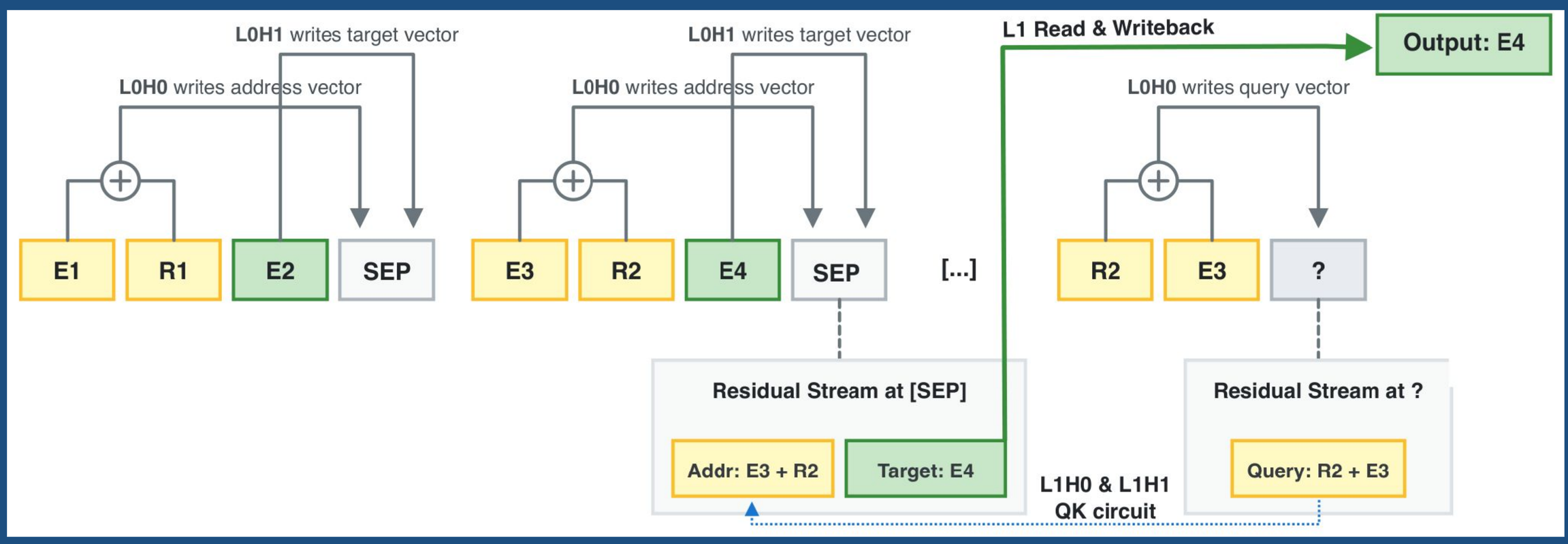

A 2-layer transformer binds entities by adding an entity vector + a relation vector, an address that is causal and linearly decodable, yet SAEs fail to surface it as a clean feature.

Additive Relational Bindings in Transformers What Sparse Autoencoders Miss

Sebastian Hoenig[1] · Su Park[1] · Kushal Jain[1] · Bart Bussmann[1] · Patrick Leask[2] ([1]Independent [2]Durham University)

1 TL;DR

- Layer 0 builds an additive entity + relation address; Layer 1 retrieves it by query–key matching.
- The address is additive (FVE 0.998), causally decisive (Δ +24.6), and linearly decodable.
- SAEs keep 100% accuracy yet almost never surface the joint address as a clean feature (best F1 ≃ 0.091).

2 Motivating question & task

- Q1. How does a transformer bind an entity to its attributes (e.g. “Alice lives in Paris”)?
- Q2. Can sparse autoencoders recover that binding as individual features?
- Setup: a 2-layer attention-only transformer retrieves the E_2 whose (entity, relation) address matches the query; 100% on held-out.

3 Additive & causal address

- Address ≃ global mean + entity vector + relation vector; the additive model explains 99.8% of variance (cosine 0.9996) and generalizes to held-out pairs (FVE 0.998, cosine 0.9995).
- Relocating + corrupting it flips the answer to the distractor (Δ +24.60): a location-independent retrieval key (Fig 3).

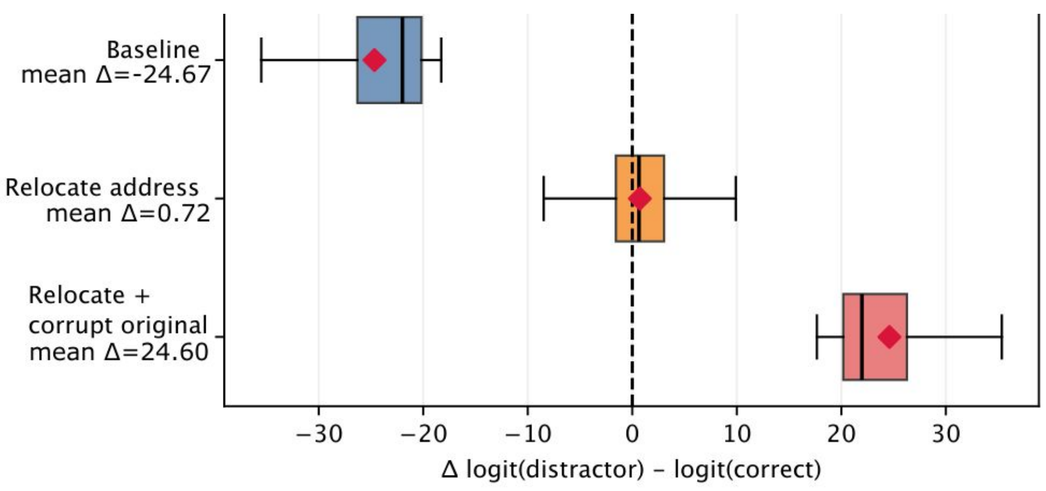

Fig 3. Causal interventions on the address representation. Relocating the target-compatible address to a distractor separator makes the model uncertain; relocating it while corrupting the original target address flips the prediction to the distractor.

4 Matching & SAE recovery

Layer 1 retrieves by query–key matching: same-head QK scores pick the correct fact with 100% top-1 accuracy (+103 margin); cross-head fails (Fig 2).

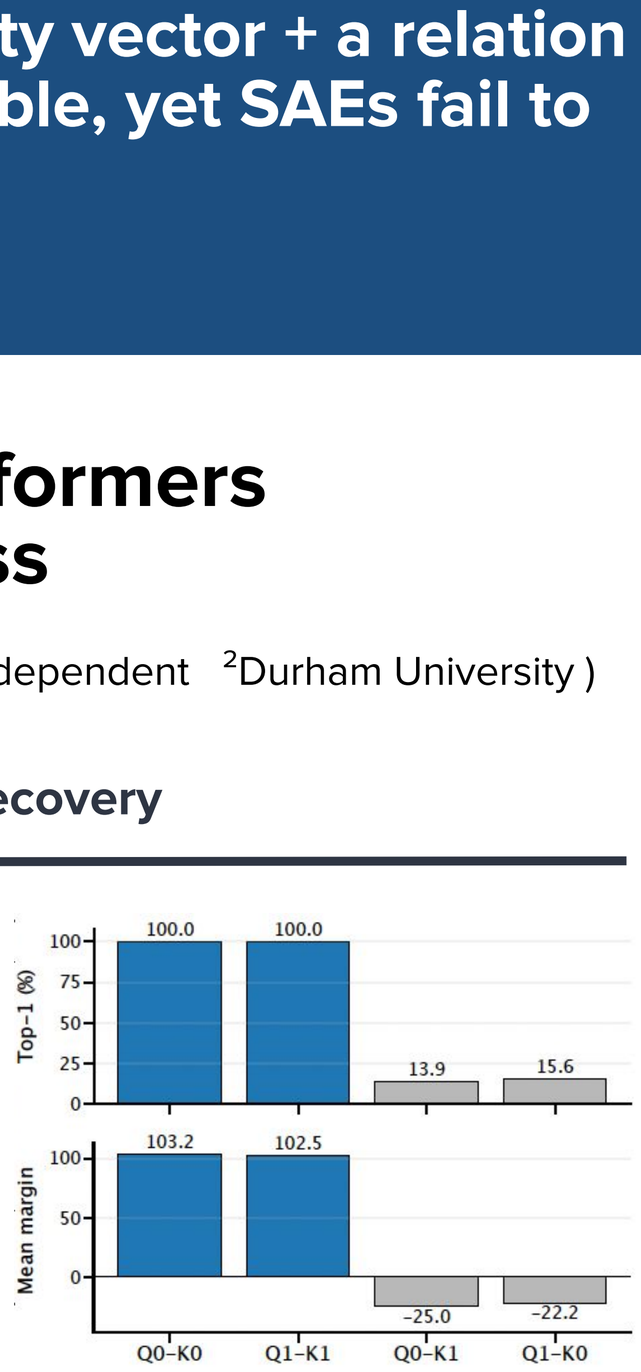

Fig 2. Layer 1 query-key matching. Same-head comparisons select the correct separator with 100% top-1 accuracy and large positive margins, while cross-head comparisons fail. Here Q0-K0 denotes scores computed from the query and key vectors of Layer 1 head 0, and Q1-K1 for Layer 1 head 1.

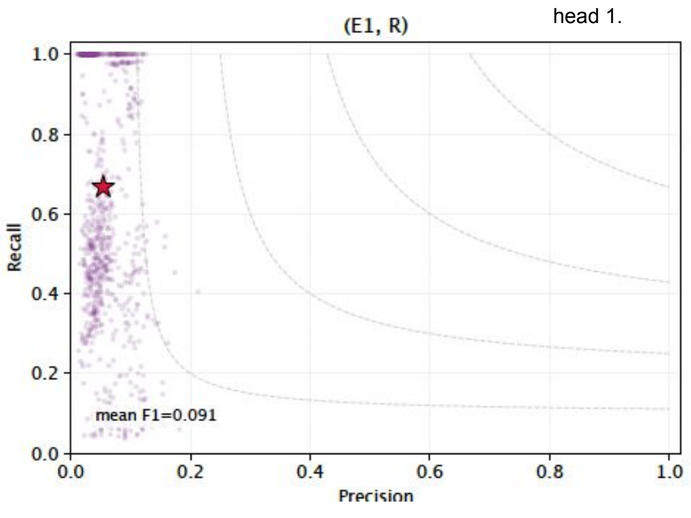

Fig 4. SAE feature recovery of the joint address (E1,R) from Layer 0 head-0 output at [SEP] positions. Each point shows the precision and recall of the best-F1 SAE feature for one entity–relation pair; the red star marks the per-panel mean.

BatchTopK SAEs preserve 100% task accuracy, but the best feature for the joint (E_1,R) address scores only F1 0.091: high recall, low precision (Fig 4).

5 Why SAEs miss it · Discussion

- Holds across all 30 sweep configs: joint F1 0.06–0.09 at the head output, 0.16–0.39 at the residual stream, vs perfect linear decoding.
- The best “joint” latent is really an entity detector: high recall, low (~0.1) precision.
- At the head output, binding is in latent magnitudes, not which latents fire (on/off readout 0.51–0.57 at k=8, vs ≃1.0).
- Feature multiplicity: in residual-stream SAEs, an entity gets disjoint E_1 vs E_2 latents (0% overlap).
- Open question: do pretrained LLMs bind via analogous additive addresses?

PAPER

CODE

hoenigsebastian@gmail.com