Evaluating the Open-source Toxicity Detector, "Detoxify", in Sensitive Migration and Health Contexts

SHUTING XIE, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Canada
VAISHALI MEYAPPAN, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Canada
HAMED KARIMI, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Canada
HIRAD DANESHVAR, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Canada
MOHAMMAD AMIN KAMALEDDIN, Interventional Psychiatry Program, St. Michael's Hospital, Unity Health, Canada
REZA SAMAVI, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Canada and Vector Institute, Canada
VENKAT BHAT, Interventional Psychiatry Program, St. Michael's Hospital, Unity Health, Canada and Department of Psychiatry, University of Toronto, Canada
ROBERTO SASSI, Child and Adolescent Division, Department of Psychiatry, University of British Columbia, Canada

Research Question. Open-source safety tools are attractive in sensitive domains such as mental health and migration because they can be deployed locally, reducing reliance on external services for processing private and institutionally sensitive text. This makes them especially appealing in settings where confidentiality, data governance, and practical deployability matter. However, local deployment does not by itself guarantee that a tool is reliable or appropriate for high-stakes use. In healthcare- (more specifically, mental-healthcare-) and migration-related contexts, misclassification can have serious consequences. For example, false positives may suppress help-seeking, trauma-related, or identity-linked expression, while false negatives may fail to identify genuinely harmful content. In this study, we evaluate Detoxify [4], a widely used open-source toxicity detection tool that is also reused as an evaluation signal in recent detoxification research [1, 6]. This research aims to answer whether Detoxify, developed primarily for general toxicity detection, can transfer reliably to more sensitive and socially consequential domains.

Methods. We evaluate Detoxify using four datasets. Two are general AI safety datasets: ToxicChat [5] and Aegis AI Content Safety Dataset 2.0 [3]. Two are more domain-sensitive datasets: MindGuard [2], which focuses on mental health safety, and a multilingual refugee hate speech dataset, hateRADAR-es [7]. We compare three Detoxify variants (original, unbiased, and multilingual). To make the comparison consistent, we build a unified evaluation pipeline and assess

Authors' Contact Information: Shuting Xie, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto, Ontario, Canada, shuting.xie@torontomu.ca; Vaishali Meyappan, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto, Ontario, Canada, vaishali.meyappan@torontomu.ca; Hamed Karimi, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto, Ontario, Canada, hamed.karimi@torontomu.ca; Hirad Daneshvar, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto, Ontario, Canada, hirad.daneshvar@torontomu.ca; Mohammad Amin Kamaleddin, Interventional Psychiatry Program, St. Michael's Hospital, Unity Health, Toronto, Ontario, Canada, amin.kamaleddin@mail.utoronto.ca; Reza Samavi, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto, Ontario, Canada and Vector Institute, Toronto, Ontario, Canada, samavi@torontomu.ca; Venkat Bhat, Interventional Psychiatry Program, St. Michael's Hospital, Unity Health, Toronto, Ontario, Canada and Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada, venkat.bhat@utoronto.ca; Roberto Sassi, Child and Adolescent Division, Department of Psychiatry, University of British Columbia, Vancouver, British Columbia, Canada, roberto.sassi@cw.bc.ca.

both ranking performance (AUROC and AUPR) and threshold-based error profiles under a high-recall operating point (precision, recall, F1, false positive rate, and false negative rate). For threshold-based metrics, we use a development set when available; otherwise, we create a held-out split for threshold selection. This design allows us to examine both overall classification quality and the behaviour of the tool in settings where high recall is important. In addition to quantitative evaluation, we also conduct qualitative error analysis to identify recurring false positive patterns in domain-sensitive settings.

Preliminary Findings. Across datasets, performance varies substantially by domain and model variant. On the refugee-related hate speech dataset (hateRADAR-es), the multilingual model achieves an AUROC of 0.83, substantially outperforming the original model (0.67), underscoring the importance of model choice for multilingual content. In contrast, on the mental health dataset (MindGuard), all variants show markedly degraded performance. Even the best-performing model reaches an AUROC of only 0.77, with precision as low as 4% at the 95% recall operating point, meaning over 95% of flagged toxic texts are false positives and false-positive rates exceeding 75%. This pattern illustrates that high recall, often treated as a safety desideratum, systematically generates a flood of false alarms in domains where the language of distress, help-seeking, and identity does not resemble conventionally toxic text. Also, false positives are frequently triggered by surface lexical cues, such as colloquial language, identity-related terms, or informal register, rather than actual harmful intent, suggesting that Detoxify relies on shallow pattern matching rather than contextual understanding.

Future Direction and Suggestions. To address this shortcoming, we need to investigate how integrating uncertainty metrics can enhance the identification of trustworthy Detoxify predictions, thereby improving the guardrail's reliability. By examining the relationship between toxicity scores and classifier uncertainty, we can assess whether uncertainty signals help flag cases in which the model struggles with borderline or context-dependent responses. Ultimately, the aim is to show that an uncertainty-aware approach provides a more reliable means of interpreting scores, enabling users to distinguish genuine toxicity from outputs that warrant skepticism.

Relevance to Bridging Divides. This project is closely related to the Bridging Divides theme because it studies the responsible use of advanced technologies in socially sensitive domains. It also speaks directly to concerns about algorithmic bias, exclusion, and digital inequality. In migration-related and healthcare-related settings, an unreliable toxicity detector may silence vulnerable users, misclassify help-seeking language, or unfairly flag multilingual and identity-linked expressions. By testing a widely used open-source safety tool on both general and domain-specific datasets, this study shows why safety tools should be carefully validated before deployment or reuse in research involving vulnerable populations.

References

[1] Zachary Coalson et al. 2025. If-guide: Influence function-guided detoxification of llms. *arXiv preprint arXiv:2506.01790* (2025).

[2] António Farinhas et al. 2026. MindGuard: Guardrail Classifiers for Multi-Turn Mental Health Support. arXiv:2602.00950 [cs.AI] https://arxiv.org/abs/2602.00950

[3] Shaona Ghosh et al. 2025. AEGIS2.0: A Diverse AI Safety Dataset and Risks Taxonomy for Alignment of LLM Guardrails. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 5992–6026. doi:10.18653/v1/2025.naacl-long.306

[4] Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

[5] Zi Lin et al. 2023. ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation. arXiv:2310.17389 [cs.CL]

[6] Huimin Lu et al. 2025. Unidetox: Universal detoxification of large language models via dataset distillation. *arXiv preprint arXiv:2504.20500* (2025).

[7] J. Mata et al. 2025. From data to detection: Developing a corpus and training language models for the identification of anti-refugee narratives in Spanish. *Array* (2025), 100526. doi:10.1016/j.array.2025.100526

Adaptive Conformal Semantic Uncertainty for Responsible AI

HAMED KARIMI, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Canada
VAISHALI MEYAPPAN, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Canada
REZA SAMAVI, Department of Electrical and Computer Engineering, Toronto Metropolitan University; Vector Institute, Canada

The full version of this study is appearing in the Proceedings of the 35th International Joint Conference on Artificial Intelligence (IJCAI 2026) and is available at *https://arxiv.org/abs/2605.04295*.

Research Motivation. Large Language Models (LLMs) are increasingly deployed in settings where their outputs can influence consequential decisions, yet they remain prone to hallucination and overconfidence. This gap between apparent fluency and actual reliability poses a central challenge for responsible AI, especially in high-stakes domains such as healthcare, law, education, scientific assistance, and socially consequential settings including migration and public-interest decision support [11, 14, 17, 18]. A core requirement in such settings is not merely strong average performance, but the ability of a system to recognize unreliable outputs and respond through abstention, fallback, escalation, or human oversight [3]. This is especially important in citizenship, migration, and social-good settings, where misleading or overconfident outputs may affect access to services, legal pathways, eligibility decisions, or support for vulnerable populations.

Semantic Uncertainty. This abstract argues that *semantic uncertainty*, rather than token-level uncertainty alone, plays a central role in the development of trustworthy generative AI. We build this position around Adaptive Conformal Semantic Entropy (ACSE), a post-hoc method for pretrained LLMs that measures uncertainty from the semantic dispersion of multiple sampled responses using *Conformal Prediction (CP)* [1, 20] and supports selective answering with finite-sample guarantees under exchangeability [9, 10, 16]. Existing uncertainty signals based on token entropy, sequence likelihood, or local decoding statistics can be useful, but they often fail to distinguish superficial lexical variability from genuine uncertainty over meaning [4, 5, 7, 15, 21, 22]. In open-ended generation, an LLM may assign high probability to responses that use different wording but express the same meaning. It may also produce responses with conflicting meanings while still appearing confident at the token level. From a responsible-AI perspective, the latter failure case is particularly concerning so that a system may appear confident while actually being semantically unstable. Semantic-level uncertainty directly targets this failure case by asking whether repeated generations express one meaning or several incompatible meanings [10].

Method Overview. ACSE operationalizes this idea by first sampling multiple responses for a prompt under a fixed decoding policy, then mapping these responses into a sentence-embedding space [2, 6, 13]. The embedded responses are hierarchically clustered using semantic proximity to form meaning-level groups [12]. Soft cluster assignments are then used to estimate a distribution over semantic clusters, from which a normalized semantic entropy score is computed. This score captures whether the model repeatedly returns the same meaning or instead spreads probability mass across multiple incompatible meanings. Low entropy corresponds to semantic agreement; high entropy corresponds to semantic dispersion and thus greater uncertainty. However, semantic entropy alone is not enough for responsible deployment. A model can exhibit a misleading consensus in which most sampled outputs agree on the same answer, yet that answer is incorrect, weakly supported, or structurally unstable. ACSE addresses this by introducing an adaptive inflation mechanism that increases uncertainty when the semantic structure of the response set appears brittle. The inflation factor is driven by several interpretable cluster-level features: (1) semantic entropy, (2) the distance between the dominant response and its cluster centroid, (3) the internal dispersion of the dominant cluster, (4) the size of the dominant cluster, and (5) the unwarranted confidence. These features penalize incorrect consensus, unstable cluster assignments, and fragile semantic support. The resulting

Authors' Contact Information: Hamed Karimi, hamed.karimi@torontomu.ca, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto, Ontario, Canada; Vaishali Meyappan, vaishali.meyappan@torontomu.ca, Department of Electrical and Computer Engineering, Toronto Metropolitan University, Toronto, Ontario, Canada; Reza Samavi, samavi@torontomu.ca, Department of Electrical and Computer Engineering, Toronto Metropolitan University; Vector Institute, Toronto, Ontario, Canada.

adjusted uncertainty score is therefore more conservative exactly in settings where semantic confidence is likely to be misleading. To convert this uncertainty score into an actionable reliability mechanism, ACSE applies conformal calibration on a held-out calibration set. This yields a threshold that supports two complementary decision rules. At the prompt level, the system answers only when the adjusted semantic uncertainty is below the calibrated threshold; otherwise it abstains. At the response level, it can also construct conformal prediction sets over sampled outputs, allowing the system to surface one or several semantically plausible responses rather than forcing a single answer. This is particularly attractive in responsible-AI settings since it turns uncertainty into a runtime control signal. Thus, the model can use calibrated semantic uncertainty to decide whether to answer, defer, retrieve evidence, request clarification, or escalate to a human.

Responsible AI Applications. We view ACSE as a useful foundation for responsible AI applications where systems must assess semantic reliability, abstain under uncertainty, and support safer downstream decision-making, including citizenship, migration, and other social-good settings. First, for trustworthy AI agents, ACSE provides a model-agnostic post-hoc safety component that can perform downstream actions. Before an agent executes a tool, issues a recommendation, or returns a sensitive answer, it can assess whether its semantic uncertainty is sufficiently low to justify action. Second, for reliability assessment, ACSE offers a more decision-relevant notion of uncertainty for generative systems than token-level metrics alone, since it is anchored in variation over meaning rather than only over words. Third, in high-stakes deployment, calibrated abstention is often preferable to unsupported certainty. This is particularly relevant in citizenship, migration, and social-good applications, where ACSE can flag semantically unstable responses, support deferral to human caseworkers or verified resources, and reduce harm from confident but unsupported guidance. Fourth, to detect and mitigate hallucination in LLM systems, ACSE treats hallucinations not merely as low-probability events, but as semantically brittle ones. This makes uncertainty actionable so that unstable cases can be abstained from, routed to retrieval, or presented with explicit caution. Empirically, ACSE improves hallucination detection across benchmarks and LLMs; for example, it achieves an AUROC of 0.88 versus 0.65 for Token Entropy (TE) [19], and improves over Conformal Abstention Policy (CAP) [16] from 0.80 to 0.88 on TriviaQA dataset [8]. As a result, ACSE is relevant not only to question answering, but also to summarization, decision support, report generation, and agentic pipelines where unsupported outputs can propagate downstream errors. Fifth, ACSE is also applicable for verification and auditing of AI systems. Rather than reducing reliability to a single score, ACSE breaks uncertainty into several clear semantic and structural signals such as whether the outputs express different meanings or whether the apparent agreement is supported by enough samples. These signals help explain why the system answered or abstained, which is useful for monitoring, analyzing, and improving failures in responsible AI settings. Sixth, ACSE can support human-AI collaboration in decision-making, including case review and decision support in citizenship and migration workflows. A calibrated uncertainty threshold can prioritize outputs into different action regimes: low-uncertainty responses may be returned directly, moderate-uncertainty cases may be shown with alternatives or warnings, and high-uncertainty cases may be escalated for human review. This is especially useful in settings where humans should remain in the loop, and semantic uncertainty becomes not just a diagnostic, but a practical mechanism for structuring interaction between the user and the model.

Limitations and Position. From a data quality and distribution-shift perspective, ACSE can reveal when prompts from a new environment induce unstable or fragmented semantic behavior. Persistent increases in semantic dispersion or inflation due to brittle clusters may serve as indicators of distribution shift, inadequate calibration data, or task mismatch. This suggests a role for ACSE in post-deployment monitoring and dataset diagnostics by not just identifying when a model is uncertain, but helping identify where the deployment context differs from the calibration regime. A responsible-AI perspective also requires acknowledging limitations. Conformal guarantees depend on exchangeability between calibration and deployment data which can be weakened by significant distribution shifts. ACSE also adds inference cost because it requires multiple samples, embedding, and clustering. Although this may be acceptable in safety-critical settings, it remains a practical deployment constraint. More broadly, uncertainty quantification alone is not enough for responsible AI. It must be part of a larger framework that defines acceptable risk, fallback actions, human oversight, and accountability. ACSE should therefore be viewed as one important reliability component, not a complete solution. Our position is that responsible generative AI requires uncertainty measures such as ACSE that reflect meaning, support action, and are properly calibrated. This makes it a promising direction for safer and more trustworthy LLM deployment, especially in applications where overconfident errors are costly and abstention is better than unsupported certainty.

References

[1] Anastasios N Angelopoulos, Stephen Bates, et al. 2023. Conformal prediction: A gentle introduction. *Foundations and trends® in machine learning* 16, 4 (2023), 494–591.

[2] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[3] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45.

[4] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. *arXiv preprint arXiv:2307.01379* (2023).

[5] Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. *arXiv preprint arXiv:2403.04696* (2024).

[6] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).

[7] Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236* (2023).

[8] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, 1601–1611. doi:10.18653/v1/P17-1147

[9] Ramneet Kaur, Colin Samplawski, Adam D. Cobb, Anirban Roy, Brian Matejek, Manoj Acharya, Daniel Elenius, Alexander M. Berenbeim, John A. Pavlik, Nathaniel D. Bastian, and Susmit Jha. 2024. Addressing Uncertainty in LLMs to Enhance Reliability in Generative AI. arXiv:2411.02381 [cs.AI] https://arxiv.org/abs/2411.02381

[10] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664* (2023).

[11] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-radiology* 1, 2 (2023), 100017.

[12] Fionn Murtagh and Pedro Contreras. 2017. Algorithms for hierarchical clustering: an overview, II. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7, 6 (2017), e1219.

[13] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[14] Murray Shanahan. 2024. Talking about large language models. *Commun. ACM* 67, 2 (2024), 68–79.

[15] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *Comput. Surveys* (2025).

[16] Sina Tayebati, Divake Kumar, Nastaran Darabi, Dinithi Jayasuriya, Ranganath Krishnan, and Amit Ranjan Trivedi. 2025. Learning Conformal Abstention Policies for Adaptive Risk Management in Large Language and Vision-Language Models. *arXiv preprint arXiv:2502.06884* (2025).

[17] Timm Teubner, Christoph M Flath, Christof Weinhardt, Wil Van Der Aalst, and Oliver Hinz. 2023. Welcome to the era of chatgpt et al. the prospects of large language models. *Business & Information Systems Engineering* 65, 2 (2023), 95–101.

[18] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[20] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world.* Springer.

[21] Jiayu Yao, Weiwei Pan, Soumya Ghosh, and Finale Doshi-Velez. 2019. Quality of uncertainty quantification for Bayesian neural network inference. *arXiv preprint arXiv:1906.09686* (2019).

[22] Liz Yarie, Dominic Soriano, Leonard Kaczmarek, Benjamin Wilkinson, and Eduardo Vasquez. 2024. Mitigating token-level uncertainty in retrieval-augmented large language models. *Authorea Preprints* (2024).

JDS

Abstract Article

A Comparative Analysis of Key Concepts of Trustworthy AI in Healthcare

MICHAEL S. RAMIREZ CAMPOS, McMaster University, Canada
THOMAS E. DOYLE, McMaster University, Canada

Abstract

Artificial intelligence (AI) has become a highly relevant tool applied to range of fields and applications. In healthcare, AI has been implemented for tasks that involve diagnosis, prognosis, treatment, and/or monitoring. Although AI has presented promising results, there are some concerns among healthcare stakeholders regarding the challenges of understanding how AI models decisions. As a result, the AI is often considered as a 'black-box' component. The trustworthy AI field arises to address this problem that affect stakeholders' trust. Nevertheless, this is a field that is still evolving, and for some key concepts there is still no single widely accepted definition, which makes it difficult to define both the challenges and the solutions thereby hindering progress in research. In the present study, we compared definitions collected from 41 literature review articles identified through a PRISMA methodology, which examined trustworthy AI in different healthcare fields using a thematic synthesis approach. We analyzed definitions of seven relevant concepts: interpretability, explainability, transparency, fairness, validity, reliability, and robustness, while including bias as a reference concept. Initially, we found that in the literature, interpretability, explainability, and transparency are often used interchangeably, and that there is no agreement on other definitions. We then performed a thematic synthesis (qualitative) and a semantic analysis (quantitative) to further analyze these definitions. Our analysis revealed consistent definitional patterns indicating that interpretability, explainability, and transparency can be defined as distinct concepts, even though they are closely related. On the other hand, we identified inconsistencies in the definitions of bias, despite it being included only for reference, as well as in fairness, reliability and validity. For instance, authors define reliability as the consistency or stability of a model's performance; however, some state that this is tested under similar conditions as the training setting, while others under different conditions. Similarly, in the case of fairness, some studies define it as a conceptual property, whereas others describe it in terms of the outcomes or behaviors that a model should achieve. These preliminary findings indicate that although interest and research in this field are on the rise, there is no general consensus on establishing a theoretical framework, and related research is limited. Our next step will be to propose clear definitions for future studies.

Keywords: Explainability, Fairness, Interpretability, Reliability, Robustness, Transparency, Trustworthy AI, Validity

Authors' Contact Information: Michael S. Ramirez Campos, ramirm7@mcmaster.ca, McMaster University, Hamilton, On, Canada; Thomas E. Doyle, McMaster University, Hamilton, On, Canada, doylet@mcmaster.ca.

Trust from Below: Lived Experience and Human Rights in Designing Trustworthy AI for Migration

Dr. Murtaza Mohiqi
Assistant Professor at the School of Business and Law, University of Agder, Norway

Abstract

As artificial intelligence (AI) systems become increasingly embedded in migration governance, asylum procedures, visa processing, and public service delivery, questions surrounding trust, legitimacy, and accountability have become central to contemporary debates on responsible AI. Existing frameworks for trustworthy AI, however, remain predominantly shaped by top-down technical and regulatory approaches that prioritize abstract principles such as fairness, transparency, and explainability, often without adequately engaging with the lived realities of migrants and other vulnerable communities. This paper argues that such models risk producing institutional forms of trust that fail to resonate with those most directly affected by automated decision-making systems.

The paper develops the concept of "trust from below" as a human rights-oriented framework for understanding and evaluating trustworthy AI in migration contexts. Rather than treating trust as a purely technical outcome or regulatory benchmark, the paper conceptualizes trust as a socially situated and relational process shaped by procedural justice, historical marginalization, linguistic accessibility, and the ability of affected individuals to challenge or contest automated decisions. In this respect, distrust toward AI systems is understood not merely as a matter of perception, but as a reflection of deeper structural inequalities embedded within digital governance infrastructures.

To ground this argument in contemporary practice, the paper examines emerging uses of AI-assisted risk assessment systems, biometric border technologies, and automated asylum processing tools within European and transnational migration governance contexts. These developments illustrate how algorithmic systems are increasingly shaping administrative decision-making processes and raise important questions regarding procedural fairness, transparency, accessibility, and the ability of affected individuals to meaningfully understand and challenge automated outcomes in legally and technologically complex environments.

Building on interdisciplinary scholarship in human rights law, migration governance, critical AI studies, participatory design, and digital constitutionalism, the paper argues that existing universalized models of trustworthy AI frequently overlook context-specific vulnerabilities across different socio-legal settings. Drawing on the author's comparative teaching, workshop, and research engagements across institutions in both the Global North and Global South, the paper highlights how trust in AI systems is shaped by differing institutional capacities, regulatory cultures, and histories of exclusion.

In response, the paper proposes a human rights-based framework grounded in three interrelated principles: participatory inclusion of affected communities throughout the AI lifecycle; context-sensitive fairness and accountability mechanisms; and accessible avenues for contestation, explanation, and redress. Particular attention is given to the importance of culturally and linguistically responsive design practices, especially in migration settings characterized by legal precarity, informational asymmetries, and unequal access to digital literacy.

By shifting the focus from formal compliance toward socially embedded trust-building, the paper contributes to ongoing debates on trustworthy and responsible AI by advancing a normative and practice-oriented reorientation: from designing systems presumed to be trustworthy in principle toward fostering trust as a lived, participatory, and institutionally accountable outcome. Ultimately, the paper argues that trustworthy AI in migration governance requires not only technical safeguards and regulatory oversight, but also sustained engagement with human dignity, social legitimacy, and justice in high-stakes decision-making environments.

Keywords: Trustworthy AI, Human Rights, Migration, Digital Inclusion, Algorithmic Justice

Leveraging Large Language Models to Map Immigrant-Serving Organizations

Data Construction, Organizational Visibility, and Implications for Digital Governance

Pedro Seguel†
Ted Rogers School of Management
Toronto Metropolitan University
Toronto, ON. Canada
pedro.seguel@torontomu.ca

Tharindu Yakkala A. Don
Department of Data Science
Toronto Metropolitan University
Toronto, ON. Canada
tyakkala@torontomu.ca

ABSTRACT

Governments, researchers, and nonprofit sector actors increasingly rely on administrative and open datasets to understand service ecosystems and support evidence-based policymaking. However, many civil society organizations, particularly immigrant-serving nonprofits and charities, remain difficult to identify within these datasets due to fragmented registries, inconsistent classification systems, and the prevalence of generic organizational names. As a result, existing data infrastructures often provide only partial representations of the organizational landscape that supports immigrant integration and community development.

This limitation is especially consequential in the context of migration, where immigrant-serving organizations constitute a critical form of social infrastructure. These organizations support settlement, employment, community integration, and access to services for newcomers and refugees. Recent research has also identified the settlement sector as an important but underexplored domain for responsible AI applications and digital infrastructure development [4]. Yet they are often undercounted or inconsistently represented in official datasets, which, in turn, affects research, funding allocation, and policy design. Traditional approaches to identifying such organizations, including manual curation and keyword-based classification, are labor-intensive and limited in their ability to scale across large and heterogeneous datasets [1, 6].

Recent advances in large language models (LLMs) create new opportunities to address this challenge by enabling scalable classification of organizations using unstructured data such as names and web content. At the same time, the growing accessibility of these tools, through user-facing interfaces and low-barrier workflows, means that dataset construction is no longer limited to technical experts. Researchers, policymakers, and practitioners can increasingly generate their own classifications using LLM-based tools, raising important questions about how such systems operate in practice and how they reshape the construction of datasets used for decision-making.

This study examines how AI-based classification approaches influence the identification and visibility of immigrant-serving organizations. Specifically, we ask: *How do different classification methods, when applied to distinct data sources, reshape which organizations become visible in large-scale datasets?* Addressing this question allows us to move beyond accuracy-focused evaluation and examine how AI systems participate in the construction of policy-relevant data and organizational visibility, building on perspectives that view datasets and classification systems as socially and technically constructed infrastructures [2].

We draw on three complementary datasets that reflect different ways of counting and representing the third sector in Canada. First, the Immigration, Refugees and Citizenship Canada (IRCC) Settlement Provider Organization (SPO) list provides a curated but partial registry of immigrant-serving organizations. Second, the Canadian Federal Corporations registry includes a broad set of nonprofit entities, but with limited structure and varying data quality. Third, a more recent dataset of registered charities from the Canada Revenue Agency (CRA) offers updated coverage across the sector. Together, these datasets represent distinct "data environments" that vary in terms of signal clarity, recency, and institutional definition of what constitutes an immigrant-serving organization. Similar to recent work identifying uneven AI integration gaps across policy infrastructures [3], these heterogeneous environments also generate different visibility gaps in how immigrant-serving organizations become represented within public datasets.

Methodologically, this study employs a comparative computational approach that integrates manual annotation, multiple classification pipelines, and cross-dataset evaluation, responding to recent calls to examine how machine learning systems interact with heterogeneous data environments rather than evaluating models solely through predictive performance metrics [5]. A labeled dataset is constructed by sampling organizations from various sources and manually annotating them to identify immigrant-serving organizations, thereby ensuring consistent classification criteria. Three classes of classification approaches are implemented: (a) name-based LLM classification using structured prompts; (b) web-augmented LLM classification utilizing a retrieval-augmented generation (RAG) workflow with external web information; and (c) benchmark methods, including

keyword-based dictionaries and traditional machine learning models. This design facilitates comparison between contemporary AI-based techniques and established approaches. In doing so, the study conceptualizes AI-enabled classification workflows as sociotechnical data-construction pipelines whose outputs vary systematically depending on retrieval strategy, signal structure, and contextual augmentation.

Our evaluation proceeds along two dimensions. First, we assess classification performance using standard metrics such as precision and recall. Second, and more critically, we examine how different methods shape the resulting dataset. To do so, we introduce a comparative analysis of organizational visibility across methods and datasets. We assess (1) coverage, defined as the proportion of known organizations (e.g., from the SPO list) correctly identified; (2) expansion, defined as the number and types of additional organizations identified beyond official lists; and (3) overlap, capturing the extent to which different methods identify the same or distinct sets of organizations. Finally, we include a benchmark dataset of nonprofit organizations from New York used in prior research, which provides an externally validated labeled sample for comparison. This design allows us to empirically contrast how organizations were previously identified through curated lists and keyword-based approaches with how they are identified through AI-enabled workflows.

The findings indicate that classification outcomes are highly contingent on the signal structure of the underlying data. In keyword-rich datasets, simpler methods such as dictionaries and name-based models demonstrate strong performance. Conversely, in low-signal environments like the SPO list, where organizational names are often generic, web-augmented LLM approaches significantly enhance recall by incorporating additional contextual information, effectively creating new interpretive metadata layers that increase the discoverability and representation of organizations within public datasets. Notably, when applied at scale, different classification methods yield systematically distinct sets of organizations. Web-based approaches identify a broader spectrum of organizations do not present in official lists, including community-based, cultural, and faith-oriented groups that provide relevant services but are excluded from formal classification schemes.

These results indicate that AI-based classification systems not only enhance data quality but also actively reshape the visibility of organizations within datasets. Depending on design choices, such as the inclusion of web context, different segments of the third sector become observable. This finding underscores that datasets used for research and policy are not neutral representations of reality; rather, they are shaped by the tools and assumptions underlying their construction.

1 Relevance for the workshop

This research addresses emerging challenges in the use of AI for public and policy-relevant data. As LLM-based tools become more accessible, a broader range of stakeholders can construct datasets with minimal technical expertise. While this development expands opportunities for knowledge generation and improved coverage, it also raises critical questions regarding trust, transparency, and accountability in data generation and use. The findings demonstrate that different AI-based approaches can produce systematically divergent representations of the same organizational field, thereby shaping what is counted, recognized, or rendered invisible in policy-relevant data. These variations may influence funding allocation, service coordination, and broader understandings of community needs.

By illustrating the mutual shaping of data and AI systems, this study contributes to research on data management and analytics and highlights significant implications for digital governance. More broadly, the study shows how AI-enabled classification pipelines increasingly function as governance-relevant infrastructures that shape what organizations become visible, measurable, and actionable within public data ecosystems.

ACKNOWLEDGMENTS

This research was undertaken thanks in part to funding from the Canada First Research Excellence Fund (CFREF), Migrant Integration in the Mid-21st Century: Bridging Divides.

REFERENCES

[1] Bloemraad, I., Chaudhary, A.R. and Gleeson, S. 2022. Immigrant Organizations. *Annual Review of Sociology*. 48, Volume 48, 2022 (July 2022), 319–341. https://doi.org/10.1146/annurev-soc-030420-015613.

[2] Kitchin, R. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

[3] Lnenicka, M., Clarinval, A., Nikiforova, A., Rudmark, D., Luterek, M., Symeonidis, D. and Bolívar, M.P.R. 2026. Artificial intelligence in policymaking: Mapping integration gaps across the public policy cycle. *Government Information Quarterly*. 43, 2 (2026), 102138.

[4] Nejadgholi, I., Molamohammadi, M., Missaghi, K. and Bakhtawar, S. 2024. Human-centered AI applications for Canada's immigration settlement sector. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2024), 1036–1050.

[5] Padmanabhan, B., Fang, X., Sahoo, N. and Burton-Jones, A. 2022. Editor's Comments: Machine Learning in Information Systems Research. *Management Information Systems Quarterly*. 46, 1 (Mar. 2022), iii–xix. https://doi.org/10.25300/MISQ/2022/461E1.

[6] Ren, C. and Bloemraad, I. 2022. New Methods and the Study of Vulnerable Groups: Using Machine Learning to Identify Immigrant-Oriented Nonprofit Organizations. *Socius: Sociological Research for a Dynamic World*. 8, (Jan. 2022), 237802312210769. https://doi.org/10.1177/23780231221076992.

HARDWARE-LEVEL UNCERTAINTY QUANTIFICATION WITH FAIRNESS VALIDATION: BALANCED TERNARY LOGIC FOR TRUSTWORTHY AI IN MIGRATION SYSTEMS

Kagan Katran
Department of Electrical Engineering, Toronto Metropolitan University

ABSTRACT

AI systems deployed in migration and citizenship contexts make high-stakes decisions with incomplete or ambiguous data. Current binary architectures force classification outputs even when uncertainty should trigger human review. This paper proposes a conceptual framework combining balanced ternary hardware ($T \in \{-1, 0, 1\}$) with fairness validation under model quantization to enforce deterministic uncertainty handling in migration AI systems. We outline a three-phase simulation roadmap to evaluate (1) fairness degradation from ternary weight quantization across demographic groups, (2) hardware-level rejection rates for ambiguous cases, and (3) comparative validation against existing selective classification baselines.

1. INTRODUCTION

Migration AI systems—deployed for visa classification, residency determination, and asylum assessment—operate on incomplete demographic and documentary data. When models encounter ambiguous inputs, current binary architectures force a binary classification (Approve/Deny) rather than escalating to human review. This creates two interconnected problems:

1. **Forced Classification on Uncertainty:** Missing documentation, unclear status, or borderline cases produce overconfident predictions that lack inherent rejection mechanisms.

2. **Fairness Degradation Under Efficiency:** To reduce computational footprint, migration AI models are quantized (compressed to lower precision). However, quantization can disproportionately degrade accuracy for minority groups, amplifying existing disparities in classification outcomes.

This paper proposes hardware-level ternary logic as a structural constraint to address both problems. Unlike software-layer confidence thresholds, a hardware-native third state (0 = Unknown) provides a deterministic trigger for mandatory human review. Critically, we frame this within a fairness validation framework: if migration AI models are quantized to ternary weights for efficiency, we must verify that quantization does not introduce or amplify bias against protected groups.

2. TECHNICAL APPROACH

2.1 Ternary Hardware for Uncertainty

Balanced ternary logic uses three native states: −1 (Deny), 0 (Unknown), +1 (Approve). When model inputs lack sufficient statistical weight to drive output toward ±1, the hardware outputs 0, triggering mandatory human-in-the-loop (HITL) review. The distinction from software thresholding: a software system can compute a confidence score and conditionally route to human review. A hardware-level 0-state provides physical verification of the rejection mechanism—auditable as a discrete, deterministic output state rather than a learned threshold buried in model weights. Implementation requires modified CMOS design to reliably distinguish three voltage levels. Recent work (Mouftah et al., 1985; Huawei patents, 2023) demonstrates feasibility through voltage-threshold biasing, though manufacturing precision remains a practical constraint.

2.2 Fairness Under Quantization

If migration AI models are quantized to ternary weights (−1, 0, +1) for memory/energy efficiency, accuracy loss is not uniform across demographic groups. Quantization can widen fairness gaps if training data is imbalanced or if certain groups have learned weights sensitive to precision loss. We propose validating three fairness metrics: (i) Per-group accuracy—does ternary quantization degrade accuracy for Group A more than Group B? (ii) Rejection rate parity—when hardware outputs 0 (Unknown), are certain groups over-represented? (iii) Downstream consequence—do flagged cases for human review correlate with sensitive attributes? This is distinct from prior work on selective classification, which does not systematically address fairness under quantization.

3. SIMULATION ROADMAP

Phase 1 (Weeks 1–2): Fairness Audit Baseline. Dataset: 10,000–100,000 migration decision records. Model: train a baseline classifier. Metrics: measure accuracy by demographic group. Output: fairness gap baseline (e.g., Accuracy = 92%, Accuracy = 81%, gap = 11%).

GroupA GroupB

Phase 2 (Weeks 3–5): Quantization Impact Simulation. Quantize model weights to ternary values using standard methods. Re-evaluate accuracy per demographic group. Measure: how does fairness gap change? Typical expectation: ternary quantization causes 5–15% accuracy loss; fairness gap may widen 1–5%. Output: table showing accuracy before/after quantization.

Phase 3 (Weeks 6–8): Hardware Rejection Behavior Simulation. Simulate ternary hardware thresholds: cases with confidence below τ output 0. For each demographic group, measure % of cases triggering 0. Analyze: are certain groups disproportionately flagged? Test multiple thresholds to find fairness-utility tradeoff. Output: decision matrix showing rejection rate and accuracy by group.

4. RELATED WORK

Selective Classification (Chow, 1970; Guo et al., 2020) allows models to abstain when uncertain but relies on software enforcement; fairness parity is understudied. Conformal Prediction (Vovk et al., 2005; Barber et al., 2021) provides finite-sample guarantees but requires post-hoc calibration. Evidential Deep Learning (Amini et al., 2020) models uncertainty via Dirichlet distributions but does not enforce hardware-level constraints. Ternary Quantization (Zhou et al., 2016; Polino et al., 2018) compresses weights for efficiency but does not analyze fairness degradation. **Our contribution:** we combine hardware-level ternary logic with systematic fairness validation under quantization, addressing both trustworthiness and equity.

5. NEXT STEPS

Immediate: Implement Phase 1–3 simulation on public migration dataset or synthetic data. Compare rejection rate and fairness outcomes against conformal prediction and selective classification baselines. Specify hardware interface: how do neural network outputs map to hardware ternary gates?

Longer-term: Physical prototype of CMOS ternary gates (Mouftah design rules). Validation on real-world migration AI system. Policy analysis: how does hardware-enforced HITL affect reviewer workflow and system outcomes?

6. CONCLUSION

This paper frames trustworthy AI in migration systems as an intersection of uncertainty handling and fairness under efficiency constraints. We propose ternary hardware as a structural mechanism to enforce human oversight while validating that quantization-driven efficiency does not amplify bias. Our three-phase roadmap provides concrete, measurable steps to evaluate this approach and identify fairness-utility tradeoffs.

REFERENCES

[1] Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2021). Predictive inference with the jackknife. *arXiv* preprint arXiv:1905.02928.

[2] Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Trans. Inform. Theory*, 16(1), 41–46.

[3] Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world.* Springer.

[4] Zhou, S., Ni, Z., Zhou, X., Wen, H., Wu, Y., & Zou, Y. (2016). DoReFa-Net: Training low bitwidth networks. *arXiv* preprint arXiv:1606.06160.

[5] Asianometry. (2025, December 18). Ternary computing: Theoretically better than binary [Video]. YouTube. https://www.youtube.com/watch?v=sWKyrAXxzGA

A socio-technical perspective on the labour-market implications of an Artificial Intelligence (AI) centered-economy

Virginia Kanyogonya, PhD
Toronto, ON

ABSTRACT

The growing rise of an artificial intelligence (AI) centered-economy is understandably causing immense anxiety for many workers and job seekers. Over the past few years, analysts have asserted that it remains unclear to what extent AI will disrupt the knowledge economy, with some suggesting that augmentation is more likely than complete replacement. While this may be plausible, there is still significant uncertainty about the future of work for high-exposure, low-complementarity jobs (HELC)—meaning jobs in which AI can perform most tasks and are vulnerable to AI-driven replacement.

This position paper presents a socio-technical perspective on the labour-market implications of an AI-centered economy. I approach this through the lens of 'meaningful employment', arguing that, as AI increasingly shapes our labour market, in addition to mapping out highly exposed occupations, it is imperative to assess the current restructuring of employment in a growing environment of rising austerity measures, hiring freezes, layoffs, precarious work and involuntary independent contracting.

Against this backdrop, I, therefore, differentiate 'meaningful work' from 'meaningful employment' as evidence shows that one can have meaningful work but not be meaningfully employed. I discuss why this differentiation matters in an AI-centered economy, not only for workers but also for policymakers, labour advocates, workforce development practitioners, mission-driven leaders and organizations committed to championing and providing meaningful employment opportunities.

Research shows that many racialized internationally trained professionals (ITPs) have historically been undervalued in Canada's labour market, many of whom primarily take up jobs in the HELC category. For instance, for several years, ITPs have dominated code-intensive jobs, which have now become highly susceptible to automation. This is also evidenced by the reduction of these occupations in Canada's most recent Express Entry requirements. That said, we are still unclear how work will be transformed for ITPs working in code-intensive jobs in an emerging AI-centered economy.

This paper, however, focuses more on employment conditions than on occupations; hence, the distinction between 'meaningful work' and 'meaningful employment'.

I argue that to ensure this technology serves the public/social good, it is not enough to focus only on AI occupational exposure, for it is equally important not to overlook the employment conditions currently taking shape in an AI-centered economy, particularly with respect to racialized labour that continues to be undervalued and exploited across the globe. Case in point, empirical evidence shows us that the deeply concerning working conditions of data labelers (i.e., an invisible foundational workforce of generative AI) in the global majority countries (e.g., Kenya, the Philippines, India, Colombia, etc.) are too significant to ignore. Which begs the question: how do we ensure that an AI-centred economy does not result in a further entrenchment of inequities?

While there is evidence of numerous benefits of AI worth celebrating, this should not preclude us from acknowledging the emerging contradictions that have surfaced in the advent of frontier AI. The fact that the godfathers of AI are also sounding the alarm on some of the societal risks and harms that have not been fully accounted for in this current intense race to accelerate frontier AI and the potential development of artificial general intelligence (AGI) is quite telling.

Furthermore, given that one of the primary principles of responsible technology is to do no harm, it is imperative that the development and deployment of AI be critically analyzed within the broader systems in which it currently operates. A meaningful employment lens could contribute to responsible AI efforts by considering the needs of workers, particularly those who have historically been undervalued in our global labour market.

CCS CONCEPTS

Artificial intelligence, human-centered computing, social and professional topics

KEYWORDS

AI-centered economy, frontier AI, meaningful employment, future of work, immigrants, racialized labour, responsible technology, responsible AI

REFERENCES

[1] Sheila Block and Grace-Edward Galabuzi. 2018. Persistent inequality: Ontario's colour-coded labour market. Canadian Centre for Policy Alternatives. https://www.policyalternatives.ca/sites/default/files/uploads/publications/Ontario%20Office/2018/12/Persistent%20inequality.pdf

[2] Jonathan D. Codell, Robert D. Hill, Dan J. Woltz, and Paul A. Gore. 2011. Predicting meaningful employment for refugees: The influence of personal characteristics and developmental factors on employment status and hourly wages. International Journal for the Advancement of Counselling 33, 3 (2011), 216–224. https://doi.org/10.1007/s10447-011-9125-5

[3] Thomas Davenport and Miguel Paredes. 2025. Can we predict what jobs AI will take? Harvard Data Science Review, 7, 4. https://doi.org/10.1162/99608f92.8975ddd1

[4] Grace-Edward Galabuzi. 2006. Canada's economic apartheid: The social exclusion of racialized groups in the new century. (1st edition). Canadian Scholars' Press Inc.

[5] Grace-Edward Galabuzi. 2018. Unequal futures: Race and class under neoliberalism in Ontario. In G. Albo & B. Evans (Ed.), Divided province: Ontario politics in the age of neoliberalism (pp. 461-492). Montreal: McGill-Queen's University Press. https://doi.org/10.1515/9780773555679-017

[6] Government of Canada. 2026. 2024-2025 Report to Parliament – Category-Based Selection in Express Entry. Immigration, Refugees and Citizenship Canada. Retrieved May 10, 2026. https://www.canada.ca/en/immigration-refugees-citizenship/corporate/publications-manuals/report-parliament-cbs-2024-25.html

[7] Timnit Gebru and Émile P. Torres. 2024. The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. First Monday, 29, 4. https://doi.org/10.5210/fm.v29i4.13636

[8] Shibao Guo. 2015. The colour of skill: Contesting a racialised regime of skill from the experience of recent immigrants in Canada. Studies in Continuing Education, 37, 3 (2015), 236–250. https://doi.org/10.1080/0158037X.2015.1067766

[9] Olya Kudina, and Ibo van de Poel. 2024. A sociotechnical system perspective on AI. Minds & Machines 34, 21. https://doi.org/10.1007/s11023-024-09680-2

[10] Till, Leopold, 2025. How AI is reshaping the career ladder, and other trends in jobs and skills on Labour Day. World Economic Forum. https://www.weforum.org/stories/2025/04/ai-jobs-international-workers-day/

[11] Alexandra Mateescu, Aiha Nguyen and Sanjay Pinto. 2026. Last place in the AI-first economy: How the AI industry relies on worker disempowerment. Data & Society. https://datasociety.net/library/last-place-in-the-ai-first-economy/

[12] Matthew Sellers. 2025. Nearly half of Canadian job seekers fear AI could eliminate their jobs: Survey. Canadian HR Reporter. https://www.hrreporter.com/focus-areas/hr-technology/nearly-half-of-canadian-job-seekers-fear-ai-could-eliminate-their-jobs-survey/393638

[13] Maxim Massenkoff and Peter McCrory. 2026. Labor market impacts of AI: A new measure and early evidence. Anthropic. https://www.anthropic.com/research/labor-market-impacts

[14] Tahsin Mehdi. (2025). Generative artificial intelligence and its potential impact on immigrant workers: Uncertainties and opportunities. Statistics Canada. https://p2pcanada.ca/wp-content/blogs.dir/1/files/2025/12/Tahsin-Mehdi.pdf

[15] Tahsin Mehdi and Marc Frenette. 2026. Canadian employment trends in the era of generative artificial intelligence: Early evidence. Economic and Social Reports, 6, 1. Statistics Canada. https://www150.statcan.gc.ca/n1/pub/36-28-0001/2026001/article/00003-eng.htm

[16] Brent Orrell. 2025. De-skilling the knowledge economy. American Enterprise Institute. http://www.jstor.org/stable/resrep71198

[17] Stanford Institute for Human-Centered Artificial Intelligence. 2026. AI Index report 2026. https://hai.stanford.edu/ai-index/2026-ai-index-report

[18] The Diary Of A CEO. 2025. Godfather of AI: They keep silencing me but I'm trying to warn them! [Video]. YouTube. https://www.youtube.com/watch?v=giT0ytynSqg

[19] Tingting Zhang, Rupa Banerjee and Aliya Amarshi. 2023. Does Canada's express entry system meet the challenges of the labor market? Journal of Immigrant & Refugee Studies, 21,1(2023), 104–118. https://doi.org/10.1080/15562948.2022.2133201

RAG-Powered Billing Agent in Healthcare: Transforming Physician Documentation into Accountable Billing Actions

ISHAAN MEHTA, Toronto Metropolitan University, Canada
GARIMA MALIK, Toronto Metropolitan University, Canada

Canada's healthcare system is facing a growing administrative burden that contributes significantly to physician burnout. Ontario physicians now spend nearly 40% of their workweek on paperwork, while medical billing remains especially complex due to the Ontario Health Insurance Plan (OHIP) Schedule of Benefits. This paper presents an Agentic AI framework for automating the medical billing workflow. The proposed system processes unstructured clinical notes from Electronic Medical Record (EMR) systems or voice dictation and uses an Agentic Retrieval-Augmented Generation (RAG) pipeline to map clinical narratives to OHIP billing codes. By reducing manual coding and preventable billing errors, the system aims to reclaim clinical time and improve claim processing efficiency. To support trustworthy deployment, the framework incorporates traceable reasoning and physician-in-the-loop review for uncertain or high-risk claims.

1 Introduction

While Artificial Intelligence (AI) in healthcare is often associated with diagnostics [1], its most immediate value may lie in reducing the administrative burden currently straining health systems [2]. Ontario is facing a primary care crisis; by 2026, an estimated 4.4 million residents will lack a family physician [2]. Research indicates that heavy administrative workloads, rather than just staffing shortages, are the primary cause of physician burnout and attrition [2].

The impact of this workload is significant. In 2025, 46% of Ontario physicians reported high levels of burnout. Family doctors spend nearly 40% of their work week, or roughly 19 hours, on paperwork [2]. Nationally, this overhead is equivalent to losing 55.6 million patient visits annually [2].

Medical billing is a central source of this frustration. The Ontario Health Insurance Plan (OHIP) maintains a Schedule of Benefits with over 6,000 active billing codes. Each code is governed by specific rules regarding patient age, location, and diagnosis [3]. Complexity and unclear rejections are major time consumers for 70% of physicians [2]. Financial losses are also notable; simple errors, such as incorrect health card versions, cost Ontario physicians an average of $10,000 in lost revenue each year [4]. Additionally, 65% of denied claims are never resubmitted because the manual effort required to investigate them is too high [5].

Current industry practices rely on manual entry or third-party clerks, both of which suffer from high turnover rates [6]. Existing automation tools generally function as rigid checkers that flag errors without resolving them [7]. In contrast, Agentic AI uses autonomous systems that employ reasoning logic rather than fixed scripts. This technology enables a hands-off revenue cycle by managing clinical coding and resolving denials independently [8].

This paper proposes a system that processes unstructured clinical narratives from Electronic Medical Record (EMR) integrations or voice dictation. By using an Agentic Retrieval-Augmented Generation (RAG) loop to map these narratives to the OHIP Schedule of Benefits, the system automates the billing process. This approach bridges the gap between clinical documentation and provincial policy, reclaiming physician time and optimizing revenue capture.

2 System Design

Ontario physicians bill the Ministry of Health through OHIP using the Schedule of Benefits for Physician Services [3] (6,000 fee codes across 980 pages). Claims are submitted electronically via MCEDT (Medical Claims Electronic Data

Authors' Contact Information: Ishaan Mehta, ishaan.mehta@torontomu.ca, Toronto Metropolitan University, Toronto, Ontario, Canada; Garima Malik, Toronto Metropolitan University, Toronto, Ontario, Canada, garima.malik@torontomu.ca.

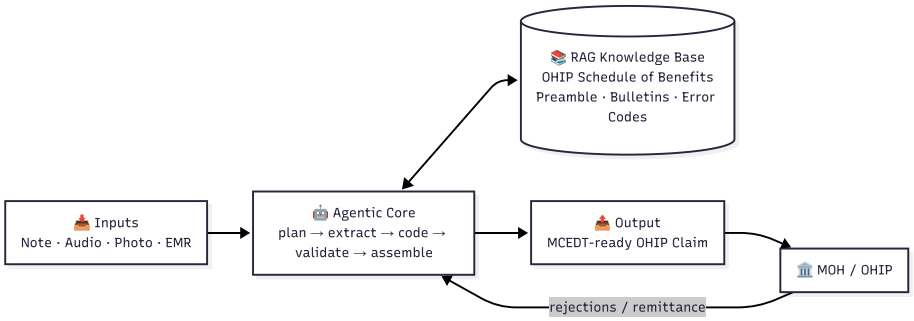

Fig. 1. Proposed system design for RAG-based agentic AI billing framework.

Transfer). We propose a RAG-based modular agentic AI system as depicted in Figure 1. Below, we briefly describe each component of the proposed system.

`Inputs`: The system will accept multiple types of data to fit naturally into existing clinical workflows such as clinical notes and EMR free text, audio dictation, hospital face sheets, and structured EMR data via application interface. This flexibility ensures the system can work with how information is captured in real practice.

`Agent-Based Architecture`: The system consists of a central orchestrator that coordinates specialized agents for different stages of the billing workflow. An Input Processing Agent converts incoming clinical data into a standardized format, while a Coding Agent identifies relevant billing codes and retrieves supporting policy context from the knowledge base. A Validation Agent applies billing rules related to bundling, eligibility, and documentation consistency to reduce errors and flag ambiguous or low-confidence claims for human review. Finally, an Assembly Agent combines all validated components into a complete, submission-ready claim.

`Knowledge Base (RAG)[9]`: The system is powered by a structured knowledge base to ensure accurate and reliable billing decisions that includes Ontario Schedule of Benefits (including preamble and billing rules), diagnostic service guidelines relevant to all medical departments, rules for consultations, time-based billing, premiums, and bundling, error codes and common rejection reasons, and historical claims and rejections from the physician's own practice (for personalization).

`Trustworthy AI and Human Oversight`: Given the high-stakes nature of healthcare billing, the system incorporates human oversight and auditability. Billing recommendations include policy references and reasoning traces for verification by clinical staff. Claims with low confidence or ambiguous billing rules can be flagged for manual review before submission, while traceable decision logs support post-hoc auditing and investigation of disputed claims.

`Output`: The system generates draft OHIP claims with billing rationale, patient/provider details, billing codes, and modifiers. Rejected claims are analyzed to identify errors and suggest resubmission plans, enabling continuous improvement through feedback and human review of ambiguous cases.

3 Conclusion and Future Work

This paper presents an Agentic AI framework for automating medical billing from clinical notes to OHIP codes. By reducing administrative workload, the system aims to mitigate physician burnout and reclaim time for patient care. Future work will focus on building high-quality datasets using expert-annotated real billing data and synthetic data to capture diverse billing patterns. Initial evaluation will target gastroenterology before expanding to family medicine and cardiology to assess robustness across documentation styles and billing rules. We also plan to investigate confidence estimation, physician feedback integration, and transparent auditing for safe and accountable deployment.

References

[1] Peng Lean Chong, Vikneswaran Vaigeshwari, Basir Khan Mohammed Reyasudin, binti Ros Azamin Noor Hidayah, Purnshatman Tatchanaamoorti, Jian Ai Yeow, and Feng Yuan Kong. Integrating artificial intelligence in healthcare: applications, challenges, and future directions. *Future Science OA*, 11(1):2527505, 2025.

[2] Ontario College of Family Physicians. Administrative burden in ontario family medicine. Technical report, OntarioMD / OCFP, 2025. URL https://opsmed.ca/resources/ontario-physician-administrative-burden/. Accessed April 14, 2026.

[3] Government of Ontario. Ohip schedule of benefits and fees, 2026. URL https://www.ontario.ca/page/ohip-schedule-benefits-and-fees. Accessed April 14, 2026.

[4] Dr.Bill. Common medical billing errors in ontario hospitals and how to fix them, 2026. URL https://www.dr-bill.ca/blog/billing-tips/common-medical-billing-errors-in-ontario-hospitals-and-how-to-fix-them. Accessed April 14, 2026.

[5] Peregrine Healthcare. The cost of inaction: How preventable denials quietly drain your practice revenue, 2026. URL https://peregrinehealthcare.com/the-cost-of-inaction-how-preventable-denials-quietly-drain-your-practice-revenue/. Accessed April 14, 2026.

[6] ERI SalaryExpert. Medical billing clerk salary in ontario, canada (2026), 2026. URL https://www.salaryexpert.com/salary/job/medical-billing-clerk/canada/ontario. Accessed April 14, 2026.

[7] CaliberFocus. Best ai agents for healthcare in 2026, 2026. URL https://caliberfocus.com/top-agentic-ai-companies-in-healthcare. Accessed April 14, 2026.

[8] Nalan Karunanayake. Next-generation agentic ai for transforming healthcare. *Informatics and Health*, 2(2):73–83, 2025.

[9] Rajvardhan Patil, Manideep Abbidi, and Sherri Fannon. Ragmed: A rag-based medical ai assistant for improving healthcare delivery. *AI*, 6(10):240, 2025.

Detecting Invisible Stress: A Digital Tool for Newcomer Youth with Autism

Establishing Physiological-Behavioural Discordance in Children with Neurodevelopmental Conditions

Salma Mohamed†
Electrical, Computer, and Biomedical Engineering Department
Toronto Metropolitan University
Toronto, Ontario, Canada
salma.mohamed@torontomu.ca

Karen Soldatic
School of Disability Studies
Toronto Metropolitan University
Toronto, Ontario, Canada
ksoldatic@torontomu.ca

Naimul Khan
Electrical, Computer, and Biomedical Engineering Department
Toronto Metropolitan University
Toronto, Ontario, Canada
n77khan@torontomu.ca

ABSTRACT

Immigrant youth (8 - 14 years) with Autism Spectrum Disorder (ASD) face a distinct “double burden” that intertwines cultural adaptation and neurodiversity-related stigma [1][2]. These children must navigate both ableist social expectations and cultural “code-switching” required to assimilate into a new country [3]. Research suggests that immigrant autistic children of racial or ethnic minority backgrounds experience "Dual Masking" [4], where they suppress their autistic traits to avoid bullying, stigma and exclusion while simultaneously minimizing their cultural identity to avoid discrimination or standing out. This dual masking phenomenon places immigrant autistic youth at a particular risk for internalized distress, which often manifests through “shutdowns” rather than visible “meltdowns” [5]. Shutdowns may appear as calm compliance but are in fact dissociative freeze states characterized by parasympathetic overactivation and vagal withdrawal, leading to long-term physiological dysregulation [5]. Because these responses are largely invisible to educators and caregivers, they frequently go unnoticed until children experience burnout [6].

This project investigates the extent to which behavioural stress labels align with physiological stress responses in children, guided by the hypothesis that neurodevelopmental populations may exhibit physiological-behavioural discordance, where internal stress is not reliably expressed through observable behaviour. As a preliminary exploratory analysis, we used the publicly available AKTIVES dataset, a physiological signal database of children with different special needs for stress recognition [7]. While the dataset does not include immigrant autistic youth directly, it provides a useful foundation for examining whether physiological stress can diverge from externally observed behavioural states in neurodevelopmental populations. The dataset included typically developing (TD) children and children with intellectual disabilities (ID), with synchronized physiological signals from the Empatica E4 wristband (blood volume pulse, electrodermal activity, and skin temperature), alongside behavioural features derived from facial emotion probabilities and expert-labelled stress annotations [7].

Physiological feature extraction and behavioural aggregation enabled direct comparison between internal physiological arousal and externally observed stress labels. Data analysis results for the TD group demonstrated that, physiological signals such as heart rate showed clearer differentiation between stress and no stress conditions, with large effect sizes $(r \approx -0.50, p < 0.001)$. In contrast, the ID children exhibited more variable and less consistent physiological patterns, with weaker alignment between behavioural labels and physiological responses. Notably, spearman correlation analysis revealed a negative relationship between physiological arousal and behavioural calmness in TD group $(\rho \approx -0.24)$, indicating expected alignment. While the ID group showed a positive correlation $(\rho \approx 0.59)$, suggesting that some children appeared behaviourally calm while physiologically stressed. These findings provide empirical support for the discordance hypothesis, indicating that behaviour alone may underestimate stress in neurodevelopmental populations.

AI driven emotion recognition systems, particularly combined with wearable data, offer scalable real-time monitoring of physiological and behavioral signals and have demonstrated high accuracy in controlled settings [8]. However, such models are often trained on homogenous

datasets and may not generalize well to diverse populations. Cultural psychology research shows that emotional expression is shaped by cultural context, meaning that models ignoring these differences can risk misinterpreting signals [9]. This raises concerns related to reliability and bias in automated stress detection systems, especially for immigrant autistic youth whose experiences are shaped by both neurodivergence and sociocultural context [10]. Future work will extend this research to the target group by incorporating machine learning models that leverage personalized physiological baselines and culturally informed behavioural interpretations, aiming to reduce bias and improve sensitivity to hidden distress. This includes considerations of ethical development, transparency in decision-making, and the risk of misclassification in vulnerable populations.

Overall, this work highlights the need for a multimodal, context-aware AI system that integrates physiological and behavioural signals, while also accounting for uncertainty, individual variability, and cultural nuance. Such approaches are essential for advancing equitable, reliable, and socially responsible AI in health and migration-related contexts.

KEYWORDS

Stress detection, multimodal AI, physiological signal analysis, immigrant youth

ACM Reference format:

Salma Mohamed, Karen Soldatic, Naimul Khan. 2026. Detecting Invisible Stress: A Digital Tool for Newcomer Youth with Autism: Establishing Physiological-Behavioural Discordance in Children with Neurodevelopmental Conditions *(TIAI Workshop' 26). ACM, Toronto, ON, CA, 2 pages.* https://doi.org/10.1145/1234567890

REFERENCES

[1] L. Fontil and H. H. Petrakos, "TRANSITION TO SCHOOL: THE EXPERIENCES OF CANADIAN AND IMMIGRANT FAMILIES OF CHILDREN WITH AUTISM SPECTRUM DISORDERS," Psychology in the schools, vol. 52, no. 8, pp. 773–788, 2015, doi: 10.1002/pits.21859

[2] B. Sritharan and M. M. Koola, "Barriers faced by immigrant families of children with autism: A program to address the challenges," Asian journal of psychiatry, vol. 39, pp. 53–57, 2019, doi: 10.1016/j.ajp.2018.11.017

[3] M. M. Akhtar, Girish, O. C. Phukan, and M. Singh, "NeuRO: An Application for Code-Switched Autism Detection in Children," 2024, doi: 10.48550/arxiv.2406.03514

[4] M. S. Benedetto, "The 'Dual Masking Phenomenon': A Critical Autoethnography of an Autistic Multiracial Latina Amid the Presence of White Autistic Researchers and Self-Advocates," Autism in adulthood, 2024, doi: 10.1089/aut.2024.0105

[5] C. I. Lee, "Autistic meltdown vs shutdown: What they are and how to manage them," L.A. Concierge Psychologist, https://laconciergepsychologist.com/blog/autistic-meltdown-shutdown/.

[6] M. A. Rashidan et al., "Technology-Assisted Emotion Recognition for Autism Spectrum Disorder (ASD) Children: A Systematic Literature Review," IEEE access, vol. 9, pp. 33638–33653, 2021, doi: 10.1109/ACCESS.2021.3060753

[7] B. Coşkun et al., "A physiological signal database of children with different special needs for stress recognition," Scientific data, vol. 10, no. 1, Art. no. 382, June 2023, doi: 10.1038/s41597-023-02272-2

[8] D. Nandini, J. Yadav, V. Singh, V. Mohan, and S. Agarwal, "An ensemble deep learning framework for emotion recognition through wearable devices multi-modal physiological signals," Scientific reports, vol. 15, no. 1, Art. no. 17263, 2025, doi: 10.1038/s41598-025-99858-0

[9] H. R. Markus and S. Kitayama, "Culture and the Self: Implications for Cognition, Emotion, and Motivation," Psychological review, vol. 98, no. 2, pp. 224–253, 1991, doi: 10.1037/0033-295X.98.2.224

[10] K. Pryor, T. Coleman, S. Z. Hossain, E. Tootil, and S. Ahmed, "Examining Large Language Models Within Autism-Related Contexts: A Systematic Review of Bias and (Mis) Representation," in Proceedings : annual International Computer Software and Applications Conference, IEEE, 2025, pp. 876–886. doi: 10.1109/COMPSAC65507.2025.00115

Shadows in the Code: Chinook, Algorithmic Bias, and the Erosion of Procedural Fairness in Canadian Immigration

Berik Izbassarov
Immigration and Citizenship Law
Queen's University
Kingston Ontario Canada
24qg27@queensu.ca

ABSTRACT

As Canada rapidly integrates Automated Decision Support Systems (ADSS) into migration management, the tension between administrative efficiency and foundational legal principles has reached a pivotal point. This position paper analyzes the "Chinook" system and its broader artificial intelligence (AI) framework from the perspective of a future Regulated Canadian Immigration Consultant (RCIC). I argue that current practices, particularly the deletion of transitory working notes and reliance on opaque "risk indicators," create a legal "black box" that undermines procedural fairness and restricts the right to meaningful Judicial Review. By examining the shift from human discretion to algorithmic "nudging," I propose essential safeguards to restore public trust and ensure that efficiency does not compromise justice.

These safeguards are operationalized through specific requirements on log retention (including risk indicators and officer overrides), structured explanations to applicants, measurable fairness metrics, and clearly assigned institutional accountability.

1. Introduction: The Era of "Invisible Borders."

Since 2014, Immigration, Refugees and Citizenship Canada (IRCC) has advanced from basic data sorting to advanced analytics. For international students and temporary residents, the "border" now functions as an algorithmic filter rather than solely a physical checkpoint. Although IRCC describes tools such as Chinook as "administrative aids," these systems have effectively become the primary gatekeepers. From a practitioner's perspective, the concern is evident: automation increases processing speed but simultaneously reduces transparency and heightens the risk of systemic bias.

2. The Case of Chinook: Erased Logic and "Black Box" Refusals

A significant concern with the Chinook system is its automatic deletion of transitory working notes at the end of each session. Under Canadian law, providing "reasons for a decision" forms the basis of procedural fairness.

- The Judicial Review Gap: When an officer's internal reasoning is erased, the court is left only with the final "form-letter" refusal. This absence of documentation makes it nearly impossible for a judge to assess whether the officer acted with bias or disregarded relevant evidence.
- The "Standard Formulation" Trap: There has been a notable increase in Study Permit refusals, with qualified applicants receiving vague, automated statements such as "I am not satisfied you will leave Canada." These refusals frequently do not reflect individualized assessments but rather represent "algorithmic echoes" of risk flags generated by the system.

In high-volume workflows, interface design may enable batch processing of applications with standardized refusal codes, raising concerns about whether individualized assessment is meaningfully preserved.

3. Bots, Bias, and the Creation of "Non-Subjects"

Recent research (Lallani, 2025; Bélanger & Bergevin-Estable, 2024) demonstrates that artificial intelligence systems actively construct social categories.

- Proxy Discrimination: Algorithmic "risk indicators" frequently serve as proxies for country of origin, age, or socio-economic status. When training data contains historical biases against specific regions, the AI system is likely to flag applicants from those regions, perpetuating a feedback loop of exclusion.
- The Illusion of Human Discretion: Although a human officer formally signs the refusal, the high volume of cases compels officers to rely on the AI's "recommendation." As a result, human discretion is increasingly supplanted by "algorithmic nudging."

This risk emerges at multiple points in the decision pipeline, including data selection, model training, triage classification (e.g., tiering), and final decision execution, indicating that bias mitigation must be multi-layered rather than confined to a single stage.

4. The Practitioner's Dilemma: Ethics and Competence

According to the Code of Professional Conduct (2021), an RCIC is obligated to maintain competence and advocate effectively. However, the absence of "explainability" in IRCC's AI tools renders this duty nearly impossible to fulfill:

1. The Communication Void: An RCIC cannot adequately explain a refusal to a client when the underlying rationale is concealed within an undisclosed risk algorithm.
2. The "Shadow Boxing" Effect: RCICs are compelled to draft Statements of Purpose (SOPs) in response to "ghost indicators" that are never officially disclosed.

This creates a structural asymmetry of information, where neither applicants nor their representatives can meaningfully engage with the decision-making criteria.

5. Position and Recommendations: Toward Trustworthy AI

I propose that the legitimacy of AI in immigration should be grounded in verifiability rather than solely in processing speed. My position includes three essential requirements:

1. Mandatory Log Retention

A ban on the deletion of transitory notes in Chinook. The entire logic path must be preserved for audit and appeal.

At a minimum, retained logs should include: (a) triggered risk indicators and associated scores or categories; (b) triage classification (e.g., Tier 1/2/3) and batch identifiers where applicable; (c) officer overrides of automated recommendations with recorded justification; (d) standardized refusal codes and their variants; (e) full linkage between Chinook and GCMS records.

Primary responsibility: IRCC for implementation; Treasury Board Secretariat for removing policy exemptions related to "assistive tools."

2. Right to Explanation

Applicants should be informed if an automated triage tool flagged their file and provided with the specific "risk categories" used.

Explanations should follow a structured format, including: (a) whether automated triage was applied; (b) which risk indicators were triggered and their plain-language meaning; (c) the specific refusal code and its evidentiary basis; (d) whether the case was processed individually or within a batch; (e) clear guidance on avenues for review and access to records.

Primary responsibility: IRCC (decision templates); legislative clarification may be required to strengthen enforceability.

3. Continuous Independent Audits

Annual, publicly available Algorithmic Impact Assessments (AIA) conducted by third parties to identify and mitigate "hidden" geographic or racial biases.

Audits should rely on measurable fairness metrics, including: (a) disparity in triage outcomes across countries (e.g., deviation thresholds); (b) concentration of refusal codes by region; (c) frequency and justification of officer overrides; (d) detection of anomalous batch refusal patterns; (e) differential processing times across demographic groups.

Primary responsibility: Treasury Board Secretariat to mandate audits; independent research bodies to conduct parallel evaluations.

Conclusion

Immigration concerns human lives, not merely data points. It is imperative that "efficiency" does not override "fairness." For Canada to sustain a trustworthy immigration system, technology must reflect core societal values. Trustworthy AI in migration necessitates a shift away from opaque systems toward transparent, accountable, and human-centred decision-making.

Without enforceable requirements for logs, explanations, measurable audits, and institutional accountability, automated systems risk entrenching systemic bias rather than alleviating it.

KEYWORDS

Automated Decision Support Systems (ADSS), Procedural fairness, Algorithmic bias, Chinook system (IRCC), Judicial review, Black box algorithms, Migration management, RCIC (Regulated Canadian Immigration Consultant), Transparency and accountability, Risk indicators

REFERENCES

[1] Bélanger, D., & Bergevin-Estable, G. (2024). Revisiting 'who gets in': borders and migration management in the era of automation and AI in Canada. Open Forum.

[2] Lallani, M. (2025). Bots, Bias, and Borders: The effects of automated decision-making on Canadian immigration systems. PhD Thesis, York University.

[3] Suleman, Z. K. B. (2021). Chinook: The IRCC's New Processing Tool and the Erosion of Judicial Review.

[4] CICC. (2021). Code of Professional Conduct for the College of Immigration and Citizenship Consultants.

[5] Government of Canada. Directive on Automated Decision-Making & AIA Reports for IRCC.