Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Ubiquitous Computing Systems (Keiichi Yasumoto (Professor))		
Student ID	2511327	Submission date	2026 / 7 / 22
Name	CHOUMAN HUSSEIN		
Thesis title	Manifestation Unit Protocol : A Representational Layer for Mechanistic Interpretability		

Abstract

Mechanistic interpretability has produced a rich inventory of component-level analyses — selectivity tables, probing scores, sparse-autoencoder feature lists, and circuit diagrams — that characterise what the internal components of a neural network encode and how they interact. These analyses are increasingly capable, yet their outputs remain difficult to use: they are typically locked inside the notebooks that produced them, non-composable across methods, not queryable in natural language, and not directly actionable for downstream audit or intervention. This thesis argues that the central obstacle to use is not measurement but representation — the layer between component-level analyses and the people who need their results — and that this layer can be studied, and improved, as a bottleneck in its own right. To fill it we introduce the Manifestation Unit protocol.

A Manifestation Unit is a typed tuple (E, S, R, D, G) — covering identity, semantics, relations, dynamics, and intervention guidance, and extended with attention-head primitives (T) for transformers — whose fields are populated automatically from model activations, rendered into natural language by deterministic templates, and accessed through a hybrid retriever combining exact matching on entity identifiers with semantic search over the remaining fields. The protocol is positioned as complementary infrastructure over the existing interpretability toolkit rather than as a competing analysis method, and four requirements — that the representation be composable, queryable, reusable, and actionable — derived from documented limitations of current interpretability outputs organise its evaluation, with faithfulness enforced as a cross-cutting guarantee that every generated statement traces to a stored statistic. Instantiated across a beta-VAE on CelebA, a CNN on CIFAR-10, and GPT-2 small, the protocol supports its central claims under matched-budget controls: structured retrieval reaches Precision@5 of 89.6% and 96.8% on the VAE and CNN, against 18.0% and 20.4% for a BM25 lexical-retrieval baseline with access to identical statistics and 3.7% and 2.8% for an unstructured prose serialisation of the same content, isolating the contribution of typed structure from that of content; on the CNN, retrieved components causally mediate the target behaviour, satisfying both sufficiency (a 76.3% prediction-flip rate under amplification, versus 6.7% for matched random selection) and necessity (a 38.3-point accuracy drop under ablation, versus 3.6 points); generated responses exceed 94% factual accuracy across configurations; and on GPT-2 the schema absorbs attention-head primitives without modification, recovers five of seven canonical name-mover heads through retrieval, and yields a minimal two-field core (semantics and relations).

We are deliberately explicit about the limits of these results. All experiments are post-hoc analyses of small, public-domain models. Causal mediation is established for the CNN only; on GPT-2 the framework's contribution is at the retrieval layer — set-overlap with a known circuit and behavioural lift over random head sets — and not strict per-head causal mediation, which a path-patching analysis shows the current schema does not deliver. Composability is claimed as a design property, evidenced by architecture-independent extension, rather than as a demonstrated composition of analyses from multiple methods. Within these boundaries, the Manifestation Unit protocol shows that a shared, typed representation layer can make existing interpretability findings queryable, faithful, and actionable through natural language, and that this layer is a tractable and independently valuable target for the field.