

Interpreting Diffusion Language Models Over Denoising Time



Scan for the paper

Jingcheng Yang^{1*}, Shangtong Wu^{2*}, Yifan Kang³, Tianhu Xiong¹, Mingyuan Wu¹, Haozhen Zheng¹, Alex Schwing¹, Klara Nahrstedt¹

¹University of Illinois Urbana-Champaign ²University of Oxford ³MIT ^{*}Equal contribution

KEY IDEA Diffusion LMs build an answer over many **denoising steps**. We trace what the model computes at each step, find that it works in clear phases, and show that boosting the right features at the **right moment** takes sorting from 0 to 128 correct answers out of 128.

WHY TIME MATTERS

Diffusion LMs write in steps

A diffusion LM fills in every position of a masked answer over many steps, so its computation unfolds over **time**.

Autoregressive LM

one token at a time, left to right



Diffusion LM

every position at once, refined over steps

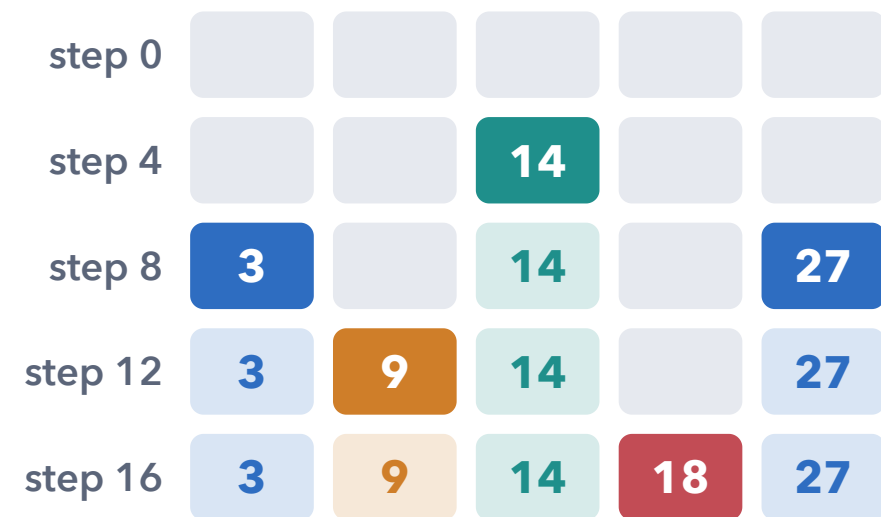


Figure 1. The same answer, written two ways.

OUR APPROACH

From sparse features to interventions over denoising time

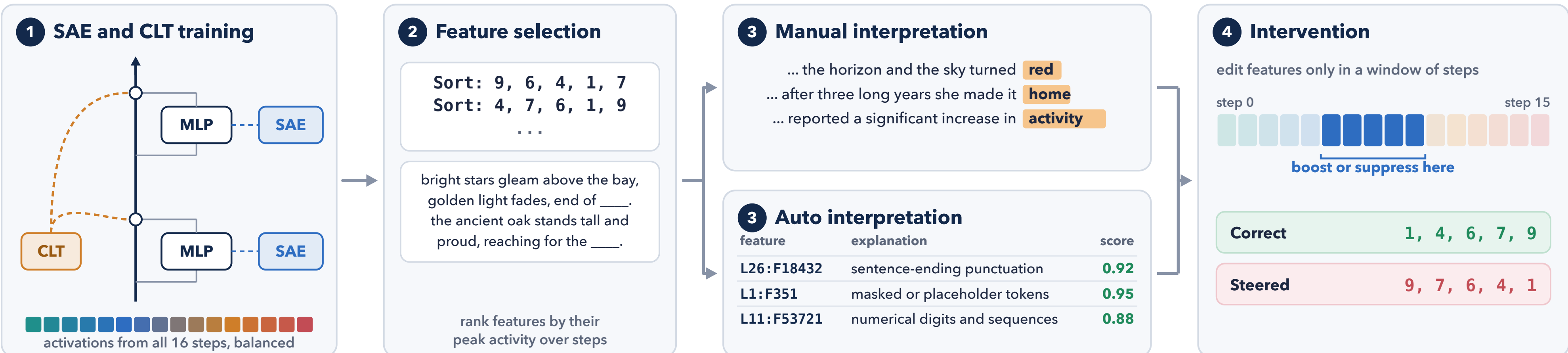


Figure 2. Train SAEs and cross-layer transcoders (CLTs) on every denoising step, select task features, interpret them by hand and with an LLM judge, then edit them during denoising.

FEATURE QUALITY

Features hold up at every step

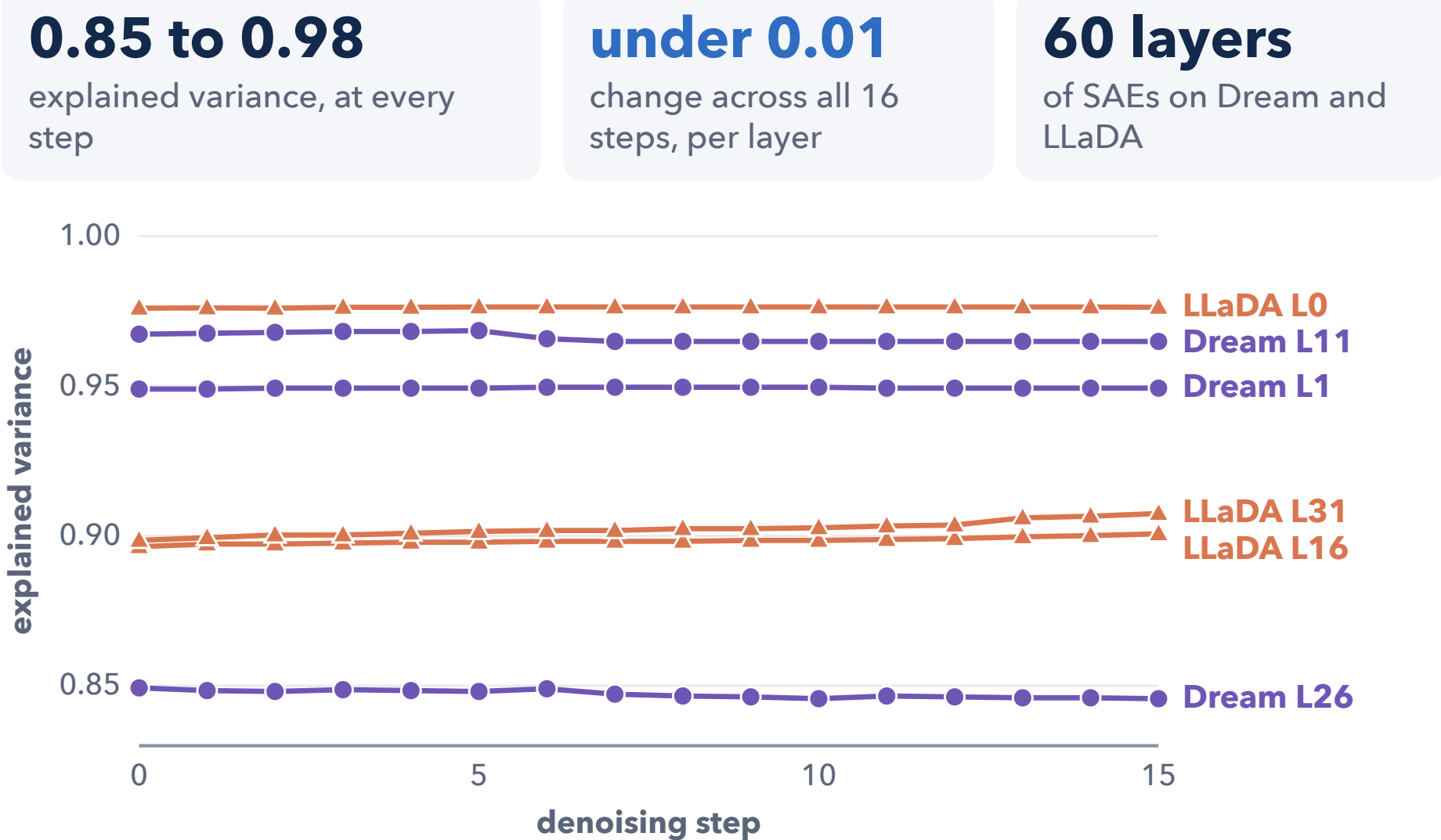


Figure 3. Explained variance of one SAE per layer, trained equally on every step (redrawn from paper Fig. 8). Depth sets the quality; the step barely changes it.

OUR TOOL

The Diffusion Tracer

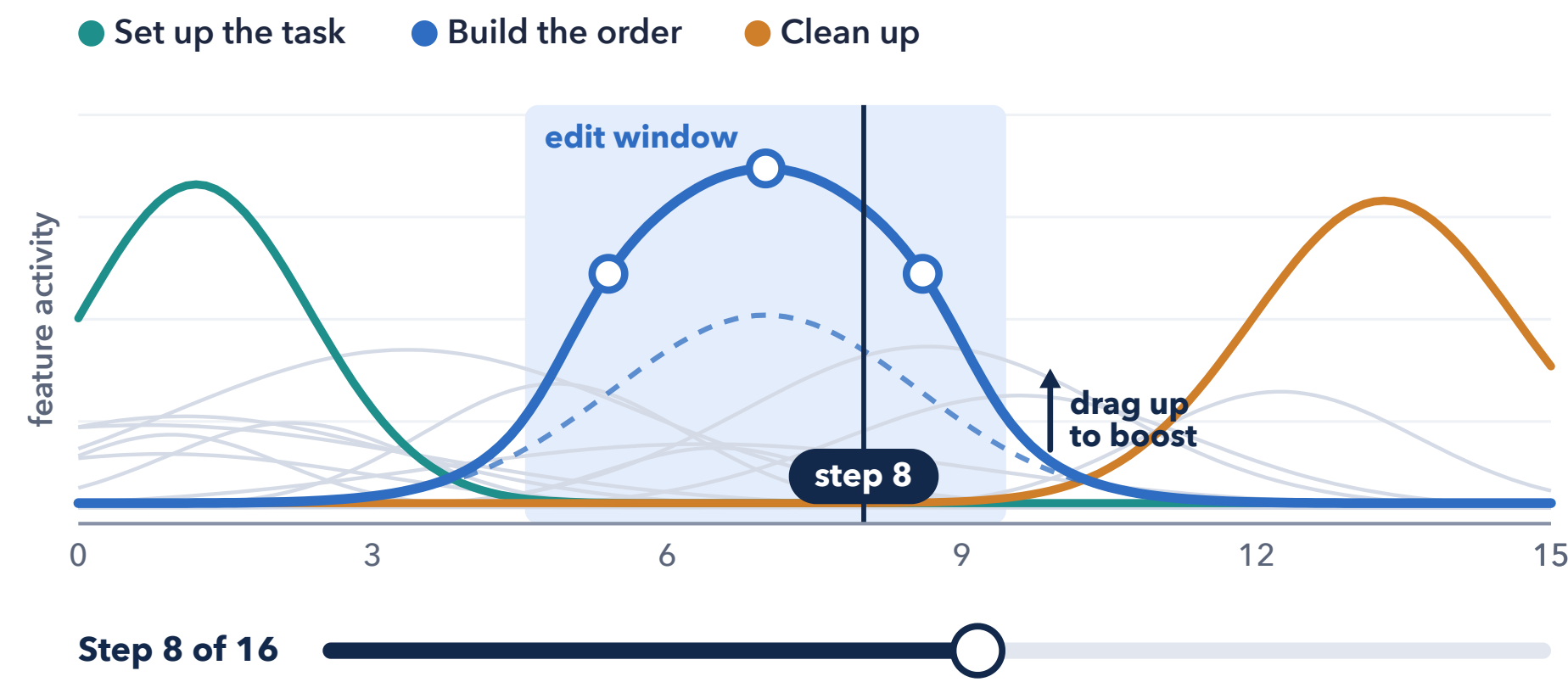
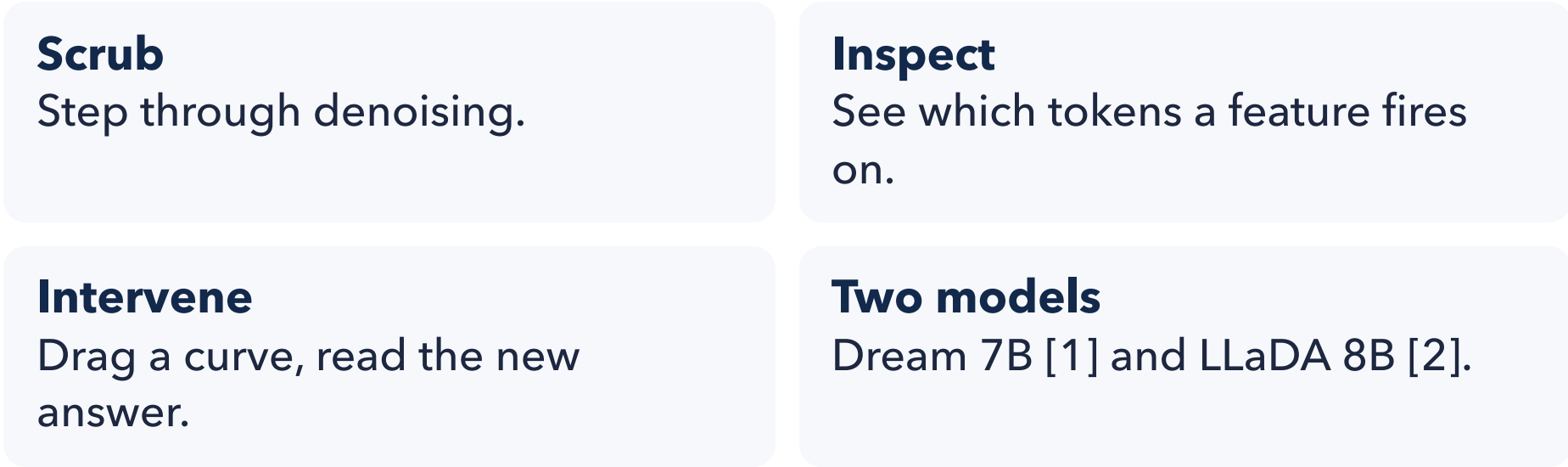


Figure 4. Feature activity over the 16 steps; dragging the blue curve boosts it only inside the edit window. Illustration of our interactive tool.



FINDING 1

The model works in phases

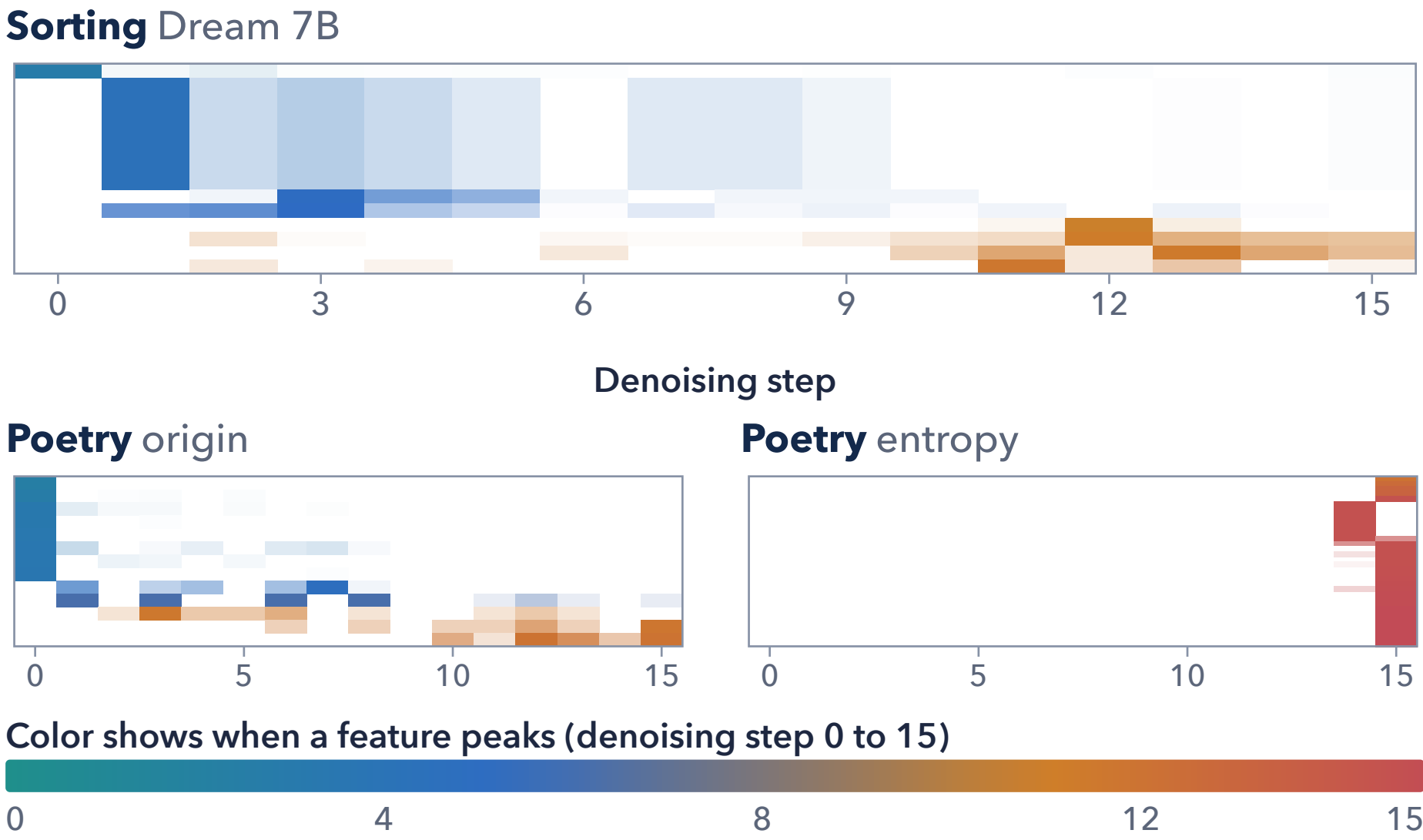
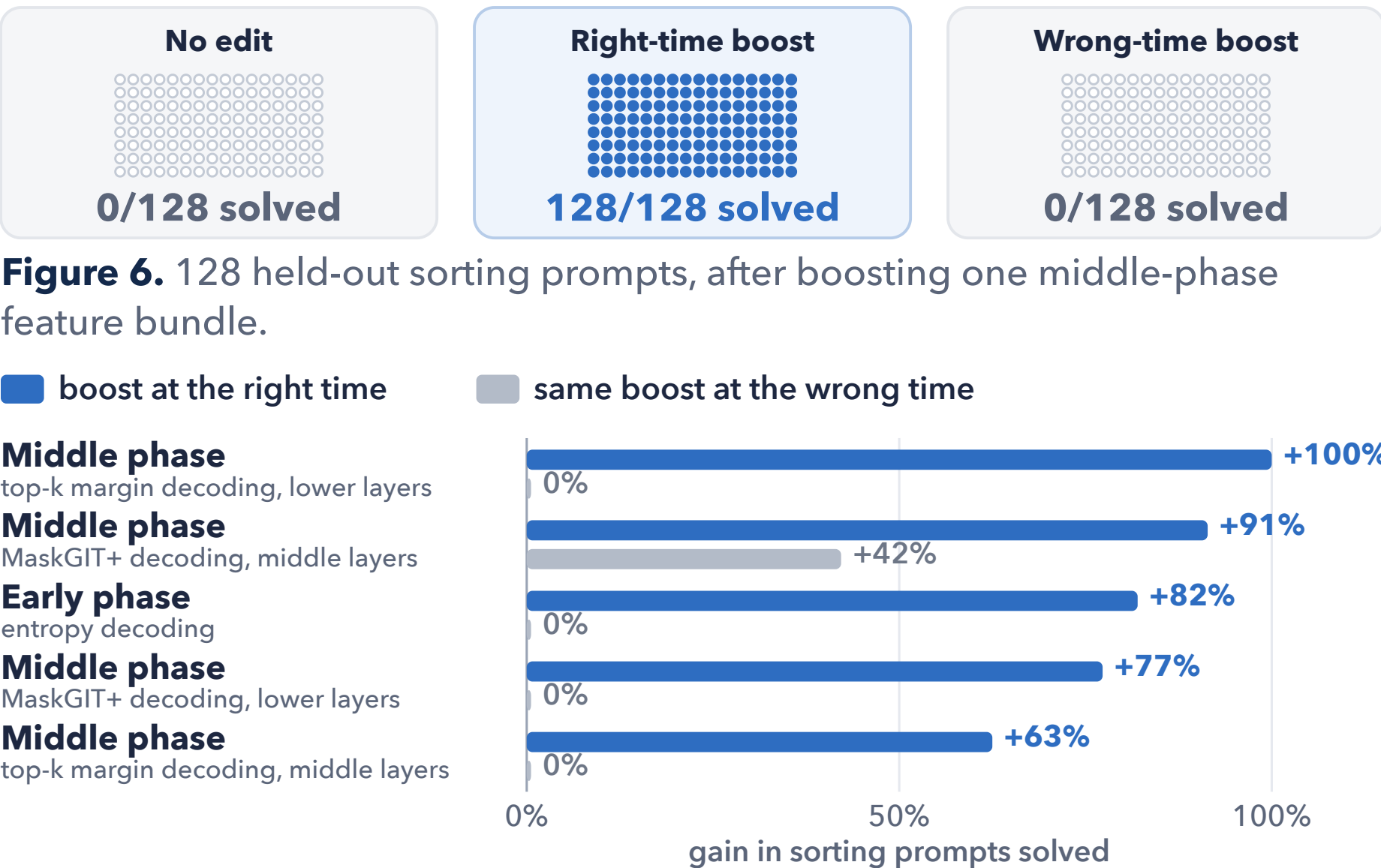


Figure 5. Real traced SAE features; each row is one feature. Early features set up the task, middle ones build the answer, late ones clean up. Phases appear under all five decoding rules we tried; the rule shapes their timing.

FINDING 2

Right feature, right time



FINDING 3

Edits ripple across layers and time

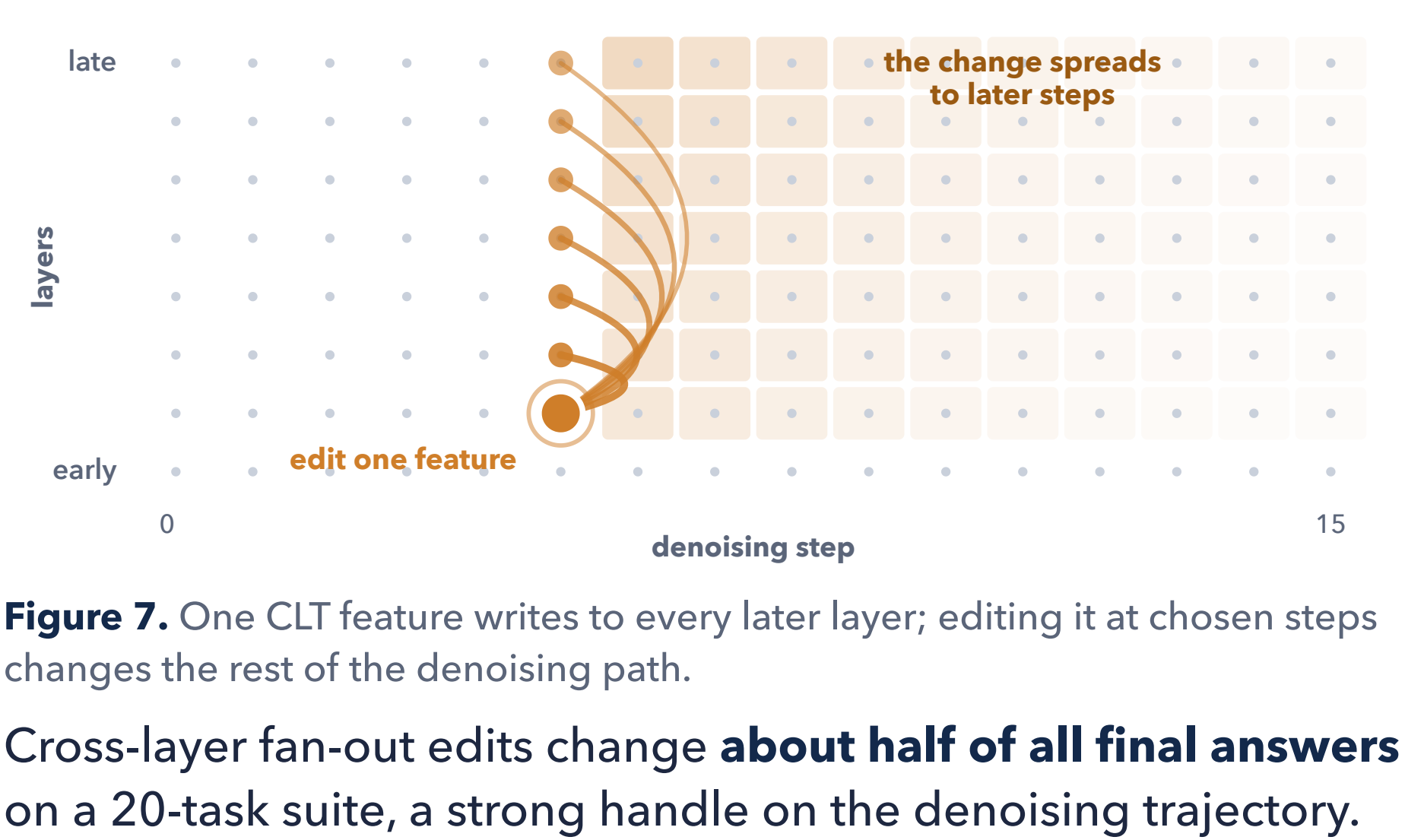


Figure 7. One CLT feature writes to every later layer; editing it at chosen steps changes the rest of the denoising path.

TAKEAWAYS

Time is the missing axis

- 1 Denoising time is a new, readable axis for interpreting diffusion LMs.
- 2 The model works in **phases**: set up, build, then clean up.
- 3 **Timing is a control knob**: the right feature at the right step can fix an answer.
- 4 Our features and tools cover two open diffusion LMs, **Dream 7B** and **LLaDA 8B**.

Open-source release, coming soon
Training code, SAE and CLT checkpoints for Dream and LLaDA, and the Diffusion Tracer.