PhD Research Proposal Form
China Scholarship Council (CSC)
2025

A remplir en français ou en anglais en fonction de la langue qui sera utilisée pour la thèse

FIELD
Sciences humaines et sociales

(eg: Mathematics, Physics, Sociology, ….)

Thesis subject title: *Exploration de la génération de récits complexes par les grands modèles de langue à travers le prisme de la mémoire épisodique*

Name of the French doctoral school/Ecole doctorale: ED540 (Lettres, Arts, Sciences humaines et sociales)

Name of the Research team/Equipe de recherche: Lattice UMR8094
Website: https://www.lattice.cnrs.fr

Name of the Supervisor/Directeur de thèse: Pascal Amsili & Olga Seminck
Email: Pascal.Amsili@ens.fr olga.seminck@cnrs.fr

Lab Language/ Langue de travail: Français et/ou Anglais

Research Proposal Abstract/Présentation du sujet:

Les grands modèles de langue (Large Language Models LLMs) ont montré une capacité impressionnante à générer du texte, y compris des récits. Cependant, les récits générés par les LLMs présentent souvent des lacunes par rapport aux récits humains, notamment en termes de complexité, de suspense et de profondeur émotionnelle. La mémoire épisodique, qui permet aux humains de se souvenir d'événements personnels vécus dans un contexte spatio-temporel spécifique, joue un rôle clé dans la narration humaine en facilitant l'intégration d'expériences passées et la construction de récits cohérents et émotionnellement engageants. L'objectif de ce projet est d'explorer comment améliorer les LLMs pour qu'ils soient capables de produire des récits dans lesquels les personnages disposent d'une mémoire épisodique, influençant leurs actions et interactions en fonction de leurs souvenirs. Ce projet vise notamment à accroître la qualité et la complexité des récits générés, en renforçant leur structure narrative, leur profondeur émotionnelle et leur cohérence temporelle.

Les méthodes envisagées pour ce projet de recherche débuteront par l'annotation des données littéraires, en particulier l'identification des passages où la mémoire épisodique joue un rôle clé dans la narration. Cela inclura le développement d'un guide d'annotation détaillé pour assurer une

identification précise et cohérente de ces passages. En parallèle, nous concevrons une métrique d'évaluation permettant de mesurer dans quelle mesure la mémoire épisodique influence les récits, en quantifiant l'importance de ces éléments dans la structure narrative.

Pour les textes, nous travaillerons avec des corpus tels que *Chapitres*, qui offrent une richesse de récits littéraires classiques et modernes, permettant d'identifier un large éventail de passages où la mémoire épisodique est mise en jeu. Une fois l'annotation réalisée, nous utiliserons ces données pour créer des données structurées sous forme de fiches synthétiques, extraites et résumées à partir des passages identifiés. Pour cette tâche, nous pourrons utiliser des outils comme *French BookNLP* afin de faciliter l'extraction d'informations pertinentes et la structuration des données, par exemple l'identification des personnages.

Ensuite, plusieurs expérimentations seront menées pour explorer différentes approches de génération de récits. Nous commencerons par des expériences de *prompt engineering*, en ajustant les invites utilisées pour guider la génération de texte par le LLM. Par la suite, nous testerons des approches de *fine-tuning*, en entraînant le modèle uniquement sur le texte des passages identifiés, afin de renforcer sa capacité à intégrer la mémoire épisodique dans la génération de récits. Enfin, nous évaluerons l'impact de la combinaison des données textuelles annotées et des fiches synthétiques, pour enrichir le modèle de la mémoire épisodique de manière plus structurée.

Les résultats obtenus seront évalués à l'aide de la métrique d'évaluation développée précédemment, afin de mesurer l'amélioration de la qualité et de la complexité des récits générés en fonction de l'intégration de la mémoire épisodique.

References:

Fountas, Z., Benfeghoul, M. A., Oomerjee, A., Christopoulou, F., Lampouras, G., Bou-Ammar, H., and Wang, J. (2024). Human-like episodic memory for infinite context llms. arXiv preprint arXiv :2407.09450.

Georgiou, A., Can, T., Katkov, M., and Tsodyks, M. (2023). Using large language models to study human memory for meaningful narratives. bioRxiv, pages 2023–11.

A. Leblond, Corpus chapitres, 2022. doi:10.5281/zenodo.7446728.

Mélanie-Becquet, F., Barré, J., Seminck, O. C., Plancq, C., Naguib, M., Pastor, M., and Poibeau, T. (2024). Booknlp-fr, the french versant of booknlp. a tailored pipeline for 19th and 20th century french literature. In Conference on Computational Literary Studies (CCLS 2024).

Piper, A., So, R. J., and Bamman, D. (2021). Narrative theory for computational narrative understanding. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 298–311.

Tian, Y., Huang, T., Liu, M., Jiang, D., Spangher, A., Chen, M., May, J., and Peng, N. (2024). Are large language models capable of generating human-level narratives ?
Tulving, E. (2002). Episodic memory : From mind to brain. Annual review of psychology, 53(1) :1–25.

Zhang, X., Seminck, O., and Amsili, P. (2024). Remember to forget : A study on verbatim memorization of literature in large language models. In Proceedings of the 2024 Conference on Coputational Humanities Research.

Type of PhD :

1.Full PhD

- Joint PhD/cotutelle (leading to a double diploma) : ~~YES~~ or **NO**
- Regular PhD (leading to a single French diploma) : **YES** or ~~NO~~

~~2. Visiting PhD (students enrolled at a Chinese institution who come to ENS for mobility period) :~~ ~~YES~~ or **NO**

PLEASE SEND THE DOCUMENT TO
Direction des Relations internationales : dri@ens.psl.eu