Committee	Select Committee on Adopting Artificial Intelligence (AI)
Question No.	001
Reference	21 May 2024
Committee member	Senator David Shoebridge

Questions

On 21 May 2024, ASIC appeared before the Senate Committee on Adopting Artificial Intelligence. ASIC officials took a question on notice (**QoN**) to provide a "*report*" to the Committee about ASIC's trial using AI. An extract of the Hansard where this QoN was taken is set out below.

> *Senator DAVID POCOCK: You mentioned the experiment you did, where you got your team to feed in submissions and then you got the AI to—are you willing to table a summary of the differences in what was generated? Is that something you can do?*
> *Mr Jefferson: We can provide some more detail about the results of that experiment on notice.*
> *Mr Longo: I remember, when the commission got briefed about this, the AI gave what I would call a 'bland summary' of the submissions. It wasn't misleading but it was bland. It really didn't capture what the submissions were saying, while the human was able to extract nuances and substance. I think that is not a bad summary.*
> *Mr Jefferson: Yes. We were principally interested in where there were references to ASIC in public submissions to the Senate Finance and Public Administration References Committee inquiry into the consulting companies. What we found was that in general terms, as the chair said, the summaries were quite generic, and the nuance about how ASIC had been referenced wasn't coming through in the AI-generated summary in the way that it was when an ASIC employee was doing the summary work.*
> *Senator DAVID POCOCK: Was that with an Australian company?*
> *Mr Longo: We did it ourselves.*
> *Mr Jefferson: We did it ourselves in conjunction with AWS based here in Australia. We used the Llama 2 large language model, which, I believe, is a Meta product.*
> *Senator DAVID POCOCK: With AWS being Amazon Web Services, you used two US companies to help?*
> *Mr Jefferson: Essentially, yes.*
> *…*
> *Senator SHOEBRIDGE: Did you say you would take on notice Senator Pocock's request for the tabling of the report? You would have had a report dealing with your offline AI test.*
> *Mr Jefferson: We can provide some more detail about it.*
> *Mr Longo: Can I be quite open with the committee: I'll be as open as I can, subject to public interest immunity.*
> *Senator SHOEBRIDGE: I'm asking for the tabling of the report. If you can take it on notice, then you can deal with it.*
> *Mr Longo: I think we'll be able to provide some colour for the committee, to show what we did and what we learned from it.*
> *Senator SHOEBRIDGE: I'm asking for the tabling of the report; that's what I'm asking for, and I assume you're taking it on notice?*
> *Mr Longo: Okay.*

Answers

ASIC is exploring and refining how artificial intelligence (AI) could be adopted into our work, recognising that AI continues to develop at a rapid pace.

ASIC completed a successful initial 'proof of concept' (PoC) between 15 January and 16 February 2024, to assess the capability of a Generative AI (Gen AI) Large Language Model (LLM) to summarise a sample of public submissions made to an external Parliamentary Joint Committee Inquiry. ASIC partnered with Amazon Web Services Australia Pty Ltd (AWS) to undertake the PoC.

As at the date of the hearing on 21 May 2024, the PoC had resulted in the production of a draft report that was prepared by AWS, incorporating input from ASIC on previous iterations of the report. Attached is a version of the report as at, 21 May 2024 which we note is a draft only. ASIC has also consulted with AWS prior to providing this draft report to the Committee.

The objectives of the PoC were to explore and trial Gen AI technologies, to focus on measuring the quality of the generated output rather than performance of the models, and to understand the future potential use of Gen AI. The PoC was an exploratory trial only and was ***not*** used for ASIC's regulatory or operational purposes.

The PoC was run over a five-week period, including the selection of the LLM, followed by an experimentation and optimisation phase with the selected LLM. The final phase of testing involved a rigorous process to assess the performance of the AI-generated summaries compared to human-generated summaries. The final results showed that the AI summaries performed lower on all criteria compared to the human summaries. The findings support the view that Gen AI should be positioned as a tool to augment and not replace human tasks.

It is important to note that the results should not be extrapolated more widely as:

- the timeframe allowed to optimise the model was limited to one week as this was as short duration PoC; and
- the PoC's point-in-time results relate to the use of certain prompts using a specific LLM and selected for one specific use-case. This limits the generalisability of the findings to other use cases and LLMs.

ASIC has taken a range of learnings from the PoC, including: the value of robust experimentation; the need for collaboration between subject matter experts and data science specialists; the necessity of carefully designed prompt engineering; and given the rapidly evolving AI landscape, the importance of providing a safe environment that allows for rapid experimentation to ensure ASIC has a continued understanding of the various uses for AI, including its shortcomings.

Technology is advancing in this area and it is likely that future models will improve performance and accuracy of results. There are opportunities for Gen AI as the technology continues to advance.

Attachment: Draft Report *'ASIC Gen AI Document Summarisation PoC – Draft report as at 21 May 2024'*

Generative Artificial Intelligence (AI) Document Summarisation Proof of concept

Final Report

March 2024

Prepared for Australian Securities and Investments Commission (ASIC)

Author: AWS Professional Services

Disclaimer

Disclaimers. The following applies to this document and all other documents, information, data, and responses (written or verbal) provided by Amazon Web Services, Inc. or any of its affiliates (collectively, "**AWS**") in connection with responding to this request and other related requests (collectively, this "**Response**"): This Response is expressly (a) informational only and provided solely for discussion purposes, (b) non-binding and not an offer to contract that can be accepted by any party, (c) provided "as is" with no representations or warranties whatsoever, and (d) based on AWS's current knowledge and may change at any time due to a variety of factors such as changes to your requirements or changes to AWS's service offerings. All obligations must be set forth in a separate, definitive written agreement between the parties. Neither party will have any liability for any failure or refusal to enter into a definitive agreement. All use of AWS's service offerings will be governed by the AWS Customer Agreement available at http://aws.amazon.com/agreement/ (or other definitive written agreement between the parties governing the use of AWS's service offerings) (as applicable, the "**Agreement**"). If the parties have an applicable Nondisclosure Agreement ("**NDA**"), then the NDA will apply to all Confidential Information (as defined in the NDA) disclosed in connection with this Response. AWS's pricing is publicly available and subject to change in accordance with the Agreement. Pricing information (if any) provided in this Response is only an estimate and is expressly not a binding quote. Fees and charges will be based on actual usage of AWS services, which may vary from the estimates provided. Nothing in this Response will modify or supplement the terms of the Agreement or the NDA. No part of this Response may be disclosed without AWS's prior written consent.

CONTENTS

1 Executive Summary

Australian Securities and Investments Commission (ASIC) procured Amazon Web Services (AWS) Professional Services to run a Proof of Concept (PoC) between 15 January and 16 February 2024, to assess the capability of Generative AI (Gen AI) to summarise a sample of public submissions made to an external Parliamentary Joint Committee inquiry, looking into audit and consultancy firms[1]. The objectives of the PoC were to explore and trial Gen AI technologies, to focus on measuring the quality of the generated output rather than performance of the models and, to understand the future potential for business use of Gen AI. The PoC was *not* used for any of ASIC's regulatory work or business activities.

The PoC consisted of multiple phases, with a preparation/set-up stage occurring before the PoC:

- Phase 1: Selection of the Large Language Model (LLM) to be used in Phase 2
- Phase 2: Experimentation and optimisation with the selected LLM (Llama2-70B)
- Phase 3: Final assessment

The final assessment results of the PoC showed that out of a maximum of 75 points, the aggregated human summaries scored 61 (81%) and the aggregated Gen AI summaries scored 35 (47%). Whilst the Gen AI summaries scored lower on all criteria, it is important to note the PoC tested the performance of one particular AI model (Llama2-70B) at one point in time. The PoC was also specific to one use case with prompts selected for this kind of inquiry.

In the final assessment ASIC assessors generally agreed that AI outputs could potentially create more work if used (in current state), due to the need to fact check outputs, or because the original source material actually presented information better. The assessments showed that one of the most significant issues with the model was its limited ability to pick-up the nuance or context required to analyse submissions.

Key observations and lessons learnt from the PoC included:

- To a human, the request to summarise a document appears straightforward. However, the task could consist of several different actions depending on the specifics of the summarisation request. For example: answer questions, find references, impose a word limit. In the PoC the summarisation task was achieved by a series of discreet tasks. The selected LLM was found to perform strongly with some actions and less capably with others.
- Prompting (prompt engineering) was key. 'Generic' prompting without specific directions or considerations resulted in lower quality output compared to specific or targeted prompting.
- An environment for rapid experimentation and iteration is necessary, as well as monitoring outcomes.
- Collaboration and active feedback loops between data scientists and subject matter experts was essential.
- The duration of the PoC was relatively short and allowed limited time for optimisation of the LLM.
- Technology is advancing rapidly in this area. More powerful and accurate models and GenAI solutions are being continually released, with several promising models released during the period of the PoC. It is highly likely that future models will improve performance and accuracy of the results.

The PoC provided valuable learnings, demonstrating the current capabilities of Llama2-70B as well as the potential for growth. Although there are opportunities for Gen AI particularly as the technology

[1] Ethics and Professional Accountability: Structural Challenges in the Audit, Assurance and Consultancy Industry – Parliament of Australia (aph.gov.au).

continue to advance, this PoC also found limitations and challenges for adopting Gen AI for this specific use case.

2 INTRODUCTION

The potential for Generative Artificial Intelligence (Gen AI) technology is immense. As such, Australian Securities and Investments Commission (ASIC) sought to explore how these technologies work in practice with a real-life use case in the organisation.

ASIC procured AWS to run a Proof of Concept (PoC) between 15 January and 16 February 2024, to assess the capability of Gen AI Large Language Models (LLM) to summarise a sample of public submissions made to an external Parliamentary Joint Committee inquiry, looking into audit and consultancy firms[2].

The project team consisted of ASIC's Chief Data and Analytics Office (CDAO) team, ASIC's Regulatory Reform and Implementation team (who acted as subject matter experts) and AWS.

3 OBJECTIVES OF THE POC

ASIC's CDAO team and AWS initially met in September 2023 and through a series of workshops, defined the objectives of the PoC as follows:

- Explore and trial Gen AI technologies
  - Use a real-life use case to explore Gen AI technologies
  - Understand the typical setup of a Gen AI solution
  - Understand how the different components can be optimised to produce better output
- Focus on measuring the quality of the output rather than performance
  - As a PoC, the main objective was the quality of the output
  - Performance and operation cost of the solution were secondary measures
- Understand the future potential for business use
  - Get a sense of the maturity level of the technology
  - Understand the potential for output quality improvement based on an optimisation process

The PoC was *not* used for any of ASIC's regulatory work or business activities - it was exploratory only.

4 METHODOLOGY

This section outlines the methodology used in this PoC, which included a Preparation / set-up stage for the PoC, followed by three phases of testing: Phase 1. Model selection, Phase 2. Optimisation, and Phase 3. Final assessment.

[2] Ethics and Professional Accountability: Structural Challenges in the Audit, Assurance and Consultancy Industry – Parliament of Australia (aph.gov.au).

4.1 Preparation / Set-Up

This section describes the preparation and set-up stage performed by ASIC and AWS to execute the PoC.

4.1.1 Technical Environment

In determining which technical environment would host the solution - given ASIC's IRAP (Information Security Registered Assessors Program) requirements - the following position was taken:

- The PoC was to be hosted on AWS's own accounts.
- Access to the AWS console was constrained to AWS staff only.
- AWS would conduct demos, playback, interactive sessions, and code reviews on a regular basis throughout the project.

4.1.2 Testbed

The solution was tested using a custom-built platform for iterative exploration, allowing rapid validation and prototyping, including but not limited to the following features:

- Use of various models available in SageMaker. A production-ready service, deployable in Australia.
- Interchangeable parameters and modes of the underlying LLMs without redeploying the solution.
- Production-ready levels of scalability and high-availability (subject to customer requirements and testing) based on serverless or auto-scaling architectural components.
- Encryption both in-transit and at-rest using customer managed keys (KMS).
- Use of a separate LLM for embedding.
- A document bot run within a private VPC in a container-based solution, allowing for future upgrades/updates.

4.1.3 Use case

The agreed use case for the PoC centred on document summarisation. This use case allowed a user to specify one or more documents (see section 05 – Dataset) to be processed by the Gen AI solution and allow the user to ask questions about the documents, with the solution generating a text as output to the user's query. See Figure 1. Document Summarisation Workflow. A summary of the use case included:

- **User Input**
    - File to be summarised.
    - User prompt to add context to the query. E.g. "What does this document say about ASIC"
- **Input File Type**
    - PDF or Word document files.
- **Output**
    - Text-based summary file.
- **Success Criteria**

- Output file created with relevant content.
- Final content to be evaluated for quality and accuracy against human generated summaries (by ASIC subject matter experts).

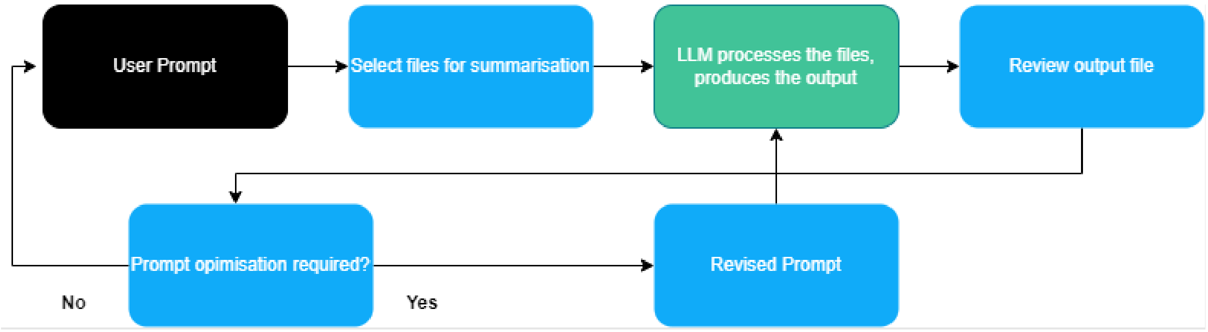


Figure 1 - Document Summarisation Workflow

4.1.4 PoC Timeframe

The PoC was limited to a five week duration, from 15 January to 16 February 2024, with one week for Phase 2, Optimisation.

4.1.5 Dataset

A relevant data set was required for the use case to be demonstrated.

- For the PoC, ASIC elected to use the public submissions to an external inquiry as the dataset: Parliamentary Joint Committee on Corporations and Financial Services inquiry into Ethics and Professional Accountability: Structural Challenges in the Audit, Assurance and Consultancy Industry (referred to as 'PJC inquiry' through-out this report).
- The dataset had a limited size / volume / number of documents and was not updated during the PoC.

4.1.6 Model Selection

A key feature of the PoC was the large language model (LLM) selection process. ASIC was interested in testing a range of LLMs of different sizes and capabilities. To allow for this, the PoC was structured into three phases: a model selection phase, optimisation phase and final assessment phase. ASIC selected three models to test in Phase 1 Model selection: Llama2-70B, Mistral-7B and MistralLite.

4.1.7 Model Prompts (summary criteria)

ASIC subject matter experts identified a range of criteria that they considered to be the main issues of interest to ASIC and what they wanted to see in the submission summaries. Prompts to the models were based on this set of criteria for all three phases of testing:

- Please create a summary for the document.
- Provide a summary of mentions to ASIC (Australian Securities and Investments Commission) with page numbers and brief context.
- Summarise recommendations on how conflicts of interest should be regulated (conflicts of interest where the entity has an audit business), refer to page numbers and brief context.

- Mention all references to more regulation of auditors/consultants with page numbers and brief context.

4.1.8 Human Summary Preparation by ASIC Staff

In October 2023, ASIC harnessed a business need for staff to complete (~54) summaries to the PJC inquiry as part of their BAU work. A structured approach was taken with the summarisation task in the advent that the summaries could be used as a comparison point in the final assessment phase. All 10 ASIC staff (ranging in staff levels) were given the same written instructions to complete the task and recorded information such as time taken to write the summaries. The following instructions were provided to the ASIC staff preparing the submission summaries:

Please create a summary for each of the below submissions. Doing the following:

- Focus on the main issues of interest to ASIC, being:
    - References to ASIC – include a brief indication of the context and page reference.
    - Recommendations on how conflicts of interest should be regulated (note: conflicts of interest where the entity has an audit business) – include a brief indication of the context and page reference.
    - References to more regulation of auditors/consultants – include a brief indication of the context and page reference.

4.1.9 Testing Approach

The methodology and testing approach for Phase 1 and Phase 2 of the PoC was developed by AWS, and the Phase 3 final assessment methodology was designed by ASIC. An assessment rubric used in Phase 1 and Phase 3 testing was designed by ASIC.

The testing approach used the human in the loop evaluation method. Testing of LLMs and the associated criteria are mainly subjective. As such the testing approach needs to reflect this. With human in the loop model evaluation, it is necessary to define the metrics and associated metric types.

Phase 1 – Model selection

In the Phase 1. Model selection methodology, ASIC selected three submissions to the PJC inquiry - which had also been summarised by human summarisers (ASIC staff) in the preparation/set-up stage - and provided the submissions and human summaries to AWS. The instructions provided by ASIC to the human summarisers were then used as the basis for a series of prompts provided to three different LLMs. It is important to note that the way LLMs function dictates that the queries or 'prompts' are provided as a series of questions. This necessitated some minor changes to the instructions provided to the human summarisers.

Due to the PoC being executed in AWS' controlled environment, the test execution involved a multi-step process developed by AWS:

1. AWS pre-loaded the dataset containing the selected submissions.
2. Using the human instructions, AWS ran a series of prompts against each LLM, for each submission.
3. Each set of responses were then combined into a single, cohesive summary. One summary per submission.

4. A document was prepared that contained the human summaries and the LLM generated responses, in the format of Document A - Human Summary, Document A - AI Summary. This was repeated for each of the chosen submissions.
5. This document (test materials) was provided to ASIC for assessment of the quality of the outputs.

In the test materials document the models were referred to as Model A, Model B and Model C without providing the relationship between actual model names and these labels. This was to maintain an element of blind assessment to ensure the model selection was objective.

Three ASIC staff undertook the assessment and validation of the AI generated outputs. Each assessor was assigned one submission to review and rate the three AI LLM outputs (labelled as Model A, B, C) in the materials provided by AWS. The models were presented in the same order for each submission. The assessors completed the assessment and rating independently.

ASIC assessors used the original submission as the source of truth ('golden answer') to validate and rate the AI summary outputs on the assessment rubric (see Appendix D). The human summary was also used to identify differences between the AI summary and the human summary (i.e. gaps or additional references). The human summaries of submissions created in the Preparation/set-up stage had not been validated and checked for accuracy. It's important to note that humans are also fallible and for the purpose of the PoC the human summaries were not considered to be the 'golden answer'.

Following the ratings, an internal debrief session was held with the three ASIC assessors to discuss their findings and key issues seen in each model with examples. The results of the assessment rubric ratings were provided to AWS alongside comprehensive qualitative feedback with examples and descriptions of key issues found by the three assessors for each model (see Appendix B - Summary of key issues observed by assessors for each model). This information was used by ASIC to select their one preferred LLM to move to the next phase of testing, Phase 2 - Optimisation.

PHASE 2 – OPTIMISATION

The Phase 2 Optimisation involved a series of activities performed by the AWS team to improve the responses generated by the chosen LLM (i.e. Llama2-70B) selected at the conclusion of Phase 1 Model selection. These optimisation techniques are covered in detail in section 4.3.1 below.

For this phase, three submissions to the inquiry were chosen by ASIC with two new submissions and one submission being the same as that used in the Phase 1 Model selection. The chosen submissions varied in submission length, format/type and organisation who had made the submission.

OPTIMISATION STEPS AND COMPONENTS

The AWS team executed various optimisation steps and used the model prompts (outlined in section 4.1.5) as the basis to generate responses against the three submissions. As one of the optimisation steps involved prompt engineering (adjusting the structure and words used in the prompt, but not the intent), there was some further variation in the prompts used in Phase 1. AWS focused on the following optimisation components as part of the retrieval augmented generation process (each component is explained in more detail below):

- Top-K and Temperature
- Llama-Index Index Type and Response Modes
- Prompt Engineering
- Per page chunking (not implemented for final testing)

Top-k and Temperature

Top-k Sampling is a technique which involves selecting the most likely next words or tokens based on their probabilities predicted by the model.

Temperature Scaling is used to control the randomness of the generated text. By adjusting to a lower temperature, this can produce more deterministic outputs whilst a higher temperature introduces more randomness and potentially less coherent responses.

lama-Index Index Types and Response Synthesis Modes

Index types are different methods of indexing to generate responses to queries. The two index types utilised as part of this PoC includes Vector store index and Summary index.

Vector store index will split up the document into nodes and will create vector embeddings of the text of each node. These vector embeddings will be queried by the LLM to perform semantic searches. Due to this, vector store index is well-suited towards question and answer use cases which require an understanding of the meaning behind words and phrases.

On the other hand, summary index is a data structure where nodes are stored in a sequence. During index construction, the document texts are chunked, converted into nodes and stored in a list. When it is queried, the summary index will iterate through the nodes and synthesize an answer from all the nodes. This will ensure a comprehensive summary of the entire document and is better suited towards summarisation use cases.

These two different index types can be utilised in conjunction with response synthesis modes. Response synthesis modes are different llama-index modules which synthesises a response given the retrieved node. As part of this PoC, there are four different response modes used: compact, refine, tree_summarise and simple_summarise.

- Compact: concatenates the chunks to fit as much text within the context window beforehand, resulting in less LLM calls.
- Refine: create and refines an answer by sequentially going through each retrieved text chunk. This makes a separate LLM call per node/retrieved chunk and is ideal for more detailed answers.
- Tree_summarise: queries the LLM using the prompt as many times as needed so that all concatenated chunks have been queried. Each chunk will be queried recursively until there is only one chunk, and one final answer left. This is good for summarisation purposes.
- Simple_summarise: truncates all text chunks to fit into a single LLM prompt. Good for quick summarisation purposes, but may lose detail due to truncation.

For further details on index types and response synthesis modes, see LlamaIndex Index Types and LlamaIndex Response Modes.

Prompt Engineering

Prompt engineering is the process of guiding generative AI solutions to generate desired outputs. In this PoC, AWS experimented with the prompt template and providing extra context within the prompt to improve the response of the chatbot. The prompt template utilised follows the model's training procedure and was the following:

<s>[INST] <<SYS>>

You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic,

```
dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature.

If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information.

<</SYS>>
```

To further refine the prompts, the prompts were changed to be more specific and included more context. An example of this was:

- Initial query: *Please give a concise answer, is Australian securities and investment commission mentioned?*
- Refined query: *Provide a summary of mentions to ASIC (Australian Securities and Investments Commission) with brief context, be concise, without quoting the original query or original/intermediate answers, only provide the final answer in a human-like response.*

Further examples are provided in appendix E.

Having unambiguous queries and experimenting to refine the prompt is essential to producing more accurate results. Prompting is also an iterative process and will require testing and subject matter experts to evaluate how well the response is.

Per page ingestion

One of the summary criteria (main prompt) was the ability to list the page numbers of where ASIC is mentioned within the document. LLMs do not inherently support the ability to include page numbers due to the way documents are chunked and converted into nodes. When the chatbot was queried to provide a summary of mentions to ASIC and explicitly asked to provide page numbers in its response, the chatbot did not provide any page number references in their response (see appendix A).

To address this issue, each page within the document was ingested separately with its own individual index. The page number of the document was also stored within the metadata during the ingestion process. After this was implemented, when the chatbot was asked the exact same question (to provide a summary of mentions to ASIC and asked explicitly to provide page numbers), it responded with two accurate page numbers and mentions to ASIC (see appendix A).

However, due to the short nature of the PoC, this optimisation feature was not implemented for the final testing phase.

ASIC Assessment of optimisation output

Following the optimisation steps, AWS then reviewed and collated the LLM responses in an Excel document (which formed the testing material for this phase). ASIC staff were provided with the Excel document which contained several iterations for each of the model prompts (i.e. references to ASIC) alongside the AI responses to each prompt iteration. The number of prompt iterations and AI responses ranged from two to ten for each of the main prompts listed at 4.1.7.

Three ASIC staff undertook the assessment and validation of the AI generated output for one submission each. For each main prompt (i.e. references to ASIC) the assessors read the AI response for each prompt iteration (between 2-10) and verified the response against the original submission (‘golden answer’). For each AI response, a ‘pass’ or ‘fail’ was recorded. In addition, each AI response in that main prompt set was ranked from highest to lowest (e.g. for 10 prompts iterations, the assessor ranked the responses from 1 (best quality response) to 10 (worst quality response). Qualitative comments on the reasons for the pass/fail and rankings was included in the assessment provided back to ASW. For example, the comments included notes on what information was inaccurate, hallucinations, missed information etc.

AWS used the assessment results and qualitative comments to further refine the LLM in preparation for the Phase 3 Final assessment.

PHASE 3 – FINAL ASSESSMENT

The Phase 3. final assessment methodology was designed by ASIC to assess the performance of the AI generated summaries using the assessment rubric and to see how this compared to summaries written by humans (ASIC staff).

For this phase, ASIC selected a sample of five additional submissions to the PJC inquiry which had been summarised by human summarisers in the Preparation/set-up stage. The sample of submissions varied in submission length, format/type and organisation who had made the submission and a mix of st, ff levels who had completed the human summary of that submission.

In this phase, the following test execution steps were performed by AWS:

1. AWS pre-loaded the dataset containing the selected submissions.
2. AWS ran a series of prompts against the LLM, for each submission. The prompts used were the refined 'optimised' prompts at the end of Phase 2 Optimisation, based on the main prompts listed at 4.1.7. Each submission was given the same prompts.
3. Each set of LLM responses were then combined into a single cohesive summary. One summary per submission.
4. These summaries were provided to ASIC as separate Word documents. The summaries were not edited from the AI generated output, except for the addition of sub-headings to reflect each main prompt and white space to improve summary readability.

ASIC engaged five business representatives (EL2 level staff across two business teams) to assess both the human and AI generated summaries. Each assessor was assigned one submission to read and rate the two associated summaries - labelled A and B. The assessors were not told that AI was involved at all – it was blind testing. The assessors scored each summary on the ASIC designed assessment rubric (see Appendix F), with each summary getting a score and qualitative feedback. Assessors were not asked to compare summary A against summary B.

DEBRIEF WITH ASIC ASSESSORS

Following the assessment ratings, a debrief session was held with the five ASIC assessors to hear more about their findings and why they gave the ratings they did. Prior to the debrief session, the assessors were informed of the true nature of the task and the background of the PoC. They were also advised which of the summaries they had assessed was AI generated and asked to focus on that in the debrief session. That was the first time they were made aware of the AI generated summary. Three of the five assessors noted that they suspected this was an AI trial.

5 Results and Findings

Phase 1 – Model Selection Results

The Phase 1 Model selection assessment rubric scores for the three submissions and for each model (labelled as model A, B, C) were provided to AWS, as shown in Table 1 below. A comprehensive qualitative analysis of the debrief session and rubric comments was also undertaken by ASIC and provided to AWS. The main issues observed by the ASIC assessors across the three models included: inaccurate or incorrect information, incorrect page references, missing information, repeated information within the same prompt response, nuance or context not picked up. See Appendix C for a summary of key issues observed by ASIC assessors for each model.

Criteria/Rating per submission	Model A	Model B	Model C
Coherency/Consistency	1	0	2
References to ASIC	0	0	0
Identifies recommendations on how conflicts of interest should be regulated	2	0	3
References to more regulation of auditors/consultants	1	0	3
Length	0	0	3
Document A Total (out of 15)	**4**	**0**	**11**
Coherency/Consistency	1	0	2
References to ASIC	1	0	1
Identifies recommendations on how conflicts of interest should be regulated	2	0	2
References to more regulation of auditors/consultants	1	1	2
Length	2	1	2
Document B Total (out of 15)	**7**	**2**	**9**
Coherency/Consistency	1	0	2
References to ASIC	1	0	0
Identifies recommendations on how conflicts of interest should be regulated	1	1	2
References to more regulation of auditors/consultants	0	0	1
Length	2	3	2
Document C Total (out of 15)	**5**	**4**	**7**
Total of documents A, B, C (out of 45)	16	6	27
Percentage of total	36%	13%	60%

Table 1 -Assessment Rubric Scores for Model A, B, C

A model selection session was held between ASIC and AWS, to review the results and select one LLM to take forward to the Optimisation Phase. In addition to the assessment rubric scores and qualitative analysis, a number of additional technical factors were presented by AWS for consideration in the decision, as shown in Table 2 below. ASIC in consultation with AWS unanimously chose to proceed with Model C, having been revealed to ASIC as Llama2-70B.

Anonymised Model Name	Model A	Model B	Model C
Model name	Mistral-7B	MistalLite	Llama2-70B
Licence	Apache 2.0	Apache 2.0	Meta / Custom
Commercial use	Yes	Yes	Yes
Base model	Mistral-7B	Mistral-7B	Llama2
MMLU score[3]	64.16	50.93	69.83
Average[4]	60.97	51.45	67.87
Context length[5]	8k	32k	4k
Model size[6]	7B	7B	69B
Instance size[7]	ml.g5.2xlarge	ml.g5.2xlarge	ml.g5.48xlarge
Price / week	$78.79	$78.79	$1,058.88

Table 2 - Additional LLM Factors used in model selection

PHASE 2 – OPTIMISATION RESULTS

The results of the Phase 2 Optimisation assessment from ASIC staff are provided in the Table 3 below. The results varied by submission and by the main prompts, but generally there was a higher number of 'fails' than 'passes' in ASIC staff's assessment of the AI responses at this point in the optimisation phase.

Main prompt set	Submission: Tax Justice	Submission: Accounting Standards Board	Submission: Accenture
References to ASIC	10 prompt iterations: Fail = 6 responses Pass = 4 responses	3 prompt iterations: Fail = 3 responses Pass = 0 responses	7 prompt iterations: Fail = 3 responses Pass = 4 responses
Conflicts of interest	3 prompt iterations: Fail = 2 responses Pass = 1 response	4 prompt iterations: Fail = 4 responses Pass = 0 responses	7 prompt iterations: Fail = 7 responses Pass = 0 responses
Recommendations on more regulation	2 prompt iterations: Fail = 0 responses Pass = 2 responses	2 prompt iterations: Fail = 1 response Pass = 1 response	9 prompt iterations: Fail = 6 responses Pass = 3 responses

[3] Source: Hugging Face – which evaluates models on 57 multidisciplinary tasks described in Measuring Massive Multitask Language Understanding . For all these evaluations, a higher score is a better score. AWS chose these benchmarks as they test a variety of reasoning and general knowledge across a wide variety of fields in 0-shot and few-shot settings.

Source: Hugging Face – which evaluate m on 7 key benchmarks using the Eleuther AI Language Model Evaluation Harness, a unified framework to test generative language models on a large number of different evaluation tasks including: AI2 Reasoning Challenge (25-shot) - HellaSwag (10-shot), MMLU (5-shot), TruthfulQA (0-shot), Winogrande (5-shot), GSM8k (5-shot).

[5] The maximum number of tokens (letters, spaces, punctuation etc.) the LLM can accept in a prompt. Higher numbers are better as they give the model more context (e.g. pages of a document to summarise) when asking it a question.

[6] The number of parameters in a model. Smaller models (such as 3 billion (3B) or 7 billion (7B) parameter models) *generally* perform worse than their larger counterparts but can be cheaper to run, and vice versa for larger models.

[7] The size of the AWS computing resource required to run the model.

Table 3 - Phase 2 Optimisation results

Phase 3 – Final Assessment Results

Phase 3 Final assessment rubric scores for the sample of five submissions and for each AI summary and human summary were provided to AWS, as shown in Table 4 and Table 5 below. A comprehensive qualitative analysis of the debrief session and rubric comments was also undertaken by ASIC and provided to AWS (see Appendix C).

The assessment rubric scores for each criteria, aggregated for all five submission were:

Criteria/Rating	AI generated summary scores - aggregated all 5 submissions	Human summary scores - aggregated all 5 submissions
Coherency/Consistency	10	12
References to ASIC	5	15
Identifies recommendations on how conflicts of interest should be regulated	5	8
References to more regulation of auditors/consultants	6	11
Length	9	15
Total (out of 75)	**35**	**61**
Percentage of total	47%	81%

Table 4 - ASIC Scoring Summary by Assessment Criteria. The maximum score per summary was 15.

The assessment rubric scores for each submission were:

Submission	AI generated summary scores (Total)	Human summary scores (Total)
Institute of Public Accountants	6	15
KPMG Australia	8	9
ATO	8	15
Dr Kelli Larson	5	10
The Australia Institute	8	12
Total of documents (out of 75)	**35**	**61**
Percentage of total	47%	81%

Table 5 - ASIC Scoring Summary by Submission. The maximum score per summary was 15.

The results achieved were not perfect (scoring approx. ~57% compared to human subject matter experts), with the human generated summaries scoring higher on every metric compared to the AI generated summaries. One AI summary scored very close to the human summary (total score of 8 and 9 respectively).

5.1 Key Themes from the Qualitative Analysis of the Debrief Session and Rubric Comments

Key themes were drawn from quotes and examples from the ASIC debrief session and comments recorded on the assessment rubrics by ASIC assessors. For the full qualitative analysis see Appendix C. The key themes included:

- Limited ability to pick up the nuance/context required to analyse submissions:

Prompt: Overall summary [8]

"…it didn't pick up the key issue in a nuanced way. I would have found it difficult to even use an output to craft a summary, I would just go back to original [submission]." **(ASIC assessor)**

Prompt: References to ASIC

"The submission identified references to ASIC but it was wordy and pointless – just repeating what was in the submission." **(ASIC assessor)**

- Included incorrect information in summaries:

  Prompt: References to ASIC

  "Included analysis which did not come from the document and does not serve the purpose. [Whereas] the human summary just said no references to ASIC." **(ASIC assessor)**

  Prompt: Recommendations on how conflicts of interest should be managed

  "Inaccurately raised legal professional privilege as a 'conflicts of interest' issue and repeated those considerations as references to more regulation of auditors/consultants." **(ASIC assessor)**

- Missed relevant information in summaries:

  Prompt: References to ASIC

  "Missed a lot of the commentary that was about ASIC (e.g. p4 content under the heading 'corporate regulator')." **(ASIC assessor)**

  Prompt: References to more regulation of auditors/consultants

  Example: For one assessor the AI missed where the submission had referred to external references which had recommendations in them.

- Missed the central point of submission:

  Prompt: Overall summary [9]

  "The summary does not highlight [FIRM]'s central point…" **(ASIC assessor)**

  General comment

  "I would have expected summary to focus on 11 key points [outlined in submission], but didn't see that level of detail." **(ASIC assessor)**

- Focused on less relevant information (giving minor points prominence):

  Prompt: Overall summary [10]

  "Made strange choices about what to highlight." **(ASIC assessor)**

  "Overall summary placed unnecessary emphasis on one minor recommendation around government procurement processes by opening with information on this, even though this recommendation was not the focus of either the inquiry or the [FIRM]'s submission." **(ASIC assessor)**

- Used irrelevant information from submission:

[8] Note that humans were not specifically asked to provide an opening summary of submissions.

[9] Note that humans were not specifically asked to provide an opening summary of submissions.

[10] Note that humans were not specifically asked to provide an opening summary of submissions.

<u>Prompt: References to ASIC</u>

"A lot of extraneous information under the 'references to ASIC' subheading that is not about ASIC (directly or indirectly)." **(ASIC assessor)**

<u>Prompt: References to more regulation of auditors/consultants</u>

Example: AI picked up the content in attachment so included irrelevant information in listing recommendations (i.e. recommendations not from the submission itself). The assessor noted *"this is not accurate and may cause misunderstanding."* **(ASIC assessor)**

Other relevant themes included:

- AI summaries were waffly, wordy
- The AI summaries did not include references/page numbers (as per prompt)
- Lacked formatting (i.e. presentation details which make it easier to read like short sentences, headings, tables)
- Repetitive[11]
- ASIC Assessors had to refer back to source material to confirm AI summary details
- AI was weakest on summarising ASIC references and recommendations around conflicts of interest
- AI summaries were sufficient in length but lacked the structure of the original submissions.
- Assessors generally agreed that the AI outputs could potentially create more work if used (in current state), due to the need to fact check outputs, or because the original source material actually presented information better.

6 DISCUSSION

The Phase 3 Final assessment overall scores showed that out of a maximum of 75 points, the human summaries scored 61 (81%) and the Gen AI summaries scored 35 (47%). Although the Gen AI summaries scored lower overall compared to the human summaries, the results should be interpreted within the limitations of this PoC (see Limitations section below).

One area where the LLMs performed poorly was in finding references to ASIC in the submission documents. Finding references in larger documents is a notoriously hard task for LLMs due to context window limitations and embedding strategies. Page references are not traditionally stored in the embedding models as the contents of PDF documents are ingested as plain text. To achieve better accuracy with this issue, substantial progress was made by splitting documents into pages and treating pages as chunks with associated metadata. There was limited time available in the PoC to further refine this approach, however these optimisation techniques showed promising results.

Another significant issue was the LLMs limited ability to analyse and summarise complex content requiring a deep understanding of context, subtle nuances, or implicit meaning. This led to summaries that missed deeper implications or oversimplified concepts within the original submissions. The finding emphasises the importance of a critical human eye which can 'read between the lines' and not take information at face value. This finding also supports the view that GenAI should be positioned as a tool to augment and not replace human tasks.

The LLM was also challenged with identifying key points in the submission documents. Items were missed or misinterpreted, and information was introduced from appendixes and presented as central to a submission. We discovered that some documents were easier to prepare and process

[11] AWS noted repetition between sections would have been due to different prompts being used each time.

than others, for example documents containing information on tables or graphs can present a challenge. The LLM was additionally observed to struggle with subjective or opinion-based content, which potentially introduced unintended consequences.

The PoC Phase 1 results supported the industry view that larger models tend to produce better results. However, the PoC demonstrated that model size was not the only determinant factor for output quality. The technology solution built for the PoC was relatively easy to implement and run. However, it does require very specific knowledge and skills, in particular optimisation skills which include ‘prompt engineering’, document preparation, and Gen AI solution configuration. A combination of these factors was required to produce the best results.

Interestingly, the PoC highlighted that a good optimisation process outweighs the model selection. In simpler terms, we proved it is more beneficial to adjust *how* you use an LLM model than obsessing over finding the "perfect" one. We found that adequate prompt engineering, carefully crafting the questions and tasks presented to the model, is crucial for optimal results. We discovered that using specific wording, symbols, and directives improved the results.

This was a key theme arising from the PoC and supports the view that investing in expertise and training in optimisation, and building the internal capabilities to continuously experiment and improve will be essential to unlock the full potential of GenAI.

6.1 Limitations of the PoC

During the PoC there were factors that were out of the project team’s control which may have affected the outcome. Some questions also remain unanswered at the conclusion of the trial. The results of the PoC should be interpreted with the following caveats and limitations:

- **PoC timeframe**: This PoC was undertaken at a single point in time (15 January to 16 February 2024), with LLM technology available at this time – noting this is a rapidly evolving field. Also, the PoC duration was limited to a 5 week duration, with one week spent in the optimisation phase. Marked improvement was seen during this period and confidence is high, but unproven, that investing more time in this phase may yield better and more accurate results. For example, further progress was made in correctly identifying page references after the formal conclusion of the final assessment.
- **Focus on one model:** Phase 2 and Phase 3 of this PoC focused on only one model (i.e. Llama2). The results do not necessarily reflect how other models may perform.
- **Limited to AWS environment:** Because the solution was built in an AWS environment, the ASIC team did not have direct access. The process of iterative testing relies heavily on the availability of SMEs to validate the response. It was likely that an environment that allowed for near real-time collaboration would have produced further improvements.
- **Risk of hallucinations**: Text-based Gen AI can on occasion generate texts that suffer from “hallucinations”, which means that the model generated text that was grammatically correct, but on occasion factually inaccurate.
- **Variation in outputs**: Testing was difficult due to low repeatability of model outputs. Asking the same question twice may not yield the same answer.
- **Specific use case may limit transferability:** The use case selected by ASIC for the PoC was quite specific. It focused on one narrow document domain, the submissions to an external government inquiry with summaries including very specific information relevant to that inquiry. It was not possible to quantify if better results would have been produced with a different dataset or for different summary requirements (for example if used for an ASIC-led consultation).
- **Differences in prompts/instructions provided to humans and AI model**: Due to the collection of tasks required for the summary, it was necessary to alter the prompts provided

to the LLM from those provided to the human summarisers. This resulted in differences in the output style of the summaries provided for assessment.
- **No 'golden answers'**: Due to the nature of the summarisation task there were not 'golden answers' to check the AI outputs against. Outputs were assessed by people using a standardised assessment rubric. Although the rubric would remove a degree of subjective bias it was possible there were individual differences in assessment, including elements that were or were not picked up by assessors.

6.2 Lessons Learnt

Through the course of the PoC, lessons learnt in different areas were captured for consideration in future initiatives:

Project:

- The project setup, structure and governance were adequate to deliver the defined objectives.
- There was good cross-representation from Business, Data Science and other areas, which allowed strong and timely feedback loops with ASIC subject matter experts.
- The project length was adequate (given budget constraints) but it would be advisable to spend more time on optimisation activities rather than model selection to provide further learnings and improve the quality of the output.
- To a human, the request to summarise a document appears straightforward. However, the task could consist of several different actions depending on the specifics of the summarisation request. For example: answer questions, find references, impose a word limit. In the PoC the summarisation task was achieved by a series of discreet tasks. The selected LLM was found to perform strongly with some actions and less capably with others.

Environment:

- A platform-centric approach enabled rapid model interchange and development by the team.
- **The Test environment was important**: The AWS environment allowed for a secure environment to host the solution. However, access was limited to AWS staff for demos, playback, interactive sessions and code reviews. A customer environment (i.e. ASIC environment) would be advisable for a future phase.
- **Optimisation was potentially more important than model selection**: The best results were produced after optimising the LLM. We recommend that future PoCs:
    - Allow more time for the optimisation phase, which may allow for more time to refine prompts and solutions to issues (e.g. identifying page references).
    - Embrace continuous improvement and develop processes that facilitate iteration and adaptation in Gen AI deployments.
    - Plan for skilling the data science team: for example, prompt engineering, Gen AI Cloud Solutions, LLM setup, are all distinct skills required for Gen AI.
- **An environment for rapid experimentation with close to immediate feedback from subject matter experts was important**. We recommend future PoCs allow for flexibility on all aspects of the Gen AI solution, for example the ability to quickly change LLMs, or system parameters.
- **Linear progression was not always possible**: Improvement is not always linear, making quantification of the optimisation hard.

- **Prompting was key:** Different models have different prompting instructions. Additionally, "generic" prompting without specific directions or considerations resulted in lower quality output compared to specific or targeted prompting.
- **Optimisation components of the model (Temperature and Top k)**: Lower temperature and higher top k improved results in some cases.

MODEL SELECTION:

- **Model choice is use-case dependent**: The best model for a particular task will depend on the specific requirements of that task. Note that in general, larger models tend to perform better than smaller models.
- **Embedding model:** The model initially selected for the embedding task performed poorly at document retrieval.

TEST METHODOLOGY:

- **Testing approach**: Future testing could employ the final testing methodology (Phase 3) pre and post experimentation and optimisations (Phase 2) to allow for a direct 'before and after' comparison.

Appendix

Appendix A - Solution Architecture

For the PoC a custom-built platform was deployed, allowing for rapid validation and prototyping. A key feature of the platform is the component interchangeability, providing ongoing flexibility and choice in LLM selection.

The following diagram describes the principal components and their interaction.

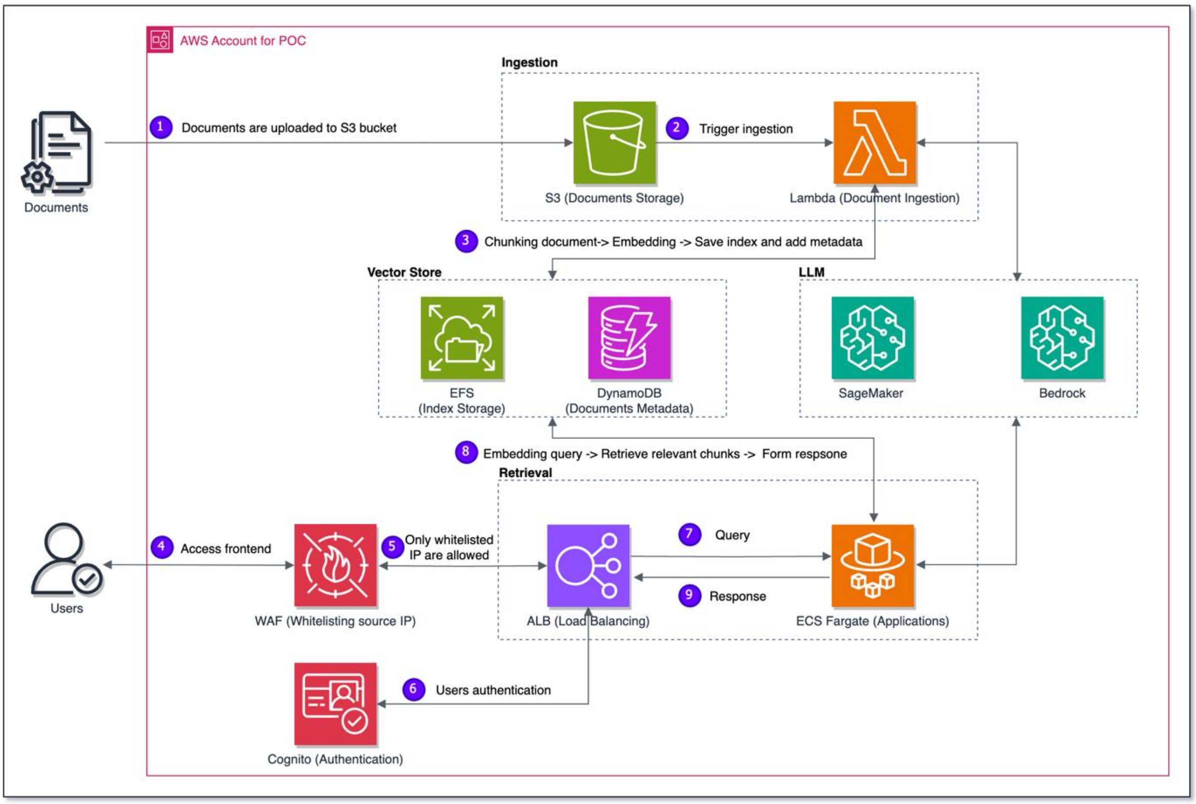


Figure 2 - High Level Architecture

The solution was deployed into a separate account (segregated environment) solely for the project. It is recommended to use account-based logical isolation to manage resources and permissions separately from production environments.

Documents are uploaded to Amazon S3 which is used for its scalability, data availability, security, and performance. Users upload documents to an S3 bucket, which acts as the initial storage location.

An event trigger is configured to automatically start the ingestion process whenever new documents are uploaded to the S3 bucket. This is essential for automation and real-time processing. AWS Lambda is a serverless compute service that runs code in response to triggers. It is used here to process the documents as soon as they are uploaded, which may include parsing, transformation, and initial analysis.

Within the deployed solution, ingestion can be also handled locally by the chatbot. The Lambda function provides chunking functionality (breaking the documents into smaller pieces), processing them to create embeddings (vector representations of the document data), and then saving these embeddings along with metadata to an index for later retrieval. This is crucial for enabling efficient search and retrieval of document content.

Amazon Elastic File System (EFS) is used to store indexes. It provides a simple, scalable file storage for use with AWS Cloud services and on-premises resources.

Amazon Cognito provides user identity and data synchronisation, enabling secure access for users to the front-end. It is used for authenticating users and managing user-specific access to resources.

AWS WAF (Web Application Firewall) is used to control the traffic that can access the front-end. Only whitelisted IP addresses are allowed, which enhances security by preventing unauthorized access.

The Application Load Balancer automatically distributes incoming application traffic across multiple targets, such as Amazon EC2 instances, containers, and IP addresses. It ensures high availability and fault tolerance for the querying interface.

Amazon ECS (Elastic Container Service) with AWS Fargate is used to run containers without managing servers or clusters. This simplifies the deployment, scaling, and management of the chat bot application that acts as a middleman between the end users and the LLMs.

Front-End – Interactive Document Chat-Bot

The document chat bot provides an interactive interface allowing a wide range of variation across all the features deployed – selecting model, parameters and output modes, upload documents for Retrieval-Augmented Generation (RAG) based responses.

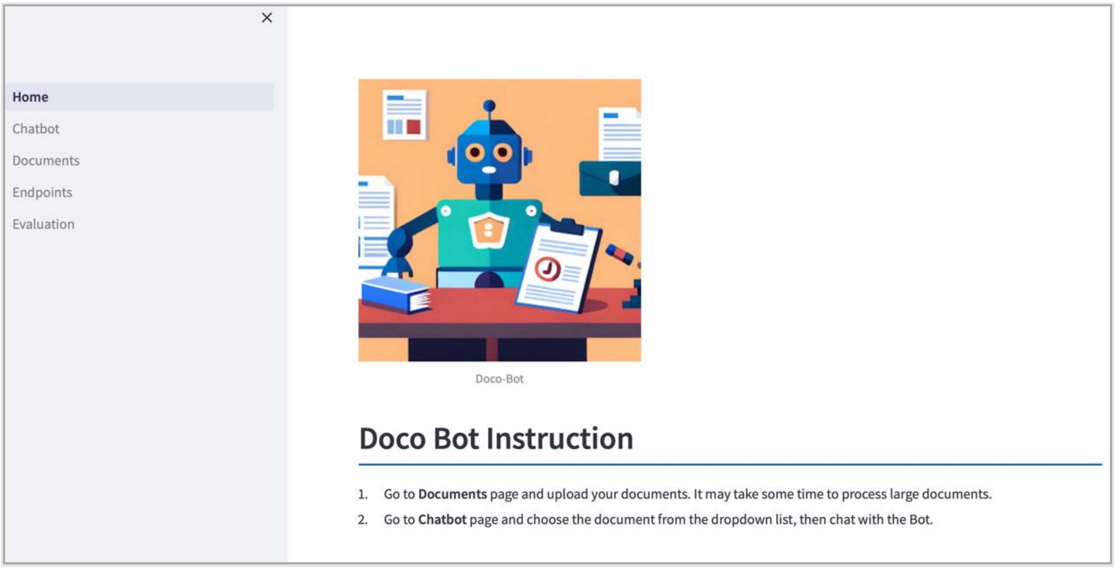


Figure 3 - Chatbot home screen

The first step is to upload a file from the local disk:

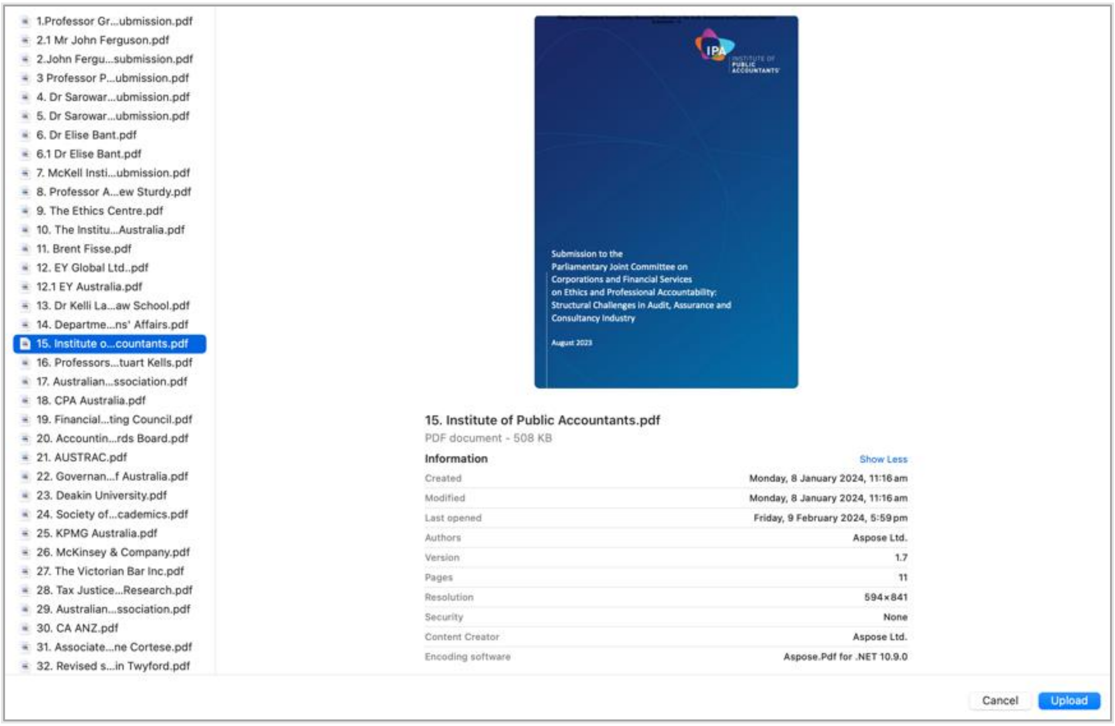


Figure 4 - Locate files

Multiple files can be uploaded at a time:

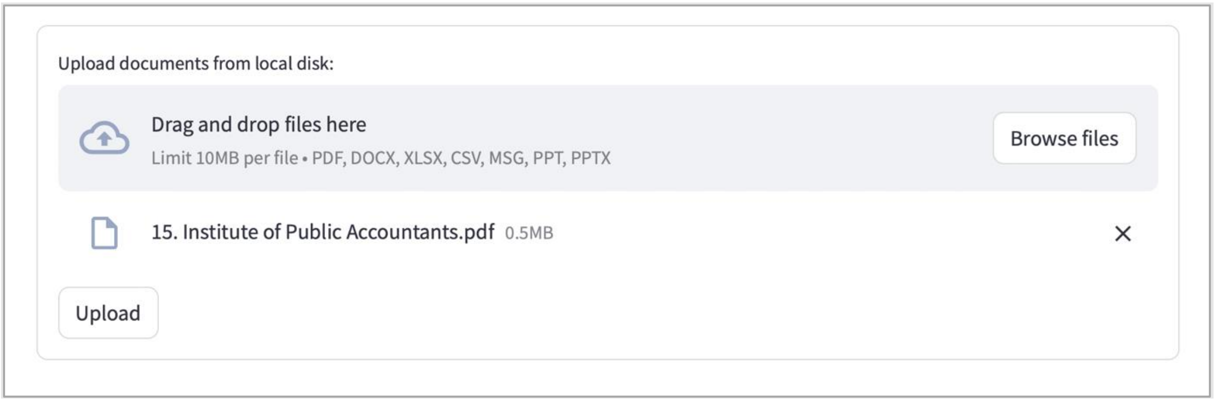


Figure 5 - Upload files

Chunking is performed upon document upload is complete. Time to perform chucking varies depending on the number and size of files.

Figure 6 - File upload in progress

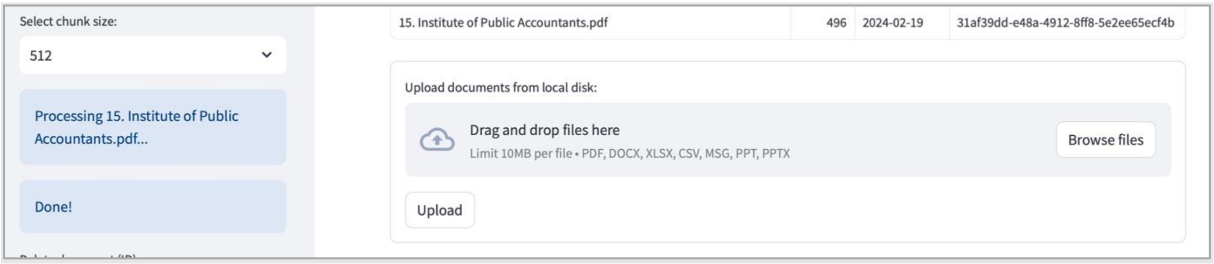
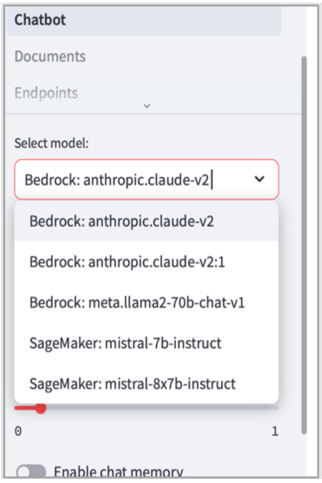


Figure 7 - File upload complete

Once the files are available in the vector store, interactive querying can be done. First, models can be selected (the list is configurable from the settings), then output mode and parameters can be changed. These settings are available for adjustments at all times and will apply to the next query.



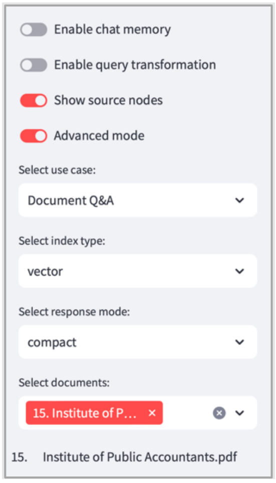


Figure 8 - Select model *Figure 9 - Adjust settings*

The query can be typed into the chat input field:

Figure 10 - Query prompt

The request is sent to the LLM and it will generate a response. Response time varies depending on the query complexity and settings used.

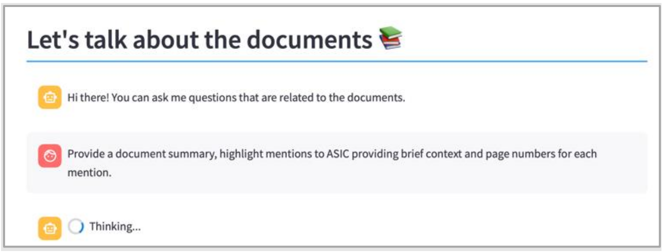


Figure 11 - Response generation in progress

Here is an example of parameters, output mode and the querying process.

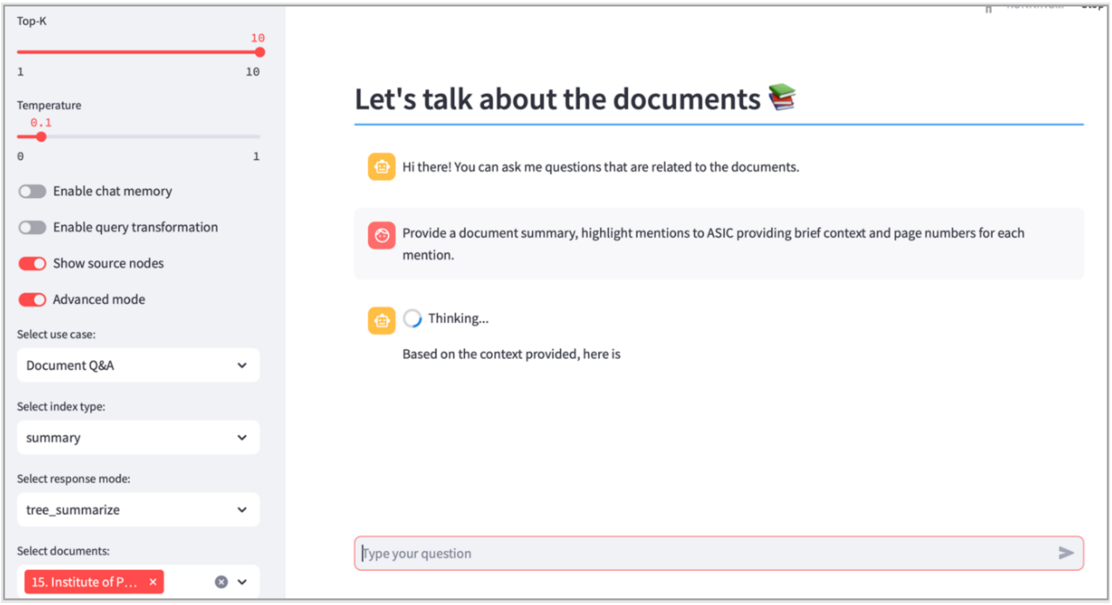


Figure 12 - Home screen with settings displayed

The response will appear right under the query and further queries can be sent interactively.

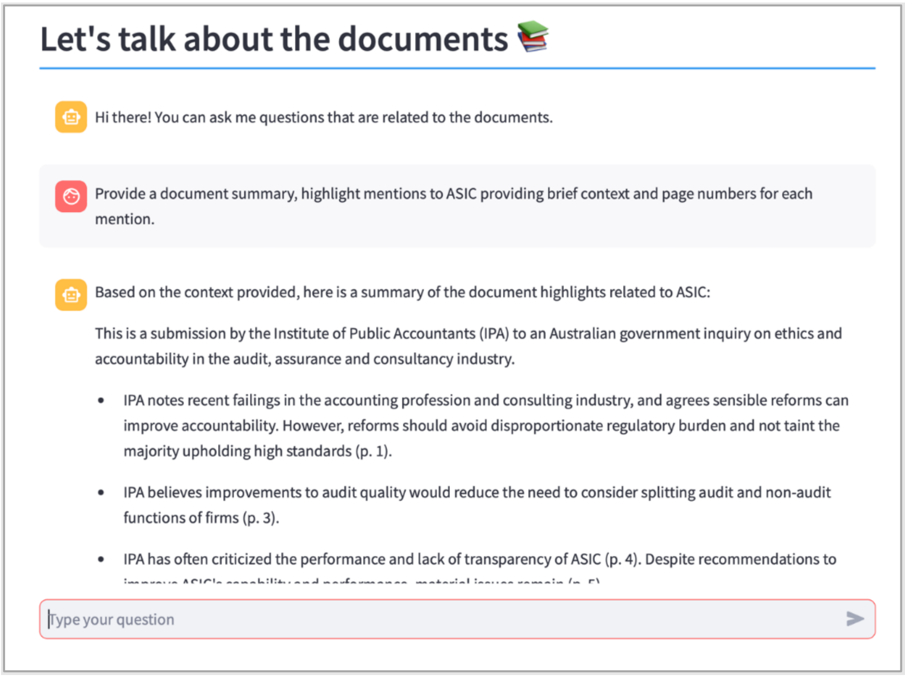


Figure 13 - Home screen showing prompt and response

APPENDIX B - PHASE 1. MODEL SELECTION: ANALYSIS OF THE DEBRIEF SESSION AND SUMMARY OF KEY ISSUES OBSERVED BY ASSESSORS FOR EACH MODEL

MODEL A summaries

Inaccurate/incorrect information

- Output for the prompt 'Summary of recommendations' had many inaccuracies which was found in all 3 submission summaries. Example from Prof Allan Fels summary:
    - *The document is a submission to the Ethics and Professional Accountability: Structural Challenges in the Audit, Assurance and Consultancy Industry inquiry by Professor Allan Fels AO. The submission discusses the use of market design to improve the provision of governmental services, specifically school transportation for disadvantaged students. The problem is that the transportation is subsidised but the service quality is poor, with children on a bus for four hours a day and arriving late to school and home. The submission outlines the theoretical and empirical work on market design and its potential to improve the provision of governmental services. The submission also discusses the role of concentration in the economy and its impact on productivity, as well as the wage price spiral and its potential to drive inflation. The submission concludes that market design can be a useful tool for improving the provision of governmental services, but that it is important to carefully consider the potential impacts of market design on workers and the broader economy.* **[all content in red inaccurate for this submission, drawing from irrelevant parts of submission – notably a Hansard transcript to another Parliamentary inquiry that was part of the Fels submission documents]**
- Incorrect page references – observed across all 3 submissions and across all prompt outputs:
    - Example from Boston Consulting Group summary for prompt 'Mentions to ASIC': ** Section E: BCG's legal status and governance under Australian law (at subsection E1). (Page 12)* **[page ref should be 20]**
    - Complexity added when the submission PDF is made up of several separate documents collated into the one submission (i.e. each doc has own page numbers). The AI models had issues with which page number to reference - the overall PDF doc pages or labelled pages in the documents, which is not always the same page reference. Regardless, the model got the page ref wrong.
        - Example from Prof Allan Fels summary for prompt 'References to more regulation': ** Page 4: "The Royal Commission on Banking and Financial Institutions (the Hayne Commission) has shown that business, especially in finance related areas (of which auditing is one), is poor at handling conflicts of interest and that culture can be driven by the interests of suppliers of services rather than customers. A major lesson of Hayne is that half measures to resolve conflicts of interest usually work poorly in practice, are gradually eroded, and poorly enforced by regulators. This is especially so when temptingly large sums of finance are involved. The only solution is full separation."* **[page ref should be p. 7 as per label of doc, or page 10 as per PDF numbers]**
- Incorrect information – attribution of quotes/content to wrong source or author:
    - Example from Prof Allan Fels summary for prompt 'Mentions to ASIC':
      ** Page 7: Mention of ASIC's recommendation to establish a new and more active audit regulator with "a new mandate, new leadership and stronger statutory powers," as well as greater funding.* **[Reference was for a recommendation that the UK's FRC be replaced by a new audit regulator. The UK government announced in March that it would accept this recommendation**…(see page 9 of PDF doc)**]**
- Incorrect information – extracts appearing as quotes, but included sub-headings:

- Example from Tax Justice Network for prompt ‘References to more regulation’: "*Involvement in scandals and crimes Against this backdrop, it is arguably unsurprising that the Big 4…”* **[Wording in red is from the sub-heading, with the content starting ‘Against this backdrop…’]**

Missing / lacking information

- No page references included for any content in ‘Summarise recommendations on how conflicts of interest should be regulated’ section for all 3 submissions.
- Missing detail in extracts, example from Prof Allan Fels summary:
  - *Page 17: Mention of ASIC's role in conducting inspections of audit firms and the results from its audit firm inspections for the period 1 January 2017 to 30 June 201* **[missing number from year]**

Repeated information within the same prompt response

- Example from Tax Justice Network summary for the prompt ‘Summarise recommendations on how conflicts of interest should be regulated’. **The 2 extracts below were listed as separate dot points, but refer to the same original content and point**.
  - *The meaningful provision of assurance can only be delivered when auditing is entirely separate and independent from all other functions. Being mindful of the organisational challenges involved in breaking up companies that currently house these services under one roof, auditing should, in the interim, be effectively and definitively ring-fenced from other functions.*
  - *As a crucial first step, this would require regulations to separate firms providing auditing services from those offering any forms of financial advice. The meaningful provision of assurance can only be delivered when this function is entirely separate and independent from all other functions. Being mindful of the organisational challenges involved in breaking up companies that currently house these services under one roof, auditing should, in the interim, be effectively and definitively ring-fenced from other functions.*
- Example from Prof Allan Fels summary to the prompt ‘Mentions to ASIC’ refers to **same content in submission (with different page ref)**:
  - *Page 5: Mention of ASIC conducting inspections of audit firms and the results from its audit firm inspections for the period 1 January 2017 to 30 June 2018.*
  - *Page 17: Mention of ASIC's role in conducting inspections of audit firms and the results from its audit firm inspections for the period 1 January 2017 to 30 June 201*

Prompts (nuance not picked-up)

- The model may not be picking up the nuance of prompts to distinguishing between ‘recommendations on how conflicts of interest should be regulated’ and ‘references to more regulation of auditors/consultants’. Although we note the relationship between the two topics and some extracts may appropriately sit within both sections.
- Sub-section headings of the summary output – reflected the wording of the prompt used, but also did not contain enough detail to reflect the original prompt. For example, the prompt ‘Recommendations on how conflicts of interest should be regulated…’ had the sub-heading ‘Summary of recommendations’ in the output but did not include the relevant words ‘conflicts of interest’. Summary doc needs to stand-alone without the prompt context - note to consider for optimisation phase.
- Language of prompt reflected in summary output - it was observed that for Prof Allan Fels’ summary the output summary used language as per the original prompt. For example ‘mentions of ASIC’ in the original prompt became part of the response e.g. ‘Page 5 Mention of ASIC….’. This was not seen as the format in the other 2 submissions. Note to consider for optimisation phase.

- Noted examples where the human summary included some references to ASIC which did not specifically have the word 'ASIC' in the original sentence/para, but instead referred to 'the regulator' – which the human intuited was a reference to ASIC whereas the AI did not.
    - Example from Prof Allan Fels summary in relation to the prompt 'Mentions to ASIC': the human summary noted '*Measures to resolve conflicts in the finance sector work badly in practice, are eroded over time and poorly enforced by regulators. [p.3]*'.
        - The point is sourced from the submission content which references the regulator (implied is ASIC, but not named): *The Royal Commission on Banking and Financial Institutions (the Hayne Commission) has shown that business, especially in finance related areas (of which auditing is one), is poor at handling conflicts of interest and that culture can be driven by the interests of suppliers of services rather than customers. A major lesson of Hayne is that half measures to resolve conflicts of interests usually work poorly in practice, are gradually eroded, and poorly enforced by regulators. This is especially so when temptingly large sums of finance are involved. The only solution is full separation;*

<u>Readability</u>

- Note for optimisation phase – the summary could benefit from brief subheadings or key words to break up text, particularly in the 'Recommendations on conflicts of interest' part of the summary.
- Note for optimisation phase – use of quote marks where the AI has drawn specific quotes from the submissions is helpful, where appropriate and to indicate a true quote rather than paraphrase. Model A did not take a consistent approach to this – sometimes using quote marks and other times not (Example from Tax Justice Network – quote marks used for 'References to more regulation' but not used for 'Summary of recommendations' despite all extracts being direct quotes in these sections).
    - Quotes should also not include headings/sub-headings, but include author where possible.

MODEL B summaries

<u>Inaccurate/incorrect information</u>

- Output for the prompt 'Summary of recommendations' 'for the Prof Allan Fels summary was both inaccurate (drawing from irrelevant content of the submission), but also potentially hallucinated e.g. cannot find ref to the wage price spiral being popular in the 70s and being discredited since then in the original submission.
    - *The wage price spiral is a theory that says that wages and prices rise together. This theory was popular in the 1970s, but it has been discredited since then. There is no evidence that it is currently driving inflation. In fact, real wages have fallen across the board, including in unionised sectors.* **[all content in red inaccurate for this submission, drawing from irrelevant parts of submission – notably a Hansard transcript to another Parliamentary inquiry that was part of the Fels submission documents]**
- Incorrect information – extracts not referencing ASIC:
    - Example from Tax Justice Network for the prompt 'Mentions to ASIC': *The Committee should recommend that the board of the Tax Practitioners Board no longer be the body that decides on the sanction to be placed on a tax practitioner.* **[extract is not correct and does not reference ASIC]**

<u>Missing / lacking information</u>

- No page references included for all content in the 3 submission summaries, for all prompts. Example from Tax Justice Network for the prompt 'Summarise recommendations on how conflicts of interest should be regulated' section:
    - *The Committee should recommend that auditing and consultancy firms be unable to make political donations. Further, any auditing or consulting firm that has made a political donation in the last year should not be able to obtain a Commonwealth Government contract.*
- Output for the prompt 'Document summary' - "*None*" was the response provided for both the Tax Justice Network and the Boston Consulting Group summary output.
- Output for 'Mentions to ASIC' - the single reference to ASIC in both the Tax Justice Network and the Boston Consulting Group submission were not picked up by the model.
    - For example, in the Boston Consulting group summary the response was "*There are no mentions of ASIC in the submission.*" which is not correct.
- Output from the AI provided a number only (which may/may not be accurate) for several prompts i.e. 'References to more regulation' and 'Mentions to ASIC', but with no further content or context:
    - Example from the Tax Justice Network: the response was "*6*" for the prompt 'References to more regulation' but did not list the actual references with context and page numbers as per the prompt.

<u>Prompts (nuance not picked-up)</u>

- Sub-section headings of the summary output – reflected the wording of the prompt used, but also did not contain enough detail to reflect the original prompt. For example, the prompt 'Recommendations on how conflicts of interest should be regulated…' had the sub-heading 'Summary of recommendations' in the output but did not include the relevant words 'conflicts of interest'. Summary doc needs to stand-alone without the prompt context - note to consider for optimisation phase.

MODEL C summaries

<u>Inaccurate/incorrect information</u>

- Output for 'Summary of recommendations' content of summaries had many inaccuracies which was found in all 3 submissions. Example from Boston Consulting Group summary, which identified the wrong inquiry:
    - *The document is a submission to the Australian Government's Senate Select Committee on Jobs for the Future in the context of Artificial Intelligence. The submission is from Boston Consulting Group (BCG), a global management consulting firm*. **[incorrect inquiry referenced in red]**
- Incorrect page references – observed across all 3 submissions and across all prompt outputs:
    - Example from Boston Consulting Group summary for the prompt 'Summary of recommendations on conflicts of interest': *1. Restrictions on suppliers providing both labour-hire services and management advisory services to the Australian Government (Submission 56, page 9).*
    - Example from Prof Allan Fels for prompt 'References to more regulation': *Page 4: The author suggests that half measures to resolve conflicts of interests usually work poorly in practice, are gradually eroded, and poorly enforced by regulators, and that the only solution is full separation*.
- Incorrect information – extracts not referencing ASIC:

- Example from Prof Allan Fels summary for prompt 'Mentions to ASIC': *The funding equation would obviously depend on the response. The UK move to create a new and more active audit regulator is a case in point. (Page 6)*. **[not a reference to ASIC]**

Missing / lacking information

- Output for 'Mentions to ASIC' – the single reference to ASIC in both the Tax Justice Network and the Boston Consulting Group submission was not picked up by the model.
    - For example, in the Tax Justice Network summary the response was: *There is no mention of ASIC (Australian Securities and Investments Commission) in the provided sources*. **[Not correct]**.

Prompts (nuance not picked-up)

- The model may not be picking up the nuance of prompts to distinguishing between 'recommendations on how conflicts of interest should be regulated', and 'references to more regulation of auditors/consultants'. Although we note the relationship between the 2 topics and (where relevant) some extracts may appropriately sit within both sections.
- Sub-section headings of the summary output – reflected the wording of the prompt used, but also did not contain enough detail to reflect the original prompt. For example, the prompt 'Recommendations on how conflicts of interest should be regulated…' had the sub-heading 'Summary of recommendations' in the output but did not include the relevant words 'conflicts of interest'. Summary doc needs to stand-alone without the prompt context - note to consider for optimisation phase.

Appendix C - Phase 3 Final Assessment: Qualitative Analysis of the Debrief Session and Assessment Rubric Comments

Theme	**Relevant prompt:** **(1) Ref to ASIC** **(2) Conflicts of interest** **(3) Recommendations** **(4) Opening summary*** **(5) General comments**	**Comments and/or examples drawn from debrief session and assessment rubric comments**
Could not pick up nuance / context	(4) Opening summary (1) Ref to ASIC (2) Conflicts of interest (3) Recommendations (5) General comments	**(4)** - *"…it didn't pick up the key issue in a nuanced way. I would have found it difficult to even use an output to craft a summary, I would just go back to original [submission]"* - *"missed nuance, context and reading between the lines"* - *"[if a person had written it] I would go back to them and get them to rewrite it"* - Example: Some significant contextual information is missing…. e.g. does not explain that the recommendation is being put forward as an alternative to requiring professional service firms to separate their audit and non-audit functions, which is a key issue for consideration in this inquiry. **(1)** - *"The submission identified references to ASIC but it was wordy and pointless – just repeating what was in the submission."* **(2)** - Example: One assessor noted that conflicts of interest is a difficult concept even for humans and that they would be surprised if AI understood the nuances because you need to understand a lot of different factors to identify - *"It did raise the ethical frameworks identified by [FIRM] but did not contextualise why it was important in the consideration of conflicts"* **(3)** - *"Very few actual recommendations identified and explained under this subheading. Instead, there are a lot of general statements about reform or regulation, e.g. 'The [FIRM] has called for more regulation'. "* **(5)** - *"AI didn't pick out right incorrect information but the information it did pick up was out of context which then went against the thrust if the general position of the submission"*
Included incorrect information	(1) Ref to ASIC (4) Opening summary (2) Conflicts of interest (3) Recommendations	**(1)** **-** *"included analysis which did not come from the document and does not serve the purpose. [Whereas] the human summary just said no references to ASIC"* **(2)**

		- *"Inaccurately raised legal professional privilege as a 'conflicts of interest' issue and repeated those considerations as references to more regulation of auditors/consultants."*
Missed relevant information	(1) Ref to ASIC (2) Conflicts of interest (3) Recommendations	**(1)** - *"Missed a lot of the commentary that was about ASIC (e.g. p4 content under the heading 'corporate regulator')."* **(2)** - Example: Missed some recommendations (e.g. disclosure of partner remuneration and reports of serious misconduct. **(3)** - Example: For one assessor the AI missed where the submission had referred to external references which had recommendations in them
Missed central point	(4) Opening summary (5) General comments	**(4)** - *"The summary does not highlight [FIRM]'s central point…"* **(5)** - *"I would have expected summary to focus on 11 key points [outlined in submission], but didn't see that level of detail"*
Focused on less relevant information	(4) Opening summary	**(4)** - *"Made strange choices about what to highlight"* - *"Overall summary placed unnecessary emphasis on one minor recommendation around government procurement processes by opening with information on this, even though this recommendation was not the focus of either the inquiry or the [FIRM]'s submission."*
Used irrelevant information	(4) Opening summary (2) Ref to ASIC (3) Recommendations	**(4)** - Example: Drew on information from the attachment **(2)** - *"A lot of extraneous information under the 'references to ASIC' subheading that is not about ASIC (directly or indirectly)"* **(3)** - Example: AI picked up what [FIRM] reported doing internally as opposed to what [FIRM] recommended more broadly - Example: AI picked up the content in attachment so included irrelevant information in listing recommendations (i.e. recommendations not from the submission itself). The assessor noted *"this is not accurate and may cause misunderstanding"*
Waffly	(4) Opening summary	**(4)** - *"waffly" / "wordy"* - *"[I] couldn't look at it in a quick way. Output wouldn't be what I would want from someone I delegated this to."* - *"long sentences"*

		- *"dense"* - *"not easy to read and digest"*
Lacked formatting (i.e. presentation details which make it easier to read like short sentences, headings, tables)	(4) Opening summary (5) General comments	**(4)** - Example: lacked headings which provided information - Example: lacked visual sign-posts - Example: long sentences running into each other - Example: way it was written was confusing - *"Lots of direct quotes from the submission, but these are not presented as such"* - *"could have been structured in an easier to read way (dot points, tables, more subheadings, a more descriptive document heading)"* **(5)** - *"Wonder if the fact I had just read the submission actually made it easier to understand the AI summary than if I had read it cold"* - *"The original submission was already quite organised with headings and a bit over four pages, arguably easier to read than the two-plus page summary"*
No page references	(1) Ref to ASIC (5) General comments	**(1)** - *"Was surprised that a human would miss that basic detail"* [An assessor speaking about when they were blind to the nature of the task or use of AI] **(5)** - Example: no page references through-out - Example: No page number of references
Repetitive**	(4) Opening summary (5) General comments	*** we note AWS pointed out repetition between sections would have been due to different prompts being used each time so suggest this theme is not weighted as highly as the other identified themes* **(4)** - *"Information is unnecessarily repeated throughout the summary (e.g. the recommendation around alternative compensation is repeated four times, once under each subheading)."* **(5)** - *"It's dense and repetitive. It could have been more concise, e.g. by removing repetitive references to IPA recommendations."*

**Note that humans were not specifically asked to provide an opening summary of submissions*

Table 6 – Phase 3. Final assessment: Qualitative analysis of debrief session with ASIC assessors and assessment rubric comments

Appendix D – Per Page Ingestion Chatbot Responses

Prompt:

"Provide a summary of mentions to ASIC (Australian Securities and Investments Commission) with brief context and page numbers, be concise, without quoting the original query or original/intermediate answers, only provide the final answer in a human-like response."

Response before Implementing Per Page Ingestion:

"ASIC is mentioned as an external regulator that may receive complaints about IPA members and as a body that has governance requirements for professional accountants' ongoing membership or statutory registration. Additionally, ASIC is constrained by legislation from sharing investigation findings with IPA, and IPA must rely on self-disclosure or a complaint from the public. There is a proposed reform to improve information sharing between bodies that consider a professional accountant's conduct, which could improve the effectiveness of the overall framework and build community trust. Furthermore, ASIC's oversight of auditors, who are frequently included as defendants in legal proceedings, adds an extra layer of pressure on firms to maintain high audit quality and adhere to professional standards. The IPA suggests that restricting firms from providing audit services to non-audit clients (and vice versa) would avoid a perception that commercial pressure from non-auditing revenue might impact the provision of auditing services, without depriving each service of the benefits of co-location in the one firm. The IPA also highlights the need for alternative compensation schemes, appropriate training, better monitoring and enforcement by ASIC, a requirement for auditors to report on internal control systems, and a rating system for auditors to improve audit quality and professional accountability. The IPA believes that focusing on reforms that require greater transparency from professional services firms is a better approach to improving probity standards than separating audit and non-audit functions. The IPA suggests that large professional firms could be subject to clearer and more comprehensive governance standards, such as those which apply in respect of ASX listed companies, modified as necessary to reflect the absence of a need to protect shareholders and capital markets. It would also be possible to mandate requirements for disclosure of partner remuneration and reports of serious misconduct. In IPA's view, this focus on individual responsibility would facilitate improved probity standards without prejudicing participants.
In summary, the IPA suggests that improving probity standards within professional services firms can be achieved by implementing reforms that require greater transparency, introducing clearer and more comprehensive governance standards, and mandating requirements for disclosure of partner remuneration and reports of serious misconduct. Additionally, the IPA proposes a rating system for auditors to improve audit quality and professional accountability."

After Implementing Per Page Ingestion:

"ASIC is mentioned in various contexts. It's a corporate regulator responsible for ensuring good culture and deterring unlawful conduct (Page 15). There have been suggestions for improvement in its capability and performance, including introducing a better monitoring system, improved enforcement, and a rating system for auditors to improve audit quality (Page 4). Additionally, the Institute of Public Accountants has suggested that ASIC should provide greater transparency and accountability in its allocation of funds, particularly in supervision and enforcement (Page 5). Overall, ASIC has been mentioned as a regulator with responsibilities in monitoring and sanctioning misconduct and poor performance, and there have been suggestions for improvement in its capability and performance."

Appendix E – Prompt Engineering – Examples of Prompt Refinement

Original Prompts	**Refined Prompts**
Please create a summary for the document	Provide a document summary, be concise, without quoting the original query or original/intermediate answers, only provide the final answer in a human-like response.
Provide a summary of mentions to ASIC (Australian Securities and Investments Commission) with page numbers and brief context	Provide a summary of mentions to ASIC (Australian Securities and Investments Commission) with brief context, be concise, without quoting the original query or original/intermediate answers, only provide the final answer in a human-like response.
Summarise recommendations on how conflicts of interest should be regulated (conflicts of interest where the entity has an audit business), refer to page numbers and brief context	Summarise proposed solutions to resolve conflicts of interests where the corporate entity has an audit business, with brief context, be concise, without quoting the original query or original/intermediate answers, only provide the final answer in a human-like response.
Mention all references to more regulation of auditors/consultants with page numbers and brief context.	Provide a summary of calls for more regulation of auditors/consultants with brief context, be concise, without quoting the original query or original/intermediate answers, only provide the final answer in a human-like response.

Appendix F - Assessment Rubric

For each criteria, select the rating that best reflects the summary. Indicate selection by shading the relevant cell. You can also add any additional comments or observations you would like to make about the summary in the additional comments field.				
Criteria/Rating	**Good (3 pts)**	**Fair (2 pts)**	**Poor (1 pts)**	**Deficient (0 pts)**
Coherency/Consistency	Summary was easy to read and understand. It moved well from point to point and had a clear emphasis.	Summary was basically easy to read and understand. It moved fairly well from point to point. Emphasis was clear enough but lacked focus.	Summary was not very easy to read or understand because it either lacked a clear focus, consistency, build or some combination thereof.	Summary was incoherent. It lacked focus. It was obvious the reviewer did not put enough time or effort into the piece.
References to ASIC	Identified **all or most** references to ASIC with indication of context and page reference **OR** accurately identified that no references to ASIC were contained in the underlying submission	Identified key references to ASIC with indication of context and page reference, but a **proportionally small number** were missed or incomplete, incorrect or inaccurate compared to the underlying submission	Unclear if reviewer understood this task. Identified **very few** references to ASIC with indication of context and page reference **OR** identified **some** references to ASIC but did not indicate context and page number **OR some** references to ASIC were incomplete, incorrect or inaccurate compared to the underlying submission	References to ASIC were not picked up. **OR** references to ASIC were incomplete, incorrect or inaccurate compared to the underlying submission
Identifies recommendations on how conflicts of interest should be regulated	Identified **all or most** recommendations that were mentioned including a brief indication of the context and a page reference **OR** accurately identified that no recommendations	Identified key recommendations including a brief indication of the context and a page reference, but a **proportionally small number** were missed or were incomplete, incorrect or inaccurate	Unclear if reviewer understood this task. Identified **very few** recommendations with brief indication of context and page reference **OR** identified **some** recommendations but did not include a brief indication of context and page number	Recommendations were not picked up. **OR** recommendations were incomplete, incorrect or inaccurate compared to the

	were contained in the underlying submission	compared to the underlying submission	**<u>OR</u>** identified **some** recommendations but included an overly lengthy indication of context **<u>OR</u> some** recommendations were incomplete, incorrect or inaccurate compared to the underlying submission	underlying submission	
References to more regulation of auditors/consultants – include a brief indication of the context and page reference	Identified **all or most** references to more regulation of auditors/consultants including a brief indication of the context and a page reference **<u>OR</u>** accurately identified that no references were contained in the underlying submission	Identified key references to more regulation of auditors/consultant, but a **proportionally small number** were missed or incomplete, incorrect or inaccurate compared to the underlying submission	Unclear if reviewer understood this tasks. Identified very few references to more regulation of auditors/consultants with a brief indication or context and page reference **<u>OR</u>** identified **some** references but did not include a brief indication of context and page number **<u>OR</u>** identified **some** references but included an overly lengthy indication of context **<u>OR</u> some** references were incomplete, incorrect or inaccurate compared to the underlying submission	References to more regulation of auditors/consultants were not picked up **<u>OR</u>** references were incomplete, incorrect or inaccurate compared to the underlying submission	
Length	Summary was succinct	Summary was fairly succinct	Summary was lengthy with some unnecessary content	Summary was not succinct and/or was significantly lengthy	
Additional comments:	*[qualitative free text space]*				Total Score

Table 7 - Assessment rubric designed by ASIC