PhD Proposal :

Reinforcement Learning with Human Feedback in the medical field

LIPN, UMR CNRS 7030

1 Context and objectives

Patient profiles describe a comprehensive summary of health-related information for an individual patient. This includes personal information (age, ethnicity, sex), a thorough record of clinical data (medications, side effects, allergies, adverse reactions), a comprehensive medical history, omics data, laboratory results as well as current health status. This paradigm shift from disease centered to patient-centered care will have a profound impact in medicine. To achieve this mission, our strategy consists in the development of cutting-edge machine learning (ML) techniques including large language models (LLMs) based on Transformers architecture [10, 9] and reinforcement learning with human feedback (RLHF) [3, 13, 2]. These algorithms ingrained in large language modeling are capable to seamlessly incorporate data from various sources. In this Phd proposal, we aim to use generative models that explore the potential of large language models (LLMs) in order to achieve efficient classification in the medical field. The main objective of this Phd proposal consists in designing a digital twin patient model that has the ability to:

1. Generate patient profiles using a large language model generator. Complete patient profiles will provide valuable information for establishing baseline vital signs and laboratory values. Monitoring changes in these parameters can be indicative of the pathology, prompting timely interventions. In this first objective, the twin patient model will be driven by a generative model, which will be a large language model (LLM) that will be able to disclose (or uncover) new patient profiles from a preliminary existing subset of patient clinical profiles. The prompt is a multimodal piece of information that invokes a sequence of existing profiles. This sequence is fed to the Conditional Generative Adversarial Networks, CGAN-based transformer in order to generate another multimodal piece of information that extends the set of the original patient profiles.

2. Predict patient responsiveness to the specific treatments at different point in time given observation (or monitored) data. In other words, this module will be supplied to track the effectiveness of the treatment at different stages of the pathology growth. In this second objective, we invoke Reinforcement Learning with Human Feedback (RLHF) since it is the most appropriate model to apply.

2 State-of-the-art

Generative models and attention-based transformer models have recently gained significant popularity. The success of the paper "attention is all you need" [15] marked the beginning of a new

era in deep learning. Generative Adversarial Networks (GAN) [5] is a machine learning framework where two networks, the generator and the discriminator, are competing against each other during training, hence the "adversarial" part. On one side, the generator learns to generate synthetic samples close to the original distribution without actually being able to see the training data. In the generative process, random noise vectors $\mathbf{z}$ are first sampled from a prior distribution $p_{\mathbf{z}}$ before being transformed by the generator network. This noise injection introduces variability and diversity in the generated samples, allowing the generator to produce a wide range of outputs. On the other side, the discriminator learns to discriminate between the real training data and the synthetic samples. The adversarial aspect reads as follows: the discriminator (D) aims to maximize the distance between the real and generated data distributions because it has an interest in being able to discriminate them. The generator network (G) aims to minimize the distance between the real and generated data distributions by making the generated data look as authentic as possible, in order to fool the discriminator. Conditional Generative Adversarial Networks (CGAN) [11] is an extension of the vanilla GAN [5]. It allows a certain degree of control over the generated samples by setting a condition the generation must meet. This condition is generally incorporated in the form of a vector $\mathbf{y}$ concatenated with the noise vector $\mathbf{z} \sim p_{\mathbf{z}}$ at the input of the generator G and provided as an input to the discriminator D. Thus, this action corresponds to conditioning the distributions of D and G on this vector concatenation.

Generative models, particularly Large Language Models (LLMs), showcase a remarkable capability to handle multimodal medical data, seamlessly bridging the gap between different types of information such as text, strings of bits, and images. These models leverage sophisticated architectures, like transformer-based structures, that are inherently versatile in capturing complex relationships within diverse data modalities. This flexibility is particularly evident in applications where medical data comes in various forms, such as in natural language understanding, image captioning, or even generating content that blends both text and visuals. By comprehensively understanding and modeling the intricate interplay between modalities, LLMs offer a powerful tool for tasks that demand a holistic approach to multimodal data, showcasing their adaptability and effectiveness in handling the complexities inherent in real-world, heterogeneous datasets. The multimodal data representation has emerged as a pivotal approach for integrating diverse forms of data. In the literature, there are methods such as CLIP [12], and GPT 4 [7] that offer a solution for multimodal data representation but use different encoders to extract information from the image and text before finding a common projection space. Large Language Models (LLMs) and Generative Pre-training Transformer (GPT-3), which is at the heart of ChatGPT, Llama-2-7b [14], Phi-2 [18], Mistral-7b [6], Falcon-7b [1], Meditron-7b [4], BioBERT [8] GatorTron [17] are examples of generative models that have already been developed and support our choice of generative models.

3 Research axes

Often, certain variables in the patient profile are not available, necessitating the completion of patient profiles. In this context, using generative models, especially with multimodal data, is a significant challenge. The inclusion of knowledge from medical professionals can enhance the learning process, leveraging their expertise to enrich the models based on reinforcement learning. In the context of this Phd proposal, we propose to explore the following research axes:

3.1 Axis 1: Generation of profile sequences

The model that we want to develop is driven by a CGAN-based transformer. It combines the architecture of a transformer, known for sequence-to-sequence tasks, with the principles of Conditional Generative Adversarial Network for generating targeted multimodal data profiles data. This model accepts a preliminary subset of patient profile sequences and discloses a new set of profile sequences. This new set is supplemented to the initial set until a certain amount of patient profiles are generated. This aligns with recent advancements in multimodal generative models where these latter are designed to comprehend and generate information across diverse modalities [16, 12]. The aim of these models, consists in creating a numerical representation (embedding) for each profile within a sequence; encapsulating essential information about the profile and its neighbouring context. Combining this CGAN with a language model invokes the cognitive attention mechanism that captures complex profiles dependencies in parallel, and quantifies profiles importance by assigning them soft weights (attention weights). Large language models have already shown promise in various technological areas, however, their exploitation in health science remains scarce. This is one of the most compelling motivations among others behind this proposed approach.

3.2 Axis 2: Prediction using Reinforcement Learning and Human Feedback

We invoke Reinforcement Learning with Human Feedback (RLHF) since it is the most appropriate model to apply. RLHF combines reinforcement learning, where an agent (patient profiles generation but trained differently) learns by interacting with an environment (dynamics and reward), *with input from human feedback who are medical professionals.* The goal is to leverage human expertize and guidance to improve the RLHF model learning process, especially in situations where obtaining a reward signal directly from the environment is challenging or impractical [13]. These signals can be for example: (i) doctors receive a high positive (or a negative) reward if the patient's health improves (or deteriorates) due to the treatment or intervention, or (ii) the algorithm rewards the doctor for accurate and timely diagnoses, especially in cases where early detection positively impacts the patient's outcome. The human feedback is used to construct a reward model, which the agent uses to guide its learning. We do not want to keep medical professionals in the loop each time a response is provided, therefore an offline database generated via the interaction between the LLM and medical doctors annotations is made available. This database could contain patient records, treatment outcomes, and other relevant information. This action ensures that the objectives or decisions of a reinforcement learning algorithm align with the goals and values of healthcare professionals. The reward model helps the agent understand which actions are more likely to lead to desirable outcomes based on caregivers input.

3.3 Tentative planning

- **Milestone 1:** We estimate an approximate duration of 12 months
  - State of the art and specification of a proposal for Axes 1, 2
  - Implementation of the first version of CGAN and LLM based on medical Dataset
  - Scientific publication
- **Milestone 2:** We estimate an approximate duration of 8 months
  - Development of a RHLF-LLM based on the first version produced in the first milestone. Generalization performance and robustness testing of models

  - Scientific publication

- **Milestone 3:** We estimate an approximate duration of 8 months
  - Integrate an offline database generated via the interaction between the LLM and medical doctors annotations is made available. Generalization performance and robustness tests
  - Scientific publications

- **Milestone 4:** Organization and dissemination of the code in open source and writing of the dissertation. We estimate an approximate duration of 8 months.

3.4 Co-Supervisors

-Thesis Director : Hanane Azzag (Associate Professor, HdR) from LIPN UMR CNRS 7030
-Co-supervisor: Zaineb Chelly Dagdia (Associate Professor, HdR, University of Paris-Saclay, Versailles Campus (UVSQ)) and a member of the DAVID laboratory, ADAM team.

References

[1] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, Étienne Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, and G. Penedo. The falcon series of open language models, 2023.

[2] M. G. Azar, M. Rowland, B. Piot, D. Guo, D. Calandriello, M. Valko, and R. Munos. A general theoretical paradigm to understand learning from human preferences, 2023.

[3] S. S. Bo An and R. Wang. Deep reinforcement learning for quantitative trading: Challenges and opportunities. *IEEE Intelligent Systems*, vol. 37, no. 2, pp. 23-26, 2022.

[4] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M.-A. Hartley, M. Jaggi, and A. Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[6] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023.

[7] A. Koubaa. Gpt-4 vs. gpt-3.5: A concise showdown. 2023.

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, Sept. 2019.

[9] A. Madane, M. D. Dilmi, F. Forest, Hanane Azzag, Mustapha Lebbah, and J. Lacaille. Transformer-based conditional generative adversarial network for multivariate time series generation. In *International Workshop on Temporal Analytics. The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2023.

[10] A. Madane, F. Forest, Azzag, Hanane, Lebbah, Mustapha, and J. Lacaille. One-pass generation of multivariate time series through conditional multivariate modeling. In *IEEE World Congress on Computational Intelligence (IEEE WCCI 2024)*, Pacifico Yokohama, Yokohama, Japan, 2024.

[11] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[12] A. Radford, J. W. Kim, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[13] H. Sun. Reinforcement learning in the era of llms: What is essential? what is needed? an rl perspective on rlhf, prompting, and beyond, 2023.

[14] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

[15] A. Vaswani and al. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[16] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua. Next-gpt: Any-to-any multimodal llm, 2023.

[17] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, C. A. Harle, G. Lipori, D. A. Mitchell, W. R. Hogan, E. A. Shenkman, J. Bian, and Y. Wu. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records, 2022.

[18] Y. Zhu, M. Zhu, N. Liu, Z. Ou, X. Mou, and J. Tang. Llava-phi: Efficient multi-modal assistant with small language model, 2024.