Callaghan
Innovation
Te Pokapū
Auaha

GovGPT Lightweight Solution Design

Richie Atkinson

Version I 0.2

Date I 02/10/2024

Table of Contents

Introduction and Context

GovGPT is designed to be a proof-of-concept/proof-of-technology for a New Zealand Government assistant, providing information for small to medium businesses. The current scope of information is limited to this domain area, and the solution will not provide guidance or advice – simply the information with full references.

To achieve the objective set by our key stakeholders, we have partnered with Whāriki Māori Business Network to ensure our balance of information is suitable for small to medium businesses in New Zealand. We have also worked closely with an expert team from Microsoft ANZ to ensure that we are using the best-of-breed solutions available in the Microsoft ecosystem.

The key objectives of this project were:

- Showcase what a small, agile team can do
- Showcase that Government can quickly develop an AI tool with a defined use-case
- Provide information from a ring-fenced data source that is correct and consistent, using Retrieval Augmented Generation (RAG)
- Do it well, do it fast, do it cheap
- Be transparent about how we've done it
- Use low-code or no-code as much as possible
- Give New Zealanders an introduction to how they could interact with Government in the future
- Start small: Target a specific set of domain knowledge (in this case, Small Businesses) and develop a proof-of-concept / proof-of-technology
- Make it multi-lingual
- Make it conversational (since removed from objectives due to time and tool limitations)
- Make it voice capable (since removed from objectives due to time and tool limitations)

Aside from the two objectives taken out of scope, all these objectives have been met.

Solution/Architecture Overview

Current Solution & Solution Selection Process

Current Solution

The current solution is based on a modified version of Azure-Samples/azure-search-openai-demo GitHub repo. This solution provides the conversational aspects and does have the option to provide voice input and output if we do bring that into scope in the future. Please see the Consultation and Future Considerations sections for more details.

At a high level, this solution has been modified as follows:

- System prompt updated to one suitable for our use-case (see below, "The System Prompt")

- Look and feel modified to suit our use-case (customised logo, etc.)
- Removal of Dev Tools functionality
- Removal of “Ask, then Answer” functionality focusing on the “Chat with your data” functionality
- Login requirements disabled

9(2)(b)(ii) - Commercial Information

Solution Selection Process

The Microsoft AI Studio platform was selected as we had an existing relationship with Microsoft, they were able to provide proof-of-concept/proof-of-technology seed funding under our Enterprise Agreement, and the platform offers a significantly faster way to iterate through different LLMs to find one which supports our use case.

OpenAI GPT-4o was settled on partially because it is faster again to iterate OpenAI tools as Microsoft have a significant amount of low- or no-code software libraries available to make the development faster and simpler, but also because the multi-modal nature of the model makes it ideal for consuming the type of data we are indexing.

Alternatives Considered

For the back end, we investigated 9(2)(ba)(i) - Obligation of confidence, however after looking at the speed and ease of deployment within our timeframe, we have placed these into a backlog for future investigation.

For the front-end, 9(2)(ba)(i) - Obligation of confidence however we found that it presented a very sterile “ChatGPT” persona and would not meet our overall needs as a conversational companion.

As part of the iterative development process, 9(2)(ba)(i) - Obligation of confidence, and while this is functional and meets most of the objectives, it is not as customisable and is a very simplistic version of what we have ultimately settled on.

Front- and back-end systems are rapidly developing, including low- and no-code tool sets, so we will be following and adapting as the proof-of-concept proceeds.

The System Prompt

The System Prompt has gone through several iterations, and the current prompt can be discovered by asking GovGPT what the current prompt is.

While this does contravene the standard approaches to exposing System Information, we felt it was important from a trust and transparency perspective that we make the prompt available to people using the tool. In addition, the thinking process is also available to end users via the citations panel.

The initial prompt was simplistic and gave a reasonable outcome, however based on feedback from domain experts within Callaghan Innovation and the input of our Data Scientist, we have since refined this significantly.

We have made a conscious decision to not limit the assistant too much, as doing so will reduce the efficacy of it (i.e. it would likely make it less helpful/able to provide helpful outputs).

Further refinements will be possible (and likely necessary) for the duration of the proof-of-concept.

Architecture Overview

Security, Privacy and Transparency

Security and privacy have been one of the cornerstones of this project. Our CISO has been embedded in the project team since the beginning of the project and has provided advice and guidance as we have progressed.

This project has been approached with a “privacy first” lens, and while this will make the system more difficult to refine and fine-tune, it means that New Zealanders can be sure that their data is not being stored and used to refine the model. To achieve that objective, we have deliberately made the choice to not allow users to refine the answer output (as this requires recording the user’s prompt).

Upon launch, we also intend to have security and privacy messages displayed for users of the system, and for the overall terms and conditions to be available as required.

9(2)(ba)(i) - Obligation of confidence [redacted]

To ensure that no non-public data has been ingested, we have kept strict separation of corporate (Callaghan Innovation) data and data being ingested. This has been accomplished using docker containers (in the form of GitHub Dev Containers) to store any data related to this project, including the ingestion/scraping scripts and application source code, plus deployment tools.

We have also made the deliberate choice to be very transparent around this project. This includes:

- Allowing users to request the prompt from the system – while this not considered best practice, it is a key aspect of keeping the system transparent and giving more New Zealanders insight into how LLMs work
- Making the ‘thought process’ and other elements visible to users should they want to see it – this is achieved using the built in Citations tools which are included in the root project
- Making it clear to users what is and what isn’t available

Data Ingestion / Indexing

Data has been sourced directly from Government websites. A list of these websites can be found in the appendix of this document. Any data ingested and indexed is public information and no information stored behind any type of secure authentication has been indexed.

Ingestion of data has been accomplished via a variety of methods, including using python scripts to download and reformat HTML, and in some cases (more common than would be ideal) saving webpages as PDFs. This work was largely manual, and for the purposes of this proof-of-concept is not currently automated, though if we reach a critical mass of indexed sites, automation will be straightforward and required.

Full detail on how data is indexed can be found in the GitHub repo, however at a high level, the ingested data is split into 2048-token chunks and run through Azure AI search to complete semantic indexing prior to being made available to the GPT-4o deployment.

See the Future Considerations section for more information around potential avenues for improvement.

Avoiding “Hallucinations” and False Information

We have elected to use a retrieval augmented generation (RAG) for this solution. RAG uses specific indexed data to generate its responses to prompts, which leads to a much higher level of accuracy. In addition to this, we have also provided the solution with ring-fenced data, and it has been instructed to ignore any data in its base model.

Restricting the model to this data ensures that we won’t see false information presented as fact to users, however it also highlights any inaccuracies in the indexed data – presenting ‘a mirror to your data’ and reflecting it back.

Additionally, we have refined the creativity, minimum semantic re-rank score and seed based on testing, research, and discussions as follows:

- Creativity: 0.02
- Min. Semantic Re-rank Score: 1.5
- Seed: Fixed at 1000

These settings ensure that there is enough creativity allows for it to have a more “human-like” interface (such as greeting the user with “Kia ora!”) while keeping the information relevant. Fixing the seed ensures that answers are consistent and not randomly weighted in other ways.

All other settings have been left at their default OpenAI values, and the root repository used as the basis for the solution also includes some enhancements which guide the assistant to perform as a RAG solution.

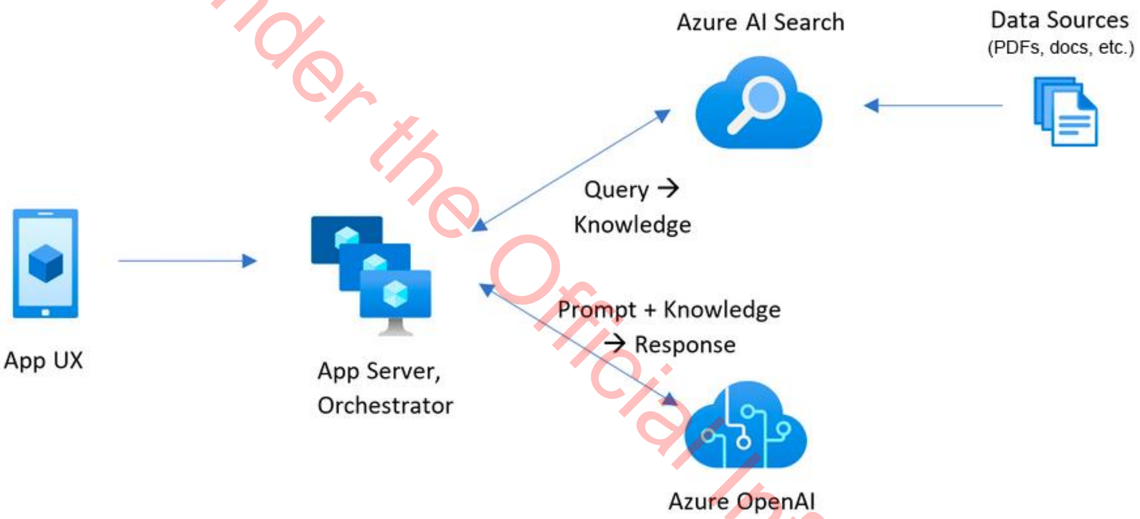
Platform

Callaghan Innovation used a separate subscription within the Microsoft Azure platform for this solution. This was partly to do with the need for secrecy in the beginning stages, but also specifically to avoid the chance of accidentally pulling in information which was not publicly available.

The solution is delivered entirely from within the Microsoft Azure platform, and includes the following services:

- Azure App Service Plans
- Azure App Services / Web Apps
- Azure Deployment Services
- Azure Blob Storage
- Azure Cognitive Services (Azure AI Search, Azure Document Intelligence)
- Azure Monitoring Services (Azure Log Analytics, Azure Application Insights)
- Azure OpenAI (9(2)(b)(ii) - Commercial Information

This diagram, from the GitHub repo, is a high-level representation of the overall architecture (which remains unchanged):



Development and Deployment

Development Process

As noted, GovGPT is based on a modified open-source project from the Azure team at Microsoft.

Given the very small size of our team, prior to the go live of the proof-of-concept, we have taken a very agile/iterative development process to reduce overhead and ensure that features for our minimum viable product were integrated and demonstrated to key stakeholders.

As we are now moving into a proof-of-concept phase, the development process will become more structured.

Development tasks will be assigned by the Technical Lead to developers and managed via Microsoft Planner. These tasks will have business and security prioritisation as well as technical complexity assigned to them.

Developers will use the built-in test tools to ensure that their changes are ready for production deployment, and following the change process detailed below, these changes will be pushed into production on a weekly cadence.

Versioning will be based loosely on the Semantic Versioning Standard, though not enough that we will reference the standard directly.

Development branches will be created for specific tasks or groups of tasks by a developer (i.e. one new branch per week, or one branch per significant feature). No development will happen directly on the main branch.

Developers will also document their changes as part of a daily Git Push – this is designed so that a) another developer or someone from our internal Digital team could pick the work up and understand it, and b) so that we maintain the transparency that is one of our core objectives.

Change Process

As with Development, the change process has been quite ad-hoc and for the proof-of-concept phase we are moving to a more structured approach.

Weekly Change Advisory meetings will be held, with business, security, and technical approvals required. Changes will need to be well documented and discussed by the developer working on the change.

Once all changes proposed to be merged for the week have been discussed and approved or declined, the approved branches will be merged to the main branch and deployed during a maintenance window using the Azure DevOps command line tool.

Changes will be communicated via the AI Activator Community, or via Callaghan Innovation's news channels if they are significant enough to warrant that (this will be a business decision).

Business Overview

Consultation with Partners

Whāriki Māori Business Network

Whāriki were brought on as a partner for this project from the very beginning and have provided key knowledge around what sources we should ingest into the solution, as well as key Te Ao Māori taonga and advice.

9(2)(b)(ii) - Commercial Information

[redacted]

Whāriki also provided their expertise and voice to our demo video as well as several grounding questions which their customers have asked in the past.

Microsoft

Microsoft ANZ have also provided partnership in the form of expert advice and guidance, and a proof-of-concept demo funding package which allowed us to get this work off the ground. It is also important to acknowledge that the open-source project this tool is based on was also provided by Microsoft.

Microsoft have also provided key advice on security and privacy for both the LLM and the application stack.

Complications

Ingestion

We found that while many Government websites are great from a human accessibility perspective, the back-end code which make this accessibility possible also renders the page very difficult for a machine to scrape and ingest.

This is predominantly where a JavaScript container is used to call streaming data, and common tools used for scraping the data files (which are largely Python-based) are unable to parse the JavaScript content.

Where we ran into this problem for the proof-of-concept, a “brute force” approach has been taken, wherein the page we wanted to scrape was printed to PDF and indexed in that format. This would not be a sustainable option for a wider base of information.

9(2)(b)(ii) - Commercial Information [redacted]

[redacted]

[redacted]

Future Considerations

In the short term

We will continue to iterate on the system, and the proof-of-concept will run for approximately 3 months so we can collect usage information (in the form of telemetry from Zoho and Azure Websites – no ‘real’ user data will be collected). Starting 1 October 2024, a set of proposed changes and enhancements will be deployed (based upon a roadmap developed by the GovGPT team) – these will be deployed in accordance with the change and release plan and communicated publicly via the AI Activator community.

Productionising GovGPT

As this proof-of-concept evolves, there will naturally be questions asked about how this can be productionised. These questions are likely to be technical, business, and security/privacy related, however the intent of GovGPT at this time is to test the waters and show that an AI companion is something that the New Zealand Government could do – the current objectives have a very specific scope, and we do not intend to go beyond that unless directed to.

9(2)(f)(iV) - Confidential advice Govt [redacted]

Ultimately this proof-of-concept is designed to last up to 3 months to collect telemetry and public feedback to provide our CE and ELT the relevant information required to consider the next steps and engage with their peers.