NeurIPS 2026

The Alignment Tax Concentrates in Output Projections



Ming Liu · Independent Researcher

github.com/mlzoo/transformer-component-analysis

Alignment helps safety — and taxes structure

```
{ "tool": "get_weather",
  "args": { "city": "Paris",
            "unit": "C" } }
```

Tool calls · JSON · SQL · code

Base model → **Instruct model**

Structured-output loss ↑

= the alignment tax

Prior work

Which layers?

This work

Which matrices inside a layer?

Weight patching: revert one matrix at a time

Instruct model, layer ℓ



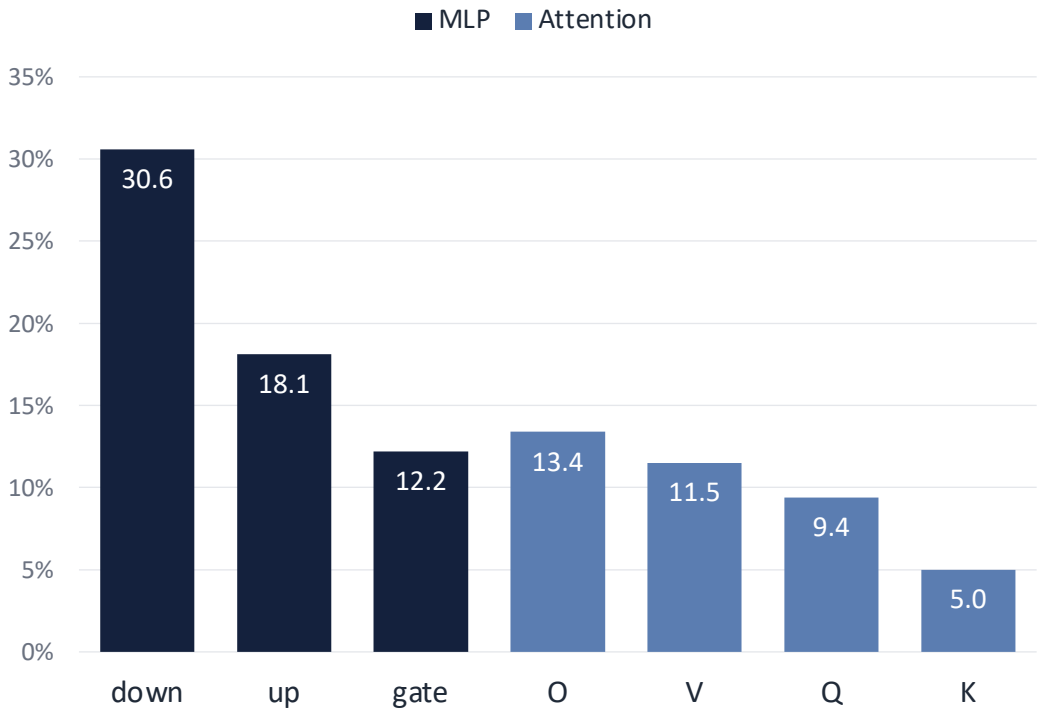
Harm score h(matrix) = Loss(Instruct) − Loss(Instruct with that matrix reverted)

Loss = cross-entropy on gold structured tokens · 209 probes, 18 task types · repeated for all 7 × L matrices

Llama-3.1-8B **Qwen2.5-7B** **Mistral-7B** + Yi-1.5-9B · 14B · 72B

W_down dominates — at equal size

Share of alignment-induced change (mean of 4 model families)



59–63%

of the change sits in the MLP

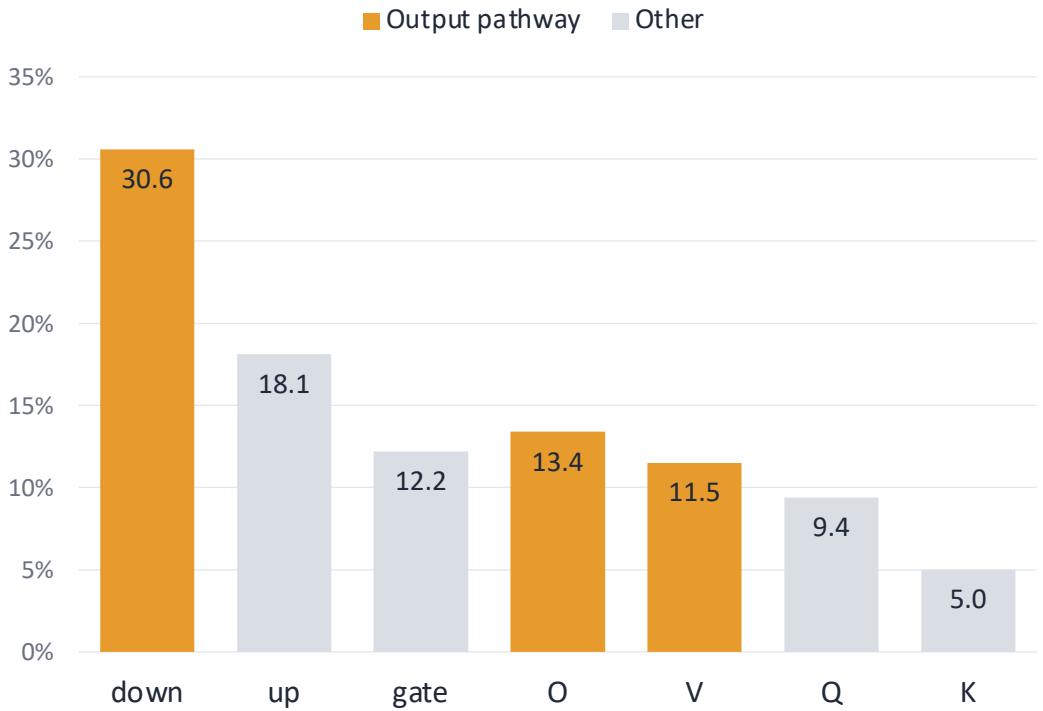
down, up, gate:

- same parameter count
- same size of weight change

→ **sensitivity, not capacity**

The output pathway carries most of the change

Share of alignment-induced change (mean of 4 model families)



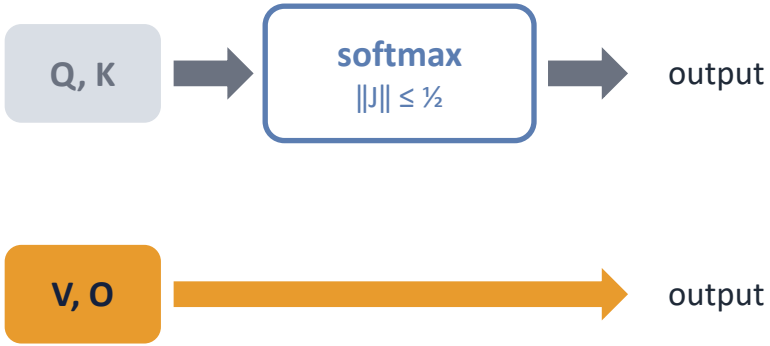
49–65%

of the change sits in V, O and down

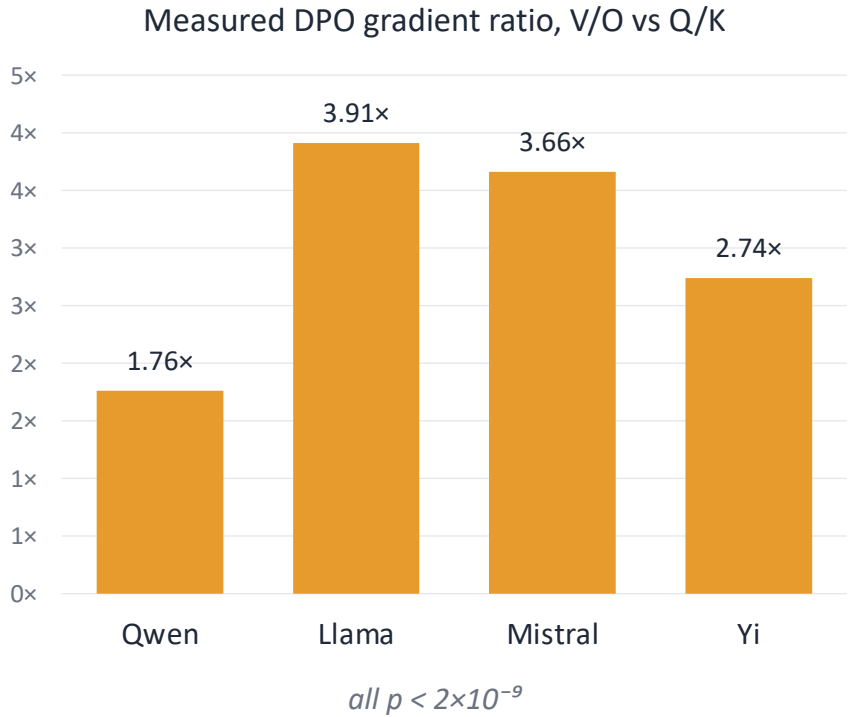


Consistent across Llama, Qwen, Mistral, Yi

Why V/O > Q/K? Softmax shrinks the gradient



Gradients to Q/K pass through the softmax and are attenuated; gradients to V/O flow directly.



Surgical Alignment Reversal (SAR)

Training-free: undo only the most harmful matrices

1

Score

Harm score for every matrix in every layer

2

Select

Take the top 5% by harm score

3

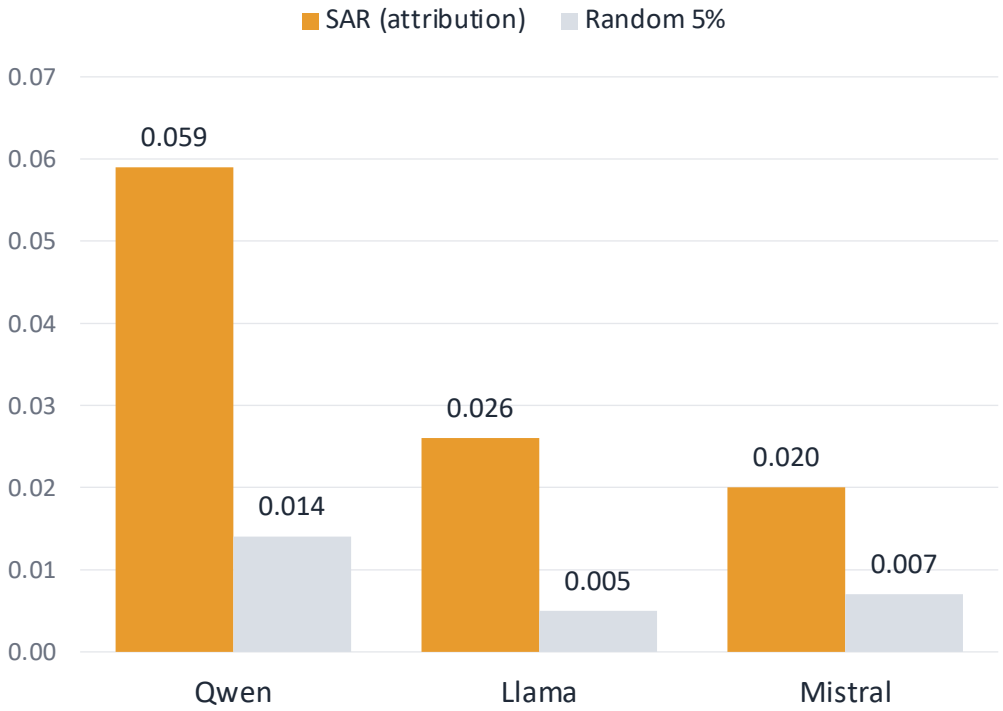
Revert

Replace them with base-model weights

Why it can work: the MLP ordering also appears under plain-text fine-tuning, but which layers alignment harms is task-specific. SAR uses that signal.

SAR beats random rollback on every model

Loss reduction on structured outputs, 5% budget



2.8–5.3×

better than random selection

62% / 47%

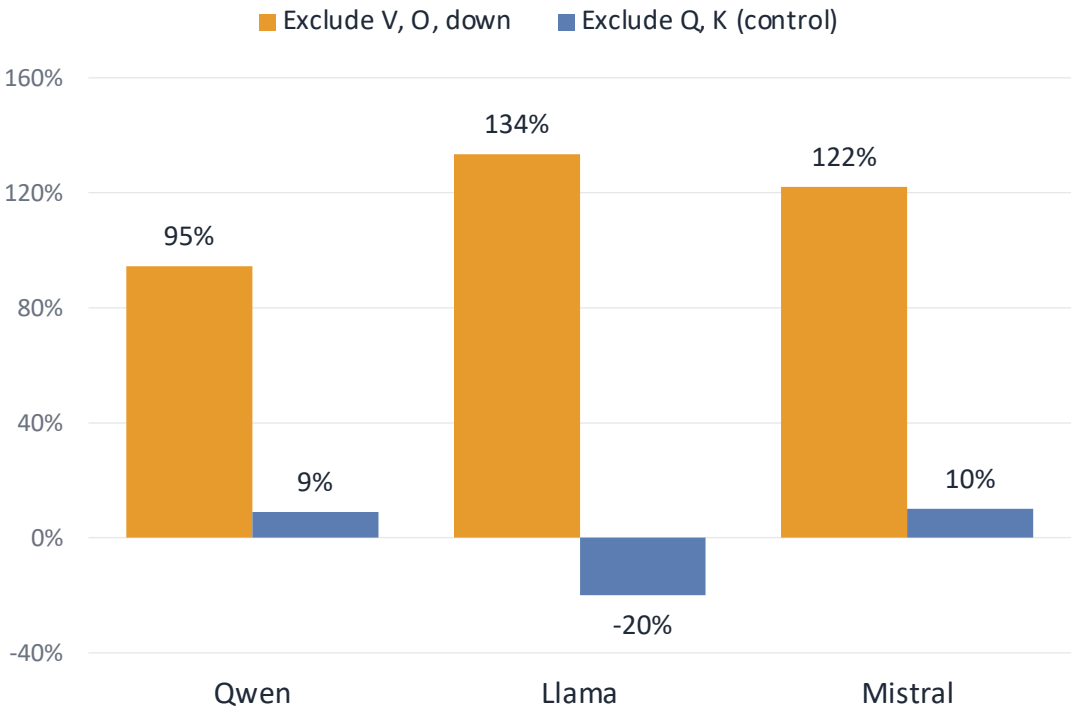
of Qwen's tax recovered (in-domain / new dataset)

MMLU $|\Delta| \leq 0.3$ pp

Refusal rate $|\Delta| \leq 1.4$ pp

Same story at training time: OC-DPO

Tax reduction when matrices are left out of LoRA during DPO



15 / 15

model × seed runs point the same way

Q K **V** **O** gate up **down**

Leave these out → tax mostly gone

Small-scale isolation test (20 pairs, seed 42 shown)

Takeaways

Where	Alignment tax concentrates in output projections, especially W_down.
Why	Architecture shapes the ordering; alignment decides which layers get hurt.
So what	It is actionable: SAR at inference time, OC-DPO at training time.

Replicates at 14B · gradient pattern holds at 72B · SwiGLU decoder models

github.com/mlzoo/transformer-component-analysis