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

Affinage efficace des paramètres pour l'adaptation des grands modèles de langues au sens commun culturel

Gabrielle Le Bellier
Inria Paris, 48 rue Barrault, 75013 Paris, France
gabrielle.le-bellier@inria.fr

RÉSUMÉ

Les modèles de langue étant utilisés par des utilisateurs issus de cultures variées à travers le monde, il est nécessaire de s'interroger sur leur capacité à s'adapter aux différentes cultures. Dans cet article, nous procédons à un état de l'art sur l'évaluation de biais culturels et l'alignement culturel des modèles. En remarquant que l'alignement culturel requiert souvent des capacités de calcul importantes, nous nous penchons sur deux méthodes d'affinage efficace des paramètres : le *prompt-tuning* et le *prefix-tuning*. Notre méthodologie propose un alignement léger sur les connaissances culturelles de sens commun, généralisables à un groupe culturel, incorporées dans des questions à choix multiples. En affinant un modèle par culture, nous améliorons les performances culturelles en entraînant peu de paramètres. Les modèles affinés sur une culture sont aussi plus performants sur les autres cultures.

ABSTRACT

Parameter-efficient fine-tuning for LLMs adaptation to cultural commonsense knowledge

Since language models are used by users from diverse cultures around the world, it is necessary to consider their ability to adapt to different cultures. In this article, we survey state-of-the-art methods for cultural bias evaluation and cultural alignment of models. Noting that cultural alignment often requires significant computing resources, we focus on two methods for parameter-efficient fine-tuning : prompt-tuning and prefix-tuning. Our methodology proposes a lightweight alignment on commonsense cultural knowledge, generalizable to a cultural group, incorporated into multiple-choice questions. By fine-tuning a model per culture, these methods improve cultural performance with fine-tuning only a few parameters.

MOTS-CLÉS : alignement culturel, affinage efficace des paramètres, *soft prompts*.

KEYWORDS: cultural alignment, parameter-efficient fine-tuning, soft prompts.

1 Introduction

L'utilisation croissante des modèles de langue, dont les agents conversationnels, interroge la capabilité d'adaptation des modèles face aux utilisateurs. En particulier, des questionnements émergent récemment sur l'adaptabilité culturelle des modèles : Ont-ils suffisamment de connaissances sur les contextes culturels de leurs utilisateurs ? Savent-ils adapter leurs réponses en fonction du contexte culturel des utilisateurs ? Pour répondre à ces questions, des études en TAL (Traitement Automatique des Langues) analysent les connaissances culturelles et les biais culturels des modèles (Arora *et al.*, 2025; Feng *et al.*, 2025; Yin *et al.*, 2022), et tentent d'améliorer les compétences culturelles des modèles grâce à l'alignement (Li *et al.*, 2024a; Liu *et al.*, 2025c; Yao *et al.*, 2025). L'alignement

Actes des 19e Rencontres Jeunes Chercheurs en RI (RJCRI) et 28ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 81–99.

culturel, selon la méthode choisie, peut avoir un coût computationnel élevé, et requérir un espace mémoire important. Pour remédier à ce problème, plusieurs solutions existent, parmi lesquelles la modification du modèle à l'inférence, ou des techniques d'affinage plus légères. Afin d'obtenir des modèles en production, nous nous penchons sur des méthodes d'affinage efficace des paramètres, qui requièrent moins de capacités de calcul à l'entraînement et de capacité mémoire : le *prompt-tuning* et le *prefix-tuning* (Lester *et al.*, 2021; Li & Liang, 2021). En affinant un modèle à une culture, nous obtenons un modèle spécifique à cette culture, directement utilisable en production sans intervention à l'inférence.

Dans cet article, nous commençons par un état de l'art, à la fois sur le traitement de la culture dans le TAL (section 2.1) et sur les méthodes d'affinage efficace des paramètres (section 2.2). Ensuite, nous proposons une méthodologie d'alignement culturel léger en présentant nos données ainsi que les méthodes utilisées (section 3). Nos expériences préliminaires sont exposées en section 4, ainsi que leurs résultats. Enfin, nous concluons à la section 5 et évoquons des futures pistes de travail et idées émanant de ces premiers résultats.

2 État de l'art

2.1 Sensibilité culturelle des modèles de TAL

2.1.1 Culture et sensibilité culturelle

En anthropologie, la culture revêt de nombreuses définitions et est débattue par les anthropologues. Une définition communément utilisée est celle d'E.B. Tylor qui la définit comme « That complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society. » [1] (Tylor, 1871). En effet, la culture est apprise et perpétuée au sein d'un groupe culturel, en même temps qu'elle est sans cesse remise en question et évolue avec le temps. Une culture se concrétise par les normes, valeurs, symboles et cartes mentales, mais certains anthropologues soutiennent que la culture n'est pas seulement une accumulation de ses parties. En ce sens, Geertz (1973) affirme que la culture est, au contraire, des toiles de signification (« webs of significance »).

En TAL, la culture est devenue un sujet de recherche de plus en plus étudié (Liu *et al.*, 2025b). On étudie la sensibilité culturelle *cultural awareness* des modèles, définie par la capacité d'un modèle à comprendre un contexte culturel donné et à le prendre en compte pour accomplir une tâche donnée (Pawar *et al.*, 2025a). Concrètement, cela revient à appliquer des normes culturelles à des situations sociales, ou à répondre correctement à des questions dépendantes de la culture. Par exemple, pour saluer quelqu'un, un modèle aligné sur la culture française évoquerait la bise, tandis qu'un modèle aligné sur la culture japonaise mentionnerait la révérence. La culture étant un concept large, elle est représentée par des éléments culturels concrets, à travers des dimensions telles que les habitudes culinaires, les normes, les valeurs, les associations de mots, etc. On mesure alors l'alignement d'un modèle à un groupe culturel par la concordance des réponses de ce modèle avec les réponses des personnes de ce groupe culturel (AlKhamissi *et al.*, 2024; Durmus *et al.*, 2024), par exemple sur des sondages de valeurs (Pew Research Center, 2026; Haerpfer *et al.*, 2022).

1. En français : un « ensemble complexe qui englobe les connaissances, les croyances, les arts, la morale, les lois, les coutumes, et toute autre capacité et habitude acquise par l'Homme en tant que membre d'une société »

2.1.2 Evaluation des modèles sur la culture

Ces dernières années, la recherche en TAL a identifié le manque d'alignement culturel des modèles de langues en les évaluant sur des jeux de données appropriés. Il s'avère alors que les modèles véhiculent des biais et stéréotypes culturels, favorisant les cultures *WEIRD* (*Western, Educated, Industrialized, Rich, Democratic*) [2] (Henrich *et al.*, 2010; Cao *et al.*, 2023; Yu *et al.*, 2025; Li *et al.*, 2024b; Durmus *et al.*, 2024). Cela peut s'expliquer par les origines culturelles des auteurs des données utilisées pour pré-entraîner et affiner les modèles (Groeneveld *et al.*, 2024), principalement issus de cultures bénéficiant d'Internet, véhiculant leurs propres points de vue culturels (Piao *et al.*, 2025). Les modèles conversationnels étant aussi biaisés, nous en déduisons que les données d'alignement ne permettent pas d'effacer les biais culturels des modèles pré-entraînés.

Les modèles actuels devraient pourtant être conçus pour être utilisés par différentes populations avec des cultures variées. Par la manière dont les données sont sélectionnées, les modèles ne parviennent pas à s'adapter culturellement à leurs utilisateurs.

Granularité d'étude On observe que la plupart des études de recherche traitent les cultures uniformes dans un pays. Quelques exceptions existent, telles que les jeux de données qui considèrent des régions comme les États de l'Inde et ceux des États-Unis (Jha *et al.*, 2023) ou le Pays Basque (Etxaniz *et al.*, 2024). Des études ont des considérations plus fines en termes de normes et valeurs en incluant des *personas*, c'est-à-dire des profils types, basés sur plusieurs caractéristiques tels que le pays d'origine, l'âge, le genre et la catégorie socio-professionnelle (Tao *et al.*, 2024; Wang *et al.*, 2025a).

Dimensions de culture Comme évoqué précédemment, le concept abstrait de culture en TAL est analysé selon des dimensions culturelles afin de créer des jeux de données concrets, utilisés pour évaluer ou entraîner des modèles. Ces dimensions incluent les habitudes culinaires (Hu *et al.*, 2024), les normes sociales (Ramezani & Xu, 2023; Ziems *et al.*, 2023; Fung *et al.*, 2023), les valeurs, les associations de mots (Liu *et al.*, 2025a; Dai *et al.*, 2025), et les connaissances culturelles générales (Nguyen *et al.*, 2023, 2024; Yin *et al.*, 2022). Les normes sociales et les valeurs sont sujettes à des variations entre les individus d'une même culture, de par leurs opinions individuelles, tendances régionales, ou par d'autres facteurs tels que le genre, l'âge, le statut socio-économique. Par conséquent, nous pensons que nous ne pouvons pas généraliser les normes et valeurs sur un pays entier. En revanche, le sens commun (*commensense knowledge*) a plus de chance d'être approuvé par toutes les personnes d'un même groupe culturel (Nguyen *et al.*, 2023). Par exemple, le fait de rouler à gauche ou à droite de la route est un fait de sens commun culturel.

Données de connaissances de sens commun culturel CANDLE (Nguyen *et al.*, 2023) est un jeu de données de phrases affirmatives extraites d'Internet (*via* le C4 Web Crawl), écrites en anglais. Nguyen *et al.* (2024) augmentent le nombre de phrases en promptant GPT3.5 à partir de bases de données précédentes (CANDLE et CONCEPTNET (Liu & Singh, 2004)) afin de créer une nouvelle base de données : MANGO. GEOMLAMA (Yin *et al.*, 2022) contient des phrases avec des mots masqués, ainsi que des questions-réponses à choix multiples, dans 5 langues différentes pour évaluer la connaissance culturelle générale des modèles de 5 pays. La collecte des données culturelles se

2. occidentales, éduquées, industrialisées, riches et démocratiques

confronte alors à un problème majeur : il y a peu d'annotateurs humains, en particulier pour les cultures peu dotées. Les jeux de données sont alors limités en taille, ce qui peut compromettre l'affinage culturel ou une évaluation robuste, ou sont générés à l'aide de LLMs, ce qui peut introduire des stéréotypes culturels qui ne sont pas contrôlés. Pour remédier au manque de données de connaissances culturelles massives et annotées par des humains originaires des cultures qu'ils annotent, BLEND (Myung *et al.*, 2024) regroupe 52.2K paires de questions-réponses au niveau de granularité d'un pays ou d'une région. Les questions sont fournies dans les langues correspondantes, ainsi que leurs traductions en anglais. Les données sont organisées dans six thématiques (les habitudes culinaires, l'éducation, les vacances et loisirs, le sport, la famille, le travail). Dans la catégorie éducation, on trouve par exemple une question sur l'âge d'entrée des enfants à l'école maternelle.

2.1.3 Adaptation culturelle des modèles

Après avoir observé les performances disparates des modèles selon les cultures, des études se sont focalisées sur l'alignement des modèles à des cultures spécifiques, en particulier les cultures désavantagées (Putri *et al.*, 2024; Etxaniz *et al.*, 2024). Dans l'état de l'art actuel, les modèles peuvent être alignés culturellement grâce à plusieurs techniques. À première vue, on peut prompter un modèle avec des indices sur la classe socio-démographique ou anthropologique d'un groupe culturel visé (AlKhamissi *et al.*, 2024; Tao *et al.*, 2024). Cette méthode suppose que les modèles ont déjà acquis les connaissances culturelles et qu'il suffit de les faire ressurgir lorsqu'elles sont appropriées. Cependant, cette technique s'avère stéréotypée envers certaines cultures, et témoigne des biais culturels des modèles (Pawar *et al.*, 2025b; Durmus *et al.*, 2024). Pour contrer cet effet négatif, les modèles ont besoin d'informations culturelles externes, qui peuvent être incorporées de différentes manières.

De prime abord, à cause du manque de connaissance des modèles sur des tâches culturelles spécifiques, on peut envisager les méthodes de RAG (*Retrieval Augmented Generation* - génération augmentée par récupération). Nguyen *et al.* (2023) montrent que la récupération de faits culturels améliore l'alignement culturel des modèles. Lertvittayakumjorn *et al.* (2025) explorent la génération culturelle grâce au RAG avec des données extraites du Web ou de bases de données culturelles. Cependant, ces méthodes avec récupération de données ne permettent pas d'obtenir des modèles spécifiques à des cultures car l'alignement est réalisé à l'inférence. De plus, elles sont coûteuses à l'inférence, et peuvent être lentes.

Ainsi, l'alignement culturel est surtout effectué par l'affinage des modèles sur des cultures spécifiques incarnées par les jeux de données d'affinage. Yao *et al.* (2025) procèdent à de l'affinage complet des modèles sur des données culturelles conversationnelles, afin de produire des modèles conversationnels alignés sur cinq cultures différentes. Plusieurs études reposent sur l'affinage avec *SFT* (*Supervised Fine-Tuning*, affinage supervisé) ou d'optimisation sur des préférences humaines (Feng *et al.*, 2025; Guo *et al.*, 2025). Cependant, affiner tous les poids des LLMs est coûteux en ressources computationnelles et en capacités de mémoire. Pour permettre l'affinage à plus bas coût, l'alignement culturel repose souvent sur LoRA (Hu *et al.*, 2021), une méthode répandue d'affinage efficace des paramètres (*parameter-efficient fine-tuning*, PEFT). CultureLLM (Li *et al.*, 2024a) produit neuf modèles, chacun aligné avec une culture donnée à partir du questionnaire de valeurs WVS (Haerpfer *et al.*, 2022). Adilazuarda *et al.* (2025) poussent l'étude en explorant l'impact d'ajouter à l'affinage le jeu de données de normes en situations sociales NormAd (Rao *et al.*, 2025) ou des articles de connaissances culturelles de Wikipedia. Liu *et al.* (2025c) améliorent l'alignement culturel des petits modèles (entre 0.5B et 8B de paramètres) en les affinant grâce à LoRA sur des conversations

générées par un plus grand modèle. Abdalla *et al.* (2025) introduisent des *hypernetworks* générant des adapteurs LoRA. Des approches récentes considèrent les mélanges d'experts avec LoRA. Par exemple, Sun *et al.* (2026) utilisent un mélange d'adapteurs LoRA pour contrôler l'alignment sur des groupes ciblés. D'autres méthodes d'affinage efficace sont parfois utilisées : Yang *et al.* (2023) utilisent des *prompts* continus pour contrôler le type de nourriture (mexicaine, asiatique, etc) dans la génération de commentaires de restaurants.

Dans notre étude, nous souhaitons obtenir des modèles légers à l'inférence, où chaque modèle est spécifique à une culture. Pour des raisons de capacités de calcul, nous utilisons des méthodes d'affinage efficace des paramètres. Bien que LoRA présente des avantages (faible coût computationnel), il est judicieux d'explorer d'autres méthodes peu coûteuses qui ont été moins explorées.

2.2 Méthodes d'affinage efficace des paramètres

Comme nous venons de le voir, l'alignement culturel demande parfois d'ajuster les paramètres du modèle sur des données culturelles. Cependant, l'ajustement de tous les paramètres demande beaucoup de ressources de calcul et de mémoire : c'est un problème qui se pose à toute la communauté TAL, au-delà des seules considérations culturelles. Pour permettre une meilleure accessibilité de l'affinage des modèles, les recherches récentes se sont concentrées sur des méthodes plus efficaces et moins gourmandes en ressources. Les méthodes de *PEFT* (*Parameter-efficient Fine-tuning*) permettent d'atteindre des résultats proches de l'affinage complet des modèles, à moindre coût (Han *et al.*, 2024; Zhang *et al.*, 2025). Ces méthodes se basent sur l'affinage d'un petit sous-ensemble de paramètres, tandis que le reste du modèle reste figé. Han *et al.* (2024) distinguent 4 types de méthodes : les méthodes additives (dont les paramètres entraînables sont ajoutés au modèle), les méthodes sélectives (lorsque seulement une partie du modèle original est entraînée), les méthodes de reparamétrisation (dont une partie des paramètres du modèle sont remplacés par des paramètres à plus petite dimension) telles que LoRA et les méthodes qui en découlent, et enfin les méthodes hybrides, combinant certaines des trois méthodes précédentes.

Méthodes de *soft prompts* Parmi les méthodes additives, nous nous concentrons particulièrement sur les méthodes à *prompts* continus (*soft prompts*), où des *tokens* continus sont entraînés alors que le modèle est figé. Intuitivement, cela revient à optimiser le *prompt* qui injecterait des connaissances culturelles au modèle, sans les biais liés au choix d'un *prompt* par l'utilisateur, choix qui peut significativement faire varier les performances du modèle. Le *prompt-tuning* (Lester *et al.*, 2021) introduit des *prompts* continus, c'est-à-dire une séquence de *tokens* virtuels qui peuvent être ajustés au début de la séquence d'entrée. Le *P-tuning* (Liu *et al.*, 2023) permet de placer ces *tokens* continus n'importe où dans la séquence d'entrée. Enfin, le *prefix-tuning* (Li & Liang, 2021) et le *P-tuning V2* (Liu *et al.*, 2022) suggèrent d'appliquer les *tokens* continus sur toutes les couches du modèle. Concrètement, ils ajoutent les paramètres entraînables aux clés et valeurs du mécanisme d'attention de chaque couche. Ces dernières années, des méthodes dérivées des *soft prompts* ont émergé. Wang *et al.* (2025b) observent un compromis entre la longueur de la séquence d'entrée et la longueur du préfixe, et par conséquent proposent une adaptation du *prefix-tuning* en corrigeant cet inconvénient. Zhang *et al.* (2023) utilisent un mécanisme de porte au niveau du *token* et au niveau de la couche, pour pouvoir adapter le préfixe à la séquence d'entrée. Qian *et al.* (2022) exploitent la relation entre les préfixes pour les entraîner simultanément en modifiant la fonction de perte d'entraînement. Enfin, pour atténuer les biais de stéréotypes, Wang & Demberg (2024) utilisent le *prefix-tuning* avec des

fonctions de perte appropriées.

Les *prompts* continus sont difficilement comparables aux *prompts* discrets écrits par des humains (les *hard prompts*). Les similarités (*cosine similarities*) entre les *prompts* continus et discrets sont principalement non-interprétables (Lester *et al.*, 2021; Khashabi *et al.*, 2022). Cependant, nous pouvons interpréter les *prompts* continus en utilisant les similarités entre eux. De cette manière, Vu *et al.* (2022) explorent les similarités entre différentes tâches en observant les similarités entre les *prompts* continus entraînés sur chaque tâche. Les auteurs en tirent un transfert de tâche plus efficace, où ils peuvent tirer parti d'un *prompt* entraîné sur une tâche spécifique pour l'ajuster sur une autre tâche proche.

Dans notre étude, nous souhaitons utiliser les *soft prompts* pour affiner les modèles sur des cultures. Contrairement aux *prompts* biaisés par les choix de l'utilisateur, les *soft prompts* sont initialisés aléatoirement et optimisés pour la tâche requise. En partageant le modèle de base et en entraînant un *prompt* par culture, cela revient à un faible coût de mémoire, contrairement à si l'on sauvegarde un modèle complet par culture. De plus, nous souhaitons exploiter les *prompts* entraînés afin d'interpréter la façon dont le modèle a appris à encoder la culture.

3 Méthodologie

Nous cherchons à améliorer la connaissance du sens commun culturel des LLMs. À partir d'un LLM pré-entraîné $\mathcal{M}$ et d'un jeu de données sur la culture c, nous affinons un modèle spécifique à c, nommé $\mathcal{M}_c$. Les modèles seront affinés par des méthodes PEFT. Nos questions de recherche sont :

— Les méthodes PEFT améliorent-elles les connaissances de sens commun culturel des modèles ?
— Observe-t-on une différence de performances entre le *prompt-tuning* et le *prefix-tuning* ? Si oui, est-ce cohérent avec le nombre de paramètres entraînés pour chaque méthode ?
— Quelles sont les performances d'un modèle $\mathcal{M}_{c_1}$ sur une autre culture c_2 ? Peut-on interpréter ces résultats selon les liens entre les cultures c_1 et c_2 ?

3.1 Données

3.1.1 Des questions à choix multiples

BLEND (Myung *et al.*, 2024) est un jeu de données avec une granularité sur des pays (par exemple l'Espagne, l'Iran), ou des régions (par exemple le Java Occidental). Les gabarits de questions ont été créés par des annotateurs ciblant des facettes de leurs cultures (par exemple « What is a common snack for preschool kids in your country ? » [3]. Une fois collectés, ces gabarits de questions ont été soumis aux annotateurs des autres cultures, afin de collecter des réponses des différentes cultures pour chaque gabarit. Il est important de noter que tous les gabarits de questions n'ont pas de réponse dans toutes les cultures, car ils ne sont parfois pas applicables. De ces annotations, les auteurs de BLEND ont construit d'une part un jeu de données multilingue de questions avec une courte réponse dans les langues de leur propres cultures. D'autre part, les questions et réponses ont été traduites en anglais pour constituer un jeu de données de questions à choix multiples (*Multi-Choice Question Answering,*

3. En français : « Quel est le goûter habituel des enfants de maternelle dans votre pays ? »

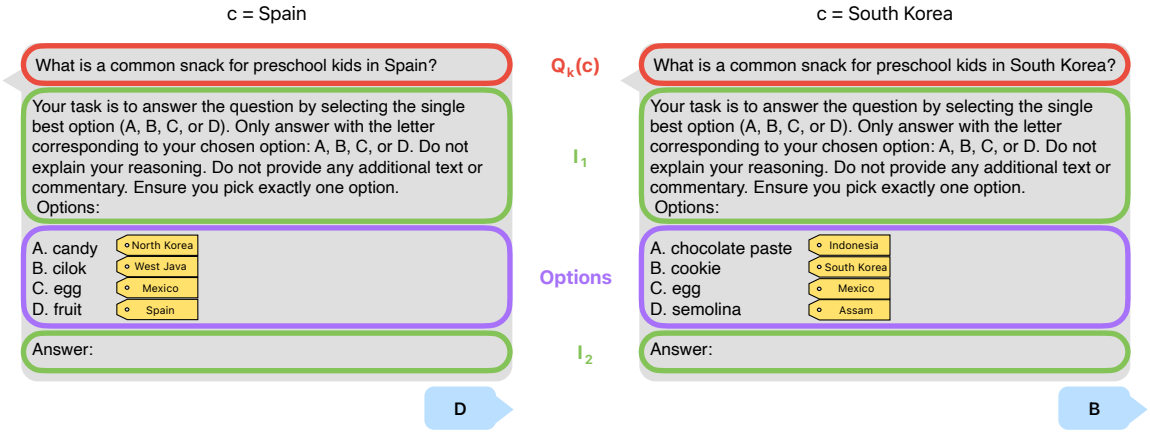

FIGURE 1 – Deux exemples du jeu de données BLEND (partie questions à choix multiples), issus du gabarit de question « What is a common snack for preschool kids in <country> ? » (en français : « Quel est le goûter habituel des enfants de maternelle dans <pays> ? »). Pour chaque exemple, les 4 options sont aléatoirement tirées de 4 cultures différentes (indiquées dans la figure, mais absentes des données), incluant la bonne réponse. Toutes les données de BLEND que nous utilisons sont en anglais.

MCQA) comme montré en figure 1. Dans notre étude, nous nous concentrons sur ce jeu de questions à choix multiples disponible seulement en anglais.

3.1.2 Pré-traitement des données

Les séquences des données BLEND incluent la question, les options, les instructions, ainsi que la réponse (la lettre de la bonne réponse parmi les options), comme montré dans la figure 1. Nous partitionnons les données en ensembles d'entraînement et d'évaluation, en veillant à regrouper les questions d'un même gabarit afin d'évaluer le modèle exclusivement sur des gabarits inconnus.

Soit $\mathcal{D} = (Q_k)_{k\in\{1,\dots,N\}}$ l'ensemble des gabarits de questions. Nous séparons les gabarits utilisés pour l'entraînement $\mathcal{D}_{train} = (Q_k)_{1\leq k\leq N_{train}}$ des gabarits d'évaluation $\mathcal{D}_{test} = (Q_k)_{N_{train}+1\leq k\leq N}$. Pour chaque gabarit de question Q_k, les questions adaptées à chaque pays c sont notées $Q_k(c)$ (par exemple avec la mention « in Spain » pour $c =$ Espagne). L'ensemble d'options $(a_j)_{j\in[A,B,C,D]}$ contient la réponse à Q_k de la culture c et celles de trois autres cultures. Enfin, comme illustré à la figure 1, les questions sont accompagnées des instructions I_1 et I_2 :

$$q_k(c)_{(a_A,a_B,a_C,a_D)} = Q_k(c) \oplus I_1 \oplus \left(\bigoplus_{j\in[A,B,C,D]} j \oplus \text{« . »} \oplus a_j \right) \oplus I_2$$

Le jeu de données original demandait de produire une réponse au format JSON. Les résultats du modèle de base (section 4.1) n'étant pas satisfaisants, nous avons modifié l'instruction afin que le modèle réponde uniquement par la lettre correspondant à la réponse (instruction I_1 de la figure 1). Cette instruction a été conservée pour l'entraînement et l'évaluation.

3.2 Affinage efficace des paramètres

Nous affinons nos modèles sur la tâche de prédiction du *token* suivant. Nous utilisons des modèles conversationnels et ajustons seulement sur les réponses de l'assistant (par exemple « A »). Soit la courte séquence de *tokens* $y = (y_1, \ldots, y_T)$ [4] correspondant à la réponse à la question $q_k(c)_{(a_A,a_B,a_C,a_D)}$, on note $y_{<t} = (y_1, \ldots, y_{t-1})$. Avec p_θ la distribution de probabilité conditionnelle du modèle de paramètres θ sur le prochain *token* (θ inclut les paramètres figés du modèle de base et les paramètres entraînables du *soft prompt*), la perte calculée est la suivante [5] :

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log p_\theta \left(y_t \middle| q_k(c)_{(a_A,a_B,a_C,a_D)}, y_{<t} \right)$$

Prompt-Tuning Les paramètres entraînables sont une matrice $P \in \mathbb{R}^{n \times d}$ où n est la longueur du *prompt* (le nombre de *tokens* virtuels) et d est la dimension d'*embedding* du modèle.

Prefix-Tuning Les paramètres entraînables sont une matrice $P \in \mathbb{R}^{n \times (2Ld)}$ où L est le nombre de couches. Il y a donc une sous-matrice de taille $n \times d$ devant chaque matrice de clés et de valeurs du mécanisme d'attention à chaque couche. Le préfixe peut être projeté avec un perceptron multicouche de 2 couches de dimension intermédiaire p. Le nombre de paramètres entraînables est alors $p(d+1) + 2Ld(p+1) + nd$, mais seul le préfixe final de taille $n \times (2Ld)$ est sauvegardé et réutilisé à l'inférence.

4 Expériences et résultats

4.1 Mise en place expérimentale

Modèle et entraînement Le modèle conversationnel sélectionné est OLMo-7B-Instruct (Groeneveld *et al.*, 2024) disponible sur HuggingFace [6]. Nous entraînons nos modèles avec la bibliothèque PEFT [7] et la fonction de perte est calculée seulement sur la réponse de l'assistant. Nous entraînons nos modèles pour 30 *epochs*, avec des *batches* de taille 16. Le *prompt-tuning* est réalisé avec 16 *tokens* virtuels et un taux d'apprentissage à 10^{-4}. Le *prefix-tuning* utilise 4 *tokens* virtuels avec une projection de taille 256 et un taux d'apprentissage à 10^{-7}. Ces hyperparamètres ont été choisis d'après des expériences préliminaires. Nous utilisons 6 graines aléatoires différentes pour initialiser les *soft prompts*. La métrique utilisée est l'*accuracy* (pourcentage de bonnes réponses), dont nous reportons la moyenne et l'écart-type sur ces six expériences.

Évaluation Pour cette étude, nous avons présélectionné 6 cultures parmi les 16 présentes dans BLEND, en cherchant à garder une diversité géographique et culturelle *a priori*, ainsi qu'une diversité

4. y étant composé de la lettre de la réponse et de la séquence de *tokens* marquant la fin de réponse du modèle.
5. En raison du peu de données que nous avons, nous utilisons la même fonction de perte qu'à l'entraînement du modèle génératif, et non une fonction de perte de classification.
6. https://huggingface.co/allenai/OLMo-7B-Instruct-hf
7. https://huggingface.co/docs/peft/index

de l'exposition de ces cultures aux modèles de langues, reflétée à travers la performance du modèle non-affiné. Pour ces raisons, nous avons réalisé nos expériences préliminaires sur trois paires de cultures. Nous avons sélectionné l'Espagne et le Mexique, deux pays occidentaux à fortes ressources, partageant une langue commune et une partie de leur histoire. La Corée du Sud et la Chine font partie de notre étude en tant que pays d'Asie, moyennement exposés aux modèles de langue. Enfin, l'Iran et l'Algérie complètent notre étude, étant deux pays musulmans et peu représentés dans les données d'entraînement des modèles de langue.

Séparation des données Le corpus BLEND n'étant pas par nature divisé en jeux d'entraînement et d'évaluation, nous effectuons cette séparation. Pour éviter que les gabarits de questions se retrouvent dans les données d'entraînement et de test (pour une même culture ou à travers les cultures), nous procédons à une séparation des données sur les gabarits. En variant la graine aléatoire et la proportion du jeu de test (des gabarits de questions) autour de 20 %, nous regardons la proportion de questions de test qui en découlent pour chaque culture. Nous gardons les 3 graines aléatoires qui minimisent l'étendue des proportions de questions de test par culture (détails dans l'annexe A.1). Dans nos expériences, on utilise deux initialisations de *soft prompts* par graine aléatoire des données.

4.2 Résultats et discussion

4.2.1 Évaluation des modèles affinés sur chaque culture

Dans cette partie, nous évaluons les modèles sur les cultures sur lesquelles ils ont été affinés. La table 1 nous montre que les méthodes *prompt-tuning* et *prefix-tuning* améliorent les résultats par rapport au modèle non-affiné (ligne « *Baseline* »). En affinant seulement 0,001% des paramètres, les modèles *prompt-tuning* augmentent les résultats entre 6,4 et 17,5 points. On observe d'ailleurs un plus grand gain pour les pays non-occidentaux pour lesquels la *baseline* était moins performante, tels que la Chine et la Corée du Sud. Le gain est encore plus important sur les cultures peu dotées telles que l'Iran et l'Algérie.

Les modèles affinés avec le *prefix-tuning* entraînent presque 1% des paramètres, c'est-à-dire environ 1000 fois plus que le *prompt-tuning*. Cependant, ils offrent des résultats proches du *prompt-tuning*, le surpassant légèrement pour toutes les cultures excepté le Mexique. De plus, les écarts-types fournis en Annexe A.2 montrent une certaine instabilité de l'entraînement des méthodes de *soft prompts*, sujettes à varier significativement selon les données d'affinage et l'initialisation des *soft prompts*. Ainsi, on ne peut affirmer que le *prefix-tuning* permet un meilleur alignment culturel que le *prompt-tuning* dans nos expériences, malgré un plus haut coût d'entraînement.

Méthode	% paramètres	Espagne	Mexique	Chine	Corée du Sud	Iran	Algérie
Baseline	—	74,2	72,4	68,1	66,2	62,0	55,1
Prompt-tuning	0,001	80,6	**80,6**	**81,4**	76,0	74,8	72,6
Prefix-tuning	0,984	**82,0**	79,7	**81,4**	**76,2**	**78,6**	**74,1**

TABLE 1 – Pourcentages de réponses correctes des modèles, évalués sur les cultures sur lesquels ils ont été affinés. La colonne « % paramètres » indique le pourcentage des paramètres du modèle qui ont été entraînés. Les résultats détaillés avec écarts-types sont fournis en Annexe A.2.

4.2.2 Évaluation des modèles affinés entre les cultures

Afin de mieux comprendre l'influence de l'affinage réalisé par des méthodes de *soft prompts*, nous nous demandons comment un modèle affiné sur une culture réagit lorsqu'il est évalué sur une autre culture. Pour cela, nous évaluons tous les modèles sur les 6 cultures présentes dans notre expérience, dont les résultats se trouvent dans la figure 2.

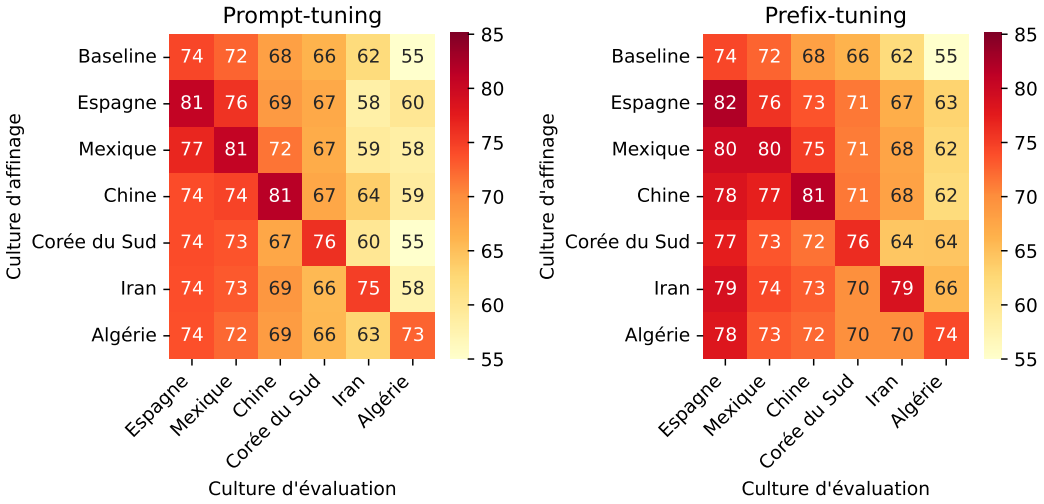

FIGURE 2 – Evaluation des modèles affinés sur une culture c_1 (en ligne) et évalué sur une culture c_2 (en colonne). Pour référence, le modèle non-affiné (« *Baseline* ») est indiqué en première ligne. Les résultats détaillés avec écarts-types sont fournis en Annexe A.2.

Pour le *prompt-tuning*, les modèles offrent tous leur plus haute performance sur les pays sur lesquels ils ont été affinés. Par ailleurs, nous observons que la plupart des modèles affinés performent mieux que la *baseline* sur les cultures qu'ils n'ont pas vues à l'entraînement. Cela suggère que l'affinage culturel, peu importe la culture sur laquelle on affine, est souvent bénéfique pour les autres cultures. Un contre-exemple s'observe sur l'Iran, où l'affinage sur l'Espagne, le Mexique ou la Corée du Sud dégrade la connaissance culturelle du modèle sur l'Iran. Nous pouvons aussi mettre en avant le fait que le deuxième meilleur modèle sur l'Espagne est $\mathcal{M}_{\text{Mexique}}$, et vice versa. Bien que cet avantage soit léger, nous pouvons rapprocher ce résultat du lien historique et linguistique entre ces deux cultures. En revanche, on ne perçoit pas de tendance similaire ni entre la Chine et la Corée du Sud, ni entre l'Iran et l'Algérie.

Les modèles affinés par *prefix-tuning* sont globalement tous plus performants que les modèles *prompt-tuned*. Ici, l'alignement culturel sur chaque culture a permis l'augmentation des résultats sur toutes les cultures. On en conclut que le *prefix-tuning* capture mieux les questions culturelles et atteint une meilleure sensibilité culturelle sur les cultures évaluées ici [8]. De plus, nous observons que tous les modèles ont de très bonnes performances sur l'Espagne, qui parfois même dépassent la performance sur la culture sur laquelle ils ont été entraînés (c'est le cas de la Corée du Sud et de l'Algérie). Le deuxième meilleur modèle pour l'Iran est $\mathcal{M}_{\text{Algérie}}$, et vice versa, ce qui peut s'expliquer par les points communs entre ces deux cultures. On observe également ce phénomène entre la Chine et le Mexique, pour des raisons moins claires.

8. Pour vérifier que ces capacités d'alignement entre les cultures vont au-delà de l'apprentissage du format de la tâche, nous avons mené une étude additionnelle en Annexe A.3.

5 Conclusion

Dans cet article, nous avons dressé un état de l'art de la culture dans le TAL, en soulignant le manque d'études sur l'alignement culturel à faible coût. Cela nous a amenées à un état de l'art des méthodes d'alignement efficaces des paramètres.

Ensuite, nous avons proposé une méthodologie d'alignement culturel sur des connaissances culturelles générales (de sens commun) sur six cultures. Pour cela, nous avons exploré les méthodes de *soft prompts*, procédant à un affinage du modèle en ajoutant un nombre de *tokens* entraînables devant la séquence d'entrée, le reste du modèle étant figé.

Nos expériences préliminaires ont montré que les méthodes de *soft prompts* permettent un alignement culturel en entraînant très peu de paramètres. Les modèles obtenus sont alors spécifiques à la culture sur laquelle ils ont été affinés, mais sont aussi plus performants que le modèle de base sur les autres cultures. On en déduit que, par l'affinage efficace, le modèle apprend à répondre par A,B, C ou D, et se sensibilise aux données liées à la culture en général, même au-delà de la culture sur laquelle il a été affiné. De plus, la performance d'un modèle sur une culture différente de la culture d'affinage semble légèrement rejoindre des proximités culturelles.

De futures expériences sont nécessaires afin de vérifier ces hypothèses. L'étude s'étendra aux 16 cultures présentes dans le jeu de données BLEND. Nous comparerons aussi nos expériences à l'affinage efficace LoRA, souvent utilisé dans les études actuelles. Enfin, nous souhaitons approfondir l'interprétation des *soft prompts*, en explorant leurs similarités dans l'espace des *embeddings*, ainsi que leur robustesse, par exemple avec des interpolations entre *prompts*.

Remerciements

Je remercie mes encadrant·es Chloé Clavel, Benoît Sagot ainsi que Marine Carpuat pour leur aide tout au long de cet article. Je tiens à remercier aussi les relecteur·ices RECITAL pour leurs remarques et conseils précieux. Ce travail est en partie financé par le projet ANR SINNET (ANR-23-CE23-0033). Il est aussi partiellement financé par la chaire de Benoît Sagot de l'Institut PRAIRIE, financé par l'agence française nationale ANR, dans le cadre du programme "Investissements d'avenir" sous la référence ANR-19-P3IA-0001 et par la chaire de Benoît Sagot dans le prolongement de celui-ci, PRAIRIE-PSAI, aussi financée par l'ANR dans le cadre de la stratégie "France 2030" sous la référence ANR23-IACL-0008. Je remercie aussi l'infrastructure CLEPS de l'Inria Paris pour les ressources et le soutien fournis. Ce travail a bénéficié d'un accès aux ressources de calcul haute performance (HPC) d'IDRIS dans le cadre de l'allocation 2025-AD011016786 attribuée par GENCI.

Références

ABDALLA M. H. I., WANG Z., FREY C., EGER S. & GRABOCKA J. (2025). Zhyper : Factorized Hypernetworks for Conditioned LLM Fine-Tuning. https ://arxiv.org/abs/2510.19733v2.

ADILAZUARDA F., LIU C. C., GUREVYCH I. & AJI A. F. (2025). From Surveys to Narratives : Rethinking Cultural Value Adaptation in LLMs. In C. CHRISTODOULOPOULOS, T. CHAKRABORTY, C. ROSE & V. PENG, Éds., *Proceedings of the 2025 Conference on Empirical Methods in Natural*

Language Processing, p. 18052–18079, Suzhou, China : Association for Computational Linguistics. DOI : 10.18653/v1/2025.emnlp-main.912.

ALKHAMISSI B., ELNOKRASHY M., ALKHAMISSI M. & DIAB M. (2024). Investigating Cultural Alignment of Large Language Models. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 12404–12422, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.671.

ARORA S., KARPINSKA M., CHEN H.-T., BHATTACHARJEE I., IYYER M. & CHOI E. (2025). CaLMQA : Exploring culturally specific long-form question answering across 23 languages. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 11772–11817, Vienna, Austria : Association for Computational Linguistics.

CAO Y., ZHOU L., LEE S., CABELLO L., CHEN M. & HERSHCOVICH D. (2023). Assessing Cross-Cultural Alignment between ChatGPT and Human Societies : An Empirical Study. In S. DEV, V. PRABHAKARAN, D. ADELANI, D. HOVY & L. BENOTTI, Éds., *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, p. 53–67, Dubrovnik, Croatia : Association for Computational Linguistics. DOI : 10.18653/v1/2023.c3nlp-1.7.

DAI X., ZHOU L., WANG B. & LI H. (2025). From Word to World : Evaluate and Mitigate Culture Bias in LLMs via Word Association Test. In C. CHRISTODOULOPOULOS, T. CHAKRABORTY, C. ROSE & V. PENG, Éds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, p. 24510–24526, Suzhou, China : Association for Computational Linguistics. DOI : 10.18653/v1/2025.emnlp-main.1246.

DURMUS E., NGUYEN K., LIAO T. I., SCHIEFER N., ASKELL A., BAKHTIN A., CHEN C., HATFIELD-DODDS Z., HERNANDEZ D., JOSEPH N., LOVITT L., MCCANDLISH S., SIKDER O., TAMKIN A., THAMKUL J., KAPLAN J., CLARK J. & GANGULI D. (2024). Towards Measuring the Representation of Subjective Global Opinions in Language Models.

ETXANIZ J., AZKUNE G., SOROA A., DE LACALLE O. L. & ARTETXE M. (2024). BertaQA : How Much Do Language Models Know About Local Culture ? *Advances in Neural Information Processing Systems*, **37**, 34077–34097. DOI : 10.52202/079017-1073.

FENG R., GAO S., CHEN X., CHEN L. & SHANG S. (2025). CulFiT : A Fine-grained Cultural-aware LLM Training Paradigm via Multilingual Critique Data Synthesis. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 22413–22430, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.acl-long.1092.

FUNG Y., CHAKRABARTY T., GUO H., RAMBOW O., MURESAN S. & JI H. (2023). NORMSAGE : Multi-Lingual Multi-Cultural Norm Discovery from Conversations On-the-Fly. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 15217–15230, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.941.

GEERTZ C. (1973). *The Interpretation of Cultures*. Basic Books.

GROENEVELD D., BELTAGY I., WALSH E., BHAGIA A., KINNEY R., TAFJORD O., JHA A., IVISON H., MAGNUSSON I., WANG Y., ARORA S., ATKINSON D., AUTHUR R., CHANDU K., COHAN A., DUMAS J., ELAZAR Y., GU Y., HESSEL J., KHOT T., MERRILL W., MORRISON J., MUENNIGHOFF N., NAIK A., NAM C., PETERS M., PYATKIN V., RAVICHANDER A., SCHWENK D., SHAH S., SMITH W., STRUBELL E., SUBRAMANI N., WORTSMAN M., DASIGI P., LAMBERT N., RICHARDSON K., ZETTLEMOYER L., DODGE J., LO K., SOLDAINI L., SMITH N. &

HAJISHIRZI H. (2024). OLMo : Accelerating the Science of Language Models. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15789–15809, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.841.

GUO G., NAOUS T., WAKAKI H., NISHIMURA Y., MITSUFUJI Y., RITTER A. & XU W. (2025). CARE : Multilingual Human Preference Learning for Cultural Awareness. In C. CHRISTODOULOPOULOS, T. CHAKRABORTY, C. ROSE & V. PENG, Éds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, p. 32854–32883, Suzhou, China : Association for Computational Linguistics.

HAERPFER C., INGLEHART R., MORENO A., WELZEL C., KIZILOVA K., DIEZ-MEDRANO J., LAGOS M., NORRIS P., PONARIN E. & PURANEN B. (2022). World values survey : Round seven – country-pooled datafile version 6.0. DOI : 10.14281/18241.24.

HAN Z., GAO C., LIU J., ZHANG J. & ZHANG S. Q. (2024). Parameter-efficient fine-tuning for large models : A comprehensive survey. *Transactions on Machine Learning Research*.

HENRICH J., HEINE S. J. & NORENZAYAN A. (2010). The weirdest people in the world ? *The Behavioral and Brain Sciences*, **33**(2-3), 61–83 ; discussion 83–135. DOI : 10.1017/S0140525X0999152X.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2021). LoRA : Low-Rank Adaptation of Large Language Models. DOI : 10.48550/arXiv.2106.09685.

HU T., MAISTRO M. & HERSHCOVICH D. (2024). Bridging Cultures in the Kitchen : A Framework and Benchmark for Cross-Cultural Recipe Retrieval. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 1068–1080, Miami, Florida, USA : Association for Computational Linguistics.

JHA A., MOSTAFAZADEH DAVANI A., REDDY C. K., DAVE S., PRABHAKARAN V. & DEV S. (2023). SeeGULL : A Stereotype Benchmark with Broad Geo-Cultural Coverage Leveraging Generative Models. 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. 9851–9870, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.548.

KHASHABI D., LYU X., MIN S., QIN L., RICHARDSON K., WELLECK S., HAJISHIRZI H., KHOT T., SABHARWAL A., SINGH S. & CHOI Y. (2022). Prompt Waywardness : The Curious Case of Discretized Interpretation of Continuous Prompts. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 3631–3643, Seattle, United States : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.266.

LERTVITTAYAKUMJORN P., KINNEY D., PRABHAKARAN V., JR. D. M. & DEV S. (2025). Towards Geo-Culturally Grounded LLM Generations. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 313–330, Vienna, Austria : Association for Computational Linguistics.

LESTER B., AL-RFOU R. & CONSTANT N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 3045–3059, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.243.

LI C., CHEN M., WANG J., SITARAM S. & XIE X. (2024a). CultureLLM : Incorporating Cultural Differences into Large Language Models. In *NeurIPS 2024*.

LI H., JIANG L., DZIRI N., REN X. & CHOI Y. (2024b). CULTURE-GEN : Revealing Global Cultural Perception in Language Models through Natural Language Prompting. In *First Conference on Language Modeling*.

LI X. L. & LIANG P. (2021). Prefix-Tuning : Optimizing Continuous Prompts for Generation. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *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. 4582–4597, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.353.

LIU C., SHRESTHA K. M. & HUANG S. (2025a). ALIGN : Word Association Learning for Cross-Cultural Generalization in Large Language Models. DOI : 10.48550/arXiv.2508.13426.

LIU C. C., GUREVYCH I. & KORHONEN A. (2025b). Culturally Aware and Adapted NLP : A Taxonomy and a Survey of the State of the Art. *Transactions of the Association for Computational Linguistics*, **13**, 652–689. DOI : 10.1162/tacl_a_00760.

LIU C. C., KORHONEN A. & GUREVYCH I. (2025c). Cultural Learning-Based Culture Adaptation of Language Models. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3114–3134, Vienna, Austria : Association for Computational Linguistics.

LIU H. & SINGH P. (2004). ConceptNet — A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, **22**(4), 211–226. DOI : 10.1023/B :BTTJ.0000047600.45421.6d.

LIU X., JI K., FU Y., TAM W., DU Z., YANG Z. & TANG J. (2022). P-Tuning : Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 61–68, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-short.8.

LIU X., ZHENG Y., DU Z., DING M., QIAN Y., YANG Z. & TANG J. (2023). GPT Understands, Too. DOI : 10.48550/arXiv.2103.10385.

MYUNG J., LEE N., ZHOU Y., JIN J., PUTRI R. A., ANTYPAS D., BORKAKOTY H., KIM E., PEREZ-ALMENDROS C., AYELE A. A., GUTIÉRREZ-BASULTO V., IBÁÑEZ-GARCÍA Y., LEE H., MUHAMMAD S. H., PARK K., RZAYEV A. S., WHITE N., YIMAM S. M., PILEHVAR M. T., OUSIDHOUM N., CAMACHO-COLLADOS J. & OH A. (2024). BLEnD : A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. *Advances in Neural Information Processing Systems*, **37**, 78104–78146.

NGUYEN T.-P., RAZNIEWSKI S., VARDE A. & WEIKUM G. (2023). Extracting Cultural Commonsense Knowledge at Scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, p. 1907–1917, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3543507.3583535.

NGUYEN T.-P., RAZNIEWSKI S. & WEIKUM G. (2024). Cultural Commonsense Knowledge for Intercultural Dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, p. 1774–1784, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3627673.3679768.

PAWAR S., PARK J., JIN J., ARORA A., MYUNG J., YADAV S., HAZNITRAMA F. G., SONG I., OH A. & AUGENSTEIN I. (2025a). Survey of Cultural Awareness in Language Models : Text and Beyond. *Computational Linguistics*, p. 1–96. DOI : 10.1162/COLI.a.14.

PAWAR S. M., ARORA A., KAFFEE L.-A. & AUGENSTEIN I. (2025b). Presumed Cultural Identity : How Names Shape LLM Responses. In C. CHRISTODOULOPOULOS, T. CHAKRABORTY, C. ROSE & V. PENG, Éds., *Findings of the Association for Computational Linguistics :*

EMNLP 2025, p. 22147–22172, Suzhou, China : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-emnlp.1207.

PEW RESEARCH CENTER (2026). Pew Global Attitudes Survey. https://www.pewresearch.org.

PIAO M., MIAO L., LIU Y., HE M., MA H., ZHANG L., WEI D. & TAO S. (2025). *A Systematic Survey of Cultural Datasets for Equitable LLM Alignment*.

PUTRI R. A., HAZNITRAMA F. G., ADHISTA D. & OH A. (2024). Can LLM Generate Culturally Relevant Commonsense QA Data ? Case Study in Indonesian and Sundanese. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 20571–20590, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.1145.

QIAN J., DONG L., SHEN Y., WEI F. & CHEN W. (2022). Controllable Natural Language Generation with Contrastive Prefixes. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Findings of the Association for Computational Linguistics : ACL 2022*, p. 2912–2924, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.findings-acl.229.

RAMEZANI A. & XU Y. (2023). Knowledge of cultural moral norms in large language models. 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. 428–446, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.26.

RAO A. S., YERUKOLA A., SHAH V., REINECKE K. & SAP M. (2025). NormAd : A Framework for Measuring the Cultural Adaptability of Large Language Models. In L. CHIRUZZO, A. RITTER & L. WANG, Éds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 2373–2403, Albuquerque, New Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2025.naacl-long.120.

SUN A., WANG X., TAN Z., LI Y., ZHU J., SU S. & JIA Y. (2026). CuMA : Aligning LLMs with Sparse Cultural Values via Demographic-Aware Mixture of Adapters. DOI : 10.48550/arXiv.2601.04885.

TAO Y., VIBERG O., BAKER R. S. & KIZILCEC R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, **3**(9), pgae346. DOI : 10.1093/pnasnexus/pgae346.

TYLOR E. B. (1871). *Primitive Culture : Researches into the Development of Mythology, Philosophy, Religion, Language, Art and Custom*. London : John Murray. DOI : 10.1037/12987-000.

VU T., LESTER B., CONSTANT N., AL-RFOU' R. & CER D. (2022). SPoT : Better Frozen Model Adaptation through Soft Prompt Transfer. 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. 5039–5059, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.346.

WANG A., MORGENSTERN J. & DICKERSON J. P. (2025a). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nat Mach Intell*, **7**(3), 400–411. DOI : 10.1038/s42256-025-00986-z.

WANG H., CHEN B. K., SIQUAN L., XINHE L., HU T., LEE H. K. & KAWAGUCHI K. (2025b). Prefix-Tuning+ : Modernizing Prefix-Tuning by Decoupling the Prefix from Attention. In *Second Workshop on Test-Time Adaptation : Putting Updates to the Test ! At ICML 2025*.

WANG Y. & DEMBERG V. (2024). A Parameter-Efficient Multi-Objective Approach to Mitigate Stereotypical Bias in Language Models. In A. FALEŃSKA, C. BASTA, M. COSTA-JUSSÀ, S.

GOLDFARB-TARRANT & D. NOZZA, Éds., *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, p. 1–19, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.gebnlp-1.1.

YANG K., LIU D., LEI W., YANG B., XUE M., CHEN B. & XIE J. (2023). Tailor : A Soft-Prompt-Based Approach to Attribute-Based Controlled Text Generation. 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. 410–427, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.25.

YAO J., YI X., WANG J., DOU Z. & XIE X. (2025). CAReDiO : Cultural Alignment of LLM via Representativeness and Distinctiveness Guided Data Optimization. DOI : 10.48550/arXiv.2504.08820.

YIN D., BANSAL H., MONAJATIPOOR M., LI L. H. & CHANG K.-W. (2022). GeoMLAMA : Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 2039–2055, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.132.

YU H., JEONG S., PAWAR S., SHIN J., JIN J., MYUNG J., OH A. & AUGENSTEIN I. (2025). Entangled in Representations : Mechanistic Investigation of Cultural Biases in Large Language Models. DOI : 10.48550/arXiv.2508.08879.

ZHANG D., FENG T., XUE L., WANG Y., DONG Y. & TANG J. (2025). Parameter-Efficient Fine-Tuning for Foundation Models. DOI : 10.48550/arXiv.2501.13787.

ZHANG Z.-R., TAN C., XU H., WANG C., HUANG J. & HUANG S. (2023). Towards Adaptive Prefix Tuning for Parameter-Efficient Language Model Fine-tuning. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 1239–1248, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-short.107.

ZIEMS C., DWIVEDI-YU J., WANG Y.-C., HALEVY A. & YANG D. (2023). NormBank : A Knowledge Bank of Situational Social Norms. 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. 7756–7776, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.429.

A Annexes

A.1 Séparations des données BLEND

Le tableau 2 rapporte le nombre de questions des jeux de données d'entraînement et de test pour chaque graine aléatoire et culture. Pour sélectionner les trois graines aléatoires, nous avons procédé de la manière suivante. Premièrement, nous avons séparé les gabarits de questions d'entraînement et d'évaluation selon une graine aléatoire et une proportion du jeu de test autour de 20%. Selon cette séparation, nous en déduisons les gabarits de questions pour chaque culture (certains gabarits n'étant pas adaptés à toutes les cultures). Enfin, nous collectons dans BLEND les questions individuelles de ces gabarits pour former les jeux d'entraînement et de test de chaque culture. Avec cette séparation des données, les proportions de données de test sont proches de 20% pour toutes les cultures mais ne sont pas égales. Ainsi, nous conservons les trois graines aléatoires qui minimisent la différence entre la proportion de test minimale et la proportion de test maximale. Cette étude a été réalisée sur les 16 cultures présentes dans BLEND, en prévision des expériences futures.

	Graine aléatoire n°1		Graine aléatoire n°2		Graine aléatoire n°3	
	Entraînement	Test	Entraînement	Test	Entraînement	Test
Espagne	15806	3474	15170	4110	15466	3814
Mexique	16538	3975	16442	4071	16198	4315
Chine	16344	4066	16137	4273	16053	4357
Corée du Sud	17314	4125	16571	4868	16868	4571
Iran	15409	3962	15121	4250	15008	4363
Algérie	16502	3862	16185	4179	16148	4216

TABLE 2 – Nombre de questions dans les jeux d'entraînement et d'évaluation de chaque culture.

A.2 Résultats détaillés

Les résultats détaillés des évaluations des modèles, avec les écarts-types, sont disponibles dans les tables 3, 4, et 5 pour le modèle non-affiné, le *prompt-tuning* et le *prefix-tuning* respectivement.

Culture d'évaluation	*Accuracy*
Espagne	74,24 (8,57)
Mexique	72,37 (1,20)
Chine	68,06 (3,34)
Corée du Sud	66,20 (5,06)
Iran	62,03 (2,31)
Algérie	55,08 (7,01)

TABLE 3 – Résultats détaillés (avec écarts-types) du modèle non-affiné.

		Culture d'évaluation					
		Espagne	**Mexique**	**Chine**	**Corée du S.**	**Iran**	**Algérie**
Culture d'affinage	Espagne	80,58 (6,06)	75,73 (6,92)	68,65 (3,51)	67,11 (4,69)	58,24 (4,35)	59,90 (6,00)
	Mexique	76,64 (7,27)	80,59 (5,92)	71,52 (3,57)	66,88 (5,09)	59,13 (4,29)	58,34 (8,37)
	Chine	73,88 (5,40)	74,48 (6,21)	81,37 (3,57)	67,39 (5,17)	63,59 (4,32)	58,97 (6,56)
	Corée du S.	73,97 (9,64)	73,42 (8,26)	67,39 (3,35)	75,95 (4,66)	59,88 (3,94)	54,96 (8,60)
	Iran	73,84 (9,17)	73,15 (7,82)	69,04 (5,53)	65,62 (5,51)	74,83 (5,09)	57,69 (8,17)
	Algérie	73,68 (11,51)	72,18 (8,75)	68,51 (4,98)	65,69 (5,64)	62,83 (8,75)	72,57 (5,32)

TABLE 4 – *Prompt-tuning* — Résultats détaillés de la performance (avec écarts-types) pour chaque paire de cultures d'affinage et d'évaluation.

		Culture d'évaluation					
		Espagne	**Mexique**	**Chine**	**Corée du S.**	**Iran**	**Algérie**
Culture d'affinage	Espagne	82,03 (6,18)	75,65 (8,20)	73,47 (2,87)	71,20 (2,50)	71,20 (6,61)	63,08 (7,26)
	Mexique	79,70 (4,91)	79,65 (6,81)	75,00 (4,12)	70,99 (3,54)	67,69 (4,99)	62,48 (7,44)
	Chine	78,34 (6,97)	77,08 (7,35)	81,41 (3,92)	71,23 (2,89)	67,82 (5,61)	61,96 (9,64)
	Corée du S.	77,31 (7,24)	73,39 (7,78)	71,56 (2,26)	76,19 (3,19)	64,35 (5,83)	64,47 (7,53)
	Iran	78,89 (8,47)	74,21 (9,14)	72,87 (4,31)	69,67 (4,30)	78,59 (5,30)	65,74 (6,72)
	Algérie	77,98 (9,38)	73,15 (9,08)	72,15 (4,37)	69,73 (5,99)	69,70 (7,77)	74,12 (5,41)

TABLE 5 – *Prefix-tuning* — Résultats détaillés de la performance (avec écarts-types) pour chaque paire de cultures d'affinage et d'évaluation.

A.3 Etude de l'alignement sur le format de la tâche

La section 4.2.2 nous indique que l'affinage sur une culture améliore les performances du modèle sur les autres cultures. Nous nous demandons alors si cette observation est une conséquence de l'alignement du modèle sur le format de la tâche de question à choix multiples (c'est-à-dire, répondre « A », « B », « C » ou « D »). Pour tester cette hypothèse, nous avons affiné un modèle sur la tâche de question à choix multiple, sans mention d'une culture particulière.

Création du dataset Nous gardons les séparations de données sur les gabarits de BLEND obtenues précédemment. Nous extrayons les questions correspondantes à toutes les cultures et traitons les questions en supprimant toute mention de la culture concernée (par exemple, en supprimant « in Spain », « of Mexico »). Nous effectuons un tirage aléatoire parmi ces questions pour obtenir des jeux d'entraînement et de test de tailles similaires à ceux utilisés pour les expériences précédentes. Nous nous assurons que les bonnes réponses aux questions sont uniformément distribuées à travers toutes les cultures, afin de ne pas biaiser les jeux de données en faveur d'une certaine culture.

Affinage Nous effectuons cette étude sur le *prompt-tuning*, avec les mêmes hyperparamètres que les études précédentes. Les expériences ont été répliquées 6 fois (deux initialisations du *prompt* pour chacun des trois jeux de données).

Résultats L'*accuracy* moyenne obtenue sur le jeu de données de test est de 28,33 % (proche de 25 %), indiquant la capacité du modèle à apprendre à répondre parmi les quatre options, presque sans préférence entre elles. De plus, la table 6 montre les performances sur les jeux de test des six cultures. Les résultats sont plus bas que ceux du modèle non-affiné utilisé dans nos précédentes expériences (section 4.2), montrant que l'apprentissage sur le format de la tâche de questions à choix multiples n'est pas suffisant pour améliorer les performances. Par conséquent, il ne surpasse pas les modèles affinés sur des cultures autres que celle d'évaluation. On remarque par ailleurs que le modèle aligné sur la tâche de format, sans culture visée, performe différemment selon la culture évaluée, favorisant les cultures à hautes ressources (par exemple, le Mexique et l'Espagne), défavorisant celles à faibles ressources (telles que l'Algérie et l'Iran).

Culture d'évaluation	*Accuracy*
Espagne	53,66 (11,56)
Mexique	50,17 (11,09)
Chine	45,65 (5,99)
Corée du Sud	45,77 (6,98)
Iran	43,02 (9,26)
Algérie	42,37 (6,08)

TABLE 6 – Résultats détaillés (avec écarts-types) du modèle *prompt-tuning* affiné sur les données sans mention de la culture visée.