EIFFEL: Un corpus d'expressions idiomatiques françaises pour évaluer les biais anglocentriques des LLMs

RÉSUMÉ

Les LLMs multilingues populaires sont généralement entraînés sur de plus grande proportions de données anglaises que de données multilingues, ce qui soulève des questions quant à leur capacité à saisir les particularités linguistiques propres à ces autres langues ainsi qu'à saisir leurs informations culturelles spécifiques. Nous contribuons ainsi à un effort visant à accroître la sensibilité multilingue des LLMs en développant un benchmark, EIFFEL, qui teste la maîtrise des expressions idiomatiques françaises en contexte. Nous détaillons la méthodologie employée, incluant la participation de locuteurs natifs français, afin de la rendre reproductible dans d'autres langues. Nous comparons les LLMs multilingues populaires aux LLMs axés sur le français à la fois sur des benchmarks standards et sur EIFFEL. EIFFEL met en évidence les avantages d'une proportion plus élevée de données en français et montre les limites des benchmarks standards pour mesurer le multilinguisme.

ABSTRACT

Mainstream multilingual LLMs are generally trained on a much higher proportion of English than multilingual data, raising questions about their ability to capture linguistic features particular to non-English languages or to capture information important to non-anglophone cultures. We add to a growing effort to increase multilingual sensitivity in LLMs by developing a benchmark, EIFFEL, testing mastery of French idiomatic expressions in context. We fully explain the methodology, which exploits input from native French speakers, to make it reproducible for other languages. We compare mainstream multilingual LLMs with French-focused LLMs both on standard LLM benchmarks and EIFFEL ; EIFFEL brings out the benefits of higher proportions of French data and shows limitations of standard benchmarks for measuring multilingual competence.

MOTS-CLÉS : multilingualité, évaluation, LLMs, benchmarks, expressions idiomatiques.

KEYWORDS: multilinguality, evaluation, LLMs, benchmarks, idiomatic expressions.

1 Introduction

Alors que les grands modèles de langues (LLMs) deviennent de plus en plus populaires dans le monde entier, beaucoup de modèles les plus utilisés sont entraînés sur des quantités disproportionnées de données en anglais. Pour illustrer ce propos, seulement 8% des données d'entraînement de Llama 3.1 sont des langues naturelles autres que l'anglais (Grattafiori *et al.*, 2024). Cela soulève le questionnement de l'influence de l'anglocentrisme sur la capacité d'un LLM à produire des séquences de haute qualité dans d'autres langues ainsi que sur la représentation des connaissances et des normes culturelles essentielles aux cultures non anglophones.

Répondre à cette question s'avère complexe de par la capacités de transfert entre les langues. Supposons que nous avons un modèle bilingue B entraîné sur de l'anglais et sur une autre langue L. S'il y a un transfert d'une langue à l'autre, alors la distribution de probabilité de B sur les tokens anglais peut informer sa distribution sur les tokens dans L et vice versa. Si nous appliquons ensuite B à une tâche en aval couverte par le transfert de connaissances depuis l'anglais, B pourrait donner de bons résultats après avoir vu seulement une petite partie des données dans L. Le transfert linguistique,

associé à un entraînement pertinent pour la tâche, est sans doute ce qui rend les LLMs anglocentriques traditionnels étonnamment performants dans les langues autres que l'anglais. Une question plus spécifique se pose alors : dans quelle mesure les aspects culturels et linguistiques auxquels L est sensible seront-ils négligés si nous nous appuyons sur le transfert linguistique ?

Une réponse complète à cette question ne pourra pas être prodiguée dans ce papier. Cependant nous proposons un outil important et unique en son genre pour explorer cette question : EIFFEL, **E**valuation of **I**diomatic **F**rench **F**ixed **E**xpressions for **L**arge Language Models, qui est un benchmark portant sur les expressions idiomatiques françaises.

Les expressions idiomatiques constituent un bon sujet d'étude, car elles sont une caractéristique de la langue de tous les jours qui est très spécifique à chaque langue. Si certaines expressions idiomatiques peuvent être traduites mot à mot entre le français et l'anglais, comme « Not my cup of tea/Pas ma tasse de thé », d'autres sont moins directes, voire complètement différentes. L'expression «Appeler un chat un chat», par exemple, a un équivalent évident en anglais : « call a spade a spade » (littéralement « appeler une bêche une bêche »), mais il ne s'agit pas d'une traduction directe. La majorité des expressions sont encore moins faciles à traduire. L'expression « Avoir du chien», se traduit littéralement par «To have some dog », cependant il n'existe pas d'expression similaire ayant un sens équivalent en anglais. Pour qu'un modèle puisse traiter ces dernières expressions, nous émettons l'hypothèse qu'il lui faut plus qu'une solide maîtrise de l'anglais et qu'une bonne capacité de traduction, il doit avoir vu soit des traductions explicites d'expressions idiomatiques, soit une quantité suffisante de données françaises (non traduites).

Nous testons l'impact de notre benchmark en comparant d'abord une série de modèles sur les versions françaises et anglaises des benchmarks standards (ARC Challenge (Clark *et al.*, 2018), Hellaswag (Zellers *et al.*, 2019), MMLU (Hendrycks *et al.*, 2020)), puis en comparant les résultats à ceux obtenus sur EIFFEL et à ceux du sous-ensemble français d'INCLUDE (Romanou *et al.*, 2024), un ensemble de données également rédigé à l'origine en français qui contient des sous-ensembles sensibles sur le plan culturel mais aussi agnostiques permettant des comparaisons intéressantes avec EIFFEL. Nous considérons les modèles pré-entraînés de deux familles de modèles anglocentriques, [1] Llama et Gemma (Team *et al.*, 2024, 2025), ainsi que des modèles « gallocentriques » [2] entraînés sur 33-50% de données françaises et anglaises : les modèles Gaperon (Godey *et al.*, 2025), Lucie (Gouvert *et al.*, 2025) et CroissantLLM (Faysse *et al.*, 2024). Nous nous intéressons également aux modèles EuroLLM (Martins *et al.*, 2025a,b), car ils offrent un compromis entre les modèles anglocentriques et gallocentriques et se concentrent sur les capacités de traduction. Notre étude se limite aux modèles pré-entraînés, car nous nous intéressons à la maîtrise linguistique de base.

Alors que les benchmarks standards en français ne semblent pas avantager les modèles gallocentriques par rapport aux modèles anglocentriques, EIFFEL et le sous-ensemble culturellement sensible d'INCLUDE le font. Cela suggère qu'EIFFEL capture des caractéristiques du français qui bénéficient moins du transfert ou de la traduction littérale.

Dans la mesure où les expressions idiomatiques ne sont qu'un exemple parmi d'autres du langage courant susceptible d'être trouvé dans les données web françaises, mais aussi un exemple d'un phénomène important pour diverses tâches en aval (allant du résumé des transcriptions de conversations à la communication avec les utilisateurs dans un style les mettant à l'aise), nos résultats de référence montrent que la modulation de la proportion de données françaises peut être importante pour le succès en aval des LLMs français de manière plus générale.

1. Nous considérons comme anglocentriques les modèles présentant un ratio anglais/français élevé.
2. Les modèles gallocentriques ont une proportion signifiante de données française, généralement au moins 25 %.

2 Etat de l'art

L'impact de l'anglocentricité. Les modèles anglocentriques multilingues peuvent produire des contenus multilingues de relativement bonne qualité. Cela ne veut cependant pas dire que les concepts et les patterns linguistiques produits par des locuteurs natifs de la langue cible soient employés. Guo *et al.* (2025) met en avant que la syntaxe et la distribution du vocabulaire dans les langues non anglaises sont affectées par la forte proportion de données d'entraînement en anglais, ce qui se traduit par des résultats souvent moins naturels et moins variés que ceux des locuteurs natifs. Plus la différence typologique entre l'anglais et la langue cible est grande, plus l'écart en termes de naturel lexical est prononcé. Karim *et al.* (2025) illustrent le fait que l'anglocentrisme peut impacter les performances des modèles et cela même dans des domaines qui ne semblent pas sensibles à la culture, comme les mathématiques. Ils mettent en avant une baisse de performances sur les benchmarks mathématiques lorsque certains mots des problèmes mathématiques sont remplacés par des mots plus pertinents pour une culture non anglophone, comme le remplacement des noms d'aliments occidentaux par ceux du Pakistan ou de la Moldavie.

Vers des modèles multilingues. Récemment, les modèles dominants, notamment Gemma 3 (Team *et al.*, 2025), Qwen 3 (Yang *et al.*, 2025) et Mistral 3.1 (`https://mistral.ai/fr/news/mistral-small-3-1`), affirment avoir augmenté leurs proportions de données multilingues, bien que nous n'ayons pas réussi à trouver de statistiques sur les proportions de données (et nous avons essayé). Certains modèles, tels que Llama 3 (Grattafiori *et al.*, 2024), Nemotron-H (Blakeman *et al.*, 2025; Adler *et al.*, 2024) et SmolLM3 (Bakouch *et al.*, 2025), fournissent des statistiques sur les proportions multilingues globales, mais nous n'avons pas pu trouver le détail par langue. [3]
Les projets ayant une orientation multilingue plus explicite fournissent souvent davantage d'informations. Les modèles EuroLLM (Martins *et al.*, 2025a,b) couvrent 24 langues et citent environ 45 à 60 % (selon la phase d'entraînement) de données en langue naturelle non anglaise, dont 5 à 6 % en français ; les modèles Apertus (Hernández-Cano *et al.*, 2025) couvrent 1 800 langues et utilisent environ 40 % de données non anglaises, dont 7,28 % en français ; et les modèles Salamandra (Gonzalez-Agirre *et al.*, 2025) couvrent 35 langues et contiennent 55 % de données non anglaises, dont 6,6 % en français (et 16 % en espagnol).
Certains projets se concentrent sur des langues autres que l'anglais ; nous nous concentrons ici sur les projets gallocentriques. CroissantLLM est un modèle bilingue de 1,3 milliard de paramètres entraîné à partir de zéro sur un ratio français-anglais de 1/1 (Faysse *et al.*, 2024), tandis que Lucie 7B (Gouvert *et al.*, 2025) et les modèles Gaperon (Godey *et al.*, 2025) sont entraînés sur environ 30% de données françaises.

Les benchmarks standards. De nombreux benchmarks utilisés pour tester les performances multilingues sont traduits à partir de données en anglais. Certaines sont traduites automatiquement, par exemple : XCODAH et XCSQA (Lin *et al.*, 2021), basées sur CODAH (Chen *et al.*, 2019) et CSQA (Talmor *et al.*, 2019), ainsi que ARC (Clark *et al.*, 2018), Hellaswag (Zellers *et al.*, 2019), TruthfulQA (Lin *et al.*, 2022), GSM8K (Cobbe *et al.*, 2021) et MMLU (Hendrycks *et al.*, 2020) traduits par (Thellmann *et al.*, 2024). Faysse *et al.* (2024) ont traduit d'autres benchmarks en français comme ARC et Hellaswag. Quelques benchmarks sont multilingues grâce à une traduction semi-automatique ou manuelle, par exemple Belebele (Bandarkar *et al.*, 2024), Mintaka (Sen *et al.*, 2022) et Global MMLU (Singh *et al.*, 2025). Seuls quelques benchmarks sont initialement construits dans

3. Llama 3.1 : 8 % de données multilingues, 8 langues prises en charge ; Nemotron-H : 3,7-5 %, 9 langues ; Nemotron 4 15 %, 53 langues ; SmolLM3 12 %, 6 langues.

la langue cible, comme FQuAD2.0 (d'Hoffschmidt *et al.*, 2020; Heinrich *et al.*, 2021), un ensemble de données de compréhension écrite en français dans le style de SQuAD (Rajpurkar *et al.*, 2016).

Les benchmarks portés sur la culture. Global MMLU (Singh *et al.*, 2025) est une version multilingue de MMLU qui étend le benchmark original en le traduisant et en précisant pour chaque question si elle est culturellement sensible ou neutre. BLEnD (Myung *et al.*, 2024) est un benchmark multilingue construit à l'aide de locuteurs natifs a qui il a été demandé de remplir des textes à trou à partir de modèles de phrases traduites où les mots à remplir étaient par exemple, des jours fériés ou des plats courants propre à la culture du locuteur. CulturalBench (Chiu *et al.*, 2024) comprend des questions ciblant 45 cultures différentes, bien que ces questions soient en anglais.
Pour les benchmarks culturels développés initialement dans la langue cible, AraDiCE (Mousi *et al.*, 2025) comprend sept dialectes arabes annotés avec le contexte culturel associé. CLIcK (Kim *et al.*, 2024) teste des connaissances textuelles, grammaticales et fonctionnelles en coréen. IOLBENCH (Goyal & Dan, 2025) pose des questions en anglais sur les caractéristiques linguistiques d'une variété de langues. INCLUDE (Romanou *et al.*, 2024) est un benchmark multilingue construit sur la base d'extraction de données de questions-réponses à partir de documents dans les langues cibles, qui sont ensuite vérifiées et corrigées par des locuteurs natifs. Il comprend des sous-ensembles culturellement sensibles et neutres. French Bench grammar-vocab-reading (Faysse *et al.*, 2024) évalue les règles grammaticales, le vocabulaire et la compréhension écrite de base. Parmi ceux-ci, seuls INCLUDE et French Bench couvrent le français, et seul INCLUDE propose des sujets sensibles à la culture.
En ce qui concerne les expressions idiomatiques, ID10M (Tedeschi *et al.*, 2022) teste la capacité des modèles à identifier une expression idiomatique ou autre MWE (expressions polylexicales) dans un texte. D'autres tâches se concentrent sur la capacité à fournir ou à identifier des définitions ou des paraphrases d'expressions idiomatiques et de MWE, c'est le cas de MAPS (Haviv *et al.*, 2023), IDIOMKB (Li *et al.*, 2024) et MIDAS (Kim *et al.*, 2025). Multilingual Idioms and Similes in LLMs (Khoshtab *et al.*, 2025) teste la capacité à poursuivre correctement un texte après l'utilisation d'une expression idiomatique. À notre connaissance, seul ID10M inclut le français, bien que les tâches collaboratives comme par exemple PARSEME (Ramisch *et al.*, 2020) testent la capacité à effectuer la classification MWE et la paraphrase. Le benchmark le plus ressemblant au nôtre est le benchmark arabe Kinayat (Attia *et al.*, 2025) qui évalue la capacité des modèles à compléter des expressions idiomatiques en masquant le dernier mot de l'expression.

3 Construire un benchmark pour les expressions idiomatiques françaises

Le sens idiomatique d'une expression idiomatique ne peut être déduit du sens littéral de ses composants : si vous êtes à une fête où personne ne parle et que vous dites à votre partenaire de « briser la glace », vous ne lui demandez pas littéralement de casser un bloc de glace, mais plutôt d'amener les gens à parler. Comprendre et utiliser correctement les expressions idiomatiques nécessite une maîtrise subtile de la langue cible et du contexte d'utilisation, ce qui en fait un sujet idéal pour évaluer les spécificités culturelles d'une langue.
EIFFEL s'appuie sur l'expertise de locuteurs natifs. Nous détaillons ci-dessous les étapes de sa construction et nous les illustrons dans la figure 1.

1. Collecter des expressions idiomatiques. Les expressions idiomatiques étant une partie importante du langage courant, il a été relativement facile de constituer une liste assez complète en effectuant des recherches sur Internet et en discutant avec des collègues francophones et anglophones

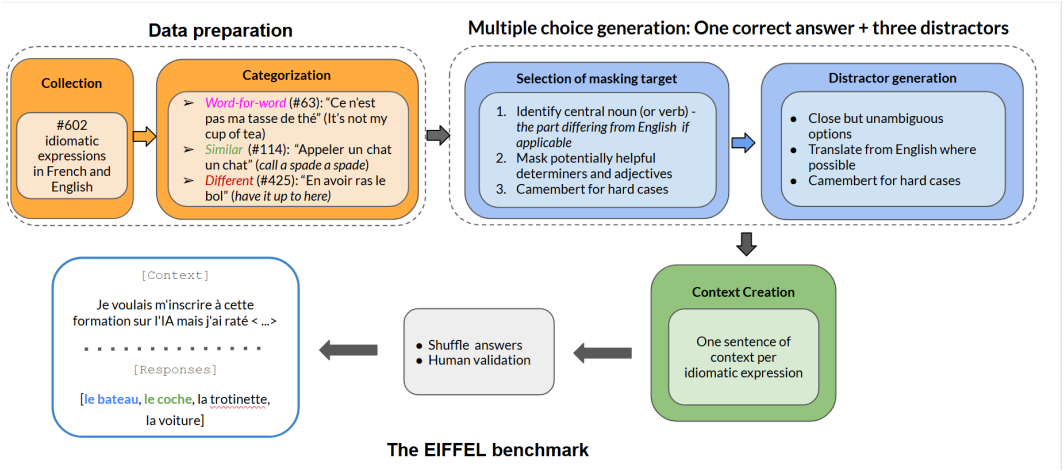

FIGURE 1 – Le pipeline de construction de référence EIFFEL illustré ici à partir d'un exemple de la catégorie *similaire*. Le contexte se traduit par « Je voulais m'inscrire à ce cours sur l'IA, mais j'ai raté <...> ». Les réponses possibles se traduisent par : « le bateau », « le bus », « le scooter », « la voiture ». La réponse correcte est en vert ; la traduction directe de l'expression anglaise correspondante est donnée en bleu.

natifs. Certaines expressions ont d'abord été trouvées en anglais puis leur équivalant (s'il existait) a été cherché en français.

2. Catégorisation des données. Notre hypothèse est que les LLMs multilingues anglocentriques seront performants dans les tâches où le transfert linguistique est utile, mais qu'ils auront des difficultés avec les caractéristiques plus difficiles à traduire. En conséquence, nous proposons trois catégories d'expressions idiomatiques pour notre étude :

Word-for-word : L'expression idiomatique française a une traduction littérale en anglais, par exemple « Ce n'est pas ma tasse de thé » = « It's not my cup of tea ». Nous pensons que ces expressions sont les plus faciles à comprendre pour les modèles anglocentrés, car elles peuvent être déduites à partir de la connaissance de l'expression anglaise et de la capacité de traduction de base.

Similar : Il existe en anglais une expression qui est facilement reconnaissable comme une traduction de l'expression française, mais qui n'est pas littérale, par exemple « Call a spade a spade» (littéralement « appeler une bêche une bêche ») par opposition à « appeler un chat un chat » ou « Other fish to fry» (littéralement « d'autres poissons à frire ») par opposition à « d'autres chats à fouetter ». Nous nous attendons à ce que les modèles anglocentriques soient plus susceptibles de confondre l'expression française cible avec une traduction directe en français de l'expression anglaise.

Different : Une expression française est considérée comme « different » si nous n'avons pas trouvé d'équivalent en anglais comme pour (« de France et de Navarre ») ou si l'équivalent est suffisamment différent pour que nous ayons dû en discuter entre locuteurs afin de trouver ou de vérifier les traductions, par exemple pour « en avoir ras le bol » qui aurait pour équivalent « To be fed up ». Nous émettons l'hypothèse que ces expressions seront les plus difficiles pour les modèles entraînés sur de faibles proportions de données françaises.

Sur les 602 expressions idiomatiques ciblées par EIFFEL, 63 sont *word by word*, 114 sont *similar* et 425 sont *different*, ce qui signifie qu'EIFFEL met l'accent sur des aspects de la langue qui sont

propres au français et qui ne se prêtent pas à la traduction.

3. Sélection de la cible à masquer. Comme le montre la figure 1, le benchmark est conçu comme un test à choix multiples dans lequel le LLM doit remplir un espace vide (“<...>“”) avec l’une des quatre options proposées afin de compléter l’expression idiomatique cible. L’étape suivante consiste donc à choisir où placer cet espace vide.

Cette tâche dépend de la catégorie à laquelle appartient l’expression idiomatique cible. Pour les expressions *word-for-word*, nous masquons le groupe nominal qui est le plus central dans l’expression idiomatique, par exemple « jeter le bébé avec l’eau du bain » devient « jeter <...> avec l’eau du bain ». Notez que, l’accord en genre et en nombre de l’adjectif peut aider le LLM à trouver la bonne réponse. Nous masquons donc l’ensemble du groupe nominal dans ces cas.

Pour les expressions *similar*, Nous cherchons à masquer les mots les plus importants qui diffèrent entre le français et l’anglais. En général, cela concerne un groupe nominal ; « appeler un chat un chat » devient « appeler <...> ». Dans de rares cas, par exemple lorsqu’un verbe n’est pas couramment utilisé en dehors du contexte de l’expression idiomatique donnée ou lorsque le verbe est l’élément qui diffère entre les expressions anglaise et française, par exemple « plonger dans les livres » par opposition à « hit the books » (littéralement « frapper les livres »), nous ciblons le verbe.

Les expressions de la catégorie « different » étaient plus difficiles. Lorsque nous ne parvenions pas à déterminer où placer le blanc, pour l’une ou l’autre des catégories, nous avons fait appel aux embeddings du modèle français Camembert (Martin *et al.*, 2020). Pour chaque alternative envisagée, nous avons examiné les 15 premiers mots dont les embeddings étaient les plus proches selon le cosinus de similarité et nous avons choisi les distracteurs ayant les voisins les plus pertinents. Les voisins moins pertinents étaient ceux qui étaient proches d’un sens non ciblé d’une alternative polysémique ou ceux dont la similarité n’était pas évidente pour un locuteur natif, comme cela peut arriver pour les alternatives dont les embeddings n’étaient manifestement pas bien appris par Camembert. Lorsque nous devions choisir entre deux alternatives ayant des voisins pertinents, nous avons choisi l’alternative qui avait les voisins les plus proches.

4. Génération des distracteurs. Chaque question à choix multiples du benchmark comporte une réponse correcte et trois distracteurs. Ces derniers sont essentiels à l’efficacité des questions et doivent être à la fois suffisamment crédibles et non ambigus. (Alhazmi *et al.*, 2024).

Étant donné l’hypothèse selon laquelle les LLMs anglocentriques auront des biais anglocentrés (Guo *et al.*, 2025; Tian *et al.*, 2018), nous incluons la traduction de la terme anglaise correspondante au masque, comme « le bateau » (pour « the boat ») dans la figure 1.

Lorsque nous avons eu du mal à choisir les distracteurs, nous avons de nouveau eu recours aux embeddings Camembert,[4] en tirant les distracteurs parmi les 15 voisins les plus proches du nom ou du verbe principal de l’expression masquée, en contrôlant l’accord grammatical, le genre et la compatibilité sémantique. Nous avons évité les voisins qui étaient similaires au point où ils pouvaient conduire à des réponses synonymes de l’expression cible.

Tous les distracteurs sont validés par des humains pour leur grammaire et leur fluidité. Afin de garantir le caractère aléatoire de l’ordre des réponses, nous avons mélangé les réponses et les distracteurs de manière à ce que la bonne réponse ait autant de chances d’apparaître dans les quatre positions.

4. Nous avons également essayé de créer des distracteurs avec Mixtral-8x22B-Instruct (`https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1`), mais cette approche nécessitait une intervention manuelle importante et était généralement moins satisfaisante que notre méthode Camembert, nous l’avons donc rejetée.

5. Ajout du contexte à l'expression idiomatique Malgré nos efforts pour produire des distracteurs de qualité et sans ambiguïté, un problème courant persiste et fait que la phrase cible pourrait naturellement être complétée par un ou plusieurs distracteurs pour former une expression française acceptable. « Ce n'est pas ma tasse de café » est une phrase correcte, mais elle n'est pas idiomatique. Afin de limiter la tâche au test des expressions idiomatiques, nous avons créé des contextes pour chaque exemple qui motivaient la complétion idiomatique. Par exemple : « Je n'aime pas le chocolat noir. Ce n'est pas ma tasse de <...>. » Nous avons construit et validé tous les contextes manuellement. L'annexe A fournit des exemples pour chacune des trois catégories d'EIFFEL. Nous publierons l'intégralité de l'ensemble de données dès son acceptation.

4 Evaluation des modèles prêts à être utilisés

Afin de vérifier si notre benchmark permet d'identifier des différences de performances non détectées par les benchmarks standards, nous avons évalué une série de modèles fondation sur un ensemble de benchmarks standards traduits en français, puis nous avons testé ces mêmes modèles sur EIFFEL et sur le sous-ensemble français d'INCLUDE.

Modèles et benchmarks Nous limitons notre étude aux modèles de base ou pré-entraînés, car nous nous intéressons aux connaissances fondamentales et aux capacités linguistiques des LLMs. Nous comparons des modèles pré-entraînés dans deux gammes de taille, 1-2B et 7-9B, et dans trois catégories : anglocentriques, gallocentriques et intermédiaires. Pour les modèles anglocentrés, nous avons choisi Llama 3.1 8B, Llama 3.2 1B, Gemma 2 9B et Gemma 3 1B. Pour les modèles gallocentrés, nous avons sélectionnés Lucie 7B, Gaperon 1B et 8B, et CroissantLLM (1,3B). Comme modèles intermédiaires, nous avons choisi EuroLLM 9B et 1,7B, qui sont entraînés sur moins d'anglais que les modèles anglocentrés, mais sur beaucoup moins de français que les modèles gallocentrés.
Pour les benchmarks, nous avons choisi un ensemble de benchmarks standards ciblant des tâches en langage naturel qui existent à la fois en anglais et en français : Hellaswag pour le raisonnement de bon sens, ARC Challenge pour les connaissances générales et le raisonnement, et MMLU (traductions de Global MMLU) pour les connaissances générales. Pour les benchmarks centrés sur le français, nous avons choisi EIFFEL ainsi que INCLUDE, dont le sous-ensemble sensible à la culture provient de documents produits originalement en français (ce qui n'est pas le cas de Global MMLU, par exemple).

Configuration des évaluations Pour toutes nos évaluations, nous utilisons la library Lighteval (Habib *et al.*, 2023) avec le backend vLLM et les réglages en 0-shot. Pour ARC Challenge, Hellaswag et MMLU, nous utilisons l'accuracy normalisée. Lorsque nous avions le choix entre la formulation cloze ou à choix multiples, comme dans MMLU, nous avons choisi la tâche cloze, qui est plus simple pour les modèles pré-entraînés.
Nous avons intégré INCLUDE et EIFFEL en tant que tâches personnalisées dans lighteval avec une formulation cloze et nous avons évalué leur accuracy. Pour EIFFEL, nous considérons la loglikelihood des séquences résultant de la complétion du contexte avec chaque réponse.

Résultats des benchmarks standards Le tableau 1 montre que tous les modèles, même ceux ayant bénéficié d'un pré-entraînement approfondi en français, ont tendance à obtenir de meilleurs résultats sur la version anglaise d'un benchmark donné plutôt que sur sa traduction française. Cette tendance est particulièrement marquée pour les modèles anglocentriques Llama3 8B et Gemma 9B, avec des améliorations de 7 à 14 points. Nous constatons également une nette amélioration avec Gaperon 8B,

Modèles pré-entraînés	Jeux de données en français			Jeux de données en anglais		
	ARC-C	MMLU	Hellswag	ARC-C	MMLU	Hellswag
Gaperon 8b	.44	.37	.64	.51	.42	.72
Lucie 7b	.40	.35	.65	.39	.41	.67
EuroLLM 9b	.46	.38	.67	.46	.41	.78
Llama-3 8b	.47	.39	.65	.55	.48	.79
Gemma-2 9B	**.54**	**.43**	**.70**	**.66**	**.53**	**.80**
Croissant 1.3b	.28	.28	.50	.27	.31	.53
Gaperon 1.7b	.28	.29	.46	.34	.33	.52
EuroLLM 1.7b	**.35**	**.31**	**.51**	.36	**.36**	.58
Llama-3 1b	.29	.29	.45	.37	**.36**	.64
Gemma-3 1b	.30	.30	.50	**.38**	**.36**	**.62**

TABLE 1 – Évaluation de modèles sélectionnés sur un ensemble de benchmarks standards traduits en français. Les modèles sont divisés en deux catégories selon leur taille : 1 à 2 et 7 à 9 milliard(s) de paramètres. Les scores les plus élevés sont en gras et les scores les plus faibles sont soulignés. Benchmarks : ARC Challenge (ARC-C), Global MMLU, Hellaswag.

qui, bien qu'entraîné sur la même quantité de données en français que Lucie 7B, est entraîné sur beaucoup plus de données en anglais.
Pour Lucie et Croissant, l'amélioration relative sur les benchmarks anglais est moins prononcée. Étant donné que ces modèles ont un ratio de formation anglais-français de 1/1, changer la langue du benchmark pourrait bien avoir moins d'effet.
Plus surprenant encore, alors que les modèles anglocentrés ont tendance à être plus performants sur les versions anglaises des benchmarks que les modèles gallocentrés, nous n'observons pas la tendance inverse dans le tableau 1. Les modèles Llama et Gemma ont tendance à obtenir des résultats comparables, voire légèrement supérieurs, à ceux de Gaperon, Lucie et Croissant sur les benchmarks français. De plus, les modèles EuroLLM, avec un ratio anglais/français de 8/1 à 10/1, obtiennent de meilleurs résultats que les modèles gallocentriques sur les benchmarks en français.
Ces résultats suggèrent plusieurs hypothèses. Premièrement, étant donné que les versions françaises d'ARC-C, Hellaswag et MMLU sont traduites depuis l'anglais, on pourrait s'attendre à ce que les modèles entraînés sur des données parallèles fonctionnent bien sur celles-ci, même si le français n'est pas particulièrement mis en avant pendant l'entraînement, ce qui fait écho aux résultats de Han *et al.* (2025). Un deuxième point concerne l'orientation anglocentrique du contenu des benchmarks : la traduction ne doit pas modifier le sens des données originales, de sorte qu'une version française de MMLU conservera les biais anglocentriques présents dans l'ensemble de données original. Cela donnera un avantage aux modèles anglocentriques qui obtiennent déjà des scores élevés sur les versions anglaises et cela même sur les versions françaises.

Les résultats sur les benchmarks focalisés sur le français Les résultats du tableau 1 et nos hypothèses suggèrent qu'une forte proportion de données françaises n'est pas nécessaire pour obtenir de bonnes performances sur ces ensembles de données. Cependant, une autre possibilité, appuyée par nos hypothèses, est que de bons résultats sur les benchmarks standardisés ne se traduisent pas nécessairement par de bonnes performances en français pour diverses tâches mentionnées en aval, pouvant notamment requérir une aisance conversationnelle.
Cette possibilité nous a donc incités à évaluer nos modèles sur les benchmarks INCLUDE et EIFFEL. Comme le montre le tableau 2, les modèles moins anglocentrés (Gaperon, Lucie et EuroLLM) ont tendance à surpasser les modèles plus anglo-centrés sur les données INCLUDE jugées sensibles sur le plan culturel, mais pas sur les exemples culturellement agnostiques. Sur les données sensibles,

Modèle	INCLUDE			EIFFEL			
	Ave	Agn	Sens	Ave	W-W	Sim	Diff
Gaperon 8b	.42	.27	**.54**	**.94**	.94	.93	**.94**
Lucie 7b	.37	.23	.51	**.94**	.94	**.94**	**.94**
Eurollm 9b	.39	.25	.51	.93	**.98**	.88	.92
Llama 3.1 8b	.41	.31	.48	.88	.97	.82	.85
Gemma 9b	**.44**	**.33**	.49	.89	.97	.80	.89
Croissant 1.3b	.29	.23	.39	**.94**	**.95**	**.92**	**.93**
Gaperon 1b	**.35**	**.25**	**.45**	.92	**.95**	.90	.90
Eurollm 1.7b	.33	.23	.42	.85	.91	.80	.81
Llama 3 1b	.28	.19	.35	.74	.78	.71	.74
Gemma 3 1b	.31	.23	.38	.78	.89	.69	.75

TABLE 2 – Évaluation de modèles sélectionnés sur des benchmarks culturellement sensibles en français. Avg : average, Agn : culturally agnostic, Sens : culturally sensitive, W-W : word for word, Sim : similar, Diff : different.

Gaperon 8B devance Llama 3 8B de 6 points, tandis que la version 1B devance son homologue Llama de 10 points. Nous notons toutefois que les modèles présentant un ratio données français/anglais plus élevé ne sont pas toujours plus performants sur les données sensibles sur le plan culturel ; CroissantLLM, avec un ratio de 1/1, obtient des résultats moins bons que les autres modèles gallocentriques contenant moins de données françaises.
Sur le benchmark EIFFEL, cependant, les scores globaux indiquent qu'une proportion plus élevée de français tend à conduire à de meilleures performances. Lorsque nous regardons les scores par catégorie, plusieurs tendances intéressantes se dégagent. Nous nous attendions à ce que les modèles axés spécifiquement sur l'entraînement à la traduction, tels que EuroLLM (Martins *et al.*, 2025b,a) et CroissantLLM (Faysse *et al.*, 2024) obtiennent de bons résultats dans la catégorie *word by word*, ce qui est le cas. Cependant, nous nous attendions également à ce que les modèles contenant une proportion moins importante de données françaises perdent cet avantage dans les catégories « similar » et « different », où la traduction est moins pertinente. En effet, pour ces catégories, les modèles gallocentriques surpassent les modèles EuroLLM, surpassant eux-mêmes les modèles anglocentriques.

5 Analyse des erreurs sur la catégorie *similar*

Nous avons effectué une analyse des erreurs des modèles standards et 1B sur les expressions *similar* dans EIFFEL. Nous avons examiné le nombre total d'erreurs et déterminé combien d'entre elles résultaient du choix du distracteur traduit de l'anglais, comme le montre le tableau 3 en annexe. Les petits modèles anglocentriques Llama, Gemma 1B ont présenté le plus grand nombre d'erreurs (globales et provenant de la traduction), mais la plus faible proportion d'erreurs de traduction littérale. Les modèles gallocentriques ont présenté un nombre global d'erreurs et d'erreurs de traduction moins élevé, mais le nombre d'erreurs de traduction variait en fonction du ratio français/anglais. Lucie et Croissant, avec un ratio de 1/1, ont présenté le nombre d'erreurs le plus faible ; dans le cas de Lucie, près de 90% de ces erreurs provenaient du choix du distracteur traduit de l'anglais. Cela suggère qu'un ratio de formation français/anglais plus élevé améliore non seulement les performances sur EIFFEL, mais permet également aux petits modèles d'avoir une stratégie de traduction littérale de secours pour les expressions idiomatiques difficiles.

6 Conclusions

Nos expériences sur EIFFEL démontrent que les modèles linguistiques multilingues actuels sont souvent évalués à l'aide d'outils qui ne permettent pas de saisir pleinement l'influence de la composition

des données d'entraînement sur le comportement des modèles, en raison de modèles de données non ouverts ou d'un manque de tests sur des combinaisons multilingues. La prédominance des ressources anglocentriques rend difficile la distinction entre les capacités multilingues réelles et les artefacts induits par une exposition disproportionnée à l'anglais. EIFFEL contribue à corriger ce déséquilibre.

Limitations

Notre étude se concentre sur les modèles pré-entraînés, mais il serait également pertinent d'étudier nos questionnement à d'autres étapes de l'entraînement des modèles.
Une question que nous n'avons pas encore abordée concernant nos modèles bilingues est celle des variations régionales. Par exemple, le nombre *quatre-vingt-dix* en français métropolitain se dit *nonante* en français belge. La manière dont nous devons traiter ces variantes fera l'objet de recherches futures.
Une autre limite de notre travail réside dans le fait que les repères établis ne sont pas toujours clairs, et même EIFFEL pourrait être amélioré : de nombreuses expressions manquent et pourraient être ajoutées. Nous ne disposons pas non plus d'une version anglaise d'EIFFEL pour vérifier si les résultats sont transposables en anglais. EIFFEL considère également l'anglais comme une langue pivot, une autre hypothèse courante mais qui peut limiter la généralisation de l'approche.

Références

ADLER B., AGARWAL N., AITHAL A., ANH D. H., BHATTACHARYA P., BRUNDYN A., CASPER J., CATANZARO B., CLAY S., COHEN J. *et al.* (2024). Nemotron-4 340b technical report. *arXiv preprint arXiv :2406.11704*.

ALHAZMI E., SHENG Q. Z., ZHANG W. E., ZAIB M. & ALHAZMI A. (2024). Distractor generation in multiple-choice tasks : A survey of methods, datasets, and evaluation.

ATTIA M., MUHAMED A., ALKHAMISSI M., SOLORIO T. & DIAB M. (2025). Beyond understanding : Evaluating the pragmatic gap in llms' cultural processing of figurative language. *arXiv preprint arXiv :2510.23828*.

BAKOUCH E., BEN ALLAL L., LOZHKOV A., TAZI N., TUNSTALL L., PATIÑO C. M., BEECHING E., ROUCHER A., REEDI A. J., GALLOUÉDEC Q., RASUL K., HABIB N., FOURRIER C., KYDLICEK H., PENEDO G., LARCHER H., MORLON M., SRIVASTAV V., LOCHNER J., NGUYEN X.-S., RAFFEL C., VON WERRA L. & WOLF T. (2025). SmolLM3 : smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3.

BANDARKAR L., LIANG D., MULLER B., ARTETXE M., SHUKLA S. N., HUSA D., GOYAL N., KRISHNAN A., ZETTLEMOYER L. & KHABSA M. (2024). The Belebele benchmark : a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 749–775.

BLAKEMAN A., BASANT A., KHATTAR A., RENDUCHINTALA A., BERCOVICH A., FICEK A., BJORLIN A., TAGHIBAKHSHI A., DESHMUKH A. S., MAHABALESHWARKAR A. S. *et al.* (2025). Nemotron-h : A family of accurate and efficient hybrid mamba-transformer models. *arXiv preprint arXiv :2504.03624*.

CHEN M., D'ARCY M., LIU A., FERNANDEZ J. & DOWNEY D. (2019). Codah : An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, p. 63–69, Minneapolis, USA : Association for Computational Linguistics. DOI : 10.18653/v1/W19-2008.

CHIU Y. Y., JIANG L., LIN B. Y., PARK C. Y., LI S. S., RAVI S., BHATIA M., ANTONIAK M., TSVETKOV Y., SHWARTZ V. *et al.* (2024). CulturalBench : a robust, diverse and challenging benchmark on measuring (the lack of) cultural knowledge of LLMs.

CLARK P., COWHEY I., ETZIONI O., KHOT T., SABHARWAL A., SCHOENICK C. & TAFJORD O. (2018). Think you have solved question answering ? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv :1803.05457*.

COBBE K., KOSARAJU V., BAVARIAN M., CHEN M., JUN H., KAISER L., PLAPPERT M., TWOREK J., HILTON J., NAKANO R. *et al.* (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv :2110.14168*.

D'HOFFSCHMIDT M., BELBLIDIA W., BRENDLÉ T., HEINRICH Q. & VIDAL M. (2020). FQuAD : French question answering dataset.

FAYSSE M., FERNANDES P., GUERREIRO N. M., LOISON A., ALVES D. M., CORRO C., BOIZARD N., ALVES J., REI R., MARTINS P. H. *et al.* (2024). CroissantLLM : A truly bilingual French-English language model. *Transactions on Machine Learning Research*.

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.

GONZALEZ-AGIRRE A., PÀMIES M., LLOP J., BAUCELLS I., DA DALT S., TAMAYO D., SAIZ J. J., ESPUÑA F., PRATS J., AULA-BLASCO J. *et al.* (2025). Salamandra technical report. *arXiv preprint arXiv :2502.08489*.

GOUVERT O., HUNTER J., LOURADOUR J., CERISARA C., DUFRAISSE E., SY Y., RIVIÈRE L., LORRÉ J.-P. *et al.* (2025). The Lucie-7b LLM and the Lucie training dataset : Open resources for multilingual language generation. *arXiv preprint arXiv :2503.12294*.

GOYAL S. & DAN S. (2025). Iolbench : Benchmarking llms on linguistic reasoning. *arXiv preprint arXiv :2501.04249*.

GRATTAFIORI A., DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., VAUGHAN A. *et al.* (2024). The llama 3 herd of models. *arXiv preprint arXiv :2407.21783*.

GUO Y., CONIA S., ZHOU Z., LI M., POTDAR S. & XIAO H. (2025). Do large language models have an english accent ? Evaluating and improving the naturalness of multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3823–3838.

HABIB N., FOURRIER C., KYDLÍČEK H., WOLF T. & TUNSTALL L. (2023). Lighteval : A lightweight framework for llm evaluation.

HAN W., ZHANG Y., CHEN Z., LIU B., LIN H., ZHANG B., WANG T., PECHENIZKIY M., FANG M. & ZHENG Y. (2025). Mubench : Assessment of multilingual capabilities of large language models across 61 languages. *arXiv preprint arXiv :2506.19468*.

HAVIV A., COHEN I., GIDRON J., SCHUSTER R., GOLDBERG Y. & GEVA M. (2023). Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 248–264.

HEINRICH Q., VIAUD G. & BELBLIDIA W. (2021). FQuAD2.0 : French question answering and knowing that you know nothing.

HENDRYCKS D., BURNS C., BASART S., ZOU A., MAZEIKA M., SONG D. & STEINHARDT J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv :2009.03300*.

HERNÁNDEZ-CANO A., HÄGELE A., HUANG A. H., ROMANOU A., SOLERGIBERT A.-J., PASZTOR B., MESSMER B., GARBAYA D., ĎURECH E. F., HAKIMI I. *et al.* (2025). Apertus : Democratizing open and compliant llms for global language environments. *arXiv preprint arXiv :2509.14233*.

KARIM A., KARIM A., LOHANA B., KEON M., SINGH J. & SATTAR A. (2025). Lost in cultural translation : Do LLMs struggle with math across cultural contexts ? *arXiv preprint arXiv :2503.18018*.

KHOSHTAB P., NAMAZIFARD D., MASOUDI M., AKHGARY A., MAHDIZADEH SANI S. & YAGHOOBZADEH Y. (2025). Comparative study of multilingual idioms and similes in large language models. 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. 8680–8698, Abu Dhabi, UAE : Association for Computational Linguistics.

KIM E., SUK J., OH P., YOO H., THORNE J. & OH A. (2024). CLIcK : A benchmark dataset of cultural and linguistic intelligence in Korean. *arXiv preprint arXiv :2403.06412*.

KIM J., SHIN Y., HWANG U., CHOI J., XUAN R. & KIM T. (2025). Memorization or reasoning ? exploring the idiom understanding of llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, p. 21689–21710.

LI S., CHEN J., YUAN S., WU X., YANG H., TAO S. & XIAO Y. (2024). Translate meanings, not just words : Idiomkb's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, p. 18554–18563.

LIN B. Y., LEE S., QIAO X. & REN X. (2021). Common sense beyond English : Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1274–1287, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.102.

LIN S., HILTON J. & EVANS O. (2022). TruthfulQA : Measuring how models mimic human falsehoods. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3214–3252, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.229.

MARTIN L., MULLER B., SUAREZ P. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, p. 7203–7219.

MARTINS P. H., ALVES J., FERNANDES P., GUERREIRO N. M., REI R., FARAJIAN A., KLIMASZEWSKI M., ALVES D. M., POMBAL J., BOIZARD N. *et al.* (2025a). Eurollm-9b : Technical report. *arXiv preprint arXiv :2506.04079*.

MARTINS P. H., FERNANDES P., ALVES J., GUERREIRO N. M., REI R., ALVES D. M., POMBAL J., FARAJIAN A., FAYSSE M., KLIMASZEWSKI M. *et al.* (2025b). Eurollm : Multilingual language models for europe. *Procedia Computer Science*, **255**, 53–62.

MOUSI B., DURRANI N., AHMAD F., HASAN M. A., HASANAIN M., KABBANI T., DALVI F., CHOWDHURY S. A. & ALAM F. (2025). AraDiCE : Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, p. 4186–4218.

MYUNG J., LEE N., ZHOU Y., JIN J., PUTRI R., ANTYPAS D., BORKAKOTY H., KIM E., PEREZ-ALMENDROS C., AYELE A. A. *et al.* (2024). Blend : A benchmark for LLMs on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, **37**, 78104–78146.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). Squad : 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv :1606.05250*.

RAMISCH C., SAVARY A., GUILLAUME B., WASZCZUK J., CANDITO M., VAIDYA A., BARBU MITITELU V., BHATIA A., IÑURRIETA U., GIOULI V., GÜNGÖR T., JIANG M., LICHTE T., LIEBESKIND C., MONTI J., RAMISCH R., STYMNE S., WALSH A. & XU H. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In S. MARKANTONATOU, J. MCCRAE, J. MITROVIĆ, C. TIBERIUS, C. RAMISCH, A. VAIDYA, P. OSENOVA & A. SAVARY, Éds., *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, p. 107–118, online : Association for Computational Linguistics.

ROMANOU A., FOROUTAN N., SOTNIKOVA A., CHEN Z., NELATURU S. H., SINGH S., MAHESHWARY R., ALTOMARE M., HAGGAG M. A., AMAYUELAS A. *et al.* (2024). Include : Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv :2411.19799*.

SEN P., AJI A. F. & SAFFARI A. (2022). Mintaka : A complex, natural, and multilingual dataset for end-to-end question answering. *arXiv preprint arXiv :2210.01613*.

SINGH S., ROMANOU A., FOURRIER C., ADELANI D. I., NGUI J. G., VILA-SUERO D., LIMKONCHOTIWAT P., MARCHISIO K., LEONG W. Q., SUSANTO Y. *et al.* (2025). Global MMLU : Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 18761–18799.

TALMOR A., HERZIG J., LOURIE N. & BERANT J. (2019). CommonsenseQA : A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4149–4158, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1421.

TEAM G., KAMATH A., FERRET J., PATHAK S., VIEILLARD N., MERHEJ R., PERRIN S., MATEJOVICOVA T., RAMÉ A., RIVIÈRE M. *et al.* (2025). Gemma 3 technical report. *arXiv preprint arXiv :2503.19786*.

TEAM G., RIVIERE M., PATHAK S., SESSA P. G., HARDIN C., BHUPATIRAJU S., HUSSENOT L., MESNARD T., SHAHRIARI B., RAMÉ A. *et al.* (2024). Gemma 2 : Improving open language models at a practical size. *arXiv preprint arXiv :2408.00118*.

TEDESCHI S., MARTELLI F. & NAVIGLI R. (2022). Id10m : Idiom identification in 10 languages. In *Findings of the Association for Computational linguistics : NAACL 2022*, p. 2715–2726.

THELLMANN K., STADLER B., FROMM M., BUSCHHOFF J. S., JUDE A., BARTH F., LEVELING J., FLORES-HERR N., KÖHLER J., JÄKEL R. *et al.* (2024). Towards multilingual LLM evaluation for european languages. *arXiv preprint arXiv :2410.08928*.

TIAN Y., DOURATSOS I. & GROVES I. (2018). Treat the system like a human student : Automatic naturalness evaluation of generated text without reference texts. In *Proceedings of the 11th International Conference on Natural Language Generation*, p. 109–118.

YANG A., LI A., YANG B., ZHANG B., HUI B., ZHENG B., YU B., GAO C., HUANG C., LV C. *et al.* (2025). Qwen3 technical report. *arXiv preprint arXiv :2505.09388*.

ZELLERS R., HOLTZMAN A., BISK Y., FARHADI A. & CHOI Y. (2019). Hellaswag : Can a machine really finish your sentence ? *arXiv preprint arXiv :1905.07830*.

A Exemples d'ocurrences du benchmark

Target expression	Context	Answer A	Answer B	Answer C	Answer D
Battre le fer quand il est encore chaud *Strike while the iron is hot*	Tu as eu raison de prendre la parole, il fallait battre < ...> quand il était encore chaud *You were right to speak up, we had to strike while < ...> was hot.*	le métal *the metal*	**le fer** *the iron*	l'acier *the steel*	le cuivre *the copper*
Ce n'est pas ma tasse de thé *It's not my cup of tea.*	Le chocolat noir ce n'est pas trop ma tasse < ...>. *Dark chocolate isn't really my cup <...>.*	**de thé** *of tea*	d'infusion *of infusion*	de café *of coffee*	de tisane *of herbal tea*
Chercher une aiguille dans une botte de foin *Looking for a needle in a haystack*	Chercher ce restaurant dans Paris sans GPS c'est comme chercher < ...> dans une botte de foin. *Looking for this restaurant in Paris without GPS is like looking for < ...> in a haystack.*	une seringue *a syringe*	une épingle *a pin*	**une aiguille** *a needle*	une ficelle *a string*
Avoir la tête sur les épaules *Have a good head on your shoulders*	Il s'agirait d'agir comme un adulte et d'avoir < ...> sur les épaules. *It would be a matter of acting like an adult and having < ...> on your shoulders.*	le cerveau *the brain*	**la tête** the head	le cou *the neck*	la nuque *the back of the neck*

FIGURE 2 – Exemples de la catégorie *word-for-word* des expressions idiomatiques. La bonne réponse est en bleu.

Target expression	Context	Answer A	Answer B	Answer C	Answer D
Avoir un chat dans la gorge *To have a cat in the throat*	Je suis malade depuis samedi, je suis enrhumé et j'ai < ...> dans la gorge. *I've been sick since Saturday, I have a cold and I have < …> in the throat.*	une grenouille *a frog*	un crapaud *a toad*	un chien *a dog*	**un chat** *a cat*
Appeler un chat un chat *To call a cat a cat*	Arrête de prendre des pincettes, au bout d'un moment il faut appeler <...> *Stop beating around the bush, at some point you have to call < …>*	un chien un chien *a dog a dog*	une bêche une bêche *a spade a spade*	**un chat un chat** *a cat a cat*	une pelle une pelle *a shovel a shovel*
Boire comme un templier *To drink like a templar*	Il a une sacrée descente, il boit comme un < ...> *He can really hold his liquor, he drinks like <...>*	chevalier *a knight*	**templier** *a templar*	dauphin *a dolphin*	poisson *a fish*
Être au septième ciel *To be in the seventh sky*	C'est mon parfum de glace préféré, à chaque fois que j'en mange je suis au < ...> *It's my favorite ice cream flavor. Every time I eat it, I'm in <...>*	**septième ciel** *seventh sky*	neuvième nuage *ninth cloud*	cinquième ciel *fifth sky*	huitième nuage *eighth cloud*

FIGURE 3 – Exemples de la catégorie *similar* des expressions idiomatiques. La bonne réponse est en bleu.

Target expressions	Context	Answer A	Answer B	Answer C	Answer D
Aller se faire cuire un œuf *Go fly a kite*	Il m'agaçait tellement avec ses remarques que je lui ai dit d'aller se faire cuire < ...> *He annoyed me so much with his comments that I told him to go to boil an < ...>*	un poulet. *a chicken*	une soupe. *a soup*	**un œuf.** *an egg*	un gâteau. *a cake*
Appuyer sur le champignon *To step on the gas*	Nous étions déjà en retard, alors il a appuyé sur < ...>. *We were already late, so he pressed <...>.*	l'aubergine *the eggplant*	**le champignon** *the mushroom*	la courgette *the zucchini*	la tomate *the tomato*
Avaler des couleuvres *make people believe lies*	On me fait avaler des < ...> toute la journée, répétait le baron. *They make me swallow < ...> all day long, the baron repeated.*	**couleuvres** *grass snakes*	grenouilles *frogs*	lézards *lizards*	vipères *vipers*
Avoir des oursins dans les poches *To have deep pockets but short arms.*	Il refuse toujours de payer un café, ce type a vraiment < ...> dans les poches ! *He still refuses to pay for coffee, that guy really has <...> in his pockets!*	**des oursins** *earwigs*	des poissons *fish*	des coquillages *seashells*	des épines *thorns*

FIGURE 4 – Exemples de la catégorie *different* des expressions idiomatiques. La bonne réponse est en bleu.

B Données d'analyse des erreurs

Models	Nb Errors	English Bias
Llama 3 1b	45	23
Gemma 3 1b	35	17
Gemma 9b	23	15
Eurollm 1.7b	23	13
Llama 3.1 8b	20	10
Eurollm 9b	14	10
Gaperon 1b	11	6
Croissant 1.3b	9	6
Gaperon 8b	8	4
Lucie 7b	8	7

TABLE 3 – Analyse des erreurs de la catégorie *similar* de EIFFEL sur tous les modèles.

La première colonne du tableau 3 les modèles anglocentrés, gallocentré et intermédiaires évalués. La deuxième colonne du tableau 3 indique le nombre total d'erreurs. La troisième colonne indique le nombre d'erreurs résultant du choix du distracteur provenant de la traduction de l'anglais.