# Reverse-Engineering LLMs: From Black Boxes to Potential Scientific Discovery Engines

*Tanya Chowdhury*
*November 1, 2025*

Large Language Models (LLMs) are remarkable at extracting structure from large-scale data—but the nature of this structure remains hidden. My research seeks to uncover the internal computational units and emergent patterns embedded within transformer networks, not only to explain model behavior, but to *leverage these learned representations for scientific discovery.* My work spans functional and mechanistic interpretability, united by a goal of converting black-box systems into engines that surface biologically or clinically meaningful hypotheses.

**Functional Interpretability:** *Axiomatic Understanding of Large Language Models.* At UMass Amherst, my early work focused on the interpretability of neural ranking models, which had seen widespread deployment without a corresponding understanding of their decision-making. We introduced *RankLIME* [1], extending LIME to pointwise, pairwise, and listwise rankers for local, scalable attribution in real-world systems. Building on this, we developed *RankSHAP* [4], an axiomatic framework grounded in Shapley values from coalitional game theory, that ensures attributions for ordering functions are consistent, faithful, and aligned with human intuition. Experiments across ranking architectures and datasets, and a user study, validated its practical reliability. Together, these contributions clarified when common attribution methods fail—and provided the community with tools tied to principled guarantees. These ideas translated directly to computational biology during an internship at *Genentech*, where I designed an axiomatically grounded attribution method for PNET, a popular sparse, domain-knowledge-enhanced network used in prostate cancer discovery. This work established that interpretability can meaningfully assist biomedical insight generation, strengthening the link between machine learning and translational science.

While these axiomatic approaches clarified what faithful explanations should look like, they left open a deeper question: what internal representations actually give rise to these attributions? This realization led me toward mechanistic interpretability—probing how ranking features are encoded within the model itself.

**Mechanistic Interpretability:** *What Statistical Features Do Ranking LLMs Learn?* Building on the axiomatic understanding of output behavior, we next turned inward—probing LLM rerankers (LLaMA2/3, Mistral, Pythia) to understand how information retrieval (IR) signals are represented within their activations [2]. We trained lightweight linear probes over MLP activations and found that models encode *compositional statistics*—nonlinear mixtures of classical IR features such as tf–idf, term overlap, and semantic alignment—rather than simple, independent proxies. These emergent compositions remain stable across architectures and datasets, indicating that LLMs converge toward a shared representational basis for ranking even when trained in isolation. Further, feature localization analyses revealed that lexical cues dominate in lower layers, while semantic and query-document interaction features intensify mid-depth, suggesting a systematic progression from surface-level to abstract matching. This functional view is complemented by a behavioral study of LoRA fine-tuning [5], where we showed: (i) ranking proficiency emerges within a few hundred steps, (ii) even rank-1 LoRA adapters capture most performance, and (iii) mid-layer MLPs—particularly up/gate projections—carry disproportionate responsibility for ranking. These findings reveal where task-specific computation lives and motivate a focus on mid-depth MLP mechanisms.

**Hedonic Neurons:** *Reverse-Engineering Computation via Synergy.* While probing reveals *which* statistical features are represented, it remains limited by its dependence on input activa-

tions—capturing only features elicited by specific examples rather than the model's full computational landscape. Mechanistic interpretability instead seeks to uncover the *emergent features themselves*: the latent, reusable subroutines that drive model behavior. Yet directly analyzing such mechanisms at scale is challenging, as each LLM layer contains millions of interdependent weights.

To address this, I adopted a weight-based lens and asked whether groups of neurons could be modeled as self-organizing computational units. Drawing on coalitional game theory, I proposed *Hedonic Neurons* [3], a framework that treats neurons as agents whose utilities encode **synergy**—the non-additive usefulness of co-activation. Using the PAC-Top-Cover algorithm, I identify stable coalitions that function as cohesive computational subroutines, showing substantially higher synergy than clustering or SAE baselines, exhibiting 3–5× greater out-of-distribution degradation when ablated, and aligning with intuitive IR constructs such as IDF weighting, overlap, and semantic matching. Tracking these coalitions across depth reveals that deeper layers refine and specialize existing features rather than inventing new ones—a pattern suggesting modular, hierarchical computation. By shifting from input-specific activations to weight-grounded coalitional structure, Hedonic Neurons provide a scalable, principled path to reverse-engineer the internal building blocks of LLM reasoning.

**Unified Research Theme.** Across RankSHAP, probing, LoRA studies, and Hedonic Neurons, my contributions form a consistent methodology:

> *Isolate and understand the internal computational units that emerge inside LLMs—then use them to surface new knowledge.*

**Future Research Directions.**

*Reverse-engineering emergent features.* My long-term goal is to establish a systematic science of reverse-engineering large language models—focusing on uncovering the emergent features and statistical regularities encoded within their weight matrices, particularly in MLP submodules. I aim to infer the underlying *rules and patterns* that these weights capture about naturally occurring data. This involves developing weight-based interpretability methods that recover the emergent priors LLMs internalize during training—the implicit relationships, abstractions, and invariances that allow them to generalize across domains. By isolating and formalizing these priors, I hope to both deepen our understanding of real-world data structure and extract actionable insights about the phenomena they model. Over time, this line of work will build toward a principled framework for decoding how learning systems compress, organize, and represent knowledge.

*Discovery for genomic science.* Looking ahead, I aim to apply this reverse-engineering framework to LLMs fine-tuned on genomic, molecular, and clinical data—domains where latent model structure often mirrors biological processes. By decoding the statistical and mechanistic priors embedded in such models, I hope to reveal how they internalize the logic of gene regulation, mutation co-occurrence, and cellular response, surfacing hypotheses that complement and guide experimental biology. My experiences at *Genentech* and *Bristol Myers Squibb* have shown that rigorous interpretability can do more than explain predictions—it can drive new lines of inquiry. I plan to continue collaborating with pharmaceutical and academic biologists to translate model-level discoveries into actionable insight. As a cancer survivor, my long-term goal is to use mechanistic interpretability to illuminate the tumor microenvironment and advance precision therapies that make cancer a tractable and predictable disease.

## References


[1] Tanya Chowdhury, Razieh Rahimi, and James Allan. Rank-lime: local model-agnostic feature attribution for learning to rank. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 33–37, 2023.

[2] Tanya Chowdhury, Atharva Nijasure, and James Allan. Probing ranking llms: A mechanistic analysis for information retrieval. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 336–346, 2025.

[3] Tanya Chowdhury, Atharva Nijasure, Yair Zick, and James Allan. Hedonic neurons: A mechanistic mapping of latent coalitions in transformer mlps. *arXiv preprint arXiv:2509.23684*, 2025.

[4] Tanya Chowdhury, Yair Zick, and James Allan. Rankshap: Shapley value based feature attributions for learning to rank. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[5] Atharva Nijasure, Tanya Chowdhury, and James Allan. How relevance emerges: Interpreting lora fine-tuning in reranking llms. *Presented at WExIR at SIGIR 2025*, 2025.