Advancing Data Collection Methods for Service Provision with Generative Intelligence

Gabriela Ferraro

School of Cybernetics
Australian National University

Contents

1 Introduction

The mental healthcare system is fragmented, increasingly complex, and difficult to navigate [Romero-Lopez-Alberca et al., 2021]. In this context, the development of navigation tools and health-system navigators has been implemented as a preferred approach to overcome these problems. In the past four years, the number of navigation tools available for mental health has increased from 7 to over 100 in Australia (Woods C. et al, 2024). Likewise, the role of navigators has become central to care coordination in this sector. However, the dyad of care navigator and local navigation tool has not yet been operationalised in the Australian Federal Government. Similarly, the proliferation of navigation tools has made it extremely difficult to reach provider organisations and request the data and updates needed to feed digital navigation platforms. As a result of this deadlock, major interest has been placed in the role of generative intelligence and agentic AI models to readily collect, screen, and report on the main characteristics of provider organisations and services available in local areas.

The aim of this study is to analyse the clarity and accuracy of information available on the webpages of mental health services and organisations in a region (the Australian Capital Territory) to inform the usability of web scraping and generative intelligence as a method to identify and update information on mental healthcare service provision.

Generative intelligence technologies have demonstrated impressive capabilities in performing tasks such as automated text generation, and information extraction [Li et al., 2024]. These technologies relied on Large Language Models (LLMs), which are trained on vast amounts of textual data to produce human-like responses based on input prompts.

In this study we use generative intelligence on three tasks:

- **Data gathering**: use web scraping to gather websites' content, and an LLM to extract information about the main characteristics of service provision.

- **Service coding**: use an LLM to describe services and the Description and Evaluation of Directories and ServicEs (DESDE) classification system to codify services.

- **Service discovery**: use an LLM to create queries and a search engine to find URLs to be use for subsequent web scraping.

For benchmarking, we use a gold standard consisting of URLs on mental healthcare in the ACT (May 2025) with features corresponding to the main characteristics of service provision (e.g., service name, service location, among others). Moreover, we evaluate the three aforementioned tasks in terms of their Technology Readiness Levels (TRLs). This approach allows us to assess their level of maturity, identify gaps between research and deployment, and determine the next steps required for advancement.

We found that generative intelligence has the capacity to enhance the efficiency, scalability, and consistency of identifying relevant services across jurisdictions by utilizing a hybrid approach that integrates human expertise with generative intelligence driven automation. Despite the impressive capabilities of generative intelligence, this technologies still face significant challenges related to reliability, bias, explainability, and their ethical use. Therefore, this report includes a dedicated section on the responsible use of generative intelligence. It analyses potential risks and outlines corresponding mitigation and monitoring strategies to ensure the safe and ethical use of these technologies.

2 Methodology

This section outlines the methodological approach adopted for the study. We describe the methods and workflow used to address each of the tasks introduced in the previous section, detailing how they were implemented. The section also presents the dataset employed for benchmarking, along with the evaluation metrics to assess performance and validate results.

2.1 Data Gathering

Data gathering refers to the process of collecting information about the main characteristics of service provision. In M-Chart-Phase data gathering is carried out using a Smart Survey. This requires phones calls or one-to-one meetings between service providers and M-Chart personnel. An effective approach to enhancing the collection, updating, and scalability of service provision data involves the partial automation of these processes. This can be achieved through the use of web scraping techniques to systematically retrieve and extract content from relevant websites, in combination with generative intelligence tools to interpret and extract information from the retrieved data.

Traditional web-scraping is done by extracting data from websites data using rules. It is usually done using established libraries for this purpose and relies on HTML and CSS markup such as tags and classes. Since it relies on rules, it is designed to extract specific data and might require manual updating if a website structure changes. Some of the advantages of traditional web-scraping is that it works well when the data format is consistent; making it fast and efficient in those situations. Some of the limitations of this approach include: fails easily when websites changed and struggles with JavaScript content – a popular programming language for dynamic websites.

Web-scraping powered generative intelligence use LLMs to interpret and extract data from websites, especially unstructured or semi-structured content. It works by proving the LLM model URLs and prompts with the information to be extracted. Some of the advantages of this approach include working with narrative data, handling semantic prompts (e.g., "extract the service provider name"), and it better handling of layout changes, as it does not depend only on HTML tags. Some of the limitations of this approach are: it is require more computational resources, it is usually slower, and could generate inaccurate data.

In this study, we use generative intelligence–assisted web scraping to gather data from service websites, specifically:

- `service name`
- `service provider`
- `service address`
- `target age group` (e.g., adolescents)
- `specific target group` (e.g., carers)
- `mental health disorder` (e.g., eating disorder)

Using a gold-standard corpus of service URLs, we developed a reproducible web extraction pipeline. This pipeline automates the retrieval, parsing, and processing of website content. Each page is retrieved via HTTP and parsed with BeautifulSoup[1] to generate a cleaned text representation of the website. The resulting text is submitted to an LLM model under a single-shot prompt that enumerates the target information (e.g., service name, provider, address) and requests a brief "Key description summary" offering an overview of the website's content. The model returns a record per URL with the requested fields and summary.

2.2 Service Coding

Service coding refers to the task of assigning a DESDE code to a service. The Description and Evaluation of Services and Directories for Long-Term Care (DESDE-LTC) is an international framework for standardizing the description and classification of long-term care services [Romero-Lopez-Alberca et al., 2021]. This allows for a more accurate and meaningful evaluation of services, and international service comparisons. Currently, the process of mapping services to DESDE codes is done manually by practitioners familiar with the DESDE code system.

A DESDE code consists of four segments, each of which represents a particular service characteristic:

1. Target population / age group
2. Diagnostic or problem area
3. Main Type of Care — the functional service category
4. Additional qualifiers

Table 1 illustrates the breakdown of the code AX-F00F99-R4-J. In summary, the code describes a residential (24-hour non-acute) mental health service for adults (aged 18–65) with mental or behavioural disorders, provided in a justice or forensic context.

In this study, we focus on codifying the first three segments of the code. The forth segment was not considered in this research as it usually requires detailed information about a service, and this information is rarely in service providers websites, e.g., outpatient, residential, day care, among others).

To map services to DESDE code we employed an LLM and pre-established DESDE taxonomies. The taxonomies used are `aga`, `target population`, and `mental health disorder`. These taxonomies can be found in Appendix B.

The information extracted via web scraping is paired with the established DESDE taxonomies to form schema-constrained prompts. These prompts instruct the LLM to assign each code in the taxonomy a confidence score between 0 and 1, indicating the model's certainty in the classification. When a confidence score falls below a specified threshold — in this case, 0.9 — an alert is triggered and logged. This alert signals that a code assignment is below the desired confidence level and should be reviewed by a human expert to verify the

[1] Beautiful Soup is a Python library designed for parsing HTML and XML documents `https://pypi.org/project/beautifulsoup4/`

Segment	Meaning	Explanation
AX	Adults (18–65 years old)	The service is targeted at adults.
F00–F99	Mental, behavioural, and neurodevelopmental disorders	Refers to the ICD-10 block F00–F99, indicating that the service focuses on mental health.
R4	24-hour non-acute residential care	The prefix **R** indicates a residential service, and **R4** specifically refers to non-acute, 24-hour staffed care facilities (e.g., supported housing or rehabilitation units).
J	Forensic or justice-related context	The suffix **J** denotes that the service is delivered in connection with the justice system (e.g., for offenders or people under legal supervision).

Table 1: Breakdown of DESDE code `AX-F00F99-R4-J`.

service classification. This prompt design anchors the model to a controlled vocabulary from which the model must score the available codes. This means that the LLM model can only choose codes from the pre-established taxonomies (see Appendix C for examples of these prompts).

2.3 Service Provider Discovery

Service provider discovery is the systematic process of finding and collecting online sources (URLs) that describe health services within a defined scope and region (e.g., mental health services in the ACT). This task constitute the entry point of the data collection pipeline: all subsequence stages - web scraping, extraction of service attributes, and DESDE coding - are constrained by what is initially identified as useful URLs. Consequently, the completeness and validity of downstream tasks depend on the coverage, specificity and fidelity of the discovery step.

To date, in M-Chart, this process has been carried out manually, relying on internet searches, *ad hoc* discovery methods, and individual analysis. Although this approach has yielded some success, it is time and resource intensive and may overlook opportunities due to its limited scalability and lack of systematization. Therefore, automated methods that can assist on systematization of the discovery step are highly desirable.

In this section, we explore a semi-automated method for discovering relevant service URLs. This approach has the potential to significantly enhance the long-term impact of M-Chart by enabling the continuous updating of information about already identified services and facilitating scalability across other jurisdictions by discovering new URLs.

We study the feasibility of discovering service providers URLs using an semi-automatic approach that combines (i) human-crafted seed queries, (ii) query expansion methods powered by LLMs, (iii) and automated search. In what follows, each of these steps are explained in detail.

2.3.1 Seed Queries

Seed queries are the initial, manually crafted search terms or phrases that act as starting points for broader or automated search. Thus, seed queries act as a foundation, defining the initial search intent and domain boundaries. The seed queries are later use to trigger query expansion using generative intelligence.

Guidelines to craft good seed queries focused on keeping the seed query as simple and specific as possible. The following guidelines have been taken into consideration [Manning et al., 2008]:

- Clear and focused topic
- Uses domain relevant terms
- Reflects real world queries (how people search)
- Can be varied (formal or colloquial)

As a starting point, seed queries include general search terms related to mental health in the ACT region and queries that target specific populations (e.g., adolescents) and service types (e.g., counselling).

Examples of general seed queries include:

- *mental health services in the Australian Capital Territory*
- *public mental health support in Canberra*
- *mental health helplines in the ACT*
- *government mental health programs the ACT*

Examples of seed queries targeting specific populations:

- *youth mental health services in the ACT*
- *LGBTQ+ mental health services in the ACT*
- *mental health resources for refugees in in the ACT*
- *mental health services for elderly in the ACT*

Examples of seed queries targeting types of services:

- *counselling services in the ACT*
- *psychiatric care in the ACT*
- *mental health crisis intervention in the ACT*
- *therapy options in the ACT*
- *online mental health support in the ACT*

This manual exercise resulted in a set of forty-eight seed queries (see Appendix A).

The next step, query expansion, builds upon these seeds to increase coverage, helping discover more relevant results that might not match the exact wording of the seed query.

2.3.2 Query Expansion using LLM

Query expansion is a technique used in information retrieval to improve search results by modifying or enhancing the original query [Manning et al., 2008]. This technique, often used in Internet search, is about adding related terms, synonyms, subtopics, or clarifications to the initial query so the search system better matches potential answers.

Overall, the goal of query expansion is to increase recall, thus finding more relevant results. In M-Chart, query expansion is use to increase the chances of finding potential new services in a systematic way.

Traditionally, query expansion techniques rely on dictionaries and other lexical resources to augment queries. Query expansion powered by generative intelligence mimic traditional techniques by rephrasing, expanding, and clarifying queries by prompting LLMs. The following are example of prompts for query expansion:

- *Expand this search query with related keywords, concepts, and subtopics*
- *Give alternative search queries to explore this same topic*
- *Expand this search query with semantic variations, not just synonyms*
- *What specific versions of this query could I ask for better results?*
- *Frame the query as if from a specific persona or perspective, e.g., What would a policy maker search for?*)
- *Format the output as a list of search-ready queries*

The query expansion process consists of expanding the forty-eight seed queries iteratively with an LLM, which generates ten variants per query. This yield 539 queries in total. These queries are then use to drive the URL-finding step with search loops.

Examples of expanded general queries:

- *list of mental health providers in the ACT*
- *directory of mental health services in the ACT*
- *mental health facilities open to the public in the ACT*
- *community-based mental health services in in the ACT*

Examples of expanded queries targeting specific populations:

- *youth mental health support services in the ACT*
- *LGBTQ+ mental health support groups in the ACT*
- *mental health services for older adults in the ACT*
- *refugee mental health support in the ACT*

Examples of other expanded queries:

- *telehealth mental health services in the ACT*
- *inpatient vs outpatient mental health care in the ACT*
- *group therapy for anxiety in the ACT*
- *peer support mental health programs in the ACT*
- *urgent care for mental health crises in the ACT*
- *free psychological counselling in the ACT*

2.3.3 Automated Search

Automated search refers to the process of searching the web programmatically. It allows systems to send search queries to search APIs or LLM, receive URLs as results, and then use those results for trigger web scraping. The workflow is as follows:

- List queries resulted from the query expansion procedure
- Send each query it to an LLM or search API
- Receive structured data like titles, URLs, and snippets

After processing the expanded queries, we obtained 705 unique URLs. When searching with an LLM, many prompts did not return a URL, which reflects a fundamental limitation of base LLMs: they are text generators without intrinsic access to live webpages. To address this limitation, we employed a search API, which yielded 4900 URLs - substantially larger set than the obtained through direct LLM querying.

2.4 Benchmark Dataset

To evaluate the performance of the proposed approaches, a benchmark dataset was used as a reference for comparison. This dataset includes carefully curated examples representative of service provision in the Australian Capital Territory. Details about the dataset structure, as well as the criteria applied during its collection processes can be found in [Romero-Lopez-Alberca et al., 2021].

2.5 Evaluation Metrics

To evaluate the performance and reliability of the approaches proposed in this study, a set of standard evaluation metrics was employed. The outcomes of the automatic approaches can be categorized into four possible results: true positives, true negatives, false positives, and false negatives, which serve as the basis for calculating the performance metrics used in this evaluation.

- True Positive (TP): A true positive occurs when the system correctly identifies a positive instance.
- True Negative (TN): A true negative occurs when the system correctly identifies a negative instance.

- False Positive (FP) (also known as a Type I error) occurs when the system incorrectly predicts a positive outcome for a negative instance.
- False Negative (FN): A false negative (also known as a Type II error) occurs when the system incorrectly predicts a negative outcome for a positive instance.

Based on these outcome categories, the following performance metrics can be derived.

Precision: the proportion of extracted values that are correct, defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity): the proportion of correct values that were successfully retrieved, defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Macro averages: these metrics reflects how the automatic approach performs on all classes equally, including the ones with few occurrences. Hence, it treats all classes equally, which is considered fairness across classes.

$$\text{Macro Precision} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FP_i}$$

where N is the number of classes, TP_i is the number of true positives for class i, and FP_i is the number of false positives for class i.

$$\text{Macro Recall} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i} \tag{1}$$

where, N is the total number of classes, TP_i is the number of true positives for class i, and FN_i is the number of false negatives for class i.

Macro recall calculates the arithmetic mean of recall values across all classes, treating each class equally regardless of its size.

Weighted averages: Same as macro average, but weight each class by its support (number of true examples). With this metric, common classes get more influence, but rare classes are not completely ignored. Thus, it treat frequent classes more strongly.

$$\text{Weighted Average } \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

where, $\bar{x}_w$ is the weighted average, x_i is the i-th observed value, w_i is the weight associated with x_i, and n is the total number of observations.

We also use the Levenshtein distance to quantify the differences between text strings output by the automatic system and the benchmark. This metric, also be referred to as edit distance, is a string metric for measuring the difference between two sequences. The Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. This is useful when nearly identical but no exact (e.g., street vs str.) sequence of characters are compared.

$$D_{i,j} = \begin{cases} 0, & \text{if } i = 0 \text{ and } j = 0 \\ i, & \text{if } j = 0 \text{ and } i > 0 \\ j, & \text{if } i = 0 \text{ and } j > 0 \\ \min \begin{cases} D_{i-1,j} + 1, \\ D_{i,j-1} + 1, \\ D_{i-1,j-1} + \text{cost}(s_i, t_j) \end{cases}, & \text{otherwise} \end{cases} \tag{3}$$

Example `street` and `str`

$$s = \texttt{street}, \quad t = \texttt{str}$$

To transform `street` into `str`:

- Delete the 4th character: `stre` $\rightarrow$ `str`
- Delete the 5th character: `stree` $\rightarrow$ `stre`
- Delete the 6th character: `street` $\rightarrow$ `stree`

Thus, three deletions are required.

$$D(\texttt{street}, \texttt{str}) = 3$$

Therefore, the Levenshtein distance between `street` and `str` is $\boxed{3}$.

The smaller the value, the more similar the two strings are. In our experiments, we use the normalize Levenshtein distance into a similarity score between 0 and 1.

3 Evaluation Results

This section presents the outcomes of the evaluation conducted to assess the effectiveness of the proposed approaches. The results are derived from experiments performed on the benchmark dataset, using the evaluation metrics introduced in the previous section. All experiments were conducted using OpenAI's GPT-5 LLM, accessed via the OpenAI API.[OpenAI, 2023].

3.1 Data Gathering Evaluation

We conduct a systematic comparison between the automatically extracted records and the gold standard reference dataset. This evaluation specifically examines the alignment of: `service name`, `service provider`, and `service address` of 165 records in terms of Precision and Recall, see Table 2.

For each case, the number of correctly matched instances (TP) and the number of mismatches (FP/FN) are listed. Since the evaluation set is balanced across the three fields, precision and recall are numerically equivalent in this context. This means that every mismatch counts simultaneously as a FP and a FN (since one incorrect prediction replaces one true value).

Results highlights that performance is highest for `Service name` extraction, while `Service address` remains the most challenging case.

	TP	**FP/FN**	**Precision**	**Recall/Sensitivity**
`Service name`	41	124	0.248	0.248
`Service provider`	23	142	0.139	0.139
`Service address`	9	156	0.054	0.054

Table 2: Data gathering evaluation in terms of Precision and Recall.

An error analysis indicates some differences are cause by FN instances, for example, a provider alias or acronym is used in the benchmark, while the provider full name is detected by the automatic system, e.g., *CAMHS* vs. *Child and Adolescent Mental Health Service*. While these two instances refereed to the same service provider, the automatic comparison method fails to classify them as true positives. Another source of false negatives is the difference in the level of abstraction or specificity used to represent service provider names (cf., *Canberra Health Services* vs. *ACT Eating Disorders Clinical Hub*). The automatic method extracts a more specific provider name, while the benchmark lists the broader parent organisation, leading the comparison to incorrectly flag the instance as a mismatch. A similar examples are: *Marymead CatholicCare* vs. *Marymead CatholicCare Canberra & Goulburn*; *Headspace Braddon - Low Intensity Stream* vs. *Headspace Canberra*; *Australian National University* vs *ANU counselling*.

Another way to capture the degree of similarity between records, is to use the Levenshtein distance. Before comparison, strings were normalised to case-fold, trim, and unify dashes, see Table 3.

Metric	`Service name`	`Service provider`	`Service address`
Distance mean	14.93	14.99	29.42
Distance median	14	15	26.0
Similarity mean	0.54	0.43	0.44
Similarity median	0.5	0.40	0.35

Table 3: Levenshtein distance between benchmark and automatically extracted information.

`Service name` shows the strongest alignment of the three, but still only middling: mean distance $\approx$ 14.93 and mean/median similarity 0.54/0.5, which means that about half the pairs are <0.5 similarity. `Service provider` performance is weaker, with a distance of $\approx$ 15.0 and similarity 0.43/0.40. This suggests frequent token or alias differences between sources. To improve scores, further steps include using token-set comparators for names and providers (this method ignores word order).

Lastly, we compared the `service address` in the benchmark dataset with the automatic extraction. We found that 41.82% (69/165) of the times the LLM is not able to retrieved an address from the scrapped information. In these cases, the LLM returns a either "not provided" and "N/A". These cases were excluded from the similarity calculus. Hence, only addresses were considered for the similarity calculation. `Service address` match with a distance $\approx$ 30 and with similarity 0.44/0.35, which indicates substantial divergence between address strings. Some of these differences reflect formatting and abbreviation variants (e.g., *Crc* vs *Crescent*, unit and building names). A way to overcome some of the mentioned limitations is to geocode addresses before comparison.

Overall, the normalised similarities are below a typical strong match cut-off (e.g., 0.80), with addresses driving most of the mismatch.

3.2 DESDE Code Generation Evaluation

Across 165 services, Precision and Recall were computed for measuring the the accuracy and coverage of the automatic coding approach. We do this by comparing the automatically assigned codes with the benchmark dataset.

The automatic DESDE coding aligns well with the benchmark for `specific target group` and `mental health disorder` but performs poorly for `target age group`. For `specific target groups`, the system matches the benchmark in 146 of 165 cases (weighted precision 0.783 and weighted recall 0.885), and for disorder it matches in 145 of 165 cases (weighted precision .772 and weighted recall 0.879), indicating the model reliably captures who the service is for (e.g., carers, any population) and the service clinical focus (e.g., intellectual disabilities, eating disorders). In contrast, `target age group` matches only 36 of 165 cases (weighted precision 0.048, weighted recall 0.218), pointing to systematic confusion with the age-band taxonomy. We also looked at Macro precison and recall

Error analysis indicates the LLM often collapses to the same code for every service, highlighting a failure to condition on input content and weak grounding in the age taxonomy. These results indicate that the approach has practical utility for scaling service classification in these dimensions. However, the low performance on *target age group* shows that feasibility is not consistent across all coding categories.

Next steps involve designing fallback strategies for automated coding—for example, defaulting to the code GX ("all ages") when no explicit age cues are detected. In this context, a hybrid decision approach could be design, combining outputs from both the generative intelligence model and a rule-based approach that use explicit textual cues. The final code assignment could be determined through a voting or weighting scheme that integrates these complementary sources of evidence.

The relatively low macro precision and recall indicate that performance varies considerably across the different DESDE codes in the taxonomies. Because macro-averaged metrics assign equal weight to each category, they are particularly affected by misclassifications in less frequent or more ambiguous service types. In this case, the model performs reliably

	Macro P	**Macro R**	**Weighted P**	**Weighted R**
`Age`	0.027	0.125	0.048	0.218
`Target`	0.080	0.091	0.783	0.885
`Disorder`	0.098	0.111	0.772	0.879

Table 4: Macro, and weighted precision and recall for Target age (`Age`), Target population (`Target`), and mental health disorder (`Disorder`).

for common categories such as *general* or *adult services*, but struggles to correctly identify under-represented codes, including those referring to specific age groups or niche service types. This pattern reflects both class imbalance within the dataset and semantic overlap between certain codes (e.g., "All age" and"Adults 18–65"), which can lead to confusion during classification.

3.3 Service Discovery Evaluation

The performance of service discovery using an LLM to retrieve URLs is evaluated at two levels: domain and URL, through systematic comparison between the gold standard records and the automatically retrieved URLs.

Domain-level evaluation assesses whether the system can identify the correct website that hosts relevant information, reflecting broad coverage. URL-level evaluation assesses whether the LLM can find the exact webpage in the gold standard, which reflects precision and usability.

Domain-level evaluation: In total, we observed 298 unique domains across both sources: 15 in common, 115 appearing only in the gold standard, and 168 only in the automatic set. We calculate precision and recall, that is, the proportion of retrieved items that are relevant (precision) and the proportion of relevant items that are successfully retrieved (recall). This gives recall 11.5% (15/(15+115)), and precision 8.2% (15/(15+168)). Roughly one in twelve automatic domains is present in the benchmark.

URL-level evaluation: resulted in a small overlap between the gold standard and the automatic set, with only 2 matching URLs.

Further manual analysis is conducted to assess the relevance of domains that appear only in the automatically retrieved set, as these websites may contain information about services not captured in the gold standard. The top-40 domains were manually scored as relevant vs non-relevant by a health service specialist. It was found that 52.5% (21 URLs) is relevant, and 47.5% are not relevant. Among the reasons for no relevance, 26.3% (5 URLs) is out-of-scope location, 15.8% (3 URLs) are national-only online/helpline services, and 15.8% are too generic. These results indicate that, although the discovery of relevant URLs using LLMs demonstrates limited recall and precision when evaluated against the gold standard, it nonetheless yields a complementary set of domains, a substantial proportion of which are relevant but not present in the gold-standard reference.

A set of URLs only in the automatic records is also manually assessed by the health service specialist. Given 37 URLs, 46% (17/37) are relevant, 27% (10/37) are not relevant, 27% have access or availability failures (e.g., site cannot be reach (13.5%), can't access site (8.1%), Page not found (5.4%), and other are not relevant due to location mismatch, not been about services, among others. Overall, this analysis shows that while a meaningful proportion of the retrieved URLs (46%) are relevant, a substantial portion is either irrelevant or inaccessible,

Preliminary results on service discovery using a combination of LLMs and a search API indicate that, of the 4,900 URLs retrieved, only 1,608 are unique. The distribution is highly skewed: the top five domains alone account for a substantial share of the dataset (`pmc.ncbi.nlm.nih.gov`: 549; `pubmed.ncbi.nlm.nih.gov`: 323; `www.act.gov.au`: 250; `www.sciencedirect.com`: 219, `www.canberrahealthservices.act.gov.au`: 153). This reflects a strong over-representation of academic and medical publication sources (e.g., PubMed, ScienceDirect, PMC), despite explicit prompting to exclude them. At the same time, the retrieved set also captures relevant domains associated with local health service providers. Many other domains appear just once or twice, reflecting the exploratory breadth of search queries. These findings suggest that more targeted prompt engineering could help reduce noise by constraining the retrieval process, particularly to avoid the over representation of academic sources. Next steps also include designing a discovery method that combines API search for URL finding with an LLM to rank URLs for relevance.

4 Responsible AI guidelines

The capabilities of AI have increased rapidly in recent years, especially generative intelligence technologies [Bengio et al., 2025]. This technology is impacting society in different ways and at different actors – government, companies, researchers and civil society – who have distinct concerns, values and priorities. As with other new technologies, there is no global regulation. However, there are several 'guidelines' on the responsible use of AI (see [Bengio et al., 2025, Parliament and Council, 2024, UNESCO, 2021, OECD, 2019, High-Level Expert Group on Artificial Intelligence, 2019]). The proliferation of guidelines reflects both the complexity of AI's impact and the desire of different actors to shape its future responsibly.

In general, AI guidelines focus on the design, development and deployment of AI systems in a way that is ethical, transparent, and aligned with human rights and societal values. Most guidelines focus on risk management and risk mitigation and provide a framework for AI governance. The following sections elaborate on the responsible use of AI in M-Chart, by analysing the use of generative intelligence technologies, their potential risks, and proposing risk mitigation strategies to ensure safe and responsible use of th technologies.

4.1 Risk Analysis

AI risk analysis is about identifying, assessing, and managing the potential harms, failures, or unintended consequences that could arise from use of AI systems [Bengio et al., 2025].

Given the rapid pace of AI development, risk analysis should be treated as a dynamic and continuous process, regularly revisited and updated to reflect emerging challenges and changes in technology (e.g., tools upgrades, policy chances, guidelines updates, etc.).

Risk analysis can result on classifications such as low, medium, or high risk, depending on various factors including the specific use case, the operational context, and the effectiveness of controls in place. To accurately assess the level of risk associated with the use of AI in M-Chart, it is essential to conduct a thorough examination of how the technology is implemented and applied in practice. This involves evaluating not only the technical aspects of the AI system—such as its accuracy, robustness, and data sources—but also its potential impacts on individuals and communities. The analysis should consider who is affected, in what ways, and whether appropriate mechanisms exist to mitigate unintended consequences or ensure accountability.

In this study generative intelligence tools powered by LLMs are used for the following tasks: web-scraping, text generation for query expansion and Internet search. In what follows, the risk associated with each task is analysed in detail.

Web scraping There is a risk that web scraping techniques could infringe copyright agreements. In the scope of M-Chart, web-scraping is use to extract factual information about health services. According to the Copyright Act 1968[2], the extraction of factual information such as business names, addresses, and other publicly available data, does not constitute copyright infringement. The scraping of public domain content (e.g., registries like the Australian Cancer database) are not protected by copyright either. Moreover, under the Australian law there are "fair dealing exceptions" to copyright infringement. The Copyright Act states that copyright material may be used for research and study purposes. However, when the data scraped is used commercially (e.g., training an AI model) it may violate copyright.

Text generation - Query expansion LLMs generate text by predicting tokens based on patterns, not by copying and pasting from original data sources. Whether using a model trained on copyright material to cause infringement is still a gray area [Parliament and Council, 2024]. If the text generated is new, original, and not a close copy of any source, it is generally not infringing on copyright [Parliament and Council, 2024]. In M-Chart, LLMs are use to generate query variations for service discovery. In this context, it is unlikely that queries generated by LLMs would infringe copyright as it is not expected that LLMs generated queries include large verbatim text from copyrighted sources (e.g., paragraphs).

Internet search The relationship between internet search and copyright law is complex and evolving. Traditional search engines return a list of links ranked by relevance using keywords, backlinks, and user behaviour indicators. With the adoption of generative intelligence in search engines, the search experience is changing from a list of links to directly generating answers to user queries. Some of the challenges of AI powered search include creating new strategies to handle accuracy and bias issues linked to LLMs. Again, as for text generation, it is currently under intense debate whether using models trained on copyrighted material implies copyright infringement.

In addition to the tasks outlined above, additional potential risks have been identified.

[2] `https://www.legislation.gov.au/C1968A00063/latest/text`

Reliability issues: Generative intelligence tools rely on machine learning models that use statistics to make predictions. Therefore, these tools lack understanding, reasoning, and intentionality at their core. However, they are designed to simulate the mentioned traits, hence, to emulate human intelligence. This is problematic because it takes away responsibility. It is currently not clear who is accountable for the outputs produced by generative intelligence technologies.

Automation Bias: There is a tendency of people to over-rely on automated systems, even when the system is wrong or provides incomplete information. This cognitive bias can lead humans to accept suggestions or outputs from AI tools without sufficiently questioning their accuracy or appropriateness, for example, users may believe the system is more accurate or capable than it really is.

Environmental harm: systems powered by LLMs cause environmental harm due to their significant energy consumption, water use, and carbon emissions. In M-Chart, LLMs are use for inference (e.g., generating text and answering queries). While training LLMs is far more energy-intensive than inference per run, inference happens constantly and at scale, so the aggregate impact can be larger over time. In the context of this project, even inference would mostly happened periodically, for example, when finding services for a new jurisdiction, the inference tools used are run at scale.

4.1.1 Who is Affected? and in What Ways?

In the context of M-Chart, the use of generative intelligence technologies could potentially affect Service Discovery in two ways:

- Information about service providers could be incomplete
  * Some services might not be found using the proposed methods. This could have implications on service gaps identification and could potentially lead to ineffective interventions.
- Information about services could be inaccurate
  * It could impact the identification of service gaps that could potentially lead to misinformed decisions.
  * It could undermine service providers reputation and trust.

4.1.2 Risk Mitigation and Monitoring Strategies

To address the risks identified throughout the analysis, this report proposes a set of mitigation strategies designed to reduce the likelihood and potential impact of those risks. These strategies aim to strengthen system reliability, ensure responsible use of generative AI tools, and maintain human oversight where necessary. By implementing these measures, the project can proactively manage risk throughout its lifecycle. The proposed mitigation strategies are detailed below:

Keeping a human-in-the-loop: The human-in-the-loop concept refers to a system design approach in which human judgment is embedded within automated processes to ensure that critical decisions remain under human control. In the context of the M-Chart project, incorporating a human-in-the-loop approach is essential to the integrity of the service discovery pipeline. While automation enables scalability, human expertise is useful for interpreting ambiguous cases and validating service relevance.

Developing Easy to Test Systems: Implementing rigorous testing and validation processes ensure not only that AI tools behave as intended, but also help to identify errors and improvement opportunities, especially when outputs are validated against known benchmarks (e.g., M-Chart for mental health and dementia in the Australian Capital Territory).

Accountability: This remains one of the most unclear and contested aspects of generative AI, particularly when it comes to determining who is responsible for the outcomes produced by these systems. According to [Parliament and Council, 2024], currently, LLMs reliability is shared among the model developers, application builders, deploying organizations, and regulators. Each has different but complementary responsibilities to ensure these systems perform safely and fairly. The organisation deploying M-Chart is responsible for ensuring that systems are safe, fair, and subject to human control.

Minimising Environmental harm: It is possible to minimise the environmental harm cause by generative intelligence technologies by using smaller, optimized models for simple tasks (e.g., quantization[3]). Other avenues for reducing environmental impact include adopting data centers powered by renewable energy; designing efficient inference pipelines (e.g., batching[4], caching[5]); and choosing hardware-efficient architectures (e.g., LoRA, sparsity-aware models).

4.1.3 Risk Analysis Outcome

Based on the results of the risk analysis conducted, the activities undertaken within the scope of this project that involve the use of generative AI tools should be classified as ***low risk***. This assessment is grounded in a careful examination of the context and intended use throughout the project lifecycle. In particular, the risk analysis considered factors such as the nature of the tasks being supported by AI, the level of human oversight, the absence of sensitive or personal data processing, and the limited potential for harmful or unintended outcomes. The following points provide additional support for this conclusion:

- Outputs are always reviewed by humans who are trained or experts in the healthcare service domain.

[3] Quantization transforms floating-point values (e.g., 32-bit or 16-bit floats) into lower-precision formats such as 8-bit integers. This significantly reduces the machine learning memory footprint and computational demands.

[4] Batching improves efficiency by processing multiple inputs at once, for example: sending 100 prompts to an LLM in one batch maximizes hardware use, speeding up inference.

[5] Caching avoids recomputing results by storing outputs so they can be quickly reused if the same input appears again—avoiding redundant computation.

- Used in non-critical contexts such as web-scraping and paraphrasing.
- No personal data is processed or stored.
- Used in non-commercial products.
- Failures and inaccuracies carry minimal to medium consequences.
- Used in non-critical decision making. The intent of the M-Chart tool is to help health-care planners and professionals to navigate the mental health ecosystem; however it is not design to assist decision making directly.

In summary, this risk analysis has identified and assessed the potential risks associated with the use of AI tools within the scope of M-Chart Phase. Given the evolving nature of AI technologies and complex regulatory landscapes, it is recommended that this risk assessment be revisited periodically to ensure continued alignment with best practices.

5 Discussion and Conclusion

This study is to inform the usability of web scraping as a method to identify and update information on mental healthcare service provision in local areas. We use generative intelligence assisted web scraping to gather data from service websites — specifically: `service name`, *service provider*, *service address*, *target age group*, *specific target group*, and *mental health disorder* as they describe key characteristics of service provision.

We rated web scraping for data gathering at Technology Readiness Level 3 (TRL)—analytical and experimental proof-of-concept because the core functions have been demonstrated on limited datasets and privacy compliance has been overlooked. In short, feasibility is shown, but reliability, scalability, and governance have not been proven in an operational environment.

We also employed generative intelligence to map services to DESDE codes, which is a service classification system. We place the service coding (taxonomy classification) component in the TRL 6–9 trajectory: currently, it resembles TRL 6–7 (an integrated prototype demonstrated in a relevant environment) because it maps services to a constrained, well-defined taxonomy with reproducible prompts, and measurable accuracy. Once it operates reliably at scale with sustained performance under real workload and governance controls, it can be considered TRL 8-9.

The service provider discovery task constitutes the entry point of the data-collection pipeline: all subsequent stages—web scraping, extraction of service attributes, and DESDE coding—are constrained by what is initially identified on websites. Consequently, the completeness and validity of downstream tasks depend on the coverage, specificity, and fidelity of the discovery step. It worth mentioning that the success of data collection is bounded by the accuracy and completeness of information published on provider websites. Moreover, content quality is heterogeneous, some sites provide rich, structured detail while others are might be sparse and outdated — introducing variability in the extraction coverage. More importantly, the process is shaped by the assumption of clarity, which refers to the degree to which information, system design, documentation, or interfaces are easy to understand, unambiguous, and transparent for their intended users. Hence, we rate the service discovery component is at TRL-3—analytical and experimental proof-of-concept—because while its feasibility has been demonstrated, it remains limited in scope, highly dependent on the quality of provider websites, and has not yet been systematically validated for compliance.

Future work will focus on developing a pipeline that leverages agentic AI models to automate the end-to-end process, to enhance scalability. This will require rigorous validation against gold-standard datasets and may incorporate human-in-the-loop to ensure reliability.

Overall, the integration of LLMs into information-processing pipelines represents a promising avenue to identify and update information in health-system navigators. At the same time, it creates an opportunity to critically re-examine and redesign checkpoints for human oversight, which remain essential given the current risks and limitations associated with generative intelligence technologies.

References

[Bengio et al., 2025] Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., Newman, J., Ng, K. Y., Okolo, C. T., Raji, D., Skeadas, T., South, T., Strubell, E., Tramèr, F., Velasco, L., Wheeler, N., Acemoglu, D., Adekanmbi, O., Dalrymple, D., Dietterich, T. G., Felten, E. W., Fung, P., Gourinchas, P.-O., Heintz, F., Hinton, G., Jennings, N., Krause, A., Leavy, S., Liang, P., Ludermir, T., Marda, V., Margetts, H., McDermid, J., Munga, J., Narayanan, A., Nelson, A., Neppel, C., Oh, A., Ramchurn, G., Russell, S., Schaake, M., Schölkopf, B., Song, D., Soto, A., Tiedrich, L., Varoquaux, G., Yao, A., Zhang, Y.-Q., Albalawi, F., Alserkal, M., Ajala, O., Avrin, G., Busch, C., de Leon Ferreira de Carvalho, A. C. P., Fox, B., Gill, A. S., Hatip, A. H., Heikkilä, J., Jolly, G., Katzir, Z., Kitano, H., Krüger, A., Johnson, C., Khan, S. M., Lee, K. M., Ligot, D. V., Molchanovskyi, O., Monti, A., Mwamanzi, N., Nemer, M., Oliver, N., Portillo, J. R. L., Ravindran, B., Rivera, R. P., Riza, H., Rugege, C., Seoighe, C., Sheehan, J., Sheikh, H., Wong, D., and Zeng, Y. (2025). International ai safety report. Technical Report DSIT 2025/001, International Scientific Panel on Advanced AI Safety. Accessed June 2025.

[High-Level Expert Group on Artificial Intelligence, 2019] High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy ai. `https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1`.

[Li et al., 2024] Li, J., Yang, Y., Bai, Y., Zhou, X., Li, Y., Sun, H., Liu, Y., Si, X., Ye, Y., Wu, Y., Lin, Y., Xu, B., Ren, B., Feng, C., Gao, Y., and Huang, H. (2024). Fundamental capabilities of large language models and their applications in domain scenarios: A survey. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11116–11141, Bangkok, Thailand. Association for Computational Linguistics.

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

[OECD, 2019] OECD (2019). Recommendation of the council on artificial intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. OECD Legal Instruments, OECD/LEGAL/0449.

[OpenAI, 2023] OpenAI (2023). Chatgpt. https://chat.openai.com/chat. Accessed: [current date], https://chat.openai.com/chat.

[Parliament and Council, 2024] Parliament, E. and Council (2024). Artificial intelligence act (regulation (eu) 2024/1689). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689. Official Journal of the European Union.

[Romero-Lopez-Alberca et al., 2021] Romero-Lopez-Alberca, C., Alonso-Trujillo, F., Almenara-Abellan, J. L., Salinas-Perez, J. A., Gutierrez-Colosia, M. R., Gonzalez-Caballero, J. L., Pinzon Pulido, S., and Salvador-Carulla, L. (2021). A semiautomated classification system for producing service directories in social and health care (desde-and): Maturity assessment study. *Journal of Medical Internet Research*, 23(3):e24930.

[UNESCO, 2021] UNESCO (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

A Appendix A

This appendix include the forty-nine queries that were manually crafted and used as seed queries for discovering relevant websites (URLs) of mental health services in the Australian Capital Territory.

M-Chart expert initial queries:

1. Mental health services in the Australian Capital Territory
2. Public mental health support ACT
3. Mental health helplines ACT

Age taxonomy derived queries:

4. Mental health services Canberra Health Services in the Australian Capital Territory
5. Mental health services for children and adolescents in the Australian Capital Territory
6. Mental health services for First Nations in the Australian Capital Territory
7. Mental health services support for adults aged 18–65 in the Australian Capital Territory
8. Mental health services support for people of any age in the Australian Capital Territory
9. Mental health services support for children and adolescents (0–17) in the Australian Capital Territory
10. Mental health services support for younger children (up to 11) in the Australian Capital Territory
11. Mental health services support for adolescents (12–17) in the Australian Capital Territory
12. Mental health services support for adolescents & young adults (12–25) in the Australian Capital Territory
13. Mental health services support for young people transitioning to adulthood (16–25) in the Australian Capital Territory
14. Mental health services support for young adults (18–25) in the Australian Capital Territory
15. Mental health services support for older adults (65+) in the Australian Capital Territory
16. Mental health services support for adults transitioning to older age (55–70) in the Australian Capital Territory
17. Mental health services support for children transitioning to adolescence (8–13) in the Australian Capital Territory

Specific target groups taxonomy derived queries:

18. Mental health services support for people who have come into contact with the justice system in the Australian Capital Territory
19. Mental health services support for armed forces veterans and their families, typically funded by DVA in the Australian Capital Territory
20. Mental health services support for people who identify as male in the Australian Capital Territory
21. Mental health services support for people who identify as female in the Australian Capital Territory
22. Mental health services support for people who identify as LGBTIQIA+ in the Australian Capital Territory
23. Mental health services support for carers of people with mental health disorders in the Australian Capital Territory
24. Mental health services support for people who use alcohol and/or other drugs in the Australian Capital Territory
25. Mental health services support for Aboriginal and Torres Strait Islander peoples in the Australian Capital Territory
26. Mental health services support for people with culturally and linguistically diverse backgrounds in the Australian Capital Territory
27. Mental health services support for students of Australian National University in the Australian Capital Territory

Mental health taxonomy derived queries:

28. Mental health services support for general mental, behavioral and neurodevelopmental disorders in the Australian Capital Territory
29. Mental health services support for schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders in the Australian Capital Territory
30. Mental health services support for mood [affective] disorders in the Australian Capital Territory
31. Mental health services support for recurrent major depressive disorder in the Australian Capital Territory
32. Mental health services support for anxiety and other nonpsychotic mental disorders in the Australian Capital Territory
33. Mental health services support for hoarding disorder in the Australian Capital Territory

34. Mental health services support for eating disorders in the Australian Capital Territory
35. Mental health services support for pre or postnatal mental and behavioral disorders in the Australian Capital Territory
36. Mental health services support for disorders of adult personality and behavior in the Australian Capital Territory
37. Mental health services support for people with a medical disorder as primary or co-morbidity in the Australian Capital Territory
38. Mental health services support for people who have attempted suicide in the Australian Capital Territory
39. Mental health services support for health hazards related to socioeconomic and psychosocial circumstances in the Australian Capital Territory
40. Mental health services support for problems related to upbringing in the Australian Capital Territory
41. Mental health services support for carers of dependent relatives or those needing respite care in the Australian Capital Territory
42. Mental health services support for problems related to legal circumstances in the Australian Capital Territory
43. Mental health services support for physical and mental strain related to work in the Australian Capital Territory
44. Mental health services support for intellectual disabilities in the Australian Capital Territory
45. Mental health services support for the family of the primary service user in the Australian Capital Territory
46. Mental health services support for problems related to housing and economic circumstances in the Australian Capital Territory
47. Mental health services support for problems related to employment and unemployment in the Australian Capital Territory
48. Mental health services support for problems related to education and literacy in the Australian Capital Territory
49. Mental health services support for mental and behavioral disorders due to psychoactive substance use in the Australian Capital Territory

B Appendix B

This appendix include three DESDE taxonomies: age, specific group target population, and mental health disorders.

Code	Description
AX	Support for adults aged 18 to 65 years old
GX	Support for people of any age
CX	Support for children and adolescents (from 0 to 17 years old)
CC	Support for younger children (up to 11 years old)
CA	Support for adolescents (from 12 to 17 years old)
CY	Support for adolescents & young adults (from 12 to 25 years old)
TA	Support for young people transitioning to adulthood (from 16 to 25 years old)
AY	Support for young adults (from 18 to 25 years old)
OX	Support for older adults (65 years old plus)
TO	Support for adults transitioning to older age (from 55 to 70 years old)
TC	Support for children transitioning to adolescence (from 8 to 13 years old)

Table 5: Age taxonomy codes and their corresponding descriptions.

Group	Description
Offenders	S. s. for people who have come into contact with the justice system
Veterans	S. s. for armed forces veterans and their families
Male	S. s. for people who identify as male
Female	S. s. for people who identify as female
LGBTIQIA+	S. s. for people who identify as LGBTIQIA+
Carers	S. s. for carers of people with mental health disorders
Alcohol and Drug users	S. s. for people who use alcohol and/or other drugs
First Nations	S. s. for Aboriginal and Torres Strait Islander peoples
CALD	S. s. for people with culturally and linguistically diverse backgrounds
ANU students	S. s. for students of Australian National University
Any population	Does not specify any target group or population

Table 6: Specific target group taxonomy and their corresponding descriptions, where S. s. refers to Support specifically.

Code	Description
F00–F99	General mental, behavioral and neurodevelopmental disorders
F20–F29	Schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders
F30–F49	Mood [affective] disorders
F33	Recurrent major depressive disorder
F40–F48	Anxiety and other nonpsychotic mental disorders
F42.3	Hoarding disorder
F50	Eating disorders
F53	Pre or postnatal mental and behavioral disorders
F60–F69	Disorders of adult personality and behavior
ICD	Support for people with a medical disorder as primary or co-morbidity
T14.91	Support for people who have attempted suicide
Z55–Z65	Support for health hazards related to socioeconomic and psychosocial circumstances
Z62.8	Problems related to upbringing
Z63.6	Carers of dependent relatives or those needing respite care
Z65.3	Problems related to legal circumstances
Z56.6	Physical and mental strain related to work
F70–F79	Intellectual disabilities
e310x	The family of the primary service user
Z59	Problems related to housing and economic circumstances
Z56	Problems related to employment and unemployment
Z55	Problems related to education and literacy
F10–F19	Mental and behavioral disorders due to psychoactive substance use

Table 7: Mental health taxonomy codes and their corresponding descriptions (ICD-based categories).

C Appendix C

C.1 Data Gathering Prompt Example

[Website content

Service type. Respond only with the type.

Service name. Respond only with a name

Provider. Respond only with provider name.

Service address. Respond only with address.

Key description

Target population. Respond only with target population.

Role: You are an assistant that extracts structured information from websites.]

C.2 Service Coding Prompt Examples

Target age prompt example

[Age taxonomy

Target Population: Young people

Service Description: Free Planned Support program at Belconnen Youth Centre providing tailored, holistic help for young people to find and maintain employment, pursue education or training, navigate community/government services (including Centrelink), and build social connections and skills to achieve their goals.

Based on the above information, classify the service target age into the age taxonomy categories. Give each classification a confidence score between 0 and 1.Rank all codes by the highest score. Return results in JSON format: a list of records, where each record has: code, label, score.]

Target age prompt example

[Target group taxonomy

Target Population: Individuals new to custody or appearing before the ACT Law Courts where there are concerns about their mental health.

Service Description: ACT Mental Health Court Assessment and Liaison Service links community mental health with the ACT Law Courts, providing mental health assessments and advice (including for the Drug and Alcohol Sentencing List) for people in custody or before the court. Referrals must come from police or a magistrate/judge; it does not provide direct clinical care but can refer to services.

Based on the above information, classify the service target age into the age taxonomy categories. Give each classification a confidence score between 0 and 1.Rank all codes by the highest score.Return results in JSON format: a list of records, where each record has: code, label, score.]

Mental health disorder prompt example

[Mental health taxonomy

Target Population: Individuals experiencing disordered eating (eating disorders)

Service Description: The Short Term Recovery Intervention for Disordered Eating (STRIDE) is a voluntary, short-term CBT program offering up to 10 individual sessions with supervised provisional psychologists for people with disordered eating. Operated from the City Community Health Centre, referrals go through the Eating Disorders Clinical Hub. STRIDE is not a medical or crisis service (call Access Mental Health or Triple Zero (000) in emergencies).

Based on the above information, classify the service target age into the age taxonomy categories. Give each classification a confidence score between 0 and 1.Rank all codes by the highest score.Return results in JSON format: a list of records, where each record has: code, label, score.]

C.3 Query Expansion Prompt Example

[Expand the following search query into 10 diverse and relevant queries.

Seed: a seed query

Return only a numbered list of queries.

Role: Generate diverse web search queries targeting the Australian Capital Territory]