Towards Robust Human–AI Decision-Making via Learning-to-Defer

Yannis Montreuil

National University of Singapore
School of Computing
15 Computing Drive, Singapore 117418
yannis.montreuil@u.nus.edu

Abstract

AI systems often fail on challenging or out-of-distribution inputs—a critical limitation in domains such as healthcare, finance, and autonomous driving. Learning to Defer (L2D) addresses this by training models not only to predict but also to decide when to defer to external experts. This thesis develops a unified and robust framework for L2D that advances its theoretical foundations, reliability, and applicability. It characterizes Bayes-optimal routing policies, establishes surrogate-consistency guarantees, and introduces a unified adversarial framework for attacking and defending L2D with Bayes-optimal robustness. It further proposes the first top-k deferral methods in both two-stage and one-stage settings. Empirical studies validate these ideas in multi-task learning and extractive question answering with large language models. Ongoing work explores token-level routing in LLMs, online adaptation with dynamic experts, and partial deferral.

Introduction

Artificial intelligence has achieved remarkable success across vision, language, and healthcare. Yet even state-of-the-art systems remain *unreliable*: for certain inputs, they fail predictably. In safety-critical domains—such as medical diagnosis or autonomous driving—even a single mistake may carry unacceptable consequences. Responsible deployment thus requires mechanisms that ensure *instance-level reliability*, not just high average accuracy.

A natural remedy is to include humans in the decision loop: the system must *learn when to act autonomously and when to defer to an expert*. This *routing problem*—allocating each input to the most reliable agent—lies at the core of Human–AI collaboration (Madras, Pitassi, and Zemel 2018).

Learning to Defer (L2D) offers a principled solution: a model jointly learns (i) a predictive function and (ii) a deferral policy deciding when to delegate. Unlike heuristic confidence thresholds, L2D explicitly optimizes the trade-off between model performance and expert reliance, promoting reliability across the full decision space.

However, existing L2D methods remain limited. They lack *reliability guarantees*, often failing to ensure trustworthy outcomes for all routed decisions; they are not *robust* to distribution shifts or adversarial inputs; and their *statistical foundations*—consistency and calibration—are poorly understood. Most approaches are also static, ignoring *sequential dynamics* where data and expert behavior evolve over time.

This work develops a *unified and robust framework for Learning to Defer*, addressing these shortcomings by:

1. Establishing **theoretical foundations**: Bayes-optimal routing and consistent surrogate objectives.
2. Enhancing **reliability and robustness**: under shift, corruption, and heterogeneous experts.
3. Extending **applicability**: from static classification to sequential and large-scale decision settings, including healthcare and language models.

Current Research

Learning-to-Defer extends selective prediction (Chow 1970) by allowing models to delegate uncertain inputs to external experts (Madras, Pitassi, and Zemel 2018). Despite its promise as a routing mechanism for Human–AI systems, current formulations often lack reliability guarantees, provide limited statistical foundations, assume deferral to a *single expert*, and degrade under distributional shift or adversarial perturbations.

To address these gaps, I first focused on *robustness*. In my recent work (Montreuil et al. 2025a) published at ICML25, I introduced the first framework to both *attack* and *defend* L2D systems. I proved theoretical guarantees showing that defended systems recover the Bayes-optimal policy even under adversarial perturbations, and demonstrated empirically that robustness is crucial in practice, where malicious actors may attempt to redirect queries to weaker agents, inflate costs, or overload human experts.

While robustness secures L2D against adversaries, another limitation is the assumption of a single "best" expert. In many critical domains, this is unrealistic: for example, in oncology, reliable treatment decisions typically require the joint expertise of radiologists, pathologists, and oncologists rather than reliance on a single doctor. To overcome this, I developed the first L2D frameworks for *multi-expert routing* (Montreuil et al. 2025c,b). My contributions include (i) a two-stage allocation mechanism for distributing queries across experts and (ii) a one-stage score-based formulation

of top-k deferral with Bayes-optimality and surrogate consistency guarantees.

Beyond robustness and multi-expert extensions, I have also broadened the *applicability* of L2D. In an ICML 2025 paper (Montreuil et al. 2025d), I designed a two-stage framework for multi-task learning, showing how deferral can improve reliability across heterogeneous prediction tasks. More recently, I submitted work on extractive question answering with large language models (Montreuil et al. 2025e), where I proposed a lightweight router that decides whether to answer directly or defer to a larger LLM, thereby balancing efficiency with reliability.

Open Questions and Research Plan

While my past work has focused on robustness, multi-expert extensions, and applicability, several important challenges remain. My ongoing projects are exploring three concrete directions.

I am investigating *LLM routing at the token level*, where a router decides dynamically during generation whether to use a lightweight or a larger language model. To the best of my knowledge, existing approaches typically perform routing at the *sequence level*, deciding whether an entire query should be handled by a small or a large model (Mao et al. 2023). The motivation for token-level routing is that early tokens are often generated with low confidence, and due to the autoregressive nature of language models, this uncertainty propagates and amplifies across the sequence, degrading overall quality. This creates an opportunity: by selectively deferring the generation of initial tokens (or other low-confidence positions) to a larger model, we may substantially improve the reliability of lightweight models while maintaining efficiency.

Furthermore, I am currently studying *online learning to defer* with a varying number of experts, where the system must adapt in real time as experts enter or leave the pool. This setting is motivated by practical deployments in which expert availability is not fixed: for instance, in healthcare, doctors may only be accessible at certain times, and in large-scale machine learning systems, cloud-based experts may be intermittently available due to resource constraints. Classical L2D formulations assume a static set of experts, which limits their applicability in these dynamic environments. My goal is to design algorithms that can efficiently update routing policies as the pool of experts evolves, while maintaining theoretical guarantees on regret and consistency.

Beyond these ongoing efforts, several open theoretical and methodological questions remain. One fundamental issue is the *non-realizability of current L2D formulations*, which can lead to suboptimal solutions even when using consistent surrogates (Mozannar and Sontag 2020; Mao, Mohri, and Zhong 2025). Another limitation is that existing one-stage methods apply only to classification and regression. Expanding L2D to more general learning paradigms—including structured prediction, ranking, multi-label classification, and even unsupervised settings—is an open direction with significant potential impact. These extensions would not only expand the scope of L2D but also deepen our understanding of how to design reliable Human–AI systems across diverse domains.

My plan for the next six months is to complete the ongoing projects on token-level LLM routing, online multi-expert deferral, and partial L2D. In the longer term, these efforts will form the theoretical and applied pillars of my dissertation, advancing L2D toward a general-purpose framework for robust and reliable Human–AI decision-making.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2023-01-041-J) and by A*STAR, and is part of the programme DesCartes which is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

Chow, C. 1970. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1): 41–46.

Madras, D.; Pitassi, T.; and Zemel, R. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31.

Mao, A.; Mohri, C.; Mohri, M.; and Zhong, Y. 2023. Two-Stage Learning to Defer with Multiple Experts. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Mao, A.; Mohri, M.; and Zhong, Y. 2025. Mastering Multiple-Expert Routing: Realizable H-Consistency and Strong Guarantees for Learning to Defer. In *Forty-second International Conference on Machine Learning*.

Montreuil, Y.; Carlier, A.; Ng, L. X.; and Ooi, W. T. 2025a. Adversarial Robustness in Two-Stage Learning-to-Defer: Algorithms and Guarantees. In *Forty-second International Conference on Machine Learning*.

Montreuil, Y.; Carlier, A.; Ng, L. X.; and Ooi, W. T. 2025b. One-Stage Top-k Learning-to-Defer: Score-Based Surrogates with Theoretical Guarantees. arXiv:2505.10160.

Montreuil, Y.; Carlier, A.; Ng, L. X.; and Ooi, W. T. 2025c. Why Ask One When You Can Ask k? Two-Stage Learning-to-Defer to the Top-k Experts. arXiv:2504.12988.

Montreuil, Y.; Heng, Y. S.; Carlier, A.; Ng, L. X.; and Ooi, W. T. 2025d. A Two-Stage Learning-to-Defer Approach for Multi-Task Learning. In *Forty-second International Conference on Machine Learning*.

Montreuil, Y.; Yeo, S. H.; Carlier, A.; Ng, L. X.; and Ooi, W. T. 2025e. Optimal Query Allocation in Extractive QA with LLMs: A Learning-to-Defer Framework with Theoretical Guarantees. *arXiv preprint arXiv:2410.15761*.

Mozannar, H.; and Sontag, D. 2020. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.