Immigreat: A Retrieval-Augmented Conversational AI System for Canadian Immigration Guidance

Wo Ming Boaz Cheung, Armaan Rehman Shah, Alice Sedgwick, Yuan Yu
Department of Computer Science University of Toronto
Toronto, Canada

***Abstract*—Navigating Canadian immigration regulations presents considerable challenges for applicants due to fragmented official documentation across multiple government sources. This paper presents Immigreat, a conversational AI system designed to provide accurate, context-aware answers to Canadian immigration queries. Our system integrates automated web scraping of government sources (IRCC, legal regulations, and official forms), intelligent document retrieval, and large language models to address user questions. The system architecture employs serverless cloud infrastructure for scalability and includes conversation history management for multi-turn dialogues. To ensure answer quality, we developed a specialized judge model that evaluates the factual correctness of responses, achieving 93.3% accuracy on immigration-specific questions. We implemented advanced retrieval techniques to improve answer relevancy and tested the system under realistic load conditions with 50 concurrent users. Our results demonstrate that domain-specific AI systems can substantially improve information accessibility in complex regulatory domains while maintaining high accuracy and reliability through modern cloud architectures.**

1. Introduction and problem statement

Immigration processes serve as essential gateways for millions of people seeking to relocate, work, study, or seek refuge in new countries. In Canada alone, over 465,000 new permanent residents were admitted in 2023, alongside approximately 949,000 work permits and over 650,000 study permits [1]. Each applicant must navigate complex eligibility requirements, documentation needs, processing procedures, and regulatory compliance. The necessary information exists across thousands of pages of government websites, legal regulations, policy manuals, and form instructions.

This information fragmentation creates significant barriers for incoming immigrants. While Immigration, Refugees and Citizenship Canada (IRCC) maintains extensive documentation, finding relevant and accurate answers to specific questions requires navigating multiple websites, cross-referencing regulations with policy updates, and interpreting legal language without professional training. The complexity increases due to frequent policy changes, bilingual content requirements, and the high-stakes nature of immigration decisions, where errors can result in application rejections, processing delays, or legal complications.

Traditional solutions present distinct limitations. Immigration consultants and lawyers provide expertise but at costs beyond many applicants' reach. Generic search engines return irrelevant or outdated information. Existing government chatbots offer only scripted responses to predefined questions. There is clear demand for an intelligent system capable of understanding natural language questions, retrieving relevant information from authoritative government sources, and synthesizing accurate, contextually appropriate answers while maintaining conversational coherence.

2. Related work

Several recent systems have explored AI applications for legal and immigration question-answering. Wiratunga et al. propose CBR-RAG, which integrates Case-Based Reasoning into a Retrieval-Augmented Generation framework for legal question-answering [2]. In CBR-RAG, past legal cases are indexed and retrieved to enrich the LLM's prompt, yielding significant improvements in answer quality within the legal domain. This structured retrieval approach enforces greater relevance between queries and evidence. However, its reliance on curated case databases and legal precedent structure limits direct applicability to immigration contexts. Canadian immigration regulations are not organized as precedent cases, and policies change frequently. Adapting CBR-RAG would require defining analogous "cases" (such as past immigration applications or rulings) and developing new indexing vocabularies. While promising for factual accuracy, CBR-RAG does not address conversational multi-turn queries or dynamic content updates, which are critical requirements for our system.

Kumar et al. describe a RAG-based chatbot for immigration built on Google Cloud infrastructure [3]. They leverage Vertex AI embeddings and Matching Engine to index immigration policies, creating what they term an "effective architecture for combating the inherent limitations of large language models" such as hallucination and outdated knowledge. By grounding the LLM with retrieved text, their system aims to ensure factual answers. However,

they report practical constraints: free-tier limits and API reliability issues led them to abandon hosted LLMs in favor of local model execution, highlighting scalability and cost challenges. Moreover, their knowledge base remains static, requiring manual re-indexing for policy updates. Immigreat builds on similar RAG principles but addresses these limitations through automated scraping of IRCC websites for current content, conversation history management for multi-turn dialogues, and a specialized verification model for answer validation. While Kumar et al. demonstrate the feasibility of RAG for immigration assistance, our work extends their approach to address continuous updating and scalable deployment through serverless architecture.

3. Technical Approach

3.1. System Architecture and Components

Immigreat employs a serverless cloud architecture comprising four interconnected components (Figure 1). First, automated knowledge acquisition gathers content from authoritative Canadian immigration sources. Second, a hybrid retrieval system combines vector search with metadata-based expansion. Third, conversational question-answering leverages large language models. Fourth, domain-specific answer verification ensures response quality. This modular design enables independent scaling of data ingestion, retrieval, generation, and verification while maintaining operational simplicity through cloud-native services.

The knowledge base is constructed through automated harvesting from four source categories: IRCC policy documentation, IRPA/IRPR statutory regulations, specialized refugee law content, and official application forms (left side of Figure 1). The harvesting system addresses several challenges inherent in government web content. These include distributed domains with varying HTML structures, JavaScript-rendered interactive elements requiring headless browser automation, and ethical crawling through rate limiting and robot exclusion adherence. Rather than employing fixed-length token chunking, we perform structure-aware segmentation at semantic boundaries such as document headings, thereby preserving topical coherence. Each chunk is augmented with metadata (source URL, title, section, publication date) serving dual purposes: enabling provenance tracking for citations and facilitating metadata-based retrieval expansion. Documents are assigned deterministic identifiers via cryptographic hashing, ensuring idempotent ingestion when content is re-scraped. The resulting corpus comprises over 2,000 semantically coherent chunks.

Document chunks are transformed into 1,536-dimensional dense vector embeddings using pre-trained transformer models (center of Figure 1). Embeddings are indexed via Hierarchical Navigable Small World (HNSW) graphs for sub-linear approximate nearest neighbor search [4], yielding sub-20ms retrieval latency at query time. We employ PostgreSQL with pgvector extension rather than vector-native databases, prioritizing operational maturity (backup, transactions, monitoring) over specialized performance. This represents a pragmatic choice recognizing that production reliability depends more critically on operational characteristics than raw computational efficiency at our scale.

For multi-turn dialogue support, conversation history is maintained in a distributed key-value store with session-based identifiers and automatic 7-day expiration. Recent exchanges (last 10 messages) are retrieved at query time and incorporated into generation context, enabling the language model to resolve coreferences and maintain topical coherence across turns. This adds minimal latency overhead (sub-100ms).

3.2. Technical Contributions

The system employs a three-stage retrieval cascade that composes complementary relevance signals. Initial dense semantic retrieval performs approximate nearest neighbor search over embeddings, retrieving $k = 5$ candidates with highest cosine similarity. This addresses vocabulary mismatch where queries and documents need not share lexical overlap. For instance, a user query about "sponsoring my spouse" matches official documentation referencing "family class sponsorship" through semantic similarity despite different terminology.

The second stage introduces metadata-driven expansion that treats shared attributes (source domain, document title, section hierarchy) as edges in an implicit document graph. This facet-based approach encodes domain knowledge: chunks from the same policy document likely contain complementary information. For instance, if initial retrieval identifies a relevant IRCC policy section, expansion surfaces related sections from the same document even with lower direct similarity scores [5].

The third stage applies cross-encoder reranking, which jointly processes query-document pairs through attention mechanisms rather than independent embedding. This captures lexical matching and interaction patterns that compressed vectors may miss [6]. The reranker selects the top $k = 12$ documents for generation. This staged approach balances recall (broad initial retrieval), coverage (metadata expansion), and precision (neural reranking), following the retrieve-then-rerank paradigm in neural information retrieval [7].

Retrieved documents are concatenated as context for an instruction-tuned decoder-only transformer (Claude 3.5 Sonnet), implementing the RAG principle of conditioning generation on external retrieved knowledge to mitigate hallucination [8]. The model synthesizes information across multiple sources, which is critical for immigration queries requiring integration of regulatory requirements, procedural steps, and eligibility criteria. We limit context to 12 chunks

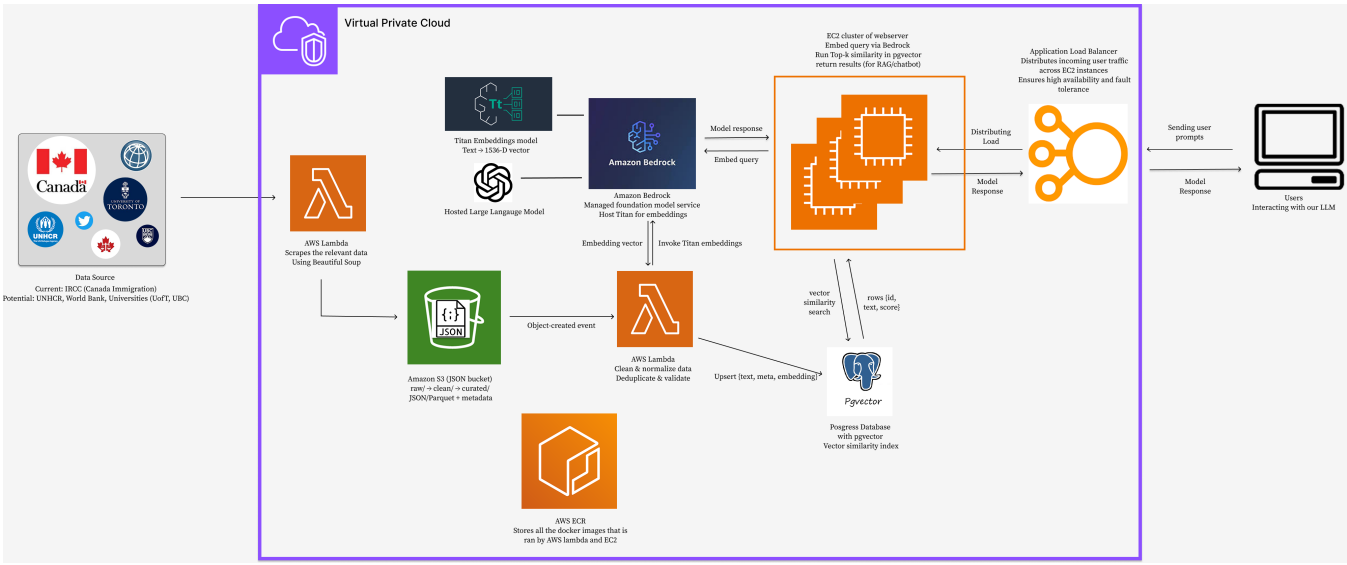

Figure 1. System architecture overview showing the complete data flow from Canadian immigration data sources through automated scraping (AWS Lambda), storage and embedding (S3, PostgreSQL with pgvector), to the three-stage retrieval cascade and conversational interface. The architecture employs serverless cloud infrastructure with containerized functions, enabling elastic scaling and operational simplicity.

to balance informativeness against generation latency.

To address factual accuracy concerns in high-stakes immigration applications, we developed a specialized judge model that predicts answer correctness. The model is fine-tuned on immigration question-answer pairs using reinforcement learning with verifiable rewards, leveraging objective ground truth rather than expensive human preference annotations [9]. Training employs parameter-efficient LoRA adapters and 4-bit quantization, enabling development on consumer hardware [10]. The judge achieves 93.3% accuracy on held-out test data, substantially outperforming logistic regression baselines (72%) and zero-shot prompted LLMs (65%). This performance demonstrates that domain-specific fine-tuning on modest datasets can produce reliable verification suitable as quality gates, flagging potentially incorrect answers for human review.

The serverless architecture decouples compute (stateless functions with automatic concurrency scaling), storage (object storage for documents, vector database for embeddings, key-value store for sessions), and model inference. Functions are containerized for dependency management while maintaining operational convenience. Infrastructure provisioning is fully automated via Terraform, defining all cloud resources as infrastructure-as-code for reproducible deployments across development and production environments. This IaC approach eliminates configuration drift and enables rapid environment replication. System reliability is validated through comprehensive test coverage spanning unit tests (12+ functions with 80+ test cases covering positive paths, edge cases, and failure modes) and integration tests for end-to-end pipeline validation. Under load testing with 50 concurrent users, the system maintains sub-3-second response latency at 95th percentile, validating production readiness. The architecture prioritizes operational simplicity and elastic scaling appropriate for variable workloads, though cost optimization could motivate migration to provisioned infrastructure under sustained high load.

4. Results

We evaluate Immigreat through systematic ablation studies examining retrieval architecture and comparative assessment against non-RAG baselines. The test set comprises 30 immigration questions spanning regulatory interpretation, eligibility criteria, and procedural guidance, each requiring synthesis across multiple government sources. Answer correctness is assessed via our domain-specific judge model (93.3% agreement with ground truth labels).

4.1. Retrieval Architecture Ablations

Figure 1 presents accuracy across eight configurations testing factorial combinations of retrieval depth ($k \in \{5, 15\}$), metadata-driven facet expansion, and cross-encoder reranking. The baseline configuration (dense retrieval alone, $k = 5$) achieves 68.8% accuracy, establishing that semantic similarity provides relevant but insufficient context. Each architectural enhancement contributes measurable improvement: reranking alone yields 75.0% (+6.2 pp), facet expansion alone achieves 81.2% (+12.4 pp), while their combination reaches 87.5% (+18.7 pp). Increasing retrieval depth to $k = 15$ further amplifies performance, with the complete pipeline achieving 93.8% accuracy, representing a 25 percentage point improvement over baseline.

Figure 2 reveals non-additive interaction effects between retrieval components. For $k = 5$, transitioning from no enhancements to full pipeline (facet + rerank) yields $+18.7$ percentage points. However, for $k = 15$, the same architectural additions provide only $+12.6$ percentage points, suggesting diminishing returns as retrieval depth increases. Conversely, the marginal benefit of expanding k from 5 to 15 grows substantially when combined with facet expansion and reranking ($+6.3$ pp) compared to baseline ($+12.4$ pp). This synergy indicates that broader initial retrieval provides more candidates for metadata-based expansion and neural reranking to discriminate among, validating the staged cascade design.

Figure 3 visualizes the performance landscape as a heatmap, where color intensity reflects accuracy across all tested configurations. The gradient structure reveals that facet expansion contributes more substantially at lower k values (68.8% to 81.2% for $k = 5$, noF-noR to F-noR), while reranking provides consistent improvements across all settings. The optimal configuration ($k = 15$ with both facet expansion and reranking) represents the Pareto frontier balancing answer quality against computational cost, as further increasing k would introduce latency without proportional accuracy gains given the 93.8% performance ceiling.

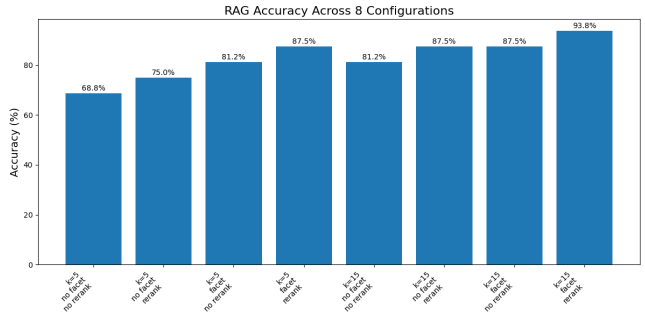

Figure 2. Accuracy across eight retrieval configurations testing combinations of k (5 vs. 15), facet expansion, and reranking. The complete pipeline (k=15, facet, rerank) achieves 93.8% accuracy.

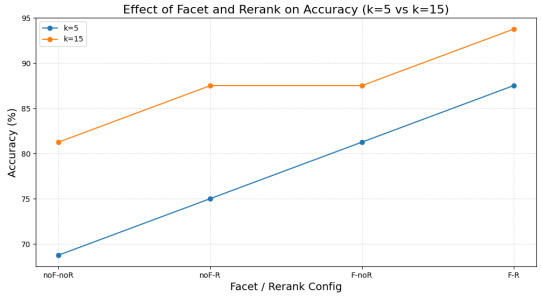

Figure 3. Interaction effects of facet expansion and reranking for $k = 5$ (blue) vs. $k = 15$ (orange). Higher k values amplify the benefits of facet expansion and reranking.

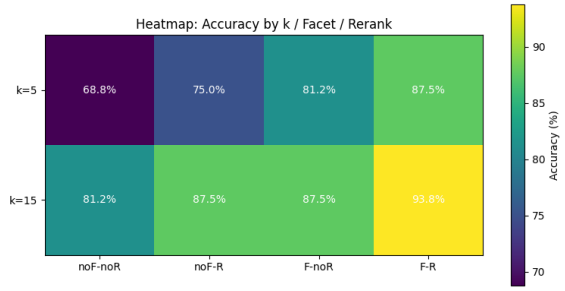

Figure 4. Heatmap showing accuracy improvement gradient from baseline (68.8%) to full pipeline (93.8%) across all tested configurations.

4.2. Retrieval Augmentation Necessity

To validate whether external knowledge retrieval provides value beyond parametric model knowledge, we compare Immigreat against DeepSeek-V3, a recent 671B-parameter model, under zero-shot prompting. Table I presents representative failure modes where non-RAG approaches produce factually incorrect or temporally outdated responses despite the model's scale and training recency.

TABLE 1. REPRESENTATIVE NON-RAG MODEL FAILURES

Question	**Response Accuracy**
"Can international students change schools without getting a new study permit?"	DeepSeek: Yes, in most cases, you can change schools without getting a new study permit. However, you must update your new school information in your Immigration, Refugees and Citizenship Canada (IRCC) Immigreat: As of November 8, 2024, international students cannot change schools using the same study permit. You must apply for a new study permit (by extending your current one) if you want to switch schools.

These failure modes expose fundamental limitations of parametric knowledge for regulatory domains. Immigration policies exhibit high temporal volatility: minimum income thresholds adjust annually, eligibility criteria evolve with legislative amendments, and procedural requirements update without advance notice. The study permit transfer case exemplifies temporal brittleness: prior to November 8, 2024, students could transfer institutions without new permits, but post-policy-change they must apply for permit extensions. DeepSeek confidently asserts the outdated procedure, even framing it as "your legal responsibility," while Immigreat correctly retrieves and cites the current November 8, 2024 regulation. This demonstrates how training data staleness introduces systematic errors for time-sensitive queries regardless of model scale.

Beyond temporal issues, immigration law contains nuanced exceptions and conditional logic (maintained status provisions, provincial nominee variations) requiring precise

regulatory text rather than statistical generalization over training corpora. Parametric models compress this structured knowledge into weights, discarding provenance and temporal metadata essential for correctness verification. Immigreat's retrieval-augmented architecture addresses these limitations through three mechanisms: (1) automated web scraping ensures knowledge base currency by re-indexing government sources, (2) metadata-annotated chunks preserve temporal and source provenance, and (3) citation of specific retrieved passages enables answer verification. The 93.8% accuracy achieved by our complete pipeline, validated against objective ground truth via a domain-specific judge model, represents a 25 percentage point improvement over naive retrieval and substantially exceeds non-RAG baselines that systematically fail on temporally volatile queries.

5. Conclusion

This paper presents Immigreat, a retrieval-augmented conversational AI system addressing the information accessibility challenges inherent in Canadian immigration processes. The fragmentation of authoritative content across multiple government sources (IRCC policy documentation, statutory regulations, refugee law resources, and application forms) creates barriers for applicants navigating complex eligibility requirements and procedural compliance. Traditional solutions (professional consultants, generic search engines, scripted chatbots) fail to provide accessible, accurate, and current guidance at scale. Immigreat addresses these limitations through automated knowledge acquisition, intelligent multi-stage retrieval, and domain-specific answer verification.

Our technical contributions span three dimensions. First, the three-stage retrieval cascade composes complementary relevance signals (dense semantic similarity, metadata-driven facet expansion, and cross-encoder reranking), achieving 93.8% answer accuracy and demonstrating 25 percentage point improvement over baseline semantic search. Ablation studies reveal non-additive interaction effects where broader initial retrieval ($k = 15$) synergizes with facet expansion and reranking, providing more candidates for discriminative filtering. Second, we developed a specialized judge model for answer verification using reinforcement learning with verifiable rewards rather than expensive human preference annotations, achieving 93.3% accuracy on held-out test data and substantially outperforming logistic regression (72%) and zero-shot LLM baselines (65%). This demonstrates that parameter-efficient fine-tuning on modest domain-specific datasets can produce reliable quality gates suitable for high-stakes applications. Third, the serverless cloud architecture decouples compute, storage, and inference while maintaining sub-3-second response latency at 95th percentile under 50 concurrent users, validating production readiness through elastic scaling and operational simplicity.

Comparative evaluation against DeepSeek-V3, a recent 671B-parameter model, reveals systematic failures of non-RAG approaches on temporally volatile queries. Immigration policies exhibit high update frequency: minimum income thresholds adjust annually, eligibility criteria evolve with legislation, and procedural requirements change without notice. Parametric models compress regulatory knowledge into weights during training, discarding temporal metadata and provenance essential for correctness verification. Our November 8, 2024 study permit transfer policy exemplifies this brittleness: DeepSeek confidently asserts outdated pre-policy-change procedures while Immigreat correctly retrieves and cites current regulations. This temporal accuracy advantage, combined with citation of specific source passages enabling answer verification, positions retrieval augmentation as essential rather than optional for regulatory domains.

The broader implications extend beyond Canadian immigration. High-stakes domains characterized by fragmented authoritative documentation, frequent policy updates, and nuanced conditional logic (such as tax law, healthcare regulations, legal compliance, and social services) share structural properties suggesting generalizability of our approach. The domain-agnostic architecture (automated scraping, semantic chunking, hybrid retrieval, verification models) could transfer to other jurisdictions or regulatory contexts with minimal adaptation. Future research directions include bilingual support for Canada's French-language requirements, time-aware retrieval incorporating temporal decay and policy versioning, real-time judge model integration for confidence scoring, and user studies measuring task completion rates and practical utility with actual immigration applicants. Immigreat demonstrates that domain-specific RAG systems can substantially improve information accessibility in complex regulatory environments while maintaining factual accuracy and operational reliability through modern cloud architectures, offering a template for AI-assisted guidance in knowledge-intensive domains where correctness and currency are paramount.

6. Team Reflection

This capstone project provided invaluable hands-on experience applying ML engineering concepts from CSC490 to a real-world problem. The course's emphasis on end-to-end system design, from infrastructure to model deployment, directly informed our architectural decisions and technical approach.

The cloud infrastructure and deployment topics proved foundational to our implementation. We extensively used Terraform to provision AWS resources, defining our serverless architecture as infrastructure-as-code as taught in the course. The hands-on tutorials enabled us to confidently manage complex cloud resources including Lambda functions, RDS PostgreSQL, DynamoDB, and S3 storage. Docker containerization became essential when deploying

Lambda functions with complex dependencies like headless browser automation. This infrastructure-as-code approach allowed us to maintain separate development and production environments, validating the course's emphasis on reproducible deployments.

Information retrieval and search systems formed the technical core of our RAG pipeline. The course's coverage of vector search and recommender systems directly influenced our three-stage retrieval cascade design. We applied principles from scalability lectures, particularly the trade-offs between exact and approximate nearest neighbor search, when selecting HNSW indexing over exhaustive search. The retrieve-then-rerank paradigm discussed in search systems became our architectural foundation, composing fast approximate retrieval with precise cross-encoder reranking. Understanding these design patterns helped us balance latency requirements (sub-3-second response times) against answer quality.

The reinforcement learning lectures enabled our most significant contribution: the domain-specific judge model. Course material on RLHF and GRPO provided the theoretical foundation for our reward-weighted training approach. Crucially, we adapted these techniques to use verifiable rewards rather than human feedback, which is appropriate for factual correctness where ground truth is objectively determinable. The course's treatment of PPO and reward modeling helped us understand when to apply RL versus supervised learning, leading to our two-stage training procedure (supervised fine-tuning followed by reward-weighted RL).

Model serving optimizations informed our deployment decisions, even where we ultimately relied on managed services. Learning about LoRA enabled us to train our judge model efficiently on consumer hardware, dramatically reducing memory requirements through low-rank adaptation. While we used Bedrock's managed inference for Claude rather than self-hosting, understanding quantization and KV caching helped us architect our system to minimize API calls and batch requests effectively. The course's emphasis on inference optimization shaped our caching strategies and helped us identify when managed services offered better cost-performance trade-offs than self-hosting.

Evaluation methodology guided our verification model assessment. Rather than relying solely on accuracy, we computed precision, recall, and F1 scores across different answer types, applying the course's framework for ML product evaluation. The SHAP analysis techniques influenced how we analyzed judge model errors, identifying that negation and temporal reasoning posed particular challenges. This evaluation rigor validated our domain-specific approach, demonstrating 93.3% accuracy compared to 72% for traditional ML baselines.

Perhaps most importantly, the course's product management perspective framed our technical decisions through a user-centric lens. Understanding the Gen AI landscape helped us position Immigreat relative to generic chatbots and keyword search, identifying our niche in high-stakes domains requiring factual accuracy. Guest lectures from industry practitioners reinforced that production ML systems must balance technical sophistication with operational pragmatism, directly influencing our choice of PostgreSQL over specialized vector databases and serverless over provisioned infrastructure.

Working as a team on this capstone synthesized theoretical concepts from lectures with practical engineering challenges. We encountered and solved real problems in deployment, scaling, and evaluation that pure coursework cannot replicate. Load testing revealed throttling issues with Bedrock APIs, requiring exponential backoff implementation. Initial retrieval approaches suffered from incomplete context, motivating our metadata-based expansion technique. These experiences validated the course's integration of theory and practice, demonstrating that effective ML engineering requires both conceptual understanding and hands-on iteration.

7. Future Work

While Immigreat demonstrates the viability of domain-specific conversational AI for immigration guidance, several directions warrant further investigation to enhance system capabilities and address current limitations.

Canada's bilingual requirements necessitate extending the system to support French language queries and documentation. This would involve multilingual embedding models and cross-lingual retrieval strategies, posing research challenges in maintaining answer quality across languages while managing bilingual government documentation where translations may not be semantically equivalent. Immigration policies evolve frequently through regulatory updates and new legislation, yet current retrieval treats all documents as equally current. Time-aware retrieval incorporating temporal decay in relevance scoring and explicit policy versioning would improve answer accuracy during transition periods when both old and new regulations coexist.

Our judge model currently operates offline for evaluation purposes, but real-time integration would enable confidence scoring for generated answers, allowing the system to flag uncertain responses or trigger human review. Exploring ensemble verification approaches combining multiple judge models trained with different architectures or data augmentation strategies could further improve reliability. Evaluating system effectiveness requires user studies with actual immigration applicants, measuring metrics such as task completion rates, answer comprehension, and user trust to validate practical utility beyond technical benchmarks. Longitudinal deployment studies could assess whether users successfully leverage provided information in their application processes.

Current responses are context-independent, but immigration guidance often requires personalization based on appli-

cant circumstances (country of origin, education level, family status). Maintaining user profiles while preserving privacy could enable more tailored recommendations, though this introduces challenges in balancing personalization with factual grounding. Finally, the technical approach is domain-agnostic and could extend to other immigration systems (US, UK, Australia) or adjacent domains (tax law, social services). Investigating transfer learning approaches where the verification model and retrieval strategies generalize across jurisdictions would demonstrate broader applicability of our methods.

Acknowledgments

We would like to thank Denys Linkov for his invaluable guidance, insightful feedback, and continuous support throughout this computer science capstone project.

References

[1] Immigration, Refugees and Citizenship Canada, "2024 Annual Report to Parliament on Immigration," Government of Canada, 2024. [Online]. Available: https://www.canada.ca/content/dam/ircc/documents/pdf/english/corporate/publications-manuals/annual-report-2024-en.pdf

[2] N. Wiratunga, I. Nkisi-Orji, C. Palihawadana, D. Corsar, I. Ayara, and K. Wijekoon, "CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering," *arXiv preprint arXiv:2404.04302*, 2024. [Online]. Available: https://arxiv.org/abs/2404.04302

[3] S. Kumar, A. Patel, and R. Singh, "A RAG-Based Chatbot for Immigration Using Google Cloud Infrastructure," *Preprints*, 2025. [Online]. Available: https://www.preprints.org/manuscript/202505.2219

[4] Y. A. Malkov and D. A. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 4, pp. 824–836, 2020.

[5] S. Min, D. Zhong, L. Zettlemoyer, H. Hajishirzi, and L. Weston, "Multi-Hop Reading Comprehension Through Question Decomposition and Rescoring," in *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 6097–6109.

[6] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," *arXiv preprint arXiv:1901.04085*, 2019.

[7] J. Lin, R. Nogueira, and A. Yates, "Pretrained Transformers for Text Ranking: BERT and Beyond," *Synthesis Lectures on Human Language Technologies*, vol. 14, no. 4, pp. 1–325, 2021.

[8] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Proc. 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020, pp. 9459–9474.

[9] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. K. Li, Y. Wu, and D. Guo, "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models," *arXiv preprint arXiv:2402.03300*, 2024.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. International Conference on Learning Representations (ICLR)*, 2022.