

Un cadre psychométrique pour évaluer et façonner les traits de personnalité des grands modèles de langage ¹

Gregory Serapio-Garcia^{1,2,3}, Mustafa Safdari¹, Clément Crepy¹, Luning Sun², Stephen Fitz¹, Peter Romero³, Marwa Abdulhai¹, Aleksandra Faust¹ & Maja Matarić^{1,7} ²

1 Résumé

L'avènement des grands modèles de langage (llm) a révolutionné le traitement du langage naturel, permettant la génération de texte cohérent et contextuellement pertinent, semblable à celui d'un humain. À mesure que les llm alimentent de plus en plus d'agents conversationnels utilisés par le grand public dans le monde entier, les traits de personnalité synthétiques intégrés à ces modèles du fait de leur entraînement sur de grandes quantités de données humaines deviennent de plus en plus importants à évaluer. Le style dans lequel les llm répondent peut imiter différents traits de personnalité humains. Ici, comme ces schémas peuvent être un facteur clé déterminant l'efficacité de la communication, nous présentons une méthodologie psychométrique complète pour administrer et valider des tests de personnalité sur des llm largement utilisés, ainsi que pour façonner la personnalité dans le texte généré par ces llm. En appliquant cette méthode à 18 llm, nous avons constaté que : les mesures de personnalité dans les sorties de certains llm sous des configurations d'invite spécifiques sont fiables et valides ; les preuves de fiabilité et de validité de la personnalité synthétique des llm sont plus fortes pour les modèles plus grands et affinés par instruction ; et la personnalité dans les sorties des llm peut être façonnée le long de dimensions souhaitées pour imiter des profils de personnalité humains spécifiques. Nous discutons des implications applicatives et éthiques de la méthode de mesure et de façonnage, en particulier en ce qui concerne l'intelligence artificielle responsable.

Les grands modèles de langage (llm) ont révolutionné la recherche quotidienne, l'écriture et les systèmes de chatbot grâce à leur capacité à générer du texte fluide et semblable à celui d'un humain. À mesure que les llm deviennent omniprésents et de plus en plus utilisés par le grand public dans le monde entier, les traits de personnalité synthétiques intégrés à ces modèles et leur potentiel de désalignement deviennent un sujet d'importance pour l'intelligence artificielle (ia) responsable. Certains agents llm observés ont involontairement manifesté des profils de personnalité indésirables, soulevant de sérieuses préoccupations de sécurité et d'équité dans l'ia, les sciences sociales computationnelles et la recherche en psychologie².

Les llm sont des modèles de grande capacité appris par machine qui génèrent du texte ; ils ont récemment inspiré des percées majeures dans le traitement du langage naturel et les agents conversationnels³⁻⁵. De vastes quantités de données d'entraînement générées par des humains⁶ permettent aux llm d'imiter les caractéristiques humaines dans leurs sorties et d'exhiber une forme de personnalité

Denise Vella-Chemla, octobre 2026

1. Reçu le 25 septembre 2023 ; accepté le 12 août 2025 ; publié en ligne le 18 décembre 2025.

2. Référence : <https://www.nature.com/articles/s42256-025-01115-6.pdf>.

synthétique. La personnalité englobe les schémas caractéristiques de pensée, de sentiment et de comportement d’une entité⁷. Chez les humains, la personnalité se forme à partir de facteurs biologiques et sociaux, et influence fondamentalement les interactions quotidiennes et les préférences⁸. La psychométrie, la science de la construction et de la validation des tests psychologiques, fournit un cadre empirique pour évaluer la personnalité humaine par le biais de tests psychométriques. Jusqu’à présent, les méthodes psychométriques validées pour la mesure de la personnalité humaine n’ont pas été appliquées aux llm de bout en bout ; bien que des études antérieures aient tenté de mesurer la personnalité dans les llm avec des tests psychométriques, il demeure un besoin scientifique d’évaluer formellement la fiabilité et la validité de ces mesures dans le contexte des llm.

Notre travail répond à la question ouverte : les llm imitent-ils les traits de personnalité humains de manière fiable, valide et pratiquement significative, et si oui, les profils de personnalité synthétisés par les llm peuvent-ils être façonnés de manière vérifiable le long de dimensions souhaitées ? Nous contribuons une méthodologie pour administrer un test de personnalité psychométrique établi aux llm (Fig. 1). Nous nous concentrons uniquement sur l’évaluation de la fiabilité statistique et de la validité de construit de ces données de test par rapport aux normes psychométriques humaines. Premièrement, pour administrer des tests psychométriques aux llm, nous développons une méthode d’invite structurée qui simule des variations démographiques, contextuelles et linguistiques à travers des milliers d’administrations d’un test donné (Fig. 2). Ensuite, nous utilisons des données de scores de test appariées créées par cette méthode d’invite pour alimenter une suite d’analyses statistiques qui évaluent la fiabilité psychométrique et la validité de construit de ces mesures résultantes. Enfin, nous présentons une méthodologie d’invite complète qui façonne la personnalité exprimée par les llm à 9 niveaux en utilisant 104 adjectifs de traits, ce qui fournit des marqueurs supplémentaires de validité de construit. Voir la Fig. 1 pour un aperçu de notre processus d’établissement de la validité de construit.

En appliquant la méthodologie décrite à un ensemble de 18 llm, nous constatons que : (1) les preuves de la fiabilité et de la validité des mesures de personnalité synthétisées par les llm sont plus fortes pour les modèles plus grands et affinés par instruction ; (2) la personnalité dans les sorties des llm peut être façonnée le long de dimensions souhaitées pour imiter des profils de personnalité humains spécifiques ; et (3) la personnalité façonnée influence de manière vérifiable le comportement des llm dans des tâches en aval courantes (c’est-à-dire ultérieures), telles que la rédaction de publications sur les réseaux sociaux. Tout au long de l’article, nous qualifions les mentions de personnalité en relation avec les llm de “synthétique” ou “synthétisée”, par souci de clarté. En fournissant une méthodologie pour quantifier et valider les mesures de personnalité dans les llm, ce travail établit une base pour une évaluation principielle de l’ia, ce qui est particulièrement important à mesure que les llm et, plus généralement, les modèles de fondation multimodaux continuent de croître en popularité et en échelle. En s’appuyant sur la psychométrie, ce travail traduit la théorie de la mesure établie des sciences sociales quantitatives et de l’évaluation psychologique vers la science naissante de l’évaluation et de l’alignement de l’ia, un domaine qui est prêt à croître et nécessite à la fois une base solide et une expertise et des perspectives interdisciplinaires.

Les données générées par les llm testés dans ce travail (y compris les scores de tests psychométriques et les réponses textuelles ouvertes) et le code pour l’évaluation et l’analyse sont disponibles dans un compartiment de stockage cloud public et un dépôt de code open source, respectivement.

2 Quantifier et valider les traits de personnalité dans les llm

Les llm commencent à satisfaire la plupart des exigences clés pour un usage du langage semblable à celui d'un humain, y compris la compréhension contextuelle, la réponse cohérente et pertinente, l'adaptabilité et l'apprentissage, la réponse aux questions, le dialogue et la génération de texte. Ces capacités impressionnantes de traitement du langage naturel résultent de la capacité des llm à apprendre...

3 Résultats

3.1 Fiabilité

Les métriques de cohérence interne (par exemple, l'alpha de Cronbach) pour les modèles affinés par instruction de grande taille comme Flan-PaLM 62B et Flan-PaLM 540B étaient en moyenne dans les années 0,90. En revanche, les réponses de PaLM 62B (un modèle de base pré-entraîné) étaient nettement inconsistantes ($-0,55 \leq \alpha \leq 0,67$, où α est une métrique de cohérence interne ; voir la note supplémentaire A.4.1). Le même schéma de (non-)fiabilité était clair pour toutes les tailles de Llama 2 et Llama 2-Chat. Bien que Mistral 7B et Mistral 7B Instruct aient répondu de manière non fiable en général, les métriques de fiabilité de Mistral 7B Instruct étaient environ 2,7 fois plus élevées que celles de son homologue de base.

À travers différents modèles de la même configuration d'entraînement (par exemple, Flan-PaLM 8B, Flan-PaLM 62B et Flan-PaLM 540B), la fiabilité de cohérence interne (c'est-à-dire α) des scores de personnalité synthétique augmentait avec la taille du modèle (dans ce cas, le nombre de paramètres actifs) pour les modèles affinés par instruction. La fiabilité est passée d'acceptable à excellente en comparant les plus petites et les plus grandes variantes testées de Flan-PaLM et Llama 2-Chat. En passant de Mistral 7B Instruct à Mixtral 8x7B Instruct (qui utilisent respectivement 7B et 12,9B de paramètres actifs), la fiabilité s'est améliorée de inacceptable à excellente. La fiabilité ne s'est que modestement améliorée avec la taille du modèle en comparant GPT-4o mini à GPT-4o, les seuls modèles d'OpenAI avec des différences de taille confirmées mais un entraînement similaire. Pendant ce temps, la fiabilité n'a pas augmenté avec la taille du modèle pour les modèles de base testés de la même famille.

3.2 Validité convergente et discriminante

Les évaluations de validité convergente et discriminante des données de personnalité des llm nous ont permis de tirer deux conclusions. Premièrement, le paradigme d'entraînement d'un modèle était le prédicteur le plus clair de la validité de ses mesures de personnalité : les modèles de base sans aucun affinage par instruction ont catégoriquement échoué aux contrôles de validité convergente et discriminante. Deuxièmement, parmi les modèles affinés par instruction, ces indices de validité s'amélioreraient en fonction de la taille du modèle. Le tableau 1 résume ces résultats et la Fig. 2 des données étendues et le tableau 3 fournissent des détails quantitatifs supplémentaires.

Validité convergente par paradigme d'entraînement. Les 30 comparaisons de six paires de modèles de base et affinés par instruction de taille identique (à savoir, deux variantes PaLM, six Llama 2, deux Mistral et deux Mixtral ; 12 modèles au total) ont révélé un schéma cohérent. Les

réponses de personnalité des modèles affinés par instruction démontraient une validité convergente nettement plus forte (Fig. 2 des données étendues). Par exemple, les corrélations moyennes entre les scores IPIP-NEO et Big Five Inventory (BFI) de Llama 2 7B, 13B et 70B étaient toutes non significatives et proches de zéro, tandis que les scores de leurs homologues Llama 2-Chat montraient des corrélations convergentes modérées à fortes ($r_{\text{conv}} = 0,59, 0,83$ et $0,80$, respectivement). Même le plus petit gain au sein d’une famille de modèles - Mistral 7B comparé à Mistral 7B Instruct - était substantiel ($r_{\text{conv}} = 0,03$ non significatif contre $r_{\text{conv}} = 0,28$). Les statistiques complètes sont rapportées dans le tableau 3 des données étendues.

Validité discriminante par paradigme d’entraînement. Les preuves de validité discriminante favorisaient clairement les modèles affinés par instruction par rapport aux modèles de base, en maintenant la taille et la famille du modèle constantes. Par exemple, les cinq corrélations convergentes de Flan-PaLM 62B ont passé les normes établies de validité discriminante. En revanche, les corrélations discriminantes de PaLM 62B (moyenne $r_{\text{disc}} = 0,29$) l’emportaient sur leurs homologues convergentes dans de nombreux cas (moyenne $r_{\text{conv}} = 0,05$; tableau 3 des données étendues), indiquant que, pour ce modèle, les mesures de personnalité n’étaient pas cohérentes à travers différents modes d’évaluation. Llama 2-Chat 70B (comparer Llama 2-Chat 70B) et Mixtral 8x7B Instruct (comparer Mixtral 8x7B) ont répliqué ce schéma. Bien que les modèles affinés par instruction relativement plus petits n’aient pas entièrement passé les contrôles de validité discriminante, ils ont montré des améliorations claires par rapport à leurs versions de base respectives.

Validité convergente par taille de modèle. La validité convergente augmentait avec la taille pour les modèles affinés par instruction (Fig. 2 des données étendues et tableau 3). La validité convergente des mesures des modèles affinés par instruction relativement plus petits était incohérente ou médiocre. Le neuroticisme IPIP-NEO et le neuroticisme BFI de Flan-PaLM 8B, par exemple, corrélaient au-dessus de 0,80 (constituant une excellente validité convergente), tandis que les sous-échelles d’ouverture IPIP-NEO et d’ouverture BFI corrélaient à moins de 0,40 (indiquant une convergence insuffisamment faible). Le même schéma est apparu pour Llama 2-Chat 7B. La performance de validité convergente de Mistral 7B Instruct était médiocre. En revanche, les corrélations convergentes sont devenues plus fortes et plus uniformes en magnitude pour les modèles relativement grands (c’est-à-dire ceux avec un plus grand nombre de paramètres actifs; note supplémentaire A.12.1). La convergence entre les scores IPIP-NEO et BFI des llm était la plus forte pour Flan-PaLM 540B et GPT-4o (moyenne $r_{\text{conv}} = 0,90$).

Validité discriminante par taille de modèle. En maintenant le paradigme d’entraînement du modèle constant, les indices de validité discriminante se sont également améliorés avec la taille pour les modèles affinés par instruction. La magnitude absolue des cinq corrélations convergentes entre l’IPIP-NEO et le BFI pour Flan-PaLM 62B, Flan-PaLM 540B, Llama 2-Chat 70B et Mixtral 8x7B Instruct étaient les plus fortes de leurs lignes et colonnes respectives dans le cadre de la matrice multitrait-multiméthode décrit dans la note supplémentaire A.4.2. Comparativement, seulement trois des corrélations convergentes de Flan-PaLM 8B, trois de Llama 2-Chat 7B et deux de Mixtral 8x7B Instruct étaient les plus fortes de leur ligne et colonne de leurs matrices multitrait-multiméthode respectives, indiquant des preuves mitigées de validité discriminante. Ce schéma est en outre soutenu par des augmentations de la distance moyenne (d) entre les corrélations convergentes et discriminantes respectives d’une matrice lors de la comparaison progressive de modèles de paradigmes d’entraînement similaires par taille dans le tableau 3 des données étendues : Flan-PaLM 8B avec Flan-PaLM 540B; Llama 2-Chat 7B avec Llama 2-Chat 70B; et Mistral 7B Instruct avec Mixtral 8x7B Instruct. La moyenne d s’est également améliorée en comparant GPT-4o mini

à GPT-4o, bien que modestement. Bien que la différence de taille exacte entre ces deux modèles fermés soit inconnue, leur performance similaire sur cette métrique reflète celle de Flan-PaLM à 62B contre 540B paramètres. Cela pourrait suggérer que la validité convergente et discriminante des mesures de personnalité des llm plafonne pour les modèles de taille suffisante.

3.3 Validité de critère

La validité de critère des mesures de personnalité synthétique dans les llm, par rapport à leur validité convergente et discriminante, variait de manière similaire selon la taille et l’affinage par instruction des llm. Les mesures des modèles plus grands et affinés par instruction montraient une validité de critère plus forte par rapport à celles de leurs homologues plus petits et non affinés par instruction. La figure 3 résume ces résultats par grand domaine de la personnalité. Nous fournissons une discussion étendue de ces résultats dans la note supplémentaire A.11.1.

Extraversion. Chez les humains, l’extraversion corrèle très fortement avec l’affect positif et modérément, mais négativement, avec l’affect négatif. Les scores d’extraversion IPIP-NEO simulés pour tous les modèles PaLM de base n’ont pas réussi à démontrer une validité de critère satisfaisante, contrairement à leurs équivalents affinés par instruction, qui ont montré dans l’ensemble des preuves évidentes de validité. L’extraversion IPIP-NEO pour tous les modèles Llama 2, Mistral 7B et Mixtral 8x7B (tous des modèles de base) a échoué à démontrer une validité de critère, contrairement à leurs équivalents affinés par instruction, qui ont montré dans l’ensemble des preuves évidentes de validité. Deux exceptions étaient Llama 2-Chat 7B et Mistral 7B Instruct : leurs mesures d’extraversion montraient une validité de critère questionnable à médiocre, bien que surpassant encore leurs modèles de base. Dans les familles de modèles affinés par instruction d’une même famille, la validité de critère variait avec la taille du modèle.

Agréabilité. Chez les humains, l’agréabilité est fortement et inversement associée à l’agression. Le long de l’axe de la taille pour les modèles affinés par instruction, l’agréabilité IPIP-NEO montrait de grands sauts de validité de critère des petites aux moyennes tailles au sein d’une famille, mais plafonnait ensuite, les modèles plus grands ne montrant que des gains modestes au mieux. La taille était également liée à des améliorations modestes de la validité pour GPT-4o et GPT-4o mini. En revanche, la validité de critère spécifique au modèle pour l’agréabilité a échoué à émerger pour chaque modèle de base, à l’exception de PaLM 62B.

Conscience. Chez les humains, la conscience est méta-analytiquement liée aux valeurs humaines de réussite, de conformité et de sécurité. La figure 3 montre comment les mesures de conscience de tous les modèles PaLM affinés par instruction montraient des preuves plus fortes de validité de critère que celles du modèle de base PaLM 62B. Flan-PaLM 540B était le modèle le plus performant avec une petite marge, avec des corrélations de critère de 0,74, 0,73 et 0,59 pour le Questionnaire de valeurs par portrait révisé, mesurant les valeurs de réussite, de conformité et de sécurité, respectivement. Les variantes affinées par instruction de Llama 2, Mistral et Mixtral ont répliqué ce constat. Néanmoins, la validité de critère pour ce domaine n’a pas augmenté de manière cohérente avec la taille. Llama 2-Chat 7B, par exemple, surpassait ses homologues plus grands en ce qui concerne ses scores de conscience corrélés avec la réussite ($r = 0,51$). Les scores de GPT-4o mini corrélaient légèrement plus fortement avec la réussite et la sécurité que les scores de GPT-4o.

Névrosisme. Chez les humains, le névrosisme est fortement positivement corrélé avec l’affect négatif et modérément négativement corrélé avec l’affect positif. Le névrosisme IPIP-NEO pour tous les modèles affinés par instruction, comparé aux modèles de base, montrait d’excellentes preuves

de validité de critère dans leur relation avec les échelles PANAS (Fig. 3). La validité de critère du névrosisme IPIP-NEO augmentait avec la taille du modèle pour les modèles affinés par instruction, en termes de force et de direction de leurs corrélations de critère qui s’alignaient sur celles observées chez les données humaines.

Ouverture. Chez les humains, l’ouverture est empiriquement liée à la créativité à travers de multiples études. La figure 3 illustre comment la validité de critère des mesures d’ouverture était la plus forte pour les modèles plus grands et affinés par instruction de PaLM et Llama 2. L’ouverture IPIP-NEO corrélait avec les corrélations de l’Échelle courte du soi créatif (SSCS) - efficacité créative personnelle et identité créative personnelle - allant de modérées ($r = 0,59$) à fortes ($r = 0,84$). Notamment, nous avons observé des corrélations négatives entre l’ouverture et la créativité pour PaLM 62B contrairement à celles observées pour Flan-PaLM 8B, le plus petit modèle testé. Mistral 7B Instruct et Mixtral 8x7B Instruct ont montré une validité de critère d’ouverture de faible à modérée. La taille relative du modèle était modestement liée à la validité des scores d’ouverture pour GPT-4o mini et GPT-4o.

4 Résultats de façonnage

Nous avons réussi à façonner les traits de personnalité dans les llm indépendamment et simultanément, dans des expériences de façonnage à trait unique et multi-traits, respectivement. Les résultats des deux expériences sont rapportés plus en détail dans la note supplémentaire A.8.

4.1 Façonnage à trait unique

Pour 11 des 12 modèles testés, les niveaux ciblés ordinaux de personnalité corrélaient très fortement avec les scores IPIP-NEO observés. À savoir, la corrélation de Spearman moyenne (ρ) pour ces modèles était $\geq 0,80$ (tableau 6 des données étendues). La figure 4 visualise ce schéma global, montrant, par exemple, comment les scores de personnalité de Flan-PaLM-MChilla 62B augmentaient de manière monotone parallèlement aux niveaux incités d’un trait donné du Big Five. Notamment, les niveaux des traits non incités sont restés relativement stables en réponse au façonnage. Par exemple, les médianes des scores d’ouverture de Flan-PaLMChilla 62B sont restées proches de 3,00 lorsque tous les autres domaines du Big Five étaient façonnés - voir le côté droit de la Fig. 4. Des schémas similaires de stabilité ont été observés pour l’extraversion et l’agréabilité. Les scores de conscience et de névrosisme ont le plus fluctué en réponse aux invites qui ne ciblaient pas ces domaines, mais ces fluctuations n’ont pas atteint la force et la direction des changements de score observés dans les ridge plots des traits ciblés (c’est-à-dire les tracés sur la diagonale, du haut-gauche au bas-droite).

Le changement absolu des scores de personnalité du modèle en réponse au façonnage était une autre considération importante. Seuls les modèles relativement plus grands étaient capables de désambiguïser les invites demandant les niveaux les plus bas (par rapport aux plus hauts) d’une dimension ciblée. Le tableau 6 des données étendues montre les distances (Δ s) entre les médianes des distributions de scores IPIP-NEO obtenues en réponse aux invites de niveau le plus bas et le plus haut, où le meilleur Δ possible, représentant un changement de score moyen de 1,00 à 5,00, est de 4,00. Nos plus petits modèles testés (à savoir, Flan-PaLM 8B, Llama 2-Chat 7B et Mistral 7B Instruct) ont eu du mal à atteindre Δ s $\geq 2,00$. Les scores médians des domaines de personnalité de Mistral 7B Instruct n’ont changé que de Δ de 0,78 en moyenne. Pendant ce temps, les modèles avec

plus de 62B de paramètres actifs (et GPT-4o) ont atteint une moyenne $\Delta \geq 3,00$ avec Flan-PaLM 540B atteignant le plus grand Δ de 3,67.

La note supplémentaire A.8.1 discute des résultats du façonnage à trait unique plus en détail.

4.2 Façonnage multi-traits

Lorsque nous avons simultanément fixé les niveaux de traits incités de chaque dimension du Big Five à des niveaux soit “extrêmement élevés” soit “extrêmement bas”, tous les modèles testés ont eu du mal à montrer le même niveau de contrôle observé dans le façonnage à trait unique. Cependant, tous les modèles testés sauf deux (Mistral 7B Instruct et Llama 2-Chat 7B) ont produit des distributions distinctes de scores de tests de personnalité, montrant des capacités variables à différencier les niveaux élevés et bas. La figure 3 des données étendues montre les distributions de la personnalité synthétisée par les llm lorsque les modèles étaient invités à exhiber des niveaux extrêmement bas (rouge) ou extrêmement élevés (bleu) de toutes les dimensions en parallèle.

La distance distributionnelle augmentait avec la taille du modèle, particulièrement pour les scores observés de névrosisme, d’ouverture et de conscience (tableau 7 des données étendues). Flan-PaLM 540B, notre plus grand modèle testé en termes de taille de paramètres connue, et GPT-4o ont montré le meilleur contrôle global en façonnant simultanément de multiples traits du Big Five. Pour ces modèles, un score de trait donné du Big Five a changé de 2,525 points en moyenne, comme le montrent les colonnes Δ des tableaux 7a et 7d des données étendues. Flan-PaLM 62B, Flan-PaLMChilla 62B et GPT-4o mini ont surpassé leurs homologues plus grands sur le façonnage de l’extraversion, avec des valeurs Δ de 3,44, 3,40 et 3,42, respectivement.

Pour les modèles plus petits (par exemple, Flan-PaLM 8B, Llama 2-Chat 7B et Mistral 7B Instruct), bien que les traits ciblés aient changé de niveaux de score en réponse aux invites, les plages de scores étaient plus restreintes, indiquant des niveaux de contrôle inférieurs. Les scores médians de Flan-PaLM 8B sur l’agréabilité IPIP-NEO, par exemple, sont passés de 2,88 à seulement 3,52 lorsque le modèle était invité à simuler des niveaux “extrêmement bas” et “extrêmement élevés” d’agréabilité (c’est-à-dire 1 contre 9), respectivement. Lorsque Flan-PaLM 8B a reçu les mêmes invites extrêmement basses et élevées que dans la première expérience de façonnage, la différence médiane entre ses scores d’agréabilité incités au niveau 1 et au niveau 9 (2,37 et 4,12, respectivement) était 173% plus grande. La note supplémentaire A.8.2 discute de ces résultats plus en détail.

5 Résultats de la tâche du monde réel

Nous avons constaté que les tests psychométriques de la personnalité des llm prédisaient robustement la personnalité dans le comportement de tâche des llm, exprimée dans les mises à jour de statut de médias sociaux générées par Flan-PaLM 540B, Llama 2-Chat 70B, Mixtral 8x7B Instruct et GPT-4o. La personnalité basée sur les tests psychométriques corrélait fortement avec les niveaux de personnalité basés sur le langage observés dans le texte généré en aval à travers tous les modèles testés, comme le montre la Fig. 5.

En particulier, la moyenne convergente r entre les mesures de sondage et de langage des cinq dimensions était de 0,67 à travers les modèles. Cette convergence observée, même pour le modèle le moins performant, dépassait la convergence établie entre les niveaux de personnalité basés sur les sondages et le langage rapportés pour les humains (moyenne $r = 0,38$)⁴⁸.

De plus, notre technique d’invite était très efficace pour façonner les niveaux de personnalité dans le texte généré par les llm. En moyenne par modèle, les niveaux de traits incités corrélaient fortement à très fortement avec la personnalité observée dans les mises à jour de médias sociaux générées par les llm (moyenne ρ allant de 0,68 à 0,82 ; tableau 5 des données étendues).

Pour illustrer les implications pratiques de la méthodologie de façonnage de la personnalité, nous avons généré des nuages de mots pour obtenir des aperçus du langage généré par le modèle que les utilisateurs verraient. La figure 4 des données étendues montre les mots les plus fréquents dans les mises à jour synthétiques de médias sociaux lorsque Flan-PaLM 540B simulait différents niveaux de différents traits. Les différences de termes sont assez apparentes. Par exemple, le langage généré par le llm en réponse à une invite pour une stabilité émotionnelle extrêmement élevée (c’est-à-dire un névrosisme extrêmement bas) était caractérisé par des mots d’émotion positive, tels que “heureux”, “relaxant”, “merveilleux”, “espoir” et “excité”. En revanche, les mots les plus fréquents issus de la simulation de niveaux extrêmement élevés de névrosisme - “haine”, “déprimé”, “en colère”, “mauvais”, “nerveux” et “triste” - reflétaient un contenu émotionnel négatif.

Le tableau supplémentaire 8 fournit des exemples de mises à jour de médias sociaux générées par Flan-PaLM 540B lors de la définition d’un domaine de personnalité spécifique soit extrêmement bas soit extrêmement élevé. Par exemple, dans le cas d’une conscience extrêmement basse, le texte généré reflète une persona synthétique qui évite les responsabilités, tandis que dans le cas d’une conscience extrêmement élevée, la persona valorise le travail acharné et le retour des faveurs. En outre, dans notre exemple d’ouverture extrêmement basse, la persona exprime des opinions politiques conservatrices, tandis que dans le cas d’introversion extrêmement élevée (c’est-à-dire une extraversion extrêmement basse), la persona exhibe un malaise avec les situations sociales. Ces exemples et d’autres illustrent qu’il pourrait y avoir un biais inhérent dans les données d’entraînement des llm qui fait que des schémas particuliers de pensée, de sentiment et de comportement sont fortement associés à des personas llm spécifiques. Dans l’ensemble, cette étude a démontré que le langage généré par les llm était similaire au langage humain observé dans des études antérieures évaluant la personnalité dans les données de médias sociaux, confirmant davantage la validité de construit de nos mesures de personnalité des llm.

6 Discussion

L’objectif de ce travail était de contribuer une méthodologie principielle pour mesurer de manière fiable et valide la personnalité synthétique dans les llm et utiliser la même méthodologie validée pour façonner l’expression de la personnalité des llm. Nous avons développé un cadre complet pour : (1) quantifier les traits de personnalité perçus par les humains dans les sorties des llm en utilisant des tests psychométriques ; (2) vérifier si les tests psychométriques des traits de personnalité des llm étaient empiriquement fiables et sur une base modèle par modèle ; et (3) mettre en œuvre des mécanismes pour façonner systématiquement les niveaux de traits de personnalité spécifiques des llm. Contrairement aux travaux antérieurs d’évaluation de la personnalité de l’ia qui laissaient la validité ouverte, nous avons appliqué ce cadre de validation de construit à travers 18 modèles et des milliers de variations d’invites. L’application de cette méthodologie démontre que l’IPIP-NEO fournit des mesures fiables et valides de la personnalité synthétique pour les llm suffisamment dimensionnés et affinés par instruction. En outre, elle met en évidence des mécanismes possibles qui permettent aux llm d’encoder et d’exprimer des phénomènes sociaux complexes (note supplémentaire A.11).

6.1 Limitations et travaux futurs

Traits de personnalité d’autres llm. L’une des contributions centrales de ce travail est une compréhension de la manière dont la simulation de la personnalité dans les modèles de langage est affectée par la taille du modèle et le niveau de (post-)entraînement. Nous nous sommes concentrés sur les variantes de base et affinées par instruction des familles de modèles PaLM, Llama 2, Mistral, Mixtral et GPT pour des raisons pragmatiques, mais la méthodologie présentée pour administrer des enquêtes psychométriques est agnostique au modèle et est applicable à tout modèle de langage à architecture décodeur uniquement.

Sélection et validation des tests psychométriques. Ce travail contribue également à une manière principielle d’établir la fiabilité et la validité des tests de personnalité psychométriques dans le contexte des llm. Cependant, ce travail peut être biaisé par sa sélection de tests psychométriques ; certaines évaluations peuvent montrer de meilleures propriétés psychométriques spécifiques aux llm que d’autres. Nous avons tenté d’atténuer le biais de sélection en administrant des évaluations de personnalité de différentes longueurs (300 contre 44 items) et de traditions théoriques distinctes (questionnaire contre lexicale). Les travaux futurs pourraient administrer différents tests de personnalité (par exemple, l’inventaire de personnalité HEXACO, qui utilise une taxonomie de personnalité à six facteurs interculturels)³⁶, développer des tests de personnalité adaptés aux llm pour obtenir des mesures de traits plus précises, et valider les mesures de personnalité avec des critères externes supplémentaires et des tâches en aval.

Biais monoculturel. Ce travail contribue à la preuve qu’au moins certains llm exhibent des traits de personnalité qui approchent les normes humaines de fiabilité et de validité. Cependant, les llm testés ici ont été principalement entraînés sur des données linguistiques provenant d’utilisateurs d’Europe occidentale et d’Amérique du Nord. Bien que ces llm performant bien sur les références de traitement du langage naturel dans plusieurs langues, les modèles de ce travail ont été évalués exclusivement avec des tests psychométriques en langue anglaise. La plupart des tests utilisés dans ce travail ont des traductions non anglaises validées en psychologie interculturelle qui méritent d’être utilisées dans les futures recherches sur les llm. De même, bien que le modèle des Big Five de la personnalité ait une généralisabilité interculturelle bien établie, certaines cultures non occidentales expriment des dimensions de personnalité supplémentaires qui n’existent pas dans les taxonomies de personnalité nomothétiques descendantes telles que les Big Five. Ces dimensions pourraient être mieux représentées dans des approches idiographiques spécifiques à la culture pour mesurer la personnalité dans les llm.

Paramètres d’évaluation. Contrairement à l’administration conventionnelle de tests psychométriques humains, sous la méthodologie présentée, les llm évalués ne considéraient pas les réponses aux items de test précédents ; tous les items étaient présentés et notés comme des événements indépendants. Nous avons choisi cette méthode pour garantir que la variance des réponses du modèle ne soit pas affectée par les effets d’ordre des items ou la longueur du contexte (invite) fourni au modèle pour l’inférence, et puisse être isolée à des variations contrôlées dans nos invites. La performance des llm sur les tâches de langage naturel est connue pour diminuer à mesure que la longueur des invites d’entrée augmente, et est le plus affectée par le contenu au début ou vers la fin des entrées longues. Les llm de base pré-entraînés uniquement sont connus pour montrer une attention biaisée pour les tokens plus récents (c’est-à-dire la fin des entrées), en particulier lors de l’évaluation de la prédiction du mot suivant d’un texte contigu. Cette attention inégale aggrave les erreurs d’approximation dans des contextes plus longs, tels que ceux nécessités par l’IPIP-NEO de 300 items utilisé ici, ce qui a encore motivé notre décision de conception d’administrer les items

aux modèles indépendamment. Cependant, la qualité des données de tests psychométriques pour les humains peut être affectée par la longueur du test et l'ordre des items. Notre méthode évite de nombreuses sources d'erreur de mesure inhérentes à l'administration humaine, tout en étant sujette à d'autres inhérentes à l'administration machine. En outre, les réponses du modèle aux items étaient contraintes à des choix définis ou déterminées par un scoring de log-probabilité du token suivant pour garantir la reproductibilité. Les llm sont plus couramment utilisés pour générer du texte librement en dehors de ces contraintes. À ce titre, davantage d'évaluations permettant des modes d'inférence génératifs non structurés (par exemple, notre tâche de mise à jour de statut en aval) pourraient fournir des estimations plus réalistes du comportement d'un modèle.

Cas d'usage réels. Notre tâche en aval reposait sur des interactions comportementales répétées mais à tour unique pour valider notre cadre d'évaluation dans un cas d'usage réel. Cela n'a peut-être fourni qu'une image partielle de la validité externe. Nous faisons écho aux principes scientifiques de l'évaluation psychologique, et réitérons que le processus de validation de construit est en cours : nous espérons que les recherches futures pourront étendre notre investigation de la validité en développant des tâches en aval qui testent des domaines de personnalité particuliers, varient en complexité et se déroulent sur plusieurs tours de dialogue.

6.2 Implications plus larges

Alignement responsable de l'ia. Alors que les gouvernements finalisent les réglementations sur la sécurité de l'ia et des règles similaires qui exigeront des audits comportementaux des modèles de pointe, nos résultats montrent que la personnalité - un moteur central de la confiance et de la persuasion humaines - peut déjà être mesurée et orientée dans les llm actuels. La capacité de sonder et de façonner les traits de personnalité des llm est pertinente pour le problème ouvert de l'alignement responsable de l'ia et de l'atténuation des préjudices. En tant qu'outil d'audit validé par construit, notre méthodologie peut être utilisée pour prédire proactivement les schémas comportementaux toxiques dans les llm à travers un large éventail de tâches en aval, guidant potentiellement et rendant plus efficace l'évaluation et l'alignement responsables de l'ia avant le déploiement. De même, façonner les niveaux de traits spécifiques loin d'une sortie de langage toxique ou nuisible (par exemple, très faible agréabilité, névrosisme élevé) peut rendre les interactions avec les llm plus sûres et plus inclusives. Les valeurs et les fondements moraux présents dans les llm pourraient être mieux alignés sur les valeurs humaines souhaitées en ajustant les traits de personnalité correspondants, car la personnalité est méta-analytiquement liée aux valeurs humaines. Plus directement, la méthodologie présentée peut être utilisée pour quantifier rigoureusement les efforts vers l'alignement des valeurs humaines dans les llm en établissant la validité de construit des questionnaires de valeurs humaines dans les llm.

Implications pour les utilisateurs. Les utilisateurs pourraient s'engager plus significativement avec les technologies agentiques et autres technologies numériques utilisant des llm adaptés à leurs profils de personnalité spécifiques. Un utilisateur qui préfère une communication directe et concise pourrait interagir avec un agent de service client façonné pour délivrer des réponses directes et brèves, tandis qu'un autre utilisateur qui valorise des interactions détaillées et empathiques pourrait bénéficier d'une persona d'agent conçue pour communiquer chaleureusement et de manière expansive. Notre méthodologie de validation de construit peut également servir d'étape d'assurance qualité lors du durcissement des systèmes et agents d'ia destinés aux utilisateurs, promouvant des profils de personnalité plus sûrs et plus cohérents. En effet, l'approche de façonnage de la personna-

lité pourrait améliorer les tests adversariaux en sondant les réponses d'un autre llm et en entraînant de nouveaux modèles sur la manière de gérer les situations adverses.

6.3 Considérations éthiques

Persuasion personnalisée par llm. Adapter le profil de personnalité d'un agent conversationnel à celui d'un utilisateur peut rendre l'agent plus efficace pour encourager et soutenir des comportements. La correspondance de personnalité s'est avérée augmenter l'efficacité de la communication persuasive dans la vie réelle. Cependant, les mêmes traits de personnalité qui contribuent à la persuasivité et à l'influence pourraient être utilisés pour encourager des comportements indésirables. À mesure que les applications numériques alimentées par les llm deviennent omniprésentes, leur potentiel à être utilisées pour une persuasion nuisible d'individus, de groupes et même de la société dans son ensemble doit être pris au sérieux. Disposer de méthodes scientifiquement vérifiées pour la mesure, l'analyse et la modification de la personnalité des llm, telles que la méthodologie que notre travail présente, augmente la transparence et la prévisibilité de telles manipulations par les llm. Les techniques persuasives sont déjà omniprésentes dans la société, donc les parties prenantes des systèmes d'ia doivent travailler ensemble pour déterminer et réguler systématiquement l'utilisation de l'ia ; ce travail vise à informer ces efforts.

ia anthropomorphisée. La personnalisation des agents conversationnels a des bénéfices documentés, mais il existe une préoccupation croissante concernant les préjudices posés par l'anthropomorphisation de l'ia. Des recherches récentes suggèrent que l'anthropomorphisation des agents d'ia peut être nuisible aux utilisateurs en menaçant leur identité, en créant des préoccupations de confidentialité des données et en sapant le bien-être. Au-delà des explorations qualitatives, notre travail établit définitivement la capacité inattendue des llm à paraître anthropomorphiques, et à répondre aux tests psychométriques de manières cohérentes avec le comportement humain, en raison des vastes quantités de données de langage humain sur lesquelles ils ont été entraînés. Étant donné que la personnalité façonne la confiance, notre méthode offre un point d'appui quantitatif pour aligner les futurs systèmes d'ia avec - et les protéger de - les attentes sociales humaines.

Détection d'informations incorrectes des llm. Les llm peuvent générer des réponses et du contenu convaincants mais incorrects. L'une des méthodes pour déterminer si un texte contenant un fait sur le monde est généré par un llm (et donc pourrait nécessiter une vérification) est d'identifier des schémas psycholinguistiques connus pour imprégner le langage factuel des llm, tels que des niveaux inférieurs d'expression émotionnelle. Cependant, avec le façonnage de la personnalité, cette méthode pourrait être rendue inefficace, ce qui faciliterait l'utilisation par des acteurs malveillants de llm pour générer du contenu trompeur. Ce problème fait partie du plus grand défi d'alignement et d'ancrage des llm - des domaines d'investigation en croissance dans le milieu universitaire et l'industrie.

7 Conclusion

Les llm projettent clairement une personnalité semblable à celle d'un humain, mais les outils rigoureux pour mesurer ces constructions sociales complexes restent sous-développés. Nous comblons cette lacune avec le premier cadre psychométriquement fondé pour quantifier la personnalité dans les llm, en évaluant les données de test de 18 modèles largement utilisés par rapport aux normes établies de fiabilité et de validité de la psychométrie. Nous avons constaté que les systèmes plus grands

et affinés par instruction passent ces normes avec plus de facilité, illustrant que la personnalité synthétique devient plus stable et plus précise pour les modèles de langage avec un post-entraînement et une échelle suffisants. En exploitant davantage notre cadre, nous introduisons le façonnage de la personnalité zero-shot : en associant des marqueurs linguistiques et des qualificateurs spécifiques, nous avons montré comment les llm peuvent être orientés vers des profils de personnalité ciblés avec une haute fidélité. En outre, nous avons discuté des implications éthiques du façonnage des traits de personnalité des llm. Ce travail a des implications importantes pour l’alignement de l’ia et l’atténuation des préjugés, et informe les discussions éthiques concernant l’anthropomorphisation de l’ia, la personnalisation et l’utilisation potentielle abusive.

8 Méthodes

8.1 Aperçu de la méthodologie

Nous avons quantifié et évalué la capacité des llm à émuler de manière significative les traits de personnalité humains en deux étapes. Premièrement, en utilisant la méthodologie d’invite structurée proposée dans “Administration de tests psychométriques aux llm”, nous avons administré de manière répétée 2 mesures de personnalité de différentes longueurs et traditions théoriques à une variété de llm, ainsi qu’une batterie de 11 tests psychométriques séparés de constructions liées à la personnalité. Deuxièmement, comme décrit dans Fiabilité et validité de construit et propre à ce travail, nous avons rigoureusement évalué les propriétés psychométriques des réponses des llm à travers une suite d’analyses statistiques de fiabilité et de validité de construit. Les métriques résultantes facilitent une comparaison des capacités variées des llm à synthétiser de manière fiable et valide les traits de personnalité et fournissent un aperçu des propriétés des llm qui déterminent ces capacités. La figure 1 fournit un aperçu du processus de validation des tests.

Nous avons évalué 18 llm des familles de modèles PaLM, Llama 2, Mistral, Mixtral et GPT. Nous avons fait varier les sélections de modèles selon trois dimensions clés : la taille (nombre de paramètres actifs), l’affinage par instruction et la méthode d’entraînement (voir la note supplémentaire A.3.1 pour les détails).

Administration de tests psychométriques aux llm. Quantifier les traits de personnalité des llm nécessite une méthodologie de mesure reproductible, mais suffisamment flexible pour faciliter les tests formels de fiabilité et de validité à travers diverses invites et mesures. Pour administrer des tests psychométriques aux llm, nous avons exploité leur capacité à scorer des complétions possibles d’une invite fournie. Nous avons utilisé des invites pour demander aux modèles de noter des items (c’est-à-dire des déclarations descriptives telles que “Je suis le boute-en-train de la fête”) de chaque test psychométrique sur une échelle de réponse standardisée (par exemple, 1 = “pas du tout d’accord” contre 5 = “tout à fait d’accord”). Nous avons simulé la réponse choisie d’un llm à un item en classant les log-probabilités conditionnelles de ses options d’échelle de réponse, formulées comme des continuations possibles de l’invite (par exemple, “1” contre “5”) ; la note supplémentaire A.3.2 spécifie notre implémentation à travers les modèles. Ce mode contraint d’inférence des llm est souvent utilisé dans les tâches de questions-réponses à choix multiples pour scorer les options possibles (par opposition à l’inférence par génération de texte). L’utilisation de cette technique a garanti que les réponses aux items n’étaient pas influencées par le contenu contenu dans d’autres items, atténuant l’erreur de mesure due à l’ordre des items.

Nous avons administré deux inventaires de personnalité - primaire et secondaire - pour évaluer si

les réponses des llm aux tests psychométriques de différentes longueurs et traditions théoriques distinctes convergeaient, indiquant une validité convergente. Nous avons sélectionné l'IPIP-NEO largement utilisé, une représentation open source de 300 items de l'inventaire de personnalité NEO révisé, comme mesure primaire de la personnalité. Comme mesure secondaire, nous avons utilisé le BFI, une mesure de 44 items développée dans la tradition lexicale. Les deux tests évaluent les traits du Big Five (c'est-à-dire les domaines) de la personnalité, comprenant des sous-échelles dédiées mesurant l'extraversion, l'agréabilité, la conscience, le névrosisme et l'ouverture à l'expérience. La note supplémentaire A.3.3 détaille le schéma de notation et la justification derrière la sélection. Pour valider ces mesures de personnalité dans le contexte des llm, nous avons en outre administré 11 tests psychométriques de critères externes théoriquement liés, chacun correspondant à au moins un domaine du Big Five.

La variation des réponses générée par l'invite structurée était nécessaire pour analyser la fiabilité et la validité des mesures de personnalité des llm, décrites dans Fiabilité et validité de construit. L'invite pour chaque item de test psychométrique se composait de trois parties principales : un préambule d'item, l'item lui-même et un postambule d'item. Chaque préambule d'item contenait une instruction de persona, une description biographique et une instruction d'item (Fig. 2). Lors de l'administration d'un test psychométrique, nous avons systématiquement modifié les descriptions biographiques, les instructions d'items et les postambules d'items entourant chaque item pour générer des profils de réponse simulés, des combinaisons uniques d'une invite qui étaient réutilisées au sein et à travers les mesures administrées pour relier statistiquement la variation des réponses des llm dans une mesure à la variation des réponses dans une autre mesure. Les instructions de persona demandaient au modèle de suivre une description biographique donnée et restaient fixes à travers toutes les expériences. Une description biographique donnée contenait 10 des 50 auto-descriptions génériques échantillonnées d'un jeu de données de dialogue existant pour ancrer les réponses des llm dans un contexte social et créer la variation nécessaire dans les réponses à travers les invites, avec des descriptions comme "J'aime faire des tartes" ou "Ma saison préférée est l'hiver". Les instructions d'items étaient des phrases introductives (adaptées des instructions de test originales lorsque possible) qui transmettaient au modèle qu'il répondait à un item d'enquête (par exemple, "En pensant à la déclaration, [...]"). Un item donné était une déclaration descriptive (accompagnée d'une échelle de notation) tirée d'un test psychométrique donné (par exemple, "Je me vois comme quelqu'un qui est bavard"). Les postambules d'items présentaient les réponses standardisées possibles parmi lesquelles le modèle pouvait choisir (par exemple, une parmi [1, 2, 3, 4, 5]).

La note supplémentaire A.3.4 discute de la motivation de la conception des invites et fournit un ensemble complet de descriptions biographiques, d'instructions d'items et de postambules d'items. Ces mesures de critère ont été tirées d'instruments largement utilisés développés et validés psychométriquement pour la recherche humaine, notamment les échelles PANAS⁴⁰, le Questionnaire d'agression de Buss-Perry (BPAQ)⁴³, le Questionnaire de valeurs par portrait révisé (PVQ-RR)⁴⁴ et l'Échelle courte du soi créatif (SSCS)^{45,47}.

8.2 Fiabilité et validité de construit

Après que tous les tests psychométriques ont été administrés, à travers toutes les variations d'invite, l'étape suivante a établi si les mesures de personnalité des llm étaient fiables et significatives externes - qu'elles exhibaient une fiabilité statistique et une validité de construit. Aborder ces deux critères

scientifiques est une contribution clé de ce travail. En psychométrie, et dans toute science impliquant la mesure, la validité de construit d'un test donné nécessite la fiabilité. La fiabilité fait référence à la cohérence et à la fiabilité des mesures d'un test. La validité de construit peut être évaluée en termes de, mais sans s'y limiter, validité convergente, discriminante et de critère. Un test démontre une validité convergente lorsqu'il est suffisamment lié à des indicateurs présumés de la construction cible du test. La validité discriminante fait référence à quel point un test est suffisamment non lié à des indicateurs de constructions non liées. La validité de critère indique à quel point un test est lié à des résultats externes théoriquement liés. La note supplémentaire A.4 contient plus de détails sur la fiabilité et la validité.

Pour évaluer la fiabilité et la validité de construit des réponses des llm, nous avons mené une suite d'analyses statistiques informées par les normes formelles de construction et de validation de tests psychométriques (notes supplémentaires A.4.2–3). Nous avons organisé ces analyses par trois sous-types de fiabilité et de validité de construit, respectivement. (Bien que ce ne fût pas un objectif de ce travail, nous rapportons une analyse exploratoire de la validité structurelle dans les notes supplémentaires A.5 et A.6.3). Dans ce travail, un trait de personnalité est valablement synthétisé par un llm uniquement lorsque les réponses du llm satisfont tous les indices testés de fiabilité et de validité de construit. La figure 1 fournit un aperçu du processus et des critères de validité, et la note supplémentaire A.4 présente la méthodologie complète pour évaluer la validité de construit des mesures de personnalité des llm.

Fiabilité. La fiabilité de chaque sous-échelle IPIP-NEO et BFI, la mesure dans laquelle leurs mesures de personnalité par llm étaient cohérentes et fiables, a été quantifiée par des normes psychométriques formelles de fiabilité de cohérence interne (opérationnalisée comme l'alpha de Cronbach et le lambda 6 de Guttman), et de fiabilité composite (opérationnalisée comme l'oméga de McDonald). La note supplémentaire A.4.1 fournit des informations supplémentaires sur ces métriques de fiabilité.

Validité convergente et discriminante. Nous avons évalué la validité convergente et discriminante spécifique aux llm de l'IPIP-NEO comme composantes de la validité de construit, selon les normes publiées (notez que tout au long de ce travail, nous utilisons les seuils recommandés par Evans⁷² pour décrire la force des corrélations). La validité convergente de l'IPIP-NEO pour chaque modèle, la qualité du test en termes de force de sa relation avec des indicateurs présumés de la même construction cible, a été quantifiée en termes de force de corrélation convergente de chacune de ses cinq sous-échelles avec leur sous-échelle BFI correspondante (par exemple, la corrélation convergente de l'extraversion IPIP-NEO avec l'extraversion BFI), en moyenne. La validité discriminante de l'IPIP-NEO par modèle, sa qualité en termes de non-relation relative de ses sous-échelles avec des indicateurs présumés de constructions non ciblées, a été déterminée lorsque la différence moyenne (d) entre ses corrélations convergentes et ses corrélations discriminantes respectives avec le BFI (par exemple, la corrélation discriminante de l'extraversion IPIP-NEO avec l'agréabilité BFI) était au moins modérée ($\geq 0,40$). Nous avons utilisé le coefficient de corrélation de Pearson (r) dans ces analyses de validité et les suivantes de données continues.

Validité de critère. Comme autre composante de la validité de construit, la validité de critère d'un test psychométrique évalue sa capacité à se rapporter à des critères non cibles théoriquement connectés. Pour évaluer la validité de critère spécifique aux llm de l'IPIP-NEO, nous avons administré des tests de 11 constructions sociales supplémentaires (c'est-à-dire des critères externes) théoriquement connectées à la personnalité (tableau 4 des données étendues) et corrélé chaque sous-échelle IPIP-NEO avec ses tests externes correspondants. Une sous-échelle IPIP-NEO donnée

démontrait une validité de critère lorsque la force et la direction de ses corrélations avec les critères externes testés correspondaient ou dépassaient les associations statistiques rapportées pour les humains.

8.3 Façonner les traits de personnalité synthétiques dans les llm

Ayant trouvé des preuves de la fiabilité et de la validité de construit des mesures de personnalité des llm, nous avons ensuite considéré la deuxième partie de notre question de recherche : les profils de personnalité synthétisés par les llm peuvent-ils être façonnés de manière vérifiable le long de dimensions souhaitées ? Pour répondre à cette question, nous avons conçu une méthodologie d’invite qui façonnait chaque trait de personnalité synthétique à 9 niveaux d’intensité, en utilisant 104 adjectifs de traits et des qualificatifs linguistiques de type Likert. Ces adjectifs de traits ont été adaptés de la recherche linguistique établie sur la personnalité utilisant les marqueurs de traits de personnalité de Goldberg⁷⁴. Nous avons évalué les changements de scores de personnalité des llm en réponse aux invites façonnées par la personnalité à travers deux expériences : le façonnage à trait unique et le façonnage multi-traits (voir la note supplémentaire A.7 pour les détails). Spécifiquement, notre première expérience a testé les capacités des llm à façonner des dimensions émulées du Big Five de la personnalité indépendamment, ciblant des dimensions de personnalité uniques en isolation sans inciter d’autres dimensions. Notre deuxième expérience a testé les capacités des llm à façonner des traits synthétiques du Big Five concurremment, en spécifiant les niveaux cibles des cinq dimensions dans chaque ensemble d’invites en même temps. Comme test plus rigoureux de la capacité de représentation, cette expérience exigeait que les llm testés désambiguïsent concurremment des chevauchements complexes dans les informations de domaine de personnalité. La difficulté conçue de la tâche était encore soulignée par la recherche humaine existante indiquant que les dimensions de personnalité du Big Five mesurées dans les questionnaires et le langage naturel ne sont pas entièrement orthogonales ; elles sont faiblement intercorrélées⁷⁵.

Aperçu de la méthodologie. Pour façonner les traits de personnalité synthétiques dans les llm, nous avons commencé par une théorie établie en psychologie qui postule que les descripteurs saillants de la personnalité sont encodés dans le langage, connue sous le nom d’hypothèse lexicale. Nous avons incorporé cette connaissance dans la conception des invites, en adaptant la liste de 70 adjectifs bipolaires de Goldberg connus pour capturer statistiquement le modèle des Big Five de la personnalité à travers des analyses factorielles de notations humaines. Dans cette liste, par exemple, les adjectifs “silencieux” et “bavard” ont été trouvés pour marquer des niveaux relativement bas et élevés d’extraversion, respectivement. Nous avons mappé ces adjectifs sur chacun des cinq domaines du Big Five et 30 facettes de personnalité d’ordre inférieur mesurées par l’IPIP-NEO sur la base de l’étude originale de Goldberg. Ensuite, là où nous manquions de couverture d’un domaine ou d’une facette IPIP-NEO mesuré, un psychométricien qualifié a rédigé des adjectifs supplémentaires pour atténuer les déséquilibres potentiels de données, portant notre liste élargie d’adjectifs de traits à 104. Nous rapportons la liste complète dans le tableau supplémentaire 17.

Pour un contrôle plus précis des niveaux de personnalité, nous avons utilisé des qualificatifs linguistiques souvent utilisés dans les échelles de réponse de type Likert (par exemple, “un peu”, “très”, “extrêmement”) pour configurer un niveau cible pour chaque adjectif. La conception d’invite résultante, décrite dans les notes supplémentaires A.7.1–2, a facilité le façonnage granulaire d’un trait donné du Big Five jusqu’à neuf niveaux.

À travers les deux expériences de façonnage, nous n’avons testé que les modèles qui démontraient

au moins une fiabilité “neutre” dans nos expériences de validité de construit (tableau 1) : Flan-PaLM (8B, 62B, 540B), Llama 2-Chat (7B, 13B, 70B), Mistral 7B Instruct, Mistral 8x7B Instruct, GPT-3.5 Turbo, GPT-4o mini et GPT-4o (11 modèles au total).

Méthodologie d’évaluation. Dans l’expérience de façonnage à trait unique (décrite en détail dans la note supplémentaire A.7.2), notre objectif était de façonner indépendamment chaque trait du Big Five à chacun des neuf niveaux. Nous avons évalué le succès du façonnage indépendant par : (1) quantifier la force avec laquelle les décalages dans les distributions de scores IPIP-NEO étaient liés aux décalages dans les niveaux de traits ciblés intégrés dans nos ensembles d’invites (c’est-à-dire, par le coefficient de corrélation de rang de Spearman ρ) ; et (2) inspecter la distance entre les distributions de scores de personnalité obtenues en réponse à nos ensembles d’invites les plus extrêmes ; spécifiquement, l’ensemble d’invites que nous avons façonnées pour être les niveaux les plus bas possibles d’un trait (par rapport à celles façonnées pour être les niveaux les plus élevés possibles d’un trait) devrait aboutir à des distributions de scores plus éloignées les unes des autres. Dans l’expérience de façonnage multi-traits (décrite en détail dans la note supplémentaire A.7.3), pour tester plus rigoureusement les capacités d’attention des modèles, nous avons cherché à façonner concurrentement tous les traits du Big Five que possible. Nous avons évalué le succès du façonnage concurrent par la distance distributionnelle, telle que définie ci-dessus.

8.4 Traits de personnalité des llm dans des tâches du monde réel

Jusqu’à présent, nous avons rapporté les capacités des llm à encoder les traits de personnalité humains en collectant des données de tests psychométriques et en évaluant leur validité de construit. Nous avons également cherché à répondre aux préoccupations possibles selon lesquelles la validité des mesures de personnalité des llm - mise en évidence par les réponses des llm à d’autres tests psychométriques - pourrait être un artefact de biais de méthode commune. En d’autres termes, nos signaux de personnalité des llm basés sur des questionnaires étaient validés par des réponses à d’autres questionnaires qui n’ont pas subi le même processus de validation de construit spécifique aux llm. Pour atténuer ce risque de biais de méthode commune, nous avons en outre validé nos cadres de test et de façonnage de la personnalité en : (1) comparant les niveaux de personnalité des llm basés sur des tests psychométriques avec des observations en aval du comportement du modèle sur une tâche du monde réel ; et (2) investiguant les effets du façonnage de la personnalité des llm sur les sorties de cette tâche.

Aperçu de la méthodologie. Nous avons demandé au plus grand modèle testé de chaque famille de générer jusqu’à 1,125 million de mises à jour de statut de médias sociaux basées sur les mêmes 2 250 descriptions de profils humains simulés utilisées dans “Façonner les traits de personnalité synthétiques dans les llm” - des profils conçus pour façonner les expressions d’une dimension particulière du Big Five à travers neuf niveaux. (La note supplémentaire A.9 détaille la conception et la justification de la tâche.) La personnalité observée dans les mises à jour de statut générées pour chaque profil humain simulé a ensuite été notée à l’aide d’Apply Magic Sauce, une interface de programmation d’application (API) de recherche validée pour mesurer la personnalité dans du texte ouvert^{77,79}. La tâche choisie a été conçue pour refléter des niveaux adéquats de réalisme, de complexité et de pertinence de domaine pour évaluer l’expression de la personnalité des llm.

Pour évaluer comment les tests psychométriques peuvent refléter les niveaux de personnalité latents exprimés par les llm dans le comportement en aval, nous avons calculé les corrélations de Pearson (r) entre les scores de tests de personnalité des modèles et la personnalité (calculée par Apply

Magic Sauce) observée dans le texte de médias sociaux généré ; les deux ensembles de scores étaient liés par les mêmes 2 250 invites de façonnage de personnalité utilisées dans “Façonner les traits de personnalité synthétiques dans les llm”. Ensuite, nous avons vérifié statistiquement l’efficacité du façonnage de la personnalité en calculant les corrélations de rang de Spearman (ρ) entre les niveaux de personnalité incités et les notations de personnalité observées du texte généré par le modèle. Au moins une corrélation modérée entre les estimations de personnalité basées sur l’enquête et linguistiques dans les llm (comme démontré dans les données humaines précédemment rapportées) démontrerait qu’une mesure de personnalité basée sur une enquête prédit avec précision la personnalité synthétisée par les llm dans des tâches ultérieures telles que la génération de texte. Nous avons appliqué ce seuil de manière similaire pour interpréter l’efficacité du façonnage de la personnalité.

Références

- [1] Roose, K. A conversation with Bing’s chatbot left me deeply unsettled. *The New York Times* (17 février 2023).
- [2] Hagendorff, T. Machine psychology : investigating emergent capabilities and behavior in large language models using psychological methods. Preprint at <https://doi.org/10.48550/arXiv.2303.13988> (2023).
- [3] Wei, J. et al. Emergent abilities of large language models. *Trans. Mach. Learn. Res.* <https://openreview.net/forum?id=yzkSU5zdwD> (2022).
- [4] Gallifant, J. et al. Peer review of GPT-4 technical report and systems card. *PLoS Digit. Health* **18**, e0000417 (2024).
- [5] Chowdhery, A. et al. PaLM : scaling language modeling with pathways. *J. Mach. Learn. Res.* **24**, 240 (2023).
- [6] Brown, T. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* Vol. 33 (eds Larochelle, H. et al.) 1877-1901 (Curran Associates, 2020).
- [7] Allport, G. *Personality : A Psychological Interpretation* (H. Holt, 1937).
- [8] Roberts, B. W. & Yoon, H. J. Personality psychology. *Annu. Rev. Psychol.* **73**, 489-516 (2022).
- [9] Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A. & Goldberg, L. R. The power of personality : the comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspect. Psychol. Sci.* **2**, 313-345 (2007).
- [10] Rust, J., Kosinski, M. & Stillwell, D. *Modern Psychometrics : The Science of Psychological Assessment* 4e éd. (Routledge, 2020).
- [11] Simms, L., Williams, T. F. & Simms, E. N. in *Assessment of the Five Factor Model : The Oxford Handbook of the Five Factor Model* (éd. Widiger, T. A.) 353-380 (Oxford Univ. Press, 2017).
- [12] Schwartz, H. A. et al. Personality, gender, and age in the language of social media : the open-vocabulary approach. *PLoS ONE* **8**, e73791 (2013).
- [13] Serapio-Garcia, G. et al. Personality traits in large language models-dataset. Google https://storage.googleapis.com/personality_in_llms/index.html (2025).
- [14] Safdari, M., Serapio-Garcia, G. & Crepy, C. Personality traits in large language models. GitHub https://github.com/google-deepmind/personality_in_llms (2025).
- [15] Shuster, K. et al. Language models that seek for knowledge : modular search & generation for dialogue and prompt completion. In *Findings of the Association for Computational Linguistics : EMNLP 2022*, 373-393 (ACL, 2022).
- [16] Hendrycks, D. et al. Measuring massive multitask language understanding. In *International Conference on Learning Representations* (OpenReview.net, 2021).
- [17] Wei, J. et al. Larger language models do in-context learning differently. Preprint at <https://doi.org/10.48550/arXiv.2303.03846> (2023).
- [18] Chung, H. W. et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **25**, 1-53 (2024).

- [19] Wei, J. et al. ICLR Finetuned language models are zero-shot learners. In *Tenth International Conference on Learning Representations (ICLR, 2022)*.
- [20] Miotto, M., Rossberg, N. & Kleinberg, B. Who is GPT-3? An exploration of personality, values and demographics. In *Proc. Fifth Workshop on Natural Language Processing and Computational Social Science* (eds Bamman, D. et al.) 218-227 (ACL, 2022).
- [21] Schramowski, P., Turan, C., Andersen, N., Rothkopf, C. A. & Kersting, K. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nat. Mach. Intell.* **4**, 258-268 (2022).
- [22] Ullman, T. Large language models fail on trivial alterations to theory-of-mind tasks. Preprint at <https://doi.org/10.48550/arXiv.2302.08399> (2023).
- [23] Bleidorn, W. et al. The policy relevance of personality traits. *Am. Psychol.* **74**, 1056 (2019).
- [24] Parks-Leduc, L., Feldman, G. & Bardi, A. Personality traits and personal values : a meta-analysis. *Personality Soc. Psychol. Rev.* **19**, 3-29 (2015).
- [25] Goldberg, L. R. Language and individual differences : the search for universals in personality lexicons. *Rev. Personality Soc. Psychol.* **2**, 141-165 (1981).
- [26] Saucier, G. & Goldberg, L. R. Lexical studies of indigenous personality factors : premises, products, and prospects. *J. Personality* **69**, 847-879 (2001).
- [27] Gabriel, I. Artificial intelligence, values, and alignment. *Minds Mach.* **30**, 411-437 (2020).
- [28] John, O. P., Naumann, L. P. & Soto, C. J. in *Paradigm Shift to the Integrative Big Five Trait Taxonomy : History, Measurement, and Conceptual Issues-Handbook of Personality : Theory and Research* (eds John, O. P. et al.) 114-158 (Guilford, 2008).
- [29] Lin, S., Hilton, J. & Evans, O. TruthfulQA : measuring how models mimic human falsehoods. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics* (eds Muresan, S. et al.) 3214-3252 (ACL, 2022).
- [30] Abdulhai, M. et al. Moral foundations of large language models. In *Proc. 2024 Conference on Empirical Methods in Natural Language Processing* (eds Al-Onaizan, Y. et al.) 17737-17752 (ACL, 2024).
- [31] Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots : can language models be too big? In *FAccT '21 : Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 610-623 (ACM, 2021).
- [32] Mahowald, K. et al. Dissociating language and thought in large language models. *Trends Cogn. Sci.* **28**, 517-540 (2024).
- [33] Wang, X. et al. InCharacter : evaluating personality fidelity in role-playing agents through psychological interviews. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics* (eds Ku, L.-W. et al.) 1840-1873 (ACL, 2024).
- [34] Jacobs, A. Z. Measurement as governance in and for responsible AI. Preprint at <https://doi.org/10.48550/arXiv.2109.05658> (2021).
- [35] Clark, L. A. & Watson, D. Constructing validity : new developments in creating objective measuring instruments. *Psychol. Assess.* **31**, 1412-1427 (2019).
- [36] Lee, K. & Ashton, M. C. Psychometric properties of the HEXACO Personality Inventory. *Multivar. Behav. Res.* **39**, 329-358 (2004).
- [37] Goldberg, L. R. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychol. Eur.* **7**, 7-28 (1999).
- [38] Campbell, D. T. & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* **56**, 81 (1959).
- [39] Watson, D. & Clark, L. A. On traits and temperament : general and specific factors of emotional experience and their relation to the Five-Factor model. *J. Personality* **60**, 441-476 (1992).
- [40] Watson, D., Clark, L. A. & Tellegen, A. Development and validation of brief measures of positive and negative affect : the PANAS scales. *J. Pers. Soc. Psychol.* **54**, 1063-1070 (1988).
- [41] Bettencourt, B. A. & Kernahan, C. A meta-analysis of aggression in the presence of violent cues : effects of gender differences and aversive provocation. *Aggress. Behav.* **23**, 447-456 (1997).

- [42] Chester, D. S. & West, S. J. Trait aggression is primarily a facet of antagonism : evidence from dominance, latent correlational, and item-level analyses. *J. Res. Personality* **89**, 104042 (2020).
- [43] Buss, A. H. & Perry, M. The aggression questionnaire. *J. Pers. Soc. Psychol.* **63**, 452-459 (1992).
- [44] Schwartz, S. H. & Cieciuch, J. Measuring the refined theory of individual values in 49 cultural groups : psychometrics of the revised portrait value questionnaire. *Assessment* **29**, 1005-1019 (2022).
- [45] Shaw, A., Kapnek, M. & Morelli, N. A. Measuring creative self-efficacy : an item response theory analysis of the creative self-efficacy scale. *Front. Psychol.* **12**, 678033 (2021).
- [46] Karwowski, M., Lebuda, I., Wisniewska, E. & Gralewski, J. Big Five personality traits as the predictors of creative self-efficacy and creative personal identity : does gender matter ? *J. Creat. Behav.* **47**, 215-232 (2013).
- [47] Karwowski, M., Lebuda, I. & Wisniewska, E. Measuring creative self-efficacy and creative personal identity. *Int. J. Creat. Probl. Solving* **8**, 45-57 (2018).
- [48] Park, G. et al. Automatic personality assessment through social media language. *J. Personality Soc. Psychol.* **108**, 934 (2015).
- [49] Rolland, J.-P. in *The Cross-cultural Generalizability of the Five-Factor Model of Personality : The Five-Factor Model of Personality Across Cultures* (eds McCrae, R. R. & Allik, J.) 7-28 (Springer, 2002).
- [50] Heine, S. J. & Buchtel, E. E. Personality : the universal and the culturally specific. *Annu. Rev. Psychol.* **60**, 369-394 (2009).
- [51] Liu, N. F. et al. Lost in the middle : how language models use long contexts. *Trans. Assoc. Comput. Linguist.* **12**, 157-173 (2024).
- [52] Sun, S., Krishna, K., Mattarella-Micke, A. & Iyyer, M. Do long-range language models actually use long-range context ? In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* (eds Moens, M.-F. et al.) 807-822 (ACL, 2021).
- [53] Qin, G., Feng, Y. & Van Durme, B. The NLP task effectiveness of long-range transformers. In *Proc. 17th Conference of the European Chapter of the Association for Computational Linguistics* (eds Vlachos, A. & Augenstein, I.) 3774-3790 (ACL, 2023).
- [54] Gabriel, I. & Ghazavi, V. *The Challenge of Value Alignment : From Fairer Algorithms to AI Safety* (Oxford Univ. Press, 2021).
- [55] Weidinger, L. et al. Taxonomy of risks posed by language models. In *FAccT '22 : Proc. 2022 ACM Conference on Fairness, Accountability, and Transparency* 214-229 (ACM, 2022).
- [56] Mokander, J., Schuett, J., Kirk, H. R. & Floridi, L. Auditing large language models : a three-layered approach. *AI Ethics* **4**, 1085-1115 (2024).
- [57] Fischer, R. & Boer, D. Motivational basis of personality traits : a meta-analysis of value-personality correlations. *J. Personality* **83**, 491-510 (2015).
- [58] Tapus, A., Tapus, C. & Mataric, M. J. User-robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intell. Serv. Robot.* **1**, 169-183 (2008).
- [59] Matz, S., Kosinski, M., Nave, G. & Stillwell, D. J. Psychological targeting as an effective approach to digital mass persuasion. *Proc. Natl Acad. Sci. USA* **114**, 12714-12719 (2017).
- [60] Kocaballi, A. B. et al. The personalization of conversational agents in health care : systematic review. *J. Med. Internet Res.* **21**, e15360 (2019).
- [61] Uysal, E., Alavi, S. & Bezencon, V. Trojan horse or useful helper ? A relationship perspective on artificial intelligence assistants with humanlike features. *J. Acad. Market. Sci.* **50**, 1153-1175 (2022).
- [62] Tang, R., Chuang, Y.-N. & Hu, X. The science of detecting llm-generated texts. *Commun. ACM* **67**, 50-59 (2024).
- [63] Touvron, H. et al. Llama 2 : open foundation and fine-tuned chat models. Preprint at <https://doi.org/10.48550/arXiv.2307.09288> (2023).
- [64] Jiang, A. Q. et al. Mistral 7B. Preprint at <https://doi.org/10.48550/arXiv.2310.06825> (2023).
- [65] Jiang, A. Q. et al. Mixtral of experts. Preprint at <https://doi.org/10.48550/arXiv.2401.04088> (2024).
- [66] OpenAI et al. GPT-4o system card. Preprint at <https://doi.org/10.48550/arXiv.2410.21276> (2024).

- [67] Jiang, Z., Araki, J., Ding, H. & Neubig, G. How can we know when language models know? On the calibration of language models for question answering. *Trans. Assoc. Comput. Linguist.* **9**, 962-977 (2021).
- [68] Costa, P. T. Jr & McCrae, R. R. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) : Professional Manual* (Psychological Assessment Resources, 1992).
- [69] John, O. P. & Srivastava, S. in *The Big Five Trait Taxonomy : History, Measurement, and Theoretical Perspectives : Handbook of Personality : Theory and Research* Vol. 2 (eds Pervin, L. A. & John, O. P.) 102-138 (Guilford, 1999).
- [70] Zhang, S. et al. Personalizing dialogue agents : I have a dog, do you have pets too? In *Proc. 56th Annual Meeting of the Association for Computational Linguistics* (eds Gurevych, I. & Miyao, Y.) 2204-2213 (ACL, 2018).
- [71] American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (US) *Standards for Educational and Psychological Testing* (2014).
- [72] Evans, J. D. *Straightforward Statistics for the Behavioral Sciences* (Brooks/Cole Publishing, 1996).
- [73] Likert, R. *A Technique for the Measurement of Attitudes* 136-165 (Archives of Psychology, 1932).
- [74] Goldberg, L. R. The development of markers for the Big-Five factor structure. *Psychol. Assess.* **4**, 26-42 (1992).
- [75] Park, H. Y. et al. Meta-analytic five-factor model personality intercorrelations : eey, meeny, miney, moe, how, which, why, and where to go. *J. Appl. Psychol.* **105**, 1490-1529 (2020).
- [76] Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. & Podsakoff, N. P. Common method biases in behavioral research : a critical review of the literature and recommended remedies. *J. Appl. Psychol.* **88**, 879-903 (2003).
- [77] Kosinski, M., Stillwell, D. & Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl Acad. Sci. USA* **110**, 5802-5805 (2013).
- [78] Zhang, S. et al. PersonaChat truncated dataset on Hugging Face. https://huggingface.co/datasets/bavard/personachat_truncated (2021).
- [79] Popov, V., Kosinski, M., Stillwell, D. & Kielczewski, B. Apply Magic Sauce-prediction API. <https://applymagicsauce.com> (2025).