L'Effacement Algorithmique de la Diversité Linguistique : Une Analyse d'Investigation de l'Intelligence Artificielle, du Français Québécois et de la Quête de Modèles Souverains

L'avènement des grands modèles de langage (LLM) représente un changement de paradigme profond, non seulement en matière de capacité de calcul, mais aussi dans l'évolution structurelle de la communication humaine à l'échelle mondiale. Alors que l'intelligence artificielle consolide rapidement sa position en tant qu'interface principale pour la recherche d'informations, la participation civique et l'expression créative, le cadre architectural sous-jacent de ces systèmes dicte silencieusement les limites du langage acceptable. Parce que ces modèles de pointe sont majoritairement conçus au sein de l'écosystème de la Silicon Valley et entraînés sur des ensembles de données anglo-américains et centrés sur l'anglais, leurs résultats présentent un effet d'homogénéisation massif. Cette standardisation algorithmique érode activement les dialectes régionaux, les idiomes culturels et les nuances structurelles des langues non dominantes. En fin de compte, ce phénomène pose une menace existentielle grave à la diversité linguistique mondiale, agissant comme un catalyseur technologique pour l'extinction numérique des langues.

Ce rapport de recherche exhaustif propose une investigation minutieuse des mécanismes techniques, culturels et sociopolitiques à l'origine de l'homogénéisation linguistique induite par l'IA. En utilisant la réalité linguistique unique du français québécois comme étude de cas principale, l'analyse illustre exactement comment les résultats standardisés de l'IA dépouillent systématiquement l'âme culturelle des dialectes régionaux, les remplaçant par des variantes « aplaties », anglicisées ou standardisées selon le modèle parisien. De plus, le rapport examine la menace mondiale des « déserts de données » sur les langues minoritaires africaines et autochtones, et évalue les paradigmes de défense émergents. Ces défenses incluent la capitalisation massive des infrastructures d'« IA souveraine » et la mise en œuvre de politiques publiques linguistiques robustes par le gouvernement du Québec et la Francophonie au sens large. Les conclusions synthétisées ici sont structurées comme l'architecture d'un article bilingue, fournissant la recherche fondamentale nécessaire à une publication d'investigation complète.

1. L'Homogénéisation Algorithmique de la Langue

La dilution de la diversité linguistique par l'intelligence artificielle n'est pas simplement un sous-produit accidentel des préférences des utilisateurs ou un bogue logiciel facile à corriger ;

c'est un artefact structurel profond intégré à chaque étape du pipeline de développement des LLM. De la curation des données et de la tokenisation à l'alignement des modèles et à la génération, l'homogénéisation algorithmique de la langue se produit par le biais de plusieurs mécanismes techniques et sociolinguistiques cumulatifs qui privilégient systématiquement les langues de prestige par rapport aux variantes régionales.

La Domination Écrasante de l'Anglais dans les Corpus d'Entraînement

Le problème fondamental sous-jacent à la dilution linguistique de l'IA est la surreprésentation massive et non corrigée de l'anglais dans les ensembles de données de pré-entraînement. Internet, qui sert de principal terrain de collecte pour les modèles génératifs de pointe, est disproportionnellement anglophone. Par conséquent, la grande majorité des modèles d'apprentissage automatique dans le monde sont principalement entraînés à l'aide de données en anglais tirées de sources occidentales.[1] Pour les 1,52 milliard de personnes qui parlent anglais, ces systèmes fonctionnent avec une fluidité presque parfaite ; cependant, ils s'appuient sur des données Internet binaires qui élargissent intrinsèquement la fracture linguistique numérique pour le reste du monde.[2]

Lorsque les modèles multilingues traitent des requêtes dans des langues non dominantes, ils ne « pensent » pas intrinsèquement dans ces langues. Au lieu de cela, ils s'appuient souvent sur un mécanisme de traduction implicite ou une stratégie de « langue pivot ».[4] Des études d'interprétabilité récentes analysant les états cachés à travers les couches neuronales intermédiaires des modèles de langage révèlent que le traitement multilingue est fortement influencé par la composition des données de pré-entraînement.[4] Les langues fortement représentées dans les données, principalement l'anglais, agissent comme des ancrages sémantiques dominants.[4] Lorsqu'un utilisateur saisit une invite dans une langue régionale, le modèle cartographie fréquemment la signification sémantique dans son espace latent à dominante anglaise, génère une réponse conceptuellement enracinée dans la logique anglaise, et traduit le résultat dans la langue cible. Ce processus dépouille systématiquement le contexte culturel, produisant un résultat structurellement et philosophiquement anglo-américain, qui ne porte qu'un masque linguistique étranger.[6]

Coûts de Tokenisation et Asymétries Structurelles

Le biais contre les langues non standard est encore plus enraciné dès la toute première étape du traitement des données : la tokenisation. Les tokeniseurs, des algorithmes qui décomposent le texte en morceaux numériques lisibles par la machine, sont optimisés pour les langues et dialectes les plus répandus dans leurs données d'entraînement. Des analyses approfondies démontrent que les formes non standard, y compris les dialectes régionaux, entraînent des coûts de segmentation nettement plus élevés.[8] Parce que le tokeniseur ne reconnaît pas le vocabulaire régional comme des mots entiers, il les fragmente en tokens disjoints plus petits (byte-pair tokens).[8]

Cela crée une double pénalité pour la diversité linguistique. Premièrement, cela dégrade la compréhension contextuelle du dialecte par le modèle, car le mot perd sa cohésion

sémantique. Deuxièmement, cela augmente le coût de calcul — et par extension, le coût financier via la tarification des API — du traitement des textes non standard. Cette réalité architecturale crée une friction systémique inhérente contre la diversité linguistique, rendant coûteux pour les entreprises d'IA de prendre en charge autre chose que des formes linguistiques standard et de prestige.[8]

La Taxe d'Alignement et l'Aplatissement Culturel

Au-delà des mécanismes de traduction, les systèmes d'IA standardisent activement la façon dont les gens raisonnent et communiquent, conduisant à ce que les chercheurs de l'Université de Californie du Sud appellent l'« homogénéisation culturelle ».[1] Parce que les modèles subissent un alignement de sécurité et de qualité en utilisant des techniques comme l'apprentissage par renforcement à partir de la rétroaction humaine (RLHF), ils sont entraînés pour produire des résultats que les évaluateurs humains jugent utiles, polis et universellement acceptables.[10] Cependant, ces évaluateurs sont souvent formés pour appliquer des critères standardisés et neutres qui pénalisent les expressions familières, l'argot régional ou les points de vue culturellement spécifiques qui s'écartent du consensus dominant.[1]

Ce processus d'alignement génère une boucle auto-renforçante de langage « aplati ». Les modèles d'IA privilégient l'efficacité, la généralisation et le consensus large.[10] Lorsque les utilisateurs humains interagissent avec ces modèles, ils font face à un compromis documenté entre la « fidélité » (la précision avec laquelle l'IA reflète l'intention spécifique de l'utilisateur) et le « coût de communication » (le temps et les efforts nécessaires pour réviser l'invite).[11] Parce que la majorité des utilisateurs privilégient la productivité, ils acceptent les résultats initiaux généralisés générés par l'IA.[11] À mesure que ce contenu généré par l'IA, hautement générique, inonde Internet, il devient les données d'entraînement pour la prochaine génération de modèles. Au fil du temps, cette boucle de rétroaction récursive restreint continuellement le spectre de l'expression humaine à un vernaculaire homogénéisé et « moyen », dépouillant la langue de sa puissance, de son émotion et de sa saveur régionale.[10]

Mécanisme d'homogénéisation	**Description technique**	**Impact sociolinguistique**
Déséquilibre des données et Scraping	Plus de 90 % des données d'entraînement des modèles de pointe proviennent de sources anglaises/occidentales, créant un espace latent dominant.[1]	Les résultats non anglais reflètent les normes culturelles anglo-américaines, érodant les épistémologies et les perspectives locales.[6]

Traduction pivot implicite	Les LLM multilingues acheminent le raisonnement complexe à travers une langue dominante (l'anglais) dans les couches neuronales intermédiaires.[4]	Perte d'idiomes spécifiques à la culture et de concepts intraduisibles ; les résultats semblent aliénants ou techniquement peu naturels.[6]
Friction de tokenisation	Les dialectes non standard et les langues minoritaires sont fortement fragmentés lors de la tokenisation en raison d'un manque de représentation.[8]	Coût de calcul accru et compréhension contextuelle réduite pour les dialectes régionaux, pénalisant les entrées diverses.[8]
Standardisation de l'alignement	Le RLHF privilégie les résultats sûrs, universels et de « langue de prestige » au détriment des variations dialectales localisées et historiquement riches.[9]	Subordination des dialectes ; les utilisateurs sont contraints de changer de code vers des langues standard pour être compris, renforçant le biais.[9]

2. L'Étude de Cas du Français Québécois

La menace de l'homogénéisation linguistique est particulièrement visible dans le contexte du français québécois. Le français québécois n'est pas simplement un sous-ensemble accentué du français métropolitain (parisien) ; il possède un vocabulaire distinct, des structures syntaxiques uniques, des idiomes historiques et un riche lexique d'expressions profondément liées à l'histoire de la province, à son environnement rigoureux et à son évolution sociale. Lorsqu'il est traité par des modèles conçus dans la Silicon Valley et entraînés principalement sur des données en français européen, l'essence culturelle québécoise est systématiquement marginalisée, entraînant un phénomène que les linguistes et les sociologues qualifient de « colonisation linguistique ».[13]

Visions Néolibérales du Monde et Effacement des Ancrages Culturels

La technologie n'est jamais neutre ; elle porte intrinsèquement la vision du monde de ses créateurs. Les recherches soulignées par Radio-Canada indiquent que l'IA générative contemporaine développe une vision du monde fortement enracinée dans un contexte « néolibéral, technophile, très masculin et sous influence américaine ».[13] Cette vision du monde a tendance à aplatir les perspectives culturelles diverses qui ne s'alignent pas avec son influence dominante. Les modèles d'IA standard produisent principalement une version du français qui adhère strictement à la standardisation parisienne, classant fréquemment les régionalismes

québécois comme des erreurs, des fautes de frappe ou des anomalies de données marginales.[13]

Un exemple principal de cette perte culturelle est l'incapacité des modèles de pointe à comprendre ou à générer de manière appropriée *les sacres* (les jurons québécois traditionnellement enracinés dans la terminologie de l'Église catholique). Ces expressions sont bien plus que de simples blasphèmes ; ce sont des marqueurs linguistiques complexes qui reflètent l'évolution historique, la rébellion et la sécularisation de la société québécoise pendant la Révolution tranquille.[13] En filtrant ces termes par le biais de directives de sécurité, ou en ne les reconnaissant pas du tout, les modèles d'IA suppriment activement une fenêtre vitale sur l'histoire culturelle commune du Québec.[13] Parce que les utilisateurs réalisent que l'IA ne peut pas comprendre leur dialecte naturel, ils modifient leur propre comportement linguistique, adoptant le français métropolitain standard simplement pour être compris par la machine. Ce changement de code forcé renforce par inadvertance la priorité accordée par le modèle au registre standard, accélérant la disparition du dialecte local.[9]

Preuves Empiriques : Les Références QFrCoRE et QFrCoRT

L'échec des LLM à traiter le dialecte québécois a été quantifié de manière empirique grâce à l'introduction récente de deux ensembles de données de référence spécialisés développés par le Groupe de Recherche en Intelligence Artificielle de l'Université Laval (GRAIL) : le Corpus Franco-Québécois des Expressions Régionales (QFrCoRE) et le Corpus Franco-Québécois des Termes Régionaux (QFrCoRT).[14]

La méthodologie de construction de ces corpus a impliqué l'extraction manuelle d'idiomes à partir de dictionnaires québécois spécialisés et de ressources lexicales en ligne.[15] Pour tester rigoureusement les modèles, les chercheurs ont utilisé GPT-4o-mini pour générer des distracteurs sémantiquement plausibles mais incorrects, mesurant la similarité à l'aide de BLEU, ROUGE et BERTScore pour s'assurer que l'IA ne pouvait pas simplement deviner en se basant sur des indices contextuels.[15]

- **QFrCoRE** comprend 4 633 instances d'expressions idiomatiques composées de plusieurs mots, telles que *"attache ta tuque avec de la broche"* (qui signifie métaphoriquement se préparer à ce qui s'en vient).[17]
- **QFrCoRT** contient 171 mots idiomatiques régionaux, tels que *"Tiguidou!"* (une expression signifiant que tout est parfait).[17]

Lorsque les chercheurs ont évalué 111 LLM distincts par rapport à ces références dans un contexte *zero-shot*, les résultats ont révélé une fracture dialectale critique et systémique.[17] Bien que les modèles aient obtenu des performances adéquates sur le français métropolitain standard, le chiffre effarant de 65,77 % des modèles a obtenu des résultats nettement inférieurs sur les idiomes québécois.[18] Seuls 9,0 % des modèles testés ont montré une compétence favorisant le dialecte régional.[18] La disparité n'est pas due à la complexité syntaxique, car les modèles ont échoué de la même manière sur des mots simples et des

phrases complètes (une différence de score moyenne de seulement 5,5 % entre les deux corpus) ; l'échec provient plutôt entièrement d'un manque de vocabulaire régional et culturel dans les données d'entraînement.[17] Alors qu'un groupe très restreint de 27 modèles frontières massifs (tels que GPT-5.1, Gemini-2.5-pro et DeepSeek-reasoner) a atteint une précision supérieure à 80 %, la grande majorité des modèles open-source plus petits et accessibles ont connu une grave dissonance socio-linguistique.[17] Radio-Canada a corroboré ces conclusions, notant qu'environ 40 % des modèles de langage existants ne comprennent tout simplement pas les *expressions québécoises*.[13]

Modèles Frontières : ChatGPT, Claude et les Erreurs de Traduction

Lorsque les utilisateurs sollicitent des modèles frontières comme ChatGPT et Claude avec des contextes québécois très spécifiques ou des idiomes inventés, l'effet d'homogénéisation devient particulièrement visible. Des expériences testant ChatGPT, Gemini et Claude sur des idiomes régionaux ou inventés révèlent des approches disparates mais imparfaites : certains modèles hallucinent des histoires entièrement fausses pour les idiomes, tandis que d'autres sur-expliquent les concepts en utilisant des formulations très académiques et peu naturelles plutôt que le dialecte conversationnel.[19]

Bien que Claude AI soit réputé pour disposer d'un mécanisme d'attention sophistiqué qui saisit très bien la linguistique et la syntaxe du français standard, il adopte fréquemment des réponses de « prestige ».[9] Si un utilisateur demande à Claude de traduire une phrase québécoise très informelle et chargée d'émotion en anglais, puis lui demande de la retraduire, le modèle renvoie rarement la phrase sous sa forme québécoise d'origine ; à la place, il produit un équivalent parisien aseptisé. Cela démontre que la représentation interne de la langue par le modèle manque de la profondeur localisée nécessaire pour maintenir la continuité dialectale, traitant la variante régionale comme une entrée à corriger plutôt que comme une sortie valide à générer.

La Prolifération des Structures Anglicisées (*Calques*)

L'effacement du vocabulaire régional est aggravé par le phénomène de la syntaxe anglicisée, où l'IA produit un vocabulaire français plaqué de force sur des structures grammaticales anglaises — communément appelées dans les cercles linguistiques *calques de l'anglais* ou anglicismes syntaxiques.[21] Parce que les LLM utilisent fréquemment l'anglais comme langue pivot implicite lors de la génération interlingue, leurs sorties en français reflètent structurellement la construction des phrases anglaises.[4]

Dans le contexte québécois, où la préservation de la langue française face à l'écrasante influence anglophone environnante est une lutte quotidienne, existentielle, politique et culturelle, les *calques* générés par l'IA accélèrent activement la dégradation linguistique. Par exemple, les modèles d'IA génèrent fréquemment des phrases comme *"prendre un cours"* (une traduction directe de l'anglais "to take a course") au lieu du français correct *"suivre un cours"*.[22] De même, les émissions de radio "open line" sont traduites littéralement par *"ligne ouverte"* plutôt que *"tribune téléphonique"*.[22]

De plus, des études sociolinguistiques sur l'évolution de l'utilisation des anglicismes dans le français québécois soulignent la prévalence croissante de verbes anglais morphologiquement non intégrés, insérés directement dans la syntaxe française.[23] Les observations incluent des phrases telles que *"Je vous rassure, on ne voulait pas sneak une proposition"* ou *"Peux-tu bring le trailer quand tu vas venir?"*.[23] Les campagnes de sensibilisation linguistique pointent vers une chronologie de dégradation générationnelle : de *"J'ai appelé ma mère"* (2000), à *"J'ai callé ma mère"* (2010), au *"J'ai call my mom"* totalement non intégré (2030).[23] À mesure que le contenu généré par l'IA normalise ces structures anglicisées dans les chatbots, les courriels automatisés et le marketing numérique, la technologie agit par inadvertance comme un cheval de Troie. Elle contourne les simples contrôles de vocabulaire et attaque directement l'intégrité structurelle et syntaxique de la langue française, laissant les locuteurs natifs avec des phrases « tellement pleines de mots anglais qu'on ne peut même plus comprendre le sens de la phrase ».[24]

Concept Source en Anglais	**Français Correct (Québec/Standard)**	**Anglicisme Généré par l'IA (Calque)**	**Impact Syntaxique et Culturel**
"To take a course"	*Suivre un cours*	*Prendre un cours*	Cartographie sémantique directe supplantant les conventions verbales françaises, dégradant les normes grammaticales.[22]
"Open line" (Radio)	*Tribune téléphonique*	*Ligne ouverte*	Traduction littérale entraînant une perte de sens contextuel et de vocabulaire précis.[22]
"To abuse" (a person)	*Maltraiter*	*Abuser* (un enfant)	Dérive sémantique sévère ; *abuser* en français signifie tromper ou exagérer, et non infliger des blessures physiques.[21]

"Fasten your seatbelt" (Metaphorical)	*Attache ta tuque avec de la broche*	*Attache ta ceinture*	Effacement total de l'idiome régional québécois au profit d'un équivalent standard parisien littéral.[17]
"To call my mom"	*J'ai appelé ma mère*	*J'ai call my mom*	Non-intégration morphologique ; verbes anglais insérés de force dans la syntaxe française sans conjugaison.[23]

3. La Menace Mondiale Pesant sur la Diversité Linguistique

La crise numérique à laquelle est confronté le français québécois n'est qu'un microcosme d'une menace beaucoup plus vaste et catastrophique pour la diversité linguistique mondiale. Sur les quelque 7 000 langues actuellement parlées dans le monde, on estime que 40 % à 43 % sont classées comme menacées ou exposées à un risque d'extinction imminent, avec une langue minoritaire disparaissant toutes les deux semaines.[25] L'intelligence artificielle, dans sa trajectoire de développement actuelle, accélère de manière agressive cette chronologie en créant d'immenses barrières numériques pour les langues non dominantes, un phénomène fondamentalement motivé par l'existence de « vides de données ».

Le Concept de Vides de Données et de Déserts de Données

Les vides de données, ou déserts de données, font référence aux écosystèmes numériques ou aux populations de marché où des ressources textuelles numérisées limitées, fragmentées ou totalement inexistantes empêchent les systèmes d'IA d'apprendre une langue, de produire des prédictions précises ou de comprendre une communauté spécifique.[27] Une fracture numérique flagrante et hautement excluante sépare les langues que l'IA peut « voir » des centaines de langues auxquelles elle reste aveugle.[28]

Les statistiques sont accablantes. Alors que les ChatGPT et Gemini du monde fonctionnent efficacement pour les 1,52 milliard de personnes qui parlent anglais, ils sont nettement moins performants pour les 97 millions de locuteurs vietnamiens, et échouent complètement pour le 1,5 million de personnes qui parlent le nahuatl, une langue uto-aztèque.[2] Même au sein des langues qui comptent des populations massives, le manque de données numérisées crée des déserts artificiels. Par exemple, le swahili compte 200 millions de locuteurs mais manque de ressources numériques et computationnelles suffisantes pour que les modèles d'IA puissent

apprendre, tandis qu'une langue comme le gallois, bien qu'ayant beaucoup moins de locuteurs, bénéficie d'une documentation approfondie et d'efforts de préservation numérique, ce qui lui permet d'obtenir de meilleures performances dans les environnements LLM.[2]

La situation est exceptionnellement grave sur le continent africain, qui compte plus de 2 000 langues, constituant près d'un tiers de la diversité linguistique mondiale. Pourtant, un pourcentage stupéfiant de 88 % des langues africaines est considéré comme « gravement sous-représenté » ou « complètement ignoré » en linguistique computationnelle.[25] La nouvelle référence SAHARA, qui a évalué rigoureusement 517 langues africaines à travers diverses tâches d'IA, a démontré que des langues parlées par des dizaines de millions de personnes — telles que le wolof, le haoussa, l'oromo, le peul et le kinyarwanda — se retrouvent systématiquement parmi les moins performantes dans les métriques de raisonnement, de génération et de classification.[28] Cet écart n'est pas dû à la complexité linguistique, mais plutôt à des décennies de sous-investissement systémique dans les ensembles de données et les infrastructures numériques.[28] De plus, de profondes disparités numériques existent même entre les langues africaines ; les modèles formés sur l'afrikaans obtiennent d'excellentes performances en traduction automatique en raison de ressources numériques abondantes, tandis que les langues autochtones comme l'isiZulu et le Sepedi sont considérablement en retard en termes de précision.[29]

Effacement Épistémique et Extinction Numérique

Lorsqu'une langue existe dans un désert de données, les conséquences vont bien au-delà du simple inconvénient pour l'utilisateur ; il en résulte un « effacement épistémique » — la suppression systémique de la manière spécifique d'une culture de connaître, de catégoriser et d'interagir avec le monde.[30] Pour les jeunes générations qui grandissent en interagissant nativement avec des téléphones intelligents, des assistants numériques et des plateformes basées sur l'IA, l'incapacité de leurs appareils à comprendre leur langue maternelle envoie un message psychologique puissant et destructeur : leur langue est obsolète, non pertinente et n'a pas sa place dans le monde moderne.[31]

Cette tragédie est illustrée de manière poignante par le sort du dialecte shoshone de l'Ouest en Amérique du Nord. Lorsque l'aînée Mae Timbimboo Parry est décédée, elle a emporté avec elle des milliers de mots décrivant les angles précis de la lumière du soleil et les traces d'animaux. Lorsque les membres de la communauté tentent d'utiliser des outils de transcription d'IA modernes pour enregistrer les connaissances écologiques traditionnelles des aînés survivants, les algorithmes renvoient des messages d'erreur, ne parvenant pas du tout à reconnaître l'audio comme étant une langue humaine.[31] De même, la langue ambo en Australie (détenant 60 000 ans de connaissances sur les plantes médicinales) et la langue beeke au Nigeria existent dans un vide numérique total, complètement invisibles pour les algorithmes modernes de synthèse vocale.[31]

Lorsque les modèles d'IA manquent de données d'entraînement suffisantes pour une langue minoritaire, ils ne se dégradent pas avec élégance. Ils hallucinent de manière extravagante,

amplifient les stéréotypes historiques et ne parviennent pas à faire la distinction entre des contextes culturels très divergents.[28] Les tentatives des entreprises technologiques de contourner le désert de données en utilisant la traduction automatique propagent souvent des erreurs. Comme le soulignent des chercheurs de Stanford, la traduction automatique peut produire une phrase qui a du sens mot à mot, mais qui reste « culturellement complètement incorrecte », ne parvenant pas à saisir la nuance de la façon dont la langue est réellement parlée.[32] Par conséquent, s'appuyer uniquement sur des LLM mondialisés pour les langues minoritaires entraîne une grave exclusion socio-économique, empêchant de fait ces populations d'accéder aux opportunités civiques, éducatives, sanitaires et économiques générées par la révolution de l'IA.[2]

Données démographiques mondiales sur les déserts de données	**Statut linguistique et capacité de l'IA**	**Conséquence épistémique**
Swahili vs. Gallois	Le swahili (200 M de locuteurs) échoue en raison de l'absence de corpus numérisé ; le gallois (beaucoup moins de locuteurs) réussit grâce à la préservation numérique.[2]	Souligne que l'exclusion de l'IA est motivée par la disponibilité des données, et non par la taille de la population ou la complexité linguistique.[2]
Référence SAHARA (Afrique)	517 langues africaines testées ; les langues majeures comme le wolof, le haoussa et l'oromo échouent systématiquement dans les tâches de raisonnement.[28]	88 % des langues africaines restent ignorées par la linguistique computationnelle, excluant des millions de personnes des économies numériques.[25]
Shoshone de l'Ouest (États-Unis)	Les outils de transcription par l'IA ne reconnaissent absolument pas l'audio comme étant un langage humain et renvoient des erreurs.[31]	Perte de connaissances écologiques hyper-spécifiques (par exemple, traces d'animaux, angles de la lumière du soleil) qui ne peuvent être traduites.[31]

Nahuatl (Uto-aztèque)	Les principaux modèles frontières échouent complètement à générer ou à raisonner dans la langue malgré 1,5 million de locuteurs.[2]	Oblige les jeunes générations à abandonner la langue au profit de l'espagnol/anglais pour accéder aux outils numériques.[31]

4. La Stratégie de Défense : IA Souveraine et Cadres Politiques

Conscients que la souveraineté linguistique est désormais profondément liée à la souveraineté numérique et infrastructurelle, les gouvernements, les organisations linguistiques et les institutions universitaires mettent en place des stratégies de défense agressives. La résistance à l'homogénéisation algorithmique se manifeste par deux vecteurs principaux : la capitalisation massive d'infrastructures physiques localisées d'« IA souveraine », et la mise en œuvre de boucliers législatifs agressifs et de politiques publiques linguistiques.

L'Essor de l'IA Souveraine au Canada et au Québec

L'IA souveraine fonctionne sur la prémisse fondamentale selon laquelle les nations doivent développer et contrôler des systèmes d'intelligence artificielle alignés sur leurs propres valeurs locales, entraînés sur des corpus de données régionaux et hébergés sur des infrastructures physiques nationales, plutôt que de s'en remettre entièrement aux hyperscalers étrangers basés dans la Silicon Valley.[34] D'ici 2026, les dépenses mondiales en systèmes d'IA souveraine devraient dépasser les 100 milliards de dollars, motivées par le besoin de contrôle stratégique et de résidence des données.[34]

Le gouvernement du Canada a lancé la *Stratégie canadienne sur la puissance de calcul souveraine en IA*, soutenue par un investissement historique de 2 milliards de dollars décrit dans le budget 2024 et le budget 2025, visant à renforcer la capacité nationale de supercalculateurs en IA.[35] Dans ce cadre, le ministère fédéral de l'Intelligence artificielle et de l'Innovation numérique a lancé un appel de propositions pour développer des centres de données d'IA à grande échelle d'une capacité supérieure à 100 mégawatts.[37] Ces installations souveraines privilégient la participation autochtone, la stricte résidence des données, l'utilisation d'énergie à faible émission de carbone et les retombées économiques sur l'écosystème.[37]

La province de Québec s'est positionnée à l'avant-garde de ce mouvement, tirant parti de manière agressive de ses vastes ressources hydroélectriques, de son climat frais et de son écosystème de recherche en IA préexistant. Le gouvernement provincial a récemment accordé 36 millions de dollars à Mila (l'Institut québécois d'intelligence artificielle) pour renforcer son réseau d'excellence et accélérer le développement éthique d'une IA alignée sur les intérêts stratégiques et linguistiques du Québec.[39] De plus, Mila a conclu un partenariat stratégique

avec 5C et Hypertec pour développer un vaste pôle de recherche souverain en IA de 250 millions de dollars au siège mondial d'Hypertec à LaSalle, au Québec.[40] Ce campus de nouvelle génération est explicitement conçu pour fournir jusqu'à 3 MW de capacité de calcul sécurisée de pointe (utilisant des GPU NVIDIA, AMD et Intel) refroidis par des technologies avancées d'immersion et de récupération de chaleur appliquée.[40]

Cette infrastructure souveraine permet aux chercheurs québécois d'entraîner et d'exécuter des inférences sur des modèles localisés sans céder des informations exclusives ou culturellement sensibles à des nuages informatiques étrangers.[40] Des organisations comme Scale AI financent conjointement des projets au sein de la *Zone économique métropolitaine* (ZEM) de la grande région de Québec, confiant des fonds pour s'assurer que les solutions d'IA développées localement donnent la priorité aux réalités économiques et sociales locales, consolidant ainsi la souveraineté numérique.[42]

En utilisant l'infrastructure de l'IA souveraine, les développeurs peuvent intentionnellement organiser des ensembles de données d'entraînement qui intègrent la littérature québécoise, le journalisme local, les archives parlementaires et les archives d'organisations culturelles comme l'ADISQ (Association québécoise de l'industrie du disque, du spectacle et de la vidéo). En outre, grâce au microréglage efficace en paramètres (PEFT) et à l'adaptation de bas rang (LoRA), les chercheurs peuvent pré-entraîner continuellement des modèles sur des dialectes régionaux avec des budgets de calcul serrés, forçant mathématiquement le modèle à tracer ses voies neuronales selon la syntaxe, le vocabulaire et les idiomes culturels québécois.[43]

Boucliers Législatifs : La Loi 96 et la *Francophonie*

L'infrastructure technologique est simultanément renforcée par des cadres politiques robustes et inflexibles. Au Québec, la modernisation de la *Charte de la langue française* (communément appelée loi 96 ou *Loi sur la langue officielle et commune du Québec, le français*) a établi des exigences juridiques strictes conçues pour isoler la langue de l'érosion numérique et corporative.[45] La législation affirme fermement le français non seulement comme langue officielle, mais comme la *langue commune* de la nation québécoise, imposant un strict « devoir d'exemplarité » à l'administration civile d'utiliser exclusivement le français dans les services publics, avec des exceptions très spécifiques et étroites pour les communautés autochtones et les immigrants récents.[47]

Fait crucial, le Québec a introduit une législation exigeant la « découvrabilité et l'accès aux contenus culturels originaux de langue française dans l'environnement numérique » en modifiant la Charte des droits et libertés de la personne.[48] Ce vaste mandat exige que les plateformes numériques, les téléviseurs et les appareils connectés s'assurent que leurs interfaces par défaut et leurs algorithmes de diffusion de contenu donnent la priorité au français.[48] Pour les agents d'IA, les outils d'entreprise alimentés par les LLM et les chatbots automatisés déployés par les entreprises opérant au Québec, cela représente une obligation de conformité stricte : les modèles ne peuvent pas simplement traduire des pensées anglaises en un français cassé et anglicisé ; ils doivent fondamentalement fonctionner dans un français

de haute qualité et culturellement précis pour éviter de lourdes pénalités réglementaires et la perte d'accès au marché.[45]

Des institutions comme l'*Office québécois de la langue française* (OQLF) interviennent également activement pour normaliser le vocabulaire de la révolution de l'IA elle-même afin d'empêcher l'importation massive du jargon technologique anglais. En définissant et en promouvant officiellement des termes tels que *algorithme de recommandation*, *voxto* et *logiciel d'intelligence artificielle*, et en fournissant des lexiques complets via la *Vitrine linguistique*, l'OQLF s'assure que le discours entourant l'IA reste profondément ancré dans la langue française.[49] Le Conseil de l'innovation du Québec soutient en outre cela en menant une réflexion collective sur les impacts sociétaux de l'IA, évaluant spécifiquement ses effets sur la croissance et le développement de la langue et de la culture nationales par le biais de rapports complets tels que *Prêt pour l'IA*.[52]

Sur la scène diplomatique mondiale, l'*Organisation internationale de la Francophonie* (OIF) a élevé la diversité linguistique dans l'IA au rang de pilier principal de son mandat. Représentant 93 États et gouvernements (comprenant 300 millions de locuteurs sur cinq continents), l'OIF mène un lobbying agressif dans les forums internationaux, y compris les Nations Unies, pour s'assurer que la diversité linguistique est codifiée dans les traités internationaux de gouvernance de l'IA.[55] Lorsque les cadres de la technologie se vantent d'ajouter des traductions automatisées de 110 langues à leurs plateformes, les conseillers en politique numérique de la Francophonie repoussent ces manœuvres de relations publiques.[57] Ils soutiennent que la traduction automatisée est insuffisante pour empêcher l'effacement épistémique ; par conséquent, la diversité linguistique structurelle profonde et l'entraînement sur des données localisées doivent devenir la « colonne vertébrale inégociable de la politique numérique ».[30]

Vecteur de défense	**Action / Initiative**	**Objectif stratégique**
Infrastructure souveraine	Pôle de recherche souverain en IA Hypertec/Mila de 250 millions de dollars à LaSalle, Québec.[40]	Fournir une capacité de calcul nationale pour entraîner en toute sécurité des modèles sur des données locales sans dépendre des géants technologiques étrangers.[40]
Stratégie fédérale de calcul	Stratégie canadienne sur la puissance de calcul souveraine en IA de 2	Construire une infrastructure nationale de supercalculateurs pour

	milliards de dollars.[35]	garantir la compétitivité mondiale à long terme et la souveraineté des données.[36]
Mandats législatifs	Loi 96 et loi sur la découvrabilité (modifiant la Charte des droits et libertés).[46]	Forcer les plateformes numériques et les outils d'IA d'entreprise à fonctionner en français de haute qualité par défaut, en imposant un devoir d'exemplarité.[47]
Contrôle terminologique	Définitions de l'OQLF pour le vocabulaire de l'IA (ex. *algorithme de recommandation*).[50]	Empêcher l'invasion lexicale du jargon technologique anglais et maintenir l'intégrité structurelle de la langue française.[51]
Diplomatie mondiale	Résolutions de l'OIF aux Nations Unies exigeant une diversité numérique structurelle.[56]	Combattre l'impérialisme numérique et s'assurer que les langues non dominantes sont intégrées dans les cadres internationaux de gouvernance de l'IA.[57]

5. Perspectives Analytiques Contrastées

Le débat sur la manière de gérer l'impact profond de l'IA sur la diversité linguistique révèle de profondes tensions philosophiques, économiques et techniques au sein de la communauté mondiale de l'intelligence artificielle.

Efficacité Mondiale vs Effacement Culturel

Du point de vue de l'ingénierie, des entreprises et de l'économie, la centralisation des modèles d'IA autour d'une langue dominante (l'anglais) est très efficace. La création d'un modèle monolithique unique et massif qui utilise l'anglais comme pivot sémantique universel permet aux entreprises hyperscale de déployer leurs technologies à l'échelle mondiale à des vitesses sans précédent. Cela abaisse radicalement les barrières à l'entrée pour le développement de nouvelles fonctionnalités, simplifie le processus d'alignement complexe (RLHF) et évite les coûts exorbitants, souvent prohibitifs, associés à la collecte, la curation et le nettoyage d'ensembles de données massifs pour des milliers de langues à faibles ressources. Les partisans de cette approche centralisée soutiennent que l'utilisation de la traduction automatisée comme pont démocratise en réalité l'accès à la connaissance.[32] Si une IA peut traduire instantanément la somme totale des connaissances médicales, scientifiques et

mathématiques humaines de l'anglais vers une langue minoritaire, elle autonomise immédiatement cette communauté, même si la nuance culturelle de la traduction est, de l'aveu général, imparfaite.[32]

À l'inverse, les sociolinguistes, les éthiciens et les défenseurs de la culture considèrent cette quête incessante d'efficacité comme une forme d'impérialisme algorithmique. Ils soutiennent que la langue n'est pas simplement un canal neutre pour la transmission de données ; c'est un cadre cognitif profond qui façonne fondamentalement la vision du monde d'une culture. Lorsqu'une IA traite un concept autochtone, africain ou régional québécois à travers le prisme philosophique et linguistique de la Silicon Valley avant de le retraduire, elle altère la signification profonde de la pensée.[6] L'efficacité, de ce point de vue, est obtenue au prix de l'effacement violent de l'épistémologie locale, garantissant que les langues minoritaires ne survivent que sous la forme de traductions syntaxiques vidées de leur substance des idéaux anglo-américains.[30]

Le Paradoxe de l'Isolement Souverain vs la Collaboration Open-Source Mondiale

La stratégie de défense de l'IA souveraine présente également un paradoxe technique très complexe. Les pays qui investissent massivement dans des centres de données souverains, des réseaux électriques à faible émission de carbone et des modèles localisés (comme la France, le Canada et la province de Québec) cherchent à protéger leurs données sensibles, leur culture et leur indépendance économique contre la monopolisation étrangère.[34]

Cependant, le développement d'une véritable IA de pointe nécessite des données et une collaboration à l'échelle planétaire. Comme le soulignent les analystes en politique de l'IA, « la souveraineté en IA fait face à un paradoxe plus profond, qui rend la collaboration mondiale essentielle, et non facultative ».[41] Une véritable autonomie en matière d'IA nécessite une participation active aux écosystèmes ouverts mondiaux. Par exemple, le système AlphaFold, lauréat du prix Nobel 2024, qui a révolutionné la découverte de médicaments, n'a été rendu possible que grâce à une base de données moléculaire mondiale en libre accès.[41] Aucune nation n'aurait pu assembler cet ensemble de données de manière isolée.

Développer un modèle de langage entièrement isolé sur des données québécoises ou françaises pourrait parfaitement préserver le dialecte, mais risque sérieusement de produire un modèle intellectuellement retardé, dépourvu des vastes capacités de raisonnement scientifique, mathématique et de codage intégrées dans des ensembles de données mondiaux plus vastes.[41] Par conséquent, les experts suggèrent que la voie à suivre la plus viable réside dans une « troisième voie » : tirer parti de modèles fondamentaux open source massifs (tels que LLaMA) et utiliser une capacité de calcul souveraine localisée pour affiner ces modèles en utilisant le microréglage efficace en paramètres (PEFT) sur des corpus régionaux de haute qualité.[41] Cette approche hybride permet à une région de télécharger l'intelligence mondiale, de l'affiner en toute sécurité derrière des pare-feu locaux à l'aide de données culturelles, et d'exécuter des inférences localement, équilibrant avec succès la puissance de calcul mondiale

et la fidélité culturelle locale.[41]

6. Termes Clés Bilingues (Glossaire)

Pour naviguer à l'intersection complexe de l'ingénierie de l'intelligence artificielle, de la linguistique computationnelle et de la préservation législative, un vocabulaire standardisé est essentiel. Le glossaire suivant fournit les équivalents anglais et français de termes hautement techniques et conceptuels pertinents pour ce domaine, reflétant la terminologie standardisée officiellement approuvée par des institutions telles que l'OQLF.[50]

Terme Anglais	Équivalent Français	Contexte / Définition
Algorithmic Homogenization	*Homogénéisation algorithmique*	Le processus par lequel les modèles d'IA réduisent la variance dans la communication humaine, standardisant les résultats pour refléter une culture dominante (généralement anglo-américaine).[1]
Sovereign AI	*IA souveraine*	Infrastructures physiques de l'IA (centres de données, GPU) et modèles développés, hébergés et contrôlés au sein d'une nation ou d'une région spécifique pour protéger les données, la sécurité et les valeurs culturelles locales.[34]
Data Desert / Data Void	*Désert de données / Vide de données*	Marchés numériques ou populations linguistiques ne disposant pas d'assez de textes numérisés (ou fragmentés) pour entraîner des modèles d'IA précis et culturellement conscients.[27]
Parameter-Efficient	*Microréglage efficace en*	Une méthode de calcul

Fine-Tuning (PEFT)	*paramètres*	utilisée pour adapter de grands modèles pré-entraînés à des dialectes régionaux à faibles ressources (comme le français québécois) en ne mettant à jour qu'environ 1 % des paramètres, permettant d'économiser des coûts de calcul.[43]
Implicit Translation Mechanism	*Mécanisme de traduction implicite*	Le phénomène algorithmique par lequel un LLM traite les entrées non anglaises en cartographiant et en raisonnant de manière interne à travers un espace latent dominé par l'anglais.[4]
Syntactic Anglicism (Calque)	*Anglicisme syntaxique (Calque)*	La traduction directe et littérale d'une phrase anglaise en français qui viole les structures grammaticales françaises traditionnelles (ex. *prendre un cours*).[21]
Recommendation Algorithm	*Algorithme de recommandation*	Un algorithme prédictif qui fournit des résultats personnalisés ; fortement réglementé par les nouvelles lois québécoises sur la découvrabilité numérique.[51]
AI Alignment	*Alignement de l'IA*	Le processus itératif visant à garantir que le comportement d'un modèle d'IA correspond aux intentions humaines, aux filtres de sécurité et aux directives éthiques ;

		souvent à l'origine de l'aplatissement dialectal.[50]
Epistemic Erasure	*Effacement épistémique*	La perte totale de la manière spécifique d'une culture de connaître, de conceptualiser et de décrire le monde en raison de l'exclusion numérique et algorithmique chronique.[30]
Voice Memo	*Voxto*	Un terme inventé par l'OQLF faisant référence à un court enregistrement vocal envoyé via un appareil mobile, standardisant le vocabulaire de la communication numérique.[51]

7. Architecture Bilingue de l'Article

La recherche exhaustive compilée dans ce rapport sert de base architecturale à un article d'investigation de fond captivant. Pour s'assurer que l'article final résonne parfaitement auprès des lectorats anglophone et francophone — tout en satisfaisant les exigences des analystes technologiques et des défenseurs de la préservation culturelle — l'accroche narrative et le plan structurel suivants sont proposés.

Accroche Narrative / Narrative Hook

(Français) : *"À l'ère numérique, une langue meurt deux fois. D'abord, lorsque le dernier aîné cesse de la parler. Ensuite, lorsque les algorithmes refusent catégoriquement de la reconnaître. Alors que les modèles d'intelligence artificielle conçus dans la Silicon Valley deviennent rapidement les gardiens ultimes des connaissances et des communications mondiales, ils standardisent silencieusement la façon dont l'humanité pense, écrit et crée. Pour le Québec, une province définie par sa lutte existentielle et séculaire pour sa survie linguistique, la menace principale n'est plus l'assimilation physique par une nation anglophone environnante, mais l'effacement invisible par un algorithme. Quand ChatGPT ne comprend pas un sacre historique, ou lorsqu'une IA d'entreprise génère un français structuré entièrement par une grammaire anglaise, ce n'est pas un simple bogue informatique : c'est la colonisation algorithmique d'une culture. Le combat pour l'avenir de la langue française s'est déplacé des rues vers les centres de données."*

(English) : "In the digital age, a language dies twice. First, when the last elder stops speaking it.

Second, when the algorithms refuse to recognize it. As artificial intelligence models built in Silicon Valley rapidly become the ultimate gatekeepers of global knowledge and communication, they are silently standardizing the way humanity thinks, writes, and creates. For Quebec, a province defined by its centuries-long, existential struggle for linguistic survival, the primary threat is no longer physical assimilation by a surrounding Anglophone nation, but invisible erasure by an algorithm. When ChatGPT cannot understand a historical *sacre*, or when an enterprise AI outputs French structured entirely by English grammar, it is not merely a software glitch—it is the algorithmic colonization of a culture. The fight for the future of the French language has moved from the streets to the server farms."

Plan de l'Article / Article Outline

1. **L'effacement invisible (The Invisible Erasure) :**
    - Ouvrir avec le phénomène du texte d'IA « aplati » et le concept de colonisation linguistique. Utiliser l'échec empirique de 65 % des LLM sur la référence QFrCoRE (par exemple, l'incapacité de comprendre *"attache ta tuque"*) comme l'incident déclencheur pour démontrer l'écart massif entre la technologie mondiale et la réalité locale.
2. **Dans la boîte noire (Inside the Black Box) :**
    - Plonger au cœur des mécanismes techniques. Expliquer les réalités architecturales du « Désert de Données », de la friction de tokenisation, et comment le « biais de pivot vers l'anglais » force les pensées non anglaises à traverser un filtre néolibéral et anglo-américain. Fournir des exemples concrets de la façon dont Claude et ChatGPT hallucinent ou sur-expliquent des idiomes, entraînant des *calques de l'anglais* qui dégradent l'intégrité syntaxique du français.
3. **Les répercussions mondiales (The Global Fallout) :**
    - Élargir la portée pour montrer que le Québec ne se bat pas seul. Discuter de la référence SAHARA et de la menace existentielle immédiate que représente l'IA pour les langues africaines et autochtones (comme le Shoshone de l'Ouest et l'Ambo). Présenter le problème comme une injustice épistémique mondiale et une extinction numérique.
4. **Bâtir la forteresse numérique (Building the Digital Fortress) :**
    - Détailler la farouche résistance. Mettre en évidence le vaste campus d'IA souveraine Hypertec/Mila de 250 M$ à LaSalle, la stratégie de calcul souverain fédérale de 2 milliards de dollars, et le bouclier législatif fourni par la loi 96 et les contrôles terminologiques de l'OQLF. Présenter l'IA souveraine non pas comme un isolationnisme politique, mais comme une infrastructure obligatoire pour la survie culturelle.
5. **Conclusion : L'avenir de la pensée (The Future of Thought) :**
    - Conclure sur les enjeux philosophiques et géopolitiques. Réitérer que la préservation de la diversité linguistique dans l'intelligence artificielle ne consiste pas simplement à traduire correctement des mots pour une interface utilisateur ; il s'agit de s'assurer que l'avenir de l'interaction homme-machine reflète la multiplicité véritable et vibrante de l'expérience humaine. Les algorithmes doivent être contraints

d'apprendre à parler nos langues, de peur que l'humanité ne soit contrainte de penser exclusivement dans la leur.

Works cited

1. AI is changing more than your writing — it may be shaping your worldview - USC Dornsife, accessed on April 17, 2026, https://dornsife.usc.edu/news/stories/ai-may-promote-cultural-homogenization/
2. How AI is leaving non-English speakers behind | Stanford Report, accessed on April 17, 2026, https://news.stanford.edu/stories/2025/05/digital-divide-ai-llms-exclusion-non-english-speakers-research
3. How language gaps constrain generative AI development - Brookings Institution, accessed on April 17, 2026, https://www.brookings.edu/articles/how-language-gaps-constrain-generative-ai-development/
4. Language Dominance in Multilingual Large Language Models - ACL Anthology, accessed on April 17, 2026, https://aclanthology.org/2025.blackboxnlp-1.7.pdf
5. PLUG: Leveraging Pivot Language in Cross-Lingual Instruction Tuning - ACL Anthology, accessed on April 17, 2026, https://aclanthology.org/2024.acl-long.379.pdf
6. Do Large Language Models Have an English Accent? Evaluating and Improving the Naturalness of Multilingual LLMs - Apple Machine Learning Research, accessed on April 17, 2026, https://machinelearning.apple.com/research/english-accent
7. Discrimination by LLMs: Cross-lingual Bias Assessment and Mitigation in Decision-Making and Summarisation - arXiv, accessed on April 17, 2026, https://arxiv.org/html/2509.09735v1
8. Which English Do LLMs Prefer? Quantifying American and British English Through a Postcolonial Lens | OpenReview, accessed on April 17, 2026, https://openreview.net/forum?id=cbh3tMZHdx
9. Sociolinguistic Dynamics on Digital Platforms - Emergent Mind, accessed on April 17, 2026, https://www.emergentmind.com/topics/sociolinguistic-dynamics-of-digital-platforms
10. A.I. Is Homogenizing Our Thoughts : r/technology - Reddit, accessed on April 17, 2026, https://www.reddit.com/r/technology/comments/1lkzzev/ai_is_homogenizing_our_thoughts/
11. AI from AI: a Future of Generic and Biased Online Content? - UCLA Anderson Review, accessed on April 17, 2026, https://anderson-review.ucla.edu/ai-from-ai-a-future-of-generic-and-biased-online-content/
12. The Homogenization of Language: How AI Could Flatten Our Words and Thoughts, accessed on April 17, 2026,

https://haixun.medium.com/the-homogenization-of-language-how-ai-could-flatten-our-words-and-thoughts-c6526804c746
13. Protéger la diversité des langues face à l'IA | Radio-Canada, accessed on April 17, 2026, https://ici.radio-canada.ca/nouvelle/2246151/diversite-langue-quebecoise--ia-intelligence-artificielle
14. Paper page - A Set of Quebec-French Corpus of Regional Expressions and Terms, accessed on April 17, 2026, https://huggingface.co/papers/2510.05026
15. [Literature Review] A Set of Quebec-French Corpus of Regional Expressions and Terms, accessed on April 17, 2026, https://www.themoonlight.io/en/review/a-set-of-quebec-french-corpus-of-regional-expressions-and-terms
16. Quebec-French Idiom Benchmark Datasets | PDF - Scribd, accessed on April 17, 2026, https://www.scribd.com/document/929479411/2510-05026v1
17. A Set of Quebec-French Corpus of Regional Expressions and ... - arXiv, accessed on April 17, 2026, https://arxiv.org/pdf/2510.05026
18. [2510.05026] Idiom Understanding as a Tool to Measure the Dialect Gap - arXiv, accessed on April 17, 2026, https://arxiv.org/abs/2510.05026
19. I invented a fake idiom to test AI chatbots — only one called my bluff | Tom's Guide, accessed on April 17, 2026, https://www.tomsguide.com/ai/i-invented-a-fake-idiom-to-test-ai-chatbots-only-one-called-my-bluff
20. "Wait this is fucking insane - Claude immediately guessed I was French" : r/ClaudeAI, accessed on April 17, 2026, https://www.reddit.com/r/ClaudeAI/comments/1hs7t3p/wait_this_is_fucking_insane_claude_immediately/
21. Dites-le en français | United Nations, accessed on April 17, 2026, https://www.un.org/en/node/67818
22. Anglicismes : Éviter les emprunts inutiles | PDF | Langue française - Scribd, accessed on April 17, 2026, https://fr.scribd.com/document/395800780/Anglicismes
23. Investigating attitudes towards a changing use of anglicisms in Quebec French | Canadian Journal of Linguistics/Revue canadienne de linguistique, accessed on April 17, 2026, https://www.cambridge.org/core/journals/canadian-journal-of-linguistics-revue-canadienne-de-linguistique/article/investigating-attitudes-towards-a-changing-use-of-anglicisms-in-quebec-french/1B1BA00CA929A3F32C4235E848008223
24. Investigating attitudes towards a changing use of anglicisms in Quebec French, accessed on April 17, 2026, https://muse.jhu.edu/article/961875
25. How the African Languages Lab empowers low-resource languages - Smartling, accessed on April 17, 2026, https://www.smartling.com/blog/african-languages-lab-empowers-low-resource-languages-with-ai
26. Embracing AI to preserve dying languages - FairPlanet, accessed on April 17, 2026,

https://www.fairplanet.org/story/embracing-artificial-intelligence-to-preserve-dying-languages/
27. How algorithmic data deserts exclude consumers - MIT Sloan, accessed on April 17, 2026, https://mitsloan.mit.edu/ideas-made-to-matter/how-algorithmic-data-deserts-exclude-consumers
28. When AI Can't Understand Your Language, Democracy Breaks Down | TechPolicy.Press, accessed on April 17, 2026, https://www.techpolicy.press/when-ai-cant-understand-your-language-democracy-breaks-down-/
29. Bridging the AI Divide: DSFSI Drives Multilingual Knowledge Access at UP's 'Abstracts into Indigenous Voices' Event | University Of Pretoria, accessed on April 17, 2026, https://www.up.ac.za/afridsai/news/bridging-ai-divide-dsfsi-drives-multilingual-knowledge-access-ups-abstracts-indigenous-voices
30. From Paris to Prompt: How La Francophonie Quietly Changed AI's Language Priorities, accessed on April 17, 2026, https://multilingual.com/la-francophonie-ai-language-diversity/
31. The AI Bias Nobody Talks About: How Training Data Is Erasing Minority Languages, accessed on April 17, 2026, https://medium.com/@afolabidare50/the-ai-bias-nobody-talks-about-how-training-data-is-erasing-minority-languages-65c81707c9fa
32. Closing the Digital Divide in AI | Stanford HAI, accessed on April 17, 2026, https://hai.stanford.edu/news/closing-the-digital-divide-in-ai
33. Linguistic diversity in AI and ML: Why it's important, accessed on April 17, 2026, https://sigma.ai/linguistic-diversity-in-ai/
34. Sovereign AI: Why Nations are Treating Compute as Critical Infrastructure. - RAISE Summit, accessed on April 17, 2026, https://www.raisesummit.com/post/sovereign-ai-compute-critical-infrastructure
35. Canadian Sovereign AI Compute Strategy - Innovation, Science and Economic Development Canada, accessed on April 17, 2026, https://ised-isde.canada.ca/site/ised/en/canadian-sovereign-ai-compute-strategy
36. Canada launches national initiative to build large-scale AI supercomputing capacity, accessed on April 17, 2026, https://www.canada.ca/en/innovation-science-economic-development/news/2026/04/canada-launches-national-initiative-to-build-large-scale-ai-supercomputing-capacity.html
37. Enabling large-scale sovereign AI data centres, accessed on April 17, 2026, https://ised-isde.canada.ca/site/ised/en/enabling-large-scale-sovereign-ai-data-centres
38. Government of Canada launches call for proposals for large scale sovereign AI data centres, accessed on April 17, 2026, https://www.dlapiper.com/en/insights/publications/2026/02/government-of-canada-launches-call-for-proposals-for-large-scale-sovereign-ai-data-centres
39. The Quebec Government Grants $36M to Mila to Strengthen Artificial Intelligence

Research and Talent, accessed on April 17, 2026, https://mila.quebec/en/news/the-quebec-government-grants-36m-to-mila-to-strengthen-artificial-intelligence-research-and
40. Mila, 5C and Hypertec Announce $250 Million LaSalle Campus and Sovereign AI Research Hub to Strengthen Canada's Role as a Global Leader in AI Innovation, accessed on April 17, 2026, https://mila.quebec/en/news/mila-5c-and-hypertec-announce-250-million-lasalle-campus-and-sovereign-ai-research-hub-to
41. Sovereignty Is Not Solitude: Open Source as Canada's Third Path in AI, accessed on April 17, 2026, https://www.cigionline.org/articles/sovereignty-is-not-solitude-open-source-as-canadas-third-path-in-ai/
42. Québec to support innovative AI projects across the Zone économique métropolitaine (ZEM), accessed on April 17, 2026, https://www.scaleai.ca/quebec-to-support-innovative-ai-projects-in-the-zone-economique-metropolitaine-zem/
43. Low-Resource Dialect Adaptation of Large Language Models: A French Dialect Case-Study - arXiv, accessed on April 17, 2026, https://arxiv.org/html/2510.22747v3
44. [2510.22747] Low-Resource Dialect Adaptation of Large Language Models: A French Dialect Case-Study - arXiv, accessed on April 17, 2026, https://arxiv.org/abs/2510.22747
45. Bill 96 Quebec Language Law Impact on Businesses - Smartcat, accessed on April 17, 2026, https://www.smartcat.com/blog/bill-96/
46. Charter of the French Language - Wikipedia, accessed on April 17, 2026, https://en.wikipedia.org/wiki/Charter_of_the_French_Language
47. Modernization of the Charter of the French language ..., accessed on April 17, 2026, https://www.quebec.ca/en/government/policies-orientations/french-language/modernization-charter-french-language
48. Québec introduces Bill requiring discoverability of French-language content on digital platforms - Dentons Data, accessed on April 17, 2026, https://www.dentonsdata.com/quebec-introduces-bill-requiring-discoverability-of-french-language-content-on-digital-platforms/
49. OQLF - Vocabulaire de l'intelligence artificielle - Gouvernement du Québec, accessed on April 17, 2026, https://www.oqlf.gouv.qc.ca/vocabulaire-intelligence-artificielle
50. vocabulaire intelligence artificielle.pdf - Symboles, accessed on April 17, 2026, https://www.oqlf.gouv.qc.ca/ressources/bibliotheque/dictionnaires/VocabulairesPDF/vocabulaire-intelligence-artificielle.pdf
51. Les 12 termes de l'année 2025 de l'Office québécois de la langue ..., accessed on April 17, 2026, https://www.oqlf.gouv.qc.ca/office/communiques/2025/20251201_douze_termes_annee_2025.aspx
52. SYNTHÈSE DE LA JOURNÉE D'ÉTUDE COMMUNS DU, accessed on April 17, 2026, https://espace.inrs.ca/id/eprint/16327/1/SyntheseCommunsIA.pdf

53. Dans l'œil de l'Obvia : notes de breffage pour les acteurs publics, accessed on April 17, 2026, https://www.obvia.ca/dans-loeil-de-lobvia
54. Intelligence artificielle | Conseil de l'innovation du Québec, accessed on April 17, 2026, https://conseilinnovation.quebec/intelligence-artificielle/
55. IP24083 | Geopolitics of the French Language - RSIS, accessed on April 17, 2026, https://rsis.edu.sg/rsis-publication/idss/ip24083-geopolitics-of-the-french-language/
56. Cooperation between the UN and the 0IF is based on common objectives - France ONU, accessed on April 17, 2026, https://onu.delegfrance.org/cooperation-between-the-un-and-the-0if-is-based-on-common-objectives
57. Mind your language: The battle for linguistic diversity in AI - UN News, accessed on April 17, 2026, https://news.un.org/en/story/2025/03/1161406