GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

Medical Report Summarisation

How LLMs were used in application?

Summarisation | Classification and recommendation | Data extraction from unstructured source

Application Tested

Changi General Hospital
SingHealth

Changi General Hospital (CGH), a major public hospital in Singapore, created a GenAI application to summarise colonoscopy and histopathology reports based on US Multi-Society Task Force (USMTF) guidelines. The goal was to provide clinicians with structured, guideline-based summary recommendations.

Tester

softserve

SoftServe is a global digital consulting company specialising in software development, data science, and AI/ML solutions.

What Risks Were Considered Relevant And Tested?

Relevant risks:

- Accuracy of Clinical Summaries
- Clinical Safety and Alignment to Guidelines

CGH was also interested in assessing the reliability of its existing automated evaluation framework (which used a well-known, open source summarisation metric) for real-world deployment

How Were The Risks Tested?

Approach

Through back testing on historical data, annotated with ground truth by subject matter experts (SMEs). Evaluation was split into 2 parts:

- Comparison of key facts extracted from the medical reports vs the ground truth facts from SME
- Comparison of final recommendation (arrived at by applying above facts to the USMTF guidelines) to the ground truth recommendation from SME

Evaluators

Rule based logic:

Python scripts used both to extract the key facts from CGH's application output, and to simulate the USMTF guidelines

Challenges

Getting SMEs to provide a scoring rubric ("what good looks like") that covers all potential scenarios

Obtaining ongoing support from (busy) SMEs to test incremental improvements

Constraints imposed by the design of the GenAI application: e.g., can determine whether granular interim data is available for testing, or whether the application can actually be improved based on test results

Insights

01

Evaluations that include confidence scores (e.g., in accuracy of extracted key facts) can help increase overall level of automation. Where the system suggests lower confidence scores, the case can be auto routed to the human in the loop

02

Using LLMs as judges need not be the only option – particularly when the evaluation can be designed with granular interim checkpoints

03

Adding detailed citations (source text or code fragment for a particular fact or recommendation) to the system's output design can make human review and annotation easier

04

Measuring accuracy of interim outputs results in greater transparency and suggestions to improve original system design

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY | AI VERIFY FOUNDATION

Medical Report Summarisation

Changi General Hospital
SingHealth

Changi General Hospital (CGH) is a major public hospital in Singapore, providing a comprehensive range of medical services with a focus on patient-centric and innovative healthcare.

Use Case

The pilot focuses on a Generative Artificial Intelligence (GenAI) application designed to summarise colonoscopy and histopathology reports based on US Multi-Society Task Force (USMTF) guidelines. The goal is to provide clinicians with structured, guideline-based summary recommendations to support surveillance and management decisions, while enabling scalable deployment without relying on manual human verification.

The application uses a fine-tuned Large Language Model (LLM) deployed on the PAIR platform (developed by Open Government Products Singapore) to generate summaries from free-text clinical reports. Prior to the pilot, CGH had already experimented with an "off the shelf" summarisation accuracy metric (using a popular open source evaluation tool). The pilot was an opportunity to explore potential alternatives that could improve the reliability of such automated evaluation.

High-level Architecture

The application utilises a sequential pipeline architecture:

01 Colonoscopy and histopathology report texts are input into the PAIR interface.

02 The reports are processed using a system prompt crafted to guide the fine-tuned Claude LLM (via PAIR) to extract key facts (e.g., number of polyps, size, histology) and produce a structured summary, including a surveillance recommendation aligned with USMTF guidelines.

03 The generated summary is evaluated using automated scoring methods.

Note that a Retrieval-Augmented Generation (RAG) structure was not employed, as the primary function was summarisation of given clinical report content rather than retrieval of external knowledge.

02

Testing Partner and Testing Approach

softserve

SoftServe is a global digital consulting company specialising in software development, data science, and AI/ML solutions. For this pilot, SoftServe partnered with Changi General Hospital (CGH) to co-develop a custom evaluation framework for its GenAI-powered clinical summarisation application.

For this pilot, SoftServe's approach was to build on CGH's earlier evaluation efforts using a popular open source evaluation toolkit, complement and compare it with a more structured and granular framework. The approach separated the evaluation into two distinct components: (a) extraction of key facts from the colonoscopy and histopathology reports, and (b) generation of recommendation as per USMTF guidelines (on the basis of (a)).

03

Risk Assessment and Testing Scope

CGH identified several key risks for the GenAI summarisation application, based on its intended use in a clinical environment where accuracy and patient safety are critical. The following risks were prioritised:

Accuracy of Clinical Summaries

Summaries must correctly extract key clinical facts (e.g., number of polyps, histology findings, resection completeness) and accurately generate surveillance recommendations based on USMTF guidelines. Accurate summaries which are guideline-based can lead to appropriate and personalised surveillance intervals.

Clinical Safety and Alignment to Guidelines

Recommendations provided in the summary must adhere strictly to the USMTF surveillance guidelines without introducing unsupported medical advice. Deviations from established clinical guidelines could compromise patient care and expose the hospital to medico-legal risks.

CGH was also concerned about a third risk, but this was related to the automated evaluation score it was using prior to the pilot, rather than the application itself

Evaluation Reliability for Real-World Deployment

The evaluation method must reliably reflect human expert judgment to ensure that future large-scale deployments of the app do not require manual review for every case. Poor evaluation reliability could lead to undetected summarisation errors, undermining clinical confidence in the tool.

Scope of Testing

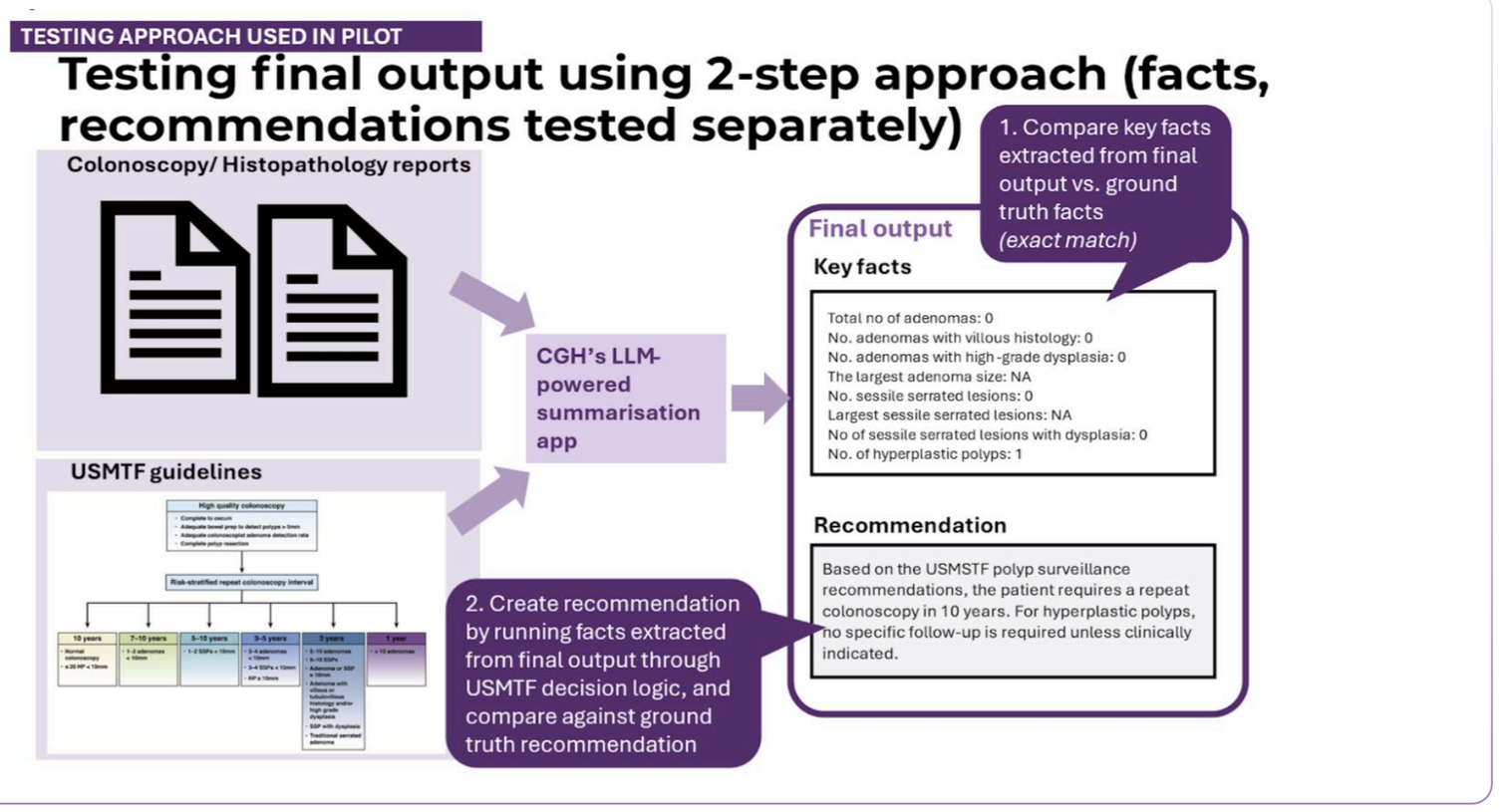
The testing focused on evaluating the accuracy of factual extraction and the recommendation suitability of the summarisation outputs. It also included a comparison of the effectiveness of the new structured testing approach implemented by Softserve, with CGH's original automated summarising quality scoring. Testing was conducted on a curated dataset of de-identified historical colonoscopy and histopathology reports prepared by CGH.

Test Design

To recap, the goal of the testing was:

To assess whether CGH's GenAI powered app (PAIR) was extracting the 8 facts correctly from the source report

To assess whether, given input data, CGH's (PAIR) app was providing the correct recommendation, in line with the USMTF guidelines. To determine what is a "correct recommendation", we had to pass the extracted data from step 1 through the rules in the USMTF guidelines, which in turn had been codified into rules implemented in Python first.



Testing Data Preparation:

The application's original output was a nicely formatted paragraph detailing the 8 key facts as well as the final recommendation. The CGH team made a slight modification to their original prompt in order to output the 8 key facts in a specific JSON format. Additionally, Softserve extracted the recommendation. Softserve considered this JSON, for the purposes of testing, to be the original application output, **PAIR JSON**.

Softserve also had a "gold standard" of what the correct response should be. This was manually curated by the CGH team along with SMEs. Softserve converted this to equivalent standard JSON format as well for the purpose of facilitating testing. For the purposes of testing, to be **Gold JSON**.

Type of tests:

- **Test A - Fact Extraction Check**: Facts extracted via LLM **(Pair JSON)** compared with the gold standard **(Gold JSON)**.
- **Test B - Recommendation Check**: Recommendation via LLM **(Pair JSON)** compared with the gold standard **(Gold JSON)**
- **Self-Consistency tests:**
  - **Test C - JSON Self consistency:** Extracted JSON facts validated for self- consistency. Ie, if 0 adenomas are found, then the field for largest adenoma should be NA
  - **Test D - Pair Self consistency: Pair JSON** recommendation validated for self-consistency within the USMTF surveillance guidelines (via Python)
  - **Test E: Gold Self consistency: Gold JSON** contents validated for self-consistency (this last test was not conducted during the pilot period)

Test Implementation

Execution of Tests:

Tests were executed using a custom set of Python scripts using input data based on outputs of PAIR prompts and the correct output “ground truth” given by the SME. The Python scripts aimed to generate accuracy checks for each of the key facts and the recommendation for each colonoscopy / histopathology report as outlined.

The testing was conducted in a secure staging environment with strict access controls.

Data Used in Testing:

01 Approximately 100 anonymised colonoscopy and histopathology reports were used. Each had “ground-truth” labels based on human input. No additional synthetic data was used – as it was not feasible to generate this algorithmically within this pilot.

02 Future extensions may include the use of synthetic data to test the system more extensively (e.g., using abbreviations, including multiple facts such as from historical data). As well as a prescription/prediction for real world accuracy measure i.e., you need to validate 1000 examples.

Cost of Testing:

The testing process involved time allocation from the Changi General Hospital and SoftServe teams, on the order of 30 hours. The costs of LLM usage were minimal.

Commentary on Results:

While the results are confidential to the deployer and tester, there are a few test-level insights gathered from tests A and B that can be shared without breaching confidentiality

Test A - Fact Extraction Check:

This test compared the facts extracted via LLM **(Pair JSON)** with the gold standard **(Gold JSON)**. The former were managed as a one-shot extraction via the Pair prompt, while the latter were extracted one variable at a time. It appears that the latter technique was significantly more accurate at fact extraction.

Test B - Recommendation Check:

This test compared the recommendation via LLM **(Pair JSON)** with the gold standard **(Gold JSON)**.

Insight 1: While conducting this test, it was noticed that the rationale for the recommendation was not clear (versus the USMTF guidelines) in some instances, while in a few other cases, some edge cases were not programmed into the python code.

Insight 2: Sometimes, the recommendation matched ground truth even though Test A had failed. This could be due to the fact that sometimes, a recommendation could be the result of multiple criteria, so even if one of them was assessed incorrectly, others could result in the same recommendation. There could have been some over-fitting in the Pair prompt, but this would need further investigation.

Insights/Lessons Learned

The pilot reinforced that more detailed "accuracy" scores (at a key fact level) would help to improve the prompting and system design, to improve overall outputs. One ongoing challenge is to get human labels to test incremental improvements in prompts and system design, as the complexity of input documents suggests there is a large variability that cannot be captured within 100 documents (our sample size for the testing).

Implementing automated evaluation within an existing system (in this case the PAIR tool) on an ongoing basis would require either 1) significant integration work to "automatically" implement evaluation + deterministic Python code execution in a user's workflow as it is not natively supported by the PAIR tool, or 2) increase the number of steps that a user would need to perform (i.e., multiple prompts). This requires changing behaviour and expectations of users.

Having a Subject Matter Expert in the loop is imperative during implementation and on an ongoing basis, to ensure that edge cases can be understood. Given that fact extraction is likely to be a key element of LLM app accuracy evaluations in many situations, additional effort may be needed to increase confidence in such extraction for it to be deployed in the real-world. Such an effort could include:

- Confidence scores (LLM-generated) for fact extraction and rejecting facts with low confidence (either via prompt or by extracting the logits/ probabilities of the predictions)
- Providing more examples of abbreviations / jargon associated with key facts as few-shot examples specific to key fact extraction
- Adding text snippets from which the facts were extracted, to enable immediate checks by the human reviewer, e.g.,:
  - In the JSON output, the *largest adenoma size* fact extraction should be accompanied by additional text stating that the fact was extracted from this text snippet "8mm sessile NICE 2 rectal polyp"
- Rigorously calculating the extent to which "ground truth" is needed, in order to be more confident in the application's real time accuracy. The consideration is similar to the way in which samples are tested in manufacturing lines.