Brief on compounded risks of AI use in Québec's public service

Eva Portelance[a,d*], Ayla Rigouts Terryn[b,d*], Afaf Taïk[c,d*]
[a]Département des sciences de la décision, HEC Montréal
[b]Département de linguistique et de traduction, Université de Montréal
[c]Département de génie électrique et de génie informatique, Université de Sherbrooke
[d]Mila - Québec artificial intelligence institute

[*]Equal contribution, corresponding authors:
{eva.portelance,ayla.rigouts-terryn,afaf.taik}@mila.quebec

The following brief identifies and discusses four families of risks that arise from AI use in public service and in the Québécois context. Each family is presented in a separate section and these sections are ordered for explanatory clarity, not importance; they are all important. A final section is dedicated to the discussion of possible precautionary measures to help address some of these risks, specifically in the context of AI use cases in Quebec public service.

On terminology. *Artificial intelligence* covers a wide range of systems. This brief focuses primarily on the class of systems driving the current wave of adoption in public services: large language models (LLMs), the generative models behind chatbots, drafting assistants, and agentic systems. Most of the risks we describe, however, are older than LLMs. They have been documented in the predictive models that public administrations have used for years to score, rank, and flag: to estimate a risk of fraud, prioritise an application, or select a file for review. We draw on that evidence throughout, because LLMs do not resolve those risks. They inherit them, and they extend their reach, since a handful of models now underpin services across the board. Some of what follows is specific to LLMs, and we say so where it is.

1 Architectural risks: The current AI service paradigm

Many contemporary generative-AI services are built around large language models (LLMs), which are one class of foundation model. Most have now heard of these models, but few understand how they actually work and why by their very nature they can introduce biases and present clear usage risks [6]. This first section will explain what they are and how they are used in agentic AI systems such as those being considered for use in public service.

1.1 Foundation models starting with LLMs

LLMs are predictive models which operate over tokens. The tokens for each model are a fixed set of words (e.g., *time, in*), subwords (e.g., *ent, ly*), and characters (e.g., *!, x, &*)[1]. LLMs are trained to predict – or maximize the probability of – the next token in a sequence, where the preceding tokens represent the context. Once a model is trained to do these predictions it can be used to generate new ones. Given an

[1]As a first distinction for the Québec context, these sets of tokens in LLMs most often lean towards frequent English words and subwords, with other languages being accounted for through characters and some common subword syllables.

input of a sequence of tokens as context an LLM outputs a probability distribution over all the possible values for the next token. A sampling algorithm is then used to pick or *generate* the next token from this distribution, Figure 1. To generate the next token after that, the process is applied recursively by adding the sampled token to the input of the next model call and sampling again and so on, which leads to compounding uncertainty. This is the generative principle behind all commercial LLMs.

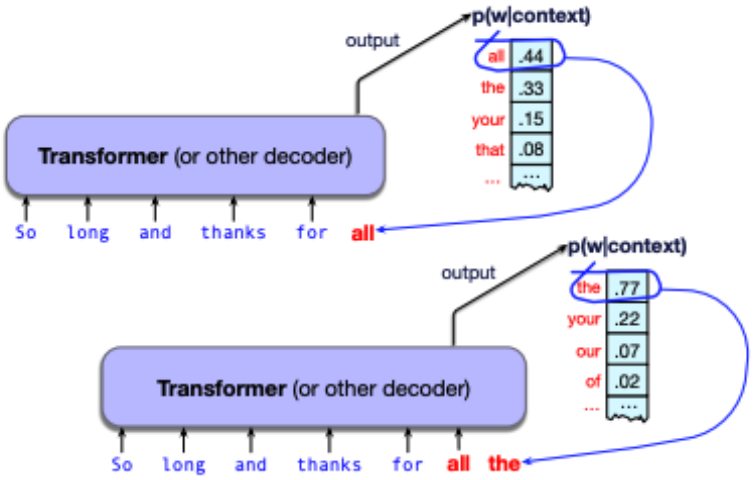

Figure 1: The LLM takes a context of tokens, *So long and thanks for*, as input and outputs a distribution over the next possible token. To generate, a sampling algorithm is then used to pick the next token from this distribution, *all*. The process is applied recursively by adding this token to the input context of the next generation and so on, compounding the uncertainty of the generation. Figure from [17].

Predictive language models have existed since the 1950s [34, 33][2], but what has allowed the newest models to generate naturalistic responses and encode so much knowledge are: (1) the architectural innovation presented by Transformer architectures [39], and (2) the sheer scale of contemporary LLMs training data and model size. Transformers can learn to represent much more complex relations between context tokens than previous models, allowing for robust links to be built across complex and sometimes very long contexts[3]. The very large scale of LLMs is however what truly gives them their power and what has lead to their use as foundation models on which most AI systems are now built.

1.2 Building on shaky foundations and Agentic AI systems

LLMs are able to encode large amounts of knowledge, linking information from across their training data. This encoded knowledge is what then serves as the foundation for their use in downstream tasks or tools. For example, conversational agents, like ChatGPT, Gemini, or Copilot, start from a foundational base LLM which is then finetuned – or further trained – on conversational and instructional data using different training techniques to make its generations seem more naturalistic in conversational settings.

LLMs can also be combined together in agentic pipelines. In agentic systems, each agent is assigned a subtask of a more complex task needing to be accomplished. A basic example is retrieval-augmented-generation (RAG) systems for question answering. LLMs can be queried and their answers are based on their encoded internal knowledge and context. If the answer however isn't there, say if the query is about current events or about private information, then a model will most likely to fail to give the correct answer

[2]The early models were known as n-gram models, or Markov models applied to language

[3]Up to 1 million context tokens can be included at the upper end of current private model offerings from companies like Anthropic and OpenAI.

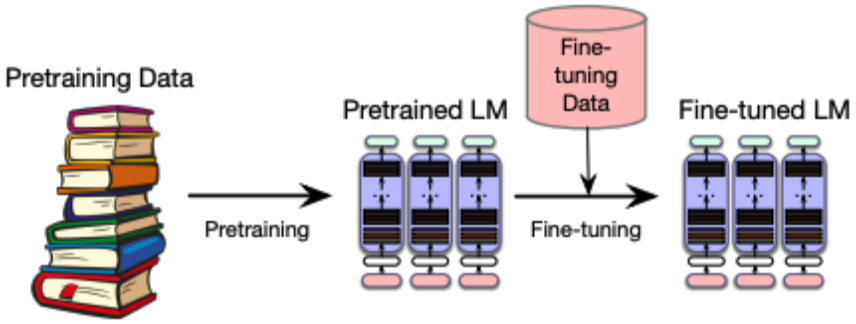

Figure 2: LLMs encode foundational knowledge from their pretraining data. They can then be finetuned on additional data for downstream tasks, while foundational knowledge and biases may stay. Figure from [17].

and hallucinate. It may seem obvious why: the answer simply isn't available in the model's context, even though there is nothing stopping the user from asking their query. An alternative to the single model solution is to have a two step system. First, a retrieval model will search for relevant information on the web or an internal database. Next this information is included in the context prompt of a generation model to increase the likelihood of generating the right answer. Though breaking down complex tasks into smaller subtasks can help accomplish them correctly, added complexity and steps can also lead to compounded prediction mistakes.

1.3 Compounding of architectural risk and the importance of context

Fundamentally, all systems that are built on LLMs are contextualised prediction models given fixed contexts. This tokenized context does not have access to broader cultural context, to the design motivations that we have in a given public service use case, or to all the background context a citizen, here user, may have. All models have are the tokens we provide in the prompt we give them and the biases in their encoded knowledge.

Let's take a concrete example of where this contextual limitation could go wrong: Santé Québec decides to automate part of the 811 phoneline triage service for people seeking medical appointments without family doctors to alleviate the current call burden. An LLM-based AI agent is prompted to ask the client for the reason for their call and to list their symptoms. The AI agent has access to existing checklists of questions to ask and routing maps depending on symptoms. However, it does not have the experience of a nurse to ask the missing probing questions in certain contexts; it does not have access to the cultural knowledge – for example that many Québécois are prideful and do not complain, misrepresenting the severity of symptoms – that a good nurse can read from an interaction; the AI agent does not have access to the background context of clients who may describe symptoms in different ways, depending on their gender, their accent or language register, or other social factors. The lack of context means that the AI agent may make different predictions for people with the same condition. This is not to say that humans do not make prediction mistakes, but that we cannot simply assume that an AI will make the same mistakes; an AI agent does not have the contextual knowledge we rely on. Trust in such systems can only be established if we understand in each use case what mistakes it does make and how they can be mitigated. The following sections examine what stands in the way of that understanding and what is at stake when it is absent.

2 Technological risks: Using privately owned models in the public system

Section 1 described risks due to how the technology works. This section turns to risks that follow from who builds and owns it. Generally, public bodies do not train their own models; they contract commercial providers. When a public body relies on a closed, vendor-hosted model, it may have limited access to the model's training data, weights, system instructions, evaluation results, and change history, while the service runs partly or entirely on infrastructure outside its control. Québec's own *Stratégie d'intégration de l'IA* [14] facilitates this adoption, but its key measures for 2024-2026 do not appear to establish generally applicable guarantees of independent technical auditing, specific contractual conditions for closed commercial models, or meaningful reconsideration of decisions materially influenced by AI [15].

Consider a Québec ministry that deploys a commercial LLM to help process incoming citizen requests: housing applications, benefit claims, complaint files. The system is evaluated before launch and performs adequately. Six months later, the provider updates the underlying model, and the update is not necessarily transparent. The updated model handles certain categories of requests differently, and the ministry has no way to determine what changed, or why.

2.1 Opacity and model drift

A public body that contracts a closed commercial model may be unable to inspect its complete training corpus, model weights, post-training data, system instructions, safety configuration, or detailed evaluation results. Sometimes this is a matter of competitive secrecy: when asked to complete disclosure forms for a U.S. government procurement coalition, vendors routinely withheld technical details they considered commercially sensitive [20]. Sometimes the information is simply not available, because many AI services are built on top of foundation models developed by others. These underlying models can also change without notice: one longitudinal evaluation found that GPT-4's measured performance on a particular task fell from 84% to 51% across service versions released a few months apart [8]. A downstream vendor or public body might not receive enough information to identify the cause or anticipate its effect on its own application.

In practice, disclosures are poor. Kuehnert et al. [20] examined 39 vendor disclosure forms submitted for U.S. government AI procurement and found that most provided only vague descriptions of training data, and over half reported no performance evaluation results at all. Policy responses are emerging. At the federal level, Canada's Directive on Automated Decision-Making requires an Algorithmic Impact Assessment for systems within its scope [38], and the Organisation for Economic Co-operation and Development's 2025 report on AI in government calls for stronger oversight [27]. But these frameworks assume the deploying body can answer questions about the model it uses, and existing access-to-information laws are not designed for algorithmic transparency: they conflict with intellectual property restrictions and trade secrets of private providers [27].

Even in a best-case scenario where a provider discloses everything it knows, the problem would not disappear. As Section 1 explained, the behaviour of foundation models in a specific situation cannot be fully predicted, even by their developers. For a public service, this means that full transparency, while necessary, is not sufficient: the model remains partly opaque even to the people who built it.

2.2 Data sovereignty and structural dependency

When a public service uses a vendor-hosted model, citizen information may be transmitted to and processed on infrastructure outside the public body's direct control. The resulting risks depend on the deployment and contractual arrangement. They include unauthorized access during transmission or processing, retention of prompts and outputs in provider logs, access by employees or subcontractors, secondary use for service improvement or model training, and exposure through a security breach. These risks can be reduced through

various strategies, but they cannot be treated as resolved merely because a provider promises not to train on client data.

A related risk concerns information already incorporated into model training. Nasr et al. [26] showed that an outsider with ordinary access to ChatGPT could induce the model to output verbatim fragments of its training data, including personal information. That particular method was patched, but the study illustrates that LLMs can memorize portions of their training data. Reliably identifying and removing a particular training example from a deployed model remains technically difficult [22].

Québec's privacy framework imposes relevant obligations, but these must be described precisely. The *Act respecting Access to documents held by public bodies and the Protection of personal information*, as amended by Loi 25, requires a privacy impact assessment for projects involving the acquisition, development, or redesign of an information system or electronic service-delivery system that involves personal information. It also imposes transparency obligations when a decision is based exclusively on automated processing of personal information. These statutory requirements should be treated as a floor rather than a complete governance framework: many consequential uses of AI assist or shape a human decision without meeting the narrow threshold of an exclusively automated decision.

These risks are compounded by the lack of options between providers and the increasing difficulty of switching between them. Even in the simplest case, a chatbot, switching is not straightforward: it means rebuilding technical integrations and re-engineering prompts for a different system. But providers are moving well beyond chatbots, toward integrated ecosystems that manage documents, automate workflows, and adapt to an organization's patterns. The more processes are built around a provider's tools, and the more the system learns about how an organization works, the more disruptive it becomes to leave. Langer and Haag describe this concentration of AI capacity as a "structural dependency particularly critical for the public sector" [21]. New providers do enter the market, but the required investments and expertise prevent them from catching up to a very small number of dominant players. A public institution that wants the best available performance must turn to these dominant providers, i.e., large American companies offering limited transparency. More transparent or locally hosted alternatives exist, although they require additional engineering capacity and generally offer lower general-purpose performance. Convincing staff to use a noticeably weaker tool is its own challenge. Section 3 examines what happens when the errors of such a system affect an entire service, and Section 4 asks whose perspective those errors reflect.

3 AI for decision-making: Introducing systemic bias

3.1 Systemic bias and algorithmic monoculture

These risks are the best documented of the four, because they do not depend on LLMs. Public administrations have used predictive models for years to score, rank, and flag, and the evidence of what goes wrong comes largely from them. Despite often being perceived as objective, AI systems are shaped by numerous design choices made throughout their development. These choices include how data are collected and cleaned, how missing values are handled, which variables are included, how different groups are represented in the training data, the type of model that is selected, and the values of its hyperparameters. In many cases, several alternative choices are technically valid and lead to models with similar levels of predictive performance. Yet these models may produce different outcomes for individual people [36]. As a result, AI systems do not uncover a single objective truth. Rather, they encode a particular set of assumptions, priorities, and trade-offs chosen by their developers [28]. A model optimized to only minimize overall error, for example, may produce different decisions than one designed with a constraint aiming to reduce disparities across demographic groups. Similarly, two models trained on the same data but using different architectures or parameter settings may disagree on whether an individual qualifies for a benefit, receives a medical referral, or is flagged for additional scrutiny [13].

In practice, this variability remains largely invisible because only one model is typically deployed, making its decisions effectively final. This phenomenon is further amplified by the widespread adoption of standardized foundation models, leading to an *algorithmic monoculture* [19]. When identical systems are deployed at scale, their biases and errors become systemic. For example, a job applicant unfairly penalized by a recruitment algorithm may be consistently excluded from employment opportunities, with no chance of being reassessed by a different system that might reach a different conclusion [11]. Similarly, if the same diagnostic model is adopted across hospitals and clinics, a patient who receives an incorrect assessment may encounter the same error throughout the healthcare system. The widespread use of a single model effectively eliminates the possibility of independent *second opinions* from alternative systems, increasing the risk that mistakes are propagated rather than corrected. This homogenization reduces society's resilience to algorithmic errors and failures.

3.2 Systemic bias and feedback loops

Vulnerability to systemic errors is severely compounded by the presence of self-reinforcing feedback loops, where algorithmic predictions actively reshape the social reality they claim to objectively measure [29]. When a model's outputs dictate real-world interventions, those interventions generate new data that reflexively validate the initial, biased prediction [23]. In predictive policing, for example, an algorithm that forecasts higher crime rates in historically marginalized neighborhoods prompts increased police deployment to those areas. Consequently, more arrests are recorded there, not necessarily due to a higher underlying propensity for crime, but because of heightened surveillance. Such results are then fed back into the model as empirical proof of its accuracy. This dynamic transforms historical bias into a self-fulfilling prophecy, capturing the system in a closed loop where errors are continuously amplified rather than corrected.

This systemic vulnerability is severely compounded in domains that lack an objective *ground truth*, where algorithmic predictions inadvertently create the very data used to validate them. In fraud detection, for instance, a model's initial mistakes can become permanently locked into future training cycles. If an algorithm incorrectly flags a legitimate transaction as fraudulent, that transaction is typically blocked or rejected, without actually verifying whether fraud would have actually occurred [30]. As the model is routinely retrained on new data, its early errors and systemic biases are fed back into the training pipeline as definitive truths. The result is that the algorithm does not learn to detect fraud independently; it learns to reinforce its own earlier mistakes.

3.3 The limits of human review

The usual answer to these risks is to keep a *human in the loop*: a trained person reviews what the system produces and catches its errors before they reach a citizen. The evidence does not support that assumption. People accept a recommendation when it confirms what they already believe and set it aside when it does not, and this holds whether the recommendation comes from an algorithm or from a human expert [32, 2]. The problem is that a model trained on past decisions and biases reproduces these same patterns, and human reviewers may therefore reinforce rather than correct existing biases. Human bias enters the data, the model encodes it, and the human review confirms it again.

The Dutch childcare benefits scandal shows what this produces. The Dutch tax authority used nationality in risk-selection and fraud-detection processes, and used it to flag families for closer scrutiny rather than to decide their cases. Flagged cases were subsequently processed by officials. The national data protection authority found that both the automated selection and the human investigations that followed were discriminatory [2]. Tens of thousands of families, many of them of foreign descent, were wrongly accused of fraud and ordered to repay sums running into tens of thousands of euros. Models reproduce biases that came from humans in the first place, and putting a human back in the loop is no guarantee against them. The scandal

did not result from a single model alone, but from the interaction of discriminatory risk selection, rigid fraud policy, weak evidence-based safeguards, and ineffective human review.

4 Linguistic and cultural risks: Québec's values and realities

4.1 Unequal data, unequal representation

The biases that spread through the dynamics described in Section 3 are not random; they reflect whose data the model was trained on. Focusing on LLMs specifically, we see that the training material is not an even sample of the world. It is dominated by content from a small number of wealthy, English-speaking, Western societies, and contains comparatively little from everyone else. As explained, the model reproduces the most probable patterns in that data, so it carries this imbalance forward. Québec French, Indigenous languages, local institutions, and many of the communities served by Québec public bodies are comparatively underrepresented in the resources used to build and evaluate these systems.

Language. The French generated by these models is often fluent, but existing evaluations identify measurable gaps in their handling of Québec-French linguistic phenomena; Québec French remains comparatively under-resourced in NLP datasets and benchmarks [18, 5]. Generic claims of "French-language capability" therefore do not establish that a model can communicate accurately and appropriately in Québec public services, and the gap is wider still for the languages of Québec's Indigenous nations. This is not only a question of style: Québec sets legal standards for the language of public communication (the Charter of the French Language, the norms of the *Office québécois de la langue française*), and any public-facing system must be evaluated against the legal, terminological, accessibility, and communicative requirements applicable to the service in question.

Facts. The model is more often wrong about places, groups, and topics that are underrepresented in its training data [25], and asking the same factual question in French rather than English can change the answer [31]. For a public service, these findings give reason to expect lower performance on underrepresented Québec-specific information. That risk must be measured for each intended application rather than inferred from general multilingual performance.

Values. Cross-cultural evaluations show that model responses do not reproduce the full distribution of views found within or across populations. Instead, responses often cluster around a limited set of national or demographic profiles, which varies with the model, prompt language, elicitation method, and topic [37, 7]. The policy concern is therefore not that a model possesses one fixed foreign culture, but that an unrepresentative default may influence communication, prioritisation, or judgment without having been explicitly selected or transparently documented.

4.2 The impact of bias

The clearest harm is that the bias falls hardest on the people least able to absorb it. Members of the public may consult general-purpose chatbots to ask how a public program applies to them before reaching an official channel. For a common case, that may work well. For an unusual one, and especially for cases where Québec's rules and regulations differ from the ones in other, better represented countries, replies will more often be wrong. The people who most need a correct answer get the least reliable one, with no public body in the loop to catch the error.

The bias also affects language itself. A single reply in French that does not quite match Québec usage is trivial. But these models now produce so much of the text people read that the words they favour are beginning to appear in everyday human speech [3]. Applied to French at scale, the same exposure could pull usage toward the models' France-leaning default and greater influence from English, and away from the norms Québec's law exists to protect. Section 3 described how effects like this compound through feedback

loops. In a public-service setting, linguistic and cultural mismatch can therefore affect not only tone, but also accessibility, the interpretation of a request, the information provided, and the exercise of administrative discretion.

Finally, these models are not passive. They tend to agree with the user they are talking to (sycophancy) [35], and they can shift what people believe (persuasion) [10]. These behaviours depend on pre-training, post-training, prompting, interface design, and user characteristics, and providers do not yet control them reliably. Research has not yet established how consistently these behaviours occur across contexts and how they affect different users. Yet these are already tools that many people rely on daily, in their work and their personal lives, and that governments are rapidly integrating into public services.

5 Taking precautions and placing guardrails

The first question is not which model to use, but whether a given task should be automated at all [1]. For some decisions the answer will be no, and that answer has to be enforceable. These tools are available to every employee from any browser, and a rule nobody can verify is not a safeguard. Where automation is warranted, model choice matters, but no model available today is a perfect solution to the problems described so far. Importantly, there is no reason to expect these issues will be solved by waiting for the next bigger, better model either. Public services will have to work with imperfect systems for the foreseeable future, so the precautions have to come from how each system is chosen, deployed, and supervised.

5.1 Evaluation before and during deployment

A model that scores well on a vendor's benchmark, usually built in English and for the majority case, can still fail on Québec's rules, on regional French, or on the people its training data represents least (Section 4). Each system therefore has to be tested in realistic settings, and the results have to be broken down by group, by region, and by language, because an average hides the failures that tend to impact the most vulnerable. The usual guardrails do not remove this need: anchoring a model in approved documents (Section 1) reduces invented facts, but it does not guarantee a better alignment with the values and languages; and rules that block certain outputs catch only the problems someone thought of in advance. Since models change rapidly, sometimes without notice (Section 2), testing has to continue after deployment, and systems have to preserve the information needed to reconstruct consequential interactions, including the model version, relevant instructions, retrieved sources, and material outputs, subject to data-minimisation, access-control, security, and retention requirements, because a decision that was never recorded cannot be audited or contested.

Québec's strategy provides for a governance framework and a risk-assessment tool for AI projects [15]. It should be conducted by evaluators who are organisationally and technically independent of the deployment decision and who have sufficient access to the provider's documentation, system, and logs. A provider unwilling to supply what an evaluation needs should be reason enough not to deploy the system. That independence requires expertise in Québec, and initiatives that build Québec-based evaluation expertise, including work on Québec French and local institutional contexts, are an important start [9]. Such collaborations should not, however, substitute for evaluation that is independent of the provider and procurement decision.

5.2 Human oversight and citizen recourse

Meaningful human oversight, i.e., human-in-the-loop, is a minimum requirement when an algorithm materially influences a consequential decision or access to a public service. A trained person should remain accountable for any decision that affects a citizen, with the time to review the output and the authority to

overrule it, rather than merely approving it. A service also has to be able to run without the model, which will at some point be unavailable, degraded, or withdrawn, and staff can only take the work back if they still know how to do it. There is evidence that reliance erodes that ability [12, 4], so staff have to keep doing the work unaided at least part of the time.

Nevertheless, even with trained experts, human review is not a catch-all (Section 3). Most systems prime the human reviewer's judgment by presenting the model's assessment first and leaving the human reviewer to confirm or modify it. Experimental evidence indicates that requiring reviewers to form an initial judgment before seeing an algorithmic recommendation can reduce anchoring and improve combined performance [1]. For sufficiently important tasks, interfaces should therefore be tested to determine whether reviewers should assess the evidence before seeing the model's recommendation, rather than being primed by it from the outset. LLMs produce fluent, confident answers whether or not they are correct, and their errors are plausible enough to survive a quick check (Section 1). A reviewer therefore has little basis on which to judge how reliable the system in front of them is, and can easily overestimate it. This compounds a problem that is present even in simpler systems: the better a system appears to perform, the less closely people check it, and the harder its output is to verify, the more readily they accept it [32]. Higher average accuracy can increase reviewers' trust and reduce scrutiny, so the errors that remain may be less likely to be detected. Evaluation must therefore measure the performance of the combined human-AI workflow, not only the model in isolation.

Errors will reach citizens whatever the review process, so citizens need protections of their own. They should be told when a service uses AI, which Québec's strategy already commits to (Measure 8.1) [15]. They should have access to a human communication channel, an accessible non-digital route where necessary, and meaningful human reconsideration of any consequential decision materially shaped by AI. For public-facing conversational systems, people should also be able to request service from a person without penalty (opt-out). Section 2 noted that current measures do not guarantee that recourse. Such recourse is meaningful only if the reviewer has access to the underlying record, the authority to depart from the system's output, and the obligation to reconsider the case on its merits.

5.3 Investing in local capacity

Québec is under-represented in the training data of the dominant models (Section 4), and the text that would correct this, in Québec French, in Indigenous languages, or about Québec's own institutions and rules, does not exist at the scale those models were trained on. Bibliothèque et Archives nationales du Québec has begun building a public bank of governmental and cultural data, now in a funded experimentation phase [24]. Such collections must be developed with clear provenance, licensing, privacy protection, and documentation of representativeness. Data involving First Nations languages, knowledge, or communities require governance developed with the relevant communities require governance developed with the relevant communities; public availability alone does not establish a right to use those materials for model training. Collections of this kind may not completely close the gap, but they are essential both for training and for evaluating systems on Québec's own terms. These efforts should go beyond just data. Québec has substantial AI research capacity, and sustaining and directing it toward the problems public services actually face would strengthen institutional resilience in the AI era.

Québec also needs models it can run itself. Section 2 noted that locally hosted alternatives may involve a trade-off in general-purpose capability and require substantial local engineering capacity. That cost is much lower for narrow tasks, and not every task calls for a large language model in the first place: a smaller model built for a specific job, and hosted locally, can perform comparably to the leading commercial systems [16]. This only works if the model can be trained on enough relevant data and kept up to date, which is why the collections described above matter. Even then, local capacity is not a cure. A system built on top of a commercial foundation model inherits the skew of that foundation, trained mostly on English and Western

data (Section 4). A model trained locally on Québec data may reduce some of that skew, but it can reproduce biases already present in Québec society and institutions. (Section 3). Running a model on Québec servers addresses only the privacy risk related to the provider's infrastructure, without solving the others (Section 2). A model the government owns still has to be evaluated for the use it is put to, supervised by people who can overrule it, and answerable to the citizens it affects.

Throughout this brief, we have highlighted how AI technology is unreliable in ways that are hard to predict (Section 1), difficult to inspect or control (Section 2), prone to systemic error (Section 3), and biased against the people and places it knows least (Section 4). There are no easy solutions, but we can take precautions and make thoughtful choices when choosing which problems could benefit from AI integration. We can be more responsible and accountable through testable design of AI solutions, independent and constant auditing of AI integrated workflows, continuously keeping people involved and capable of performing all processes, and by offering to all citizens control over when and where AI is involved in decisions that affect their lives.

About authors

Dr. Eva Portelance is an Assistant Professor in Decision Sciences at HEC Montréal, and by courtesy Adjunct Professor in Computer Science at Université de Montréal. She is also an IVADO Professor and an associate member of Mila — Québec Artificial Intelligence Institute. She is one of the authors of the seminal white paper *On the opportunities and risks of foundation models* from the Stanford Center for Research on Foundation Models which helped define the current paradigm for LLM technologies. She studies the emergence of language understanding and reasoning in AI and children.

Dr. Afaf Taïk is an Assistant Professor at the Electrical and Computer Engineering Department at Université de Sherbrooke. She is also an associate academic member at Mila - Québec Artificial Intelligence Institute. Her research focuses on responsible artificial intelligence, specifically fairness and privacy in machine learning.

Dr. Ayla Rigouts Terryn is an Assistant Professor of Translation Technology and AI in the Département de linguistique et de traduction at Université de Montréal. She holds an FRQ-IVADO Research Chair (2026-2031) on languages and AI, is an IVADO Professor, and an associate academic member of Mila - Québec Artificial Intelligence Institute. Her research analyses cultural and value biases in large language models across languages, with a focus on non-English settings and low-resource languages.

References

[1] U. Agudo, K. G. Liberal, M. Arrese, and H. Matute. The impact of AI errors in a human-in-the-loop process. *Cognitive Research: Principles and Implications*, 9(1):1, Jan. 2024.

[2] S. Alon-Barkat and M. Busuioc. Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory*, 33(1):153–169, Jan. 2023.

[3] B. Anderson, R. Galpin, and T. S. Juzek. Model Misalignment and Language Change: Traces of AI-Associated Language in Unscripted Spoken English. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):179–191, Oct. 2025.

[4] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, and R. Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, July 2025.

[5] D. Beauchemin and R. Khoury. QFrCoLA: A Quebec-French Corpus of Linguistic Acceptability Judgments. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 119–130. Association for Computational Linguistics, 2025.

[6] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. A. Creel, J. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. E. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. F. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. P. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. F. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. H. Roohani, C. Ruiz, J. Ryan, C. R'e, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. P. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. A. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models. *ArXiv*, 2021.

[7] B. Bulté and A. Rigouts Terryn. LLMs and Cultural Values: The Impact of Prompt Language and Explicit Cultural Framing. *Computational Linguistics*, 52(2):1–88, Apr. 2026.

[8] L. Chen, M. Zaharia, and J. Zou. How Is ChatGPT's Behavior Changing Over Time? *Harvard Data Science Review*, 6(2), Apr. 2024.

[9] Cohere and Mila. Cohere and Mila partner to advance Quebec French language and cultural context in AI. Press release, 27 May, 2026.

[10] T. H. Costello, K. Pelrine, M. Kowal, A. A. Arechar, J.-F. Godbout, A. Gleave, D. Rand, and G. Pennycook. Large language models can effectively convince people to believe conspiracies, 2026.

[11] K. Creel and D. Hellman. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy*, 52(1):26–43, 2022.

[12] A. Ferdman. AI deskilling is a structural problem. *AI & SOCIETY*, 41(4):3001–3013, Apr. 2026.

[13] P. Ganesh, A. Taik, and G. Farnadi. Systemizing multiplicity: The curious case of arbitrariness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 1032–1048, 2025.

[14] Gouvernement du Québec. Stratégie d'intégration de l'intelligence artificielle dans l'administration publique 2021-2026. Technical report, Secrétariat du Conseil du trésor, 2021.

[15] Gouvernement du Québec. Mesures clés 2024–2026: Stratégie d'intégration de l'intelligence artificielle dans l'administration publique 2021-2026. Technical report, Ministère de la Cybersécurité et du Numérique, 2025. ISBN 978-2-550-98807-6.

[16] S. Gupta, A. Basu, M. Nievas, J. Thomas, N. Wolfrath, A. Ramamurthi, B. Taylor, A. N. Kothari, R. Schwind, T. M. Miller, S. Nadaf-Rahrov, Y. Wang, and H. Singh. PRISM: Patient Records Interpretation for Semantic clinical trial Matching system using large language models. *npj Digital Medicine*, 7(1):305, Oct. 2024.

[17] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, chapter 7: Large Language Models. 3rd edition, 2026. Online manuscript released January 6, 2026.

[18] E. Khan, F. Saidani, O. V. Esbroeck, R. Khoury, and L. Kosseim. Low-Resource Dialect Adaptation of Large Language Models: A French Dialect Case-Study. In *The Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10723–10734, Palma, Mallorca, Spain, May 2026.

[19] J. Kleinberg and M. Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

[20] B. Kuehnert, N. Johnson, R. Dotan, and H. Heidari. Disclosure or Marketing? Analyzing the Efficacy of Vendor Self-reports for Vetting Public-sector AI. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, pages 687–713, New York, NY, USA, June 2026. Association for Computing Machinery.

[21] P. F. Langer and S. Haag. AI Sovereignty: Redundant Use of Large Language Models for Public Sector Resilience. *Public Organization Review*, May 2026.

[22] S. Liu, Y. Yao, J. Jia, S. Casper, N. Baracaldo, P. Hase, Y. Yao, C. Y. Liu, X. Xu, H. Li, K. R. Varshney, M. Bansal, S. Koyejo, and Y. Liu. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2):181–194, Feb. 2025.

[23] K. Lum and W. Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[24] Ministère de la Culture et des Communications du Québec. Le québec à l'heure de l'IA: le gouvernement mandate BAnQ pour poser les fondations d'une banque publique de données. Communiqué, 8 mai, 2026.

[25] J. Myung, N. Lee, Y. Zhou, J. Jin, R. A. Putri, D. Antypas, H. Borkakoty, E. Kim, C. Perez-Almendros, A. A. Ayele, V. Gutiérrez-Basulto, Y. Ibáñez-García, H. Lee, S. H. Muhammad, K. Park, A. S. Rzayev, N. White, S. M. Yimam, M. T. Pilehvar, N. Ousidhoum, J. Camacho-Collados, and A. Oh. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. *Advances in Neural Information Processing Systems*, 37:78104–78146, Dec. 2024.

[26] M. Nasr, J. Rando, N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. Choquette-Choo, F. Tramer, and K. Lee. Scalable Extraction of Training Data from Aligned, Production Language Models. *International Conference on Learning Representations*, 2025:82363–82435, May 2025.

[27] OECD. Governing with artificial intelligence: The state of play and way forward in core government functions. Technical report, OECD Publishing, Paris, 2025.

[28] C. O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

[29] N. Pagan, J. Baumann, E. Elokda, G. De Pasquale, S. Bolognani, and A. Hannák. A classification of feedback loops and their relation to biases in automated decision-making systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–14, 2023.

[30] J. Pombal, P. Saleiro, M. A. Figueiredo, and P. Bizarro. Prisoners of their own devices: How models induce data bias in performative prediction. *arXiv preprint arXiv:2206.13183*, 2022.

[31] J. Qi, R. Fernández, and A. Bisazza. Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore, Dec. 2023. Association for Computational Linguistics.

[32] G. Romeo and D. Conti. Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. *AI & SOCIETY*, 41(1):259–278, Jan. 2026.

[33] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[34] C. E. Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

[35] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez. TOWARDS UNDERSTANDING SYCOPHANCY IN LANGUAGE MODELS. 2024.

[36] J. Simson, F. Pfisterer, and C. Kern. One Model Many Scores: Using Multiverse Analysis to Prevent Fairness Hacking and Evaluate the Influence of Model Design Decisions. In *ACM FaccT*, 2024.

[37] Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346, Sept. 2024.

[38] Treasury Board of Canada Secretariat. Directive on automated decision-making, 2019. Amended 2023.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.