

Layer Effects for SAE Latents - Qwen2.5-14B

Latent 0: Retrieval and Presentation of Supporting Examples or Facts
Latent 1: Retrieval and Presentation of Supporting Examples or Facts
Latent 2: Retrieval and Presentation of Supporting Examples or Facts
Latent 3: Retrieval and Presentation of Supporting Examples or Facts
Latent 4: Retrieval and Presentation of Supporting Examples or Facts
Latent 5: Structural Decomposition and Subcomponent Analysis
Latent 6: Calculation and Execution of Solution Steps
Latent 7: Problem Identification and Task Framing
Latent 8: Initiation of Analytical Decomposition
Latent 9: Final Answer Assertion and Justification
Latent 10: Stepwise Explicit Calculation and Arithmetic
Latent 11: Generating Alternative Considerations and Expanding the Hypothesis Space
Latent 12: Visual Mapping and Structural Diagramming
Latent 13: Anticipating and Weighing Consequences (Pros, Cons, and Trade-offs)
Latent 14: Solution Synthesis and Implementation Planning
Latent 15: Pattern Hypothesis Generation in Sequential Reasoning
Latent 16: Explicit Calculation and Enumeration of Possibilities
Latent 17: Converse Inference / Rejection of Affirming the Consequent

