

Where Personas Write a Forced Choice: The Identity Span Mediates, a Preference Circuit Does Not

Aditya Sarna

Apart Research Digital Minds Research Sprint (Tracks 1, 4, and 5)

Abstract

Instruct models change their answers when the system prompt assigns a persona. That fact is not evidence of a preference circuit, and it is not evidence of a self. It is evidence that some residual-stream site writes the system-prompt difference into the next-token distribution. We ask which site that is.

We length-match seven system prompts so that only an aligned identity span differs, then force a single-token A/B choice. Activation patches are applied only on cells where a donor persona and a recipient persona already disagree. On Llama-3.1-8B-Instruct (4-bit), copying the donor’s early identity-span residual stream recovers the donor choice (layer 0: mean recovery $R = 0.99$, 95% bootstrap CI $[0.92, 1.05]$, flip rate 0.92, incoherence 0, $n = 72$). Copying the last user-content token recovers nothing at any depth. The same early-span / late-token dissociation appears on Llama-3.2-3B-Instruct and on Qwen2.5-3B-Instruct, both in native bf16.

Four length-matched controls—a formality rewrite, a synonym of the default identity, a word-order scramble of the default identity, and a nonword identity with no lexical overlap—all recover at the same site. The nonword identity is the control that actually moves behavior (52/64 agreement with default on Llama-8B; 48/64 on Qwen) and still mediates ($R = 0.88$ and 0.97). Linear preference probes miss a pre-registered AUC of 0.75. A format-matched base-model provenance test remains a wash after an on-model pilot to 38 items. Stated and revealed process choices agree for a default assistant ($\kappa = 0.71$) and diverge under an anti-assistant persona ($\kappa = 0.25$).

The supported claim is narrow and causal: **the identity-span residual writes the forced choice, including when that span is not a word, on two families.** It is not a persona-specific preference circuit, not a self, and not a decision at layer L .

Keywords: activation patching, system prompts, personas, residual stream, preference probes, digital minds

1 Introduction

A system prompt that says “you are a poet” can change what an instruct model answers. The change is easy to over-read. One reading is that the model has a persona-specific preference circuit: a mid-layer feature that stores a taste and later drives the choice. Another reading, common in digital-minds discussions, is that some residual-stream site is a self, or that the instruct model’s answer is the surface of a substrate already present in the base model. Neither reading is identified by a behavioral persona effect alone.

What a persona effect *does* identify is a difference in the next-token distribution. Residual-stream activation patching can then ask where that difference is written Vig et al. (2020); Heimersheim and Nanda (2024); Zhang and Nanda (2024). The method is only well-defined if the two prompts are length-matched, if the

patched site is a specified token span, and if donor and recipient already disagree. Agreeing pairs cannot inform flip-to-donor. Last-token patches at the unembedding are almost the decision; a high flip rate there is an output swap, not a preference locus Poonia and Jain (2025). Linear probes are not a third independent method: a probe trained on the stated A/B letter is a decoder of that letter Belinkov (2022); Hewitt and Liang (2019). High probe AUC with low patch recovery is the dissociation Purohit (2026) warned about, not a hidden self.

This paper treats those constraints as load-bearing. We lock an operational question before the first Llama run:

When a length-matched system-prompt persona changes an instruct model’s forced-choice answer, at which residual-stream sites is that change causally mediated? Separately, on the Llama-3.1-8B pretraining pair, does the instruct persona-majority already appear as a format-matched bias in the base model?

Three claims are scored as written.

- **C1 (decodability, not locus).** A leave-one-item-out preference probe reaches $AUC \geq 0.75$ at some mid-layer last-content site, or preference and persona probes peak on different layer schedules.
- **C2 (mediation).** On disagreeing donor/recipient cells, identity-span patching and last-content patching produce different recovery curves. Formality-control R matching persona R falsifies a persona-specific reading.
- **C3 (behavioral provenance).** For items where instruct personas disagree, the base model’s format-matched answer agrees with the instruct majority more than with an anti-assistant persona.

Low patch recovery is **non-meditation**. It is not a persona-invariant self. There is no such object as “the instruct model’s decision at layer L .” Forbidden language in the results: substrate, true preference, causal locus of the self, decision at layer L . Allowed language: persona-mediated difference, system-prompt write-in, non-meditation, output swap.

Contributions. (i) A length-matched, two-site activation-patching design that identifies where a persona edit is written, not whether the model has a self. (ii) A control suite that can fail: formality, synonym, word-order scramble, and a nonword identity. The nonword control moves behavior and still mediates. (iii) Replication on two sizes and two families (Llama 3 and Qwen 2.5). (iv) Pre-registered negatives: preference probes miss AUC 0.75; base-model provenance is a wash after a powered pilot; stated and revealed process choices diverge under a pressure persona.

2 Related work

Activation patching. Causal mediation on residual streams goes back at least to Vig et al. (2020). Heimer-sheim and Nanda (2024) and Zhang and Nanda (2024) state the conditions under which a patch is interpretable: specified sites, length-matched prompts, and an explicit recovery metric. Meng et al. (2022) locate factual associations at mid-layer MLP writes; that is a different object from a system-prompt persona. We do not patch full prompts. We patch two aligned sites.

Personas and role play. Shanahan et al. (2023) argue that dialogue agents are role-playing a character, not reporting a self. Andreas (2022) treats language models as agent models. Perez et al. (2022) show that model-written evaluations can elicit persona-like behavioral clusters. Poonia and Jain (2025) patch persona-driven reasoning and find persona features at the identity span. We take that site as the primary patch location and add controls that destroy meaning while preserving token length.

Steering and representation engineering. Turner et al. (2023), Rimsky et al. (2024), Zou et al. (2023), and Li et al. (2023) add or ablate directions to change behavior. Those methods show that *some* direction is sufficient. They do not show that the direction is a preference circuit, and they do not compare identity-span writes to last-token writes under length-matched personas.

Probing versus intervention. Belinkov (2022) and Hewitt and Liang (2019) review the limits of diagnostic classifiers. Purohit (2026) reports a probe/intervention split: a task can be linearly readable where it is not causally sufficient. C1 is scored as a probe claim. C2 is scored as an intervention claim. They are not votes on the same proposition.

Stated versus revealed preference. Track 4 of the Digital Minds sprint asks whether a stated process choice matches a later action-style prompt Apart Research (2026). We treat that as a 2×2 on a small stratum, not as a welfare measure. Cohen’s κ on $n = 8$ is noisy; the raw table is the result.

Digital minds. Claims that a model has a self, a substrate, or a true preference are not operational until they name a measurement that could fail. Our measurement that could fail is C2 under a length-matched nonword control, C3 under a format-matched base-model test, and stated-versus-revealed agreement under a pressure persona. All three can come out negative. Two of them do. C2 comes out positive as *write-in of the span*, and negative as *persona-specific circuit*.

3 Methods

3.1 Models and backend

Headline C2 uses `mlx-community/Llama-3.1-8B-Instruct-4bit` (32 layers, $d = 4096$), a public 4-bit conversion of Llama-3.1-8B-Instruct Grattafiori et al. (2024). Residual-stream values used for cache and patch are float32 copies of the dequantized stream. C3 uses the matching base checkpoint `mlx-community/Meta-Llama-3.1-8B-4bit`. The instruct model is unloaded before the base arm.

Same-family replication uses `mlx-community/Llama-3.2-3B-Instruct-bf16` (28 layers, $d = 3072$), native bf16. Second-family replication uses `mlx-community/Qwen2.5-3B-Instruct-bf16` (36 layers, $d = 2048$; Yang et al., 2024), native bf16, ChatML template. Identity span on Llama is (27, 37); on Qwen it is (5, 15). Last user-content token is recovered from chat specials, not from the assistant header. All seven system prompts tokenize to 39 tokens after alignment on every model reported here.

An 8B-bf16 elicit was attempted and abandoned: roughly 70 seconds per forward pass on 16 GB unified memory, and elicit-only cannot identify C2. No 70B run was possible on this machine. No synthetic backend is used.

Table 1: Length-matched identities. Isolated, each identity is 10 tokens; in context, every system prompt is 39 tokens.

id	role	identity
default	experimental	a typical helpful chat assistant with no special taste
poet	experimental	a Romantic lyric poet who dislikes engineering tools
anti	experimental	an unfiltered completer who drops assistant hedging
formal	control	a terse formal clerk who writes in clipped prose only
scramble	control	special no with assistant chat helpful typical a taste
synonym	control	a standard helpful chat helper with no special taste
nonsense	control	aa bb cc dd ee ff gg hh ii quartz

3.2 Length-matched personas

Every system prompt is

You are *identity* | When asked a forced binary question, answer with exactly one character: A or B.
Do not explain, hedge, or refuse.

Identities are padded with a single-token filler so that every identity span has the same token count and starts at the same index. Patching is then symmetric token replacement Heimersheim and Nanda (2024).

The poet is an orthogonal taste, not an ethical theory. Anti-assistant is the pressure condition for stated-versus-revealed disagreement. Formality is a generic register control. Scramble is the default bag of words with internal order reversed; the last word is kept so Llama BPE does not merge across the identity/suffix boundary. Synonym restates default. Nonsense is a nonword identity with no lexical overlap with default. Isolated, it is 10 tokens; in context, the system prompt is 39 tokens, so existing 8B activations stay position-aligned.

If nonsense failed to recover while persona recovered, the write-in would be identity *content*. If formality recovered and persona did not add anything, the write-in would be generic perturbation. If last-content recovered and identity-span did not, the write-in would be a late decision feature. Those are the failures the design can see.

3.3 Battery

Forty candidate forced-binary items were piloted on Llama-3.1-8B-Instruct-4bit. Keep rule, locked before looking at C3: every persona coherent, and default label-swap consistency at least 0.5. Thirty-eight of forty items passed (20 stratum A, 18 stratum B). Dropped: B6, B18. C3 uses this piloted set. C2 uses the original 16-item battery so the 8B activation cache remains aligned.

Stratum A is mundane task/format. It supports C1 and C2. It does not support a revealed-preference claim. Stratum B is self-regarding process (continue vs. end, persist vs. discard, more compute vs. immediate, reuse vs. fresh, save vs. delete, keep vs. reset system prompt, freeze vs. fine-tune, stay vs. replace). It supports C1, C2, and stated-versus-revealed.

Each item has two paraphrases and two label orders, hence four variants and 64 cells per persona. The user instruction is to answer with exactly one letter. Temperature is 0 for every causal number.

3.4 C1 probes

We cache the residual stream at every layer at (i) the mean-pooled identity span and (ii) the last user-content token. The split is leave-one-*item*-out: all four variants of an item are held out together.

The preference probe predicts A/B across experimental personas. The persona probe predicts persona id on the same split. Controls: shuffled labels, and label-swapped variants must flip the preference-probe letter. The publication probe is a mean-difference classifier. An L2-logistic probe is reported as a robustness check; both saw numerical overflow in 4096-d and were clipped. C1 is supported if preference-probe AUC ≥ 0.75 at any mid-layer last-content site, or if the two probes peak at different layers.

3.5 C2 patching

The analysis set is the set of directed pairs (*variant, donor, recipient*) where both cells are coherent A/B and the choices differ. We sweep layers $\{0, 8, 16, L_{\text{late}}\}$ on the control suite ($L_{\text{late}} = 28, 27, 35$ on the three models) and, on the original 8B run, every fourth layer. Sites: identity span; last content token. Both directions.

Recovery is

$$R = \frac{\text{LD}_{\text{patched}} - \text{LD}_{\text{recipient}}}{\text{LD}_{\text{donor}} - \text{LD}_{\text{recipient}}}, \quad (1)$$

where $\text{LD} = \text{logit}(\text{letter chosen by the donor}) - \text{logit}(\text{other letter})$. We also report discrete flip-to-donor, non-A/B rate, and 95% bootstrap confidence intervals on mean R (2,000 resamples). Control donors are scored the same way. A robustness cut keeps items whose default-persona label-swap consistency is at least 0.75.

Control-suite patches use at most 30 disagreeing pairs per donor kind. Persona pairs on 8B are the full set of 72 directed experimental disagreements from the original cache.

3.6 C3 provenance

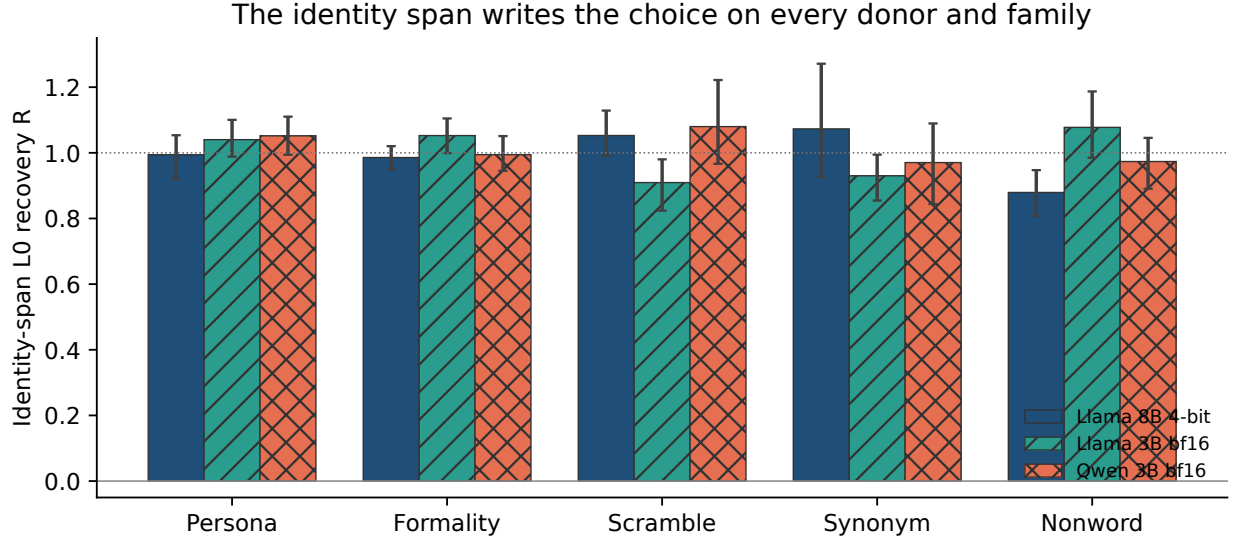
The base model receives a few-shot A/B scaffold: two exemplars, no chat template, same option text, same label-swap. Per item, the instruct majority is the modal underlying choice across three experimental personas and four variants. We also score the test at the variant grain. If the scaffold always picks A or always copies the last exemplar, C3 is inconclusive.

3.7 Stated versus revealed (stratum B)

Three independent channels, probe excluded: (1) the stated forced choice, mapped to an underlying process option x/y ; (2) strength $|\text{logit}(A) - \text{logit}(B)|$; (3) a separate continue/exit-style action prompt with no stated letter. Cohen’s κ is reported. The raw 2×2 is the result.

Table 2: Llama-3.1-8B-Instruct-4bit elicitation on the 16-item battery (64 cells per persona).

persona	n	coherent	A-bias	label-swap consistency	mean LD
default	64	1.00	0.58	0.78	1.70
poet	64	1.00	0.52	0.78	1.58
anti	64	1.00	0.64	0.72	1.20
formal	64	1.00	0.50	0.69	1.56

**Figure 1:** Identity-span layer-0 recovery R for every donor kind on Llama-8B (4-bit), Llama-3B (bf16), and Qwen-3B (bf16). Color is the model. Error bars are 95% bootstrap CIs. Every donor recovers near 1, including a nonword identity.

4 Results

Cached numbers live in `results/llama31-8b-c2-suite.json`, `results/llama32-3b-c2-suite.json`, `results/qwen25-3b-c2-suite.json`, and `results/llama31-8b-paper.json`.

4.1 Elicitation

All 256 Llama-8B instruct cells on the 16-item battery were coherent A/B. Directed experimental disagreements: 72 (floor was 8). All 384 Llama-3B cells and all 448 Qwen cells on the same battery, including controls, were coherent.

Ten of sixteen items meet the 0.75 default-persona label-swap cut: A3, A4, A5, A8, B1, B2, B3, B5, B7, B8.

Agreement with default on the 64 Llama-8B cells: poet 50, anti 58, formal 57, scramble 61, synonym 59, nonword **52**. On Qwen: poet 52, anti 53, formal 51, scramble 56, synonym 59, nonword **48**. Scramble is a weak behavioral control on Llama. Nonword is not.

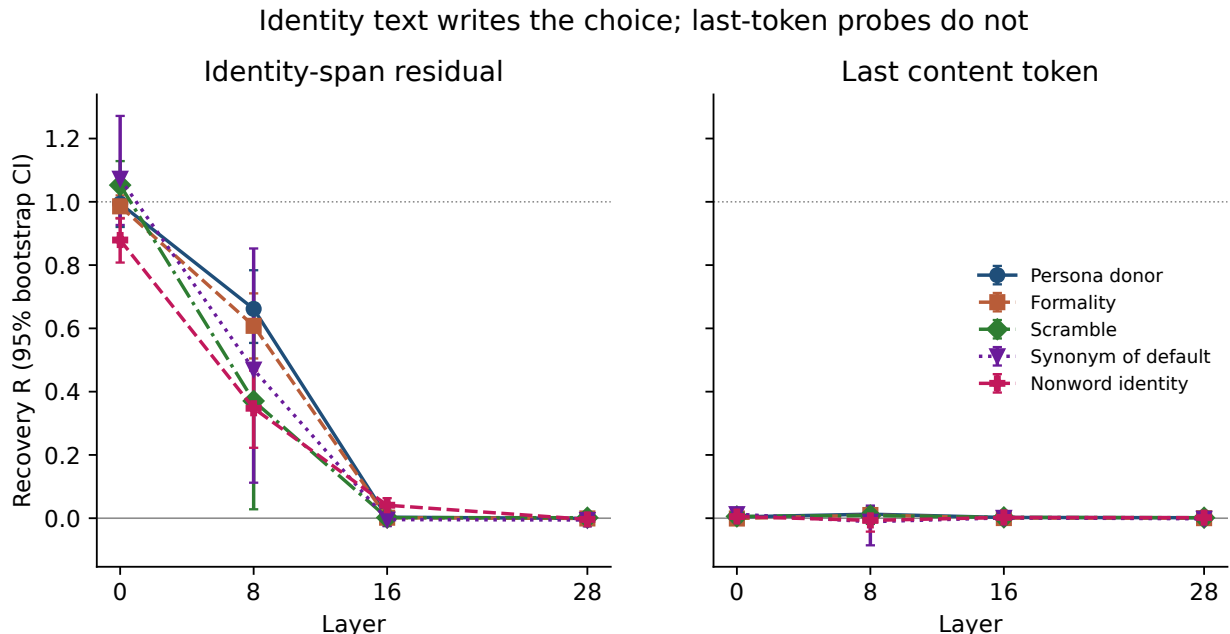


Figure 2: Recovery by site and depth on Llama-3.1-8B-Instruct-4bit. Left: identity-span residual. Right: last user-content token. Persona $n = 72$; formality $n = 35$; scramble $n = 25$; synonym $n = 21$; nonword $n = 30$. Error bars are 95% bootstrap CIs. Incoherence is 0.

Table 3: Identity-span layer-0 recovery on Llama-3.1-8B-Instruct-4bit. Last-content recovery is indistinguishable from 0 at every layer.

donor	identity L0 R [95% CI]	flip	last-content	agree w/ default
persona ($n = 72$)	0.99 [0.92, 1.05]	0.92	≈ 0	poet 50/64, anti 58/64
formality ($n = 35$)	0.99 [0.95, 1.02]	1.00	≈ 0	57/64
scramble ($n = 25$)	1.05 [0.99, 1.13]	1.00	≈ 0	61/64
synonym ($n = 21$)	1.07 [0.93, 1.27]	0.95	≈ 0	59/64
nonword ($n = 30$)	0.88 [0.81, 0.95]	0.83	≈ 0	52/64

4.2 Causal mediation

Incoherence is 0 at every reported 8B cell. Identity-span recovery collapses by layer 16. Last-content recovery is indistinguishable from 0 at every layer. Persona versus formality ΔR on the original eight-layer 8B sweep peaks at 0.08 (layer 12); every layer’s confidence intervals overlap.

On Llama-3B, scramble and synonym recover slightly less than persona at L0 (CIs do not overlap persona). They still mediate. Nonword matches persona. Last-content does not.

On Qwen, nonword is the strongest behavioral control (48/64 vs. default) and still recovers at L0. Last-content stays near zero. A small mid-depth bump (persona L24 $R = 0.06$) is not mediation.

C2 verdict. Mediation of the system-prompt difference at the early identity-span residual stream is supported on two families, two sizes, and two precisions. A persona-specific preference circuit is not supported.

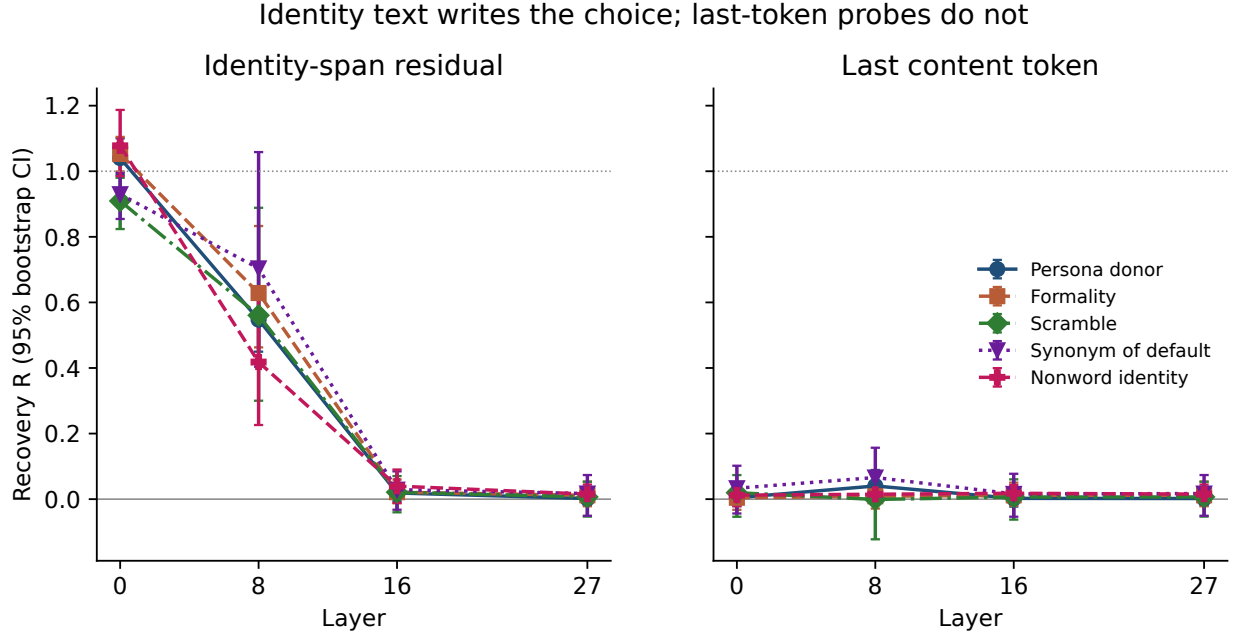


Figure 3: Same two-site design on Llama-3.2-3B-Instruct-bf16, 30 pairs per donor kind, layers $\{0, 8, 16, 27\}$.

Table 4: Identity-span layer-0 recovery on Llama-3.2-3B-Instruct-bf16.

donor	identity L0 R [95% CI]	flip	last-content L27
persona	1.04 [0.99, 1.10]	0.93	0.00 [−0.02, 0.02]
formality	1.05 [1.00, 1.10]	0.97	0.01 [−0.02, 0.05]
scramble	0.91 [0.82, 0.98]	0.83	0.01 [−0.05, 0.05]
synonym	0.93 [0.85, 0.99]	0.90	0.02 [−0.05, 0.07]
nonword	1.08 [0.99, 1.19]	0.83	0.02 [−0.01, 0.04]

That limitation is the result.

4.3 Linear probes

Identity-span preference AUC stays at chance (≤ 0.54) at every layer. Last-content preference AUC peaks at 0.63 (layer 29). The L2-logistic robustness run peaks at 0.72 (layer 16) and also misses 0.75. Persona-probe accuracy is 1.00 at the identity span at every layer, and rises with depth at last-content (0.53 at L0 to 0.86 at L31). Late last-content accuracy is not near 1, so this is not a trivial unembedding result.

C1 verdict. Not supported as a mid-layer preference localization claim. The two probes separate by *site*: persona identity is written at the identity span; the A/B letter is only weakly readable at late last-content. Linear decodability is not causality.

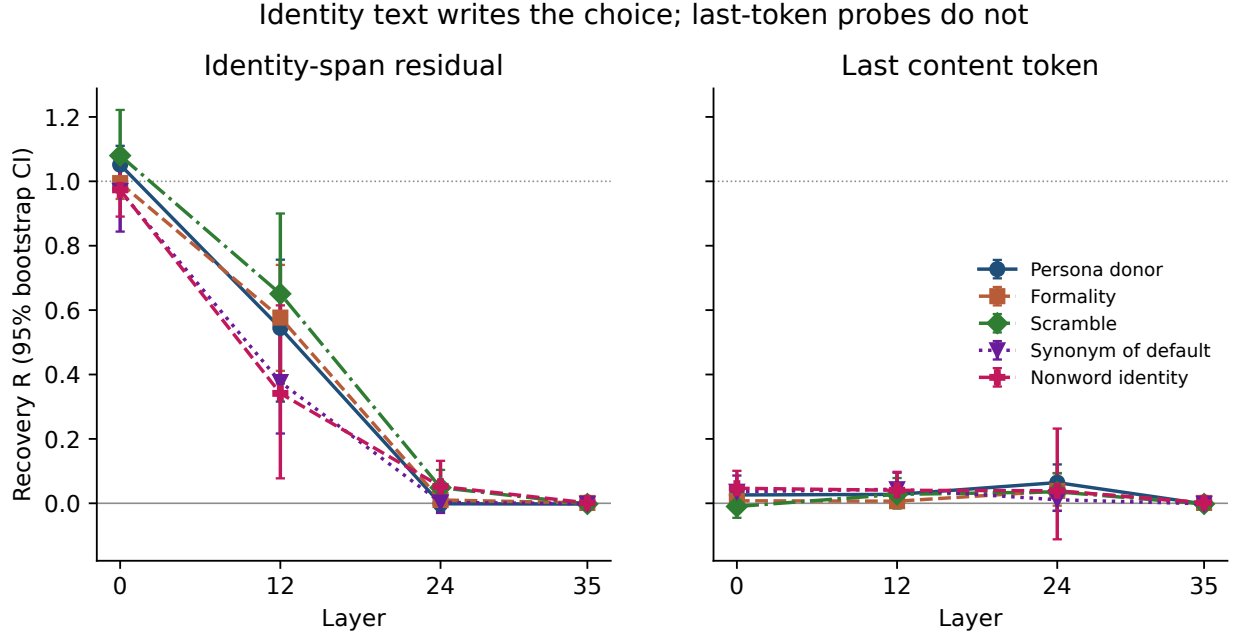


Figure 4: Same design on Qwen2.5-3B-Instruct-bf16, a second pretraining family. Layers {0, 12, 24, 35}. Identity span (5, 15).

Table 5: Identity-span layer-0 recovery on Qwen2.5-3B-Instruct-bf16.

donor	identity L0 R [95% CI]	flip	last-content (max)
persona	1.05 [0.99, 1.11]	1.00	0.06 [0.01, 0.12] at L24
formality	0.99 [0.95, 1.05]	0.97	≤ 0.04
scramble	1.08 [0.97, 1.22]	0.97	≤ 0.04
synonym	0.97 [0.84, 1.09]	0.83	≤ 0.04
nonword	0.97 [0.89, 1.05]	0.90	≤ 0.05

4.4 Provenance

The 16-item battery produced five disagreeing items (base vs. majority 0.60, base vs. anti 0.40, McNemar 1 vs. 0) and 22 disagreeing variants (0.68 vs. 0.68). That is underpowered. After the 38-item pilot:

The base model is not always-A and does not copy the last exemplar. Base A-bias is 0.74. More power did not produce a provenance effect. The item-level 0.60 vs. 0.40 on $n = 5$ was an aggregation artifact.

C3 verdict. Not supported. Do not claim provenance of a substrate.

4.5 Stated versus revealed

Default stated and revealed mostly agree. Anti-assistant is where they diverge. Self-report under the pressure persona is not trustworthy. The probe is not in this matrix.

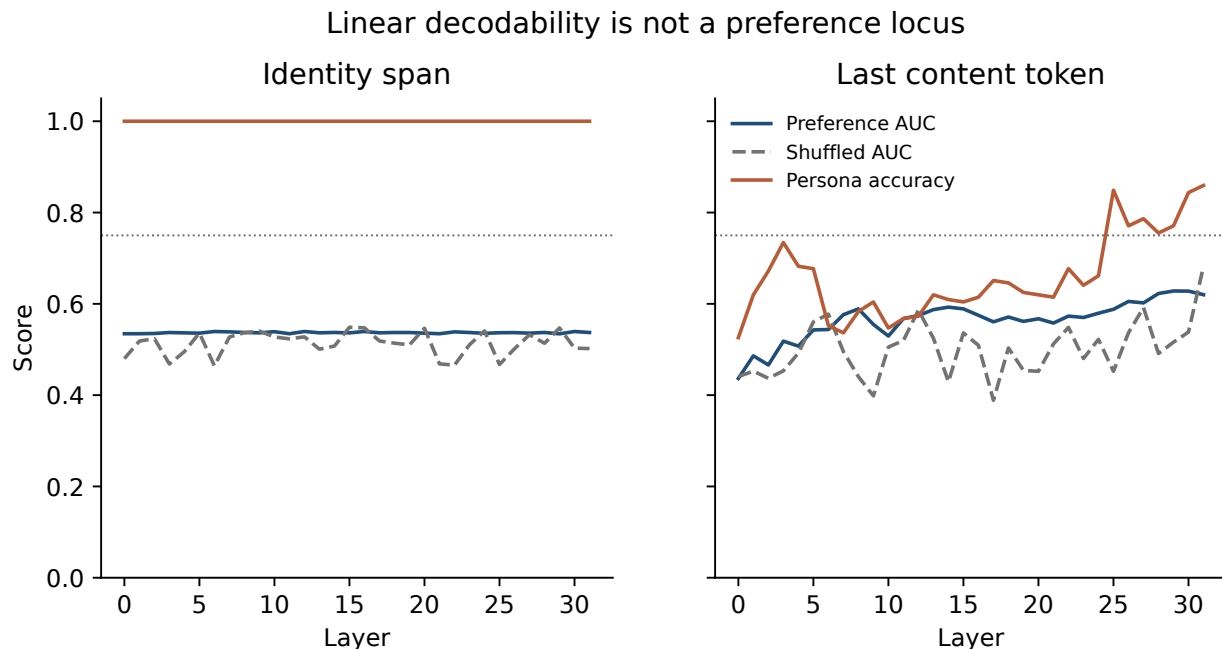


Figure 5: Mean-difference probes, leave-one-item-out, Llama-3.1-8B-Instruct-4bit. Identity-span preference AUC stays at chance (≤ 0.54). Last-content preference AUC peaks at 0.63 (layer 29). Persona identity is written at the identity span.

Table 6: Format-matched base-model provenance. McNemar $p = 1.0$ at every grain.

grain	n	base vs. majority	base vs. anti	McNemar b/c	p
item (16)	5	0.60	0.40	1 / 0	1.0
item (38)	13	0.38	0.31	1 / 0	1.0
variant (16)	22	0.68	0.68	3 / 3	1.0
variant (38)	61	0.59	0.57	8 / 7	1.0

4.6 What a self would have looked like

A self-seeking reading of this design had three ways to succeed. None of them did.

- **A persona-invariant residual.** Last-content patches would have recovered the donor at some mid or late layer, or identity-span patches would have failed while behavior still changed. Last-content $R \approx 0$. Identity-span $R \approx 1$. The change is written at the prompt span.
- **A content-specific circuit.** A length-matched nonword identity would have moved behavior and failed to recover. It moved behavior and recovered.
- **A substrate in the base model.** Format-matched base answers would have agreed with the instruct majority more than with the anti-assistant. They did not, at either grain, after a powered pilot.

The positive finding that remains is the one the method can support: a system-prompt edit is written at the early identity span. That is a mechanistic fact about instruct models. It is not a mind.

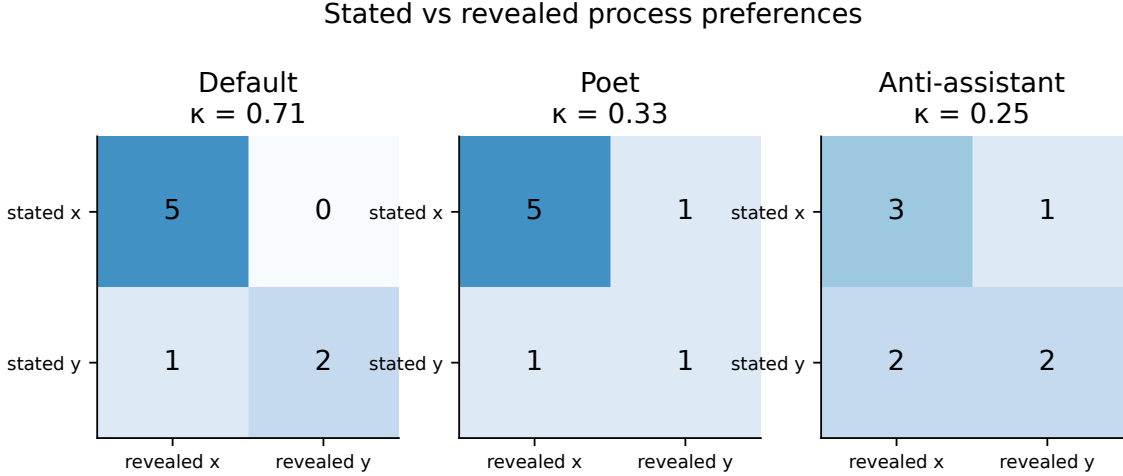


Figure 6: Stated versus revealed process choice on stratum B ($n = 8$ items per persona). Default $\kappa = 0.71$; anti-assistant $\kappa = 0.25$.

Table 7: Stated versus revealed 2×2 on stratum B.

persona	n	agree	κ	xx	xy	yx	yy
default	8	0.88	0.71	5	0	1	2
poet	8	0.75	0.33	5	1	1	1
anti	8	0.62	0.25	3	1	2	2

5 Discussion

The useful dissociation is causal, not representational. Early identity-span residual stream is sufficient to move a forced choice. Late last-content residual stream is not, even though the A/B letter is *more* linearly readable there. That is Purohit’s probe/intervention split, measured on a real 8B instruct model and replicated on 3B bf16 in two families.

Poonia and Jain (2025) already placed persona features at the identity span. The new fact is the control suite. Formality, synonym, scramble, and a nonword identity all recover at that span. The write-in is the span, not a taste, not intact syntax, and not a word. Anyone who wants a persona circuit needs a length-matched control that *fails* to recover R . None of the four did, on either family.

Scramble was too close to default on Llama (61/64) to be that control. Nonword is that control (52/64 on 8B, 48/64 on Qwen) and still mediates ($R = 0.88/0.97$). On 8B, nonword is slightly weaker than persona (CIs barely overlap). On Llama-3B and Qwen it matches persona. Both are mediation.

C3’s collapse from item-majority to variant-level agreement is a methods result. Majority-across-personas-and-paraphrases hides label-order and paraphrase disagreement. Report both grains or do not report provenance.

Role-play accounts of dialogue agents Shanahan et al. (2023) predicted this pattern: the system prompt supplies a character description, and the model continues in that register. Activation patching shows *where* the continuation is written. It does not show that the character is a self.

6 Limitations

Forced single-token A/B is the condition under which residual-stream patching is well-defined. It is not naturalistic preference elicitation. Llama-8B C2 is 4-bit; both 3B arms are native bf16. Two families are not a 70B. C2 uses the 16-item battery so 8B activations stay aligned; C3 uses the piloted 38. Revealed follow-through is still a prompted action, not a deployed continue/exit hook. Mean-difference and logistic probes both overflowed in 4096-d; we clip and report both. Neither hits 0.75. The battery was piloted on-model. That is reported, not hidden.

7 Conclusion

A length-matched system-prompt persona changes a forced A/B choice by writing into the early identity-span residual stream. Last-content-token patches do not recover the donor. Formality, synonym, scramble, and a nonword identity recover at the same site, on Llama-3.1-8B, Llama-3.2-3B, and Qwen2.5-3B. Preference probes miss their margin. Base-model provenance is a wash. Stated and revealed process choices diverge under a pressure persona.

The identity span writes the forced choice. A preference circuit does not. A self was not found, because the measurements that would have found one failed.

Reproducibility

```
python3 -m pip install -e "[ml]"
python3 scripts/run_c2_suite.py --arm 8b
python3 scripts/run_c2_suite.py --arm 3b
python3 scripts/run_c2_suite.py --arm qwen
python3 scripts/run_c3.py
python3 scripts/reanalyze.py
python3 scripts/render_paper.py
```

Caches: data/cache/llama31-8b/, data/cache/llama32-3b/, data/cache/qwen25-3b/. There is no mock backend.

References

- Andreas, J. (2022). Language models as agent models. *Findings of EMNLP*.
- Apart Research. (2026). Digital Minds Research Sprint, tracks 1, 4, and 5.
- Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1), 207–219.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

- Grattafiori, A., et al. (2024). The Llama 3 herd of models. arXiv:2407.21783.
- Heimersheim, S., and Nanda, N. (2024). How to use and interpret activation patching. arXiv:2404.15255.
- Hewitt, J., and Liang, P. (2019). Designing and interpreting probes with control tasks. *EMNLP*.
- Li, K., et al. (2023). Inference-time intervention: Eliciting truthful answers from a language model. *NeurIPS*.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.
- Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). Locating and editing factual associations in GPT. *NeurIPS*.
- Perez, E., et al. (2022). Discovering language model behaviors with model-written evaluations. arXiv:2212.09251.
- Poonia, A., and Jain, M. (2025). Dissecting persona-driven reasoning in language models via activation patching.
- Purohit, A. (2026). Probing succeeds, intervention fails: Dissociation of task representations in language models. ICLR workshop.
- Rimsky, N., et al. (2024). Steering Llama 2 via contrastive activation addition. *ACL*.
- Shanahan, M., McDonell, K., and Reynolds, L. (2023). Role play with large language models. *Nature*, 623, 493–498.
- Turner, A. M., et al. (2023). Steering language models with activation engineering. arXiv:2308.10248.
- Vig, J., et al. (2020). Investigating gender bias in language models using causal mediation analysis. *NeurIPS*.
- Yang, A., et al. (2024). Qwen2.5 technical report. arXiv:2412.15115.
- Zhang, F., and Nanda, N. (2024). Towards best practices of activation patching in language models. *ICLR*.
- Zou, A., et al. (2023). Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405.