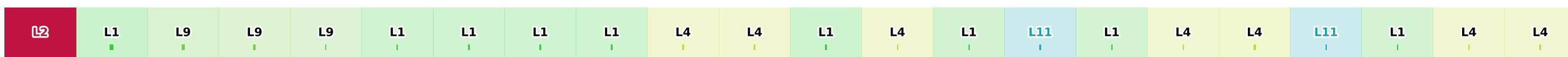


SAE Reasoning Trace - DEEPSEEK-R1-DISTILL-LLAMA-8B, Layer 6

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

<begin_of_sentence> is a sample text for visualization. The S AE will try to represent each token using latent features .



Most Active Latent Features

- | | | |
|----------------------|-----------------------|----------------------|
| L1: Problem Ident... | L9: Local Consist... | L2: **Projecting ... |
| L4: Calculation R... | L11: Metacognitive... | |