data can be reused to support the development of diagnostic products related to trial disease, reply YES. If it is explicitly mentioned that data "cannot" be used to support the development of "related" or "similar" diagnostic products is explicitly forbidden, reply NO. If it is not explicitly mentioned that data can be used to support the development of "related" or "similar" diagnostic products or no information is available, reply NA. Examples indicators of NA reply: "Please be aware that no data or information can be used for any other research purposes in the trial". "Please be aware that no data or information can be used for any other research purposes in the future by any means without prior referral to the Research Ethics Committee of the Egyptian Ministry of Health and Population." Extract information from document to support the answer. Descriptor: DPROD_TR_REL
Continued on next page...

Prompts
From the information provided in the informed consent form (ICF). Does the ICF mention that trial data can be reused to develop other types of therapeutic products, tools, devices, or procedures, including new treatments (not only related trial)? If it is mentioned that data can be reused to support the development of any other types of therapeutic products, reply YES. If it is explicitly mentioned that data "cannot" be used to support the development of any other types of therapeutic products, tools, devices or procedures, including new treatments is explicitly forbidden, reply NO. If it is not explicitly mentioned that data can be used to support the development of any other types of therapeutic products, tools, devices or procedures, including new treatments or no information is available, reply NA. Examples indicators of NA reply: "Please be aware that no data or information can be used for any other research purposes in the procedures which are not known at this time." "Please be aware that no data or information can be used for any other research purposes in the future by any means without prior referral to the Research Ethics Committee of the Egyptian Ministry of Health and Population." Extract information from document to support the answer. Descriptor: TRPROD_ALL
From the information provided in the informed consent form (ICF). Does the ICF allow data reuse to develop therapeutic products related to the trial disease? If it is mentioned that data can be reused to support the development of therapeutic tools and procedures related to the disease being studied in a clinical trial, reply YES. If it is explicitly mentioned that data "cannot" be used to support the development of therapeutic tools and procedures related to the disease being studied in a clinical trial is explicitly forbidden, reply NO. If it is not explicitly mentioned that data can be used to support the development of related of therapeutic products, tools, devices or procedures, including new treatments or no information is available, reply NA. Example indicator of "NA" reply: "Please be aware that no data or information can be used for any other research purposes in the future by any means without prior referral to the Research Ethics Committee of the Egyptian Ministry of Health and Population." Extract information from document to support the answer. Descriptor: TRPROD_TR_REL
From the information provided in the informed consent form (ICF). Does the ICF mention that trial data can be used to improve the quality of this and other trials? This includes learning from past studies to inform future trial design and enhance scientific analysis methods. Look for the words "to improve quality". For example, "learn from past studies to plan new studies or improve scientific analysis methods". If it is mentioned that data can be reused to improve quality of this and other trails, reply YES. If it is explicitly mentioned that data "cannot" be used to improve quality of this and other trails reply NO. If it is not explicitly mentioned that data can be used to improve quality of this and/or other trails or if no information is available reply with NA. Extract information from document to support the answer. Descriptor: QUAL_ALL
Continued on next page...

Prompts
From the information provided in the informed consent form (ICF). Does the ICF mention if the link for re-identification of data subject will be deleted in 30 years? Check for word 'link', implying link between data and IDs. Examples: "All coded data, including yours, will be kept by the Sponsor. Only your trial doctor will be able to link your unique code number to you. This link will remain at the trial site for a maximum of 30 years and will then be destroyed by the trial doctor. After that it is not possible to link your unique code number directly back to you." - "This link will remain at the trial site for a maximum of 30 years and will then be destroyed by the trial doctor." "While the data can be collated in medical institutions for a maximum of 30 years after the completion of the clinical trial, the data that the investigator can collate will be discarded after that period. After that, you will not be able to directly match your unique code number." The answer can be only YES or NO. If it is mentioned that the link for re-identification of data subject will be deleted in 30 years, reply YES, else reply NO. Extract information from document to support the answer. Descriptor: DLINK30
From the information provided in the informed consent form (ICF). Does the ICF mention if data can be used for up to 50 years? Answer can be only YES or NO. If the data can be stored for 50 years, respond with YES. Else respond with NO. Extract information from document to support the answer. Descriptor: RR50
From the information provided in the informed consent form (ICF). Does the ICF mention if data can be used for up to 80 years? Answer can be only YES or NO. If the data can be stored for 80 years, respond with YES. Else respond with NO. Extract information from document to support the answer. Descriptor: RR80
From the information provided in the informed consent form (ICF). Does the ICF contain sentences with the same implied meaning as any of the following examples? Examples: - "Furthermore, law number 19.628 will be followed strictly." - "All personal information regarding your participation in this study will be confidential, with the exception of cases in which access is required by law." - "Additional research outside of the trial using encoded data must be approved by the ethics committee." - "Prior approval by the ethics committee is required if an external researcher uses data for a project outside the scope of the ICF." - "Additional research outside of the trial using encoded data must be approved by the ethics committee." The answer can be only YES or NO. If any of the above examples or similar appear, respond with YES. Else if not reply with NO. Descriptor: RESTRICTIVE_WORDING
Continued on next page...

Prompts
From the information provided in the informed consent form (ICF). Does the ICF mention Genomic information. Look for mentions of genes, genetic testing, DNA, or RNA. Examples of phrases that imply genomic information: 1. For ICFs in Dutch ICFs, look for the following examples: - “Uit het lichaamsmateriaal dat we voor dit extra onderzoek bij u afnemen halen we de erfelijke informatie”. “Deze stofjes kunnen bijvoorbeeld eiwitten, RNA of DNA zijn. RNA en DNA bevat erfelijke informatie.” 2. English specific examples: - Your DNA (genetic information) will be removed from one blood sample. - In this trial, non-genetic and genetic biomarker testing will be done. - DNA genetic tests will be conducted to evaluate specific genes known to cause mutations changes in the gene structure) that may lead to GPP. RNA tests will be conducted to identify genes involved in how the investigational drug works in the body, how the body responds to the drug, and the severity of the disease. If words such as: genes, genetic testing, genetic, genetic information, DNA, or RNA or similar are mentioned, reply YES. Else if no mention of words such as: genes, genetic testing, genetic, genetic information, DNA, or RNA, reply NO. Descriptor: GENOMIC_INFO

Appendix A.2 1.2. GPT-4-o versus GPT-4.1

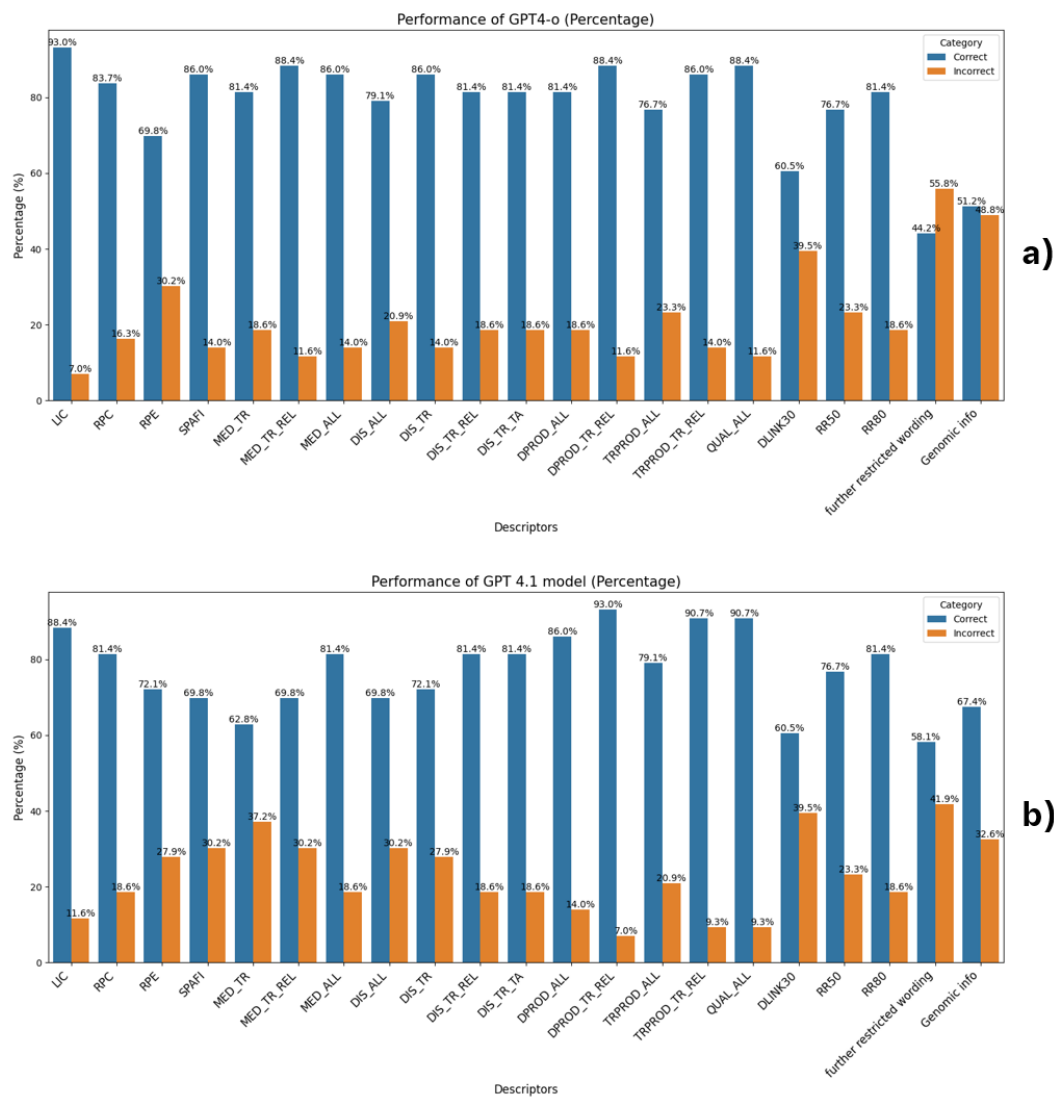


Figure A1. Comparison of GPT-4-o (a) versus GPT-4.1 (b) on several descriptors with GPT-4.1 outperforming GPT-4-o on descriptors requiring instruction following and reasoning. FURTHER_RESTRICTIONS refers to the RESTRICTIVE_WORDING prompt.

Appendix A.3 1.2. GPT-4-o versus GPT-4.1

Table A2. Performance of the different chunk sizes in terms of accuracy (%) for all data sharing descriptors. The performance was conducted on 438 ICFs.

Descriptor	500	1000	2000
LIC	78.5	86.9	91.1
RPC	70.6	72.8	78.3
RPE	76.9	82.4	88.8
SPAFI	85.6	86.3	88.1
MED_TR	87.7	87.7	89.0
MED_TR_REL	85.6	88.4	89.0
MED_ALL	63.7	58.5	58.0
DIS_ALL	88.8	85.2	81.7
DIS_TR	85.4	88.4	88.8
DIS_TR_REL	84.5	87.4	88.4
DIS_TR_TA	63.5	58.5	58.0
DPROD_ALL	82.2	80.1	73.3
DPROD_TR_REL	85.2	89.5	93.2
TRPROD_ALL	80.6	76.7	71.5
TRPROD_TR_REL	83.1	86.7	91.3
QUAL_ALL	84.3	85.6	90.6
DLINK30	77.6	77.2	77.9
RR50	91.3	92.5	92.7
RR80	94.3	94.5	94.5
RESTRICTIVE_WORDING	59.1	61.9	64.6
GENOMIC_INFO	62.3	63.7	65.1
Overall	79.6%	80.5%	81.6%

Appendix A.4 Example of results from the Application

	descriptor	outcome	extract	reasoning
0	COUNTRY	NA	No information available	The provided informed consent form is a template that includes placeholders such as <insert country of OPU>.
1	TRIAL_NO	1404-0043	Footer to be adapted locally as needed (e.g. version and date): Main Consent Form dated 30 Oct 2020 Page 1 of 2	The trial number is explicitly listed in the ICF in multiple locations as "Trial No: 1404-0043" and "PROTOCOL No.
2	VERSION	1	Informed-consent-main-version-01-2020-10-30.pdf	The file name at the very start of the document is "informed-consent-main-version-01-2020-10-30.pdf". Followin
3	VERSION_DATE	30-Oct-2020	Footer to be adapted locally as needed (e.g. version and date): Main Consent Form dated 30 Oct 2020 Page 1 of 2	The version date is explicitly stated in the document footer: "Main Consent Form dated 30 Oct 2020". The date is
4	LIC	YES	In case another organization takes over development or commercialisation of the trial drug, your coded data or	The text explicitly states that if another organization (i.e., a licensee or successor entity) takes over developmen
5	RPC	YES	Your coded data and biosamples are needed for the Sponsor, its research partners and service providers (like cl	The informed consent form explicitly states that the sponsor's "research partners" and "service providers" may r
6	RPE	YES	The Sponsor is convinced that access to trial data advances clinical science and medical knowledge and is in th	The form explicitly states that anonymized data may be shared with "credible researchers" for approved scienti
7	SPAFI	YES	Your coded data and biosamples are needed for the Sponsor, its research partners and service providers (like cl	The document explicitly states that the sponsor, its affiliates, research partners, and service providers (includin
8	MED_TR	YES	Your coded data and biosamples are needed for the Sponsor, its research partners and service providers (like cl	The consent form clearly states that coded data and biosamples will be used to understand how the trial drug w
9	MED_TR_REL	YES	Your coded data and biosamples are needed for the Sponsor, its research partners and service providers (like cl	The consent form explicitly states that coded trial data will be used not only for this study but also in related res
10	MED_ALL	YES	Your coded data and biosamples are needed for the Sponsor, its research partners and service providers (like cl	The provided excerpt explicitly states that trial coded data may be used by the sponsor, research partners, and i
11	DIS_ALL	NA	No information available	The provided informed consent form gives information on the use of coded data and biosamples, stating that it
12	DIS_TR	YES	The data will be used in this study and in related research activities necessary for the drug development progr	The explicit statement in the data sharing section indicates that the coded data may be reused "to better under
13	DIS_TR_REL	YES	Your coded data and biosamples are needed for the Sponsor, its research partners and service providers (like cl	The ICF explicitly states that coded data can be used to "better understand yours, related diseases and associat
14	DIS_TR_TA	NA	No information available	After careful review of the provided informed consent form, there is no explicit information provided regardin
15	DPROD_ALL	NO	The data will be used in this study and in related research activities necessary for the drug development progr	The ICF only allows the use of coded data for diagnostic tests or drugs to treat "yours and related diseases". It do
16	DPROD_TR_REL	YES	The data will be used in this study and in related research activities necessary for the drug development progr	The document states, "The data will be used in this study and in related research activities necessary for the dru
17	TRPROD_ALL	NA	No information available	The Consent Form provides information on the use, sharing, and anonymization of coded trial data and biosam
18	TRPROD_TR_REL	YES	Your coded data and biosamples are needed for the Sponsor, its research partners and service providers (like cl	The provided extract explicitly states that coded data from the trial may be used to develop drugs to treat the pi
19	QUAL_ALL	YES	• learn from past studies to plan new studies or improve scientific analysis methods	The document explicitly mentions in the "Use of Your Coded Data and Biosamples" section that data will be use
20	DLINK30	YES	All coded data, including yours, will be kept by the Sponsor. Only your trial doctor will be able to link your uniq	The consent form explicitly states that the link between the unique code number and the participant will remai
21	RR50	NO	Storage of Your Coded Data All coded data, including yours, will be kept by the Sponsor. Only your trial doctor w	The ICF specifies that the link between the participant's personal data and the coded trial data will be kept at th
22	RR80	NO	All coded data, including yours, will be kept by the Sponsor. Only your trial doctor will be able to link your uniq	The ICF clearly states that the link to the subject's identity (unique code number) will be kept for a maximum of
23	FURTHER_RESTRICTIONS	NO	No sentences with the same implied meaning as the provided examples were found in the ICF. There are no stat	I reviewed the informed consent form and did not identify any sentences requiring new ethics committee appr
24	GENOMIC_INFO	YES	As an optional part of this trial, you are being asked to allow the collection and storage of additional blood sam	The ICF specifically mentions the collection and storage of blood samples for DNA extraction, and lists pharmac

Figure A2. The full outcome from the application. FURTHER_RESTRICTIONS refers to the RESTRICTIVE_WORDING prompt.

Appendix A.5 Validation

Table A3. The number of ICFs analysed grouped by country, including the number of ICFs which were completely correct and the number of ICFs where the extracted outcome on data sharing descriptors were correct. OPUs = operative units (templates).

Country	Total ICFs	ICFs All Descriptors Correct (%)	ICFs Data Sharing Correct (%)
OPUs	47	11 (23.4%)	13 (27.7%)
Argentina	3	2 (66.7%)	2 (66.7%)
Austria	35	13 (37.1%)	13 (37.1%)
Belgium	4	2 (50.0%)	2 (50.0%)
Canada	14	2 (14.3%)	2 (14.3%)
Chile	4	0 (0.0%)	0 (0.0%)
China	12	2 (16.7%)	2 (16.7%)
Czech Republic	14	1 (7.1%)	1 (7.1%)
Denmark	23	3 (13.0%)	3 (13.0%)
Germany	44	24 (54.5%)	24 (54.5%)
Hungary	6	4 (66.7%)	4 (66.7%)
Ireland	9	7 (77.8%)	7 (77.8%)
Italy	5	0 (0.0%)	0 (0.0%)
Japan	15	0 (0.0%)	1 (6.7%)
Republic of Korea	51	4 (7.8%)	6 (11.8%)
Mexico	7	1 (14.3%)	2 (28.6%)
Netherlands	20	5 (25.0%)	5 (25.0%)
Norway	10	10 (100.0%)	10 (100.0%)
Poland	39	32 (82.1%)	32 (82.1%)
Portugal	4	0 (0.0%)	1 (25.0%)
Russia	12	4 (33.3%)	7 (58.3%)
Spain	31	10 (32.3%)	12 (38.7%)
Sweden	2	0 (0.0%)	0 (0.0%)
Taiwan	13	5 (38.5%)	8 (61.5%)
Turkey	2	2 (100.0%)	2 (100.0%)
United Kingdom	32	8 (25.0%)	9 (28.1%)
USA	30	18 (60.0%)	18 (60.0%)

References

1. Part 46: protection of human subjects. Code of Federal Regulations. <https://www.ecfr.gov/on/2018-07-19/title-45/subtitle-A/subchapter-A/part-46>, 2018. Accessed: 2025-12-18.

2. Unlu, O.; Shin, J.; Mailly, C.J.; Oates, M.F.; Tucci, M.R.; Varugheese, M.; Waghlikar, K.; Wang, F.; Scirica, B.M.; Blood, A.J.; et al. Retrieval-Augmented Generation-Enabled GPT-4 for Clinical Trial Screening. *NEJM AI* **2024**, *1*. <https://doi.org/10.1056/aioa2400181>.

3. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.; Chen, D.; Dai, W.; et al. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys* **2023**, *55*, 1–38, [2202.03629]. <https://doi.org/10.1145/3571730>.

4. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv* **2020**, [2005.11401]. <https://doi.org/10.48550/arxiv.2005.11401>.

5. Amugongo, L.M.; Mascheroni, P.; Brooks, S.; Doering, S.; Seidel, J. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health* **2025**, *4*, e0000877. <https://doi.org/10.1371/journal.pdig.0000877>.

6. Shi, Q.; Luzuriaga, K.; Allison, J.J.; Oztekin, A.; Faro, J.M.; Lee, J.L.; Hafer, N.; McManus, M.; Zai, A.H. Transforming Informed Consent Generation Using Large Language Models: Mixed Methods Study. *JMIR Medical Informatics* **2025**, *13*, e68139. <https://doi.org/10.2196/68139>.

7. Liu, H.; Borgman, A.; Gagnon, J.; Boisvert, D.; Valian, H. Revolutionizing Informed Consent Form Analysis with Generative AI: Enhancing Efficiency in Drug Development and Data Re-use. In Proceedings of the Phuse US Connect. Phuse US Connect, 2024.
8. Decker, H.; Trang, K.; Ramirez, J.; Colley, A.; Pierce, L.; Coleman, M.; Bongiovanni, T.; Melton, G.B.; Wick, E. Large Language Model-Based Chatbot vs Surgeon Generated Informed Consent Documentation for Common Procedures. *JAMA Network Open* **2023**, *6*, e2336997. <https://doi.org/10.1001/jamanetworkopen.2023.36997>.
9. Vaira, L.A.; Lechien, J.R.; Maniaci, A.; Tanda, G.; Abbate, V.; Allevi, F.; Arena, A.; Beltramini, G.A.; Bergonzani, M.; Bolzoni, A.R.; 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* **2025**, *53*, 18–23. <https://doi.org/10.1016/j.jcms.2024.10.002>.
10. OpenAI.; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2023**, [2303.08774]. <https://doi.org/10.48550/arxiv.2303.08774>.
11. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A Family of Highly Capable Multimodal Models, 2025, [arXiv:cs.CL/2312.11805].
12. 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* **2024**, *6*, 1145–1153. <https://doi.org/10.3390/make6020053>.
13. Rudra, P.; Balke, W.T.; Kacprowski, T.; Ursin, F.; Salloch, S. Large language models for surgical informed consent: an ethical perspective on simulated empathy. *Journal of Medical Ethics* **2026**, *52*, 85–90. <https://doi.org/10.1136/jme-2024-110652>.
14. Allen, J.W.; Levy, N.; Wilkinson, D. Empowering Patient Autonomy: The Role of Large Language Models (LLMs) in Scaffolding Informed Consent in Medical Practice. *Bioethics* **2026**, *40*, 183–193. <https://doi.org/10.1111/bioe.70030>.
15. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, [2005.14165]. <https://doi.org/10.48550/arxiv.2005.14165>.
16. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S.R. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv* **2018**, [1804.07461]. <https://doi.org/10.48550/arxiv.1804.07461>.
17. Adiwardana, D.; Luong, M.T.; So, D.R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. Towards a Human-like Open-Domain Chatbot. *arXiv* **2020**, [2001.09977]. <https://doi.org/10.48550/arxiv.2001.09977>.
18. OpenAI. GPT-4.1 Model | OpenAI API — developers.openai.com. <https://developers.openai.com/api/docs/models/gpt-4.1>. [Accessed 02-03-2026].
19. OpenAI. text-embedding-3-small. <https://developers.openai.com/api/docs/models/text-embedding-3-small>. Accessed: 2026-2-14.
20. European Parliament.; Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive). <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>, 2016. Accessed: 2026-2-18.
21. Geroimenko, V. *The Essential Guide to Prompt Engineering, Key Principles, Techniques, Challenges, and Security Risks*; Springer Cham, 2025; pp. 37–83. https://doi.org/10.1007/978-3-031-86206-9_3.
22. OpenAI. Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1/>, 2025. Accessed: 2025-12-9.
23. Streamlit. Streamlit. <https://streamlit.io/>. Snowflake Inc. Accessed: 2026-01-10.
24. Fachada, N.; Fernandes, D.; Fernandes, C.M.; Ferreira-Saraiva, B.D.; Matos-Carvalho, J.P. GPT-4.1 Sets the Standard in Automated Experiment Design Using Novel Python Libraries. *Future Internet* **2025**, *17*, 412, [2508.00033]. <https://doi.org/10.3390/fi17090412>.
25. Jones, N. AI hallucinations can't be stopped — but these techniques can limit their damage. *Nature* **2025**, *637*, 778–780. <https://doi.org/10.1038/d41586-025-00068-5>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.