CMSC 35370: AI Agent for Science **Fall 2025**

Assignment 2 Report

Student: Chenxi Peng Finish date: October 26^{th}

1 Task 1: Ingest data into Milvus

During the text vectorization stage, I use BERT to encode the chunk. BERT is a pre-trained transformer model developed by Google for natural language processing (NLP) that excels at understanding the context of words by considering both the text that comes before and after them. It utilizes the encoder part of the transformer architecture, making it particularly effective for tasks like search result ranking, sentiment analysis, and question answering rather than generating new text.

Considering the length of the original document and the feature of scientific literature, I choose 1000 as chunk size and 100 as chunk overlap.

By running the command line below, I successfully insert 26811 entities into the local Milvus database. The fields of the database are `"id", "vector", "text"`.

```
python submit/ingest.py
-dataset-path /dataset
-database-path /db
vector
-model sentence-transformers/all-MiniLM-L6-v2
-embedding-size 384
```

2 Task 2: Ingest data into Whoosh

Based on the same chunk method, I create local Whoosh database. The fields of the Whoosh are `"id", "text"`. The command line to run the script are as shown below.

```
python submit/ingest.py
-dataset-path /dataset
-database-path /db
kw
```

3 Task 3: Set up retrieval tools

Firstly, I create an agent that can use the Milvus database to answer the questions. By running the command below, we get a final response:

```
 python submit/rag.py
-generation-model openai/gpt-oss-120b
-database-path /db
-query "what is treatment for pain?"
vector
-embedding-model sentence-transformers/all-MiniLM-L6-v2
=== Final Answer ===
Pain can be managed through a range of non-pharmacological and pharmacological
approaches, with primary-care patients with musculoskeletal pain showing the best
outcomes from non-drug strategies such as self-management advice and education,
structured exercise therapy, manual therapy, and psychosocial interventions (e.g.,
cognitive-behavioural therapy and pain-coping skills) that provide short-term
relief and functional improvement; pharmacological options-including oral or topical
analgesics, local joint injections, and, where appropriate, opioid medications-offer
only modest, short-term benefit, while adjunctive modalities like ultrasound,
transcutaneous electrical nerve stimulation (TENS), laser therapy, acupuncture, and
the use of ice/hot packs or aids/devices may be employed as supplemental treatments,
and referral for surgery is considered only when conservative measures fail.
```

Then I let ChatGPT come up with 10 questions based on the documents. I create `dataset/questions.jsonl` to store the questions and corespondding keywords. By running two commands shown below, I

get the final answers for the questions, and make a table to compare them. You can see the table in the Appendix.

A comparison of retrieval outputs reveals significant differences between the Whoosh and Milvus databases, which are triggered by keyword and sentence-based queries, respectively.

When a user provides keywords, the agent queries the Whoosh database and synthesizes an answer from the top five search results. Conversely, a full-sentence query utilizes the Milvus database. A primary distinction is the response length; answers generated from Whoosh are substantially more complete and detailed. However, this verbosity sometimes includes extraneous information. In contrast, responses from Milvus are more brief and concise, directly addressing the user's query.

This presents a clear trade-off. The keyword-based approach (Whoosh) provides a greater volume of information, which may offer unexpected insights or related inspiration. Its drawback is that the user may not receive a direct answer and might need to manually synthesize the relevant information from the detailed text. The sentence-based approach (Milvus) excels at providing direct, immediate answers, allowing the user to quickly obtain the specific information they seek. The limitation, however, is a lack of depth; the answers may not contain sufficient detail, potentially requiring the user to conduct further searches.

1 Appendix

Table 1: Comparison of 10 answers between the Vector Database and Text Database.

No.	Milvus Database Answer	Whoosh Database Answer
1	The research report by Fiona Paton and colleagues tackles the problem of how to improve outcomes for individuals experiencing a mental-health crisis by systematically examining what evidence exists for the different models of care that are currently available; in other words, it seeks to answer which crisis-care models are supported by the best evidence and how they might be implemented to achieve better patient outcomes.	The core issue highlighted across the discussion is the rapidly worsening mental-health crisis, which the Fiona Paton report identifies as a pressing public-health emergency driven by insufficient services, rising demand and systemic gaps in prevention and care; consequently, the central research question emerging from the report asks: "What evidence-based strategies and policy reforms can most effectively close these gaps and reduce the burden of mental-health problems on individuals and society?"—a query that seeks to translate Paton's detailed findings on service shortfalls, demographic disparities and the social determinants of mental ill-health into actionable recommendations for improving access, quality and outcomes in mental-health care.
2	The Crisis Concordat outlines four key stages of care for people experiencing a mental-health crisis: (1) ensuring access to support before the crisis point, (2) providing urgent and emergency access to crisis services, (3) delivering high-quality treatment and care while the person is in crisis, and (4) promoting recovery and preventing future crises.	The Crisis Concordat defines a mental-health-crisis pathway made up of four sequential stages of care: (1) providing support before a person reaches the crisis point, aimed at early-intervention and prevention; (2) ensuring urgent and emergency access to crisis services, including rapid assessment in accident-and-emergency departments and police-based routes under the Mental Health Act 1983; (3) delivering high-quality treatment and care while the person is in crisis, such as intensive community or inpatient interventions; and (4) promoting recovery and preventing future crises through follow-up, rehabilitation and longer-term support. The rapid evidence synthesis found very little robust evidence for the effectiveness of pre-crisis services, and the evidence for the best ways to improve emergency access and police involvement was inconclusive, highlighting significant uncertainties across all four stages.

Table 1 (continued)

No.	Vector Database Answer	Text Database Answer
3	The rapid synthesis focused on gathering and analysing existing secondary evidence, specifically searching for and reviewing relevant clinical guidelines and previously published systematic reviews as the primary types of evidence examined.	In responding to a mental-health crisis, researchers can draw on a range of evidence types—including quantitative data such as epidemiological surveys, service-utilisation statistics, and experimental trial results; qualitative insights from interviews, focus groups, and narrative case studies; and mixed-methods or real-world evidence like electronic health records and social-media analytics—to capture both prevalence and lived experience; by employing rapid-synthesis techniques such as accelerated systematic reviews, rapid evidence maps, or living reviews, they can quickly aggregate, appraise, and summarize this diverse literature, ensuring that policy makers and clinicians receive timely, rigorously vetted recommendations that reflect the most current and comprehensive understanding of the crisis.
4	The report concluded that crisis resolution teams (CRTs) are clinically more effective than inpatient care for a range of outcomes, demonstrating better results for people in crisis, although the implementation of this model varies across the UK and few teams meet all evidence-based criteria for good practice.	A thorough report on the effectiveness of crisis teams in addressing mental-health emergencies should begin with clear objectives, describe the population served, and outline the specific interventions employed (e.g., rapid assessment, de-escalation, linkage to follow-up care). Key performance indicators—such as reduction in hospitalization rates, time to stabilization, client satisfaction scores, and incidence of repeat crises—must be presented with baseline comparisons and statistical significance where possible. Qualitative insights from service users and staff, including themes of safety, perceived empowerment, and barriers to access, enrich the quantitative data and highlight areas for improvement. The analysis should contextualize findings within existing evidence, identify which team structures or practices yielded the greatest outcomes, and offer actionable recommendations—such as enhancing multidisciplinary coordination, expanding community outreach, or investing in staff training—to strengthen future crisis response and improve overall mental-health outcomes.
5	The researchers concluded that the evidence on which services most effectively improve urgent and emergency access to crisis care is inconclusive, meaning they could not identify or recommend specific interventions that reliably enhance emergency access in settings such as accident-and-emergency departments or support police under the Mental Health Act.	In conclusion, researchers emphasize that ensuring rapid, reliable urgent emergency access is a cornerstone of effective crisis-care services, as delays can exacerbate patient outcomes and strain health systems; their studies consistently show that integrating streamlined triage protocols, interoperable communication platforms, and community-based response teams not only shortens response times but also improves coordination among stakeholders, ultimately delivering more resilient and equitable care during emergencies.

Table 1 (continued)

No.	Vector Database Answer	Text Database Answer
6	The background notes that acyclovir is cleared primarily by the kidneys—it is excreted unchanged in the urine through glomerular filtration (with some tubular secretion)—and to a much lesser extent it undergoes hepatic metabolism to the inactive metabolite 9-carboxymethoxymethylguanine.	Acyclovir is eliminated from the body almost entirely by the kidneys, with the primary routes being glomerular filtration and active tubular secretion; a modest portion is reabsorbed in the renal tubules, and only a small fraction undergoes hepatic metabolism to inactive metabolites before being excreted in the urine. Consequently, renal function is the key determinant of acyclovir clearance, and dose adjustments are often required in patients with impaired kidney function.
7	The exclusion criterion for both versions of the study was a serum creatinine level greater than 1.7 mg/dL.	In all versions of the acyclovir trials involving infants, an elevated serum creatinine level has been used as a key exclusion criterion to protect participants from the drug's potential nephrotoxicity; typically, infants whose creatinine exceeds the age-adjusted normal range (often > 1.0 mg/dL or the upper limit of the laboratory reference for newborns) are screened out, and the exact cutoff may vary slightly between study protocols, but the underlying principle remains the same—ensuring adequate renal function before enrollment.
8	The later protocol versions (2.0 onward) limit pharmacokinetic sampling to a maximum total blood volume of 1600 μL per infant (equivalent to 1.6 mL).	For infant pharmacokinetic (PK) sampling under the acyclovir protocol versions 2.0 and later, the amount of blood that can be drawn is limited by the infant's total blood volume (approximately 80–90 mL per kg of body weight). The protocol permits no more than 5 percent of that total volume to be taken in any single collection—roughly 4–5 mL per kg—provided that the cumulative amount over a 24-hour period does not exceed 1 percent of the total volume (≈ 0.8–0.9 mL per kg) and that weekly totals stay below 3 mL per kg. In practice, this means a typical PK sample for an infant on the acyclovir 2.0+ regimen is limited to about 0.5–2 mL per draw, depending on the child's weight, ensuring safety while still permitting the necessary concentration measurements.
9	In the initial (Version 1.0) protocol, every infant group—whether preterm (23–29 weeks) or term (30–42 weeks) and regardless of post-natal age—was slated to receive a dose of **500 mg per square meter of body-surface area administered intravenously every 8 hours (q8h)**.	According to protocol version 1.0, the planned IV acyclovir dose for infants is 20 mg kg□¹ per administration (which corresponds to roughly 500 mg m□²), given every 8 hours; this regimen is recommended for all infant groups (typically those ≤ 12 months of age) to ensure adequate antiviral coverage while maintaining safety.
10	The original protocol specified that "the primary objective is to assess the pharmacokinetics of intravenously administered acyclovir at a single center in infants less than 60 days of life with suspected infection," while the revised version states that "the primary objective is to assess the PK of IV acyclovir in premature infants < 35 weeks gestational age and < 45 days of life."	The primary objective of the study is to evaluate the safety, pharmacokinetics, and therapeutic efficacy of acyclovir when administered to infants, using rigorously defined study protocols that specify dosing regimens, inclusion criteria, monitoring procedures, and outcome measures to ensure reliable data collection and ethical oversight.