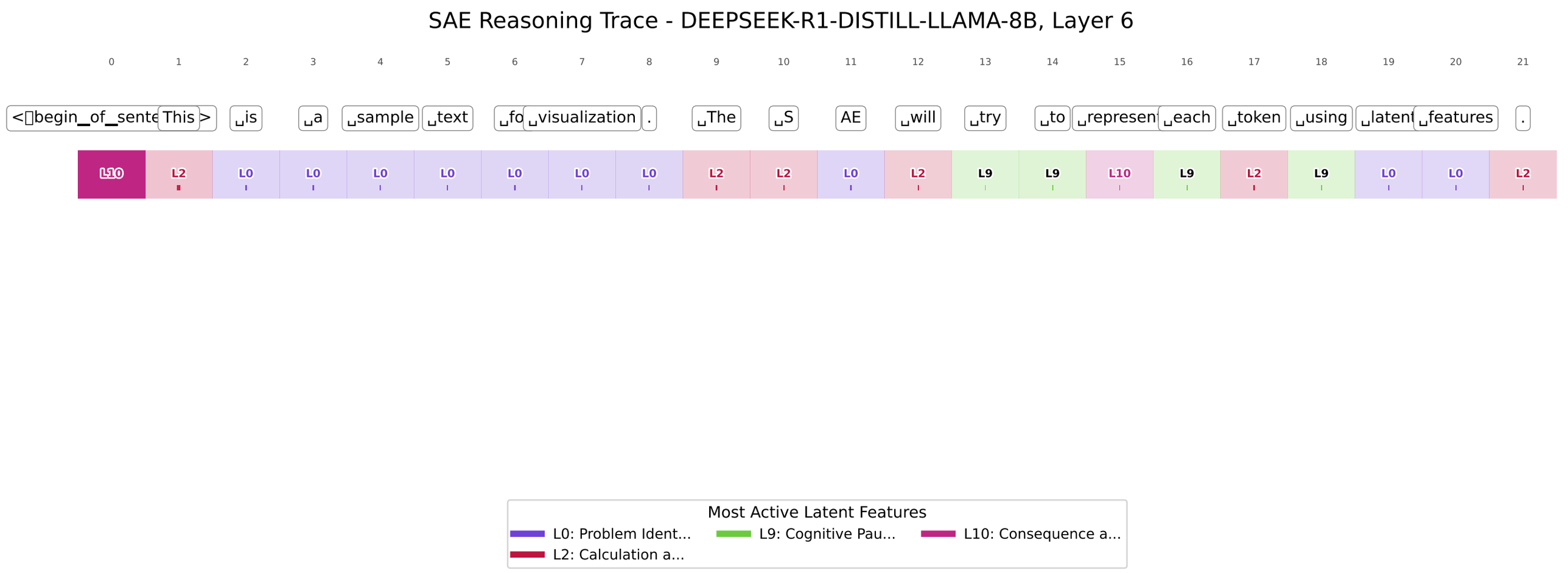
SAE Reasoning Trace - DEEPSEEK-R1-DISTILL-LLAMA-8B, Layer 6
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
<｜begin▁of▁sente
This
>
▁is
▁a
▁sample
▁text
▁fo
▁visualization
.
▁The
▁S
AE
▁will
▁try
▁to
▁represen
▁each
▁token
▁using
▁latent
▁features
.
L10 L2 L0 L0 L0 L0 L0 L0 L0 L2 L2 L0 L2 L9 L9 L10 L9 L2 L9 L0 L0 L2
Most Active Latent Features
L0: Problem Ident...
L9: Cognitive Pau...
L10: Consequence a...
L2: Calculation a...