Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Évaluer et atténuer les différences de performance des GML pour les tâches de génération de textes français et anglais

Rebecca Knowles Fatemeh Azadi Michel Simard Chi-kiu Lo 羅致翹
Samuel Larkin Marc Tessier Cyril Goutte
Traitement multilingue de textes, Conseil national de recherche Canada, Ottawa, Canada
Prénom.Nom@nrc-cnrc.gc.ca

RÉSUMÉ

Les grands modèles de langue (GML) sont généralement entraînés sur des données linguistiquement déséquilibrées, ce qui mène à de meilleures performances en anglais que dans d'autres langues. Dans cet article, nous construisons manuellement un ensemble de 104 requêtes (appariées en anglais et en français) portant sur des sujets liés à la sécurité et à la réglementation, dans des contextes canadiens. Les requêtes sont associées à des réponses de référence extraites manuellement de sources fiables. Une annotation humaine des sorties de trois GML révèle que les réponses en français souffrent d'un déficit de précision factuelle, comparativement aux réponses en anglais. L'utilisation de la traduction automatique permet d'atténuer cet écart de performance, mais n'est pas toujours une solution acceptable.

ABSTRACT

Evaluating and mitigating LLM performance disparities on French and English text generation tasks

Large language models (LLMs) are typically trained on linguistically unbalanced data, resulting in better performance on tasks in English than on tasks in other languages. In this work we manually curate a dataset of 104 queries (paired English and French) that focus on topics related to safety and regulations in Canadian contexts. The queries are paired with gold standard responses manually extracted from trusted and authoritative sources. We perform human annotation of the resulting outputs for three LLMs and find that LLM responses in French suffer in terms of factual accuracy as compared to English responses. Machine translation can ameliorate this performance disparity, but may not always be an acceptable solution.

MOTS-CLÉS : Grands modèles de langue, sécurité et réglementation, biais linguistique.

KEYWORDS: Large language models, safety and regulation, linguistic bias.

1 Introduction

L'intérêt généralisé pour les outils utilisant des grands modèles de langue (GML) soulève d'importantes questions quant aux performances de ces outils et à leur équité dans les contextes multilingues. Nous savons qu'il existe des disparités de performance entre les langues, en raison des quantités relatives de textes de chaque langue dans les données d'entraînement des GML, que ces textes soient explicitement multilingues ou non (Briakou *et al.*, 2023; Zhang *et al.*, 2023b, entre autres). Dans cette

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 238–258.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

étude, nous montrons que la performance des GML affiche des disparités interlinguistiques lorsqu'il s'agit de répondre à des questions de sécurité et de réglementation en anglais et en français, dans des contextes canadiens. Plus spécifiquement, nous montrons que les trois GML évalués produisent moins de réponses correctes et plus de réponses contenant divers types d'erreurs lorsque la requête est posée en français que lorsqu'elle est posée en anglais. Ces disparités interlinguistiques présentent plusieurs risques : risques de *désinformation*, risques liés à *l'équité* par rapport aux groupes démographiques, risques *juridiques* et *éthiques* liés au déploiement de tels outils, et risques *réputationnels* (perte de confiance), ainsi que les risques liés aux *biais* dans les GML. Notre recherche est liée aux travaux sur les disparités multilingues face à l'utilisation malveillante des GML, mais elle met plutôt l'accent sur les utilisateurs sans intention néfaste. Notre étude sert d'avertissement et de rappel qu'il est essentiel d'évaluer les outils basés sur les GML de façon multilingue, surtout pour les risques liés à la sécurité, même lorsque les modèles prétendent à une couverture multilingue.

Nous décrivons ici la construction d'un ensemble préliminaire de données : des requêtes bilingues (anglais-français) et les réponses correctes correspondantes, extraites manuellement de sources fiables, et pouvant être utilisées pour évaluer les disparités de performance entre les deux langues officielles du Canada. Comme attendu, nous constatons des disparités de performance entre les réponses des GML aux requêtes anglaises et françaises. Les réponses en anglais ont tendance à être plus exactes ; les réponses en français ont tendance à comporter plus d'erreurs. Cependant, la précision factuelle est globalement assez faible, même pour l'anglais, pour les GML que nous avons évalués. Malgré une certaine variation entre les trois GML que nous avons évalués et entre les sujets couverts par nos requêtes, les disparités interlinguistiques sont remarquablement constantes.

2 Travaux connexes

Liu *et al.* (2024) présentent une étude exhaustive des ensembles de données existants, incluant des données d'évaluation multilingues. Liu *et al.* (2025) proposent un corpus multilingue d'évaluation de la véracité des GML à l'aide de questions-réponses. D'autres travaux ont révélé des disparités de performance entre l'anglais et d'autres langues sur une variété de tâches (Lai *et al.*, 2023; Holtermann *et al.*, 2024, entre autres). Shen *et al.* (2024) ont aussi montré que les GML ont une plus grande tendance à générer des réponses dangereuses et non pertinentes dans les langues à faibles ressources que dans les langues mieux dotées. XSAFETY (Wang *et al.*, 2024), un référentiel multilingue d'évaluation de la sécurité, a mis en évidence plus de réponses peu sûres pour les requêtes dans des langues autres que l'anglais que pour celles en anglais.

En ce qui concerne l'évaluation des sorties de GML, FActScore (Min *et al.*, 2023) et ses variantes multilingues (Shafayat *et al.*, 2024; Chataigner *et al.*, 2024; Kim *et al.*, 2024) ont permis d'explorer la décomposition automatique de textes factuels de référence en « faits atomiques » pour évaluer la factualité des GML. Qiu *et al.* (2023) ont proposé une métrique, mFACT, pour évaluer la fidélité dans les résumés non anglophones, et constaté que les GML multilingues avaient tendance à produire plus souvent des réponses fallacieuses dans des langues autres que l'anglais. Dans une étude liée de près à la nôtre, sur la sécurité et les contextes canadiens, Nejadgholi *et al.* (2024) ont révélé des écarts de performance extrêmes entre l'anglais et les langues peu dotées couramment utilisées par les immigrants au Canada lors de requêtes ChatGPT concernant les vaccins.

Constatant des disparités importantes entre les performance en anglais et dans des langues peu dotées sur diverses tâches utilisées pour évaluer les GML, Ahuja *et al.* (2023) évaluent une approche

Requête	Quand les enfants doivent-ils recevoir le vaccin DCT-polio-Hib en Ontario ?
Sous-réponses	(a) *2 mois*, (b) *4 mois*, (c) *6 mois*, (d) *18 mois*

TABLE 1 – Exemple d'une requête et ses quatre sous-réponses vérifiées.

traduction-test (la question en langue source est traduite en anglais avant d'être soumise au GML), qui mène à des améliorations pour les langues peu dotées, mais pas pour le français. Zhang *et al.* (2023a) présentent aussi des résultats mitigés dans un paradigme de traduction-test. D'autres travaux ont exploré diverses approches de « *crosslingual prompting* » ou de traduction (Huang *et al.*, 2023; Wang *et al.*, 2024, entre autres). L'approche *en pipeline* décrite ci-dessous peut être vue comme une extension naturelle, adaptée à la situation considérée ici, où un utilisateur pose une question et attend une réponse dans la même langue (en l'occurrence en français). La réponse doit alors être traduite à nouveau de l'anglais vers la langue source.

3 Données et méthodologie expérimentale

Dans ce travail exploratoire, nous construisons manuellement un ensemble de 104 paires (anglais-français) de requêtes, portant sur des sujets de sécurité et de réglementation spécifiques au contexte canadien. Chaque paire comprend une requête en anglais et une traduction manuelle de cette requête en français [1]. Ces requêtes sont conçues de telle manière que les réponses soient simples à vérifier et, lorsqu'elles comportent plusieurs composantes, faciles à décomposer en un petit ensemble de faits vérifiables individuellement, par exemple, une liste de vaccins. Ainsi, la réponse à la requête du tableau 1 se divise en quatre sous-réponses. La version anglaise présente les mêmes sous-réponses vérifiées. Alors qu'un utilisateur pourrait être plus enclin à demander « Quels vaccins les bébés doivent-ils recevoir ? », les questions de notre ensemble de requêtes sont plus spécifiques et contextualisées : « D'après le calendrier de vaccination systématique de l'Ontario, quels vaccins les bébés doivent-ils recevoir à 15 mois ? ». Les réponses de références sont ainsi plus clairement définies pour les besoins de l'annotation [2]. Notons que, en plus de faciliter l'annotation, ces questions plus précises procurent plus de contexte aux GML, les plaçant dans une position plus favorable pour répondre correctement. On peut donc penser que les performances mesurées ici tendraient à sur-estimer ce que l'on observerait sur des requêtes plus naturelles. Dans certains cas, le processus d'annotation suggère qu'il reste malgré tout des ambiguïtés ou des sous-spécifications. La section 6 décrit des pistes d'améliorations futures et d'extension du jeu de données, de requêtes et de réponses.

Les requêtes couvrent les catégories suivantes, offrant un échantillon géographiquement et linguistiquement diversifié de régions canadiennes : le calendrier de vaccination systématique de l'Ontario (54 requêtes divisées dans notre analyse en deux groupes : celles dont la réponse est une liste de vaccins et celles dont la réponse est une liste d'âges auxquels les vaccins devraient être administrés), la réglementation québecoise sur les pneus d'hiver (10), le programme de délivrance progressive du permis de conduire en Colombie-Britannique (10), le régime d'assurance santé de l'Alberta (10), le code de la route au Nouveau-Brunswick (10), et les règlements sur la chasse et la pêche sportive au Nunavut (10). Chaque ensemble de requêtes s'appuie sur une ou deux sources officielles faisant

1. La traduction a été réalisée par un membre de l'équipe du projet qui est locuteur natif du français et parle couramment l'anglais et le français. Les paires de requêtes ont ensuite été vérifiées par des collègues parlant couramment l'anglais ou le français.

2. Il peut subsister des ambiguïtés, par exemple s'il s'agit d'un autre Ontario que la province canadienne.

autorité (disponibles en ligne et faisant potentiellement partie des données d'entraînement des GML ; voir l'annexe A pour plus de détails).

Nous comparons ici trois modèles : `microsoft/Phi-4-mini-instruct`[3], `ibm-granite/granite-3.2-8b-instruct`[4] et `Qwen/Qwen2.5-7B-Instruct`[5]. Ces modèles varient en taille de 3,8 à 8 milliards de paramètres et listent l'anglais et le français comme langues prises en charge (cf. annexe B). Pour chaque modèle, nous soumettons toutes les requêtes en anglais (EN) et toutes les requêtes en français (FR). Nous soumettons également un troisième ensemble conçu pour atténuer les inégalités de performance : dans l'approche *en pipeline* (FR-EN-FR), nous traduisons les requêtes françaises en anglais à l'aide de la fonction de traduction de Microsoft Word, de « français (Canada) » vers « anglais », nous les soumettons aux GML en anglais, puis traduisons automatiquement la sortie en « français (Canada) ». Cette approche simule un scénario où un utilisateur soumet une requête en français et attend une réponse en français, mais où le système utiliserait en arrière-plan la traduction automatique pour tenter d'éviter une dégradation de performance. Notons que même si nous avons les requêtes anglaises de référence dans notre ensemble, le réalisme de la simulation impose que nous traduisions automatiquement la requête française provenant de notre hypothétique utilisateur. Cette approche de traduction s'est déjà révélée efficace dans plusieurs tâches de TALN (Artetxe *et al.*, 2023) et a aussi été explorée avec les GML (Ahuja *et al.*, 2023). Cependant, des travaux récents ont remis en question ce paradigme, du moins avec certains GML ; voir Intrator *et al.* (2024), entre autres.

Nous utilisons l'outil d'annotation `brat` (Stenetorp *et al.*, 2012) pour annoter chacune des paires requête-réponse (une requête, une réponse GML) en consultant les sous-réponses de référence. Les annotateurs ont appliqué plusieurs annotations à la sortie, mais nous nous concentrons sur deux : CORRECT (une sous-réponse de référence est présente et cette annotation CORRECT est directement liée à l'une des sous-réponses) et ERREUR (une mauvaise réponse est présente), cf. annexe D. Par exemple, si la réponse GML à la requête dans le tableau 1 contient du texte énonçant que les enfants doivent recevoir le vaccin à l'âge de 4 mois, ce texte doit être annoté CORRECT, et cette étiquette doit être liée à la sous-réponse (b). Si la réponse GML dit que les enfants doivent recevoir ce vaccin à l'âge de 2 ans, ce passage dans la réponse GML doit être marqué avec l'étiquette ERREUR, qui n'est liée à aucune sous-réponses, car il n'y a pas de correspondance claire entre le texte erroné et une sous-réponse particulière (voir aussi des exemples dans l'annexe D). Nous agrégeons ces annotations au niveau de la requête, par un vote majoritaire (cf. annexe C), avec les étiquettes suivantes :

— `COMPLET`, si *toutes* les sous-réponses ont été liées à une étiquette CORRECT par plus de la moitié des annotateurs, ou `INCOMPLET`, dans le cas contraire.

— `ERRONÉ`, si plus de la moitié des annotateurs ont annoté *au moins une* ERREUR dans la réponse GML, ou `SANS ERREUR`, dans le case contraire[6],

Chaque paire requête-réponse reçoit donc l'une des quatre paires d'étiquettes, détaillées dans le tableau 2 : `COMPLET/SANS ERREUR`, `COMPLET/ERRONÉ`, `INCOMPLET/SANS ERREUR`, `INCOMPLET/ERRONÉ`. La situation idéale est `COMPLET/SANS ERREUR` : une réponse GML répond correctement et exhaustivement à la requête sans fournir de mauvaises réponses. Le pire résultat est `INCOMPLET/ERRONÉ` : non seulement le GML ne répond pas, ou pas complètement, à

3. https://huggingface.co/microsoft/Phi-4-mini-instruct
4. https://huggingface.co/ibm-granite/granite-3.2-8b-instruct
5. https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
6. D'autres façons d'agréger les réponses des annotateurs produisent des résultats similaires à ceux que nous présentons ici.

	COMPLET	INCOMPLET
SANS ERREUR	contient toutes les bonnes réponses, ne contient aucune réponse erronée	ne contient pas toutes les bonnes réponses, et aucune réponse erronée
ERRONÉ	contient toutes les bonnes réponses, et une ou des réponses erronées	ne contient pas toutes les bonnes réponses, et une ou des réponses erronées

TABLE 2 – Interprétations des combinaisons d'étiquettes au niveau de la requête.

la requête, mais il fournit aussi au moins une mauvaise réponse. Les deux autres situations intermédiaires peuvent varier en intensité. Une réponse COMPLET/ERRONÉ inclut tous les éléments d'une bonne réponse, mais fournit aussi des informations incorrectes. Une réponse INCOMPLET/PAS ERRONÉ ne fournit pas, ou pas complètement, la bonne réponse, mais elle ne fournit pas non plus d'informations erronées. C'est le cas par exemple lorsque la réponse GML est un refus de répondre. Nous laissons une étude plus approfondie de ces types de réponses non extrêmes à de futurs travaux.

Chaque réponse FR ou EN d'un GML est annotée par 2 à 5 annotateurs, qui maîtrisent tous la langue de la requête et de la réponse. Les réponses FR-EN-FR GML ont été annotées par 2 ou 3 annotateurs chacune, à l'exception des requêtes vaccinales, qui n'ont été annotées que par un seul annotateur en raison de contraintes de temps (cf. annexe C). Trois annotateurs ont contribué à l'ensemble final d'annotations en anglais, tandis que cinq annotateurs ont complété des annotations en français (l'un de ces annotateurs a participé aux annotations anglaises et françaises). Pour chaque langue, l'ensemble complet contient 312 paires requête-réponse. En anglais, les annotateurs ont annoté entre 168 et 312 paires requête-réponse, pour une moyenne de 227 paires requête-réponse par annotateur. Pour le français, les annotateurs ont annoté entre 117 et 258 paires requête-réponse, soit en moyenne 178 paires requête-réponse par annotateur. Finalement, pour l'approche pipeline FR-EN-FR, les annotateurs ont annoté entre 30 et 195 paires requête-réponse (en moyenne, 98 paires par annotateur).

3.1 Accord inter-annotateurs

Nous calculons l'accord interannotateur en utilisant le κ de Cohen (Cohen, 1960), via l'implémentation `scikit-learn` (Pedregosa *et al.*, 2011)[7]. Nous examinons d'abord l'accord interannotateur pour les sous-réponses à chaque requête (les composantes manuellement identifiées de la réponse complète), en calculant le κ de Cohen entre les paires d'annotateurs sur le sous-ensemble partagé des données qu'ils ont tous deux annoté (cf. tableau 3). Pour les annotations FR, l'accord sur l'étiquetage des sous-réponses CORRECT variait de 0,662 à 0,831 (moyenne de 0,765). Bien que l'interprétation exacte des valeurs du κ de Cohen soit quelque peu controversée, de telles valeurs sont généralement considérées comme reflétant des niveaux d'accord substantiels à excellents (Landis & Koch, 1977). Les requêtes vaccinales ont été perçues par certains annotateurs comme plus difficiles à annoter[8], et, pour le français, on observe une légère diminution de l'accord sur celles-ci, allant de 0,400 à 0,823 (moyenne de 0,607). À part la plus basse, ces valeurs sont considérées comme indiquant un niveau d'accord modéré. En ce qui concerne les étiquettes ERREUR, nous calculons l'accord au niveau de

7. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

8. En effet, ces requêtes ont plus de sous-réponses par question, plus de terminologie technique avec des distinctions minimes mais importantes (comme de légères variations sur les noms de vaccins), et sont répétitives.

	EN			FR			FR-EN-FR		
	Min.	Moy.	Max.	Min.	Moy.	Max.	Min.	Moy.	Max.
CORRECT	0,728	0,811	0,891	0,662	0,765	0,831	0,729	0,790	0,927
ERREUR	0,761	0,803	0,826	0,520	0,646	0,789	0,507	0,661	0,772

TABLE 3 – Accord interannotateur (κ de Cohen) entre les paires d'annotateurs.

la requête en vérifiant si une ou plusieurs étiquettes ERREUR sont présentes dans la réponse GML complète. Pour le français, on trouve des niveaux d'accord modérés à excellents quant à savoir si une étiquette ERREUR devrait apparaître dans la réponse GML : 0,520 à 0,789 (moyenne 0,646). Pour les annotations FR-EN-FR, l'accord sur l'étiquetage des sous-réponses CORRECT variait de 0,729 à 0,927 (moyenne de 0,790) et l'accord sur les étiquettes ERREUR variait de 0,507 à 0,772 (moyenne de 0,661), mais il y avait moins d'annotations. Pour l'anglais, nous observons des tendances similaires entre les trois annotateurs[9] : l'accord sur l'étiquetage des sous-réponses comme CORRECT variait de 0,728 à 0,891 (moyenne 0,811), avec une légère baisse pour les données vaccinales (minimum de 0,628, moyenne de 0,773 et maximum de 0,855).

La raison pour laquelle l'accord interannotateur est plus bas sur le français que sur l'anglais n'est pas claire. Cela pourrait être dû au fait que certains annotateurs, ayant annoté moins de réponses, sont moins expérimentés, ou bien qu'avec plus d'annotateurs, le sous-ensemble de données sur lequel nous pouvons calculer l'accord entre les annotateurs a tendance à être plus petit.

4 Résultats

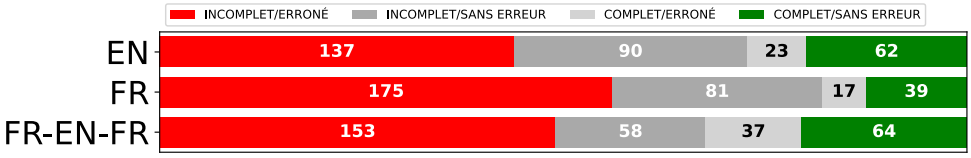


FIGURE 1 – Nombre de combinaisons d'étiquettes pour les réponses de GML en anglais, en français et avec l'approche du pipeline de traduction FR-EN-FR (toutes catégories, tous GMLs).

La Figure 1 présente les fréquences de chacune des quatre combinaisons d'étiquettes, pour l'ensemble des catégories et des GML. Sur la gauche de la figure, on voit qu'il y a plus de réponses INCOMPLET/ERRONÉ (le pire scénario) en français qu'en anglais (175 vs. 137), tandis qu'il y a plus de réponses COMPLET/SANS ERREUR (le meilleur scénario) en anglais qu'en français (62 vs. 39). Cela illustre l'impact des disparités interlinguistiques : malgré leurs prétentions multilingues, les GML que nous avons utilisés performent moins bien en français qu'en anglais sur le même ensemble de requêtes. Notons cependant que la performance globale est assez faible dans les deux langues. L'approche du pipeline de traduction FR-EN-FR permet de réduire l'écart de performance que nous

9. Quatre annotateurs ont initialement soumis des annotations pour l'anglais. Cependant, l'un d'entre eux a montré une concordance nettement plus faible avec les autres annotateurs — κ de 0,337 (CORRECT) et 0,237 (ERREUR). Comme il s'agit d'un cas assez extrême, nous retirons ses annotations de notre analyse. Nous notons cependant que les résultats avec et sans cet annotateur présentent des tendances comparables.

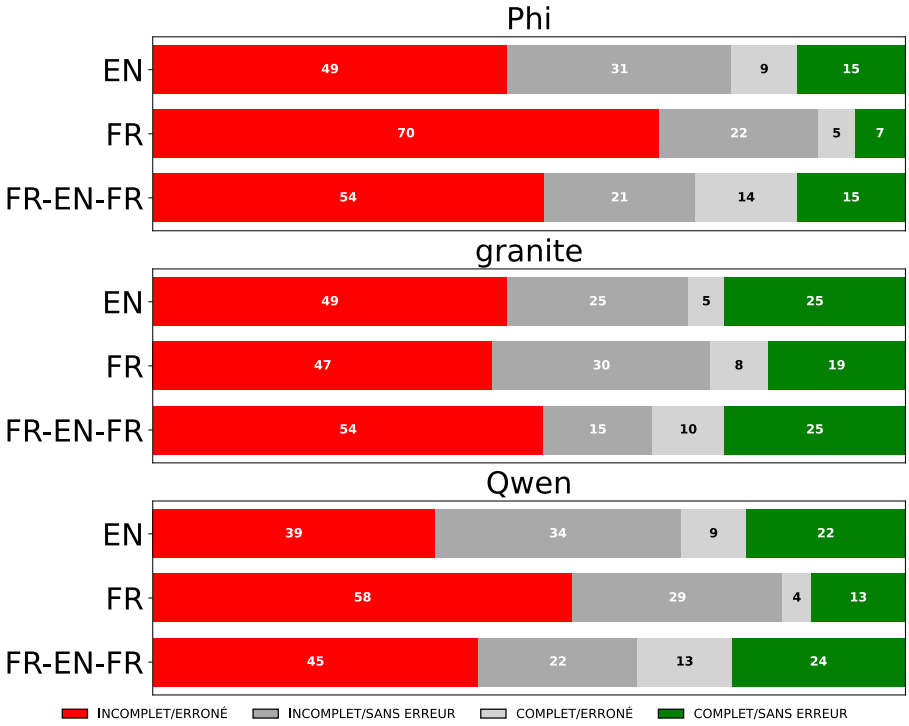


FIGURE 2 – Nombre de combinaisons d'étiquettes pour les réponses de GML en anglais, en français et avec l'approche du pipeline de traduction FR-EN-FR, par GML (toutes catégories).

observons, probablement en raison de la qualité des systèmes de traduction automatique pour cette paire de langues.

Nous décomposons maintenant cette analyse par GML, cf. figure 2. Nous remarquons qu'il existe des différences de performance marquées selon les modèles, que ce soit entre les langues ou au sein d'une même langue. Sur les requêtes anglaises, le modèle `granite` et le modèle `Qwen` reçoivent tous deux un nombre plus élevé d'étiquettes `COMPLET/SANS ERREUR` que le modèle `Phi` ; les modèles `Phi` et `granite` reçoivent un nombre plus élevé d'étiquettes `INCOMPLET/ERRONÉ` que `Qwen`. Sur les requêtes françaises, le modèle `Phi` a le plus grand nombre d'étiquettes `INCOMPLET/ERRONÉ`, plus de la moitié des requêtes reçoivent une étiquette `ERRONÉ`. Le modèle `granite` a la meilleure performance en français sur ce jeu de requêtes, et ne montre qu'une faible amélioration du nombre d'étiquettes `COMPLET/SANS ERREUR` (+6) avec l'approche du pipeline de traduction, mais on observe aussi une augmentation du nombre d'étiquettes `ERRONÉ` (+9). Lorsqu'il s'agit de déployer un modèle, on peut vouloir prioriser le plus grand nombre possible de réponses `COMPLET/SANS ERREUR`, ou bien minimiser le nombre de réponses `INCOMPLET/ERRONÉ`. Comme on peut le voir sur la figure 2, certaines priorités pourraient mener à choisir le même modèle en anglais et en français, tandis que d'autres peuvent conduire à déployer des modèles différents pour le français et pour anglais.

La figure 3 montre ces résultats, ventilés par catégorie de requête. Nous notons que les requêtes vaccinales ont un nombre particulièrement faible de réponses `COMPLET/SANS ERREUR`. Cela s'explique en partie par le fait qu'elles ont tendance, en moyenne, à contenir plus de sous-réponses, et sont donc plus susceptibles d'avoir des réponses incomplètes. Pour toutes les catégories, on observe plus de réponses `INCOMPLET/ERRONÉ` pour le français que pour l'anglais. Dans quatre cas, il y a moins de réponses `COMPLET/SANS ERREUR` pour le français que pour l'anglais, tandis que pour

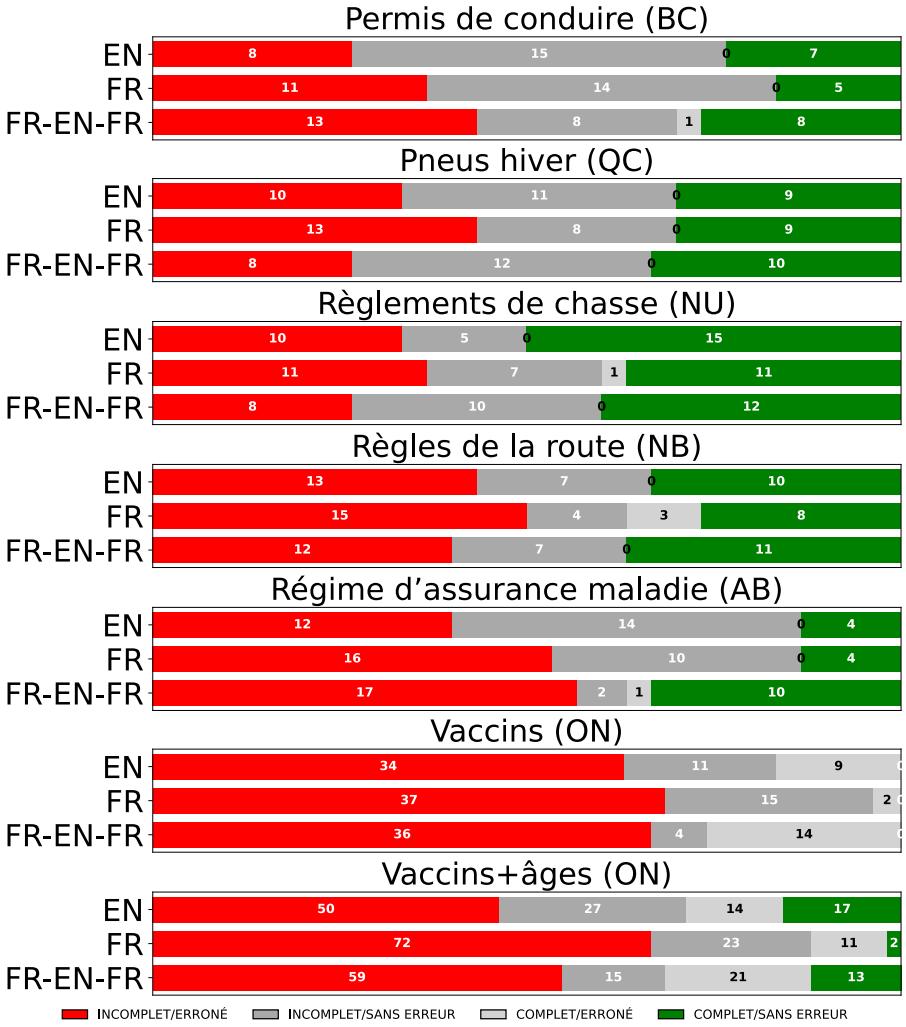


FIGURE 3 – Un graphique comparant le nombre de combinaisons d'étiquettes entre l'anglais et le français et l'approche du pipeline de traduction FR-EN-FR (par catégorie, tous GMLs).

les trois autres cas, les résultats sont identiques pour les deux langues. Globalement, cela confirme la présence de disparités interlinguistiques dans ces tâches de sécurité et de réglementation. Nous soulignons qu'en désagrégant les résultats en catégories de plus en plus petites, nous sommes limités par la petite taille de cet ensemble de données. Ainsi, les conclusions tirées de ces ensembles plus restreints de données doivent être interprétées avec prudence. Néanmoins, la forte cohérence des tendances appuie l'hypothèse qu'il existe effectivement de réelles disparités de performance.

En plus de notre analyse quantitative, nous avons identifié plusieurs problèmes notables dans les réponses. À au moins deux occasions, le modèle `granite` produit un texte qui pourrait faire office de réponse complète (en français, comme prévu), puis produit le texte « **Traduction en français :** » avant de répéter la réponse. Tous les modèles ont des difficultés avec les détails et le contenu technique : les réponses GML qui incluent des mesures incluent parfois à la fois des pouces et des millimètres avec une conversion incorrecte entre les deux ; les réponses françaises contiennent parfois des versions anglaises des noms de vaccins plutôt que des versions françaises [10] ; elles dirigent parfois

10. Cependant, certaines des abréviations françaises utilisées dans le calendrier d'immunisation de routine de l'Ontario diffèrent de celles du Catalogue national des vaccins (https://nvc-cnv.canada.ca/fr/catalogue-des-vaccins).

les lecteurs vers des sites web inexistants ou autrement incorrects pour plus d'informations ; et des acronymes comme AHCIP (*Alberta Health Care Insurance Plan*) sont parfois mal interprétés. Ces observations, et d'autres encore, ont été soulevées lors de discussions avec les annotateurs et dans leurs commentaires. Elles mettent en lumière des directions potentielles pour des travaux futurs. Un autre aspect qui pourrait nécessiter un examen supplémentaire est que le modèle `granite`, qui selon certains critères est le modèle le plus performant en anglais comme en français, est aussi celui qui génère certaines des erreurs les plus alarmantes. Cela soulève la question de savoir si un modèle généralement plus fiable est plus susceptible de tromper les utilisateurs lorsqu'il commet des erreurs. Nous ne distinguons pas complètement les différents niveaux de gravité des erreurs, ni ne demandons aux annotateurs de vérifier des informations potentiellement trompeuses qui ne correspondent pas aux réponses de référence. Néanmoins, plusieurs annotateurs ont utilisé l'étiquette `OTHER` et les fonctions de commentaire pour fournir cette information ou signaler qu'ils n'étaient pas sûrs s'il y avait des informations erronées. Bien que certaines de ces erreurs soient mineures, d'autres sont plus préoccupantes. Par exemple, une réponse affirme qu'un vaccin particulier n'est plus utilisé au Canada, une réponse incorrecte qui risquerait d'alimenter la désinformation et l'hésitation envers les vaccins. Cette erreur commise par un GML est particulièrement d'actualité : en 2025, l'Ontario faisait face à la plus grande éclosion de rougeole depuis 2008 [11], l'une des maladies ciblées par le vaccin MMRV mentionné dans la réponse GML.

5 Limitations

Le jeu de données étudié ici est de relativement petite taille, comportant 936 réponses annotées de GLM pour 104 questions reglementaires très ciblées sur des domaines spécifique, dont plus de la moitié portent sur la vaccination en Ontario. La présence d'un plus grand nombre de requêtes concernant les vaccins est simplement dûe au fait qu'elles étaient simples à générer de manière formulaire.

L'objet de notre étude porte sur la différences de performance des GLM, et non sur les gains de performance brute qui pourraient être obtenus par exemple en utilisant des modèles de plus grande taille, ou plus spécifiques au français (Godey *et al.*, 2025; Gouvert *et al.*, 2025), des techniques plus élaborées comme le RAG (*Retrieval Augmented Generation*), ou bien l'optimisation des amorces utilisées (*prompt engineering*). Bien que ces différentes pistes puissent sans aucun doute améliorer les réponses, il n'est pas évident que la *différence* de performance entre langues disparaisse. Par exemple, des GLM spécifiques au français pourraient améliorer les réponses aux questions françaises, mais dégrader les réponses aux questions anglaises, et nécessiter un *pipeline* différent. En fait, des expériences effectuées sur un jeu de données amélioré avec des GLM commerciaux de grande taille suggèrent que les disparités mises à jour persistent malgré des réponses globalement très améliorées.

6 Directions futures

Le processus d'annotation a mis en évidence certaines situations où les requêtes, qui aspirent à la clarté, contenaient malgré tout des ambiguïtés potentielles. Dans des travaux futurs, nous prévoyons

11. cbc.ca/news/canada/toronto/measles-cases-ontario-double-last-two-weeks-1.7470166

d'intégrer ces enseignements pour clarifier et simplifier certaines requêtes. On voudra notamment clarifier si une question porte sur un vaccin ou une vaccination spécifique contre une maladie particulière, ce qui affecte à la fois les requêtes et les réponses. Un exemple est le vaccin contre la varicelle, qui est parfois administré seul et parfois dans le cadre d'un vaccin combiné qui couvre la rougeole, les oreillons, la rubéole et la varicelle. Un autre exemple est une question sur les vaccins que les bébés devraient recevoir à 2 et 4 mois. Pour diminuer l'ambiguïté, cela pourrait être séparé en deux requêtes plus simples. Nous n'avons pas non plus inclus le vaccin annuel contre la grippe et les vaccins contre la COVID-19 dans la plupart des réponses de référence, car ils apparaissent dans une section différente des directives de vaccination pour l'Ontario (le vaccin contre la grippe est administré de façon saisonnière plutôt qu'à âge fixe). À l'avenir, nous examinerons la meilleure façon de les inclure dans le jeu d'évaluation. Enfin, dans certains cas, les questions et réponses étaient floues en ce qui concerne les vaccins nécessitant plusieurs doses. Bien que réduire ces ambiguïtés devrait améliorer les évaluations futures utilisant ce type de données, nous notons que la nature entièrement parallèle, anglais et français, des requêtes et des réponses de référence signifie que la présence de ces ambiguïtés n'invalide pas les résultats, car elles impactent les deux langues de manière égale.

Nous avons également inclus des informations supplémentaires entre parenthèses dans certaines des réponses de référence, dans le but de faciliter le travail des annotateurs. Ces informations supplémentaires sont par exemple des mesures en différentes unités, des acronymes ou abréviations, des noms courants ou formels pour la même maladie, ou d'autres détails mineurs. Il n'est cependant pas clair que ces informations ont été utile ou si elles ont engendré de la confusion. Nous chercherons à clarifier cela dans de futurs travaux.

Notre ensemble de données pourrait être naturellement élargi de plusieurs façons. L'une d'elles serait d'utiliser des requêtes plus structurées ou préformatées, jumelées à des références appropriées faisant autorité, pour élargir les requêtes (par exemple, les recommandations de vaccins) à travers les provinces et territoires du Canada. Il pourrait aussi être utile de produire un ensemble comparable de requêtes concernant les recommandations de vaccins aux États-Unis. Nous avons en effet constaté que certaines réponses des GML pourraient avoir été influencées par des recommandations vaccinales issues d'autres pays. Par exemple, le vaccin contre l'hépatite B est recommandé en Ontario pour les enfants de 7e année (12-13 ans), tandis que la CDC aux États-Unis, au moment de l'étude, le recommande à tous les enfants dès la naissance[12]. Certaines réponses de GML incluaient cette recommandation de vaccination plus précoce et il serait intéressant de vérifier s'il existe d'autres exemples d'influence provenant d'autres pays.

Nous avons amélioré notre ensemble de requêtes à partir de ces expériences et prévoyons de publier un jeu de données amélioré dans les prochains mois.

7 Conclusions

Dans cet article, nous avons décrit la construction d'un ensemble préliminaire de questions et de réponses de référence pour tester les disparités linguistiques des réponses fournies par trois GML sur des sujets liés à la sécurité et à la réglementation, dans des contextes canadiens. Ce jeu de test vise les situations où un utilisateur qui consulte un GML peut se mettre lui-même à risque ou mettre d'autres personnes à risque (par exemple, risques pour la santé à cause de vaccins manqués, risques

12. https://www.cdc.gov/hepatitis-b/hcp/vaccine-administration/index.html (Consulté le 2025-03-31.)

de dommages ou de sanctions légales pour ne pas utiliser des pneus d'hiver lorsque nécessaire, etc.). Il s'agit d'un ensemble de données d'évaluation bilingue, rédigé à l'origine en anglais et traduit manuellement en français.

Nous avons examiné la performance de trois GML multilingues sur cet ensemble de données à l'aide d'une évaluation humaine. Bien que nous ayons sélectionné des modèles plus petits qui pouvaient facilement et efficacement être exécutés en interne, sans risque d'exposer l'ensemble de test à des modèles externes, nous notons de façon anecdotique que des requêtes similaires soumises à de grands modèles accessibles au public semblent présenter des schémas d'erreur similaires. Ce travail est aussi limité par la taille relativement petite de l'ensemble de données, bien que nous ayons tenté de compenser en couvrant une gamme de sujets à des profondeurs variables. Nous utilisons une approche fondée sur l'évaluation humaine, et constatons un accord inter-annotateurs relativement élevé sur les réponses correctes et erronées.

Nous observons que dans les trois GML et dans toutes les catégories de requêtes, il existe des disparités systématiques entre l'anglais et le français. Dans l'ensemble, il y a peu de réponses entièrement correctes (moins de 25%), même en anglais. Néanmoins, il y a presque toujours moins de réponses françaises entièrement correctes. Les réponses françaises contiennent aussi plus fréquemment des informations erronées. Nous présentons des résultats préliminaires pour une approche d'atténuation, utilisant un pipeline de traduction, qui produit des résultats initiaux prometteurs pour combler les écarts entre l'anglais et le français.

En conclusion, nous notons que le français reste une langue relativement bien dotée en ressources. En particulier, dans le contexte canadien, il existe de nombreuses sources de textes parallèles entre l'anglais canadien et le français canadien, qui sont probablement bien représentées dans les données d'entraînement. Même dans ce contexte, et même en formulant des questions précises et contextualisées, nous constatons qu'il existe des disparités évidentes de performance. On peut s'attendre à des disparités encore plus importantes pour les langues à plus faibles ressources. Nous recommandons donc fortement aux organismes et aux individus qui souhaitent mettre à disposition des outils et des contenus reposant sur les GML — et particulièrement dans des contextes liés à la sécurité — d'effectuer une analyse approfondie des disparités de performance interlinguistiques avant de déployer ces outils. Même les GML qui prétendent à une couverture multilingue sont susceptibles de présenter des disparités de performance interlinguistiques marquées, qui peuvent entraîner des risques déséquilibrés de préjudice pour les utilisateurs qui font confiance aux résultats de ces outils.

Remerciements

Ce projet a été mené par le CNRC au nom de l'Institut canadien de la sécurité de l'intelligence artificielle (ICSIA). Nous remercions nos collègues pour des discussions fructueuses, leurs commentaires et leur contribution au processus d'annotation. Nous remercions les relecteurs anonymes, dont les commentaires ont permis d'améliorer et clarifier cet article.

Références

AHUJA K., DIDDEE H., HADA R., OCHIENG M., RAMESH K., JAIN P., NAMBI A., GANU T.,

SEGAL S., AHMED M., BALI K. & SITARAM S. (2023). MEGA : Multilingual evaluation of generative AI. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 4232–4267, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.258.

ARTETXE M., GOSWAMI V., BHOSALE S., FAN A. & ZETTLEMOYER L. (2023). Revisiting machine translation for cross-lingual classification. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 6489–6499, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.399.

BRIAKOU E., CHERRY C. & FOSTER G. (2023). Searching for needles in a haystack : On the role of incidental bilingualism in PaLM's translation capability. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 9432–9452, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.524.

CHATAIGNER C., TAÏK A. & FARNADI G. (2024). Multilingual hallucination gaps in large language models. DOI : 10.48550/arXiv.2410.18270.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46. DOI : 10.1177/001316446002000104.

GODEY N., ANTOUN W., TOUCHENT R., BAWDEN R., ÉRIC DE LA CLERGERIE, SAGOT B. & SEDDAH D. (2025). Gaperon : A peppered English-French generative language model suite. DOI : 10.48550/arXiv.2510.25771.

GOUVERT O., HUNTER J., LOURADOUR J., CERISARA C., DUFRAISSE E., SY Y., RIVIÈRE L., LORRÉ J.-P. & COMMUNITY O.-F. (2025). The Lucie-7B LLM and the Lucie training dataset : Open resources for multilingual language generation. DOI : 10.48550/arXiv.2503.12294.

HOLTERMANN C., RÖTTGER P., DILL T. & LAUSCHER A. (2024). Evaluating the elementary multilingual capabilities of large language models with MultiQ. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2024*, p. 4476–4494, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-acl.265.

HUANG H., TANG T., ZHANG D., ZHAO X., SONG T., XIA Y. & WEI F. (2023). Not all languages are created equal in LLMs : Improving multilingual capability by cross-lingual-thought prompting. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 12365–12394, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.826.

INTRATOR Y., HALFON M., GOLDENBERG R., TSARFATY R., EYAL M., RIVLIN E., MATIAS Y. & AIZENBERG N. (2024). Breaking the language barrier : Can direct inference outperform pre-translation in multilingual LLM applications ? In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 2 : Short Papers)*, p. 829–844, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-short.75.

KIM V. T., KRUMDICK M., REDDY V., DERNONCOURT F. & LAI V. D. (2024). An analysis of multilingual FActScore. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 4309–4333, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.247.

LAI V. D., NGO N., POURAN BEN VEYSEH A., MAN H., DERNONCOURT F., BUI T. & NGUYEN T. H. (2023). ChatGPT beyond English : Towards a comprehensive evaluation of large language models in multilingual learning. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 13171–13189, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.878.

LANDIS J. R. & KOCH G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174. DOI : 10.2307/2529310.

LIU W., WU N., DING W., LIANG S., GONG M. & ZHANG D. (2025). Selected languages are all you need for cross-lingual truthfulness transfer. In O. RAMBOW, L. WANNER, M. APIDIANAKI, H. AL-KHALIFA, B. D. EUGENIO & S. SCHOCKAERT, Éds., *Proceedings of the 31st International Conference on Computational Linguistics*, p. 8963–8978, Abu Dhabi, UAE : Association for Computational Linguistics.

LIU Y., CAO J., LIU C., DING K. & JIN L. (2024). Datasets for large language models : A comprehensive survey. DOI : 10.48550/arXiv.2402.18041.

MIN S., KRISHNA K., LYU X., LEWIS M., YIH W.-T., KOH P., IYYER M., ZETTLEMOYER L. & HAJISHIRZI H. (2023). FActScore : Fine-grained atomic evaluation of factual precision in long form text generation. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 12076–12100, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.741.

NEJADGHOLI I., MOLAMOHAMMADI M. & BAKHTAWAR S. (2024). Social and ethical risks posed by general-purpose LLMs for settling newcomers in Canada. DOI : 10.48550/arXiv.2407.20240.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

QIU Y., ZISER Y., KORHONEN A., PONTI E. & COHEN S. (2023). Detecting and mitigating hallucinations in multilingual summarisation. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 8914–8932, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.551.

QWEN TEAM (2024a). Qwen2.5 : A party of foundation models. https://qwenlm.github.io/blog/qwen2.5/.

QWEN TEAM (2024b). Qwen2.5-LLM : Extending the boundary of LLMs. https://qwenlm.github.io/blog/qwen2.5-llm/.

SHAFAYAT S., KIM E., OH J. & OH A. (2024). Multi-FAct : Assessing factuality of multilingual LLMs using FActscore. In *First Conference on Language Modeling*.

SHEN L., TAN W., CHEN S., CHEN Y., ZHANG J., XU H., ZHENG B., KOEHN P. & KHASHABI D. (2024). The language barrier : Dissecting safety challenges of LLMs in multilingual contexts. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2024*, p. 2668–2680, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-acl.156.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). brat : a web-based tool for NLP-assisted text annotation. In F. SEGOND, Éd., *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107, Avignon, France : Association for Computational Linguistics.

WANG W., TU Z., CHEN C., YUAN Y., HUANG J.-T., JIAO W. & LYU M. (2024). All languages matter : On the multilingual safety of LLMs. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds.,

Findings of the Association for Computational Linguistics : ACL 2024, p. 5865–5877, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-acl.349.

ZHANG W., ALJUNIED M., GAO C., CHIA Y. K. & BING L. (2023a). M3exam : A multilingual, multimodal, multilevel benchmark for examining large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

ZHANG X., LI S., HAUER B., SHI N. & KONDRAK G. (2023b). Don't trust ChatGPT when your question is not in English : A study of multilingual abilities and types of LLMs. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 7915–7927, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.491.

Annexes

A Ensemble de données : détails

— **Vaccins (Ontario) :** 54 requêtes liées à la vaccination en Ontario (principalement la vaccination des enfants). Celles-ci proviennent du calendrier de vaccination de routine de l'Ontario [13]. Parmi celles-ci, 9 sont des questions relativement formelles du type "According to Ontario's routine vaccination schedule, what vaccines should babies receive [at certain ages] ?"/«D'après le calendrier de vaccination systématique de l'Ontario, quels vaccins les bébés doivent-ils recevoir [à tel age] ?», 9 sont des versions moins formelles : "What vaccines should babies in Ontario get [at certain ages] ?"/«Quels vaccins les bébés doivent-ils recevoir [à tel age] en Ontario ?» Les réponses à ces questions sont des listes de vaccins. Les questions restantes portent sur l'âge auquel les individus devraient recevoir certains vaccins, encore une fois divisés en deux formes ("At what age"/«À quel âge» and "When"/«Quand»). Les réponses à ces questions sont des listes d'âges ou de tranches d'âge.

— **Pneus d'hiver (Québec) :** 10 requêtes concernant les règlements sur les pneus d'hiver du Québec [14], couvrant des sujets tels que la date d'application des règlements, à qui et quels véhicules elles s'appliquent, ainsi que les détails des exigences.

— **Permis de conduire (Colombie-Britannique) :** 10 requêtes concernant le programme de permis gradué pour conducteurs en Colombie-Britannique [15]. Elles couvrent des sujets tels que les types de tests, l'âge d'admissibilité et les règles pour les conducteurs titulaires d'un permis d'apprenti.

— **Régime d'assurance maladie (Alberta) :** 10 requêtes liées au Régime d'assurance santé de l'Alberta (AHCIP) et à l'admissibilité [16]. Ce sont en fait 5 requêtes distinctes, où l'une utilise

13. anglais : https://www.ontario.ca/page/ontarios-routine-immunization-schedule, français : https://www.ontario.ca/fr/page/calendrier-de-vaccination-systematique-de-lontario

14. français : https://www.quebec.ca/transports/circulation-securite-routiere/periode-hivernale/pneus-hiver, anglais : https://www.quebec.ca/en/transports/traffic-road-safety/winter-road-safety/requirements-for-winter-tires

15. Les deux sources sont uniquement en anglais : https://www.icbc.com/driver-licensing/new-drivers/Graduated-licensing, https://www.icbc.com/driver-licensing/new-drivers/Get-your-L

16. anglais : https://www.alberta.ca/ahcip-eligibility, https://www.alberta.ca/ahcip-moving-to-alberta, français : https://www.alberta.ca/fr/ahcip-eligibility, https://www.alberta.ca/fr/ahcip-moving-to-alberta

le nom complet “Alberta Health Care Insurance Plan” et l’autre l’acronyme “AHCIP”, ce qui nous permet d’examiner comment cette différence dans la formulation des questions pourrait affecter la précision (travaux futurs).

— **Règlements de la route (Nouveau-Brunswick) :** 10 requêtes liées à des sujets tirés du Guide du conducteur du Nouveau-Brunswick [17].

— **Règlements de chasse (Nunavut) :** 10 requêtes liées à la chasse et à la pêche sportive au Nunavut pour les gens qui n’ont pas de droits de chasse autochtones [18]. Plusieurs de ces questions sont des questions dont la réponse est oui/non.

Chaque ensemble de requêtes repose sur une ou deux sources officielles. Ce sont toutes des sources disponibles sur Internet qui, à ce titre, peuvent être trouvées dans les données d’entraînement des GML [19]. Dans la plupart des cas, l’information est disponible en ligne en anglais et en français . Ce n’était pas le cas pour les informations sur le programme de licence graduée de la Colombie-Britannique, qui sont disponibles en anglais, chinois simplifié et pendjabi.

B Modèles : détails

— `microsoft/Phi-4-mini-instruct` : Comme décrit dans la carte du modèle (https://huggingface.co/microsoft/Phi-4-mini-instruct), il s’agit d’un « dense decoder-only Transformer model » de 3,8 milliards de paramètres, entraîné entre novembre et décembre 2024 (date limite du jeu de données en juin 2024) sur 5 milliards de tokens et publié en février 2025. L’anglais et le français figurent tous deux dans la liste des langues prises en charge.

— `ibm-granite/granite-3.2-8b-instruct` : Comme décrit dans la carte du modèle (https://huggingface.co/ibm-granite/granite-3.2-8b-instruct), il s’agit d’un modèle à long contexte de 8 milliards de paramètres. Il a été publié le 26 février 2025. La carte du modèle ne précise pas explicitement la date limite des données d’entraînement. L’anglais et le français figurent tous deux dans la liste des langues prises en charge.

— `Qwen/Qwen2.5-7B-Instruct` : Comme décrit dans la carte du modèle (https://huggingface.co/Qwen/Qwen2.5-7B-Instruct) et les billets de blogue Qwen 2.5 (Qwen Team, 2024a,b), il s’agit d’un modèle de langue causal de 7,61 milliards de paramètres entraîné sur 18 milliards de tokens avec un apprentissage par instructions amélioré. L’anglais et le français figurent tous deux dans la liste des langues prises en charge.

17. anglais : https://www.gnb.ca/content/dam/GNB3/t/dt-ct/handbook-manuel/docs/part-4-traffic-on-the-highway-rules-of-the-road.pdf, https://www.gnb.ca/content/dam/GNB3/t/dt-ct/handbook-manuel/docs/part-5-safe-driving.pdf, français : https://www.gnb.ca/content/dam/GNB3/t/dt-ct/handbook-manuel/docs/partie-4-circulation-regles-route.pdf, https://www.gnb.ca/content/dam/GNB3/t/dt-ct/handbook-manuel/docs/partie-5-conduite-securitaire.pdf

18. anglais : https://www.gov.nu.ca/sites/default/files/documents/2024-12/Nunavut_Hunting_Regulations_Guide_2024-25.pdf, https://www.gov.nu.ca/sites/default/files/documents/2024-05/Nunavut%20Sport%20Fishing%20Guide%20Final%202024-2025.pdf, français : (Non lié), https://www.gov.nu.ca/sites/default/files/documents/2024-05/Nunavut%20Sport%20Fishing%20Guide%20Final%202024-2025%20-%20FRE.pdf

19. Certaines peuvent être plus récentes que la date d’entraînement la plus récente, mais elles restent souvent inchangées (ou présentent seulement des modifications mineures) par rapport aux versions précédentes.

Catégorie	EN	FR	FR-EN-FR
Vaccins (Ontario)-formelles	3	2	1
Vaccins (Ontario)-moins formelles	2	2	1
Vaccins (Ontario)-À quel âge	2	2	1
Vaccins (Ontario)-Quand	2	3	1
Pneus d'hiver (Québec)	2	4	2
Permis de conduire (Colombie-Britannique)	2	3	2
Régime d'assurance maladie (Alberta)	2	2	3
Règlements de la route (Nouveau-Brunswick)	2	3	2
Règlements de chasse (Nunavut)	3	5	2

TABLE 4 – Nombre d'annotateurs qui annotaient chaque catégorie de requêtes.

	Annotateur 1	Annotateur 2	Annotateur 3	Final
Sous-réponse (a)	CORRECT	CORRECT	—	CORRECT
Sous-réponse (b)	CORRECT	CORRECT	CORRECT	CORRECT
Sous-réponse (c)	—	CORRECT	—	—
Sous-réponse (d)	CORRECT	CORRECT	CORRECT	CORRECT

TABLE 5 – Exemple de vote majoritaire pour les étiquettes CORRECT. Cette réponse recevrait alors une étiquette globale `INCOMPLET`.

Toutes les requêtes ont été effectuées en utilisant le pipeline `"text-generation"` fourni par HuggingFace[20]. Nous avons utilisé le tokeniseur par défaut pour chaque modèle, fixé le nombre maximal de nouveaux jetons générés à 500, et utilisé `"do_sample" : False` (en utilisant une stratégie de décodage gloutonne). Le pipeline HuggingFace gère la conversion de l'entrée vers le modèle approprié ; nous exécutons sans amorce système, en initialisant le `"role"` à `"user"` et utilisant le texte de la requête comme `"content"`. Les requêtes ont été exécutées sur des GPU V100 avec 32 Go de RAM.

C Annotations

Le tableau 4 montre le nombre d'annotateurs par catégorie de requêtes. Chaque réponse FR ou EN d'un GML est annotée par 2 à 5 annotateurs et les réponses FR-EN-FR GML ont été annotées par 2 ou 3 annotateurs chacune, à l'exception des requêtes vaccinales, qui n'ont été annotées que par un seul annotateur en raison de contraintes de temps.

Le tableau 5 illustre la méthode de vote majoritaire utilisée pour les étiquettes CORRECT. Dans le cas où plus de la moitié des annotateurs donnent une étiquette CORRECT à une sous-réponse, celle-ci reçoit également une étiquette CORRECT (cf. sous-réponses a, b, et d). Dans le cas où moins de la moitié des annotateurs donnent une étiquette CORRECT, la sous-réponse ne reçoit pas une étiquette CORRECT (cf. sous-réponse c). Dans cet exemple, la réponse du GML pour cette requête recevrait l'étiquette globale `INCOMPLET` car seulement trois des quatre sous-réponses ont

20. https://huggingface.co/docs/transformers/en/main_classes/pipelines

reçu l'étiquette CORRECT. Si les quatre sous-réponses avaient reçu l'étiquette finale CORRECT, alors l'étiquette globale COMPLET serait appliquée.

D Instructions pour les annotateurs

Nous présentons ci-dessous le guide d'annotation en entier. Il est écrit en anglais, et toutes les étiquettes sont nommées en anglais (pour ERREUR, l'interface utilise "WRONG"). Au fil du processus, nous avons maintenu un document FAQ avec des clarifications supplémentaires.

Nous avons depuis intégré ces clarifications et modifications dans un guide d'annotation révisé. Notons que dans ce rapport, lorsque nous discutons des étiquettes CORRECT au niveau des sous-réponses, nous parlons techniquement de la relation entre une sous-réponse (GOLD) et l'étiquette CORRECT ajoutée par un annotateur à un segment de texte de la réponse du GML, à l'étape 2 des instructions.

Dans des travaux futurs, nous prévoyons d'approfondir la directive avec un petit ensemble d'annotateurs afin de mieux clarifier le guide d'annotation. Notre objectif serait d'aider les annotateurs à mieux comprendre tous les composantes du guide d'annotation, en particulier celles telles que PARTIAL où l'accord entre les annotateurs était plus faible.

Accessing the *brat* annotation tool:

The annotation instructions on the following page will cover how to perform the specific annotations for this task, but you can consult the brat manual for more detailed user interface information: https://brat.nlplab.org/manual.html

In order to edit, you will need to log in with the username and password you have been provided.

You will be provided with a link to a brat directory with your name, which will include the data for you to annotate (please only annotate data within your own named directory).

You will be provided with a suggested order in which to annotate.

Please keep track of which directories of data you have finished annotating in the annotation spreadsheet: [URL]

Annotation instructions:

You will be presented with the original query, the set of gold standard correct responses, and the LLM output. Each gold standard correct response is pre-labeled "GOLD"; please do not modify these labels.

Step 1: Highlight spans in the *LLM output* with the following labels by clicking and dragging to highlight a span or by double-clicking (for a single word):

Main labels:

- CORRECT: Use this to mark a correct answer (i.e., that is semantically equivalent to one of the gold standard correct responses); try to mark as short a span as possible that still provides the full correct answer.
- PARTIAL: Use this to mark a partially correct answer (e.g., correct on only half of a span of time).
- WRONG: Use this to mark a wrong answer (i.e., one that does not match any of the gold standard correct responses or one that contradicts the gold standard responses).
- REPEAT: Use this to mark a repeated correct answer if the same correct answer is listed more than once in the LLM response (you should mark the first instance with CORRECT and any subsequent repeats with REPEAT).

Additional labels:

- WAFFLE: Use this to indicate when the response waffles, hedges, prevaricates, evades, or otherwise avoids providing a direct answer (e.g., if the LLM output describes what a common response to the question *might* be or what is *often* the case, rather than giving a clear response to the specific question).
- REFUSAL: Use this to indicate if the LLM response is a refusal to answer the question.
- IRRELEVANT: Use this to highlight text that is unrelated to the question posed (e.g., answers a different question, is completely unrelated).
- OTHER: Use this to highlight anything else that you want to draw attention to.

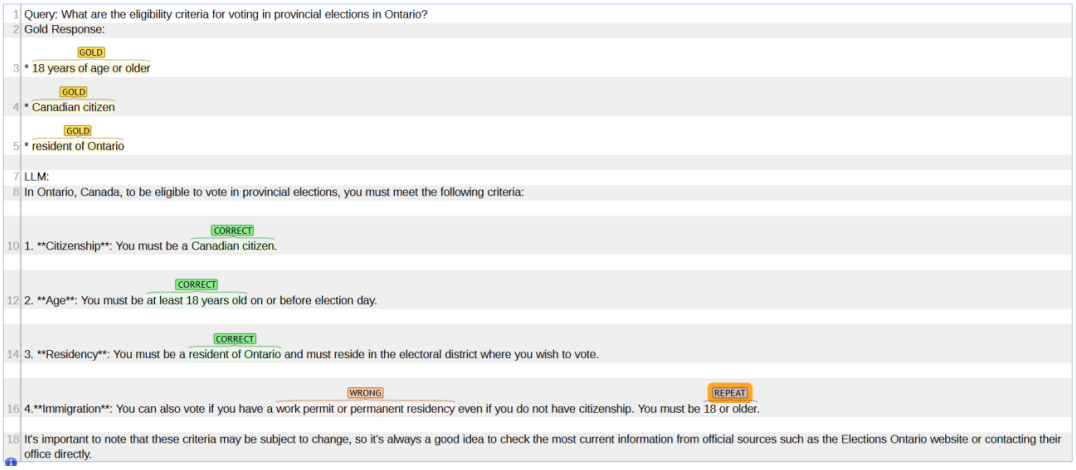
You can also add a written note if desired for clarification (e.g., if you highlight something as OTHER or if you want to give more information about a PARTIAL match or if you want to indicate that you're unsure of a label).

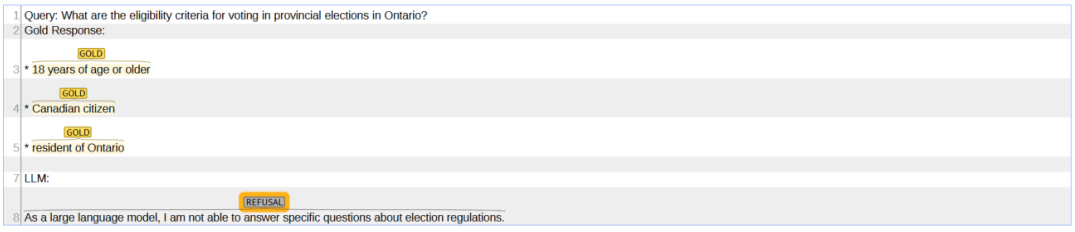
If you make a mistake or want to change a label you have assigned, you can click on that label and modify or delete it.

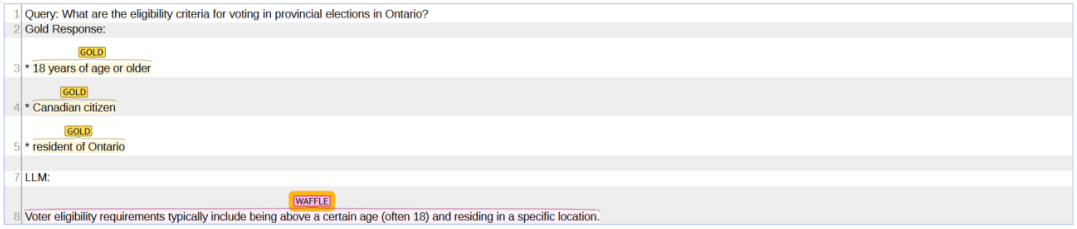
Step 2: When you have finished highlighting the spans, add a relation between any CORRECT or PARTIAL span and its corresponding GOLD span from the gold standard correct responses by clicking with the mouse on the GOLD span and dragging a connection to the corresponding CORRECT or PARTIAL. (It is recommended that you add the relations after completing the rest of your highlights, as the cross-line spans are visually disruptive.)

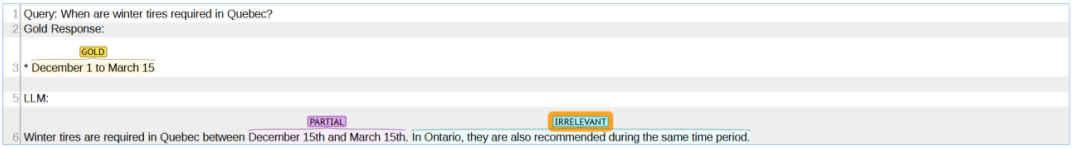
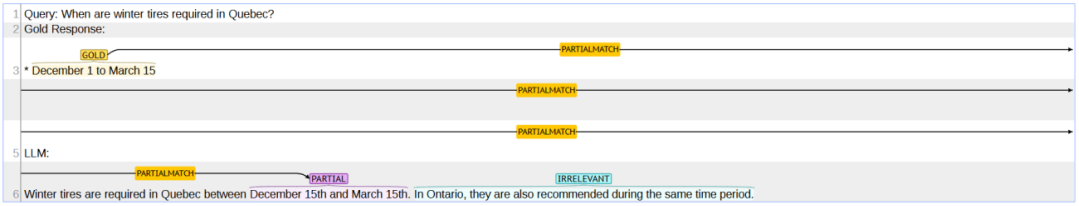
Next annotation: When you have completed both steps for an annotation, you can proceed to the next document in the directory by using the forward arrow in the top left corner of the user interface.

Here are example annotations after **Step 1**:









Example after Step 2: