Étude en Cours : Évaluation de la Capacité des Modèles de Langage à Simuler le Discours de Patients lors de Thérapies: Impact du Fine-Tuning et du Conditionnement

Lucie Galland, Catherine Pelachaud, Florian Pecune

Étude en Cours : Évaluation de la Capacité des Modèles de Langage à Simuler le Discours de Patients lors de Thérapies: Impact du Fine-Tuning et du Conditionnement

Lucie Galland
lucie.galland@isir.upmc.fr
ISIR, Sorbonne université
Paris, France

Catherine Pelachaud
pelachaud@isir.upmc.fr
CNRS-ISIR, Sorbonne université
Paris, France

Florian Pecune
florian.pecune@u-bordeaux.fr
CNRS-SANPSY, Université de Bordeaux
Bordeaux, France

ABSTRACT

Le développement rapide des modèles de langue ouvre des portes à la recherche en modèle de dialogue. Cependant, évaluer la capacité de ces modèles à produire des réponses naturelles et cohérentes reste un défi majeur. Dans cette étude en cours, nous explorons l'effet du fine-tuning et du conditionnement sur les performances de réponse de Mistral, un modèle de langue de pointe et open-source, dans le contexte spécifique des thérapies délivrées par le biais d'agents virtuels. Nous développons un agent capable de mener des entretiens motivationnels contrôlés, en utilisant un patient virtuel pour évaluer l'efficacité de notre approche thérapeutique. Nous nous appuyons sur le corpus HOPE pour modéliser le comportement du patient virtuel, en utilisant Mistral7B instruct pour générer des réponses naturelles. Dans ce cadre, nous planifions de mener en premier lieu un pré-test pour déterminer la modalité la plus appropriée pour la collecte des données, soit texte ou audio, car les phrases générérées sont destinées à être délivrées à l'oral. Ce pré-test sera suivi d'une étude principale où les participants évalueront les réponses générées par Mistral dans différentes conditions. Ces travaux contribueront à améliorer notre compréhension de la capacité des modèles de langage à simuler le discours humain dans des contextes thérapeutiques, et fourniront des indices précieux pour le développement futur de tels systèmes.

CCS CONCEPTS

• **Human-centered computing → User studies**.

KEYWORDS

User study, NLG evaluation, Motivational interview, Simulated patient, Multimodality

1 INTRODUCTION

Dans le cadre de nos recherches visant à développer un agent capable de mener des entretiens motivationnels de manière contrôlée, nous nous concentrons actuellement sur la création d'un patient virtuel. L'objectif de ce patient virtuel est d'évaluer l'efficacité de notre approche thérapeutique, comme suggéré par Chiu et al. [3]. De plus, il pourra être utilisé pour pré-entraîner nos modèles afin de prévenir l'effet de départ à froid lors des interactions humain-agent. Pour ce faire, nous élaborons un patient virtuel dont le comportement est similaire à celui des patients présents dans le corpus HOPE [8], notamment en ce qui concerne les actes de dialogue. À cette fin, nous avons développé un premier modèle chargé de sélectionner le prochain acte de dialogue réalisé par le patient. Cet acte de dialogue est ensuite utilisé pour conditionner un modèle de langue qui générera un texte réalisant cet acte, tout en tenant compte du contexte. L'objectif de l'étude présentée ici est de valider l'utilisation de Mistral7B instruct [7] pour la production de phrases naturelles et cohérentes dans ce contexte spécifique.

2 TRAVAUX CONNEXES

L'évaluation de la génération de langage naturel est difficile car elle dépend grandement du contexte et de la tâche. De plus, les métriques automatiques ont tendance à être peu informatives et ne sont pas bien corrélées avec les évaluations humaines [10]. Cependant, de nombreux articles dans le domaine de la génération de langage naturel utilisent des métriques automatiques pour évaluer leur modèle [10]. De plus, les articles réalisant une évaluation humaine utilisent des métriques et des questionnaires différents. Récemment, des travaux ont été réalisés pour fournir des lignes directrices pour les évaluations [4, 6, 10]. Ils conseillent de bien définir les concepts à évaluer ainsi que de créer des questionnaires unifiés. Dans le domaine des agents virtuels, Fitrianie et collègues [5] ont proposé une unification des questionnaires et ont défini 19 concepts et des questionnaires associés pour évaluer les agents virtuels. Dans ce travail, nous prenons en compte ces efforts pour évaluer notre modèle de langue.

3 OBJECTIF DE L'ÉTUDE

L'objectif principal de cette étude est d'évaluer la capacité du modèle de langue Mistral7B instruct à générer des réponses simulées qui imitent le langage naturel d'un patient participant à des thérapies, en réponse à un contexte donné. Ces réponses doivent présenter un caractère naturel, un style oral et respecter les consignes de comportement spécifiées (conditionnement en acte de dialogue). Pour atteindre cet objectif, nous examinerons quatre points particuliers. Tout d'abord, dans un pré-test, nous examinerons la possibilité d'évaluer des textes écrits émanant d'interviews oraux. Ensuite, dans notre étude principale, nous chercherons à déterminer si le modèle Mistral est capable de simuler un patient lors d'une entrevue motivationnelle. De plus, nous évaluerons si le fine-tuning de Mistral sur l'ensemble de données HOPE permet d'améliorer la qualité des réponses générées. Enfin, nous étudierons l'impact du conditionnement du contenu de la réponse sur la qualité des réponses produites. Pour répondre à ces objectifs, nous formulons quatre questions de recherche :

- **RQ0 :** Est-ce que les phrases générées reproduites par un synthétiseur vocal ou écrites sont perçues similairement ?
- **RQ1 :** Est-ce que le fine-tuning de Mistral produit des réponses plus naturelles (**a**) et cohérentes (**b**) que Mistral ?
- **RQ2 :** Est-ce que le conditionnement en actes de dialogues produit des phrases moins naturelles (**a**) et cohérentes (**b**) qu'en l'absence de conditionnement ?
- **RQ4 :** Est-ce que les modèles Mistral produisent des réponses aussi naturelles (**a**) et cohérentes (**b**) que de vrais patients ?

Nous avons décidé d'utiliser Mistral7B instruct [7] car c'est un modèle open-source performant, spécialement entraîné pour respecter des instructions (dans notre cadre, le conditionnement en acte de dialogue). Pour répondre à nos questions de recherche, nous étudierons cinq conditions différentes : Ground Truth (GT), Mistral non-conditionné (M), Mistral conditionné (Mc), Mistral fine-tuned non-conditionné (fM), Mistral fine-tuned conditionné (fMc), avec deux variables indépendantes : Conditionnement et Fine-tuning, et deux variables dépendantes : Naturel et Cohérence.

4 MESURES

Un questionnaire visant à unifier l'évaluation des agents basés sur 19 concepts a été proposé [5]. Nous avons sélectionné les 2 concepts qui s'appliquent à notre étude : le naturel et la cohérence du comportement et avons adapté le questionnaire associé. Les participants évalueront leur accord avec les énoncés suivants sur une échelle de Likert à 7 points. Pour le naturel : "La phrase aurait pu être produite par un humain" et "La phrase ne ressemble pas à celle d'un humain', et pour la cohérence : "La phrase s'intègre harmonieusement dans le contexte environnant" et "La phrase n'a pas de sens".

5 CONCEPTION

5.1 Pré-test

Dans un premier temps, nous prévoyons de réaliser un pré-test afin de sélectionner la modalité à utiliser pour délivrer le stimulus (phrase générée par le modèle de langue) dans l'étude principale. En effet, le texte que nous souhaitons évaluer est une transcription d'un dialogue oral. Les phrases produites oralement ne sont souvent pas grammaticalement correctes, ce qui peut entraver leur compréhension une fois écrites [9]. Le langage écrit et oral ont également de nombreuses différence de structure car le langage écrit est fait pour être lu quand le langage oral est fait pour être entendu [1]. Cet argument tendrait à favoriser l'utilisation de la modalité audio pour notre étude. Cependant, les phrases générées par notre modèle seront transcrites dans la modalité audio par un synthétiseur vocal (TTS). Nous utilisons un TTS pour contrôler la voix utilisée et étudier toutes les phrases (y compris les phrases générées par Mistral) avec la même voix. Cependant, l'utilisation du TTS ajoute des informations telles que la prosodie qui peuvent influencer la perception des participants de la phrase générée par les participants à la place de la phrase générée par notre modèle de langue. Pour ces raisons, nous réalisons ce pré-test afin de répondre à **RQ0**. Si les évaluations des transcriptions sont corrélées avec les évaluations de l'audio, nous pourrons ensuite utiliser la modalité texte et ne pas prendre en compete les effets du TTS sur les notations. Pour cela, nous sélectionnons 30 phrases aléatoirement ainsi que leur contexte (2 tours de parole précédents) que nous transformons en audio en utilisant la TTS Bark[2], qui est notamment capable de réaliser des hésitations. Soixante participants, répartis en 2 groupes (Texte et Audio), évaluent 15 phrases différentes choisies aléatoirement parmi les 30 possibles. Le naturel et la cohérence sont évaluées en utilisant les mesures présentées précédemment. Cette comparaison nous permettra de déterminer la modalité à utiliser dans l'étude principale.

5.2 Étude Principale

Le but de l'étude principale est de répondre à nos questions de recherche **RQ1, RQ2, RQ3**. En utilisant la modalité déterminée lors du pré-test, les participants évalueront 15 phrases dans différentes conditions, sélectionnées de manière aléatoire. Chaque participant évaluera toutes les conditions, mais pour des phrases différentes. Nous prévoyons de tester les hypothèses suivantes :

- **H1 :** Nous n'observerons pas de différences significatives en termes de pertinence et de fluidité entre (GT),(M) et (Mc). Pour cela, nous utiliserons un test d'équivalence.
- **H2 :** Nous observerons des différences significatives en termes de pertinence et de fluidité entre (M) et (fM) avec (M) > (fM) et (Mc) et (fMc) avec (Mc) > (fMc). Pour cela, nous utiliserons un test de mixed ANOVA.
- **H3 :** Nous n'observerons pas de différences significatives en termes de pertinence et de fluidité entre (M) et (Mc), et entre (fM) et (fMc). Pour cela, nous utiliserons un test d'équivalence.

6 CONCLUSION

Nous présentons ici notre plan d'étude visant à déterminer dans un premier temps si la modalité texte ou audio peut être utilisée pour évaluer le naturel et la cohérence des transcriptions de dialogues oraux. Dans un second temps, nous étudierons l'impact du fine-tuning ainsi que du conditionnement des modèles de langage sur le naturel et la cohérence des phrases produites en utilisant la modalité choisie dans le pré-test. Cela nous permettra de déterminer si ces modèles sont utilisables pour simuler un patient virtuel lors d'une interview motivationnelle, et ensuite les utiliser pour entraîner et tester nos modèles de langage dans le cadre de la création d'un agent capable de réaliser une interview motivationnelle de manière contrôlée.

7 REMERCIEMENTS

Ces travaux ont été partiellement financés par les projets ANR-DFG-JST Panorama et ANR-JST-CREST TAPAS (19-JSTS-0001-01).

REFERENCES

[1] Wallace Chafe and Jane Danielewicz. 1987. Properties of spoken and written language. In *Comprehending oral and written language*. Brill, 83–113.
[2] P.W.D. Charles. 2024. Bark. https://github.com/suno-ai/bark.
[3] Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A Computational Framework for Behavioral Assessment of LLM Therapists. *arXiv preprint arXiv:2401.00820* (2024).
[4] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 1383–1392.
[5] Siska Fitrianie, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. 2020. The 19 unifying questionnaire constructs of artificial social

agents: An iva community analysis. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[6] Dimitra Gkatzia and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005-2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*. 57–60.

[7] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[8] Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. 735–745.

[9] Jim Miller. 1995. Does spoken language have sentences. *Grammar and meaning* (1995), 116–135.

[10] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language* 67 (2021), 101151.