Memorization and Expressivity in Transformers: A Learning-Theoretic Perspective

Maxime Meyer

Department of Mathematics, National University of Singapore, Singapore, 117543
IPAL, IRL2955, Singapore
maxime.meyer@u.nus.edu

Abstract

Transformers have reshaped modern artificial intelligence, yet their theoretical foundations remain incomplete. This thesis investigates the approximation power and memory limitations of transformers. I combine tools from approximation theory and statistical learning theory to provide provable guarantees on expressivity, memorization capacity, and inherent architectural constraints. My contributions include the first rigorous proof of memory bottlenecks in prompt tuning and new results on the expressivity of transformers. The long-term goal of my doctoral research is to develop a principled theoretical framework that grounds the empirical behavior of large-scale transformer models in formal approximation-theoretic results.

Introduction

Transformers are among the most successful architectures in artificial intelligence, achieving state-of-the-art results across natural language processing, vision, and many other domains. Despite this empirical success, their theoretical foundations remain incomplete. Fundamental questions remain unanswered: What functions can transformers approximate? How much information can they reliably memorize from long contexts? And what are their inherent limitations as learning systems?

In the remainder of this summary, I describe my progress to date, beginning with research in learning theory that I further developed at the start of my PhD, and then turning to my main line of work on the theory of transformers.

Previous Research and Foundations

At the start of my doctoral studies in February 2025, I continued to develop work on *online learning*, a framework where a learner makes sequential predictions in rounds, receives feedback, and updates its hypothesis with the goal of minimizing *regret* compared to the best fixed strategy in hindsight. Unlike classical batch learning, online learning captures adversarial and dynamic environments, making it a powerful tool for studying the limits of sequential decision-making.

One of my first projects as a PhD student was to substantially revise and extend a manuscript on *quantum state learning*, which I had initiated during my Master's studies. Classical wisdom suggests that pure states—which can be represented by a single vector—are easier to learn than mixed states, which require a full density matrix description. In the online learning framework, however, I proved that this intuition fails: pure and mixed states are equally difficult to learn. Specifically, I showed that both classes share essentially the same *sequential fat-shattering dimension*, leading to identical regret scaling (Meyer et al. 2025a). While my coauthors helped me identify the broad research area and provided feedback on the manuscript, I was responsible for selecting the focus on pure versus mixed states, developing the proofs, and rewriting and strengthening the paper so that it was accepted at NeurIPS 2025. Soumik Adhikary contributed specifically to the analysis of smoothed online learning in Section 5.2.

In addition, just prior to the start of my PhD I began co-supervising an undergraduate student on a project about *online learning to defer*, a framework in which a predictive model can either make its own prediction or defer to one of several external experts. The central challenge is to learn a deferral strategy that minimizes long-term loss by balancing autonomy with selective reliance on experts of varying reliability. Our contribution is to place this problem in the *online learning* setting. In this formulation, the learner not only chooses whether to predict or defer, but also adapts to a potentially varying number of experts over time, making the analysis significantly richer than in the static case. My co-supervisor provided expertise in deferral models, while I contributed the online learning perspective. The resulting work, now under submission to AISTATS, illustrates both the theoretical depth of deferral in dynamic environments and my early involvement in mentoring and collaboration during my PhD.

Research Progress to Date

Since starting my PhD, my research has focused on the *theoretical foundations of transformers*. My first major project investigates *prompt tuning*, a technique that has achieved widespread empirical success in adapting pretrained language models to new tasks, but remains poorly understood theoretically. In my paper *Memorization Limitations*

of Prompt Tuning in Transformers (Meyer et al. 2025b), I analyzed the memorization capability of transformers under prompt tuning and established two central results. First, I proved that the amount of information a transformer can memorize increases at most linearly with the length of the prompt. Second, I provided the first rigorous explanation of a phenomenon widely observed in practice: the degradation of performance in long contexts. In particular, I showed that transformers possess an inherent memory limitation regarding the amount of new information they can learn from a prompt. My main contributions in this project were to formalize the problem setting, develop the theoretical proofs, and write the majority of the manuscript, while my coauthors assisted with refining the formulation and proofreading. This work, completed during the first months of my PhD, has been submitted to AAAI and is currently under review. It offers formal guarantees on the capabilities and limitations of transformers, grounding their empirical behavior in provable theoretical foundations.

I am now extending this research to investigate the *expressivity* of transformers. In ongoing work (to be submitted to ICML 2026, and in the meantime to the Math4AI workshop at AAAI), I develop new approximation-theoretic tools, including covering and packing number arguments, to characterize the classes of functions representable by transformers of given depth and width. While one of my collaborators contributes to the experimental validation, my primary role was to identify the topic, design the theoretical framework, and develop the proofs.

Together, these projects define the central trajectory of my PhD: uncovering the mathematical principles that govern transformers. By proving formal limitations (prompt tuning (Meyer et al. 2025b)), establishing approximation bounds (expressivity, in progress), and linking them to empirical observations, my research contributes to building a principled theoretical framework for modern AI systems. This work not only deepens our theoretical understanding but also lays the groundwork for interpretable and trustworthy large-scale models.

Anticipated Progress by the Workshop Date

Between now and the Doctoral Consortium in January 2026, I expect to consolidate and extend the first phase of my PhD research. I will present my NeurIPS 2025 paper on online quantum state learning and continue to build on the techniques developed there, which I believe can inform new directions in the analysis of sequential models. In parallel, I will finalize my ongoing paper on the expressivity of transformers, which will first be submitted at the Math4AI workshop at AAAI, and then to ICML 2026. I will also follow through with the AISTATS submission on online learning to defer, on which I serve as co-supervisor. Finally, I have begun working on a new project in transformer theory that grows out of the limitations identified in my prompt tuning paper. By the time of the Consortium, I aim to have advanced this project to the point of a first draft, marking the next major step in my dissertation research.

Conclusion and Future Research Plan

Looking beyond the immediate milestones, the central objective of my PhD is to uncover the mathematical principles that govern the transformer architecture. My ongoing and completed projects on memory limitations and expressivity represent the first steps toward this goal. In the next stage of my dissertation, I aim to develop a comprehensive approximation-theoretic framework that quantifies what transformers can and cannot represent, how their structure constrains their ability to generalize, and how their empirical behavior reflects these theoretical limits. By pursuing this direction, I hope not only to deepen our theoretical understanding of large-scale sequence models but also to lay the foundations for interpretable and trustworthy AI systems whose limitations are rigorously characterized. In this way, my research aspires to bridge the gap between abstract theory and the practical challenges posed by modern language models, ensuring that the next generation of AI is grounded in firm mathematical principles.

At a broader level, my research also aspires to uncover the hidden laws that govern language itself. Transformers provide a natural setting for this pursuit: when language is viewed as a random process in their embedding space, these models implicitly encode the statistical structure underlying human communication. By rigorously analyzing this space, I hope to reveal the equations that shape linguistic patterns. Such an understanding could lead to breakthroughs in machine learning theory, particularly in building AI systems whose decisions are transparent and reliable.

References

Meyer, M.; Adhikary, S.; Guo, N.; and Rebentrost, P. 2025a. Online Learning of Pure States is as Hard as Mixed States. arXiv:2502.00823.

Meyer, M.; Michelessa, M.; Chaux, C.; and Tan, V. Y. F. 2025b. Memory Limitations of Prompt Tuning in Transformers. arXiv:2509.00421.