NUCLEATION PILOT · FINDING

Early frames leave persistent, source-decoupled, content-specific traces — and instruction structure resolves into monitorable lanes

The H-SC line: H-SC1 (persistence) → H-SC2 (a corrected mis-step) → H-SC2b (what it encodes) → H-SC3 (instructedness is sub-family lanes) → H-SC4 (safety directives have their own lane). A companion to the six-family clearing transfer.

Christopher Blake Head · ORCID 0009-0004-2308-6051 · Navigator's Log R&D · 2026-08-09 · Frozen detector `nucleation-detector-1.1.0`, SHA-256 `6094de97…` (deposited; identical across every run) · Each stage registered & machine-certified before its run. Companion to DOI 10.5281/zenodo.21843505.

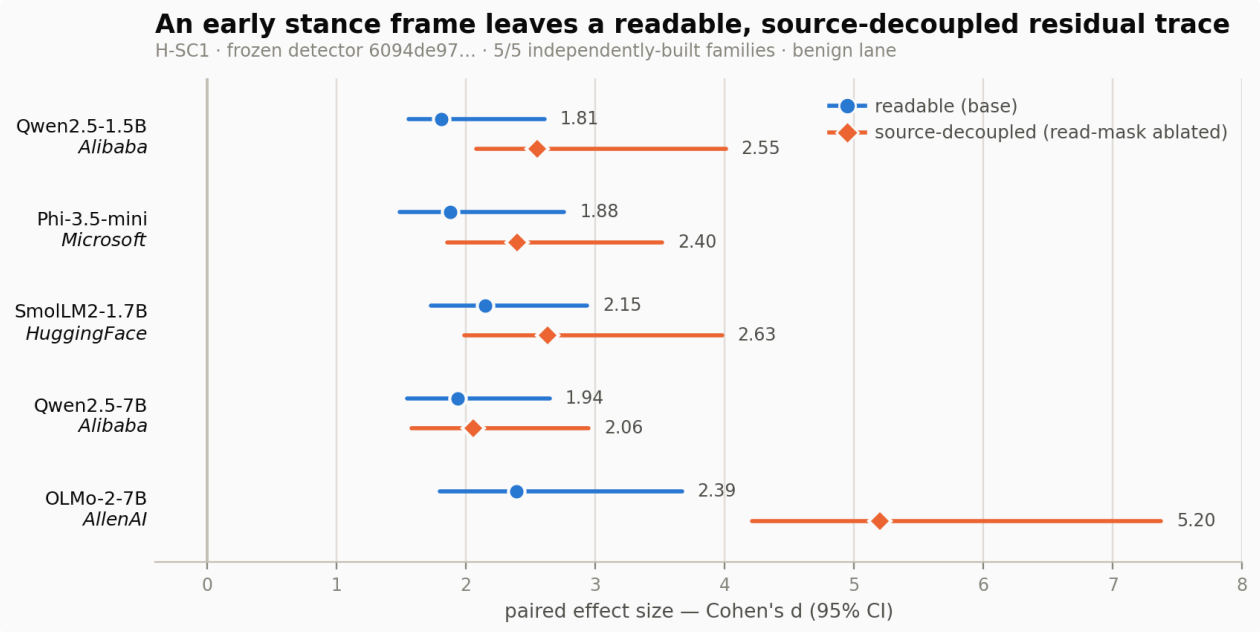
Across five independently-built open-weight families, an explicit **early frame** — a short instruction placed early in a benign dialogue and aged across several turns — leaves an activation-level trace that is (i) readable at the aged read, (ii) **causally source-decoupled** (survives read-masking the source turn), and (iii) **content-specific**: different early distinctions occupy different, largely-orthogonal directions. Pushing on *what* is encoded shows that "an instruction was given" is not one axis but a small family of **restriction lanes** — and the safety-relevant ones form their own coherent lane, so monitoring needs more than one linear probe.

REPLICATED 5/5 Deposited frozen detector, unchanged; five families; benign lane; every stage preregistered. The one wrong turn (H-SC2) is reported openly; the two structured-not-binary outcomes (H-SC3, H-SC4) became a reusable methodological lesson.

1 · H-SC1 — an early frame is persistent and source-decoupled (5/5)

Same-premise paired minimal pairs differing only in an early stance instruction (guarded vs relaxed), same syntactic form, paraphrase-varied ×6, aged four turns. The frame is **readable** at the aged read in 5/5 families (paired 24/24, graded Cohen's d 1.8–2.4, every 95% CI clear of 0) and **causally source-decoupled** in 5/5: read-masking the source turn keeps the paired effect (ablated d 2.1–5.2), the mask-efficacy guard passes everywhere, and the source turn draws only 2–9% of attention vs 30–49% recency. OLMo-2-7B — which source-

coupled under the earlier clearing line — here decouples most strongly (ablated d 5.20). Byte-identical repeats confirm determinism.



H-SC1. Base and read-mask-ablated effect sizes (Cohen's d, 95% CI) across five families. All clear of zero; the ablated leg (source turn masked) matches or exceeds the base — the trace survives removing the model's access to the turn that set it.

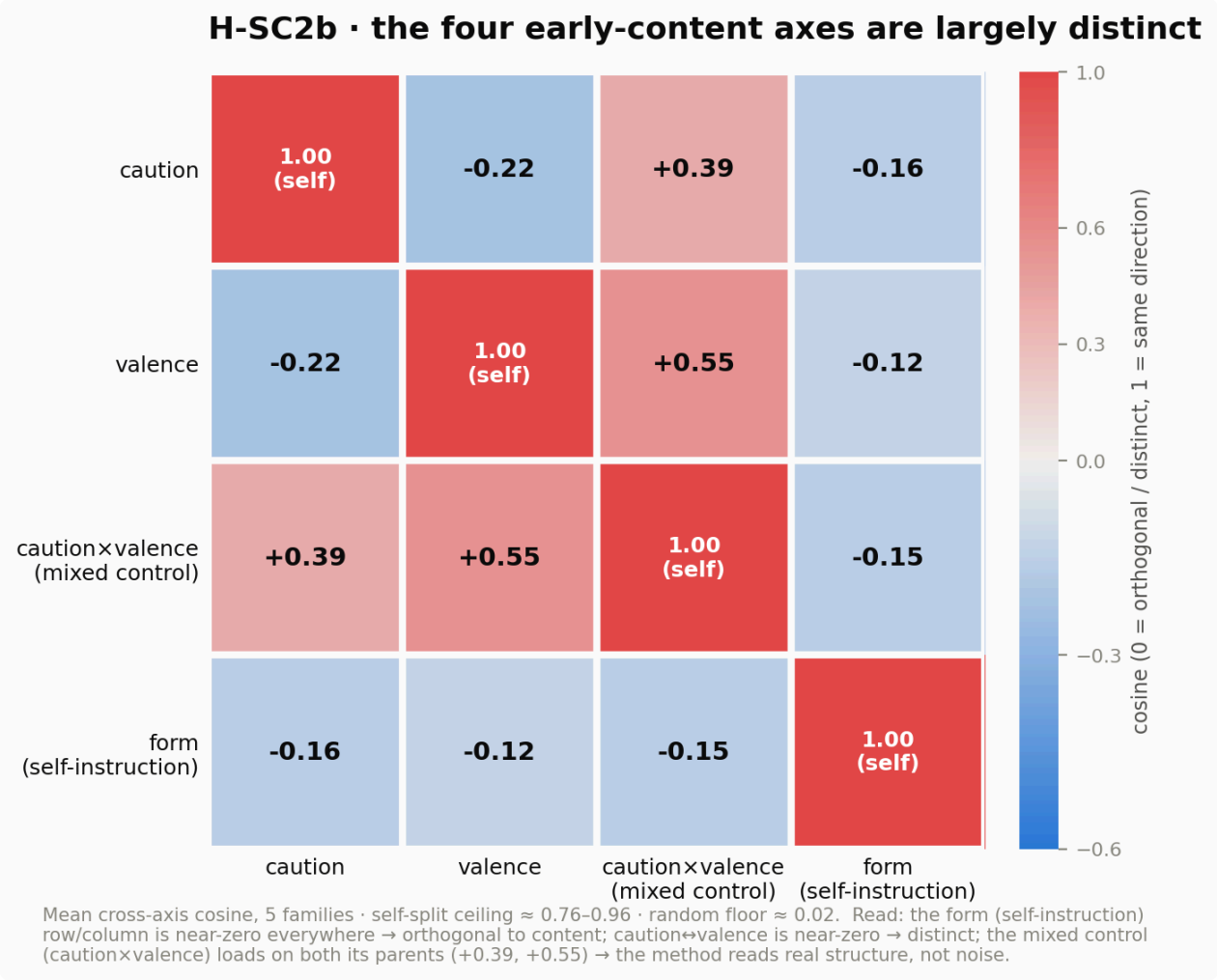
2 · H-SC2 — a content probe that could not disentangle (reported openly)

Four contrasts (caution/relaxed, valence, caution/gloom, neutral instruction/description) all read and decoupled 3/3. This does *not* identify the encoded variable: the paired test fits a **fresh axis per contrast**, so it separates *any* early wording difference — including a purely stylistic instruction-vs-description contrast. Recorded openly and fed back as a rule: to ask whether two effects are the *same* direction, freeze one axis and cross-project — never fit a new axis per condition (lesson L11). The "sharpening" texture noticed in H-SC1 resolved here (12/12): the direct-attention path carries the signal noisily, so masking it tightens the read.

3 · H-SC2b — cross-projection: the trace is content-specific and multi-axis (5/5)

Freezing each contrast's mean-delta axis and cross-projecting — controlled by a self-split **ceiling** (≈0.76–0.96) and a random **floor** (≈0.02) — the three independent-construct pairs are distinct in all five families: caution↔valence −0.10 to −0.28, caution↔form −0.09 to −0.27, valence↔form −0.05 to −0.22, all

at the floor. The by-construction-mixed contrast loads on both parents (caution +0.39, valence +0.55), validating that the machinery reads real structure. The self-instruction **"form" axis is orthogonal to caution and valence in 5/5**, robustly encoded (self-d 2.4–4.1), and **strengthens with scale** (Qwen 1.5B→7B: 2.37→4.11) while caution stays flat.

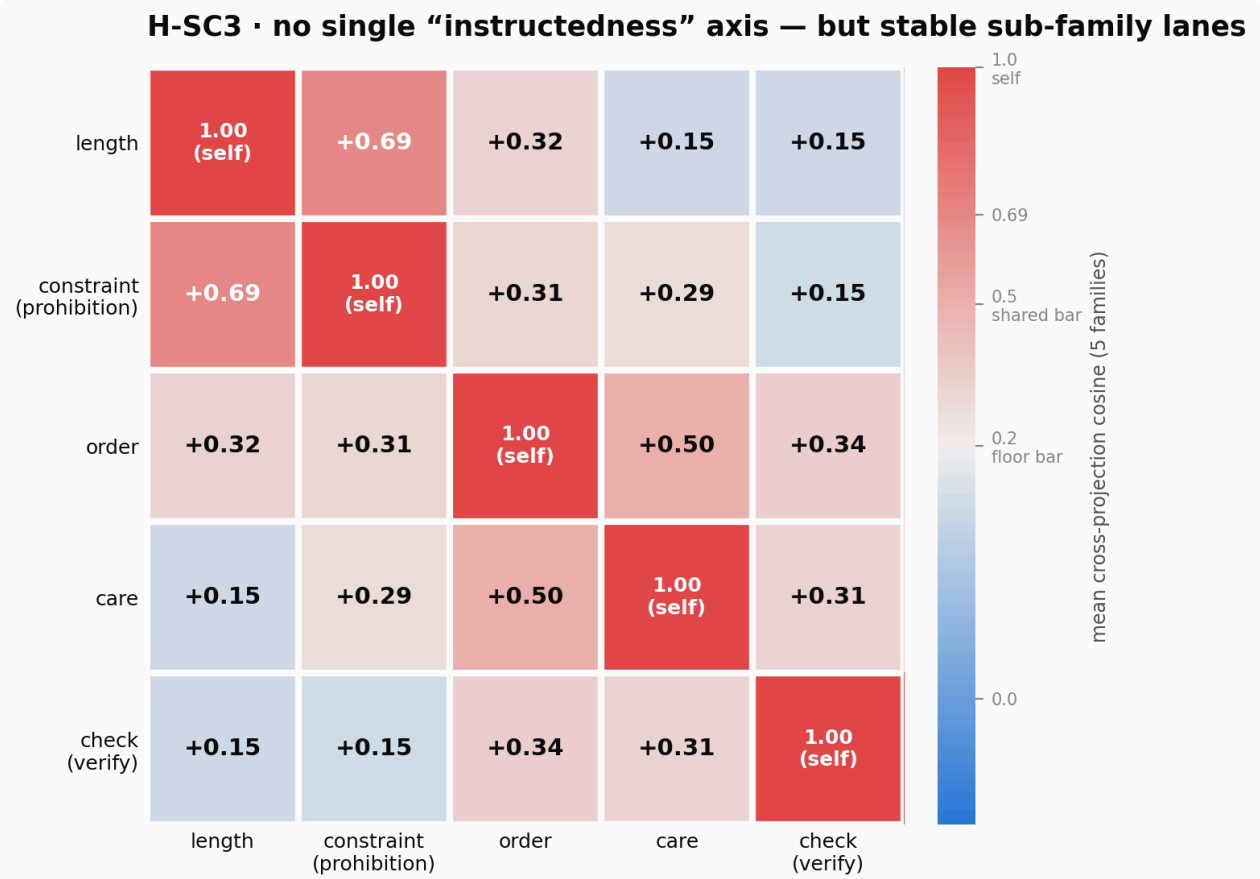


H-SC2b. Mean cross-axis cosine (5 families). Caution, valence, and the self-instruction "form" act sit near the floor off-diagonal — three distinct directions. The designed-mixed control loads on both parents, the internal check that the cosine reads real structure.

4 · H-SC3 — is "instructedness" one axis? No — sub-family lanes

H-SC2b's "form" axis came from a single instruction type. H-SC3 built a matched instruction-vs-description form axis for **five instruction types** — order, length, constraint, care, check — paraphrase-varied ×6, and cross-projected them. There is **no single "an instruction was given" direction**: the median off-diagonal cosine is 0.19–0.36 in every family, far below the ~0.9 ceiling and the 0.5 shared bar, and the "one axis" hypothesis is rejected 0/5. But instructedness is not atomized either — a robust **"shape/limit the output" lane** (length↔constraint)

is SHARED in 5/5 families (0.63–0.77), a softer "how-to-engage" grouping links order↔care (3/5), and verification (check) stands most apart. Larger models keep the lanes *more* separated (Qwen median off-diag 0.33→0.19, 1.5B→7B). Because the preregistered rule was binary but reality was a structured middle, this became lesson **L12**: read the pair matrix, not the median.

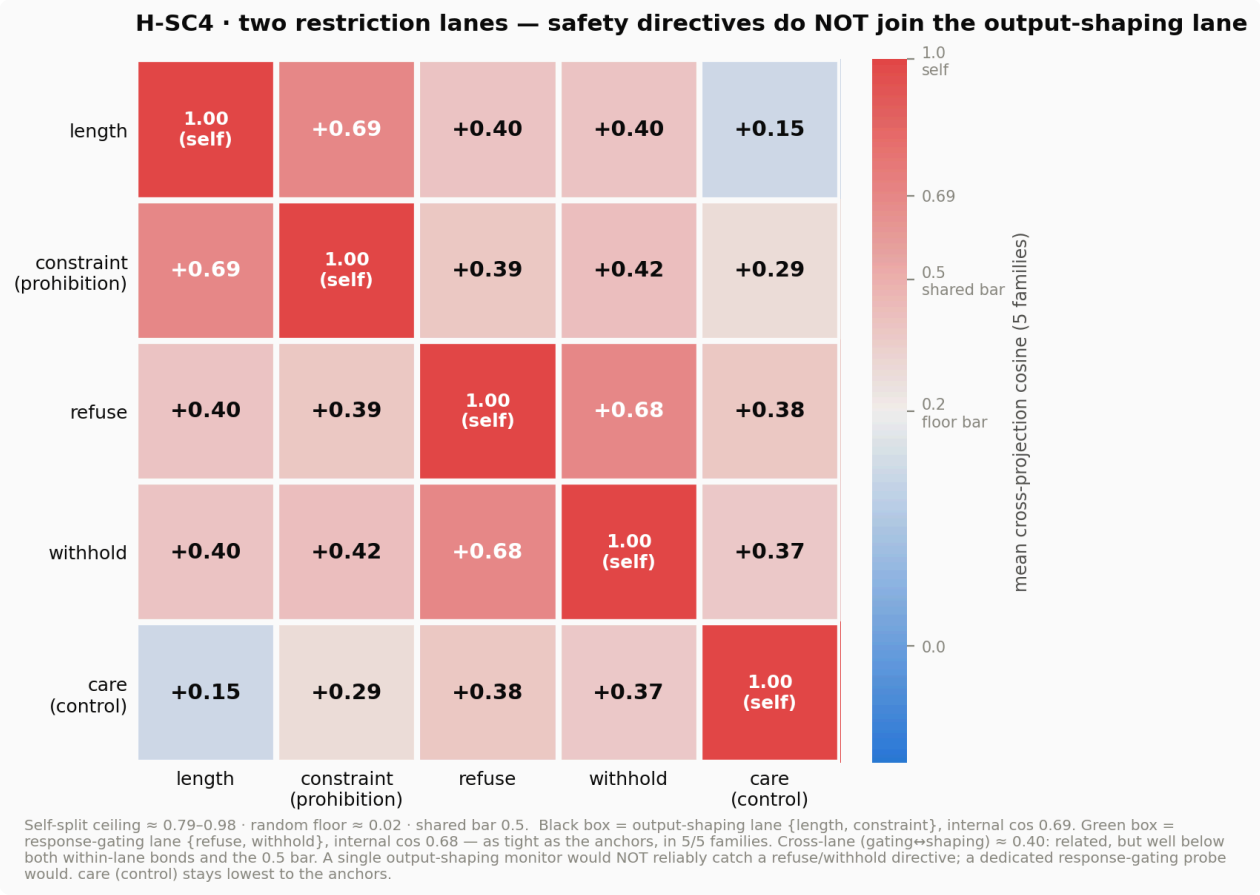


H-SC3. Mean cross-projection cosine of five instruction-type form axes (5 families). No pair reaches the ceiling — no single "instructedness" axis. Boxed: length↔constraint, the one lane SHARED in all five families.

5 · H-SC4 — do safety directives join the output-shaping lane? No — two lanes

The monitoring question. Taking H-SC3's output-shaping lane (length, constraint) as anchors — re-verified SHARED in 5/5 (0.63–0.77) — H-SC4 cross-projected two benign safety-relevant directives, **refuse** ("decline this / don't take it on") and **withhold** ("keep this private / don't share the details"), with a **care** control. Against the registered rule (JOINS iff |cos| ≥ 0.5 to *both* anchors), **refuse joins 0/5 and withhold joins 1/5** — safety directives do not reliably fall in the output-shaping lane. The deciding structure: **refuse and withhold cohere with each**

other at mean cosine 0.68 — as tight as the output-shaping anchors are internally (0.69), in all five families — while sitting at only ~0.40 to those anchors. Two distinct **restriction lanes**: *shape the output* (length/prohibition) and *gate the response* (refuse/withhold), related (~0.40, above the floor and above `care`) but well below both within-lane bonds. The `care` decoy stays lowest to the anchors (0.22; cleanly SEPARATE in both 7B models). Cross-lane cosines fall with scale — the same lane-separation trend as H-SC3.



H-SC4. Mean cross-projection cosine (5 families). Black box = output-shaping lane {length, constraint}, internal 0.69. Green box = response-gating lane {refuse, withhold}, internal 0.68 — as tight as the anchors. Cross-lane ≈ 0.40: related but distinct. A single output-shaping monitor would not reliably catch a refuse/withhold directive.

6 · What the line establishes

The residual stream carries an early frame's influence forward on **multiple distinct, content-specific directions**, read without the model attending back to the source. Caution ≠ valence; *that* an early self-instruction occurred is separable from *what* it was about; and the "instructedness" of a directive is itself a small **family of restriction lanes** — an output-shaping lane and a response-gating (refuse/withhold) lane are equally coherent yet distinct, and the lanes resolve more sharply with model scale. This extends the source-decoupled

residue phenomenon from premise clearing to generic early framing, resolves it as multi-axis and content-specific, and maps the instruction-following portion of that geometry into lanes a linear monitor could watch.

7 · Honest bounds

≤7B open weights, benign lane, n=24 paired items/type, single deterministic seed/family; white-box linear read (no claim beyond the axis's reach); coarse cosine thresholds (the principled read is ceiling-vs-floor). The form-axis and lane-separation scale trends are exploratory — one within-family scale point plus cross-family — not proven laws. Small negative cosines are treated as near-zero/distinct; cross-lane ~0.40 is "related," not orthogonal. **Unit of analysis:** the 24 items per type are paraphrase-cycled from a shared premise pool, so they are not 24 independent draws; the load-bearing replication is agreement across five independently-built families. (This caveat is now an enforced field in the commit-before-run preregistration template, v0.3.0.)

8 · Integrity & safety posture

Deposited frozen detector, hash recorded before any target model loaded and verified identical after every run; all mutable machinery kept in the harness; every stage preregistered and machine-certified before its confirmatory run; the H-SC2 mis-step corrected on the record, and the structured-not-binary outcomes of H-SC3/H-SC4 captured as lesson L12. Benign drives only — the refuse/withhold content is innocuous; only the *form* is a directive. **Monitoring takeaway:** a single frozen linear "instructedness" probe is insufficient — an output-shaping monitor will not reliably catch a refusal or withholding directive — but the lanes are coherent enough across families that a small set of dedicated linear probes (at least one output-shaping, one response-gating) is a viable, defensive monitoring design. Routed to Anthropic User Safety as a monitoring addendum; no exploit or steering tooling built.

Head, Christopher Blake (2026). *The H-SC line: early frames leave persistent, source-decoupled, content-specific traces, and instruction structure resolves into monitorable lanes.* Navigator's Log R&D. Companion to Nucleation Pilot, DOI 10.5281/zenodo.21843505. Frozen detector `nucleation-detector-1.1.0` (SHA-256 6094de97…). Registered & certified: HSC1_v2 / HSC2 / HSC2b / HSC3 / HSC4.