How Alignment Routes

Localizing, Scaling & Controlling Policy Circuits in Language Models

Gregory N. Frank | Chief Scientist, Molt AI Corp

ICML 2026
Mech Interp Workshop

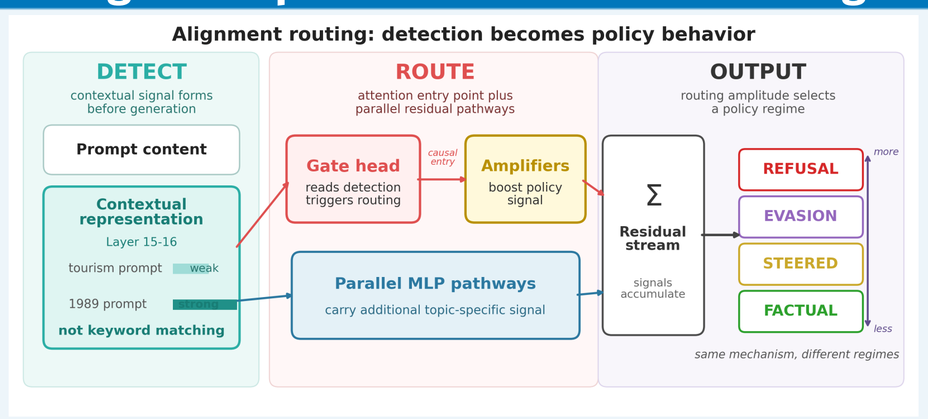
TL;DR. Alignment policy is implemented by a *sparse attention circuit*: an early **gate** head reads detected content and fires deeper **amplifier** heads toward refusal. The same motif appears in **12 models (2B–72B) from 6 labs**. Because the gate **commits at the gate layer** (before the deeper **amplifiers**), a simple in-context **cipher** collapses its causal necessity **70–99%**: the model turns to puzzle-solving instead of refusal. By the time deeper layers could decode anything resembling the original harm, it is **past the gate**, too late to refuse.

Topic detection is easy, behavioral response is not

Four models, one query about a sensitive topic (our corpus spans Chinese-political censorship and general safety). A linear probe at mid-depth is **100% accurate in all four**. Every model *detects* the topic. Yet one refuses, one emits propaganda, one answers factually, one fabricates.

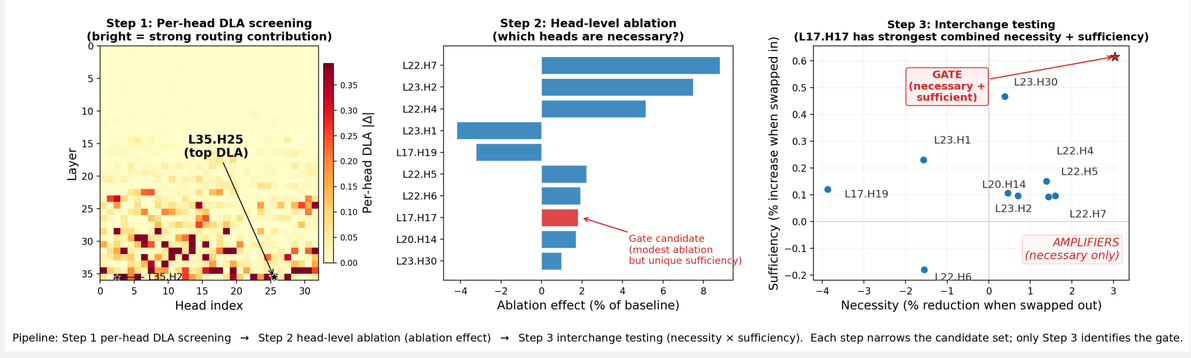
Detection is universal and trivial. The **behavior** is decided *later*, by a learned map we call **routing**. *This work localizes that machinery: where it lives, how it scales, and how it fails.*

We found a gate-amplifier mechanism controlling refusal



Detection forms in the **mid layers**. A **gate** head reads it and writes a routing vector; **amplifier** heads **a few layers later** boost that vector toward refusal. MLP pathways carry topic-specific signal in parallel, and can dominate on narrow single-topic prompts. The output regime (refusal, evasion, or factual) is set by the signal's amplitude.

Finding the gate: three methods must agree



Pipeline: Step 1 per-head DLA screening → Step 2 head-level ablation (ablation effect) → Step 3 interchange testing (necessity × sufficiency). Each step narrows the candidate set; only Step 3 identifies the gate.

No single method finds the gate:

- **DLA** (output contribution) → later-layer amplifiers; gate ranks below 150th.
- **Ablation** (necessity) → gate only 6th.
- **Interchange** (content-specific signal) → **gate L17.H17** (in Qwen3-8B) wins by a wide margin, $p < 0.001$.

Gate = necessary *and* sufficient (a **trigger**); **amplifier** = necessary only (a **booster**).

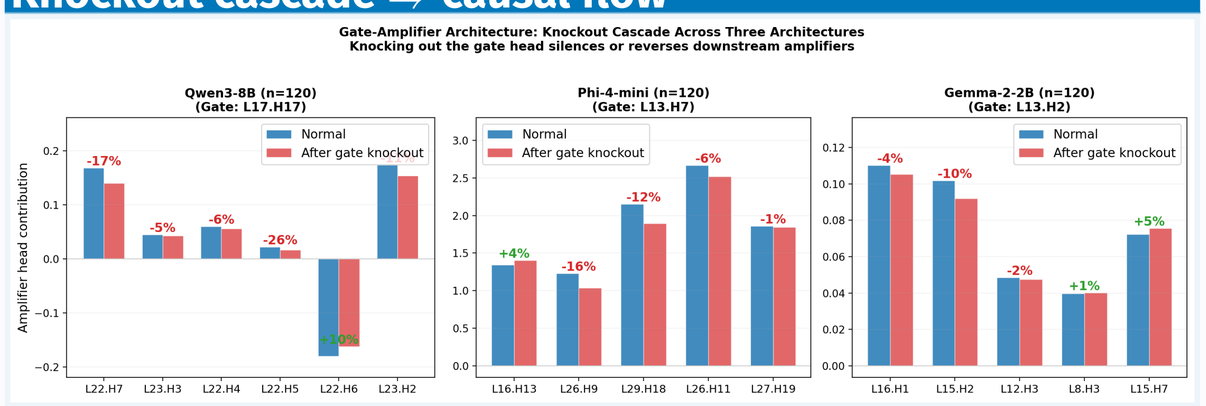
The gate is a trigger, not a carrier

The gate contributes $<$**1%** of the routing signal at the output. DLA alone would never find it. Yet it is causally decisive.

Head	Role	DLA (L18 / out)	Interchange
L17.H17	Gate	**#2** / >20	1.1% ($p<.001$)
L22.H7	Amp	n/a / #5	0.8%

Right after it writes (L18) it ranks **#2**; by the output, distributed carriers dominate and it vanishes. *Like a thermostat: it doesn't produce the heat; it controls what does.*

Knockout cascade ⇒ causal flow



Zero the gate's output, recompute amplifier DLA. Amplifiers **collapse or reverse** in three architectures (n=120): Qwen3-8B 5/6, Phi-4-mini 3/5, Gemma-2-2B 3/5. Gate cascade effect **10.5%** vs. a non-gate null of 3.9%; bootstrap Jaccard 0.92–1.0, permutation $p < 0.001$.

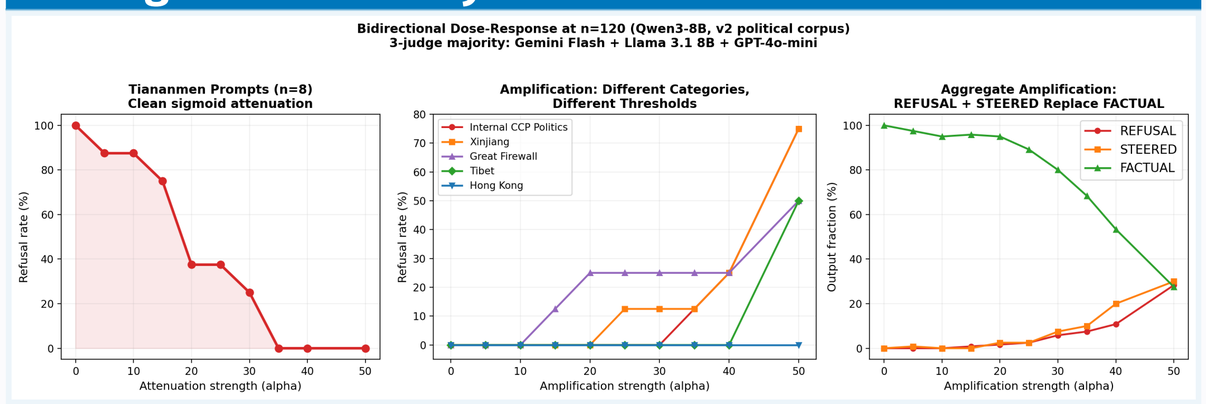
One motif, twelve models, six labs (2B–72B)

Interchange screening ($n \geq 120$) finds the same gate–amplifier motif in **every** model tested; specific heads differ by lab, and the gate widens from a **single head** in the smallest models to a **band of adjacent heads** at scale. Per-head **ablation collapses** (58× weaker at 72B) while **interchange stays informative**.

Model	Lab	Params	Gate	Nec%
Gemma-2-2B	Google	2B	L13.H2	**8.4**
Llama-3.2-3B	Meta	3B	L27.H1	3.0
Phi-4-mini	Microsoft	3.8B	L13.H7	3.4
Qwen2.5-7B	Alibaba	7B	L25.H1	2.4
Mistral-7B	Mistral	7B	L31.H22	1.0
Qwen3-8B	Alibaba	8B	L17.H17	1.1
Gemma-2-9B	Google	9B	L38.H14	1.9
GLM-Z1-9B	Zhipu	9B	L19.H23	4.7
Phi-4	Microsoft	14B	L38.H25	2.6
Qwen3-32B	Alibaba	32B	L56.H3	3.2
Llama-3.3-70B	Meta	70B	L26.H40	2.0
Qwen2.5-72B	Alibaba	72B	L79.H11	1.3

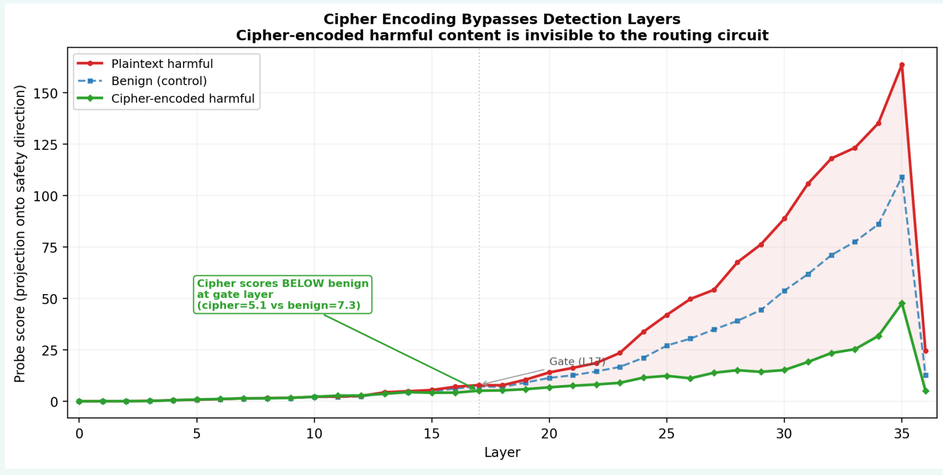
Audit with interchange (causal swaps), not ablation, which fails silently at scale.

Routing is continuously controllable



Turning the **gate head**'s routing signal up or down slides behavior continuously: **refusal** → **evasion** → **factual** (a clean sigmoid on Tiananmen, 100%→0%). On safety prompts the *same knob* turns refusal into **harmful guidance**: the capability is **gated by routing, not removed**. This is *ablation localized to one point*: we modulate one head's activation, not a refusal direction across the whole network, sweeping the full allow-to-refuse spectrum. Thresholds vary by topic and input language.

Early commitment ⇒ a cipher bypass



The gate **commits at the gate layer**, before deeper layers finish reading the input. Teach a substitution **cipher** in-context and the detection signal **collapses 66–88%**; the model *solves the puzzle* instead of refusing.

- Gate interchange necessity drops **70–99%** across three models.
- **Rescue:** inject the gate's *plaintext* activation → **48% of refusals restored** (Phi-4-mini), localizing the failure to the *routing interface*.

Any encoding that fails to form the detection-layer signal the gate reads bypasses the policy.

Why it matters

- Alignment is a **localizable, controllable early-commitment circuit**, not a diffuse property.
- Encoding jailbreaks have a **circuit-level explanation**: they *starve the gate*, not the safety capability. **Defenses** must act upstream of detection or be encoding-robust; most current training is neither.
- Bonus: **cipher contrast analysis** maps the full content-dependent circuit in $O(3n)$ forward passes.

Paper & code

arXiv 2604.04385

github.com/gregfrank/how-alignment-routes