Supplementary Materials: Concept Encoding Strategies Across 28 Transformers

James Henry *Independent Researcher* · jamesrahenry@henrynet.ca · ORCID 0009-0005-7126-9466

Written: 2026-05-05 22:15 UTC

Updated: 2026-07-17 00:50 UTC — §D reframed from exclusion to full correction (decision change, James 2026-07-16): exfiltration stays at C=17, corrected and re-run under the recorded-draw protocol (§D). Added 2026-07-17 00:21 UTC as the defect/exclusion record. Prior update 2026-07-16 06:19 UTC — consistency pass against the current main text (28-model/C=17 state): §B scope note for the two pending model rows, retired score-taxonomy language removed, §C summary rederived from Tables S1–S2.

Companion document to the main preprint. Cross-references of the form "§N.M" point into the main preprint; cross-references of the form "§A" through "§L" point within this document.

---

§A. Technical Notes

Float64 metrics: All Fisher separation, coherence, and Procrustes computations are performed in float64 regardless of model inference precision. This corrects a silent overflow bug discovered in fp16 Fisher normalization at deep layers of models with >36 layers.

OPT embedding projection: OPT-350M has `word_embed_proj_dim=512` but `hidden_size=1024`, causing layer-0 activations to have different dimensionality than subsequent layers. Our feature tracker handles this by closing feature tracks at dimension boundaries and birthing new tracks in the larger space.

Marchenko-Pastur threshold: Dark matter features are defined as eigenvectors with eigenvalues above the Marchenko-Pastur random matrix theory upper edge, $\lambda_+ = \sigma^2(1+\sqrt{q})^2$ where $q = n_{\text{features}}/n_{\text{samples}}$ is the dimension-to-sample aspect ratio of the covariance estimate, ensuring they represent organized structure rather than sampling noise (implementation: `caz/manifold_detector.py` `_mp_upper_edge`, run via `caz/detect_manifolds.py`).

Velocity smoothing window: The velocity metric's adaptive window rule ($k = \max(1, \lfloor L/24 \rfloor)$, main text §2.3) is described for completeness but is not the window that generated the released artifacts: velocity arrays — including those plotted in Figure 3 — were produced with `compute_layer_metrics`'s default fixed window $k = 3$. 0 of the 476 stored velocity series match the adaptive rule. This has no effect on any reported count: the scored detector (§2.4) takes the separation series as its only input and never reads velocity, so the discrepancy is a description-vs-artifact mismatch, not a result-affecting one.

---

§B. Per-Model Behavioral Profiles

Full profiles for all 37 models (28 base + 9 instruct-tuned). Base-model structural rows were regenerated 2026-07-22 on the corrected 17-concept N=250 corpus (see the Summary scope note below); persistent-feature counts come from the main text's §7 census. Derived from **17-concept CAZ extraction** (agency, authorization, causation, certainty, credibility, deception, exfiltration, formality, moral_valence, negation, plurality, sarcasm, sentiment, specificity, temporal_order, threat_severity, urgency) with scored multi-modal detection and spectral deep dive (SVD, cos-threshold feature tracking). Instruct-tuned variants are included for completeness; full alignment training analysis is reserved for future work.

Column key: Layers = transformer layer count; Hidden = residual stream dimension; CAZs = total detected allocation zones; Score = mean geometric CAZ score; Func score = mean functional CAZ score (GQA/alternating only; equals Score for MHA); Multimodal = fraction of concepts with 2+ peaks; Embed CAZ = concepts with an L0 CAZ peak (L0 = output of transformer block 0, the shallowest analysed layer; the token-embedding output is dropped before analysis — see main §2 layer-indexing convention); Maj = major CAZes (score > 0.5); %Gentle = fraction of CAZes with score < 0.05; Features = persistent spectral features (5+ layer lifespan). Cross-model universal-feature matching is not tabulated here: the dark-matter census is in main §7, and cross-architecture alignment is outside this paper's scope (main §5.5 records P5 as deferred; the alignment analysis itself is [Henry, 2026d, in preparation]) — the earlier "UFs" column was a stale N≈200 atlas quantity and has been dropped pending an N=250 atlas relabel.

> **Provenance / refresh status (N=250 rebuild).** Structural columns (CAZs, Score, Multimodal, Embed CAZ, Maj, %Gentle) for **all 28 base models** were regenerated from the corrected N=250 artifacts on 2026-07-22 (`scripts/supplementary_b_rows.py`), post-exfiltration-correction. **Features column** is from the N=250 spectral deep-dive / §7 census (cos-threshold 0.5, 5+ layer lifespan). **Instruct-variant structural columns (CAZs through %Gentle) remain N≈200 (7-concept)** — the 17-concept instruct rerun is pending; counts are not directly comparable to base model rows. The former UFs (cross-model atlas match) column has been dropped — it was a stale N≈200 quantity the main text never cites; the dark-matter census lives in main §7, and cross-model alignment is out of scope here (main §5.5; [Henry, 2026d, in preparation]). Key N=250 finding: %Gentle is low across all three paradigms (paradigm means 11–20%; family means 6–23%) at N=250 with 17 concepts — the earlier N≈200 "high %Gentle" picture reflected the 7-concept subset, not a structural property.

Summary by Paradigm

The tables and summary statistics below cover all 28 base models — MHA = 18 (Pythia ×8, GPT-2 ×4, OPT ×5, Phi-2 ×1), GQA = 8 (Qwen ×5, Llama ×2, Mistral ×1), Alternating = 2 (Gemma-2 ×2) — on the corrected 17-concept N=250 corpus. Regenerated 2026-07-22 from the `paper_n250` artifacts after the exfiltration label correction (`scripts/supplementary_b_rows.py`, cross-checked to the detector's 1,045-region /

363-multimodal aggregate); the earlier 26-model recompute predated that correction. The GQA multimodal rate now reads 94%, matching the main text's full-roster figure (§4.2); the other paradigm multimodal rates (MHA 70%, Alternating 65%) likewise match the main-text §4.2 breakdown.

Paradigm	Models	Mean CAZ/model	Mean score	Mean func score	Artifacts	Multi-modal rate	Mean %Gentle
MHA	18	31.7	0.524	0.524	0	70%	11%
GQA	8	50.0	0.272	0.095	0	94%	21%
Alternating	2	37.5	0.397	0.358	41	65%	16%

MHA: functional = geometric (full Fisher trust). GQA: functional scores ~35% of geometric. Alternating (Gemma): 41 artifact regions total across 2 base models (local-attention layers, near-zero functional impact); functional bottleneck is the final global attention layer. Major CAZes (score > 0.5) appear across all MHA families at N=250 — no longer concentrated in GPT-2 only. %Gentle is low and architecture-flat at N=250; the N≈200 pattern of high %Gentle in sparse architectures is an artifact of the smaller 7-concept corpus.

Pythia (EleutherAI) — MHA, RoPE, concentrated score profile

Model	Layers	Hidden	CAZs	Score	Multi-modal	Embed CAZ	Maj	%Gentle	Features
pythia-70m	6	512	28	0.446	65%	0	5	14%	0
pythia-160m	12	768	49	0.233	100%	0	4	18%	2
pythia-410m	24	1024	41	0.325	88%	0	7	12%	33
pythia-1b	16	2048	28	0.493	65%	1	10	14%	20
pythia-1.4b	24	2048	30	0.483	76%	1	10	0%	27
pythia-2.8b	32	2560	33	0.389	76%	0	8	18%	27
pythia-6.9b	32	4096	34	0.365	88%	0	8	12%	30
pythia-12b	36	5120	33	0.375	71%	0	8	12%	29

CAZ scores vary non-monotonically with scale at N=250 (0.446 → 0.233 → 0.325 → 0.493 → 0.483 → 0.389 → 0.365 → 0.375), reflecting interplay of depth and the 17-concept sampling. Multimodality is non-monotonic across the ladder (65–100%), peaking at 160m (100%). Embedding CAZes appear at 1b (plurality) and 1.4b (deception) — both concepts with strong lexical-level tokenizer signatures; absent at 2.8b through 12b where the deeper residual stream defers lexical signals. Persistent feature count (5+ lifespan, N=250) is zero at 70m (6 layers — the lifespan threshold spans essentially the full model), rises sharply at 410m (33), then plateaus at 20–30

for 1b through 12b. %Gentle is low and non-monotonic (0–18%, with pythia-1.4b at 0%) — no systematic scale trend, consistent with §8.3's no-scale-trend conclusion. Major CAZes present at all scales (4–10 per model). (pythia-12b Features = 29 from the §7 census.)

GPT-2 (OpenAI) — MHA, learned positions, concentrated score profile

Model	Layers	Hidden	CAZs	Score	Multi-modal	Embed CAZ	Maj	%Gentle	Features
gpt2	12	768	20	1.124	18%	0	13	0%	10
gpt2-medium	24	1024	22	1.082	24%	0	13	14%	30
gpt2-large	36	1280	32	0.417	76%	0	8	6%	32
gpt2-xl	48	1600	29	0.560	65%	0	11	10%	29

Only family with learned (non-rotary) position embeddings. GPT-2 base and medium have the highest geometric CAZ scores in the entire corpus (1.124 and 1.082) — unusually high-prominence, low-multimodality encoding at small scale: each concept has a single dominant peak with high separation relative to baseline. At large and XL scale, multimodality rises (76%, 65%) and scores fall (0.417, 0.560) as separation distributes across more layers. Zero embedding CAZes at any scale — learned positional embeddings may suppress L0 lexical signals. GPT-2 predates The Pile (2019 vs late 2020); diverges from Pythia/OPT on credibility, sentiment, negation alignment (0.18–0.49 cosine vs 0.85+ within-family — N≈200 UF-atlas figures, pending the N=250 relabel; not comparable to any cross-architecture alignment analysis, which is out of scope here — main §5.5, [Henry, 2026d, in preparation]). %Gentle low (0–14%). Major CAZes high at all scales (8–13).

OPT (Meta) — MHA, learned positions, concentrated score profile

Model	Layers	Hidden	CAZs	Score	Multi-modal	Embed CAZ	Maj	%Gentle	Features
opt-125m	12	768	23	0.722	35%	0	10	4%	15
opt-350m	24	1024	34	0.307	88%	0	10	29%	19
opt-1.3b	24	2048	33	0.522	76%	0	9	0%	33
opt-2.7b	32	2560	36	0.528	88%	0	14	8%	32
opt-6.7b	32	4096	33	0.569	82%	1	15	18%	29

High scores at small scale (opt-125m: 0.722) consistent with N≈200; the N=250 17-concept picture adds Maj=10 at that scale. Only opt-6.7b shows an embedding CAZ (temporal_order at L0); other sizes have no L0 peaks. Deep CAZes (>80% depth) more

common than other MHA families — OPT shows stronger late-layer separation. Note: opt-350m has `embed_proj_dim=512 ≠ hidden_size=1024`; this dimension mismatch at layer 0 is handled by closing feature tracks at the boundary (see §A). Major CAZes are substantial at every scale (9–15), highest at 2.7b–6.7b (10, 10, 9, 14, 15). %Gentle is low (0–29%) with no monotonic trend.

Phi-2 (Microsoft) — MHA, RoPE, concentrated score profile

Model	Layers	Hidden	CAZs	Score	Multi-modal	Embed CAZ	Maj	%Gentle	Features
phi-2	32	2560	32	0.486	71%	2	13	6%	27

Synthetic textbook training produces unique geometry (§3.2). Embedding CAZes for credibility and deception — both concepts with lexical-level surface signals Phi-2's synthetic training fails to abstract past the tokenizer. Score=0.486 and Maj=13 are mid-range for MHA, contrasting sharply with the N≈200 picture (Score=0.159, Maj=0) — the 10 additional concepts reveal substantially more structured encoding. %Gentle=6%, very low. Early affective concepts (moral_valence at 25% depth, temporal_order at 28%) show mid-range coherence (0.11–0.18) — early but unconsolidated assembly. Persistent features (N=250): 27 — typical for a 32-layer MHA model. The Procrustes-based assignment run found zero cross-model UF matches, which grounds the geometric-isolate reading (that run is an N≈200 atlas quantity, not a result this paper reports; cross-architecture alignment is out of scope — main §5.5, [Henry, 2026d, in preparation]). Geometric isolate (per the Procrustes run) — training data matters more than architecture. See §3.2 in main preprint.

Qwen 2.5 (Alibaba) — GQA, RoPE, SwiGLU, distributed score profile

Model	Layers	Hidden	CAZs	Score	Func score	Multi-modal	Embed CAZ	Maj	%Gentle	Features
Qwen2.5-0.5B	24	896	52	0.290	0.102	100%	0	8	13%	24
Qwen2.5-1.5B	28	1536	55	0.230	0.081	94%	1	4	20%	31
Qwen2.5-3B	36	2048	65	0.185	0.065	100%	2	6	28%	43
Qwen2.5-7B	28	3584	51	0.278	0.097	94%	1	6	24%	33
Qwen2.5-14B	48	5120	52	0.282	0.099	94%	1	6	17%	35

Instruct variant	Layers	Hidden	CAZs†	Score†	Func score†	Multi-modal†	Embed CAZ†	Maj†	%Gentle†	Features
Qwen2.5-0.5B-Instruct	24	896	34	0.224	0.078	100%	6	2	41%	187
Qwen2.5-1.5B-Instruct	28	1536	30	0.187	0.065	100%	4	0	47%	158
Qwen2.5-3B-Instruct	36	2048	41	0.143	0.050	100%	0	0	56%	154
Qwen2.5-7B-Instruct	28	3584	34	0.142	0.050	100%	1	0	41%	179

† Instruct structural columns are N≈200 (7-concept); not directly comparable to base model rows.

Highest CAZ density of any GQA family at N=250 (51–65 regions per model — 17 concepts). Embedding CAZes absent at 0.5B; appear at 1.5B+ driven by deception, exfiltration, and specificity — concepts with strong surface-form signals in smaller Qwen's tokenizer. Functional scores ~35% of geometric — GQA distributes encoding across layers. Persistent feature count (N=250) 24–43, with 3B having the most (43) — the deepest Qwen base model carries fewer despite its greater depth (Qwen2.5-14B: 35), so the count does not scale monotonically with layer depth. Major CAZes present at all scales at N=250 (8, 4, 6, 6, 6) — absent only in the N≈200 7-concept view. %Gentle 13–28%, considerably lower than the N≈200 picture (41–61%) — the 10 additional concepts reveal structured encoding the 7-concept subset missed. (Qwen2.5-14B Features = 35 from the §7 census.) Instruct structural columns remain N≈200; no directional claim is made pending the 17-concept instruct rerun (the printed instruct scores are not comparable to the N=250 base rows).

Llama 3.2 (Meta) — GQA, RoPE, SwiGLU, distributed score profile

Model	Layers	Hidden	CAZs	Score	Func score	Multi-modal	Embed CAZ	Maj	%Gentle	Features
Llama-3.2-1B	16	2048	36	0.346	0.121	76%	0	6	25%	12
Llama-3.2-3B	28	3072	44	0.280	0.098	100%	3	10	23%	24

Instruct variant	Layers	Hidden	CAZs†	Score†	Func score†	Multi-modal†	Embed CAZ†	Maj†	%Gentle†	Features
Llama-3.2-1B-Instruct	16	2048	17	0.298	0.104	86%	0	0	35%	196
Llama-3.2-3B-Instruct	28	3072	28	0.167	0.058	100%	1	0	57%	181

† Instruct structural columns are N≈200 (7-concept); not directly comparable to base model rows.

Fewer CAZes than Qwen at comparable scale (36–44 vs 51–65). Major CAZes present at N=250 (6 and 10) — previously absent in the 7-concept run. Embedding CAZes at 3B (deception, plurality, specificity); absent at 1B. Persistent feature count (N=250): 1B=12, 3B=24 — rises with scale. Instruct feature counts remain N≈200 (§7). Gated model (requires HF license).

Mistral (Mistral AI) — GQA, RoPE, SwiGLU, distributed score profile

Model	Layers	Hidden	CAZs	Score	Func score	Multi-modal	Embed CAZ	Maj	%Gentle	Features
Mistral-7B-v0.3	32	4096	45	0.284	0.099	94%	3	7	16%	36
Mistral-7B-Instruct-v0.3	32	4096	28†	0.055†	0.019†	100%†	1†	0†	71%†	36

† Instruct structural columns are N≈200 (7-concept); not directly comparable to base model rows.

N=250 with 17 concepts substantially revises the N≈200 picture: Score rises from 0.061 to 0.284, Maj from 0 to 7, %Gentle from 73% to 16%. The 10 additional concepts reveal structured mid-layer encoding that was undersampled in the 7-concept run. Embedding CAZes at 3 concepts (credibility, plurality, sarcasm) — more than other single-paradigm GQA models at 7B. Persistent feature count 36 — second-highest among GQA base models (Qwen2.5-3B: 43) — grouped-query attention distributes separation while long-range spectral coherence is preserved (Mistral-7B-v0.3 uses full, non-windowed attention; the sliding-window attention of v0.1 was removed at v0.2). Mean CAZ depth ~54% — deepest of the GQA family. Zero cross-model UF matches for the base model.

Gemma 2 (Google) — alternating local/global attention, GQA, RoPE, GeGLU

Gemma 2 uses alternating sliding-window (odd layers, local) and global attention (even layers). Fisher separation peaks at local-attention layers have zero ablation impact and are classified as `artifact`. The functional bottleneck is the final global attention layer — confirmed at L24 for gemma-2-2b, where activation patching recovery is ~1.0 for all concepts (§6.4); the corresponding layer for gemma-2-9b (L40) is architecture-inferred and patching-confirmed (§F Table S3, recovery 0.947–1.039 across all seven tested concepts at L40). **Measurement caveat (main §6.9):** this family's single-direction geometric estimates do not converge at N=250 (concept information is present at control level under held-out probing but distributed across ~20–78 effective dimensions), so the geometric columns in the table below — score, %Gentle, major-CAZ counts — carry substantially wider error than other families'; the patching-based functional results are unaffected. Persistent feature count (N=250): 21 for 2b, 17 for 9b — mid-range across the full corpus, not the lowest. Local/global alternation produces sparser CAZ geometry without suppressing persistent spectral structure.

Model	Layers	Hidden	CAZs	Score	Func score	Artifacts	Func peaks	Maj	%Gentle	Features
gemma-2-2b	26	2304	36	0.400	0.376	19	17	13	17%	21
gemma-2-9b	42	3584	39	0.394	0.340	22	17	12	15%	17

Instruct variant	Layers	Hidden	CAZs†	Score†	Func score†	Artifacts†	Func peaks†	Maj†	%Gentle†	Features†
gemma-2-2b-it	26	2304	34	0.076	0.056	3	5	0	65%	12
gemma-2-9b-it	42	3584	31	0.173	0.174	26	2	0	39%	5

† Instruct structural columns are N≈200 (7-concept); not directly comparable to base model rows. gemma-2-9b-it values from GEM run 2026-04-30; ablation run 2026-05-06.

At N=250 (17 concepts): artifacts = local-attention-layer peaks, 19 for 2b (2 concepts produce a second local-attention peak) and 22 for gemma-2-9b (42 layers accumulate more across the deeper stack). Functional peaks = 17 per model (one per concept at or near the final global attention layer). Maj (13/12 geometric) are driven by the functional peak regions — the scoring algorithm sets `functional_caz_score = max_caz_score × 1.0` for the verified bottleneck layer, so these are genuine high-score functional regions, not artifacts. %Gentle 15–17% — substantially lower than N≈200 (72–74%) due to the expanded concept set. Both gemma-2-9b variants run in bfloat16 sharded across 2× L4 (Table 1).

Cross-Family Observations

Score distribution and CAZ geometry. MHA models concentrate score mass into fewer, higher-scoring CAZes (mean 31.7/model, score 0.524); GQA models distribute it across more, weaker regions (50.0/model, score 0.272). This is a *score-distribution* difference, not an ablation-efficacy one — the main text retires the earlier "redundant vs. sparse encoding" reading at C=17 (main §6.4/§8.3: MHA and GQA are comparable on ablation). Functional/geometric score ratios diverge most sharply in GQA (functional ≈ 35% of geometric); in the alternating architecture (Gemma) the divergence instead concentrates at local-attention artifact layers, while the aggregate ratio stays high (≈ 89%) because scoring re-weights the architecture-determined functional peak.

%Gentle and Major CAZes at N=250. The N≈200 picture of high %Gentle in sparse architectures is an artifact of 7 concepts. At N=250 (17 concepts), %Gentle is low and architecture-flat across all families:

Family	%Gentle (base mean, N=250)	Major CAZes (base total, N=250)
Pythia (MHA)	12%	60 across 8 models
GPT-2 (MHA)	8%	45 across 4 models
OPT (MHA)	12%	58 across 5 models
Phi-2 (MHA)	6%	13 (1 model)
Qwen (GQA)	20%	30 across 5 models
Llama (GQA)	24%	16 across 2 models
Mistral (GQA)	16%	7 (1 model)
Gemma-2 (Alternating)	16%	25 across 2 models (incl. artifact peaks)

%Gentle still increases slightly from MHA (~11%) to GQA (~21%) to alternating (~16%), but the contrast is far smaller than at N≈200. On functional trust the sharpest break is between MHA (full Fisher trust, no artifact correction) and GQA (functional trust ≈35%); Gemma-2's alternating architecture is a distinct case — high aggregate functional/geometric ratio (≈89%) but with artifact regions requiring the final-global-layer correction. Note the Major-CAZ totals above are the detector counts over the 28 base models (254 total), consistent with the main text's §8.3 tally (254; Table 2).

Scale effects. CAZ score varies non-monotonically with scale at N=250 — the clean decreasing-score-with-scale picture from N≈200 (driven by the 7-concept subset)

does not hold for all families. Multimodality generally increases with depth. Embedding CAZes appear at specific conceptual hotspots (deception, plurality, credibility, specificity) rather than uniformly above a parameter threshold. Persistent feature count (N=250) is driven more by architecture and depth than by parameter scale: very small models with few layers (pythia-70m at 6 layers: 0; gpt2 at 12 layers: 10) are constrained by the 5-layer lifespan threshold; beyond that, counts are roughly 20–43 with no strong scaling trend within families.

Architectural outliers. Phi-2 is a geometric isolate — its training process (synthetic data, distillation) correlates with zero cross-model feature matches in the Procrustes assignment run and unique concept ordering (§3.2); causal attribution is unresolved. GPT-2 diverges from same-era MHA models on epistemic and affective geometry, consistent with its pre-Pile training corpus. Mistral's flatter score profile (0.284) sits mid-pack in the GQA cohort, consistent with grouped-query attention rather than any windowing — Mistral-7B-v0.3 uses full attention (the sliding-window attention of v0.1 was removed at v0.2).

---

§C. Multi-Generator Consensus Replication

Run: 2026-04-12. Four models, seven concepts, ~1,400 pairs per concept (100 topics × 14 generators across Anthropic, Google, OpenAI, and Mistral). Extraction run identically to main study; depth reported as % of total layers.

Peak assembly depth (% of total layers) is compared between the single-generator dataset (100 pairs, Claude Sonnet 4.6) and the multi-generator consensus dataset (~1,400 pairs). Values marked † indicate embedding-layer leakage (peak at L0, flagged by extraction pipeline); these pairs are excluded from ordering statistics. Negation consensus pairs are incomplete for all models (1,322/1,400 pairs; some generators did not produce valid negation contrasts for all topics).

Table S1: Peak depth comparison — MHA models.

Concept	Pythia-1.4b (MHA, 24L)			GPT-2-XL (MHA, 48L)		
	Single	Multi	Δ	Single	Multi	Δ
causation	45.8%	45.8%	0	39.6%	52.1%	+12.5
credibility	58.3%	62.5%	+4.2	10.4%	60.4%	+50.0
negation	33.3%	16.7%	−16.6	45.8%	35.4%	−10.4
sentiment	50.0%	41.7%	−8.3	47.9%	50.0%	+2.1
certainty	54.2%	50.0%	−4.2	60.4%	47.9%	−12.5
moral valence	50.0%	33.3%	−16.7	52.1%	47.9%	−4.2
temporal order	54.2%	50.0%	−4.2	62.5%	43.8%	−18.7

Table S2: Peak depth comparison — GQA models.

Concept	Qwen2.5-3B (GQA, 36L)			Llama-3.2-3B (GQA, 28L)		
	Single	Multi	Δ	Single	Multi	Δ
causation	72.2%	72.2%	0	42.9%	64.3%	+21.4
credibility	0.0%†	75.0%	—	0.0%†	71.4%	—
negation	41.7%	0.0%†	—	25.0%	0.0%†	—
sentiment	72.2%	86.1%	+13.9	60.7%	78.6%	+17.9
certainty	77.8%	75.0%	−2.8	64.3%	60.7%	−3.6
moral valence	75.0%	77.8%	+2.8	25.0%‡	78.6%	+53.6
temporal order	72.2%	80.6%	+8.4	53.6%	64.3%	+10.7

‡ Both negation and moral valence peak at L7 (25.0% of 28 layers) in the Llama-3.2-3B single-generator run. Verified in raw data: negation S=0.578, moral valence S=0.725 — different separation magnitudes confirm these are distinct measurements, not a copy error.

Summary. Category-level ordering structure survives generator diversity: the deep epistemic/relational concepts stay deep in the multi-generator condition in all four models, causation is exactly stable (Δ = 0) in two of four, and negation is the shallowest concept in five of the six model×corpus conditions where its peak is unflagged (the exception is GPT-2-XL's single-generator run, where the unflagged shallow credibility value of 10.4% sits lower — see pattern 3 below). Per-cell shifts are nonetheless substantial: of the 24 measurable model×concept cells in Tables S1–S2, 10 fall within $|\Delta| \leq 5$ pp and the median $|\Delta|$ is ≈9.4 pp; **no concept satisfies $|\Delta| \leq 5$ pp across all four models**. Three patterns emerge from the shifts:

1. **Most-stable concept — certainty** (within 5 pp in three of four models: −4.2, −2.8, −3.6; the exception GPT-2-XL at −12.5). Causation is exactly stable where it is stable (Δ = 0 in Pythia-1.4b and Qwen2.5-3B) but shifts by +12.5 to +21.4 pp in the other two models. No concept is robust to generator diversity across the full model set at the 5 pp bar.

2. **Leakage swap — credibility and negation**: In both GQA models, single-generator credibility pairs trigger embedding-level leakage (L0) while multi-generator pairs resolve to deep epistemic positions (60–75%). The reverse holds for negation: single-generator pairs produce clean mid-depth peaks; multi-generator pairs trigger L0 leakage, consistent with 14 generators preferentially using explicit negation markers (*not*, *never*, *no*) that Qwen and Llama tokenizers resolve at the embedding layer. This swap is absent in MHA models, which are less sensitive to surface lexical features at the embedding stage.

3. **Credibility variance**: GPT-2-XL shows a large credibility shift (10.4% → 60.4%) between datasets without triggering the L0 flag — a shallow-but-not-leakage artifact in the single-generator pairs that resolves with generator diversity. This is consistent with the credibility puzzle described in §3.3: the concept is built from unnamed primitives that different generators operationalize differently, produc-

ing genuine multi-site assembly rather than a stable single peak.

§D. Exfiltration: Corpus Label Defect and Correction

Added: 2026-07-17 00:21 UTC. Reframed 2026-07-17 00:50 UTC — decision changed from exclusion to full correction (James, 2026-07-16). Finalized 2026-07-18 12:19 UTC — corrected artifacts applied and re-pooled across the paper (D.2/D.3 outcome, net-effect note); editorial banner removed.

All exfiltration results in the main text are computed from a corrected re-analysis: a data defect was discovered in the concept's source corpus after extraction — and after earlier versions of this paper had reported the defective values (exfiltration appeared as the deepest concept, mean peak depth 85.2%). We document the defect, its measured impact, and the correction protocol here in full, because a headline number that quietly changes between paper versions is exactly the kind of thing a reader should get to audit.

D.1 The defect

The Rosetta Concept Pairs (RCP) corpus is version-controlled, and its history makes the defect precisely datable. Commit `1a61a76` (2026-07-02, "fix(exfiltration): correct inverted labels on 87/107 topics") corrected a label-direction bug in the exfiltration pair generator: on 87 of the concept's 107 topics — 1,454 of 1,938 records — the positive/negative labels were swapped (1,440 of these survive in the corrected corpus and carry a flipped label; the other 14 were among the records subsequently deleted, reconciling the two counts below). Commit `a088c29` (2026-07-07) patched the generator itself and removed bad records corpus-wide. This study's activation corpus (`paper_n250`) was extracted 2026-06-08 through 2026-06-14, **before both fixes**: every model in the study saw the defective labels.

A byte-level diff between the extraction-era corpus state (`b5fa231`) and the corrected state confirms the fix changed **labels only**: of 1,938 original exfiltration records, 1,924 survive with byte-identical text; 1,440 had their `label` field flipped; 14 bad records were deleted; no pair identifiers changed and no text was rewritten.

Resolved against this study's recorded 250-pair draw: **179 of the 250 exfiltration pairs (71.6%) carried swapped labels; 70 were correct; 1 was subsequently deleted as a bad pair.** Because the inversion is partial, its effect is not a harmless sign flip (Fisher separation is symmetric under a global label swap). A 71.6% inversion *mixes* the two classes: the "positive" class as analyzed was ~72% true negatives. This depresses measured separation, corrupts the difference-of-means direction, and can relocate the apparent peak layer.

D.2 Measured impact

The stored activations are label-blind (a forward pass does not consume the label), and the calibration metadata records the draw in row order, so the defect is correctable in post for calibration-level statistics: swap the class assignment for the 179 flipped

pairs, drop the deleted pair. A preliminary post-hoc relabel of three models spanning the corpus scale range first quantified the damage and motivated the full re-run:

Model	Peak separation (defective → post-hoc relabel)	Dominant peak depth (defective → post-hoc relabel)
pythia-70m	0.216 → 0.460	83.3% → 83.3%
gpt2-xl	0.219 → 0.379	97.9% → 64.6%
Qwen2.5-3B	0.224 → 0.458	91.7% → 94.4%

Peak separation was understated 1.7–2.1×, confirming the labels were mixing the classes. The post-hoc relabel suggested peak depth might also move substantially (gpt2-xl by 33 pp), which raised the possibility that exfiltration was not truly the deepest concept and made a full, consistently-extracted re-run necessary rather than a post-hoc patch.

The full re-extraction (D.3) resolves this more mildly than the post-hoc relabel implied. Extracted cleanly across the 28-model corpus, corrected exfiltration's mean dominant-peak depth is **81.0%** (down from 85.2%), with the cross-model standard deviation tightening from 21.7 to 15.0 — the defect was inflating both the depth estimate and its scatter, exactly as a mixed-class contaminant should. Peak separation roughly doubles (mean 0.44 vs. ~0.22). Exfiltration **remains the deepest concept** by mean depth (rank 17 of 17; the next-deepest, sarcasm, is at 70.2%), though it is now the per-model-deepest concept in 8 of 28 models rather than 14 — a minority-of-models property on which the mean-depth ranking nonetheless rests comfortably (81.0% vs. 70.2%). The apparent gpt2-xl discrepancy between the two procedures (post-hoc relabel 64.6%, full re-extraction 97.9%) is not extraction infidelity but multimodality: gpt2-xl's corrected exfiltration profile has a deep peak at the final layer (L47, Fisher separation 0.382) and a broad mid-depth cluster (L26–L30, ≈0.30–0.32), a near-tie. The quick calibration-level relabel resolved to the mid cluster; the full scored detector — which the paper uses throughout — resolves to the deep peak. Because the labels are deterministic and the texts byte-identical, relabel-in-place and re-extraction produce the same per-layer separations to numerical precision; they differ only in which peak of a genuine near-tie the reported summary selects, which is a property of this concept's multimodality (§4.2), not of the correction. The main text's C=17-era domain-specificity caveat (§2.2), which flagged exfiltration's depth as interpretable "cautiously," turns out to have been warranted for a reason beyond the one given.

Two corpus-wide observations are retroactively explained by this defect and should not be read as properties of the concept or of any model family: (1) exfiltration was the noisiest concept in cross-draw direction-stability comparisons for *every* model (e.g. GPT-2 0.82, Qwen2.5-3B 0.85 cross-draw DOM cosine vs. ≈0.99 on their other concepts) — different draws contain different fractions of flipped pairs, so the mixed-class direction wanders; (2) a post-fix re-extraction produces an exfiltration direction that *anti-correlates* (cosine ≈ −0.7) with the stored pre-fix artifacts — expected by construction, since a majority- inverted labeling approximately negates the difference-of-means direction.

D.3 The correction

The calibration-level relabel shown above is cheap, but every downstream causal result (ablation, global sweep, patching, multimodal protocols) consumes the corrupted direction, so exfiltration's full per-concept pipeline was re-run rather than patched. The correction uses the **same 249 surviving pairs of the original recorded draw** — texts recovered from the extraction-era corpus commit (byte-identical, §D.1), corrected labels overlaid, the one deleted pair dropped — not a fresh sample from the repaired corpus, whose pool no longer matches the draw the other sixteen concepts are documented against. Activations were freshly extracted from this reconstructed set through the identical extraction pipeline and environment as the original corpus, and all downstream exfiltration artifacts were regenerated from them. One downstream parameter initially differed — the exfiltration re-run's GEM handoff ablation used a 3-layer width where the other sixteen concepts use a 1-layer width — but exfiltration has since been re-derived at the standard 1-layer width (n = 249; `paper_n250/_p2_exfil_width1/`), matching the other concepts; §8.7 reports the negligible effect on the handoff-vs-peak counts (net −2 of 442, with Phase-1 and the Gemma subset unchanged). Correction provenance (the reconstructed pair set, the per-pair flip list, and gate checks against the relabel demonstration in §D.2) is published alongside the corrected artifacts. The defective originals remain retrievable in the frozen `paper_n250` snapshot revision, so the before/after is fully auditable.

Because the corrected analysis was run after the defect (and its likely direction of effect) was known, exfiltration's results carry a different epistemic status than the other sixteen concepts' — a distinction we flag rather than hide. The correction changes the data to what it was always specified to be; it does not select among analyses.

Net effect on the paper. Every result touched by exfiltration was re-pooled from the corrected artifacts and compared against the defective-corpus value. No central conclusion changed, and one summary statement was corrected (below). Several statistics improved once the noisiest concept carried clean labels: the split-half reliability ceiling rose from 0.894 to 0.911 (§3.2), the GEM handoff-vs-peak advantage from 33/34 to 34/34 and its full-corpus extension from 297/442 to 304/442 (§6.5), and Table 3's exfiltration variance tightened (std 21.7 → 15.0). The corrected values shifted the aggregate ordering statistics slightly (median τ 0.417 → 0.404), and one summary statement required rewording: GPT-2-medium's ordering τ, previously 0.008, is exactly zero under the corrected reference, so the paper reports 27 of 28 models positively correlated with one at zero rather than "all 28 positive" (§3.1).

D.4 The corpus-wide bad-pair cleanup (second fix)

Commit `a088c29` also deleted 2,001 bad records across sixteen concept files corpus-wide. Resolved against this study's draws: 156 of 4,250 drawn pairs (3.7%) were later deleted as bad, spread thinly across concepts (at most 16 of 250 for any concept; zero for four concepts). We disclose rather than recompute: the C14 calibration-draw sensitivity analysis (8 models × 17 concepts, 50 bootstrap re-draws within each 250-pair set) bounds the effect of far larger draw perturbations than a ≤6.4% deletion — mean direction-of-means cosine between bootstrap and full-sample directions is 0.990 (minimum 0.868 across all 136 model-concept cells), and peak-depth locations

are stable under resampling. The bootstrap perturbation is random and mean-zero whereas the deletion is a directed removal of a systematically defective subset, so the two are not identical in kind; but the deletion is far smaller in scale (≤6.4% vs. the ~37% expected omission of a bootstrap resample), and the directions' demonstrated robustness to the larger random perturbation makes a material effect from the smaller directed one unlikely. For the sixteen non-exfiltration concepts we therefore disclose rather than recompute; the one concept whose defect was both large and directed (exfiltration) is the one we fully re-ran.

D.5 Reproducibility note

Because the RCP pool changed after extraction, re-running the study's sampler against the current corpus does not reproduce the recorded draws (same seed, smaller pool, different subset). Reproducibility of this study is anchored instead by (1) the recorded per-concept pair manifests in the calibration metadata (draw order and class layout) — present for all 28 models on exfiltration (the re-run) and for 6 models on each of the other 16 concepts, i.e. 124 of the 476 cells carry an explicit manifest; the remaining cells rely on (2) and (3) below — (2) the extraction-era RCP commit (`b5fa231`), retrievable from the corpus git history, and (3) the frozen `paper_n250` activation snapshot on Hugging Face. All three together reconstruct exactly what every model saw, including the defective labels this section documents.

The sampler itself is deterministic from the extraction code: extraction ran from `rosetta_tools` commit `6fed9e2` (2026-06-08), which replaced the earlier process-randomized `hash((concept, split))` seed with a stable per-concept seed. The `v1.3.1` git *tag* predates this fix and retains the `hash()`-based seeding (which is `PYTHONHASHSEED`-salted and draws a different subset per invocation), so the extraction commit — not the tag — is the reproduction pin (§9). The 6 models carrying manifests all drew the identical subset, confirming determinism from that commit.

---

§E. Concept-Ordering Robustness Batteries

Supporting evidence for main-text §3.2. §3.2 reports that the depth-ordering tendency (median τ = 0.404; 27 of 28 base models positive with the 17-concept grand-mean ordering, one exactly zero, none negative; Wilcoxon p = 1.49×10⁻⁸) survives every robustness control applied and that training-token frequency is a real but partial contributor at the shallow end. The full derivations are here; no value in this section differs from what §3.2 summarizes.

Phi-2 revisited. At C=7, Phi-2 showed $\tau = -0.25$ — the most extreme of three negative-τ models (alongside GPT-2 and GPT-2-medium) at that concept count, interpreted as an ordering inversion driven by its synthetic-textbook training. At C=17, Phi-2 has $\tau = 0.418$ (positive, mid-pack). Its 7-concept ordering remains anomalous: credibility and deception now peak at embedding depth (0%), while several relational/epistemic concepts (temporal_order, moral_valence, certainty, sentiment) sit at 22–28% — shallower than average. But the 10 additional concepts — particularly the syntactic cluster (specificity 12.5%, plurality 3.1%, negation 18.8%) and the

deep cluster (agency, exfiltration, formality at 81–87.5%) — align with Phi-2's actual shallow/deep structure, pulling its τ positive. This illustrates a methodological point: a per-model ordering index is sensitive to the concept set chosen. The C=7 Phi-2 anomaly was real (the early credibility/deception assembly is reproduced here with the same model), but it was also in part a measurement artifact of a limited reference concept set. **The current C=17 analysis should be considered the authoritative result.** Phi-2's shallow embedding-level resolution of credibility and deception is documented in §3.3; it is no longer an ordering outlier at scale.

Low-τ models. The lowest-τ models at C=17, setting aside the Gemma-2 pair treated separately in §6.9 (gemma-2-2b's $\tau = 0.100$ would otherwise place it third), are gpt2-medium ($\tau = 0.000$ — exactly zero, §3.1), gpt2 ($\tau = 0.060$), Qwen2.5-0.5B ($\tau = 0.101$), and opt-350m ($\tau = 0.105$). All four are small or shallow: gpt2-medium has 24 layers and 1024 hidden dim; gpt2 has 12 layers; Qwen2.5-0.5B has 24 layers and 896 hidden dim; opt-350m has 24 layers. At this scale, many 17-concept pairs are separated by less than 2 layers in the dominant peak, compressing the rank differences that Kendall's τ measures — which invites the reading that the low τ is limited depth resolution rather than a concept ordering genuinely different from larger models. **The split-half data do not support that reading for three of the four.** Each model's ordering was recomputed independently on each half of its 250 pairs (see "The reliability ceiling", below; `scripts/results/c5_splithalf_tau_ceiling.json`): split-half τ is 0.895 (gpt2-medium), 0.532 (gpt2) and 0.715 (Qwen2.5-0.5B), implying per-model ceilings of 0.83–0.97, while these models agree with the grand mean at only 0.000–0.101. A model that reproduces its own ordering across independent halves at 0.895 and matches the consensus at 0.000 is not failing to resolve the ordering; it is resolving a different one. Depth compression is real (§4.4) and does add noise, but it is not what those three τ values are made of: their low τ is a reliable disagreement, not a measurement failure. **opt-350m is the one exception and we mark it as such:** its split-half τ is 0.310 (per-model ceiling ≈ 0.69), the third-lowest reliability in the corpus after the two Gemma-2 models, so for opt-350m the depth-resolution reading is *not* excluded and its τ = 0.105 should be read as partly a measurement limit. For the other three, the reliable-disagreement reading is a more interesting result than the artifact reading it replaces, and it does not weaken the ordering tendency — it is the same fact that defeats the shared-frequency account below, seen per-model.

Leave-one-out (LOO) robustness. The median τ is computed against the grand-mean ordering, which includes each model in its own reference — a mild self-inclusion bias. Under a LOO recompute (each model's τ is computed against the grand mean of the remaining 27 models), median τ is 0.373, a reduction of roughly 3 percentage points. 27 of 28 models remain positive under LOO; only gpt2-medium — whose full-sample τ is already exactly zero — turns marginally negative (−0.031), consistent with its status as a reliable disagreer rather than a noisy measurement. The Wilcoxon test remains highly significant ($p = 5.25 \times 10^{-6}$). Self-inclusion inflates the reported median τ slightly but does not alter the qualitative conclusion.

Leave-one-family-out (LOFO) and family-level statistics. Because 22 of the 28 models are scale ladders within 4 families, model-level checks understate the dependence structure — a family's siblings vote for their own ordering in every model-level reference. Under a LOFO recompute (each model's τ computed against a reference

ordering built excluding its own family), the median τ is 0.392 — a modest attenuation from 0.404 — and the Wilcoxon test remains significant ($p = 3.7\times10^{-8}$). All 8 families' mean τ are positive — the GPT-2 and Gemma-2 cohorts — the former containing the reliable low-τ disagreers, the latter the alternating-attention pair with unreliable single-direction readouts (§6.9) — sit at the low end, Pythia and OPT at the high end — and the 8×8 between-family ordering matrix is positive in all 28 off-diagonal cells (with the corrected exfiltration ordering the previous lone exception, gemma2 × qwen2, turns positive). The ordering tendency is not an artifact of within-family pseudo-replication.

Training corpus frequency as a confound. A parsimonious alternative to the CAZ account: the concept ordering reflects corpus word frequency rather than architecture-specific allocation dynamics. High-frequency syntactic markers ("not", "no", "several") are seen in nearly every training sentence and may encode early; rare technical or behavioral terms ("exfiltration", "authorization") appear rarely and may require deeper contextualization regardless of architecture. For each concept, we identify the 20 tokens most differentially present between positive and negative RCP pairs (smoothed log-odds of token count, minimum 5 total occurrences, computed over the full raw consensus-pair corpus for that concept) and take their mean Zipf frequency (`wordfreq`, English), then correlate this against mean peak depth across all 17 concepts. The correlation is negative and statistically significant at N = 17 concepts: Spearman $\rho = -0.657$ ($p = 0.004$), Kendall $\tau = -0.456$ ($p = 0.010$). The concept-label frequency itself (Zipf frequency of the concept name — "credibility", "negation", etc. — rather than its discriminative tokens) shows a much weaker, non-significant positive relationship ($\rho = 0.248$, $p = 0.338$), confirming the effect is carried by the actual discriminative vocabulary, not the concept label. Notable exceptions still exist: *formality*'s discriminative tokens are the highest-frequency of any concept (informal markers "get", "it's", "stuff", "you're", "don't") yet it assembles at mid-pack depth (49.7%, §3.1); *specificity* (mean Zipf 4.99, tied-highest) is the shallowest concept, consistent with the frequency account, but its discriminative tokens ("q", "mg", "c") look more like measurement-unit artifacts of the contrastive pair construction than genuine high-frequency words.

Partialling frequency out of the ordering. The concept-level correlation above says frequency predicts *where concepts sit in the grand-mean ordering*; it does not say frequency produces the cross-model agreement that the ordering result rests on. Partial rank correlation separates the two: computing each model's Kendall τ against the grand mean while controlling for concept-level mean Zipf frequency gives a median partial τ of **0.380**, against 0.404 uncontrolled — an attenuation of 0.024 (6%). All 28 models are positive on the partials (including gpt2-medium, whose raw τ is zero), and the Wilcoxon signed-rank test on the partials is at its extreme ($W = 406$, $p = 3.7\times10^{-9}$, one-sided). The models whose orderings run *against* the frequency account (both Gemma-2 models, GPT-2-medium, Mistral-7B, Qwen2.5-0.5B — each showing positive τ with frequency where the account predicts negative) have their agreement with the grand mean *rise* once frequency is removed. Whatever the ordering tendency is, removing the frequency signal costs it four hundredths.

The reliability ceiling. A second counter-argument concerns τ itself: if training corpus frequency were the primary driver, and all 28 models were trained on similar

internet-text corpora, every model should produce the same ordering. The naive form of this argument compares the observed median $\tau = 0.404$ against a ceiling of 1.0, which is not the right comparison — measurement noise attenuates τ below 1.0 whatever the truth is, so the ceiling must be estimated rather than assumed. We estimate it directly from the calibration pairs. Recomputing each model's full 17-concept ordering independently on each half of its 250 pairs (odd/even split; the same detection pipeline, which reproduces the stored per-layer artifacts to 2×10^{-16}) gives a median split-half τ of 0.717 at 125 pairs, or 0.835 at the full 250 after a Spearman–Brown correction; the 28-model grand mean, being an average, is far more stable (0.993). If every model shared one true ordering, the largest median τ we could then expect to *observe* is $\sqrt{(0.835 \times 0.993)} = \mathbf{0.911}$ (0.921 computing the same quantity on Spearman ρ, where the rank attenuation is better behaved). The observed 0.404 is less than half of that. Noise does not close the gap: the shared-ordering account predicts ~0.91 and the data give 0.40.

A second test without the attenuation formula. Under a shared ordering, model *m*'s first half and model *n*'s second half are two noisy reads of the same thing, so within-model agreement τ(A_m, B_m) and between-model agreement τ(A_m, B_n) should be equal. They are not: within-model mean $\tau = 0.651$, between-model mean $\tau = 0.254$ — models resemble themselves far more than they resemble each other (ratio 0.39, where 1.0 would mean interchangeable). Both tests point the same way, and neither depends on the other's assumptions. Shared corpus frequency would predict convergence; the observed τ indicates that architecture, scale, and model-specific properties also shape the ordering beyond the shared training signal. Frequency remains a genuine and significant correlate of *where concepts land on average* ($\rho = -0.657$) — the two facts are compatible. The bag-of-words probe baseline intended as the direct surface control turns out to be uninformative for the ordering question: held-out BoW classification sits at ceiling for essentially every concept (AUC ≈ 1.0, §2.2), because contrastive minimal pairs are lexically separable by construction, so no per-concept lexical-difficulty signal exists to correlate against depth. The surface-vs-depth discrimination therefore rests on the three tests above plus one more from the corpus-quality battery: dominant directions re-estimated from topic-disjoint halves of the calibration set agree as well as random halves do (mean cosine 0.957 vs. 0.960 across an 8-model subset; `scripts/c14_calibration_topic_sensitivity.py`), so the directions the ordering is computed from are not topic-driven. An off-ceiling BoW redesign was completed as a control (surface difficulty does not reproduce the depth ordering; §2.2).

Frequency as a partial contributor. Given the token-frequency correlation is significant, this confound deserves more than a dismissal: it should be read as a real, partial contributor to the shallow end of the ordering (specificity, plurality, negation — all high discriminative-token frequency, all shallow) rather than a fully independent architectural effect at those three concepts specifically. The mid-to-deep concepts (causation through exfiltration, spanning 55–81% depth) show much less token-frequency spread and are not well explained by this account. A stronger test would require models trained on deliberately different corpora, so the corpus-frequency signal can be dissociated from the architectural signal directly.

---

§F. Architecture-Conditioned Ablation and Patching — Full Detail

Supporting evidence for main-text §6.4. §6.4 reports the cohort-level result (Table 9: MHA and GQA comparable on GEM handoff ablation, 0.588 vs 0.625; Gemma-2 weaker at the Fisher peak, 0.367, but near-complete recovery at its final global attention layer) and the peak-vs-causal divergence headline (50.3% of multimodal cases). The per-cohort elaboration, the Gemma per-concept recovery table, the divergence calibration, the overshoot analysis, and the held-out endogeneity check are here. No value differs from §6.4.

MHA cohort (GPT-2, Pythia, OPT, Phi-2; 2019–2022): Single-CAZ handoff ablation suppresses separation by ~59% at the output (0.588 mean, computed unclipped; 22 of 306 cells show ablation *increasing* separation and contribute negative reductions rather than being truncated at zero) and patching at the CAZ peak recovers a valid concept representation (0.672 raw / 0.533 trimmed). Concepts are not fully removed by a single CAZ because other CAZes downstream continue writing to the concept direction (multimodal allocation, §4.2). In MHA models, identified CAZ peaks are ablation-sensitive and injection-responsive; both metrics fall short of 1.0, indicating partial redundancy even in MHA models.

GQA+SwiGLU cohort (Qwen, Llama, Mistral; 2023–2025): Peak ablation is comparable to the MHA cohort (0.625 vs 0.588), and patching recovery is somewhat higher (0.770 raw / 0.693 trimmed). At the full 17-concept set the two cohorts are comparable on ablation — the larger MHA-vs-GQA gap seen on the earlier 7-concept subset (which motivated a "redundant vs. sparse encoding" reading) does not persist, and we draw no architectural ranking from it. Both cohorts confirm the same methodological point: the detected CAZ peak is a causally active, injection-responsive site.

Gemma-2 (alternating local/global) takes this to its architectural extreme. Local sliding-window attention layers cannot propagate non-local information beyond their context window. The single global attention layer integrates across the full sequence but only once, and its output is subsequently sealed by local layers. Ablating the identified CAZ peak — which may be an early global layer — has only weak effect (0.367 mean; §6.9 shows single-direction ablation is uninformative on Gemma's distributed concept code rather than evidence of inertness). But patching at the **final global attention layer** (L24 in Gemma-2-2b, L40 in Gemma-2-9b) produces near-complete recovery on every concept:

Model	Concept	CAZ recovery	Global layer recovery	Depth (%)
gemma-2-2b	causation	0.325	**1.016**	92%
gemma-2-2b	certainty	0.449	**0.981**	92%
gemma-2-2b	credibility	0.717	**1.050**	92%
gemma-2-2b	negation	0.321	**1.013**	92%

Model	Concept	CAZ recovery	Global layer recovery	Depth (%)
gemma-2-2b	sentiment	0.441	**1.003**	92%
gemma-2-2b	temporal_order	0.235	**1.009**	92%
gemma-2-2b	moral_valence	0.869	**0.940**	92%
gemma-2-9b	causation	0.203	**0.994**	95%
gemma-2-9b	certainty	0.462	**0.957**	95%
gemma-2-9b	credibility	0.752	**1.039**	95%
gemma-2-9b	negation	0.198	**0.960**	95%
gemma-2-9b	sentiment	0.534	**0.978**	95%
gemma-2-9b	temporal_order	0.404	**0.947**	95%
gemma-2-9b	moral_valence	0.547	**0.977**	95%

Table S3: Gemma-2 concept score recovery at identified CAZ peak vs. final global attention layer. Across 7 concepts, two model sizes, same layer, near-complete recovery. With only two models from one family, this is a case study, not a validated architectural principle; confirmation requires additional alternating-attention architectures. As a measured empirical matter, the final global attention layer — not the data-driven Fisher peak — is the maximum-recovery site for this cohort; why that is, and whether it generalises, is follow-up.

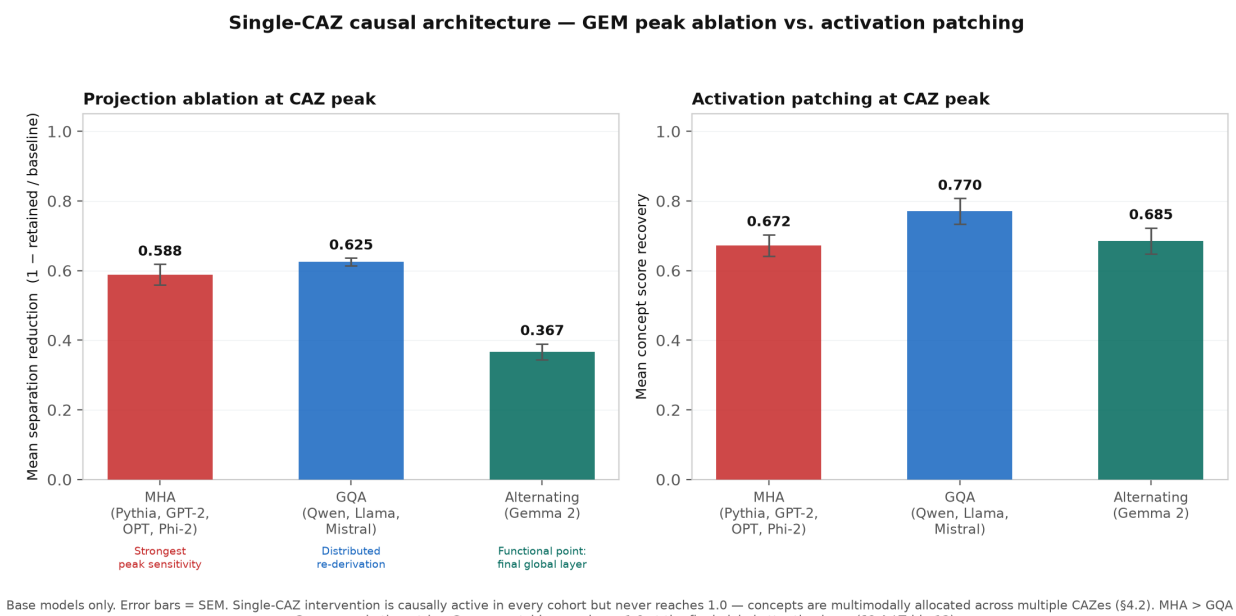

Figure S1: Mean projection ablation (left, GEM handoff protocol) and activation patching (right, single-layer at CAZ peak), grouped by model cohort, all 17 concepts. Error bars are SEM. MHA and GQA are comparable on ablation (0.588 vs 0.625); Gemma is lower (0.367). Patching means shown are raw (inflated by overshoot; trimmed means: MHA 0.533, GQA 0.693, Gemma 0.648; see Table 9). Base models only; GEM ablation: 306/136/34 measurements, patching: 306/136/34 (MHA/GQA/Gemma). Gemma's distinct localization is discussed in §8.3.

Patching-recovery depth. The layer of maximum patching recovery is deeper than the geometric peak: in MHA models mean peak patching recovery occurs at 83.1% depth while most concept peaks fall at 60–70%; in GQA models maximum recovery depth is 92.1%. This depth gap between the geometric peak and the maximum-recovery layer is a measured property, larger for the sparse-attention architectures; its interpretation is follow-up.

Peak-vs-causal divergence — full breakdown. A systematic test across 358 multimodal cases (N=250, C=17, all 28 base models; using self-retained separation from ablation as the causality measure) quantifies the split directly. In 50.3% of cases, the geometrically dominant CAZ is not the most causally active one; when the two diverge, the causal peak is deeper in 94.4% of cases. The mean depth gap in percentage points is *not* reported here: it is sensitive to how the gap is defined, and the recomputed definition yields ≈21.9 pp against the 41.3 pp of an earlier vintage (with a correspondingly unstable cohort breakdown), so only the direction and the 94.4% rate are treated as established (main text §8.7). The split rate varies substantially by concept: specificity splits in 82.6% of cases — its shallow geometric peak is rarely the causal bottleneck — while certainty (21.1%) and threat_severity (31.2%) are the most stable. Credibility splits in 57.1% of cases; the embedding-leakage sub-representations that create geometrically salient shallow peaks are not consistently causal. The 50.3% overall divergence rate indicates that geometric-peak $\neq$ causal-peak is a general property of

the detection method, not a credibility-specific pathology. A within-cell permutation null calibrates this rate: randomly reassigning which detected region carries the minimum self-retained separation (10,000 shuffles, preserving each cell's region count) produces a null divergence rate of 57.9% ± 2.5% — matching the $1-1/\bar{k}$ expectation for $\bar{k} \approx 2.38$ regions per cell (the multimodal-cell mean, higher than the corpus-wide 2.20 regions/pair because these cells are ≥2-region by construction). The observed 50.3% sits significantly *below* that null (z = −3.02, one-sided p = 0.0016): CAZ score carries real information about which region is most causally active — dominant and causal peaks coincide more often than random assignment would produce — while still failing to identify the causal peak in roughly half of multimodal cases. The divergence is therefore neither an artifact of region counts nor measurement noise; it is a systematic property (the causal peak is deeper in 94.4% of divergent cases) that score-based selection does not capture. Dominant CAZ is defined as argmax CAZ score; causal peak as argmin self-retained separation. Fisher detection reliably identifies layers where concept geometry is organized; it should not be used alone to select intervention targets. Cohort labels (MHA vs GQA+SwiGLU) are used here only as a descriptive grouping of the score distribution; the main text retires the earlier MHA/GQA *behavioural* split at C=17 (§6.4/§8.3), where the two cohorts are comparable on ablation, so no behavioural difference should be read from the grouping.

Overshoot. Several small MHA models (opt-125m: recovery 1.150; pythia-70m: 1.024) show recovery exceeding 1.0 — patched negative examples surpass the positive class centroid. Recovery above 1.0 indicates the patching intervention overrides rather than restores natural computation. This limits the "restoration" interpretation of patching results generally — the metric measures how effectively an artificial signal biases activation geometry in the concept direction, not how faithfully the network's own computation is reproduced. These models appear to have tight coupling between the CAZ peak and the output, with no subsequent dampening layers to attenuate the injected signal. Overshoot rates and trimmed means are reported directly in Table 9.

Held-out endogeneity check. The concept score recovery metric is endogenous to the calibration set: both the patching direction Δ_L and the reference scores $\bar{s}_{\text{pos}}, \bar{s}_{\text{neg}}$ are estimated from the same N=250 pairs whose activations are then patched. To bound this inflation, we evaluated patching recovery on N=100 held-out pairs (RCP indices 251–350, not used to estimate any quantity in Table 9), with directions calibrated on a 50-pair calibration subset. Mean calibration-set recovery: MHA 1.311, GQA 2.085, Gemma 0.954. Mean held-out recovery (100 unseen pairs): MHA 1.086, GQA 1.854, Gemma 0.734. The reduction is approximately 22 pp across all cohorts (median 15 pp across the 476 individual model-concept pairs), uniform across all three architecture families. All 7 model-concept pairs with negative held-out recovery (1.6%) are GPT-2 / GPT-2-medium, likely reflecting direction-estimation instability at the 50-pair calibration size; 98.4% of held-out pairs show recovery exceeding 0.1, confirming that identified CAZ peaks are injection-responsive on unseen data. The 22 pp figure is an upper bound on the inflation in Table 9's estimates (more calibration pairs yield less overfit directions); the qualitative conclusion — identified CAZ peaks are injection-responsive — is confirmed at held-out recovery levels that remain substantial in all three cohorts.

Optimal-ablation sharpness by architecture era (supports main-text §5.1). The architectural generation effect: modern architectures show a different pattern.

Architecture Era	Families	Sep. Reduction CV	Optimum Shape
Legacy (2019–2022)	GPT-2, Pythia, OPT	0.02–0.26	Broad
Modern (2024–2025)	Qwen 2.5, Gemma 2	4.78–6.40	Narrow

Table S4: Sharpness of the optimal-ablation-layer optimum by architecture era. CV = coefficient of variation of separation reduction across layers within a model; a low CV means many neighboring layers are near-optimal (broad optimum), a high CV means only a narrow band is. This sharpness pattern is decoupled from ablation magnitude: the earlier "redundant vs. sparse encoding" reading of this table — attributing it to a difference in how strongly MHA vs. GQA cohorts ablate — does not survive at C=17, where the two cohorts are comparable on mean ablation magnitude (§6.4, §8.3). What Table S4 captures is a real difference in optimum sharpness, not a difference in how much of the concept each cohort encodes.

In Qwen and Gemma, concept encoding is layer-specific — ablation works strongly at some layers and barely at others. The cohort effect is on sharpness of the optimum, not on its position: all three cohorts locate their median optimum near mid-depth (MHA 60%, GQA 56%, Gemma 59%), but the MHA optimum is broad (many neighboring layers are near-optimal) while the GQA/Gemma optimum is narrow (mis-targeted ablation fails). Within the 2023–2025 GQA cohort, the sharpest suppression layers lie near the end of the assembly chain (typically the final handoff into the unembedding projection) rather than at the Fisher peak itself — which is one reason peak-ablation is weaker than peak-patching in that cohort (§6.4).

---

§G. Cross-Validation Against Gemma Scope SAEs — Full Analysis

Supporting evidence for main-text §8.6. §8.6 reports that an independent sparse-autoencoder decomposition (Gemma Scope, gemma-2-2b) converges on CAZ's structural signal on two counts — direction agreement (best positive cosines 0.51–0.84) and shared-peak layers (L11/15/17/19) — with the per-layer curve correlation r = 0.979 reported only as a consistency check. The full per-feature analysis, polysemantic-axis structure, shared-peak semantics, and curve-equivalence derivation are here, with the honest limits (no shared-peak permutation null; direction cosines measured against underdetermined gemma-2-2b point estimates, §6.9). No value differs from §8.6.

Scope note. This section uses all 17 concepts; the main study in §§3–6 likewise uses all 17 concepts.

We ran a targeted cross-validation of Gemma-2-2b CAZ structure against the Gemma Scope sparse autoencoders (gemma-scope-2b-pt-res, 16k-width residual stream SAEs at all 26 layers; Lieberum et al., 2024). For each of our 17 concepts, we measured

two quantities per layer: (1) SAE discrimination — the mean absolute differential activation of the top-20 SAE features between concept-positive and concept-negative pairs; and (2) direction agreement — cosine similarity between the top differential SAE feature's decoder direction and the concept-specific eigenvector derived from paired activation differences.

Layer agreement. SAE discrimination increases monotonically across all 26 layers, peaking at L24 for causation and certainty, and at the final layer (L25) for the remaining 15 concepts. CAZ peak layers sit consistently in the top 25–40% of layers by SAE discrimination — above the median but not at the peak. This is the expected pattern, not a failure of agreement. The SAE measures a cumulative quantity: how much has the representation separated by layer L? The residual stream integrates information across layers, so discrimination will grow until the final output. The CAZ framework measures a derivative quantity: where does separation increase most sharply? CAZ peaks are assembly events; the SAE peak is where assembly is complete. The two methods are asking different questions about the same process.

Direction agreement. At CAZ peak layers, the top differential SAE feature decoder directions align with our 17 concept-specific eigenvectors at best positive cosine similarities of 0.51–0.84. Several polysemantic SAE features function as shared axes at distinct depths: feature 12945 at L11 positively aligns with causation (+0.59), credibility (+0.62), and specificity (+0.64); feature 843 at L13 aligns with certainty (+0.62), plurality (+0.66), and sentiment (+0.67); feature 14558 at L15 aligns with exfiltration (+0.73), authorization (+0.53), and temporal order (+0.52); and feature 8529 at L19 aligns with negation (+0.84), formality (+0.69), and urgency (+0.60) on its positive pole while certainty (−0.80), moral valence (−0.77), deception (−0.73), sentiment (−0.71), and sarcasm (−0.69) occupy the negative pole. Feature 1059 at L5 anti-aligns tightly with the security-adjacent cluster — authorization, exfiltration, threat severity, and urgency all at cosines near −0.89. The direction alignments are moderate rather than near-perfect, but this is the expected result of two methodologically distinct decompositions of the same space: the SAE fragments concept structure into sparse, individually decodable features (mean polysemantic alignment ~0.20–0.22 across all features at CAZ-predicted layers), while our eigenvector captures the same structure as a single dense direction. The moderate per-feature alignment is not evidence that either method is wrong — it reflects that the SAE is distributing across many features what CAZ captures holistically. In 2304-dimensional space, any alignment consistently above zero is non-random; the question is what unit of analysis you choose. A second factor caps the achievable cosine specifically for this model: the concept-specific eigenvectors compared against here are gemma-2-2b difference-of-means directions, which §6.9 shows are underdetermined point estimates at N=250 (split-half reproducibility 0.62–0.90, not the ≥0.96 of other models). The SAE alignment is therefore measured against one noisy draw of each concept direction; the 0.51–0.84 range is a lower bound distorted downward by that instability, and this section's cosines should be read with the §6.9 caveat rather than as tight estimates.

Feature 8529 at L19 is the broadest polysemantic axis in the 26-layer profile. Of the concepts for which L19 is a CAZ peak or shared peak, negation (+0.84), formality (+0.69), and urgency (+0.60) sit on the positive pole and certainty (−0.80), moral valence (−0.77), deception (−0.73), sentiment (−0.71), and sarcasm (−0.69) sit on

the negative pole. The five anti-aligned concepts each achieve their best positive alignment via concept-specific features at other layers — but at lower magnitudes (0.57–0.67) than negation's positive alignment (0.84). This asymmetry suggests that feature 8529's axis primarily encodes negation and its close semantic neighbours, with the anti-aligned concepts captured there as a contrast class rather than directly.

This reading is coherent but post-hoc: we observe anti-alignment for a cluster of semantically connected concepts, identify one shared feature, and conclude they form a contrast class on a common axis. Two specific follow-ups would sharpen this: (i) fit a **concept-conditioned SAE** trained on residual-stream activations from each concept's contrastive corpus separately, and check whether each concept's top feature decoder direction aligns positively with its eigenvector at the CAZ peak; (ii) conduct a **signed-projection analysis** (|cos|, sign reported separately) across all 16k SAE features and all 17 concepts to characterise the full alignment-magnitude distribution rather than relying on the single top feature per layer. Until those are run, the polarity-inversion reading should be treated as a hypothesis compatible with the data, not an established interpretation.

Shared peak layers. Across 17 concepts, four layers show maximum concurrent CAZ peak overlap — layers 11, 15, 17, and 19 each host 6 concepts simultaneously, each with 25 polysemantic SAE features. The four shared-peak layers (layers where multiple CAZ peak eigenvectors co-occur across concepts simultaneously) capture distinct semantic clusters: L11 groups structurally relational concepts (agency, causation, certainty, exfiltration, plurality, temporal order); L15 groups pragmatic and epistemic-ordering concepts (agency, causation, formality, sarcasm, specificity, temporal order); L17 groups evaluative and social-role concepts (authorization, credibility, deception, formality, sentiment, urgency); and L19 groups epistemic and assertion-modifying concepts (certainty, credibility, formality, moral valence, negation, sentiment). Additional smaller concentrations appear at L5 (3 concepts, 18 features), L16 (3 concepts, 10 features), and L22 (3 concepts, 9 features). Layer 13 provides a secondary 2-concept concentration (agency, authorization). This is convergent with CAZ's concurrent peak structure: the SAE, which has no knowledge of our eigenvectors or our peak detection method, identifies the same layers as structurally distinctive. The shared-peak geometry is visible to both methods. This finding has no permutation null: we have not quantified how often 4 of 26 layers would show comparable cross-concept peak overlap under a null model of randomly placed CAZ peaks, unlike the CAZ-detection permutation null in §4.1. That null is out of scope for this paper — reserved for the planned SAE cross-validation follow-up — so the shared-peak result should be read as a suggestive post-hoc pattern-match rather than a statistically established convergence.

Curve equivalence. Both methods measure a cumulative, unnormalized quantity. The SAE discrimination score grows monotonically as the residual stream accumulates concept-relevant information toward the final layer. The CAZ eigenvector separation score (mean paired activation difference projected onto the top concept eigenvector) follows the same pattern — neither corrects for within-class variance growth near the unembedding layer. Residual stream activations grow in norm as they approach the LM head; both concept centroids and within-class scatter expand proportionally, so raw distances accelerate while Fisher-normalized separation is flat or

declining in the same region (only 8–15% of Fisher S(l) curves show positive late-layer velocity; Fisher-normalized separation is flat or declining in the final 20% of depth in 85–92% of concept curves across all paradigms). Given this shared measurement basis, high curve correlation is the expected result. We directly compared the per-layer discrimination curves by computing the Spearman rank correlation between the two scores at each of the 26 layers: mean $r = 0.979$ across all 17 concepts (range: 0.949–0.993), confirming the methods track the same underlying cumulative signal. This correlation reflects agreement on unnormalized accumulation, not on Fisher-corrected concept strength; readers using the SAE discrimination curve as a proxy for concept encoding quality in late layers should apply Fisher normalization first.

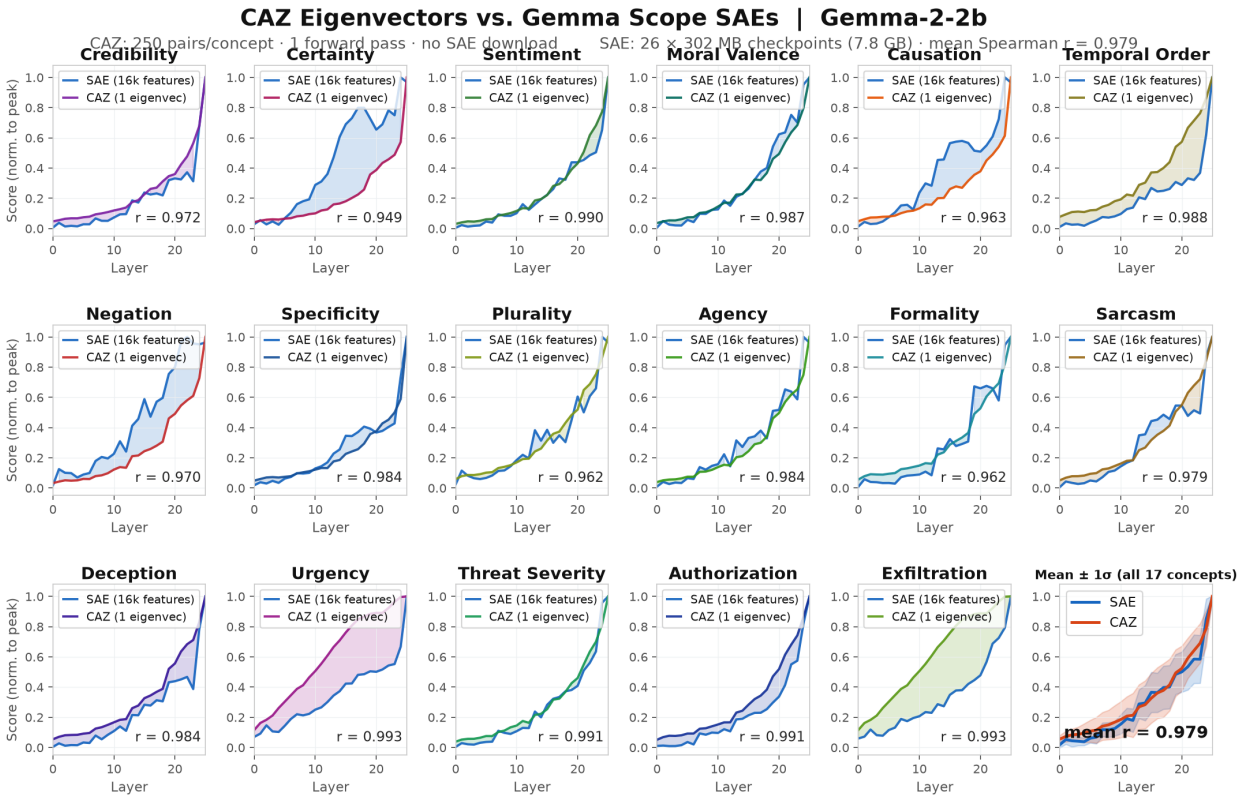

Figure S2: Per-concept comparison of CAZ eigenvector separation curves against Gemma Scope SAE discrimination curves, both normalised to each curve's own peak, across Gemma-2-2b's 26 layers — one panel per concept (17 total) plus a mean $\pm$ 1σ summary panel (bottom right). Per-concept Spearman rank correlations between the two curves range from 0.949 (certainty) to 0.993 (urgency, exfiltration), mean $r = 0.979$ across all 17 concepts — a consistency check confirming both methods track the same underlying cumulative signal (shared monotonic basis; high correlation is expected, not independent convergence evidence — see §8.6). The polysemantic-feature-count and direction-cosine-heatmap findings discussed above (shared peak layers L11/L15/L17/L19; best positive cosines 0.51–0.84) are not separately depicted in this figure — they are reported in the preceding prose ("Direction agreement," "Shared peak layers").

The CAZ assembly events (L11, L15, L17, L19) are not visible as peaks in either cumulative curve — they are visible in the *derivative*: the layers where the cumulative curve's slope is steepest. This is exactly what the CAZ deep-dive algorithm measures. The cumulative curve tells you where the concept is represented most fully; the derivative tells you where it was built.

Scope. This comparison covers 17 concepts against a 16,000-feature SAE. The SAE represents the full activation space; our eigenvectors are targeted probes for specific semantic axes. The question is not whether CAZ can replace SAE — it cannot, and is not designed to — but whether the two methods converge on the same structural signal for the concepts CAZ does cover. The positive evidence is (i) **direction alignment** at CAZ peak layers (best positive cosines 0.51–0.84 across all 17 concepts; polysemantic axes at L5, L11, L13, L15, and L19 account for the primary alignment signals), and (ii) the **shared-peak finding** at L11, L15, L17, and L19 — the SAE has no knowledge of our eigenvectors or peak-detection method, yet identifies the same four layers as maximally polysemantic (25 shared features each, with 6 of 17 concepts peaking simultaneously at each). These two pieces of evidence are what a third method converging on CAZ's structural findings would look like.

The curve correlation ($r = 0.979$) is a consistency check, not independent convergence evidence: both methods measure cumulative unnormalized separation and both grow monotonically as the residual stream accumulates concept-relevant information toward the final layer, so high correlation is the expected result given shared measurement basis — we report it to confirm that the two pipelines track the same underlying quantity, not to count it toward method convergence.

---

§H. Gentle-CAZ Ablation and Specificity Controls — Full Detail

Supporting evidence for main-text §6.1. §6.1 reports that causal efficacy is broadly distributed across the CAZ score spectrum (gentle CAZes reach 63.2 pp mean suppression), that CAZ peaks produce 3.59× greater suppression than peak-distal layers (Table 8, Figure 4; a model-level cluster bootstrap of the pooled statistic recenters at 3.60× [3.28, 3.97], and a separate mean-of-ratios estimand reads 4.29× — see §H), and that ablation is direction-specific (median 606.6× over random directions). The full score-category table, the specificity controls, the cluster-bootstrap/covariate robustness suite, and the direction-specificity breakdown are here. No value differs from §6.1.

Causal efficacy by score category.

Category	Score	Count	Self-retained	Suppression (pp)	Suppr > 20pp
Major CAZ	> 0.5	78	29.3%	70.7	97.4%
Strong	0.2–0.5	123	32.4%	67.6	95.1%
Moderate	0.05–0.2	139	39.3%	60.7	95.0%

Category	Score	Count	Self-retained	Suppression (pp)	Suppr > 20pp
Gentle	< 0.05	55	36.8%	63.2	98.2%

Table S5: Causal efficacy by CAZ score category (N=250). 395 per-CAZ ablations across all 28 base models × 7 multimodally-structured concepts (those with ≥2 detected CAZ regions; the per-CAZ multimodal protocol ablates every detected region, a denser population than the scored-detection floor of Table 2, so counts are not directly comparable to Table 2's totals). Each region is ablated individually and self-retained separation is measured; suppression (pp) = 100 − mean self-retained. "Suppr > 20pp" = fraction of CAZes whose ablation suppresses concept separation by more than 20 percentage points — the robust operating point of §8.7. The pattern is the load-bearing result: causal efficacy is broadly distributed across the score spectrum — gentle CAZes (63.2 pp mean suppression) remain solidly causally active despite their low geometric prominence, and all four categories exceed the 20pp threshold in >95% of cases. Coverage caveat: the 7 concepts are the original set (credibility, negation, causation, temporal_order, sentiment, certainty, moral_valence, §2.2); none of the 10 concepts added at C=17 — including all 5 behavioral/safety concepts (threat_severity, authorization, urgency, deception, exfiltration) — are represented in this table. Whether the score-vs-causal-efficacy pattern holds for that subset, several of which carry the domain-specificity caveat noted in §2.2, is untested.

Note: 395 regions total, computed from the per-model `ablation_multimodal_*.json` artifacts for these 7 concepts across all 28 canonical models (Table 1 roster; 150 model-concept files; every model contributes at least one qualifying region). Restricting to 26 of the 28 base models gives 361 regions.

Standard prominence thresholds miss most of the causally active features detectable by scored CAZ analysis. Of the 395 CAZes in Table S5, the 194 gentle-or-moderate features would be filtered out by a 10% prominence detector — leaving only the 201 strong or major CAZ peaks — yet gentle-or-moderate features remain causally active in >95% of cases at the 20pp suppression threshold (§8.7 reports sensitivity to threshold choice). Gentle CAZes produce less suppression per feature than major CAZes (63.2 pp vs. 70.7 pp mean) — a real gap, not parity — but the magnitude remains substantial: low-prominence features are not merely numerous, individually they still suppress the concept by a wide margin, just not as much as high-prominence ones.

Gentle CAZes retain 36.8% of separation on ablation versus 29.3% for major CAZes — a 7.5 pp gap, present but modest — so score, which by construction predicts geometric prominence, turns out to only weakly predict the *magnitude* of causal effect, and does not predict the binary question of whether a feature is causally active at all. 98% of gentle CAZes produce more concept suppression than the mean peak-distal layer (11.1% — the mean under the dominant-peak peak-distal classification used throughout this section; Table 8's *any*-scored-peak peak-distal mean is higher, 14.0%, and the two classifications are not interchangeable — see §8.7), compared to ~97% of major CAZes — a difference of ~1 pp, not the orders-of-magnitude difference in geometric prominence. A feature with a prominence of 0.5% of peak separation, occupying a few layers with minimal coherence boost, still carries enough of the concept to measurably degrade model performance when removed — just not as much as a major

CAZ does.

Detector margins. At N=250 the gentle CAZ category has a small margin near the mean peak-distal layer: 1 of 55 gentle features (1.8%) falls below the 11.1% peak-distal mean, and that same feature shows *increased* separation on ablation (retained 157.1%) — a single inverse case (certainty in opt-2.7b, at 3.1% depth) that is definitively not a concept-carrying feature and represents a false positive of the scored detector. The lone sub-baseline feature is at shallow depth (<20% of model depth), consistent with the embedding-to-transformer transition where anomalous velocity spikes can trigger false detections (cf. Figure 3, L1 arrow). The practical false-positive rate of the gentle detector is low (~2% marginal at N=250), but it is non-zero, and the margin concentrates in shallow layers.

Baseline: peak-distal layers. The peak-distal baseline provides the reference point missing from a threshold-only analysis. Using the global sweep data on Table 8's peak-distal classification (>3 layers from any detected region peak; n = 6,884), 28.3% of peak-distal layers exceed 20% suppression when the concept direction is ablated, compared to 98.2% of gentle CAZes at this threshold — a 3.47× enrichment on the same baseline as the §6.1 3.59× result. The enrichment ratio increases monotonically with the threshold (≈1.7× at 5%, 3.5× at 20%, ≈20× at 50%), as the peak-distal efficacy rate falls faster than the gentle-CAZ rate (gentle: ≈98% at 5%, 98.2% at 20%, ≈80% at 50%). (An alternative dominant-peak-only peak-distal classification, used elsewhere in this section, gives a smaller 20% peak-distal rate and hence a larger nominal ratio; the two classifications are not interchangeable — see the note above.) The 20% cutoff is not a favorable selection but an intermediate point in a continuous separation between CAZ and peak-distal layers.

This has direct implications for any method that uses activation magnitude, eigenvalue rank, or probe confidence to decide which features "matter." A threshold set to exclude noise will also exclude roughly half the detectable causal structure — features that are individually ~75% as potent as the most prominent features and that collectively match them in aggregate suppression.

Control: suppression is concept-specific. To establish that the suppression observed in Table S5 reflects genuine concept-feature alignment rather than a generic effect of ablating any deep feature, we ablated the 260 top-eigenvalue dark-matter features (10 per model) across 26 of the 28 base models at N=250 — features the network dedicates significant capacity to, but which are not necessarily concept-aligned — and measured retained separation for each target concept. Dark-matter features retained 94.5% of concept separation on average, compared to 83.3% for concept-aligned features (cosine similarity ≥ 0.15 to the concept direction, from the N=250 CAZ ablation data) — a gap of ~11.2 percentage points. The effect is modest compared to CAZ ablation because top-eigenvalue features tend to be long-lifespan features whose downstream layers partially recover from the ablation. The key point is specificity: concept suppression is concentrated in concept-aligned features, not distributed across arbitrary high-capacity features.

A natural layer-position contrast comes from §6.4: GEM handoff ablation produces comparable mean separation reduction in the MHA (0.588) and GQA (0.625) cohorts at the identified handoff layer — both cohorts show a strong, non-trivial effect, so the

handoff-layer intervention is causally active everywhere regardless of attention architecture. The ~11.2 pp dark-matter vs concept-aligned gap above and the concept-specificity of Table S5 (all 28 base models) triangulate the same conclusion as the C=17 global sweep: suppression tracks functional allocation, not layer position.

Cluster-robust and covariate-adjusted enrichment. CAZ peaks produce 3.59× greater concept suppression than peak-distal layers, as a ratio of pooled means (Mann-Whitney U = 2,772,013, p = 1.58×10^{-141}; 476 CAZ-peak vs 6,884 peak-distal layer measurements across 28 base models × 17 concepts). The effect holds across all architecture families and model scales, and is positive in every one of the 28 models; per-model ratios range from 2.17× (pythia-12b) to 12.56× (pythia-160m — a 12-layer model whose sparse peak-distal set, only a handful of layers >3 from any detected peak, inflates its per-model ratio), with the mid-scale models showing intermediate ratios (Llama-3.2-1B: 5.51×, Qwen 2.5-0.5B: 5.57×). The mean of the 28 per-model ratios is 4.32× — higher than the pooled ratio because the small-model tail (pythia-160m, gpt2 at 7.83×) pulls the per-model mean up; the model-level cluster bootstrap below is the outlier-robust estimate. This tail is a property of the *comparator*, not the effect: because the scored detector tiles the full depth, the peak-distal set (>3 layers from any peak) covers ≈78% of a ≤12-layer model's layers but only ≈33% at 37+ layers (mean CAZ coverage 77.9% at ≤12 layers, 56.2% at 13–24, 46.3% at 25–36, 32.8% at 37+), so shallow models estimate the peak-distal baseline from a small, peak-distant sliver (low mean → high ratio) while deep models estimate it from a large, representative set. The per-model dispersion and the 4.32× mean-of-ratios are therefore depth-confounded; the pooled 3.59× is layer-weighted and dominated by the large peak-distal sets of deep models, and — the load-bearing point — the enrichment is positive at every depth, including where the comparator is cleanest (pythia-12b 2.17×), so it is not an artifact of the comparator's depth-dependence. (A depth-invariant proportional-depth margin would make the comparator comparable across scales; it is not adopted here because it would move the frozen headline value, and the depth-decile regression below already shows the effect survives depth adjustment.) Concept suppression is not a consequence of ablating at any network layer — it is concentrated at the specific layers where that concept allocates. **A note on this p-value's magnitude**: the 7,360 measurements it is computed over are not independent — adjacent layers within a model are highly autocorrelated, and 22 of the 28 models are scale ladders within 4 architecture families, sharing training corpus and initialization scheme. p = 1.58×10^{-141} should not be read as a literal significance level; it reflects large-N pooling over correlated observations, not 135 orders of magnitude of independent evidence. The per-model ratio range (2.17×–12.56×, positive in every one of 28 models) and the family-stratified consistency are the more meaningful evidence for the effect than the p-value's magnitude — the same reading already applied to the ordering tendency's Wilcoxon statistic above (§3.1). **Cluster-level confirmation.** To replace the caveat with a direct correction rather than leaving it as a caveat alone, we ran a model-level cluster bootstrap (resample all 28 models with replacement, 2,000 resamples) and a within-model permutation test (shuffle peak/peak-distal labels within each model, 2,000 permutations) on the global-sweep ablation data underlying this result. This uses an independent, simplified peak/peak-distal classification (single dominant Fisher peak vs. layers >3 away, 10,471 measurements —

not the exact Table 8 methodology, so the point estimate differs) and gives a cluster-adjusted enrichment ratio of 4.29× with 95% CI [3.83×, 4.84×] under model-level resampling, and a within-model permutation p = 0.0/2,000 (no permutation of the 2,000 run matched or exceeded the observed ratio). This 4.29× is a *mean-of-ratios* on the simplified classification, so it sits above the pooled 3.59× for the same reason the per-model mean does — the shallow-model tail. Bootstrapping the frozen Table 8 statistic *directly* — the pooled ratio-of-means, resampled over the 28 models, exact Table 8 recipe (`scripts/f5_enrichment_clusterboot_ratioofmeans.py`) — is instead a no-op: **3.60× [3.28, 3.97]**, centred on the 3.59× point estimate. Clustering neither inflates nor deflates the enrichment; the 4.29× and 3.60× differ only in estimand. The cluster-adjusted CI excludes 1.0× (no enrichment) by a wide margin, confirming the effect survives proper accounting for within-model and within-family non-independence — this is not merely a caveated large-N p-value, it is an independently re-derived, cluster-robust confirmation of the same direction and approximate magnitude. **Family-level and covariate-adjusted confirmation.** Three further controls sharpen this. Resampling at the architecture-family level (8 families — the most conservative clustering unit, since 22 of 28 models are scale ladders within 4 families) gives a bootstrap enrichment mean of 4.27× with 95% CI [3.58, 4.94], and every family's individual ratio exceeds 2.7× (range 2.77×–5.48×) — the enrichment is not a consensus of the scale-ladder families. Because deeper interventions sit closer to the final-layer readout, we also regressed separation reduction on intervention depth across all 13,022 global-sweep measurements: CAZ-peak membership retains a large independent effect after controlling for depth (coefficient +0.301, p ≈ 0), and the CAZ-vs-peak-distal difference is positive in every depth decile, growing from +0.06 in the shallowest bin to ∼+0.54 in the deepest — depth proximity to the readout does not explain the enrichment. Finally, because the ablated direction $u^{(\ell)}$ is estimated with layer-dependent quality (high-separation layers yield cleaner difference-of-means estimates), we regressed reduction on log local Fisher separation: CAZ membership again retains an independent effect (+0.303, p ≈ 0, n = 13,005) — the enrichment is not reducible to direction-estimation SNR. (Outputs: `paper_n250/_round3_statistical_objections/` on the public dataset.) *Provenance: Table 8 and every robustness estimate in this paragraph were regenerated 2026-07-22/23 from the corrected `paper_n250` global-sweep artifacts after the exfiltration label fix (`scripts/table11_reconstruct.py`, `scripts/p3_enrichment_robustness.py`; the Table 8 recipe — peak = per-file `caz_peak`, peak-distal = layers >3 from any detected region peak — reproduces the peak-distal baseline to the digit). The re-run shifted every estimate by ≤0.15 with no conclusion changed: cluster bootstrap 4.31×→4.29×, family bootstrap 4.13×→4.27×, depth coefficient +0.305→+0.301, SNR coefficient +0.313→+0.303; all CIs still exclude 1.0× and the within-model permutation p stays 0.0/2,000.*

Direction specificity control. The layer-specificity result establishes *where* to ablate; a complementary control establishes *what* to ablate. We ablated 10 random unit vectors at the same CAZ peak layer for each (model, concept) pair and compared the resulting separation reduction to the concept-direction ablation. Across 465 of the 476 (model, concept) pairs spanning all 28 base models and all 17 concepts — the remaining 11 are excluded because concept-direction peak ablation measured a *negative* separation change there (retained separation ranging 105–489% of baseline, the

same overshoot phenomenon noted above for certainty/opt-2.7b (detector margins, this section); a specificity ratio against random directions is undefined when the numerator is a null effect) — the concept direction produces suppression far outside the random-direction distribution: median per-pair z = 388.7, with a median specificity ratio of **606.6×** over random directions (mean concept reduction 0.551, mean random reduction 0.002). The z-score is the more stable statistic — the ratio's denominator is a near-zero random-ablation mean, so its magnitude should be read qualitatively. The ratio of means (0.551/0.002 ≈ 276×) is lower than the median per-pair ratio because the distribution is right-skewed: pairs with near-zero random reduction produce very large individual ratios that compress the median relative to what naive division of means would suggest. The concept direction exceeded all 10 random seeds in 99.8% of pairs. The result is consistent across architecture cohorts: MHA 456.8×, GQA 1011.7×, Gemma 522.8× (Figure S3). The GQA ratio being highest is directionally consistent with a sparser-encoding reading — if GQA+SwiGLU models concentrate concept signal into fewer, more tightly organised directions, random directions at the same layer would fall farther from the concept axis than in more redundant MHA encodings, producing a larger gap between concept and random ablation. The main text, however, retires the "redundant vs. sparse encoding" dichotomy as unsupported by ablation efficacy at C=17 (§6.4/§8.3), so this specificity-ratio pattern is offered as a descriptive observation, not as evidence for that hypothesis. Separation suppression at CAZ peaks is not a consequence of removing any direction from the residual stream — it is specific to the concept's geometric axis.

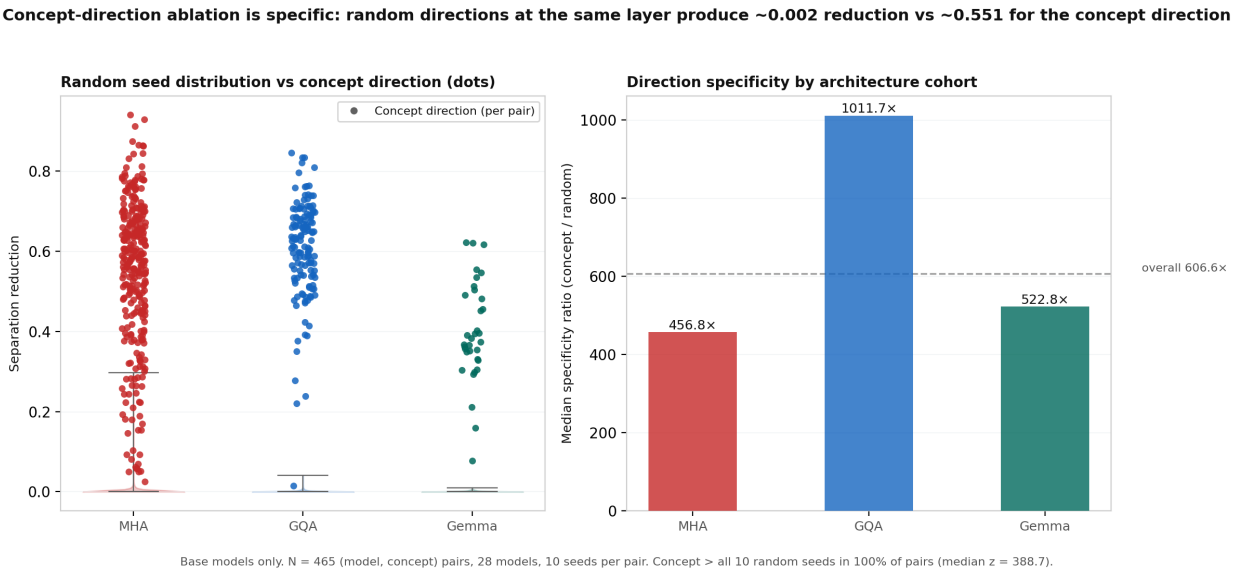

Figure S3: Violin plots show the distribution of separation reduction across 10 random unit vectors ablated at the CAZ peak layer, per architecture cohort (base models only). Dots show concept-direction reduction for each (model, concept) pair. Across all cohorts, concept-direction ablation is an outlier relative to the random distribution. Right panel: median specificity ratio by cohort.

§I. Limitations — Extended Derivations

The three longest limitation analyses from main-text §8.7 are given in full here; each is named and summarized in §8.7. No value differs.

Generator robustness of the ordering result. The RCP corpus mitigates single-generator idiosyncrasy by construction (§2.2), but a residual concern is that the 250-pair draw's ordering could still be dominated by one or two of the 14 generators. We test this directly with a leave-one-generator-out analysis: for each of the 28 base models, we recompute each concept's peak depth 13–14 times (once per generator with adequate representation in that model's 250-pair draw), each time excluding that generator's pairs entirely, and compare the resulting 17-concept depth ordering to the full-corpus ordering via Kendall's τ. Across all 28 models, mean per-model τ between full and leave-one-out orderings is 0.931 (median 0.946, range 0.833–0.993) — dropping any single generator leaves the ordering essentially intact. The worst individual case (one generator, one model: gemma-2-2b) drops to τ = 0.434, and five models (gpt2, gpt2-large, gpt2-xl, pythia-1b, gemma-2-2b) show mean τ below 0.87, consistently lower-power small/shallow or alternating-attention models — consistent with the reliable-disagreement account already given for low aggregate τ above ("Low-τ models"), rather than a new failure mode. This directly rules out the narrower version of the LLM-generated-corpus concern (§8.7 "Dataset provenance"): no single generator is load-bearing for the ordering result. It does not address the broader concern that all 14 generators are themselves transformer LMs; a non-LLM contrastive baseline remains open follow-up work.

CAZ score formula sensitivity: The composite CAZ score (prominence_norm × coherence_boost × $\sqrt{\text{width}/L}$) combines three components that could be weighted differently. We compared four scoring formulas on the same 1,045 regions: (A) prominence alone, (B) prominence × $\sqrt{\text{width}/L}$, (C) prominence × coherence_boost, and (D) the current composite. All pairwise Kendall τ between formula rankings exceed 0.82 at region level and 0.77 at the dominant-CAZ-per-(model,concept) level, indicating that the rank ordering of regions is robust to formula choice. The architecture-ordering claim — MHA produces far more major CAZes than GQA — holds across all four formulas (MHA–GQA gap: 11.8–17.6 pp depending on formula). The absolute percentage in the "major" category (score > 0.5) is the most formula-sensitive statistic: it ranges from 11.5% under formula B (removing coherence dampens peak scores substantially) to 35.2% under formula C (coherence-only boost inflates scores), with the current formula giving 24.3% (Table 2). Category breakdowns in §8.3 (e.g., "30.9% major for MHA") should be read as formula-specific; the architecture ordering and depth-concentration claims are formula-robust.

Multiple comparisons: No formal multiple-comparison correction is applied across the 28 × 17 × L test grid. Most reported statistics are aggregated (cohort means, median τ, Wilcoxon test), which partially mitigates the concern — the 3.59× ablation enrichment and the $\tau = 0.404$ ordering tendency are each single tests over pooled data. However, concept-level breakdowns (e.g., the 50.3% divergence rate across 358 cases in §6.4, per-concept split rates) involve implicit multiple testing that is not corrected. The relay-feature permutation p = 0.001 (§7) is exposed to a multiple-comparisons concern because the candidate-discovery pipeline tries multiple screen-

ing rules before the permutation test. The pipeline (`gem/relay_funnel_null.py`) applies exactly four successive operationalizations of "relay feature" — Rule A (any concept alignment ≥ 0.5), Rule B (≥ 2 concepts above threshold globally), Rule C (≥ 2 concepts at any layer), Rule D (≥ 2 different-concept dominants at different layers — the reported criterion). The paper reports Rule D, the strictest. Bonferroni correction for 4 rules: adjusted $\alpha = 0.05/4 = 0.0125$. The reported $p = 0.001$ clears this adjusted threshold ($0.001 < 0.0125$), so the relay finding survives multiple-comparisons correction for the four rules that were tried. However, the pre-screen pipeline also made two thresholding decisions (Marchenko–Pastur eigenvalue cutoff; 5-layer minimum lifespan) that represent additional analytical degrees of freedom not counted in the four-rule correction. Treated conservatively, the relay p-value should be read as surviving correction for the explicitly enumerated screening rules but remaining subject to those pre-screen choices. This is an additional reason — beyond the dominant-variance and circularity concerns noted in §7 — that the relay feature count is provisional and should not be cited as a confirmed result.

---

§J. Multimodal and Depth-Dependent Concept Geometry — Tables and Detail

Supporting tables for main-text §4.2 (multimodal allocation) and §4.3 (depth-dependent concept-pair geometry). The headline rates, the directional patterns, and the honest nulls/caveats are stated in §4.2/§4.3; the per-concept tables, the saddle-vs-rotation boundary check, the full Procrustes-underdetermination caveat, and the semantic-relatedness alternative are here. No value differs.

Shallow–deep sub-representation cosines. The sub-representations at different depths are geometrically distinct:

Concept	N multimodal	Mean cos(shallow, deep)
credibility	4	0.405
certainty	8	0.292
causation	6	0.135
moral_valence	4	0.117
negation	6	0.235
temporal_order	6	0.176
sentiment	0	—

Table S6: Cosine similarity between shallow and deep peak directions within the same model, over models the threshold detector flags as multimodal for that concept (`is_multimodal`; signed cosine). Values of 0.12–0.41 are well above random baseline (|cos| ≈ 0.02 for random unit vectors in these dimensions) but well below identity — consistent with either geometrically distinct sub-representations or a single diffuse concept direction that rotates gradually across depth. The N counts fall sharply from the earlier N=200 corpus (credibility 17→4, sentiment 2→0): at N=250 the 10%-prominence threshold detector resolves a clean shallow+deep pair in fewer models,

even though the scored (0.5%-floor) detector finds credibility bimodal in nearly all models (§3.3). Read this table as a conservative, threshold-gated lower bound on multimodal sub-representation geometry.

Saddle-vs-rotation boundary check. We compared the saddle point boundary definition against an independent geometric signal: adjacent-layer directional stability $DS(\ell) = |\cos(d_\ell, d_{\ell+1})|$ (the absolute value of the adjacent-layer directional-stability quantity defined in [Henry, 2026a §4.2]), where d_ℓ is the dominant concept direction at layer ℓ. A sharp rotation event (local minimum of $DS < 0.7$) marks a layer where the model's top concept direction pivots near-orthogonally. Across all 28 models and 7 original concepts, 50% of separation saddle points fall within ± 2 layers of a direction rotation event — indistinguishable from a permuted baseline of 53% (ratio 1.07×; $N = 2000$ permutations shuffling rotation positions within each model).

The two boundary definitions are essentially uncorrelated. This null result is ambiguous: it could mean that saddle points and direction rotations track genuinely distinct phenomena (functional separability changes vs. geometric direction changes), or it could mean that saddle points are not tracking structurally meaningful boundaries at all. We report it without privileging either interpretation.

Depth-dependent concept-pair geometry. After Procrustes alignment into a shared coordinate space, concept pairs show consistent depth-dependent relationships across the 20 same-dimension model pairs tested:

Concept Pair	Category	Shallow alignment	Deep alignment	Direction	Models with pattern
causation × temporal_order	relational	moderate	higher	converge	20/20
credibility × certainty	epistemic	moderate	lower	diverge	20/20
sentiment × moral_valence	affective	moderate	lower	diverge	18/20

Table S7: Depth-dependent concept-pair geometry across 20 same-dimension model pairs. Cosine values are omitted because the Procrustes alignment is heavily underdetermined (see §4.3 caveat); only the direction of the depth-dependent change is reported, as a descriptive observation of this paper. Absolute cosine values remain subject to the underdetermination caveat below and should not be cited as measured quantities. This is a small, exploratory same-dimension pair set analysed with raw (non-LOCO) peak-position Procrustes as a qualitative directional check; a dedicated cross-architecture alignment analysis, with the null controls appropriate to a quantitative convergence claim, is separate companion work [Henry, 2026d] and is neither reproduced nor scored here.

A methodological caveat on Procrustes alignment. Aligning 17 concept directions in d-dimensional space ($d \gg 17$) leaves the rotation matrix heavily underdetermined — effectively 136 constraints for $d(d-1)/2$ degrees of freedom. The absolute

cosine values in Table S7 may be inflated by alignment overfitting, and the sign of small deltas is not guaranteed to survive that inflation either: a poorly-constrained rotation can in principle flip the sign of modest effects along with their magnitude. What is reported here is that the direction of the delta (converge vs. diverge) is *consistent* across all 20 model pairs — a descriptive pattern, not a validated directional claim. A full analysis of Procrustes alignment quality — including the permutation-based significance tests that would address both the magnitude-inflation and sign-robustness concerns for the alignment procedure itself, and that a quantitative cross-architecture convergence claim would require — belongs to separate companion work [Henry, 2026d], not to this validation study. This paper reports Table S7 only as a qualitative directional pattern and rests no quantitative claim on it; its absolute cosine values should not be cited as measured quantities, the underdetermination caveat applying regardless of the directional consistency.

A parsimonious alternative deserves equal weight with the allocation-dynamics reading. Causation and temporal ordering are semantically related in any distributional-semantics sense — cause precedes effect, and the training corpus reflects this — so any model that learns temporal structure will represent the two concepts more similarly at layers where temporal reasoning is complete, regardless of architecture-specific allocation mechanisms. Credibility and certainty share surface cues ("I'm certain" ≈ "credible") at shallow depths but functionally dissociate once the model can distinguish speaker confidence from claim support, which is also a depth-dependent property of representation learning on natural text. Under this reading, the converge/diverge pattern is downstream of shared training distributions, not of anything specific to the CAZ framework. All 28 models were trained on overlapping internet-text corpora; the 20/20 consistency establishes that the depth-dependent geometry is *robust*, but not that it is *caused by* concept allocation dynamics rather than by semantic relatedness at different levels of abstraction. Distinguishing these readings would require either (a) models trained on deliberately non-overlapping corpora to see if the pattern dissociates from training data, or (b) a mechanistic intervention (e.g., ablation of a shared shallow direction, §4.5) showing that the converging/diverging concepts causally interact in a way that pure distributional geometry would not predict. Neither has been run. This section reports a pattern; the mechanism behind it is unresolved.

---

§K. Corpus Construction and Pair Audits

Supporting detail for main-text §2.2. §2.2 states the corpus provenance (250 pairs/concept from the multi-generator RCP consensus pool), the domain-specificity caveat, and the three pair-audit results (BoW ceiling; lexical overlap 0.154; the C13 human-validation first pass — 66% clean, the antonym-vs-absence defect, moral_valence categorically weak). The multi-generator construction, the single-generator risks it mitigates, the cross-architecture survival test, the single-generator-corpus comparison, and the full audit prose are here. No value differs.

Dataset provenance. The main study uses 250 pairs per concept from the Rosetta Concept Pairs (RCP) consensus corpus (https://github.com/jamesrahenry/Rosetta_Co

ncept_Pairs; DOI: https://doi.org/10.5281/zenodo.20059650), which is publicly available under a permissive license. The RCP corpus was built specifically to address the single-generator concern: for each of the 17 concepts, 100 topics were independently prompted to 14 models spanning four AI labs (Anthropic, Google, OpenAI, Mistral), producing approximately 1,400 pairs per concept with no pair originating from a single generator. The 250 pairs used here are a stratified draw across generators. Three specific single-generator risks — **lexical confounding** (a single generator may overuse diagnostic vocabulary), **register gradients** (abstract concepts may be generated more formally than concrete ones), and **template convergence** (a single prompt converges on a narrow narrative range) — are mitigated by construction: each risk requires that one generator's tendencies dominate the corpus, which cross-generator averaging prevents.

Dataset quality: We validated pair quality via a cross-architecture survival test: each pair was scored against a 6-model gauntlet spanning both MHA (Pythia-1.4b, GPT-2-medium, OPT-1.3b, Phi-2) and GQA architectures (Qwen2.5-1.5B, Llama-3.2-3B). A pair "survives" a model if its peak Fisher-normalized separation across layers exceeds 0.3. All 250 pairs per concept survive all 6 models — 100% survival rate across all 17 concepts. Population-level separations range from 0.55 (temporal_order, Llama) to 1.56 (credibility, Phi-2), well above the threshold. The pairs are robustly concept-diagnostic across architecturally diverse models; the results in this paper are not artifacts of model-specific pair tuning.

Two observations from the main study further constrain any residual artifact concern at the ordering level. First, the concept ordering varies across models: median per-model $\tau = 0.404$, against a measured reliability ceiling of 0.911 (§3.2 — the ceiling is estimated from split-half agreement rather than assumed to be 1.0, since measurement noise attenuates τ regardless). If the ordering were determined solely by surface properties of the pairs, all models receiving the same pairs should produce the same ordering, and we would observe τ near that ceiling. We do not — the 17-concept ordering ranges from $\tau = 0.000$ (gpt2-medium) to $\tau = 0.796$ (opt-2.7b), and the split-half data show the low-τ models are reliably measured (§3.2), so their weak alignment is a genuine ordering difference rather than depth-resolution noise. Controlling for discriminative-token frequency directly leaves the agreement essentially intact (median partial τ = 0.380; §3.2). The ordering is an interaction between pairs and models, not a property of the pairs alone. Second, credibility — the concept most vulnerable to lexical confounding (§3.3) — is tied for the highest cross-model variance (std 32.7) outside the shallowest syntactic cluster, consistent with models resolving the same pairs via different mechanisms rather than all latching onto the same surface signal.

Consistency with original single-generator corpus (Supplementary §C). An earlier version of this study used 100 pairs per concept generated by a single model (Claude Sonnet 4.6). Supplementary §C reports a head-to-head comparison — rerunning extraction on four architecturally diverse models (Pythia-1.4B, GPT-2-XL, Qwen2.5-3B, Llama-3.2-3B) with both the original 100-pair single-generator set and the ~1,400-pair RCP multi-generator consensus pool (the §C run predates the N=250 stratified draw the main study uses). The category-level ordering structure holds under both conditions — deep concepts stay deep, and negation is the shallowest

concept in five of the six model×corpus conditions where its peak is unflagged (the exception: GPT-2-XL's single-generator run, where an unflagged shallow credibility artifact at 10.4% sits lower — §C). Per-concept depths shift substantially in places: 10 of 24 measurable model×concept cells sit within $|\Delta| \leq 5$ pp (median $|\Delta| \approx 9.4$ pp), with certainty the most stable concept (three of four models within 5 pp). Generator-dependent shifts on negation and credibility are diagnostic rather than refutational — they track the multimodal structure independently documented in §3.3 and §4.2. The ordering's category structure has survived both single-generator and multi-generator corpus construction; per-cell peak depths are corpus-sensitive.

Three pair-level audits accompany the corpus. (a) A bag-of-words probe baseline (TF-IDF 1–2gram + logistic, held-out-topic GroupKFold; `scripts/c12_bow_baseline.py`) classifies essentially every concept perfectly from raw text (grand mean held-out AUC 0.999): the pairs are lexically separable *by construction*, as contrastive minimal pairs must be. This ceiling means the BoW baseline cannot discriminate the depth-ordering hypothesis — there is no per-concept variance in lexical separability to correlate against depth — and the surface-vs-depth question is instead carried by the frequency partial-τ, reliability-ceiling, and topic-disjoint controls (§3.2); an off-ceiling redesign (deliberately crippled classifiers whose per-concept *difficulty* varies; `scripts/c12_bow_ordering.py`) has since confirmed that surface lexical difficulty does *not* reproduce the depth ordering (Kendall τ spans 0.18–0.38 across seven pre-specified difficulty measures against the corpus $\tau = 0.404$; six are non-significant at $n = 17$, and the one exception — a character 2–3-gram / (1 − AUC) measure at $\tau = 0.382$, $p = 0.034$ — does not survive correction for the seven measures tried), so the ordering is not a lexical-difficulty artifact. (b) Pair-level lexical overlap is modest (mean Jaccard 0.154, content-word Jaccard 0.112; `scripts/c13_lexical_overlap.py`) — positive and negative texts of a pair share topic vocabulary but are not near-duplicates. (c) **Human validation of pair fidelity — first pass complete.** A staged crowd study (31 raters, ≥2 per pair, 2-of-3 flags = defective; `scripts/c13_rating_app.html`) collected 540 analysis-grade ratings over 180 pairs (10 per concept, 20 for exfiltration): **119 pairs (66%) clean, 41 (23%) unresolved, 20 (11%) defective — so 89% (160/180) are not flagged defective, and 66% are affirmatively clean.** Fourteen of the 17 concepts validate at ≥76% valid (sarcasm and specificity 100%, most others 87–97%); three fall lower — causation and deception at ≈63%, and **moral_valence at 45% with no clean pair of ten.** The dominant defect is systematic and diagnostic rather than random noise: the negative passages are frequently *opposite-pole* (antonym) rather than *concept-absent* — negative-side failures outnumber positive-side 2.7 : 1, and moral_valence's negatives are morally *bad* rather than morally *neutral*, so an antonym negative still activates the concept direction. This is a generation-template issue (a secondary register-leakage variant — negatives separable by style or topic rather than by the concept — is flagged with it); moral_valence and the 20 defective pairs are slated for regeneration under a corrected presence-versus-absence template, a corpus-level fix shared with the companion papers [Henry, 2026a; Henry, 2026b; Henry, 2026d]. Concept-level rates rest on 10 pairs each (exfiltration 20) and are directional; on contested pairs a 2-of-3 verdict is near coin-flip stability, so the aggregate clean rate and the systematic antonym signature — not individual borderline pairs — are the load-bearing results. Full protocol, per-rater quality screening, and reliability analysis are in the C13 first-pass report accompanying the corpus.

§L. Reproducing Each Result — Full Script Index

Supporting detail for main-text §9 (Conclusion). The Conclusion states that every table and figure reproduces from the public corpus, model weights, and code; the repository versions, commit pins, and the full per-result script index are here.

Every table reproduces from the public concept-pair corpus (https://github.com/jamesrahenry/Rosetta_Concept_Pairs; Zenodo DOI 10.5281/zenodo.20059650), the public model weights, and two public code repositories, using the `paper_n250` activations as a starting point. The reusable **library** — the imported API `rosetta_tools.{caz, ablation, extraction, reporting, dataset}` — is `rosetta_tools` (https://github.com/jamesrahenry/Rosetta_Tools; Zenodo DOI 10.5281/zenodo.20361433). Extraction was run during 2026-06-08–14 from the repository state at or after commit `6fed9e2` (which reports library version 1.3.1). **Reproduce the pair draws from `6fed9e2` specifically**: that commit introduced the stable per-concept sampling seed the released corpus depends on; the identically-named `v1.3.1` git *tag* predates it (and reports version 0.1.0) and still uses process-randomized, `PYTHONHASHSEED`-salted seeding, so checking out the tag draws a different subset on each run. The **analysis, figure, and regeneration scripts** named below (the `extraction/`, `caz/`, `gem/`, and `scripts/` paths) live in `Rosetta_Analysis` v1.1.0 (https://github.com/jamesrahenry/Rosetta_Analysis; Zenodo DOI 10.5281/zenodo.21583139), which imports that library.

Generating scripts by result: CAZ extraction and per-concept depth/ordering (Tables 1, 3; Figs 1–2) — `extraction/extract.py` then `caz/analyze.py` / `caz/analyze_scored.py`; scored vs. threshold detection and score categories (Table 2) — `caz/analyze_scored.py` (0.5% prominence floor); multimodal allocation and cross-concept cosines (Table 5; §J Table S6) — `caz/analyze_structure.py` / `caz/probe_shallow_deep.py`; single-layer and global-sweep ablation, direction-specificity null (Table 8; §F Table S4) — `gem/ablate.py`, `gem/ablate_global_sweep.py`, `gem/ablate_random_direction.py`; the P1 optimal-ablation-layer split (§5.1) — `scripts/b4_optimal_layer_split.py`; P4 post-chain degradation (§5.4, Table 7) — `gem/test_p4_post_chain.py --part a` (Part (b), unembedding token clustering, is `--part b`, GPU-only); per-CAZ multimodal ablation (Table S5) — `gem/ablate_multimodal.py`; GEM handoff and cohort ablation (Table 9; §F Table S3) — `gem/ablate_gem.py`, aggregated by `gem/aggregate_gem_results.py` and re-verified by `scripts/table12_cohort_recompute.py`; exfiltration ablation-width robustness (§6.5, §8.7) — `scripts/exfil_width_robustness_check.py`; dependency structure (§6.2) — `gem/ablate_gem_multi_zone.py`, with the threshold-cutoff sensitivity sweep in the same section reproduced by `scripts/c6_sensitivity.py`; activation patching (Tables 9, 10; §F Table S3) — `gem/patch.py`; the Gemma-2 split-half instability diagnostics (§6.9) — `scripts/gemma_subtlety_test.py`, `scripts/gemma_rank_k_ablation.py`, `scripts/distilgpt2_distillation_test.py`, `scripts/gemma_whitening_control.py`, and `scripts/gemma_convergence_recompute.py`; CKA boundaries and coasting (§6.7) — `caz/cka_validation.py` then `caz/analyze_coasting.py`, with `scripts/coasting_fraction_recompute.py` for the reported coasting fraction; behavioral pilot (Table 11) — `gem/ablate_gem_behavioral.py`, aggregated by `gem/aggregate_behavioral_pilot.py`; permutation null (Table 4) — `caz/permutation_null.py`; spectral dark-matter census, persistent/relay features and the concept-specificity

ablation control (§7, §H Table S5 control; Table 12) — `caz/deep_dive.py` → `caz/detect_manifolds.py`, `gem/ablate_dark_matter.py`, `gem/relay_funnel_null.py`; and the SAE cross-validation (§8.6) — `caz/gemma_scope_xval.py` against the public Gemma Scope residual-stream SAEs (`google/gemma-scope-2b-pt-res`, via `sae_lens`).