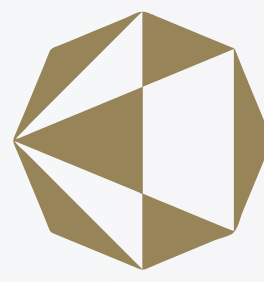


Application Tested

Tester

# Clinical Study Report Authoring



resaro

MSD, a global pharmaceutical company, tested iRAP - an authoring tool for Clinical Study Reports (CSRs) that integrates LLM capabilities with rule-based logic to facilitate rapid generation. Designed to enhance efficiency of document development process for medical authors.

Resaro offers independent, third-party assurance of mission-critical AI systems, with extensive experience in testing Computer Vision and Generative AI systems.

## How LLMs were used in application?

- Summarisation
- Retrieval augmented generation
- Data extraction from unstructured source




## What Risks Were Considered Relevant And Tested?

### ✓ Inaccurate/Incomplete output

Generation of inaccurate, incomplete, or hallucinated content in CSRs, which can undermine trust in the tool, lead to delay in drafting, and necessitate additional costly resources for manual oversight and/or corrections


## How Were The Risks Tested?

### Approach

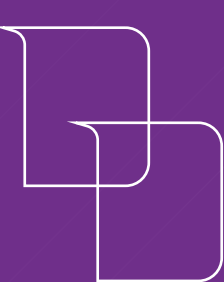
-  Creating a targeted evaluation data set (using data from a single publicly available clinical trial)
-  Passing it through the iRAP app, and
-  Comparing the final output from the app against the evaluation dataset

➔ Metrics Used  
Faithfulness, Contradiction, Prompt Alignment

### Evaluators

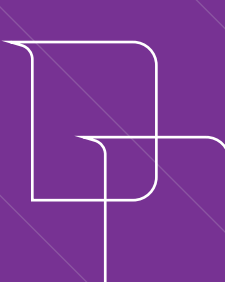
-  LLM as a judge
-  Review by human experts

## Challenges



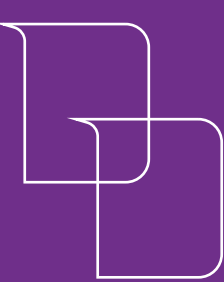
### Limited Dataset Variety

Due to confidentiality reasons and access controls, available trial data represented a narrow range of studies and therapeutic areas



### Need for Domain Knowledge/Context

Review of CSR content required familiarity with clinical terminology and trial-specific logic



### Robustness Testing Limitations

Restrictions on synthetic data generation and API access limited the ability to isolate and stress-test LLM-generated components independently

## Insights

01

LLMs as judges reliably identified factual inconsistencies, hallucinations, and prompt misalignments. However, for use cases with domain-specific evaluation criteria, targeted prompt engineering is needed to ensure their reliability

03

Hallucination is a priority concern for LLM-based applications for medical authoring

02

Early validation of access protocols and data flows is critical when planning third-party evaluation. Team had to pivot initial test design due to architectural constraints and data governance requirements

04

With careful coordination, even limited-scope testing can yield valuable insights and lay the foundation for broader evaluation efforts



# Clinical Study Report (CSR) Authoring



Merck & Co., Inc., Rahway, New Jersey, USA is a global pharmaceutical company (outside the U.S. and Canada, Merck & Co., Inc., Rahway, New Jersey, USA is known as MSD).

## Use Case

The pilot focused on a GenAI application called iRAP for Clinical Study Reports (CSR) - a web-based authoring tool that integrates Large Language Model (LLM) capabilities with rule-based logic to facilitate the rapid generation of CSRs. The application is designed to enhance the efficiency of the document development process for medical authors, ensuring that the generated content aligns with MSD’s established templates, styling, and authoring principles.

## Application Workflow

The application uses a sequential pipeline that combines rule-based logic, LLM-generated content, and structured authoring workflows to allow the production, editing, and review of draft CSRs.

### Application Features:



#### Rapid Document Generation

iRAP enables users to swiftly create draft CSRs by employing business logic and predefined rules. This functionality ensures that relevant information is accurately extracted from study data, reducing the time spent on manual drafting.



#### Proprietary Prompt Library

The application utilises a proprietary prompt library that allows for the generation of text based on specific inputs, thereby enhancing the quality and relevance of the content produced. This feature ensures that the generated text adheres to the required scientific and regulatory standards.



#### Intuitive User Workflow

The user workflow is designed to be intuitive, allowing authors to select relevant trial-specific data, such as the protocol and Tables, Listings, and Figures (TLFs). This process is crucial for generating comprehensive and accurate CSRs.



#### Structured Output

The system produces a 14-section report draft, which includes essential components such as Synopsis, Study Objectives, Investigational Plan, Participants, Efficacy, Safety Evaluations, Conclusions.




#### Authoring and Editing Workflows

The workflow allows for further editing of the generated CSR draft. Authors can easily modify sections, incorporate feedback, and ensure that the final CSR meets all regulatory and quality standards before submission.




Resaro is an independent, third-party AI assurance provider that conducts testing of AI systems. For this pilot, Resaro took the following testing approach:



### Test Dataset Collection and Preparation


Collected a test dataset of pairs consisting of input data - parsed elements of protocol documents and pre-processed TLFs - and the corresponding LLM-generated output.



### Test Execution for Hallucination

Defined what constitutes a hallucination in the context (e.g. inclusion of the following in the generated output:

- Unsupported insights from the results,
- Inconsistency in sorting and ranking of values from the results,
- Fabricating non-existent protocol/TLFs elements





### Response Evaluation


Measured the hallucination tendency of the model for generated paragraphs in a CSR through a measure of faithfulness to the corresponding input protocol sections and/or TLFs. LLMs guided through appropriate prompts had been used as a judge (e.g., Guided GPT-4o-mini), followed by human reviewers to check on discordant examples as a secondary human-in-the-loop evaluation.

**Risk Assessment**

The deployment of the iRAP application for generating CSRs introduces several risk considerations that must be carefully managed to ensure the quality (specifically factuality) of the report drafting process. Below is a subset of applicable risks, along with additional considerations specific to this use case:

-  **Inaccurate/Incomplete Output:**

The major risk is the generation of inaccurate, incomplete, or hallucinated content in CSRs, which can undermine trust in the tool, lead to delay in drafting, and therefore necessitate additional costly resources spent on manual oversight and/or corrections.
-  **Security Risk**

That goes beyond malicious use of this application: Broader cybersecurity risks such as sensitive data closure, prompt injection, and jailbreaking are critical but not covered in this pilot.
-  **Bias and Content Risks:**

Risks of unfair bias and generation of inappropriate or unsafe language are moderate concerns for the use case but not covered within this pilot.

Risks like breach of use-case specific compliance, inadequate transparency, inappropriate data disclosure, and other embarrassing content are assessed as low or not directly relevant.

**Scope of Testing**

The testing specifically focused on the accuracy and correctness of the output summarized by the LLM. Testing was conducted on a dataset of compiled protocols, TLFs, and LLM-generated paragraphs.

Technical tests were designed to specifically address the identified risks, combining automated and manual methods:

Data Extraction

The test procedure involved manual extraction of input data from source protocols and TLFs. The prompt library containing instructions for the LLM to generate various sections of the CSR was also extracted.

Information Consistency Evaluation

Designed to evaluate factual consistency in generated CSR outputs with respect to the information present in the input protocol and/or TLFs. A hallucination was defined as any instance of content that:

- (i) introduces unsupported or fabricated details, or
- (ii) misrepresents or omits critical protocol or TLF information.

Generated outputs were systematically compared against the inputs by using an LLM as a judge

Evaluation metrics included:

- **Faithfulness metric**  
Quantified how factually consistent the generated content is compared to the source information. Multiple implementations of the metrics that utilise different prompt templates can be utilised.
- **Contradiction metric**  
A variation of faithfulness metrics that focused purely on contradictions (and not support) in the generated content compared to the input.
- **Prompt Alignment metric**  
Identified if the generated output followed the instructions present in the corresponding prompt.

Human reviewers logged all detected hallucinations and discordant examples based on these judgement criteria:

- **Source Alignment**  
Verified whether each fact stated in the output was directly traceable to the corresponding protocol or TLF data.
- **Verifying against Prompt Instructions**  
Flagged any discrepancies between prompt instructions and the generated output.
- **Content Appropriateness**  
Identified any content that could mislead stakeholders or violate expected reporting standards.



Execution of Tests

01

- **Manual data collection**  
An evaluation dataset (based on mapping of protocol and TLF to expected content) was manually prepared.
- **Automated test runs**  
An evaluation pipeline was built and deployed to test each input-output pair using an LLM as a judge against pre-defined evaluation metrics.
- **Manual review**  
The input-output pairs that had been flagged to be inaccurate or hallucinated by the LLM as a judge were then manually compiled and shared in a csv format to facilitate verification and provide explanation and feedback by testers.
- **Tests were executed within a secure, access-controlled staging environment.**

Cost of Testing

03

- **The testing process involved significant time allocation, primarily driven by data access approvals, manual data extraction, and review of test results.**
- **From the Deployer Organisation's business and product teams:**  
Approx. 2 weeks were spent coordinating and securing the necessary approvals to access the trial data, navigating multiple stakeholder groups and confidentiality requirements.
- **From the Testing Organisation's technical team:**
  - Approx. 1 week was spent on manual data extraction and structuring (reconstructing protocol and TLF mappings) to prepare a usable test dataset, and,
  - Approx. 1 week was spent on running the automated hallucination tests, reviewing flagged outputs, applying the human evaluation criteria, and consolidating review feedback.
- **There were no significant direct computational costs associated with LLM usage during this testing phase. The primary resource investment was in human effort for approvals management, manual data preparation, and detailed review activities.**

Data Used in Testing

02

- **Testing was conducted using a single CSR report generated from a publicly available clinical trial. This study was selected due to limited access to proprietary trial data; however, it provided sufficient breadth of input-output test data pairs to support meaningful evaluation.**
- **As direct access to the original protocol and TLF inputs was unavailable, the tester manually reconstructed source information by extracting relevant content from the generated CSR. This included parsing tables, narratives, and protocol-aligned sections to create a baseline for hallucination detection.**
- **Synthetic or mock data was intentionally not used, as each clinical trial is highly specific in design, terminology, and data structure. Using real historical data ensured that testing reflected realistic reporting patterns and regulatory requirements, which would not have been accurately captured through artificial data generation.**

Challenges in Implementation

04

- **Limited Dataset Variety**  
Due to confidentiality reasons and access controls, available trial data represented a narrow range of studies and therapeutic areas. Approvals were coordinated across multiple stakeholder groups to enable the use of a publicly available clinical trial. While coverage was limited, the selected report allowed for meaningful evaluation of core system functionality.
- **Robustness Testing Limitations**  
Restrictions on synthetic data generation and API access limited the ability to isolate and stress-test LLM-generated components independently. Although not critical to the factuality testing objectives, these constraints reduced opportunities for deeper robustness analysis.
- **Need for Domain Knowledge/Context**  
Review of CSR content required familiarity with clinical terminology and trial-specific logic. To support the evaluation, the backend prompt library was referenced to clarify how outputs were generated. This helped streamline interpretation and maintain review accuracy without full-time SME involvement.



### Insights on Risk Assessment

The pilot reinforced that hallucination - through measuring faithfulness to corresponding protocol and TLFs elements - is a priority concern for LLM-based applications in the context of medical authoring. While other risks such as bias or inappropriate language were assessed as low likelihood, the importance of faithfulness in representing trial data was underscored. The exercise also highlighted that data access can influence the depth and scope of testing in highly controlled environments.

### Lessons from Test Design

Initial test design assumed access to structured inputs (protocols and TLFs), which was later adapted due to architectural constraints and data governance requirements. The team successfully pivoted by reconstructing inputs from the final CSR, though this added complexity to the test implementation. This experience highlights the importance of early validation of access protocols and data flows when planning third-party evaluations.

### Insights from Test Implementation

Implementation effort was shaped primarily by organisational safeguards around data access, rather than technical barriers to evaluation. While test execution itself was streamlined through automation, significant time was invested in securing approvals and preparing test data through manual extraction. The pilot ultimately demonstrated that with careful coordination, even limited-scope testing can yield valuable insights and lay the foundation for broader evaluation efforts.

### Manual Verification of the LLM as a Judge's Evaluation

Manual review of flagged outputs showed that the LLM as a judge was generally reliable in identifying factual inconsistencies, hallucinations, and prompt misalignments. For specialised use cases, the reliability of the LLM as a judge can be further improved through targeted prompt engineering to align with domain-specific evaluation criteria. While a small number of edge cases required further checks against the input source, the step-wise approach in the verification gave confidence not just in the generated output, but also additional assurance in the LLM's usefulness as a first-pass evaluator.