Large language models and explainability for healthcare

Oriana Presacan
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
oriana.presacan@stud.etti.upb.ro

Bogdan Ionescu
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
bogdan.ionescu@upb.ro

***Abstract*—This paper outlines my current research on applying machine learning, explainable artificial intelligence, and large language models to healthcare. First, I present a deep-learning pipeline for autophagy analysis that couples YOLOv8 detection, U-Net++ segmentation, a vision-transformer classifier, and class-activation maps to track and interpret autophagic events in fluorescence-microscopy images. Second, I investigate how decoding strategies—greedy, beam, and stochastic sampling—affect factual accuracy and clinical usefulness in domain-adapted large language models, providing guidance for safe deployment on medical benchmarks. Together, these studies highlight how xAI can both clarify cellular dynamics and tune language-model outputs for healthcare applications.**

***Index Terms*—AI, LLM, explainability, machine learning**

I. Introduction

Artificial intelligence (AI) has moved from research prototypes to a general-purpose capability that undergirds search, recommendation, language interfaces, and code generation. This shift was catalyzed by deep learning's representation power [1] and, more recently, by the transformer architecture, whose self-attention enables efficient modeling of long-range dependencies in text [2]. Within this paradigm, LLMs trained on web-scale corpora exhibit strong in-context learning and zero-/few-shot generalization [3]. Empirically, model performance improves predictably with increases in compute, data, and parameters [4], though data-efficient scaling highlights the importance of high-quality tokens over sheer parameter count [5]. Beyond pretraining, alignment techniques such as instruction tuning further adapt models to human preferences and task specifications [6].

At the same time, current LLMs remain probabilistic next-token predictors with limitations that matter in practice. They can produce fluent but incorrect content ("hallucinations") and struggle with robustness and calibration [7]. Data provenance and bias raise concerns about fairness, transparency, and potential harms at scale [8], while training and inference costs have environmental implications [9]. Mitigations include retrieval-augmented generation to ground outputs in external sources [10], preference-based optimization, and more rigorous, task-relevant evaluation protocols. Thus, the present landscape combines rapidly advancing capability with active work on reliability, efficiency, and governance—aimed at making LLMs not only more powerful but also safer and more useful in real-world deployments.

II. Current Work

The section summarizes some of my current research projects undertaken as part of my first PhD year.

A. Deep learning and XAI for Autophagy

Autophagy is an intracellular "clean-up" mechanism that preserves cellular equilibrium by breaking down and repurposing damaged organelles and macromolecules. Accurately quantifying this process is essential but its rapid, multifaceted behavior makes manual microscopy analysis impractical. We created an automated pipeline that applies modern deep-learning techniques to fluorescence-microscopy images of. The workflow combines object detection, segmentation, state classification, cell tracking, and interpretability tools. A bespoke tracker follows cells through division and shape shifts without needing labeled trajectories. To make model reasoning transparent, we employ class-activation maps and t-SNE embeddings, while domain biologists confirmed the biological plausibility of the outputs. The results illustrate how explainable AI can reduce labor, improve insight, and advance autophagy research.

B. LLM Decoding in Healthcare

LLMs rely on decoding algorithms (rules that determine which token to select next) to turn their learned probabilities into full sentences. Greedy decoding always selects the highest-probability token, producing fluent yet often repetitive text. Beam search explores several high-scoring sequences in parallel, improving global coherence but frequently yielding safe, generic wording. Stochastic approaches such as top-k sampling, nucleus (top-p) sampling, and temperature scaling inject controlled randomness, boosting diversity at the cost of occasional factual drift. In our study, we measure how each strategy affects accuracy on medical-domain LLMs and benchmarks, providing guidance on choosing decoding parameters for health-care applications.

*2^{nd} Workshop on Artificial Intelligence for Multimedia,
AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest,
Bucharest, Romania, 15 July, 2025*

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. *et al.*. Askell, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[4] J. Kaplan, S. McCandlish, T. B. Brown, A. Radford, R. Child, J. Wu, and D. Amodei, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.

[5] J. Hoffmann, A. Botev, M. Fairbank, D. De las Casas, E. Buchatskaya, L. Hendricks, and A. *et al.*. Zhai, “Training compute-optimal large language models,” *arXiv preprint arXiv:2203.15556*, 2022.

[6] L. Ouyang, J. Wu, X. Jiang, D. Almeida, S. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Kapoor, and K. *et al.*. Slama, “Training language models to follow instructions with human feedback,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[7] Z. Ji, Y. Lee, J. Fries, T. Yu, D. Su, D. Xu, P. Henderson, M. Witbrock, D. Weld, and J. Han, “A survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[8] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021, pp. 610–623.

[9] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in nlp,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3645–3650.

[10] P. Lewis, E. Perez, A. Piktus, V. Karpukhin, N. Goyal, F. Petroni, A. Rogers, M.-A. Simion, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.