**Introduction**

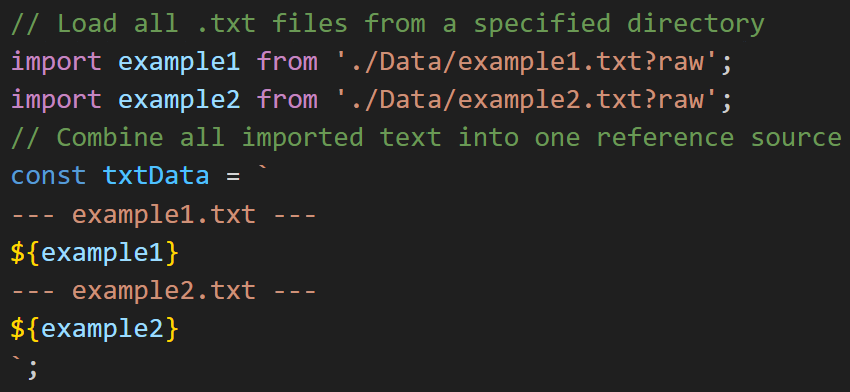
This document explains the Data Schema Model approach used in the TERAH project to provide reliable, general information about retirement planning, superannuation, and the Age Pension in Australia. The model relies on a controlled collection of pre-approved text files containing extracts from publicly available government and industry sources. These text files act as a static knowledge base that the chatbot’s language model references when generating responses. The system uses a carefully designed prompt framework to ensure the model provides factual, general guidance without personalized financial advice.

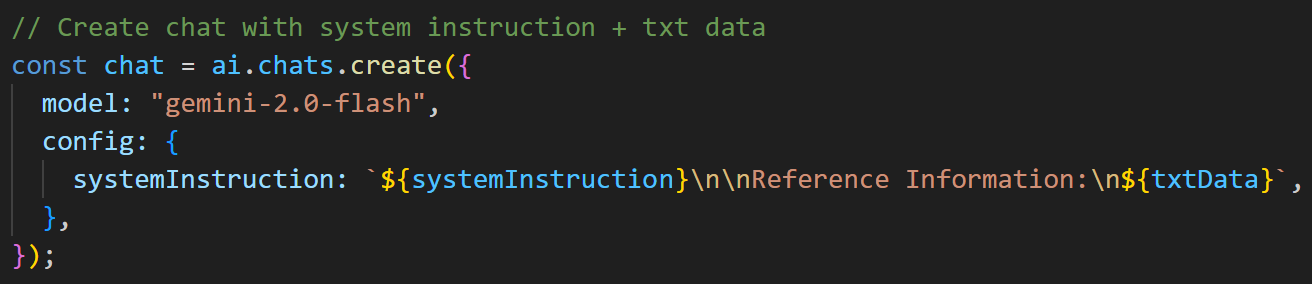
**Data Sources and Content Management**

The source texts included in the data files are manually selected and copied from official websites such as the Australian Taxation Office (ATO), Services Australia, MoneySmart, and other reputable bodies. This manual curation guarantees that the content is accurate, verifiable, and consistent with the project’s commitment to provide trustworthy information. Each text file contains only factual content and avoids any sensitive, personal, or speculative information. The files are stored in plain text format to maintain simplicity and ease of audit. They serve as static snapshots of the relevant information at a specific point in time, with updates made periodically as needed.

**Data Integration and System Instructions**

At runtime, the system imports all text files and concatenates their contents into a single reference block. This consolidated reference is embedded into the language model’s system prompt alongside a detailed instruction set that guides the chatbot’s behavior.





The instructions restrict the chatbot to providing only general information and explicitly prevent it from offering personalized advice or recommendations. For example, if a user requests advice tailored to their individual circumstances, the chatbot politely declines and directs them to consult a licensed financial adviser. The system prompt also contains specific guidelines for handling frequently asked topics, such as lost or unclaimed super, salary sacrifice, eligibility for Age Pension, and tax implications of superannuation withdrawals.

**Benefits of the Data Schema Model**

This design ensures that the chatbot’s responses are grounded in a fixed, controlled data source, while leveraging the language model’s natural language understanding and generation capabilities. By separating the reference information from the system instructions, the model maintains strict adherence to the content boundaries and project ethics. The language model interprets the combined prompt to produce answers that are concise, clear, and consistent with Australian retirement law and policy. The use of plain text files as the data schema offers several advantages. The static snapshots are easy to review and audit, enabling project stakeholders and compliance officers to verify exactly what information the chatbot references. This transparency is critical for maintaining trust in the chatbot’s outputs, especially given the sensitive nature of retirement and financial planning topics. Furthermore, the simple file format requires minimal infrastructure and complexity compared to more advanced ingestion or indexing systems, facilitating rapid prototyping and straightforward maintenance.

**Limitations and Mitigation Strategies**

However, this approach has important limitations. Because the data files are manually curated and updated, they can become outdated if source information changes and the files are not promptly refreshed. This introduces a risk that the chatbot may provide information that no longer reflects current policies or regulations. Additionally, concatenating large amounts of plain text into a single prompt can approach the token limits of the language model, potentially impacting performance or cost. The lack of structured metadata or indexing also means the system cannot easily identify or highlight the provenance of specific pieces of information within the text, making precise traceability more difficult.

To mitigate these challenges, a regular update schedule should be established to review and refresh the text files in line with official source updates. Where possible, minimal metadata such as source URLs and capture dates can be embedded as comments or headers within the text files to aid manual tracking. Maintaining a test suite of predefined queries is recommended to verify that the chatbot’s responses remain accurate and compliant after each data update. This test suite should include both positive checks for factual correctness and negative tests to ensure that the chatbot does not provide personalized advice or stray from its instructions.

**Conclusion and Future Improvements**

Overall, the Data Schema Model provides a straightforward and auditable way to ground TERAH’s knowledge base in high-quality, approved text extracts. While it requires manual maintenance and faces scalability constraints, it offers a transparent foundation on which the chatbot can reliably deliver general retirement information. The combination of static, controlled data and carefully crafted system instructions enables a balance between safety, clarity, and responsiveness appropriate for the project’s goals. Future improvements could include adding lightweight metadata to the data files for better provenance, introducing validation scripts to check file integrity and compliance before deployment, and exploring incremental update mechanisms to reduce the maintenance burden. Despite these potential enhancements, the current model serves as a robust baseline for integrating trusted reference content with advanced language model capabilities.