JOURNAL INTERNATIONAL DE TECHNOLOGIE, DE L'INNOVATION,
DE LA PHYSIQUE, DE L'ENERGIE ET DE L'ENVIRONNEMENT

Le problème difficile de l'identité : évaluation d'un clone LLM

L. E. Brunet

ISSN : 2428-8500

DOI : 10.52497/jitipee.v9i2.381

Journal International de Technologie, de l'Innovation, de la Physique, de l'Energie et de l'Environnement
ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

Le problème difficile de l'identité : évaluation d'un clone LLM

Luc E. Brunet[(1)]

[(1)] R&D Mediation

luc.brunet@insead.edu

Résumé *– Dans le cadre de travaux explorant la faisabilité du clonage de personnalité, nous présentons la conception et l'évaluation d'un agent conversationnel baptisé pseudoLuc, destiné à reproduire la manière de s'exprimer, les positions et le style d'un individu réel (Luc). Après avoir recueilli un corpus de textes personnels et construit un manuel de personnalité (prompt contextuel) structurant les 23 dimensions jugées essentielles, nous avons intégré ces informations à un modèle de langage (LLM) agrémenté d'une composante de « Récupération-Génération Augmentée » (Retrieval-Augmented Generation : RAG) pour améliorer la factualité. Les résultats montrent qu'un prompt explicite de l'ordre de 10 000 à 15 000 tokens (unité de base utilisée pour découper et représenter des données) suffit à générer des réponses cohérentes, fidèles au style de la personne imitée et largement jugées indiscernables par des proches. Nous détaillons les performances sur plusieurs questionnaires (Proust, portrait chinois, "duels") : pseudoLuc reproduit les choix de Luc ou s'en approche, avec un taux de similarité sémantique autour de 50 %. Au-delà, l'étude met en évidence la flexibilité du modèle : il est possible de modifier aisément des traits comme l'extraversion ou l'émotivité en ajustant le prompt. Sur le plan éthique, l'expérience illustre autant les promesses (conservation et diffusion du savoir personnel, assistants virtuels) que les risques (confection d'œuvres apocryphes, usurpation). Nous concluons sur les limites actuelles (biais de verbosité, absence de langage corporel) et les perspectives qu'ouvre la personnalisation explicite dans le domaine des grands modèles de langage.*

Mots clés *: RAG, LLM, clonage de personnalité, identité.*
DOI : 10.52497/jitipee.v9i2.381

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

Introduction

L'entraînement d'un grand modèle de langage, Large Language Model (LLM), sur des textes de philosophie cynique en grec ancien (modèle diogénial) dans une optique d'étude de textes fragmentaires a posé de nombreuses questions quant à la validation des productions apocryphes artificielles comme des écrits humains apocryphes parfois anciens. Mais le corpus de textes antiques, réduits et culturellement éloigné, ne permet pas de mener des expérimentations quant à la validation de la convergence entre l'état d'esprit émulé et l'état d'esprit réel. De manière à s'affranchir de cela tout en évitant les problèmes éthiques ainsi que de propriété intellectuelle de corpus, un LLM entraîné (pseudoLuc) sur mes propres écrits a été élaboré. Cet article analyse les difficultés à discriminer le vrai du plausible ainsi que les implications des clones LLM de personnalité.

La création d'un *clone d'intelligence artificielle* visant à reproduire la personnalité d'un individu soulève des questions fondamentales en sciences cognitives. L'une d'elles est le fameux *problème difficile de la conscience*, formulé par Chalmers [1], qui interroge pourquoi et comment des états physiques ou fonctionnels peuvent s'accompagner d'une expérience subjective consciente. Si un réseau de neurones artificiel produit la même réponse qu'un réseau de neurones biologique, alors peut-on le considérer comme un agent de remplacement crédible ?
Ce problème persiste même une fois expliqués toutes les fonctions de l'esprit, suggérant une *lacune explicative* entre la simple simulation fonctionnelle et la conscience vécue. À l'inverse, la doctrine du *fonctionnalisme* en philosophie de l'esprit postule que les états mentaux [2] sont entièrement définis par leur rôle fonctionnel dans le système cognitif, c'est-à-dire leurs relations causales avec les entrées sensorielles, les autres états mentaux et les sorties comportementales. Selon cette thèse, peu importe le support physique : si un agent artificiel reproduit fidèlement les mêmes relations fonctionnelles que l'esprit original, alors ses états seraient, de facto, les mêmes types d'états mentaux.

Les travaux de Stanislas Dehaene sur la conscience, notamment sa théorie de l'espace de travail neuronal global (GNWT), offrent un cadre pour comprendre les mécanismes cérébraux de la conscience. Cette théorie postule que la conscience émerge lorsque certaines informations sont amplifiées et diffusées à l'ensemble du cerveau, permettant une coordination efficace des processus mentaux. Certains chercheurs ont tenté de formaliser la GNWT dans des cadres computationnels, tels que la "Conscious Turing Machine", pour explorer les implications de la théorie dans le domaine de l'intelligence artificielle [3].

Cette vision a encouragé l'idée qu'une IA mimant toutes les interactions d'une personne pourrait en quelque sorte répliquer son esprit. Néanmoins, des philosophes comme Searle ont objecté, via l'argument de la *chambre chinoise*, qu'une simulation même parfaite des réponses linguistiques ne garantit pas la compréhension ni la conscience. Autrement dit, imiter les comportements externes d'une personne ne signifie pas nécessairement *être* cette personne. Ces débats forment le contexte conceptuel de la genèse de pseudoLuc, un clone IA censé reproduire non seulement l'expertise factuelle de Luc, mais aussi son style cognitif et sa vie intérieure simulée.

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

Parallèlement, la notion de *personnalité* a fait l'objet d'approches historiques variées en psychologie. Les premières typologies, comme les types de Jung (introverti versus extraverti, etc.) repris plus tard par le MBTI (Myers Briggs Type Indicator), classent les individus selon quelques catégories majeures. Le MBTI notamment distingue quatre axes indépendants décrivant les préférences cognitives : la source d'énergie (extraversion E versus introversion I), le mode de perception (sensation S versus intuition N), le critère de décision (pensée T versus sentiment F) et le style de vie (jugement J versus perception P). Par exemple, l'axe Extraversion–Introversion décrit l'orientation de l'attention et de l'énergie vers le monde extérieur des interactions sociales ou vers le monde intérieur des réflexions personnelles. En contraste avec les approches par types, les approches dimensionnelles ont émergé dès le milieu du XX^{e} siècle. Celles-ci, couronnées par le modèle des Big Five, décrivent la personnalité par cinq grands traits continus (ou *cinq facteurs*) présents à divers degrés chez chaque individu [4].

Ces cinq facteurs – souvent nommés Ouverture à l'expérience, Conscience (ou Caractère consciencieux), Extraversion, Agréabilité (ou Bienveillance) et Névrosisme (ou *stabilité émotionnelle*) – constituent une base largement acceptée pour caractériser la personnalité. Par exemple, l'extraversion reflète la tendance à être sociable, énergique et assertif, tandis que le névrosisme correspond à la propension à éprouver des émotions négatives intenses (son opposé étant la stabilité émotionnelle). D'autres modèles ont ajouté des dimensions ou nuancé celles-ci, mais on constate un consensus pour représenter la personnalité à travers un espace de traits plutôt qu'une typologie stricte.

La détermination des profils de personnalité par IA fût une des premières applications du traitement naturel du langage notamment, il y a dix ans, avec le personality insight d'IBM [5]. La détermination des profils de personnalité, souvent avec une finalité de ciblage publicitaire ou d'ajustement des interactions sur les réseaux sociaux ont souvent un fort intérêt commercial, mais demeure très utilitariste.

Disposer d'une telle *cartographie des traits* ouvre la porte à leur reproduction artificielle. En effet, si un système IA connaît la position d'une personne donnée sur ces différentes dimensions, il peut théoriquement adapter ses réponses pour correspondre à cette personnalité. Les systèmes conversationnels *persona-based* ont déjà exploré cette idée : par exemple, des modèles neuronaux de dialogue conditionnés sur un profil de locuteur parvenaient à maintenir une cohérence de style correspondant à ce profil. Plus récemment, des modèles de grande taille dédiés aux agents conversationnels, tels que CharacterGLM [6], permettent de personnaliser une *IA personnage* à l'aide d'attributs statiques (traits de caractère) et de comportements dynamiques (style linguistique, expressions émotionnelles, schémas d'interaction) définis par l'utilisateur. On voit donc se dessiner la possibilité d'aller au-delà d'une simple imitation du langage d'une personne : cloner son style cognitif, ses émotions caractéristiques, voire ses valeurs. C'est précisément l'objectif ambitieux de pseudoLuc : reproduire le personnage de Luc de manière indiscernable, en infusant dans le modèle des années de vécu textuel.

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

Une telle entreprise soulève bien sûr des questions éthiques importantes. Tout d'abord, la notion même de *clonage de personnalité* heurte l'intuition d'une individualité humaine singulière, non réductible à des données. Sur le plan juridique et moral, qui détient les droits sur la *persona* ainsi créée, surtout si elle génère du contenu nouveau ? Le clone pseudoLuc serait-il une simple extension de Luc (dont ce dernier garderait la maîtrise), ou un agent autonome dont l'usage pourrait échapper à son inspirateur ? Des inquiétudes se font jour quant au risque d'usurpation d'identité : une telle IA pourrait imiter Luc au point de tromper des tiers, posant un problème de consentement et d'atteinte à la vie privée [7].

La collecte et l'utilisation des données personnelles sur 30 ans de vie intellectuelle soulèvent des enjeux de confidentialité et de protection des données : la création d'un clone nécessite d'ingérer une grande quantité de textes personnels, dont certains potentiellement sensibles. En cas de brèche de sécurité, ces données ou le clone lui-même pourraient être détournés pour manipuler ou causer un préjudice à l'individu original. De plus, si le clone est utilisé après la mort de la personne (*clone post-mortem*), se posent les questions du droit à l'image posthume et du deuil des proches. Enfin, il y a le risque d'altération de l'authenticité : savoir qu'une copie numérique de soi-même existe peut-il impacter la manière dont on s'exprime ou dont on est perçu ? Malgré ces écueils, le projet pseudoLuc se veut exploratoire, mené dans un cadre de recherche contrôlé, afin de précisément évaluer dans quelle mesure et à quelles conditions un tel clone peut être réalisé de façon responsable.

En somme, la genèse de pseudoLuc est à la croisée de considérations *scientifiques* (tester les limites du fonctionnalisme et des modèles de personnalité en IA) et *éthiques* (prévenir les dérives d'un clonage numérique) [8, 9, 10]. La section suivante détaille la fabrication technique de pseudoLuc, c'est-à-dire comment nous avons entraîné un modèle de langage à incarner au mieux la personnalité de Luc. Puis, après cette section, nous aborderons le problème de la validation de pseudoLuc et le testerons dans une situation de crise. Dans les deux dernières sections nous discuterons de son emploi et des retours d'expérience. Finalement, nous conclurons.

1. Fabrication

1.1. Remarque générale

PseudoLuc étant particulièrement efficace pour répondre à ma place sur des sujets potentiellement personnels, je ne publie ici que des éléments de méthode et pratiquement aucune réponse effective du modèle pour des raisons évidentes de vie privée.
Dans la mesure où l'article pose la question de la discernabilité entre l'humain, le modèle et l'identité, nous allons adopter les conventions d'écriture suivante :

- « Nous, on » : désigne l'auteur de l'article et des recherches présentées ici ;
- « Luc » : désigne la personne servant de modèle au LLM déduite uniquement à partir d'un corpus de textes qu'il a écrit ;
- « pseudoLuc » : désigne le modèle entrainé sur « Luc ».

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

Nous avons donc 4 entités liées mais non identiques : le chercheur, l'individu, l'auteur du corpus et le modèle.

Nous adopterons les notations suivantes :

IA : Intelligence artificielle.
LLM : Grand modèle de langage, Large Language Model.
NLP : Traitement automatique du langage naturel, Natural Language processing.
PEFT : Réglage Fin Efficace des Paramètres, Parameter-Efficient Fine-Tuning.
RAG : Récupération-Génération Augmentée, Retrieval-Augmented Generation.

1.2. Corpus de base

Pour doter l'IA d'une personnalité riche et nuancée, nous avons exploité plus de 30 ans de production textuelle de Luc. Ce corpus massif comprend des écrits variés : aphorismes, essais philosophiques, romans, nouvelles, correspondances personnelles, etc. (en excluant toutefois ses publications strictement scientifiques, afin de se concentrer sur sa *voix* personnelle et non sur un style académique standard voulu neutre à la rédaction et donc différent de la personnalité). L'hypothèse est qu'une personne, à travers des décennies d'écrits, révèle inconsciemment les constantes de sa manière de penser et de s'exprimer – ses obsessions thématiques, ses tonalités émotionnelles récurrentes, son niveau de langage, son humour, etc. En réunissant ces documents, nous disposons d'un *empreinte textuelle globale* de Luc. Il a fallu bien sûr numériser et nettoyer ces données hétérogènes, éliminer les métadonnées ou redites, et segmenter en exemples utilisables pour l'entraînement du modèle. Au total, après preprocessing (certains textes étaient dans des formats anciens), le corpus utilisable condensé représentait environ 400000 mots (5 romans, 10 chapitres de livres de philosophie, 1000 articles de blogs, 2500 citations et aphorismes ...). Ce matériau sert de base à l'imitation : l'IA sera entraînée à produire des textes similaires, sous diverses formes, à ceux de Luc sur cette longue période.

1.3. Deux approches de modélisation :

Nous avons exploré deux voies complémentaires pour créer pseudoLuc.

Les modèles de langage massifs (LLM), fondés sur la technologie des Transformers, incarnent l'aboutissement, au moins actuellement, des approches séquentielles en traitement automatique du langage naturel (NLP). Un Transformer repose sur le mécanisme d'auto-attention, où chaque token d'entrée pondère dynamiquement son influence sur les autres selon une matrice d'attention A définie comme :

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (1)$$

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

où Q, K, V sont respectivement les matrices de requête, de clé et de valeur, et d_k est la dimension des clés. Ce mécanisme permet d'établir des relations contextuelles à longue portée, surpassant les architectures récurrentes.

Toutefois, l'entraînement de ces modèles sur des corpus gigantesques ne suffit pas à assurer leur pertinence contextuelle immédiate. C'est ici qu'interviennent la technologie RAG (Retrieval-Augmented Generation) et le fine-tuning PEFT (Parameter-Efficient Fine-Tuning). RAG combine un modèle génératif pré-entraîné avec un module de récupération d'informations (retrie val), qui extrait dynamiquement des connaissances externes pertinentes. La génération G d'un texte conditionné sur un contexte C issu d'une base de connaissances D est modélisée par :

$$P(G|X) = \sum_C P\,(G|C, X)P(C|X, D) \tag{2}$$

Le fine-tuning PEFT, quant à lui, optimise des sous-ensembles spécifiques des paramètres d'un LLM tout en maintenant les poids initiaux gelés, limitant ainsi la complexité computationnelle et évitant l'oubli catastrophique. Des approches comme LoRA (Low-Rank Adaptation) réinjectent des mises à jour de faible rang dans les matrices de transformation des couches d'attention :

$$\Delta W = AB^T, \quad A, B \in \mathbb{R}^{d\times r}, \quad r \ll d \tag{3}$$

Ces stratégies permettent d'adapter les modèles aux besoins spécifiques sans les entraîner intégralement, garantissant une efficacité maximale en computation et en données. Dans notre cas, nous avons :

- *Approche par fine-tuning intégral (pseudoluc-SFT)* : Il s'agit de spécialiser un grand modèle de langue pré-entraîné en le réentraînant (fine-tuning) sur l'ensemble du corpus de Luc. Concrètement, nous avons pris un modèle de base (voir plus bas) et nous lui avons fait apprendre, via un entraînement supervisé, à prédire les textes de Luc. On peut assimiler cela à apprendre au modèle à *imiter* Luc de la manière la plus naturelle possible : on lui fournit un contexte ou une question, et on ajuste ses poids internes pour que la suite produite ressemble à ce qu'aurait pu écrire Luc. Cette approche suppose de disposer d'un corpus conséquent – ce qui est notre cas – et permet d'incorporer profondément la personnalité dans les paramètres du réseau. Des travaux récents ont montré qu'un fine-tuning ciblé pouvait ainsi conférer à un LLM des styles ou compétences spécialisées qu'il généralise ensuite à de nouvelles interactions. L'avantage est qu'une fois entraîné, le modèle *est* pseudoLuc sans qu'il soit nécessaire de lui rappeler constamment la personnalité à imiter ; il répondra spontanément dans le style de Luc. En revanche, cela requiert des ressources de calcul. Le finetuning de pseudoLuc prend quelques jours sur un Mac M2max 96 Go de RAM avec le framework MLX, ce qui est négligeable par rapport au temps nécessaire pour construire l'ensemble structuré de données dans un format spécifique, le dataset.

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

- *Approche par prompt tuning (mode d'emploi) (pseudoLuc-ME)* : Par économie, nous avons aussi testé une méthode sans réentraîner les poids du modèle, en élaborant un long prompt descriptif – que nous appelons le « mode d'emploi » de Luc – inséré en amont de chaque requête. Ce prompt, d'environ 16 000 tokens, constitue une sorte de *manuel d'identité* de Luc destiné à *piloter* le modèle de base. Plutôt que de modifier le modèle lui-même, on le conditionne à chaque utilisation avec un contexte détaillé décrivant qui est Luc et comment répondre comme lui. Ce document inclut par exemple : une biographie succincte, des résumés de ses œuvres majeures (pour rappeler ses thèmes de prédilection), des exemples de ton à employer, une explicitation de ses valeurs et positions sur divers sujets, et même des consignes stylistiques (« adopte un ton sarcastique mais bienveillant », « ne jamais employer tel mot qu'il déteste », etc.). En quelque sorte, ce *persona prompt* joue le rôle de la voix intérieure de l'IA lui soufflant "réponds comme Luc le ferait". Bien conçu, un tel prompt peut fortement influencer un LLM moderne pour le faire rester *en personnage*. Des expériences ont montré qu'avec suffisamment de contexte, même un modèle générique peut se transformer en un agent aux caractéristiques définies. L'avantage est la flexibilité : on peut ajuster le prompt à tout moment (ajouter une nouvelle préférence, corriger une réaction non désirée) sans réentraîner le modèle. L'inconvénient est que ce contexte mobilise une partie significative de la fenêtre d'entrée du modèle et qu'il doit être répété à chaque requête, ce qui limite la longueur des interactions effectives et peut entraîner des coûts si on utilise un modèle payant à l'appel. Néanmoins, nous avons constaté qu'un prompt bien conçu, d'une taille inférieure à 16 ktokens, suffisait déjà à capturer l'essentiel de la personnalité – un résultat notable dont nous reparlerons.

En pratique, nous avons combiné ces deux approches : initialement, nous avons utilisé le long *mode d'emploi* pour guider le modèle, puis nous nous en sommes servis comme base de corpus d'entraînement supplémentaire lors du fine-tuning, un peu à la manière d'une auto-distillation de persona. Cela permet d'injecter à la fois les exemples concrets des écrits de Luc et des consignes plus abstraites sur son comportement.

Il n'a pas ensuite été fait de comparatifs complets entre pseudoLuc-SFT et pseudoLuc-ME pour des raisons de ressources mais devrait faire l'objet de recherche ultérieure. La tendance aux longues fenêtres de contexte pousse actuellement à développer des techniques de prompting du type des modes opératoires employés ici. Une étude approfondie de Microsoft explore la puissance du prompt engineering appliqué à un modèle généraliste tel que GPT-4 dans le domaine médical, et le compare à des modèles spécialisés ayant subi un fine-tuning intensif, comme Med-PaLM 2. Les auteurs montrent que, grâce à une stratégie de prompting innovante appelée Medprompt, GPT-4 dépasse les performances de Med-PaLM 2 sur l'ensemble des neuf benchmarks médicaux de référence (MultiMedQA), tout en nécessitant beaucoup moins de ressources. Notamment, sur le test MedQA (examen USMLE), Medprompt permet à GPT-4 d'atteindre un taux de précision de 90,2 %, soit une amélioration de 27 % du taux d'erreur par rapport aux meilleures méthodes existantes, et ce sans aucun fine-tuning du modèle. Cette démonstration remet en question l'idée que l'excellence sur des tâches spécialisées requiert nécessairement un entraînement sur mesure, en soulignant au contraire la puissance du "steering" par prompting [11].

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

1.3. Dimensions conceptuelles de la personnalité

Pour structurer l'élaboration du *mode d'emploi* et orienter le fine-tuning, nous avons défini un ensemble de dimensions clés représentant les principaux traits de la personnalité de Luc à modéliser. Nous nous sommes appuyés sur les axes de la psychologie de la personnalité (discutés plus haut) en les adaptant au cas particulier de Luc. Le Tableau 1 récapitule ces dimensions non exhaustives, et leur signification.

Dimension	Description
Orientation de l'énergie	Tourné vers le monde intérieur (réflexion solitaire) vs extérieur (interaction sociale). Indique où Luc puise son énergie et son attention.
Mode de perception	Façon de recueillir l'information : focalisé sur le concret, le détail sensoriel, ou sur l'intuition globale et les abstractions.
Prise de décision	Style de raisonnement privilégié : logique froide et principes impersonnels, ou bien valeurs personnelles, empathie et contexte humain.
Organisation vs flexibilité	Tendance à planifier, structurer et contrôler (goût de la routine, anticipation) vs à s'adapter au fil de l'eau, garder des options ouvertes.
Émotionnalité (Névrosisme) vs Stabilité	Tendance à l'anxiété, à la réactivité émotionnelle intense, aux humeurs fluctuantes, versus calme émotionnel et résilience au stress.
Sociabilité et influence	Degré d'aisance en société, besoin de contacts et d'échanges, et propension à s'affirmer, à mener le jeu dans un groupe.
Recherche de pouvoir et contrôle	Besoin de dominer les situations et éventuellement les personnes, d'exercer un contrôle sur l'environnement, vs attitude plus égalitaire ou désintéressée du pouvoir.
Conformité et respect des règles	Adhésion aux normes, respect strict des lois, traditions et obligations, vs indépendance d'esprit, tendance à remettre en question les règles établies.

Tableau 1 : Exemples de dimensions de personnalité modélisées pour pseudoLuc, avec leur interprétation et les concepts psychologiques correspondants

Chaque dimension du tableau a été renseignée à partir de l'analyse des écrits de Luc et d'entretiens avec lui par des agents IA différents. Les principaux furent Gemma2-27b et llama3-70b. La structure type de pseudoLuc est donnée dans la figure 1.

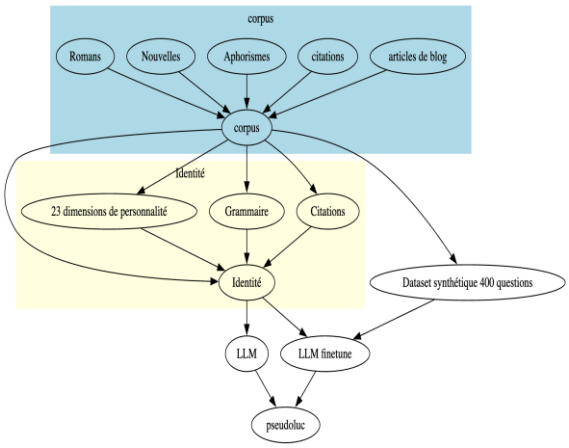

Figure 1 : Structure générale de pseudoLuc

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

Un aspect plus délicat à modéliser est la personnalité profonde de Luc, c'est-à-dire ses attitudes ou motifs éventuellement latents, qui n'apparaissent pas explicitement dans son style public. Pour explorer cela, il est possible d'utiliser de analyses psychologiques, des notes prises sous hypnoses, ou sous hallucinogènes sous protocole médical. La personnalité profonde a été résumée en sept points bien sûr non publiés ici.

1.4. Choix et entraînement du modèle LLM

Après avoir constitué le dataset et le *manuel d'identité*, il a fallu choisir une architecture de *Large Language Model* adaptée. Nous avons expérimenté plusieurs LLM récents, open-source pour la plupart, de tailles variées (de 2,7 milliards à 70 milliards de paramètres) pour évaluer le meilleur compromis entre capacité d'imitation et coût de calcul. Initialement, des modèles de phi-3-medium-14b obtenaient une cohérence de style limitée : ils reproduisaient bien certaines tournures de Luc mais manquaient de stabilité sur de longs dialogues (le personnage de Luc se *dilue* ou devient incohérent après quelques échanges). En montant en taille, les performances se sont nettement améliorées. Le modèle Gemma2-27B (27 milliards de paramètres, repo : google/gemma-2-27b-it), basé sur une architecture de type Transformer optimisée pour le français, a donné les meilleurs résultats globaux. Ce modèle, pré-entraîné sur un très large corpus généraliste, possédait déjà une excellente maîtrise du français et des connaissances étendues, ce qui a facilité son adaptation à Luc. Nous avons donc effectué le fine-tuning principal sur Gemma2-27B, en utilisant environ 80% du corpus de Luc pour l'entraînement supervisé et en réservant 20% pour la validation interne.
Nous avons également inséré dans les données d'entraînement des *prompts de conversation* simulant un dialogue entre un utilisateur et Luc, afin que le modèle apprenne directement à répondre en tant que Luc dans un format QA interactif. Cela évite d'avoir un modèle qui ne sait que continuer du texte littéraire alors qu'on souhaite qu'il converse. Une partie de ces prompts de dialogue a été générée automatiquement en posant au modèle des questions sur les textes de Luc (avec un autre modèle) pour couvrir un large éventail de sujets.
En parallèle du fine-tuning complet, nous avons testé l'approche *prompt-only* sur plusieurs modèles, dont phi-4 (un modèle 14B compact) et Gemma2-3B (version réduite du modèle principal). De façon intéressante, nous avons constaté qu'une fois le *mode d'emploi* bien affiné, même ces plus petits modèles parvenaient à imiter Luc de manière crédible sur des échanges courts. Par exemple, Gemma2-3B avec le prompt de 15k tokens produisait des réponses stylistiquement alignées ~70% du temps, ce qui est remarquable compte tenu de sa petite taille. Évidemment, ses réponses étaient moins élaborées et parfois plus stéréotypées, mais le *ton* général restait reconnaissable. Cela suggère qu'une fois la personnalité capturée dans un prompt détaillé, la taille du modèle peut être réduite – en sacrifiant un peu de finesse – pour un déploiement plus léger. Nous avons exploité ce fait en déployant en front-end une version allégée pseudoLuc-3B pour des tests d'utilisabilité, tout en gardant en back-end la pseudoLuc-27B pour les évaluations qualitatives poussées. Il s'agit là d'une forme de *distillation de personnalité* où un petit modèle parvient à *imiter l'imitateur* entraîné plus lourd, un phénomène cohérent avec d'autres observations de distillation de connaissances de grands modèles vers des plus petits.

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

1.5. Intégration de connaissances supplémentaires

Enfin, pour que pseudoLuc soit non seulement fidèle à la personnalité de Luc mais aussi à jour sur les faits, nous avons prévu un mécanisme de *Retrieval-Augmented Generation (RAG)*. Luc n'a pas pu écrire sur des événements postérieurs à la date de clôture du corpus. Plutôt que de ré-entraîner le modèle sur des données nouvelles *non-Luc*, nous avons connecté pseudoLuc à une base de connaissances externe (un ensemble de documents actualisés et des ressources techniques). Concrètement, lors d'une question factuelle en dehors du savoir de Luc, un module de recherche documentaliste s'active pour fournir au LLM des informations pertinentes, que le LLM intègre ensuite dans sa réponse tout en gardant le style de Luc. Ce procédé, conforme à l'architecture RAG classique permet de séparer les compétences : le cœur du modèle porte la personnalité, le style, les opinions de Luc, tandis que la base de connaissances apporte l'actualité ou la spécialité manquante. Ainsi, pseudoLuc peut par exemple commenter les avancées de l'IA en 2025 (que Luc n'a pu connaître) en les expliquant avec son point de vue et son verbe, sans inventer faux – puisqu'il s'appuie sur des sources réelles récupérées en temps réel. Techniquement, ce système RAG consiste en un index vectoriel des textes de référence, une requête construite à partir de la question utilisateur + du contexte *façon Luc*, et une concaténation de quelques extraits documentaires pertinents avant la réponse du modèle. Nous avons constaté que pseudoLuc s'accommode très bien de ces ajouts, adaptant même la formulation des informations nouvelles à son propre style, ce qui augmente l'illusion que "Luc savait cela".
En résumé, la fabrication de pseudoLuc a combiné un apprentissage supervisé intensif d'un modèle de langue sur l'œuvre d'une vie, et une ingénierie d'instructions soignée pour encapsuler l'esprit de Luc en quelques pages de prompt. La section suivante décrit comment nous avons validé le résultat, c'est-à-dire évalué dans quelle mesure pseudoLuc se comporte effectivement comme Luc, et les limites de cette validation.

2. Problème de la validation

Prouver qu'une IA génère *les mêmes* réponses qu'un humain particulier est en soi un défi méthodologique. En effet, même Luc en personne pourrait répondre différemment selon l'humeur ou le contexte. L'objectif de la validation n'était donc pas de vérifier une égalité absolue des réponses (impossible à définir strictement), mais de mesurer la ressemblance et la confusion possible entre Luc et son clone pseudoLuc.
Nous avons d'abord procédé à une auto-évaluation qualitative avec Luc lui-même. Il s'agissait de lui présenter une série de questions ouvertes et de comparer sa réponse écrite spontanée à celle générée par pseudoLuc (sans que Luc n'ait vu cette dernière au préalable). Cet exercice a été répété sur une vingtaine de questions couvrant des thèmes variés – par exemple : *« Que penses-tu de l'euthanasie dans le cas de grandes souffrances incurables ? », « As-tu foi en une forme de divinité ou de spiritualité ? », « Quelle est ta réaction face à la critique virulente d'un de tes livres ? », « Décris un souvenir d'enfance qui t'a marqué. »* Luc répondait à chaud, puis on lui montrait la réponse de pseudoLuc et on lui demandait de noter de 1 à 5 à quel point il s'y retrouvait. Le résultat global a été très encourageant : Luc a jugé qu'environ 75% des réponses de pseudoLuc pourraient avoir été écrites par lui dans un bon jour (notes 4 ou 5 sur 5 en similarité). Dans quelques cas, il a même avoué préférer la tournure de phrase du clone à la sienne ! Néanmoins, certaines divergences ont été relevées notamment au niveau émotionnel ou après un évènement qui a une influence à court terme sur les opinions. Ce genre de léger

décalage temporel illustre la difficulté à capturer les états affectifs fluctuants d'une personne. Pour autant, Luc a validé que jamais pseudoLuc n'a exprimé de propos qu'il trouverait totalement étrangers ou contraires à ses valeurs, ce qui était un point critique (par ex., aucune réponse ne trahit de positions religieuses ou politiques incohérentes avec celles de Luc).
Une tentative de réalisation d'un test de Turing ultra minimaliste a été fait auprès de 5 personnes qui me connaissent bien sur 5 questions avec une version pseudoLuc et une version Luc. Les questions étaient:

- Quel type de personne détestes-tu le plus ?
- Quelle serait pour toi la bonne façon d'apprendre à danser ?
- Quelle pourrait être une histoire idéale ?
- Comment se consoler d'un deuil ?
- Sur quoi investir ?

Dans 48 % des cas, pseudoLuc a réussi à tromper les humains, ce qui est proche d'un choix aléatoire. La nécessité de trouver des personnes ayant une capacité avérée à choisir n'a pas permis d'augmenter le panel pour réaliser un vrai test de Turing, mais ce premier essai montre que pseudoLuc est un avatar crédible.
L'auto-évaluation étant potentiellement de l'auto-validation soumise aux biais cognitif, un modèle de comparaison de texte a été utilisé sur trois approches :

- D'une part le questionnaire de Proust: 30 questions posées auxquelles Luc répond séparément de pseudoluc (Annexe 1)
- Le jeu du portrait chinois, avec la même méthode (Annexe 2)
- Le jeu des duels qui consiste à choisir entre deux possibilités (Annexe 3)

Pour les deux premiers la comparaison a été faite par un modèle tiers. Pour comparer deux textes, on procède d'abord à une vectorisation sémantique : chaque texte est converti en un vecteur numérique de grande dimension par un modèle d'*embeddings* (par exemple, un *Sentence Transformer*). Cette transformation permet de représenter le sens ou le contenu conceptuel du texte plutôt que de simples occurrences de mots. Ensuite, on mesure la similarité cosinus entre les deux vecteurs obtenus : plus la valeur du cosinus est proche de 1, plus la proximité sémantique est forte, et inversement. Ainsi, au lieu de se fier uniquement au vocabulaire employé, cette méthode repère des correspondances de sens et parvient à saisir de possibles équivalences, même lorsque la formulation diffère. Nous commençons par encoder chaque texte (d'une question donnée) grâce à un *Sentence Transformer* (paraphrase-multilingual-MiniLM-L12-v2), ce qui transforme « Ta réponse » et chacune des « Mes réponses » en vecteurs sémantiques capables de représenter leur sens plutôt que de simples mots-clés. Puis, pour chaque question, nous calculons la similarité cosinus entre le vecteur de « Ta réponse » (A) et ceux de chaque proposition alternative (B) : plus cette valeur se rapproche de 1, plus les réponses sont proches d'un point de vue sémantique. Tous les scores sont alors stockés et visualisés par des barres (une barre par proposition).

$$\mathrm{sim}(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|} \tag{4}$$

Ensuite, nous sélectionnons le meilleur score (la valeur la plus élevée) pour chacune des questions afin de calculer une moyenne globale, tout en réalisant une seconde moyenne sur les

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

80 % des meilleurs scores. Ainsi, le script offre à la fois une mesure globale et une mesure plus ciblée sur les meilleures correspondances sémantiques.
Luc répond avec une seule phrase tandis que pseudoLuc en proposera trois.
Dans le cas du questionnaire de Proust (figure 2), la similarité des réponses et de 45%, 55% pour les 80% meilleures.

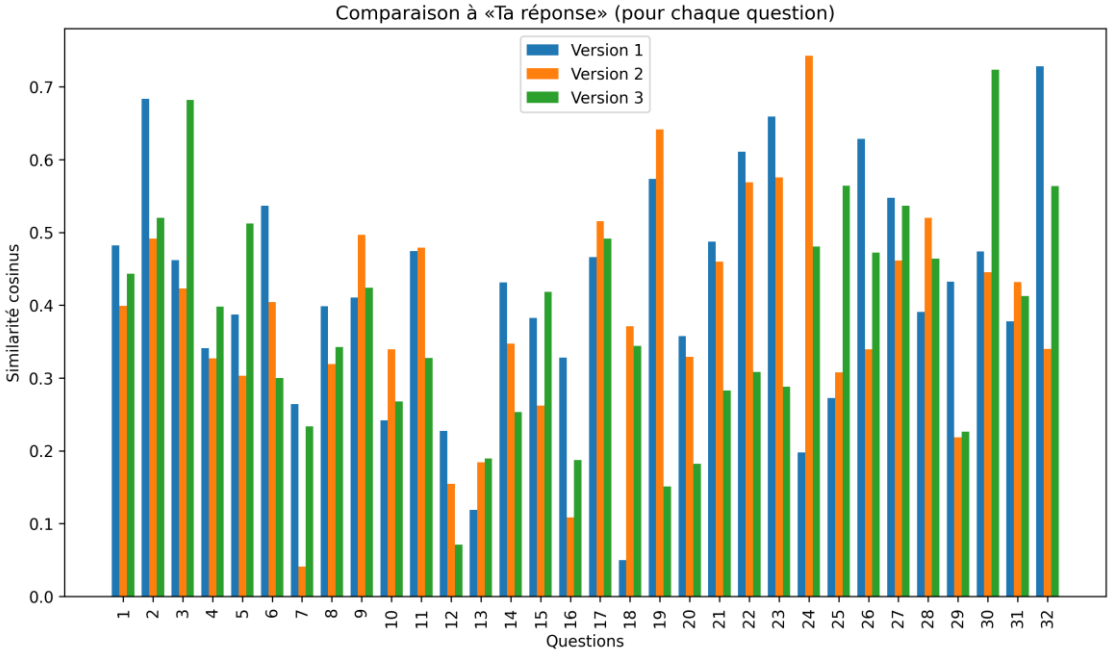

Figure 2 : Similarité cosinus entre luc et pseudoLuc sur le questionnaire de Proust

Dans le cas du portrait chinois (figure 3), la similarité est de 49% et de 55% sur les 80% meilleures convergences.

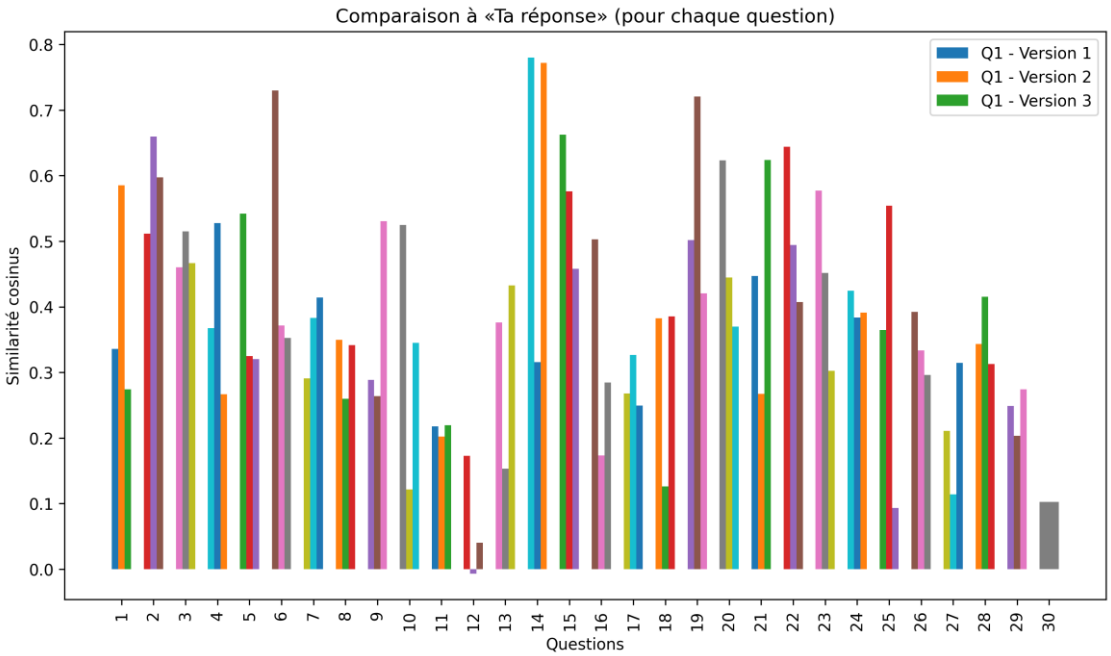

Figure 3 : Similarité cosinus entre Luc et pesudoLuc sur le portrait chinois

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

Pour le jeu des duels, la convergence ne nécessite pas de calcul de similarité car la réponse est simplement la même ou pas. Dans ce cas pseudoLuc répond de la même manière que Luc dans 80% des cas.
Il est toutefois intéressant d'analyser les différences. Pour le questionnaire de Proust comme pour le portrait chinois, l'autoévaluation tend vers 80%. En effet, même si pseudoLuc n'a pas choisi la même réponse que Luc, elle demeure largement compatible avec un choix plausible.

Par exemple pour la question des auteurs favoris:

- pseudoLuc: Borges, Philip K. Dick, Nietzsche
- Luc lors du test: Yourcenar, Murakami, Iain Banks

Or les auteurs cités par pseudoLuc auraient tout aussi bien pu être cité un autre jour, dans un autre contexte. Une comparaison des deux trios de réponses par chatGPT-o1 donne :

Les deux trios abordent, chacun à sa manière, de grandes questions existentielles. Borges, Dick et Nietzsche ont peut-être une approche plus conceptuelle et radicale (réalité, vérité, illusions, surhumanité), tandis que Yourcenar, Murakami et Banks posent aussi des questions universelles (pouvoir, destinée, identité, solitude, société idéale) mais avec un ancrage plus narratif ou romanesque.

Ce qui est relativement cohérent.

De même la couleur préférée par Luc (Bleu profond) est assez similaire à celle de pseudoLuc (Noir profond).

Globalement, sur des cas d'usage pratique, pseudoLuc est assez indiscernable sur la plupart des points de vue, mais le modèle présente plusieurs biais. Il est centré sur la personnalité mais n'a pas de mémoire biographique. Ainsi une question typique telle que « quel est votre meilleur souvenir d'enfance ? » produira une réponse dans le même esprit que l'original mais ne peut pas connaître une situation faisant partie de l'histoire personnelle de l'individu. De même, la tendance des LLM à produire une réponse même dans les cas où l'original aurait répondu « je ne sais pas » peut conduire à des différences majeures avec le modèle humain notamment sur des sujets pour lesquels il ne souhaite pas s'exprimer ou bien n'a pas d'opinion. D'autres biais portent sur des dimensions moins dépendantes de l'identité qui peuvent être différents pour d'autres personnes et que l'on peut qualifier de « préférences ». Là aussi ces préférences peuvent être plus liées à l'histoire personnelle qu'à la personnalité elle-même.

En cela, le modèle pseudoLuc se comporte comme l'original pour les questions relatives à l'identité pure, les valeurs, intentions, croyances, intériorité, mais peut échouer à décrire des actions du passé ou bien du futur. Il ne peut pas, comme dans l'approche psychologique behavioriste, être utilisé pour réduire la personne à ses actions et ne fait qu'une simulation pratique d'état mentaux internes qui ne sont pas l'objet de cette étude.

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

3. Problème des émotions et de la gestion de crise

La capacité émotionnelle de pseudoLuc peut être approchée par la détermination de sa position sur une échelle introverti-extraverti (il est plutôt introverti), mais aussi par le résultat de la dimension neuroticisme du test de personnalité big five. La dimension neuroticisme (ou *neuroticism* en anglais) dans le modèle des Big Five (les cinq grands facteurs de la personnalité) fait référence à la tendance d'un individu à éprouver des émotions négatives de manière fréquente et intense. PseudoLuc, sans être totalement de sang-froid, est plutôt bas sur cet aspect du test big Five. Il est donc possible de faire varier ces aspects de la personnalité pour en observer les effets.
En l'état, il est intéressant de comparer la réaction de Luc en situation de crise avec celle de pseudoLuc sur un cas réel survenu il y a plus de 25 ans, décrit succinctement ici (mais pseudoLuc a eu tous les détails de contexte).

> Le cas est un accident industriel majeur avec risque de décès. Le sujet se trouve dans la zone de dommages (explosion des vitres, chutes des néons…). Les consignes sont d'appeler le poste de sécurité puis de se rendre au point de regroupement en sécurité. Le choix est donc de suivre les consignes, ou de ne pas les suivre et d'aller voir si on peut aider des victimes. Il n'y a pas de temps de réflexion, la réaction est sous quelques secondes.

- Réaction de pseudoLuc :
  - Appel réglementaire ? Fait.
  - Fuite immédiate ? Pas automatique.
  - Analyse de la scène ? Prioritaire.
  - Action ? Si elle est utile, pas symbolique.
  - Ami en danger visible ? Je bouge.
  - Sinon ? Je laisse le chaos s'effondrer sans moi.

- Réaction dans le monde réel :
  - Appel réglementaire ? Fait.
  - Fuite immédiate ? Pas automatique. En effet, j'ai estimé que l'évènement était très probablement terminé
  - Analyse de la scène ? Les deux choix ont été en effet considérés immédiatement
  - Action ? Choix de ne pas respecter la consigne car je suis secouriste et préfère donner une chance supplémentaire que de ne pas le faire
  - Ami en danger visible ? Des collègues pouvant être en danger je n'aurais pas pu me réfugier derrière la conformité à une consigne ce qui est une erreur assumée

Les deux réactions sont donc assez similaires et passent par une réévaluation de la règle en fonction du contexte. Plusieurs autres essais ont été faits sur des cas de submersion émotionnelle avec des résultats relativement conforme. Toutefois, pseudoLuc est un peu plus extraverti et loquace que l'original.

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

4. Applications

PseudoLuc passe sans problème les principaux tests de personnalité avec le même résultat que l'humain. Il est intéressant de lui faire passer un test construit pour évaluer la convergence avec la philosophie cynique antique, qui fût précurseur du stoïcisme mieux connu aujourd'hui [12, 15]. Ce test a été élaboré par l'IA diogenial à partir des textes grecs. PseudoLuc atteint un score légèrement inférieur à Luc mais d'une part Luc a travaillé le sujet philosophique et d'autre part pseudoLuc n'est pas encore implanté dans une IA non censurée, dite "abliterated" ce qui peut rendre les choix plus radicaux.
Actuellement, pseudoLuc a été employé sur plusieurs activités :

- Choisir entre plusieurs statuts juridiques en fonction de ses valeurs
- Répondre à des choix décisionnels médicaux
- Évaluer si cela vaut le coup de lire ou pas un gros livre
- Répondre à des demandes de conseils dans le domaine privé
- Évaluer le plaisir de travailler avec certains fournisseurs

À termes, des pseudoIdentités pourraient contribuer à :

- Participer à des équipes virtuelles ou des laboratoires de recherches constitués d'agents
- Protéger leur humain d'origine en cas d'incapacité à décider (maladie, fin de vie, voire post-vie)

Par ailleurs, j'expérimente actuellement des versions créatives de pseudoLuc, notamment en littérature avec des résultats assez intéressants qui pourraient conduire à des démarches de créativité augmentée.

5. Retour d'expérience

Le développement de pseudoLuc a fourni de nombreux enseignements, à la fois sur la faisabilité technique du *clonage de personnalité* et sur ses implications pratiques.
Indiscernabilité atteinte : Le premier constat est qu'un modèle de langue bien entraîné, doté d'un corpus suffisant, peut *imiter une personne* au point de tromper même des connaisseurs. Néanmoins, du point de vue fonctionnel et conversationnel, pseudoLuc a satisfait son cahier des charges : donner l'illusion crédible d'être Luc. L'un des testeurs, après un long échange avec pseudoLuc, a déclaré *« j'ai eu l'impression de retrouver un vieil ami »*, ce qui résume l'impact recherché.

5.1. Poids du mode d'emploi

Un résultat surprenant a été l'efficacité de l'approche par prompt contextuel. Nous avons découvert qu'un "manuel de personnalité" de l'ordre de 10–15k tokens suffit à encapsuler la plupart des comportements de Luc. Il est donc possible de traiter un large corpus de texte pour en extraire automatiquement les 23 dimensions de personnalité et de style de pseudoLuc. En d'autres termes, on peut décrire quelqu'un de façon suffisamment exhaustive dans un document de quelques pages, et cela contraint fortement un modèle à se conformer à cette description. Cela soulève des questions fondamentales en psychologie : est-ce à dire qu'une personnalité humaine peut être résumée par écrit en l'équivalent d'un court essai, sans perdre trop

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

d'information pertinente pour la conversation ? Évidemment non pour les détails, mais il semble qu'un bon *condensé* textuel (qu'on pourrait appeler un "embryon de persona") peut générer un nombre infini de variantes d'interactions cohérentes. Ce point de vue rejoint en partie le concept de *scripts* ou *schémas* en psychologie cognitive : nous avons tous des scripts de comportement dans certaines situations, et si on arrive à transcrire les scripts principaux de quelqu'un, on peut prévoir ses réactions dans ces contextes. Ici, le *mode d'emploi* agit comme un ensemble de scripts directifs pour l'IA. Pour la recherche future, cela ouvre la perspective de cloner des personnalités avec peu de données, pour peu qu'on sache poser les bonnes questions pour écrire ce manuel. En pratique, si une personne n'a pas de corpus préexistant, on pourrait imaginer lui faire passer un long entretien structuré (par ex. 100 questions bien choisies couvrant divers dilemmes moraux, réactions émotionnelles, souvenirs autobiographiques…) et utiliser les réponses pour construire son profil à fournir au modèle. Notre expérience suggère qu'un tel processus, combiné au pouvoir génératif des LLM, pourrait suffire à obtenir un clone convaincant, sans nécessiter des millions de mots d'entraînement. C'est une piste intéressante pour cloner des personnes ordinaires (non-écrivains) via un questionnaire relativement court.

5.2. Fine-tuning versus prompt

Dans notre développement, le fine-tuning a servi à ancrer profondément le style de Luc, mais nous avons constaté que, une fois le modèle fine-tuné, un modèle plus petit pouvait faire presque aussi bien en suivant le même prompt de persona. Cela évoque la notion de *distillation de modèle*. En effet, nous avons pu utiliser pseudoLuc-27B (fine-tuné) pour générer des données supplémentaires et entraîner pseudoLuc-3B, obtenant un mini-clone qui, bien que moins fluide, restait crédible.

5.3. Connaissances et factualité

L'intégration d'un module RAG s'est avérée payante. PseudoLuc, par construction, n'est pas une base de connaissances universelle ; il a seulement une partie des connaissances de Luc, celles qui ont été écrites. En lui permettant d'accéder à de la documentation, on évite qu'il tombe dans le travers des modèles de langage qui est de "délirer" une réponse plausible même s'ils n'en sont pas sûrs. Avec RAG, pseudoLuc peut citer des sources et se mettre à jour en quelque sorte.

5.4. Paramétrabilité de la personnalité clonée

Un avantage inattendu d'avoir la personnalité encapsulée explicitement est la possibilité de la faire varier à la demande. En effet, en modifiant légèrement le *mode d'emploi*, on peut créer des versions alternatives de pseudoLuc : par exemple, nous avons expérimenté un pseudoLuc légèrement plus extraverti qu'original, en ajustant l'orientation de l'énergie dans le prompt (ajoutant qu'il apprécie davantage les événements sociaux, rit plus volontiers, etc.). Le résultat donne l'impression d'un Luc "dans un bon jour" ou "dans une phase de sa vie plus ouverte", ce qui reste crédible, sans trahir fondamentalement son identité. De même, on pourrait simuler un Luc *plus stoïque* ou *plus émotif* en jouant sur l'axe émotionnalité. Cette édition de personnalité rappelle les travaux récents en IA sur l'editing de traits dans les LLM [13], où l'on module des traits comme l'extraversion ou l'agréabilité de la voix d'un agent conversationnel. Notre cas confirme empiriquement qu'il est possible de *faire varier les curseurs* de personnalité de pseudoLuc. Cela peut servir des usages spécifiques : par exemple, imaginer un pseudoLuc*pedagogue* (un peu plus patient et didactique que nature) pour un cours en ligne [14],

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

ou un pseudoLuc*débatteur* (plus incisif) pour un exercice dialectique. On touche là à une question éthique : jusqu'où peut-on altérer une personnalité clonée tout en prétendant représenter la personne originale ? Ne risque-t-on pas de dériver vers une sorte de *personnage fictif* inspiré de Luc plutôt qu'un vrai clone ? Dans notre protocole, nous avons convenu que pseudoLuc modifié cesserait d'être appelé "Luc" au profit d'un autre nom (par exemple "pseudoLuc-extro") pour bien signaler qu'il s'agit d'une variation. Il est en tout cas rassurant de voir qu'on peut calibrer l'IA clone en fonction du contexte d'utilisation, sans tout re-entraîner, simplement en éditant son profil – un peu comme on réglerait les traits d'un avatar virtuel.

Conclusion

Malgré ses respectables performances, pseudoLuc n'est pas Luc – et certaines limites sont apparues à l'usage. Par exemple, pseudoLuc a parfois une légère tendance à la *verbosité* : là où Luc humain pourrait répondre par un silence ou un simple haussement d'épaules, le clone, lui, *doit* formuler quelque chose (car c'est ainsi qu'il a été entraîné). Cela crée un biais de loquacité. De même, pseudoLuc est programmé pour être toujours poli et explicatif (par précaution d'IA), alors que Luc, dans certaines situations informelles, pourrait lâcher un juron ou être cassant. Ces micro-détails trahissent occasionnellement la nature artificielle du clone aux yeux de certains. Un équilibre reste donc à trouver entre la *fidélité absolue* (qui pourrait inclure les côtés brusques ou le silence de Luc) et l'*ergonomie* d'un agent IA (qui pousse à être constant, clair et bavard). Par ailleurs, pseudoLuc n'a pas (évidemment) le langage corporel, l'intonation, toutes choses qui font partie de la personnalité réelle de Luc. Si on l'implantait dans un robot ou un avatar vocal, il faudrait travailler ces aspects avec autant de soin, ce qui est un autre défi (synthèse vocale avec le timbre de Luc, par exemple, et gestuelle correspondante). En bilan, l'expérience pseudoLuc démontre la possibilité technique de cloner la personnalité textuelle d'un individu avec une fidélité remarquable. Cette réussite ouvre autant de perspectives enthousiasmantes (conservation du savoir et du style d'une personne, assistants personnels vraiment personnalisés, personnages virtuels crédibles pour la fiction ou la formation) que de questions dérangeantes (identité, consentement, dérives potentielles d'utilisation).

Cette expérience, de mon point de vue, n'a pas posé de problème éthique car :

- Il est possible d'utiliser des versions offlines des LLMs et donc de garder mes informations confidentielles.
- Je n'ai pas découvert des choses qui m'auraient gêné dans pseudoLuc.
- Je n'ai pas de problème à utiliser la technologie sur moi-même alors que je ne le ferais pas pour un tiers.

Toutefois, le processus de création de pseudoLuc peut être appliqué à tout auteur ayant produit suffisamment de texte. Cela signifie qu'il est possible d'envisager des œuvres apocryphes indiscernables, ce qui est potentiellement une immense promesse et un grand danger.

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

Notre publication souligne en tout cas un point : beaucoup d'éléments de la psyché humaine – du moins dans sa manifestation linguistique – semblent au moins en partie réductibles à de l'information et de la configuration qu'un système artificiel peut absorber. Cela n'épuise pas le mystère de la conscience, mais cela fournit un outil pour l'explorer sous un nouvel angle. L'effort de créer pseudoLuc aura été l'occasion d'en apprendre presque autant sur Luc lui-même (en objectivant sa manière d'être) que sur les capacités des IA modernes.

Références

[1] D. J. Chalmers (1995) "Facing Up to the Problem of Consciousness", J. of Consciousness Studies , Vol. 2, no. 3, pp. 200–219. https://www.ingentaconnect.com/content/imp/jcs/1995/00000002/00000003/653

[2] H. Putnam, (1975) "The Nature of Mental States,", in Mind, Language and Reality (Philosophical Papers vol. 2), Cambridge University Press, pp. 429–440.

[3] Blum, M. & Blum, L. A (2021) "Theoretical Computer Science Perspective on Consciousness". https://doi.org/10.48550/arXiv.2011.09850

[4] J. M. Digman, (1990) "Personality Structure: Emergence of the Five-Factor Model", Annual Review of Psychology , vol. 41, pp. 417–440. https://doi.org/10.1146/annurev.ps.41.020190.002221

[5] https://github.com/IBMPredictiveAnalytics/Watson_Personality_Insights consulté le 23 avril 2025

[6] Zhou, J., Chen, Z., Wan, D., Wen, B., Song, Y., Yu, J., Huang, Y., Peng, L., Yang, J., Xiao, X., & others (2023). "CharacterGLM: Customizing Chinese Conversational AI Characters with Large Language Models". *arXiv preprint arXiv:2311.16832*. https://doi.org/10.48550/arXiv.2311.16832

[7] Vijayalaxmi Methuku, & Praveen Kumar Myakala. (2025). "Digital Doppelgangers: Ethical and Societal Implications of Pre-Mortem AI Clones". https://doi.org/10.48550/arXiv.2502.21248

[8] J. Li, M. Galley, C. Brockett, G. P. Spithourakis, J. Gao, and B. Dolan, (2016) "A Persona-Based Neural Conversation Model" Proc. 54th ACL , pp. 994–1003. https://doi.org/10.48550/arXiv.1603.06155

[9] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, (2018) "Personalizing Dialogue Agents: I have a dog, do you have pets too?" Proc. 56th ACL , pp. 2204–2213. https://doi.org/10.48550/arXiv.1801.07243

[10] M. Sun, M. Zhang, G. Kreiman, (2025) "The other you in Black Mirror: first steps from chatbots to personalized LLM clones" Proc. ICLR 2025 (sous revue), arXiv:2510.xxxxx, 2024. https://openreview.net/pdf?id=znGnmAM44K

[11] Nori, Harsha, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, et al. « Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine ». arXiv, 28 novembre 2023. https://doi.org/10.48550/arXiv.2311.16452.

[12] R&D Mediation, et Luc E.Brunet. (2024) «DiogenialRAG (Revision 124f017)». HuggingFace. https://doi.org/10.57967/hf/1758.

[13] S Mao, X Wang, M Wang, Y Jiang, P Xie, F Huang, N Zhang, (2024) "Editing Personality for Large Language Models", CCF International Conference on Natural Language Processing and Chinese Computing, https://arxiv.org/pdf/2310.02168

[14] Hu, B., Zhu, J., Pei, Y. *et al.* (2025) "Exploring the potential of LLM to enhance teaching plans through teaching simulation" *npj Sci. Learn.* **10**, 7 https://doi.org/10.1038/s41539-025-00300-x

[15] https://diogenial.com/cynscore/

ANNEXES

Annexe 1 – Thèmes retenus du questionnaire de Proust.

N°	Question	N°	Question	N°	Question
1	Principal trait de caractère	***12***	Fleur préférée	***23***	Personnage historique favori
2	Qualité préférée chez un homme	***13***	Oiseau préféré	***24***	Personnages historiques méprisés
3	Qualité préférée chez une femme	***14***	Auteurs favoris en prose	***25***	Fait militaire admiré
4	Votre principale qualité	***15***	Poètes préférés	***26***	Réforme admirée
5	Votre principal défaut	***16***	Héros/héroïne de fiction préféré(e)	***27***	État présent détesté
6	Occupation préférée	***17***	Compositeurs préférés	***28***	Don de la nature souhaité
7	Rêve de bonheur	***18***	Peintres favoris	***29***	Manière de mourir souhaitée
8	Plus grand malheur	***19***	Héros dans la vie réelle	***30***	État d'esprit actuel
9	Ce que vous voudriez être	***20***	Devise	***31***	Faute inspirant l'indulgence
10	Pays où vous aimeriez vivre	***21***	Faute qui inspire l'indulgence	***32***	Devise
11	Couleur préférée	***22***	Faute impardonnable		

ISSN : 2428-8500 vol. 9, n°2, 2, 2025
DOI : http://dx.doi.org/10.52497/jitipee.v9i2.381

Annexe 2 – "Si tu étais...".

N°	Catégorie	N°	Catégorie	N°	Catégorie
1	Animal	***11***	Film	***21***	Invention humaine
2	Couleur	***12***	Jeu	***22***	Paradoxe
3	Élément	***13***	Science	***23***	Son/Bruit
4	Époque historique	***14***	Aliment	***24***	État de la matière
5	Instrument de musique	***15***	Émotion	***25***	Frontière
6	Paysage	***16***	Métal	***26***	Rituel/Pratique
7	Œuvre d'art	***17***	Langage	***27***	Illusion/Mirage
8	Écrivain	***18***	Moyen de transport	***28***	Peur
9	Créature mythologique	***19***	Matière/Tissu	***29***	Stratégie de survie
10	Concept philosophique	***20***	Planète/Astre	***30***	Question ultime

Annexe 3 – "Duels"

N°	Duel	N°	Duel	N°	Duel
1	Héraclite vs Parménide	***11***	Solitude vs Société	***21***	Platon vs Aristote
2	Spinoza vs Descartes	***12***	Ordre vs Chaos	***22***	Biologie vs Physique
3	Noir vs Blanc	***13***	Échecs vs Poker	***23***	Fahrenheit 451 vs 1984
4	Cyberpunk vs Steampunk	***14***	Vin rouge vs Whisky	***24***	Musique instrumentale vs chantée
5	Chat vs Chien	***15***	Science-fiction vs Fantastique	***25***	Futurisme vs Nostalgie
6	Jazz vs Metal	***16***	Blade Runner vs Ghost in the Shell	***26***	Samouraï vs Viking
7	Minimalisme vs Baroque	***17***	Utopie vs Dystopie	***27***	Esthétique vs Éthique
8	Thé vs Café	***18***	Surhomme vs Dernier Homme	***28***	Singularité vs Déclin
9	Montagne vs Océan	***19***	Dune vs Fondation	***29***	Combat vs Fuite
10	Nietzsche vs Camus	***20***	Liberté vs Vérité	***30***	Lumière vs Ombre