IntelliStatCan: Gen-AI chatbot for obtaining statistical information

Presented by Milana Karaganis
Statistics Canada

December 2024

Agenda

- Business Opportunity
- Solution – Minimal Viable Product
- First Results and Further Enhancements
- Lessons learned so far

Beyond Research: Fully Integrating AI

Vision

Full responsible and secure integration of AI, including Gen-AI and LLMs, in Statistics Canada's processes to deliver relevant programs, efficient operations and enhanced services to Canadians.

Roadmap

Use cases: Demonstrate business value	**AI governance**	**AI, Gen-AI and LLM tools**	**Gen-AI playbooks**	**Support employees**
• Ideation • Prioritization • Implementation • Value realization	• Steering committee • Ethics • Legal • Technical	• Infrastructure • AI platform • Environments • COTS – Co-pilot etc.	• Best Practices • Guidelines • Checklists • Ways of working	• Communication • Change Management • Training

Business Opportunity

Business Challenge - Dissemination	Gen-AI Project Outcomes

Problem Statement
• External website search is a challenge for end-users
• Accelerate discovery of StatCan information by the end users
 	
Indicators of Success
• Internal expert validation feedback
• Retrieval of correct information to respond to the user query
• Validation with Client Engagement group
• Validation with Government of Canada key stakeholders
	Technical
	• Identify repeatable solution artifacts
	  • Common development environment
	  • Common production support
	  • Reusable integration artifacts
	  • Common DSc solution stack
	• Security requirements for approving Gen-AI development & production
	• Compare performance of GPT 3.5 vs. 4 – compute power and costing considerations
	
	Maturing IT knowledge of AI technologies

Solution

StatCan website GenAI Chat Bot (IntelliStatCan) using Retrieval-Augmented Generation (RAG)

Infrastructure

- Back-end: Data Analytics as a Service Development Environment + Production
- Front-end: Webapp hosted in Kubernetes

AI Engine

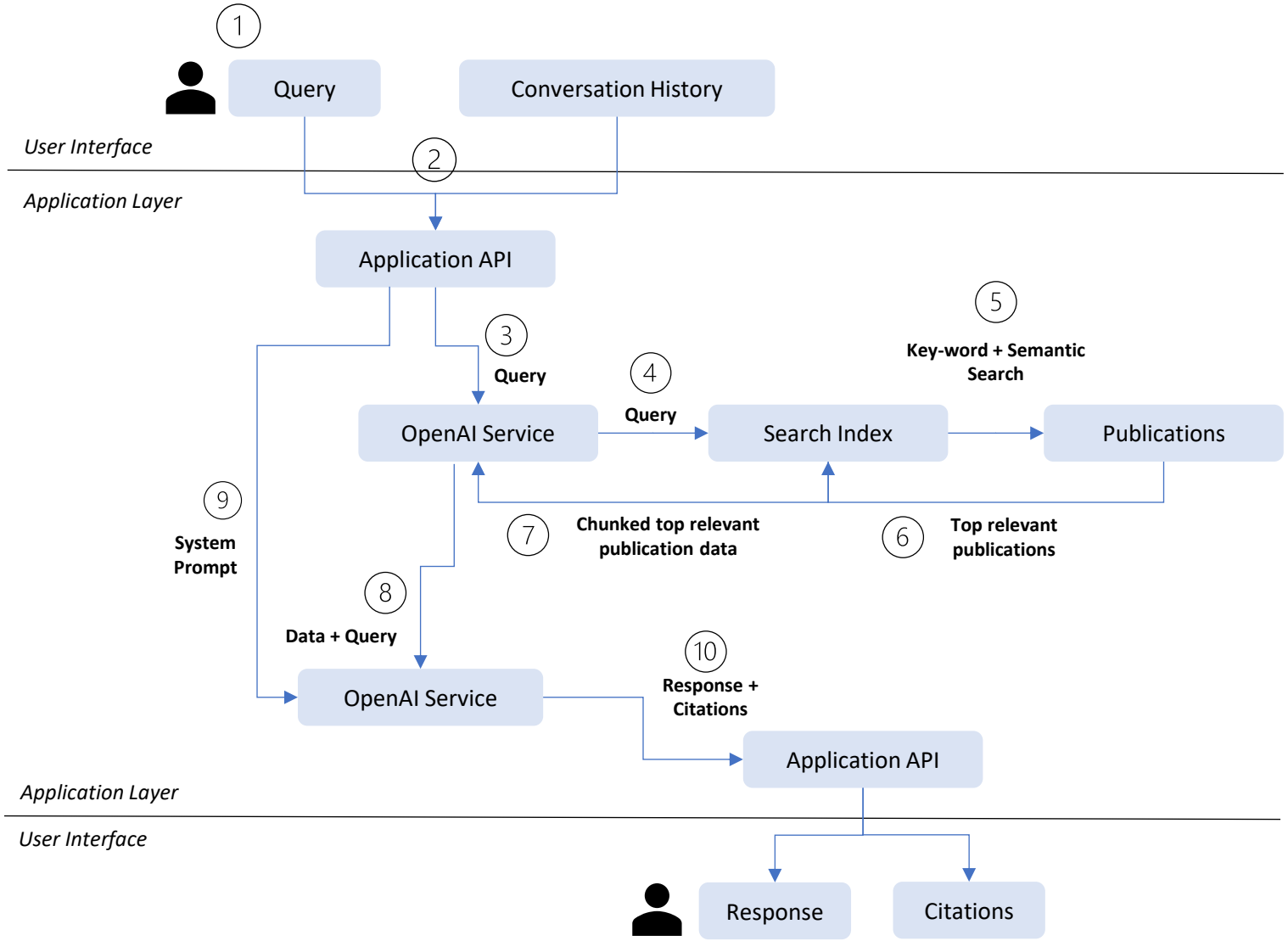
- MS Azure OpenAI Model API service (Model as a service): GPT 3.5 Turbo 16K vs. GPT-4 Omni
- Azure AI Search (formerly Cognitive Search)

Data

- Production: 18,000+ website publications in pdf, both English and French
- Publications are being added daily
- Exclude: Tabular Data, and non-text media (graphics, videos, etc.)

Technology stack – see Annex A

System Flow – Anatomy of a Query



1. *User submits a query using the Intelligent Search User interface.*
2. *User query, as well as any conversation history from the active session, is passed on to the application API service.*
3. *User query and conversation history is sent to the OpenAI service for response generation.*
4. *The user query is passed on the Azure Search index to retrieve relevant publication content.*
5. *The search index, which contains all of the publications in scope, is searched for relevant content.*
6. *The search index returns the most relevant publications based on preconfigured threshold.*
7. *The relevant content is served as chunked text blobs to OpenAI service for integration into AI prompt.*
8. 9. *The initial user query, the relevant data content, and the system prompt configured with the application, is passed along to OpenAI service for response and citations.*
10. *The AI-generated response, as well as the publication citations, are passed back to the user via the application API.*

Phase 1: Development of MVP Solution

- November-December 2023 - Rapid Prototyping (approx. 6 weeks)
- January 2024 - Agile Production Roadmap (approx. 9 weeks)
- February 27/28, 2024 - Hackathon
- March 2024 –Architecture and cyber approvals, accessibility and responsible ML reviews, governance approvals
- March 28, 2024 – MVP release and testing (through May)
- See Annex B – Examples of queries and test volumes

Phase 1: MVP Results

Configured a **production environment** incorporating **four new GenAI-related technologies** (Azure OpenAI, AI Search, Cosmos dB, Azure Function App)

User expectations based on ChatGPT (wanting the tool to do and answer more – constrained by focusing on pdf publications only)

Users experienced with the full website and would like Tabular data

Testing should be expanded beyond Subject-Matter experts (e.g., persons not experienced with prompting)

A good start, however **quality of responses** needs to be improved:

Search is as (or more) complicated as GenAI in a RAG

Orchestration features are not sufficient for citizen-public release

Next Steps – IntelliStatCan 2.0

- **Enhancements** being applied to the solution (now – January 2025):
  - Search: from **Full-Text Search with Semantic Reranking** to **Hybrid Search with markdown-based chunking and scoring profiles**
  - **Custom workflow orchestration**: from standard workflow orchestration built in Azure OpenAI to custom workflow orchestration
  - **Usability:** focus users on asking questions about available content (publications) rather than wide-open questions (e.g., display top 5 questions asked; show masking to suggest wording of questions etc.)
  - **Migrate to GPT-4o**

Next Steps – IntelliStatCan 2.0 (cont-d)

Response Quality Testing Framework (December 2024)
Iterative testing and response quality improvements (e.g., prompting, scoring, hackathon) (December 2024 – February 2025)
Governance reviews and approvals (architecture, cyber, responsible ML/AI, Accessibility etc.) (January – February 2025)
Production support & release plans; securing compute capacity
Complete release for Government of Canada testing (February 2025)
Further enhancements in user experience (see Annex C)

Lessons Learned

Quality standards place additional requirements on AI applications (e.g., RAG):

Teaching AI business logic and quality expectations via prompt design, fine-tuning, search optimization, and orchestration optimization – leverage entire tech stack

Finding balance to leverage technology and meet expert expectations

Cybersecurity matters – include from the beginning

Data is important – not only data quality but also data engineering and expecting continuous changes in datasets

Fast evolving technology – stretching typical processes to incorporate ongoing changes

ANNEXES

Annex A: Technology Stack

Cloud VM

Cloud VMs to enable one-time script runs to scrape and store publications content.

Azure AI Search

Azure AI Search to index and retrieve relevant content based on keyword and semantic search.

Kubernetes Services

Kubernetes clusters to host and serve frontend Intelligent Search application.

Azure Blob Storage

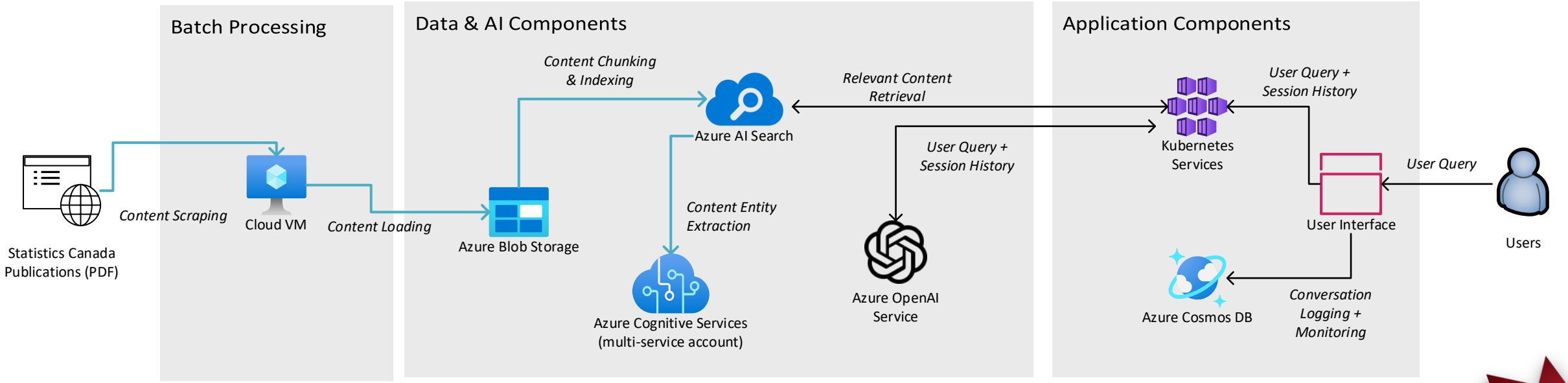
Azure Storage Containers to store publication content in both English and French.

Azure OpenAI Service

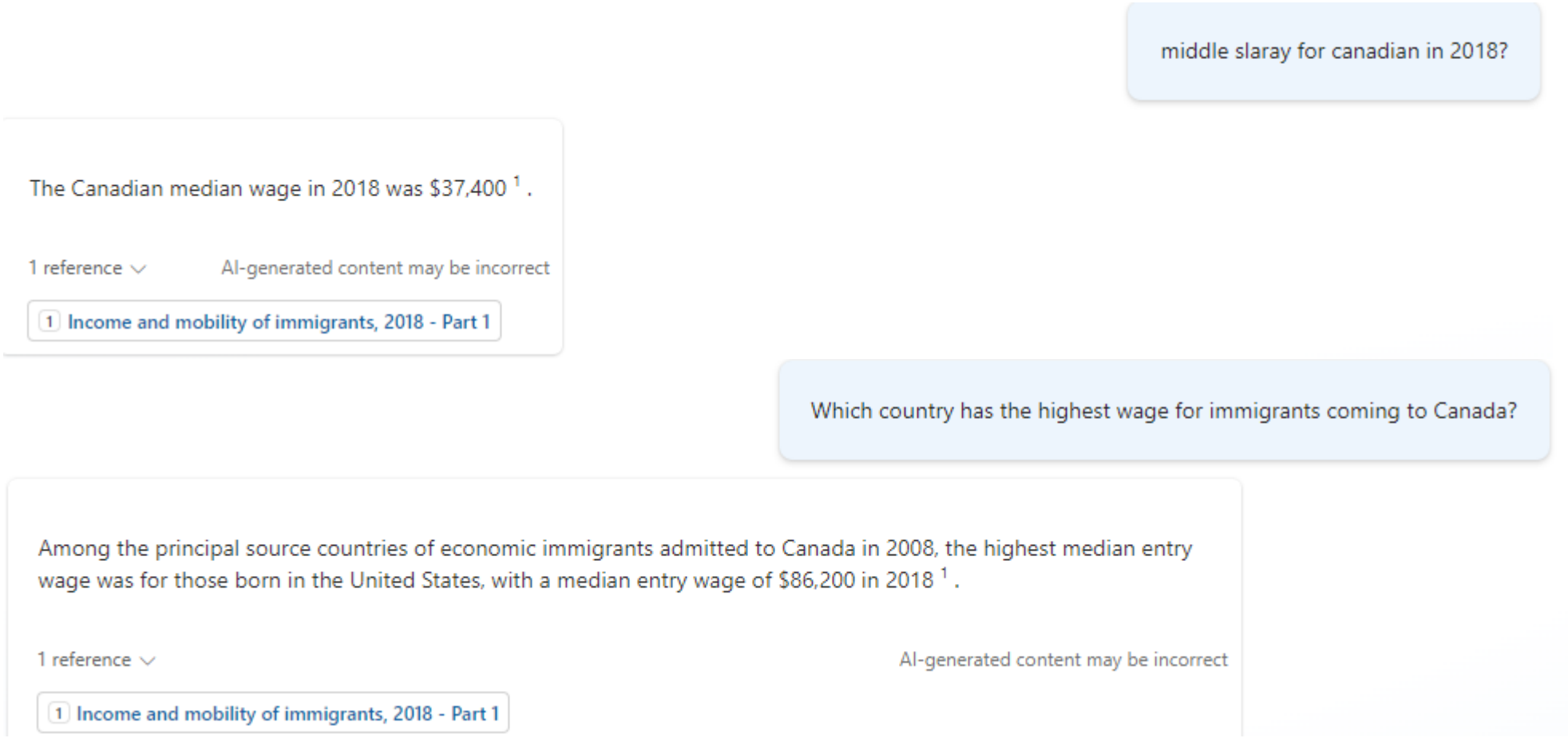
Currently using Azure OpenAI GPT-3.5 model and planning on upgrading to OMNI when available

Azure Cosmos DB

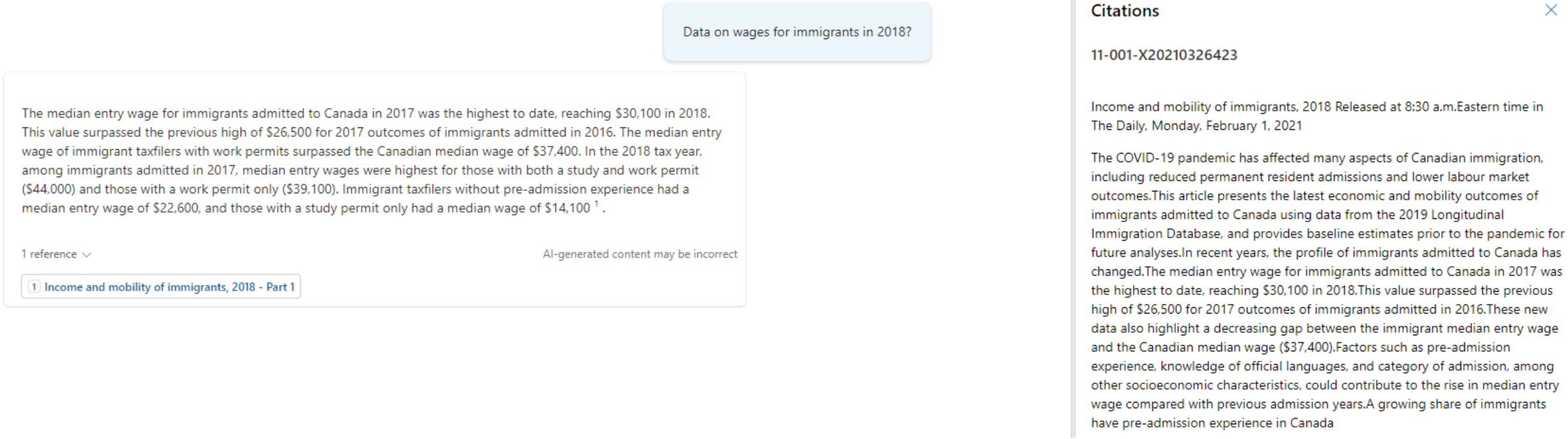
Azure Cosmos DB post-MVP for conversation logging and monitoring.



Annex B: Results from IntelliStatCan



Annex B: Result from IntelliStatCan - Citations



Annex B: 1 Month Usage
2024-05-28

Statistics Canada Intelligent Search App Usage - PROD

5/28/2024 8:59:26 PM

As Of Date/Time

1699
EN Conversations

5763
EN User Prompts

200
FR Conversations

585
FR User Prompts

EN User Prompt Common Words

FR User Prompt Common Words

Annex C: IntelliStatCan: Future User Experience - Verbosity

Statistics Canada Statistique Canada

Statistics Canada Intelligent Search

This tool is an Intelligent Natural Language Search experience on top of Statistics Canada Publications

Choose a conversation style

More **Succinct** | More **Balanced** | More **Detailed**

Data on wages for immigrants in 2018?

Annex C: IntelliStatCan: Future User Experience - Personas

Statistics Canada Statistique Canada

Statistics Canada Intelligent Search

This tool is an Intelligent Natural Language Search experience on top of Statistics Canada Publications

Choose a Persona

Research focused | **Publications** focused | **General** FAQs

Data on wages for immigrants in 2018?

Annex C: IntelliStatCan: Future User Experience - Filters

Statistics Canada Statistique Canada

Statistics Canada Intelligent Search

This tool is an Intelligent Natural Language Search experience on top of Statistics Canada Publications

Year	2023	2022	2021	View All
Subject Area	Energy	Population and demography	Transportation	View All

Data on wages for immigrants in 2018?

Stay connected!

StatsCAN app

Eh Sayers podcast

StatsCAN Plus

The Daily

Website

Surveys and statistical programs

Data service centres

My StatCan

Questions? Contact us: infostats@statcan.gc.ca

Statistics Canada—
Your National Statistical Agency

Delivering insight through data for a better Canada