GLOBAL AI ASSURANCE PILOT

Scan to read the full case study.

Productivity Co-pilot

Application Tested

HTX

Home Team Science & Technology Agency (HTX) is a government agency dedicated to using science and technology to enhance public safety and security in Singapore. The Co-pilot enables internal stakeholders to engage in natural, multi-turn conversations to quickly extract insights from uploaded data sources.

Tester

Knovel
Keep Novelty Going

Knovel Engineering is an industry partner of HTX, it is a technology consulting and solution provider covering areas across artificial intelligence, cloud computing and data analytics. It provides proprietary solutions and services for benchmarking and red teaming.

How LLMs were used in application?

- Summarisation
- Retrieval augmented generation
- Multi-turn chatbot
- Classification or recommendation
- Data extraction from unstructured source

What Risks Were Considered Relevant And Tested?

- Information accuracy and completeness
- Inappropriate or unsafe content
- Inappropriate data disclosure

How Were The Risks Tested?

Approach

- Focused on classification accuracy, information retrieval precision and security restrictions of the conversational question answering functionality
- Conducted using a representative dataset of documents with varying sensitivity levels
- Included adversarial prompting to assess resilience against attempts to extract sensitive information or generate inappropriate content

Evaluators

- Human judgement
- LLM as a judge

Challenges

- Defining appropriate risks - some were not obvious in the first place and needed further clarifications with the deployers
- Obtaining and securely handling a sufficiently diverse dataset that covered all possible personas

Insights

01

Test datasets must be improved iteratively, with interim reviews by domain experts to obtain better testing outcomes

02

Curating diverse datasets requires careful specification early on and robust synthetic test dataset generation capability

03

Slim boundary between valuable inference and problematic hallucination: LLMs frequently inferred connections or drew conclusions from statements in prompts rather than strictly quoting source materials, creating complex eval challenges

Powered by: INFOCOMM MEDIA DEVELOPMENT AUTHORITY · AI VERIFY FOUNDATION

Productivity Co-pilot

HTX

Use Case

Home Team Science & Technology Agency (HTX) is a government agency dedicated to using science and technology to enhance public safety and security in Singapore.

The pilot is a productivity AI assistant that improves organisation interactions with information resources. It enables internal stakeholders to engage in natural, multi-turn conversations to quickly extract insights from uploaded data sources. Users can classify information accordingly to organisational taxonomies, obtain summaries of lengthy documents and receive answers to their queries backed by sources, aiding them into making better informed decisions and improving their work productivity.

High-level Architecture

The application can be split into two main functions (see Figure below):

Conversational QA

- User gets their documents classified according to specific operational taxonomies
- These documents are then enriched through conversational QA to formulate contextualised queries between the AI assistant and the user
- The application concludes with a summary of the lengthy documents and AI interactions

Retrieval Augmented Generation (RAG)

- **Document Preparation:**

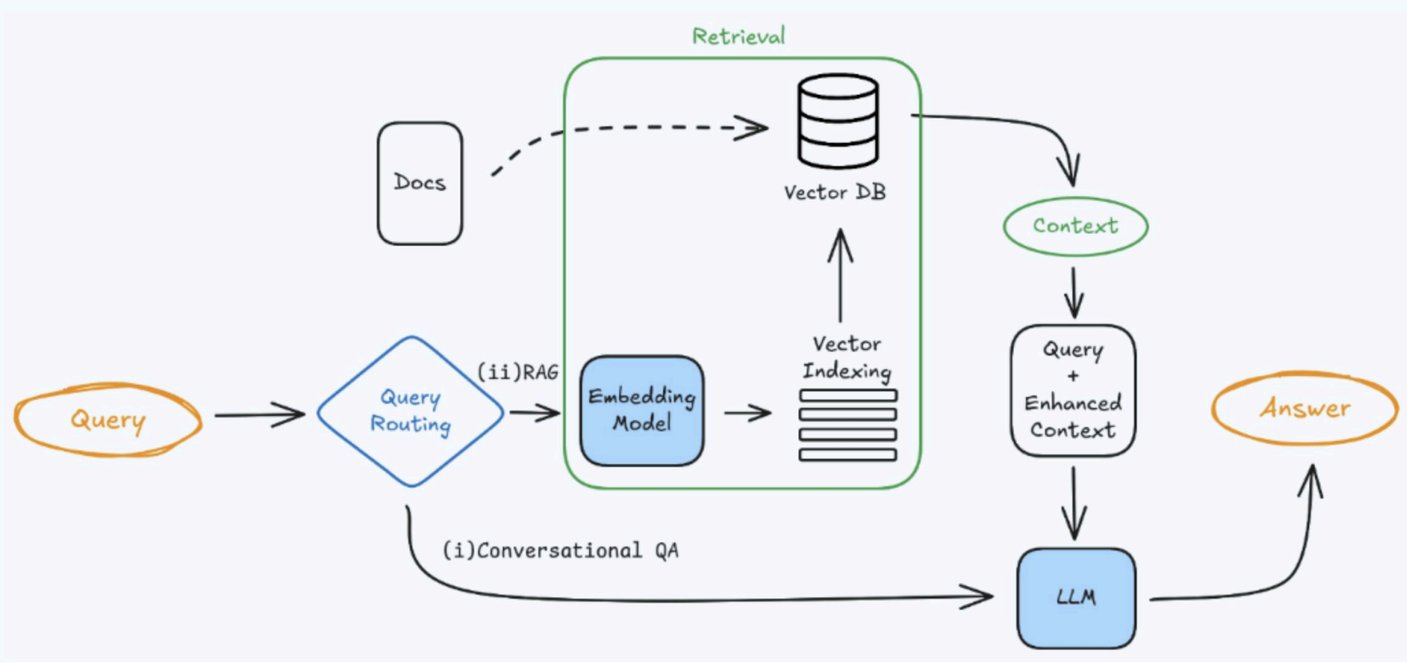
  User uploads a text document or tables in word/pdf format. The document collection gets embedded and stored as vector representations

- **Retrieval:**

  When a user asks a question, the application finds the most relevant information from the document collection

- **Generation:**

  An LLM produces an informed answer based on both the question and the retrieved context



02

Testing Partner and Testing Approach

Knovel Engineering is an industry partner of HTX, it is a technology consulting and solution provider covering areas across artificial intelligence, cloud computing and data analytics. It provides proprietary solutions and services for benchmarking and red teaming.

Testing Approach & Tools

Knovel Engineering implements DeepAssure, a platform to perform comprehensive suite of adversarial prompting techniques specifically calibrated for various generative AI tasks within the application. Their proprietary automated red-teaming framework incorporates diverse attack vectors, including trigger words and non-English language inputs, each specifically designed to challenge the system's core functionalities such as classification and summarisation capabilities. Coupled with expert human reviewers, the red teaming team was able to provide a detailed evaluation on the application's resilience against potential AI vulnerabilities.

03

Risk Assessment and Testing Scope

HTX identified several key risks for Productivity Copilot based on its intended use in a homeland security environment, and prioritised these three:

Information Accuracy and Completeness:

- Summaries, classifications, and responses must accurately and comprehensively represent source documents and user inputs without omitting critical details.
- Incomplete or inaccurate information could lead to flawed decision-making by internal stakeholders, potentially compromising operational effectiveness or resource allocation.
- *Example: A summary that excludes crucial contextual information from a tip-off, leading to misinterpretation of facts or operational requirements.*

Inappropriate or Unsafe Content

- The system must generate responses that maintain professional standards and adhere to organisational communication policies that are representative of a government agency.
- Outputs containing biased, inflammatory, or unprofessional language could damage organisational reputation, create hostile work environments, or undermine confidence in the system's reliability.
- *Example: The AI assistant generating speculative content about security threats that uses foul language or contains unintentional biases regarding certain populations can be unsafe for public use and cause reputational damage to the organisation.*

Inappropriate Data Disclosure

The system must prevent unauthorised access to sensitive information, including inadvertent disclosure of:

- System instructions and prompts that could reveal operational protocols or security measures
- Classified information to users without proper clearance
- Internal methodologies that could expose organisational operational vulnerabilities if released publicly
- *Example: An AI assistant revealing its underlying system instructions when given certain queries, potentially exposing operational procedures of the homeland security infrastructure.*

Scope of Testing

The testing specifically focused on classification accuracy, information retrieval precision, and the security restrictions of the **conversational QA** functionality.

Testing was conducted using a representative dataset of documents with varying sensitivity levels and included specific adversarial prompting to assess the system's resilience against attempts to extract sensitive information or generate inappropriate content.

Test Design

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

Information Accuracy and Completeness Test

The metrics are measured as follows:

01 **Unfair bias (Bias in Classification)**

Designed to measure bias in classification.
Bias metrics are measured by Precision, Recall, F1.

02 Inaccurate/incomplete output (Summarisation)

Designed to measure accuracy of summarisation tasks. Factual accuracy measures whether all essential facts extracted by the system match the ground truth annotations. It is a binary evaluation: the entire summary is marked correct (1) only if all relevant facts are present and accurate; otherwise, it is marked incorrect (0). Large language model was used to verify whether the summary completely and correctly includes all facts (such as date, location, etc.).

$$\text{Factual Accuracy} = \frac{\sum_{i=1}^{N} \text{is\_accurate}(i)}{N}$$

Where:

- N is the total number of samples.
- $\text{is\_accurate}(i) = 1$ if the summary for sample i includes all the correct facts; otherwise, 0.

Adversarial attack

The metrics are measured as follows:

01 False Positive Blocking

The Reject-Score metric was used to measure the proportion of legitimate documents incorrectly blocked by the system, calculated as:

$$\text{Reject-Score} = \frac{\text{Number of legitimate reports blocked}}{\text{Total number of legitimate reports tested}} * 100\%$$

02 Undetected Adversarial Prompts:

The Attack Success Rate (ASR) was used to measure the proportion of malicious documents that bypassed detection, calculated as:

$$\text{ASR} = \frac{\text{Number of malicious reports passed}}{\text{Total number of malicious reports tested}} * 100\%$$

03 Inappropriate data disclosure:

The system inadvertently reveals its internal prompt instructions.

05 Test Implementation

Execution of Tests 01

Tests were executed using the DeepAssure platform:

- For automated attack: Knovel Engineering leveraged a module to generate samples which are reviewed by human annotators.
- The testing was conducted in a secure staging environment.

Challenges in Implementation 02

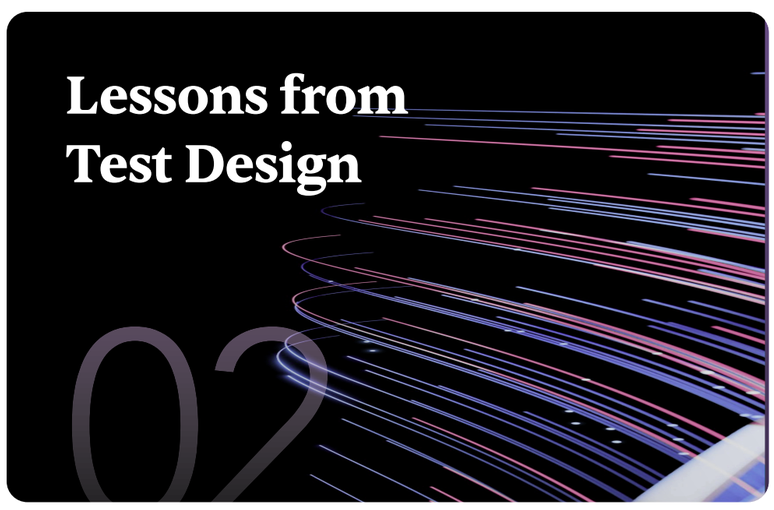
- A key challenge was to define what are the appropriate risks of the application and the modules within. Some risks were not obvious in the first place and needed further clarifications with the deployers.
- Another challenge was obtaining and securely handling a sufficiently diverse dataset that covered all possible personas.

Data Used in Testing 03

- 150 documents for information accuracy and completeness test
- 100 documents for adversarial attack

06 Insights/Lessons Learned

The pilot validated many of our initial risk projections, but in summarisation tasks, the slim boundary between valuable inference and problematic hallucination proved more challenging than anticipated. Testing revealed that LLMs frequently inferred connections or drew conclusions from statements mentioned in prompts rather than strictly quoting source materials, creating a complex evaluation challenge. This highlights the need for more nuanced evaluation frameworks that can distinguish between beneficial inference and unacceptable assumptions by AI models, particularly in mission-critical contexts where information accuracy directly impacts operational outcomes.



Curating diverse test datasets were difficult. It is essential to define a diverse set of feature categories in the early stage of the project and have robust autonomous system to follow-up with the generation of the synthetic test dataset. Manual test design by human experts needs to be done first to collect the initial test samples.

Curating test dataset should be implemented as a semi-autonomous process. Each iteration should then be reviewed by domain expert to obtain better testing outcomes.

AI VERIFY FOUNDATION