Analyze Feature Flow to Enhance Interpretation and Steering in Language Models

Daniil Laptev[1][2], Nikita Balagansky[1][2], Yaroslav Aksenov[1], Daniil Gavrilov[1]
[1]T-Tech, [2]Moscow Institute of Physics and Technology

Ever wondered how to pull the strings inside your LLM?

Decode the Black Box →

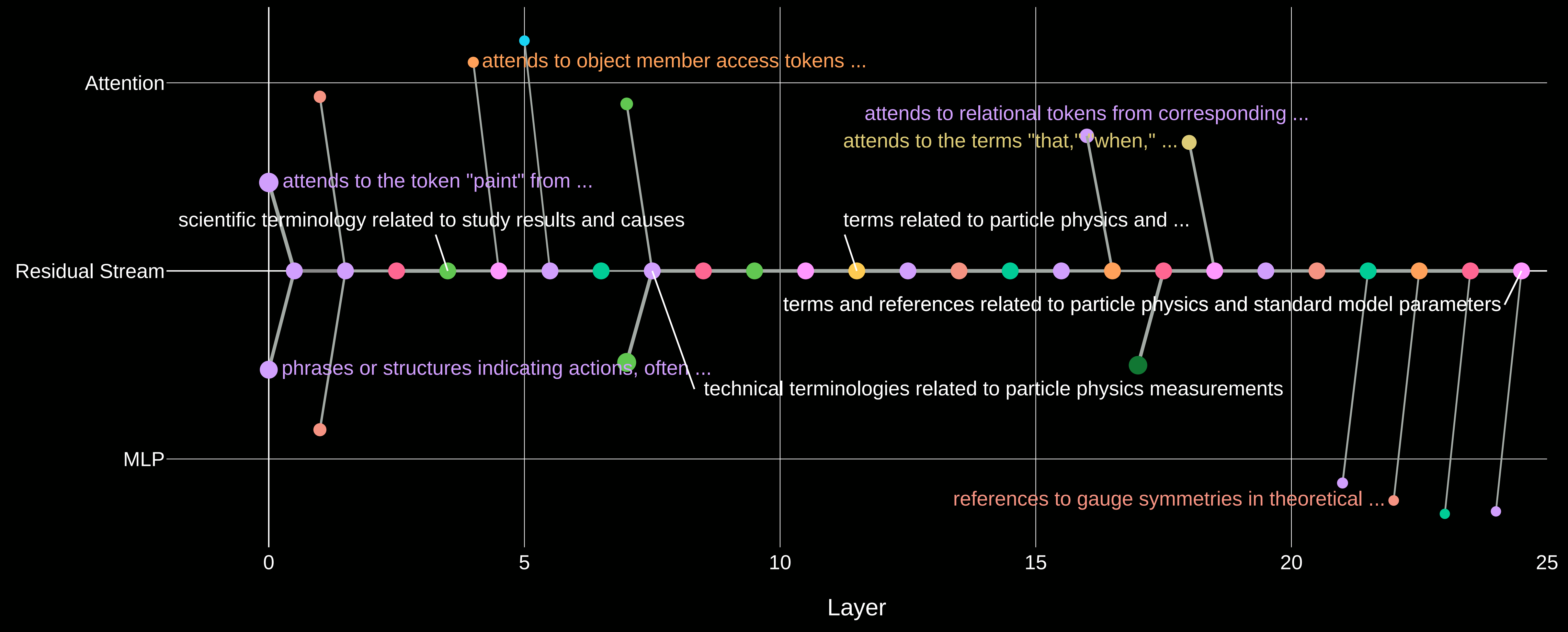
LLMs captivate us but conceal how meaning forms.

Single-Layer Tools Fall Short →

Most interpretability stops at one layer—features evolve across many.

Chain & Steer Features →

Compose layer-to-layer matches into a flow graph, then amplify or mute subgraphs



(Fig. 2) Flow graph for feature 14548 on 24th residual: semantics branch, merge, and resurface.

Match → Compute top cosine-similarity matches between SAE decoder weights

Chain → Link matches into a directed flow graph.

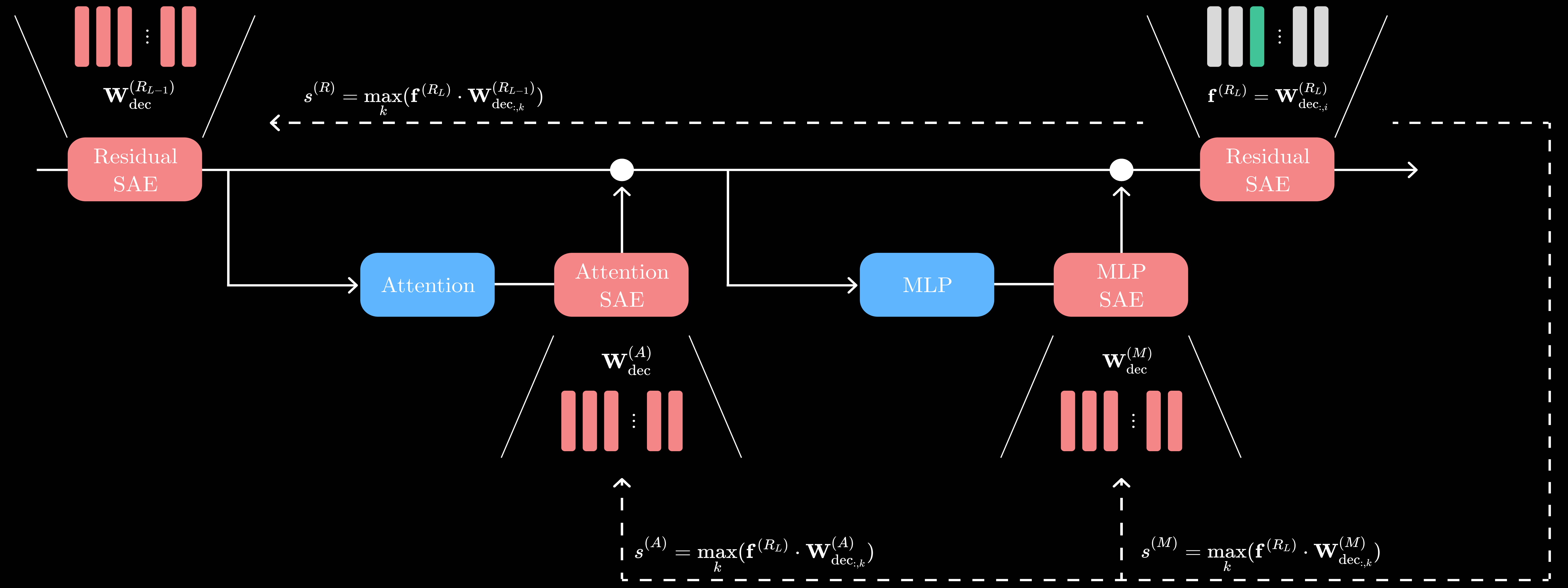
Test → Disable edges to confirm causal links.

Control → Amplify or mute subgraphs to steer output.

65% causal deactivation. Data-free cosine matching achieves a 65% success rate—on par with the Pearson-correlation baseline—versus 73% for exhaustive search.

Theme-steering boost. Full flow graph outperforms single-layer hacks and reduces its dependence on the rescaling hyperparameter.

Mapping feature lifecycles turns interpretation into control.


$\mathbf{W}_{\text{dec}}^{(R_{L-1})}$
$s^{(R)} = \max_k(\mathbf{f}^{(R_L)} \cdot \mathbf{W}_{\text{dec}:,k}^{(R_{L-1})})$
$\mathbf{f}^{(R_L)} = \mathbf{W}_{\text{dec}:,i}^{(R_L)}$
Residual SAE
Residual SAE
Attention
Attention SAE
MLP
MLP SAE
$\mathbf{W}_{\text{dec}}^{(A)}$
$\mathbf{W}_{\text{dec}}^{(M)}$
$s^{(A)} = \max_k(\mathbf{f}^{(R_L)} \cdot \mathbf{W}_{\text{dec}:,k}^{(A)})$
$s^{(M)} = \max_k(\mathbf{f}^{(R_L)} \cdot \mathbf{W}_{\text{dec}:,k}^{(M)})$

(Fig. 1) Feature flow at layer ℓ.