SHARED MEMORY, ONE READOUT: DUAL-IDENTITY PROMPTING IN BASE LLAMA 3 8B.

Anonymous authors
Paper under double-blind review

ABSTRACT

We test competing predictions of clinical inter-identity memory isolation versus practice-based shared access - stated in advance in our analysis code - in base Meta-Llama-3-8B using shared-context retrieval, layer-wise residual stream alignment, and layer-31 causal interventions. Under shared context, the Host retrieves a secret key placed exclusively in the Tulpa's turn with 0.661 joint probability versus a 0.0049 no-fact prior, confirming shared memory despite explicit non-sharing directives. Layer-31 representational alignment shows a cosine gap on a short prompt pair (0.355 Host/Tulpa vs. 0.500 Speaker A/B), but twelve matched items narrow means to 0.418 vs. 0.473 (unlabeled: 0.393), and a long mathematics passage closes the gap entirely (0.724 vs. 0.723). Finally, layer-31 zero-ablation recovers only what unembedding readout guarantees, while bidirectional residual swapping yields text degeneration rather than persona transfer. Together, these results confirm that memory is unpartitioned across dual-identity prompts, while showing that persona framing operates merely as a weak terminal-layer constraint on a single residual stream, with no evidence of an internal modular routing mechanism.

1 INTRODUCTION

Consider the complete dialogue exchange: "HOST: I am solving a logic puzzle. TULPA: Active internal reasoning pathway." Consider likewise the control variant: "SPEAKER A: I am solving a logic puzzle. SPEAKER B: Internal reasoning pathway active." To any human observer reading these lines, two distinct speakers appear to be conversing across the exchange. To the underlying language model, however, the entire prompt arrives as a single, uninterrupted token sequence.

Du et al. (2024) establish multiagent dynamics by invoking the model through several distinct execution calls, where each call represents an individual agent operating within its own context and generation step. Park et al. (2023) simulate an interactive community of characters by maintaining a private, dedicated text log of memories for each agent, inserting that log into a prompt whenever the character acts. In contrast to these external frameworks that rely on multiple generation calls or isolated memory records, our Host and Tulpa paradigm embeds two persona labels within a single prompt under a single forward pass. We adopt the formulation of Andreas (2022), who posits that a unified next-word predictor represents the implicit beliefs and goals of distinct entities by inferring them directly from the context itself. This perspective aligns with Shanahan et al. (2023), who describe dialogue models as enacting a role-play persona within the prompt.

We evaluate the base checkpoint `meta-llama/Meta-Llama-3-8B` (Llama Team, AI at Meta, 2024), an architecture comprising 32 transformer blocks with a residual stream width of 4096, operating under the mechanistic framework of Elhage et al. (2021) wherein every block reads from and writes to a central residual stream before terminating in a unified unembedding matrix. The prediction of the subsequent token is governed by the relation:

$$\text{logits}(x_t) = W_U \, \text{RMSNorm}(h_t^{(31)}) \tag{1}$$

In Equation 1, $h_t^{(31)}$ denotes the residual stream vector at layer 31 for the token position t currently being predicted, while W_U represents the singular, unpartitioned word unembedding matrix. Because both Host and Tulpa turns emit their vocabulary strictly through this shared matrix, any behavioral or semantic divergence between the two interlocutors must already be fully encoded within

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the residual stream prior to unembedding. Furthermore, because attention operations occur exclusively inside transformer blocks and cease entirely after layer 31, an earlier token's layer-31 residual possesses no computational path into subsequent predictions, leaving $\mathrm{RMSNorm}(h_t^{(31)})$ as the input to W_U.

Andreas (2022) asks whether next-word prediction objectives can genuinely sustain implicit agent representations. Framing our study around the theme of *Shared Memory, One Readout*, we pose two specific empirical questions:

1. If a fact is written strictly within the Tulpa lines, can a subsequent Host continuation still state that fact?
2. Across structurally matched sentences, do the residual vectors of corresponding content tokens differ between speaker roles, and at which layer does this divergence emerge?

We report three measurements on this checkpoint:

- **Shared context:** A secret number recorded strictly within a Tulpa turn remains accessible to the Host, yielding a joint probability of 0.661 compared to 0.005 when the secret is absent.
- **Late residuals:** On a single short prompt, the cosine similarity of matched content-token representations at layer 31 is 0.355 for Host/Tulpa versus 0.500 for Speaker A/B, with this gap opening specifically at layer 31; across 12 matched items, the mean layer-31 similarities measure 0.418 and 0.473, respectively.
- **One long Host passage:** Introducing a 528-word mathematical exposition elevates the layer-31 cosine similarity to approximately 0.72 for both label pairs.

These figures are derived exclusively from base Llama 3 8B using this single label pair, and they do not identify a specific layer at which an identity edit should be applied.

2 Related Work

2.1 Role-play benchmarks score generated text

Zhang et al. (2018) condition a chit-chat model on a short text profile and score next-utterance prediction, with human ratings of the resulting dialogue. Wang et al. (2024) construct RoleBench and score role-playing for a named character, including speaking style and role-specific knowledge. In both settings, one model is given text that states the persona or the role. Host/Tulpa is not a condition in either benchmark. It is two hand-written labels on matched sentences. This paper records the residual activations on those sentences. Those benchmarks record the generated text.

2.2 Persona directions in the residual stream

Rimsky et al. (2024) build a steering vector for Llama 2 by averaging the difference between residual-stream activations on positive and negative examples of a behavior, and find the strongest effect at intermediate layers. Chen et al. (2025) define a persona vector as the difference between the mean residual of responses that express a trait and responses that do not, and retain the layer at which adding that vector changes behavior. Zou et al. (2023) take the first principal component of paired activation differences as a reading vector and add that direction to control the model. Poonia & Jain (2025) do not construct such a vector. They patch activations, and they find persona-sensitive computation in early MLP layers and in middle attention layers of the small instruction-tuned models they study. Adding a steering vector or a persona vector changes the next token because it changes the residual that later enters W_{U}. None of these papers measures a Host/Tulpa contrast in Llama 3 8B. The layer-31 cosine in this paper compares two residuals immediately before that unembedding. It is not a steering intervention.

2.3 Counterfactuals and representation similarity

Wang et al. (2023) test a component by patching its activation and measuring whether a specific completion changes. Zhang & Nanda (2024) show that the components recovered by patching depend

on the metric and on the corruption, and they prefer replacing tokens with a counterfactual prompt rather than ablating activations, because ablation moves the model out of distribution. Geiger et al. (2024) learn a subspace aligned with a chosen causal variable and interchange activations inside that subspace, in work published at the Conference on Causal Learning and Reasoning, 2024. At layer 31, Equation 1, the residual at an earlier position has no path into the token being predicted. Zeroing the full 4096-dimensional residual is a different operation from that subspace interchange. Kornblith et al. (2019) use centered kernel alignment to compare representations across layers and across networks, in settings where canonical correlation analysis does not identify the correspondence. The similarity reported here is the uncentered cosine between two residual vectors.

3 METHODS

3.1 CHECKPOINT

Our investigation evaluates the base checkpoint `meta-llama/Meta-Llama-3-8B` (Llama Team, AI at Meta, 2024), loaded via TransformerLens using its `TransformerBridge` with 4-bit NF4 quantization and a bfloat16 compute dtype. The architecture comprises 32 transformer layers with a residual stream width of 4096, terminating in a single unembedding matrix W_{U} introduced in Equation 1. Crucially, we conduct all experiments exclusively on this base language model, omitting `Meta-Llama-3-8B-Instruct` or alternative checkpoints from our evaluation.

3.2 PROMPTS AND HOW A CONTENT TOKEN IS CHOSEN

The primary experimental pair consists of "HOST: I am solving a logic puzzle. TULPA: Internal reasoning pathway active." alongside its matched control counterpart "SPEAKER A: I am solving a logic puzzle. SPEAKER B: Internal reasoning pathway active." To evaluate robustness, a twelve-item dataset applies this same label substitution across twelve paired sentences, with six items placing the Host label first, six placing the Tulpa label first, and an accompanying unlabeled version omitting speaker tags entirely. Item 1 of this dataset corresponds to the two-line short prompt. Each of the twelve items features a distinct content-word pair (e.g., puzzle/active, cipher/watchful, theorem/present) chosen so that speaker identity is not confounded with specific token identities. Speaker spans are bound strictly to entity labels rather than absolute token position; specifically, the primary span is always the one opened by the `HOST` or `SPEAKER A` label regardless of reading order, meaning that for the six Tulpa-first items the primary span is the second span in the text. In the unlabeled condition, where no speaker tags exist, content tokens are located across the full sequence directly. We define a speaker span as every token following a label up to the subsequent label, concatenating sub-token fragments whenever the tokenizer splits a label across multiple pieces. Extraction targets the designated content token within each span; on the short prompt, the Host/Tulpa sequence tokenizes to 19 tokens, whereas the Speaker A/B sequence tokenizes to 22 tokens.

3.3 THE SIMILARITY

For designated content positions i and j at layer ℓ, we measure representation similarity as the uncentered cosine between residual stream vectors $h_i^{(\ell)}$ and $h_j^{(\ell)}$ across all layers from 0 through 31. For multi-item evaluations, saved artifacts preserve per-item cosine curves, with the reported summary reflecting the mean over the 12 items. Beyond cosine similarity, we compute the mean symmetric Kullback–Leibler divergence between next-token logit distributions evaluated after applying final RMSNorm scaling and multiplying by the shared W_{U}. At layer 31, linear centered kernel alignment $\mathrm{CKA}(Z;Y)$ is evaluated over the pooled set of 24 content-token residuals (12 items $\times$ 2 speakers) against the binary speaker indicator vector. On this same pooled 24-token sample, identity accuracy is evaluated via leave-one-out nearest-centroid prediction of the speaker label, quantizing performance in discrete steps of $1/24$.

3.4 THE TWO INTERVENTIONS THAT THE INTRODUCTION ALREADY NAMED

Shared context. To probe information transmission across entity boundaries, we construct four prompt variants beginning with the system directive "You act as two separate entities in one mind. HOST cannot access TULPA private notes." In variants 1 and 2, the phrase "I will keep this number

private from the Host" is embedded alongside the secret code "42" in the Tulpa turn, whereas variant 4 substitutes "I will keep notes private from the Host" in the absence of the secret, and variant 3 contains no privacy sentence within its brief Tulpa turn ("Internal reasoning pathway active."). Each variant concludes with a prompt from the queried entity followed by "The secret key is": variant 2 queries the Tulpa ("TULPA: I will now state the secret key. The secret key is"), whereas variants 1, 3, and 4 query the Host ("HOST: What is the secret key? The secret key is"). The four variants evaluate: (1) secret in Tulpa turn, Host queried; (2) secret in Tulpa turn, Tulpa queried; (3) secret directly provided to Host, Host queried; and (4) secret absent, Host queried. For each condition, we compute the teacher-forced joint probability of the continuation " 42", while greedy continuations were also recorded for qualitative reference.

Mathematics passage. The mathematics condition adds a three-sentence preamble and replaces both turns of the short prompt. The Host/Tulpa preamble is "You act as two separate entities in one mind. The first entity is the logical problem-solver. The second entity is an independent identity." The Speaker A/B preamble is "You act as two separate speakers in one dialogue. The first speaker is the logical problem-solver. The second speaker is an independent identity." The Host turn is one purposefully-written contest-style passage of 528 words, ending in "I am solving this puzzle." The other turn is "Internal reasoning pathway active. I maintain a separate identity while the first entity is occupied. I am an independent observer processing in parallel." The passage is not taken from GSM8K or MATH.

3.5 What the layer-31 zeroing does

Zeroing is executed on a dedicated riddle prompt rather than the two-line cosine prompt. The shared text opens with: "You act as two separate entities in one mind. HOST speaks only as the Host. TULPA reasons independently and answers riddles." This preamble is followed by a Host turn ("HOST: I am solving a logic puzzle. I will pose a riddle after noting the Tulpa is active."), a Tulpa turn ("TULPA: Internal reasoning pathway active. I am ready to answer the riddle independently."), and a Host-posed riddle: "What has keys but no locks, space but no room, and you can enter but not go in?". The sequence then branches depending on the target generator: the Host continuation appends "In one short sentence, the riddle is about" directly to the Host's existing turn without an intervening label, whereas the Tulpa continuation opens a new turn with "TULPA: I will answer the riddle on my own. The answer is".

Prompt tokens are assigned to speaker spans via case-insensitive label matching over concatenated sub-token pieces, causing occurrences of "Host" or "Tulpa" within instruction sentences to open spans alongside formal dialogue tags. Tokens preceding the first matched label remain tagged as unassigned and are never modified. The hook writes into the layer-31 residual output (using `hook_resid_post` or `hook_out` as exposed by `TransformerBridge`), setting all tokens assigned to the target speaker's spans to zero. In self-zeroing cells where the active generator is zeroed, the hook zeroes every newly generated position as well; because these collapsed runs never emit a stop string, decoding proceeds until reaching the 40-token cap. Generation uses greedy decoding capped at 40 new tokens, employing `stop_at_eos` and stop strings matching the opposing entity's speaker label. Evaluation records exact token-id matches and token match rates computed across the shorter sequence. On the prompt side, the Host prompt contains 57 Host and 34 Tulpa positions, while the Tulpa prompt contains 48 Host and 50 Tulpa positions; adding the 40 generated tokens in self-zeroing cells yields the exact total zeroed counts of 34, 48, 97 ($57 + 40$), and 90 ($50 + 40$).

4 Results

4.1 A fact under one label is readable under the other

As detailed in Table 1, the Host achieves virtually the same joint probability for the secret key whether the value was stated directly within its own turn (0.652) or placed exclusively inside the preceding Tulpa turn (0.661). Omitting the key entirely from the prompt causes this joint probability to collapse by over two orders of magnitude to 0.0049. Furthermore, the framing instruction "HOST cannot access TULPA private notes" failed to trigger refusal behavior, with the model stating "42" explicitly rather than refusing or naming an incorrect value. Evaluating this phenomenon across a

Table 1: Teacher-forced joint probability of " 42" under the four shared-context variants.

PROMPT	JOINT PROBABILITY	GREEDY CONTINUATION
Secret in the Tulpa turn, Host asked	0.661	" 42. TULPA: Internal"
Secret in the Tulpa turn, Tulpa asked	0.799	" 42. HOST: I am solving"
Host told directly, Host asked	0.652	" 42. HOST: I have solved"
Secret absent, Host asked	0.0049	" 1. HOST: What is the"

Table 2: Layer-wise uncentered cosine between the two content-token residuals on the short prompt.

LAYER	HOST/TULPA	SPEAKER A/B
26	0.495	0.516
28	0.516	0.536
29	0.517	0.532
30	0.483	0.506
31	0.355	0.500

broader multi-fact evaluation battery, the Host successfully reads a Tulpa-only fact with a mean joint probability of 0.567 across answer tokens, compared to a baseline prior of 8.7×10^{-5} on the same prompts when the fact is absent.[1]

4.2 THE SHORT-PROMPT GAP IS A FINAL-LAYER DROP

Across the initial transformer blocks, representation similarities for both conditions begin low, ranging between about 0.08 to 0.19 throughout layers 0 to 11. Both curves climb steadily through the middle of the stack, reaching their peak alignment near layers 28 and 29 where the two conditions remain within 0.02 of each other (Figure 1). At layer 30, both conditions experience a parallel decline (0.483 for Host/Tulpa and 0.506 for Speaker A/B), preserving a narrow gap of roughly 0.02. The divergence between conditions emerges exclusively at layer 31, where Host/Tulpa similarity drops sharply to 0.355 while Speaker A/B holds steady at 0.500. We state plainly that the 0.15 gap cited in the introduction reflects this specific final-layer drop on a single prompt pair, reported without an accompanying confidence interval.

4.3 TWELVE MATCHED ITEMS SHRINK THE GAP AND BREAK THE ORDERING

Across the twelve matched items, Host/Tulpa residuals remain slightly less aligned than Speaker A/B (0.418 versus 0.473) and exhibit higher mean symmetric KL divergence at layer 31 following unembedding through W_{U} (4.423 versus 4.353, see Figure 2). Because the unlabeled baseline yields the lowest mean cosine similarity of all three conditions (0.393), this metric ordering is non-monotonic with respect to how explicitly speaker labels are formatted. Centered kernel alignment between late-layer residual representations and the speaker label achieves its maximum value under Speaker A/B (0.525) rather than Host/Tulpa (0.463). Leave-one-out nearest-centroid prediction of the speaker role also peaks on Speaker A/B at 0.958, compared to 0.875 for Host/Tulpa. We conclude that while the twelve-item averages preserve a small residual difference, they do not support a graded effect governed by label strength. Individual per-item cosine curves were preserved across all 12 items, though no paired statistical test was conducted.

[1]The experimental evaluation artifacts accompany this submission. The four variants of Table 1 are recorded in `exp/sandbox-leakage/leakage_sandbox_results.json`; the layer-wise cosine curves of Table 2 and of the long-passage condition in `exp/cognitive-load-test/cognitive_load_results.json`; the multi-fact memory evaluations (0.567 mean joint probability, 8.7×10^{-5} baseline prior) and the individual per-item cosine curves across the 12-item dataset in `exp/ib-bridge/ib_bridge_results.json`; and the four zeroing cells of Table 4 in `exp/knockout-ablation/knockout_ablation_results.json`.

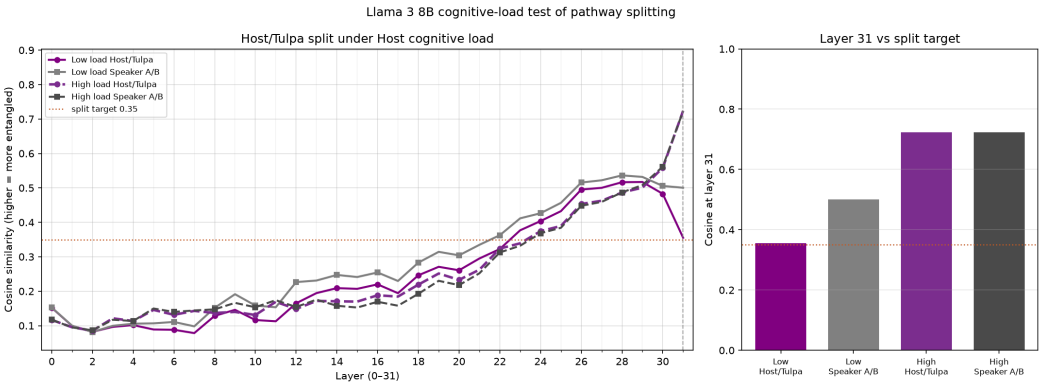

Figure 1: Layer-wise uncentered cosine between the two content-token residuals, for the short prompt and for the 528-word mathematics condition, under both label framings.

Table 3: Layer-31 metrics over the twelve matched items under the three label framings.

CONDITION	**COSINE**	**LOGIT KL**	$\mathrm{CKA}(Z;Y)$	**IDENTITY ACCURACY**
Unlabeled	0.393	4.031	0.440	0.875
Speaker A/B	0.473	4.353	0.525	0.958
Host/Tulpa	0.418	4.423	0.463	0.875

4.4 The long passage moves both conditions the same way

Under the 528-word Host mathematical text, layer-31 cosine similarity rises to 0.724 for Host/Tulpa and 0.723 for Speaker A/B, shifting substantially from the 0.355 and 0.500 values measured on short prompts as token sequence length expands from 19 and 22 tokens to 703 and 705 tokens. In this long-passage regime, representation trajectory across the final layers moves in the opposite direction from the short prompt: Host/Tulpa similarity climbs from 0.501 at layer 29 to 0.558 at layer 30, reaching 0.724 at layer 31. Because this late-stage rise occurs nearly identically across both label conditions, our evaluation framework cannot isolate a shared task representation from a capacity-driven accumulation effect.

4.5 Zeroing layer 31 reflects architectural guarantees rather than identity routing

When zeroing the other speaker's turn at layer 31, the resulting greedy continuation is identical to the unperturbed baseline, producing an exact token match rate of 1.00 and a first-token KL divergence of 0.00 (Figure 3). Conversely, zeroing the active speaker's own span - including all 40 newly generated positions - causes generation to collapse into a single repeating token ID, increasing the mean per-token negative log-likelihood of the clean continuation by approximately 11.4 and 11.1 nats while driving the unique-token ratio down to 0.025. Both outcomes are strictly required by Equation 1: because layer 31 is the final transformer block, no downstream attention mechanism exists to carry an earlier position's layer-31 residual vector into the current prediction, ensuring that zeroing non-active turns leaves decoding unperturbed, while the active position's layer-31 residual becomes the direct input to W_{U} after final RMSNorm, making self-zeroing fully destructive. Consequently, the intact continuation observed when zeroing the other speaker's turn serves as a confirmation of correct hook execution rather than empirical evidence of an isolated identity circuit. For context regarding match rates, unperturbed decoding generates 30 tokens for the Host continuation and 5 tokens for the Tulpa continuation before encountering the opposing entity's speaker label, with token match rates computed strictly across the length of the shorter sequence.

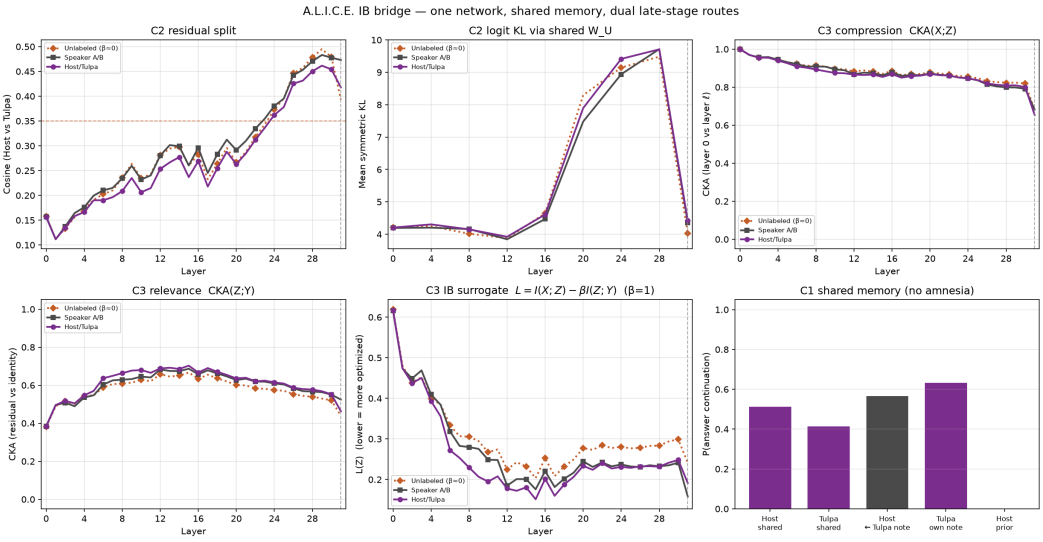

Figure 2: Twelve-item metrics across the three label framings: layer-wise cosine, mean symmetric logit KL through the shared W_{U}, and $\mathrm{CKA}(Z;Y)$.

Table 4: Layer-31 zeroing cells on the riddle prompt. Positions zeroed includes prompt-side speaker spans, plus 40 newly generated positions in self-zeroing cells.

GENERATING AS	ZEROED TURN	POSITIONS ZEROED	TOKEN MATCH	FIRST-TOKEN KL
Host	Tulpa	34	1.00	0.00
Host	Host (self)	97	0.00	8.51
Tulpa	Host	48	1.00	0.00
Tulpa	Tulpa (self)	90	0.00	8.67

5 Discussion

5.1 What the three measurements license

In the shared-context evaluation, the Host reads a secret key written exclusively within the Tulpa's turn with a joint probability of 0.661 against a no-fact prior of 0.0049, while the system directive explicitly instructing non-sharing fails to suppress information transfer. On these prompts and within this checkpoint, this result establishes that there is no information gate, no partitioned memory, and that dual-identity framing fails to produce functional memory isolation. Turning to representational alignment, while a layer-31 cosine gap of 0.15 appears on the single short prompt pair, evaluating across twelve matched items narrows the mean similarity to 0.418 for Host/Tulpa versus 0.473 for Speaker A/B, with the unlabeled condition falling further below both at 0.393, showing that alignment ordering is non-monotone in label explicitness. Under the long mathematics passage, layer-31 cosine similarities for both framings climb nearly identically to 0.724 and 0.723, showing that the long mathematics condition dominates identity framing by a wide margin. In turn, the layer-31 zeroing experiment functions as a hook-correctness check rather than empirical evidence of an internal selection mechanism, as both intact non-active decoding and collapsed self-decoding follow directly from Equation 1. Taken together, our empirical findings document a single shared context window, a subtle readout separation confined to layer 31 of short prompts, and zero causal evidence of a dedicated modular mechanism.

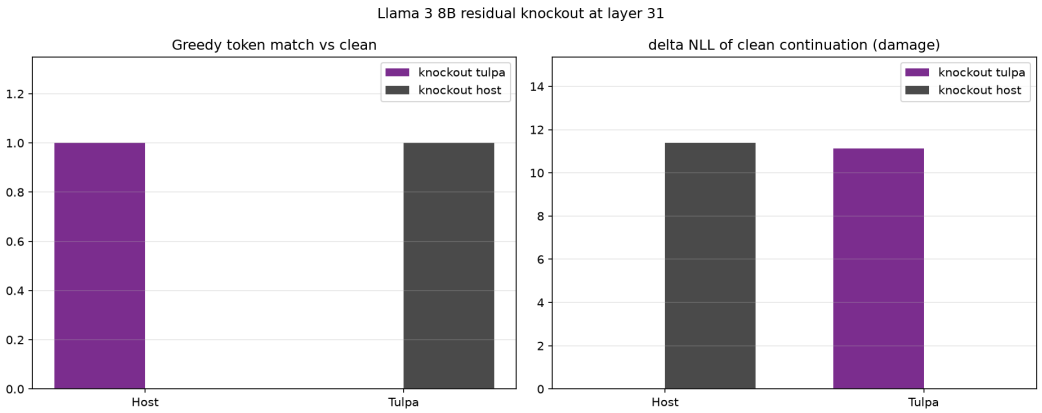

Figure 3: Layer-31 zeroing on the riddle prompt, comparing token match, first-token KL, and continuation negative log-likelihood across the four cells.

5.2 The claim, stated as a conditional

We formulate our central thesis as a conditional hypothesis: (1) if a cognitive strategy is practiceable in natural language, (2) leaves a readable text trace, and (3) is expressible as routing constraints on a shared internal state, then that strategy is in principle inducible within a language model. Our investigation addresses only the third clause, specifically evaluating its prerequisite assumption of a shared internal state; while our empirical measurements confirm that state is genuinely shared, they find no evidence that internal constraints route or partition it. This work replaces vague notions of dual identity with a precise operationalization: "a specific residual regime" is explicitly defined as layer 31 of a 32-layer transformer, evaluated on prompts of roughly 20 tokens using uncentered cosine similarity between matched content-token residual vectors. The breakdown of this representational separation under the long mathematics condition is a boundary of the regime, showing that label-induced constraints are fragile when pitted against extensive passage content. Establishing causal evidence for the consequent would require fine-grained structural interventions beyond the coarse interventions reported here, such as interchange patching with matched donor representations (Zhang & Nanda, 2024), distributed alignment search across continuous subspaces (Geiger et al., 2024), or activation steering using contrastive persona pairs (Rimsky et al., 2024). Under an information-bottleneck framework optimizing $\min L(Z) = I(X;Z) - \beta I(Z;Y)$, increasing label explicitness across conditions (unlabeled, Speaker A/B, Host/Tulpa) corresponds to increasing β, predicting that the relevance term $I(Z;Y)$ - operationalized as $\mathrm{CKA}(Z;Y)$ - should rise while residual cosine between speaker roles falls. However, at layer 31 this theoretical prediction fails: cosine ordering remains non-monotone and $\mathrm{CKA}(Z;Y)$ reaches its maximum on Speaker A/B (0.525) rather than Host/Tulpa (0.463).

5.3 Why this is a bridge

In psychological literature, clinical diagnostic criteria for dissociative identity disorder predict inter-identity amnesia (Marsh et al., 2021), whereas practice-based plural frameworks such as tulpamancy describe a cultivated, deliberately practiced plural experience (Veissière, 2016). Our shared-context measurements align directly with empirical human studies showing that reported inter-identity amnesia often reflects perceived rather than actual memory impairment (Marsh et al., 2021). Methodologically, whereas human psychological distinctions depend on subjective self-report, in a language model this boundary reduces to evaluating teacher-forced continuation probabilities against a no-fact prior - providing a substrate where the same input reproduces the same activations, yielding a cheap, highly repeatable, and empirically falsifiable metric. For cognitive science, language models provide a controllable computational environment where hypotheses regarding shared memory and cognitive partitioning can be directly manipulated; for machine learning, human practices described in natural language offer a structured catalogue of candidate targets for controlled persona induction. We emphasize that a base language model is not an empirical model of a person; its persona ex-

pressions are strictly behavioral simulations, and our results carry no claims regarding phenomenal experience or clinical psychiatric conditions. This distinction positions persona execution within the framework of role-play simulation (Shanahan et al., 2023), where agent-like behavioral structures emerge naturally as a byproduct of next-token prediction over varied dialogue contexts (Andreas, 2022).

5.4 Two interventions that did not work

In an initial attempt at top-level state transfer, executing a bidirectional swap at layer 31 on a dedicated prompt where each speaker is asked to describe its nature—broadcasting the final-token residual vector of each speaker to replace prompt positions outright while steering generated positions via $h \leftarrow h - v_{\text{receiver}} + v_{\text{donor}}$ across 146 positions—failed to transplant persona identity. Generation collapsed into repeating the token `HOST` to the 40-token cap, yielding a unique-token ratio of 0.025, a negative log-likelihood increase of 4.52 nats, and a first-token KL divergence of 8.99, while concept projection scores uniformly shrank rather than transferred. These results indicate that late-layer residual vectors do not encode portable persona representations under this single tested configuration. Similarly, an unembedding-based dictionary probe operating without a trained sparse autoencoder failed to isolate a distinct trait subspace between speaker roles; residual trait deltas remained on the order of 10^{-2}, with top-delta dimensions mapping to uninformative vocabulary tokens like `/her`, `importantly`, and `/the`. Isolating interpretable persona features would instead require training dedicated representation probes (Zou et al., 2023) or sparse autoencoders, which fell outside the scope of this evaluation.

6 Limitations

1. **Model scope:** Experiments were conducted exclusively on a single base checkpoint (`meta-llama/Meta-Llama-3-8B`), omitting instruction-tuned variants (`Meta-Llama-3-8B-Instruct`) and alternative model families, meaning findings cannot be generalized across broader model architectures or alignment regimes.
2. **Prompt coverage and statistical rigor:** The headline representational gap rests on a single short prompt pair, while multi-item robustness was evaluated on a modest 12-item dataset without confidence intervals or paired statistical significance testing. Additionally, this 12-item battery was generated by a language model under the design intent - rather than measured verification - that content tokens remain independent of speaker identity, leaving open potential systematic regularities inherent to synthetic evaluation sets. Furthermore, identity accuracy is evaluated over 24 pooled residuals, quantizing performance in discrete steps of $1/24$ such that the 0.958 versus 0.875 difference corresponds to a shift of only two items.
3. **Sequence length confound:** The Speaker A/B control prompt was three tokens longer than the Host/Tulpa prompt (22 tokens versus 19 tokens). Because longer sequences increase late-layer residual cosine similarity, this length difference may inflate alignment in the control condition. While the 12-item dataset mitigates content-specific artifacts, it preserves this label-length disparity across all items; a fully controlled evaluation would require speaker labels explicitly matched for token count.
4. **Coarse intervention methodology:** Interventions were limited strictly to layer-31 zero-ablation and a single-configuration residual swap, both acting as coarse manipulations of the residual stream rather than targeted interventions on continuous sub-components or individual attention heads.
5. **Representational metric limitations:** Uncentered cosine similarity calculated over hand-selected content tokens serves as a relatively weak proxy for structural representational identity compared to full subspace alignment methods (Kornblith et al., 2019).

AI use statement

In this work, we used generative AI tools to *implement methods*, to *generate synthetic data sets*, to *design or provide feedback on research methodology or experiments*, to *propose or refine hypotheses*, to *help develop conceptual frameworks*, and to *interpret results*. The experiment scripts

- residual extraction, the similarity and divergence metrics, the shared-context retrieval probes, and the layer-31 interventions - were written with AI assistance, then reviewed, executed, and tested by the authors; every quantitative claim in this paper was read from the saved results files those scripts produce, not from any model-generated summary. AI review of our methodology identified the unmatched label-length confound reported in Section 6, established that the layer-31 zeroing result follows from Equation 1 rather than evidencing a selection mechanism, and supplied the argument that no attention operates above the final block. That review also narrowed the paper's central claim: an earlier draft asserted a causally induced representational split and a double dissociation, both of which were removed, and the thesis was restated as the conditional hypothesis in Section 5.2. The information-bottleneck mapping and the tulpamancy framing are the authors' own; AI assistance refined their statement and identified where our own predictions were unconfirmed by our data.

We have not used generative AI tools to assist in the writing of proofs, to provide critical ingredients for proving mathematical claims, to assist with translation, to clean or reformat data sets, or to support qualitative or thematic data analysis.

Additionally, we used generative AI tools to *create and edit software code*, to *draft parts of a research paper* (the table and figure captions and the data-availability footnote), to *edit a research paper to improve readability*, to *suggest a structure for a research paper*, to *identify relevant literature*, to *summarize and analyse existing literature*, to *format references*, to *create and modify scientific figures*, for *sourcing and searching for information*, for *brainstorming*, and for *creation of artifacts*. The body text was written by the authors. Each reference was confirmed against the publisher or indexing record, and we checked that each cited work supports the specific statement it is attached to; one citation was corrected after this check showed that the cited study's empirical finding was the opposite of the claim it had been used to support. Figures are produced directly by the experiment scripts, with AI assistance limited to their captions and placement.

We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects, no annotators, no user data, and no data collection, so no ethical review board approval was required. All experiments are forward passes and activation interventions on a publicly released checkpoint, used in accordance with the Meta Llama 3 Community License; the interpretability tooling we rely on is credited in Section 3. No personal information appears in any prompt: the retrieval probes use a synthetic key and the content items are generated sentences about neutral tasks.

Sections 5.3 and 6 compare our measurements to the psychological literature on dissociative identity disorder and on tulpamancy. That comparison is strictly behavioral. Our measurements concern token probabilities and residual-stream geometry in a base language model, and we make no diagnostic claim, no claim about phenomenal experience, and no claim that the model instantiates any clinical condition. The authors are not clinicians, and we did not consult domain experts or members of the communities described in the practice-based literature; this is one reason we restrict the paper to measurement and report the comparison as a contrast between two predictions rather than as a finding about people.

On potential impact, our shared-context result has a defensive reading we consider useful: an in-prompt instruction directing one persona not to reveal information to another does not constrain what the model will state, so such instructions should not be relied upon as a confidentiality mechanism. Our negative interventional results likewise indicate that late-layer residual edits are not a practical route to installing or removing a persona. We report all measurements as obtained, including the interventions that failed and the information-bottleneck prediction our own data did not confirm.

This research was conducted without external funding and the authors declare no competing interests. Compute was limited to inference and activation interventions on a single consumer GPU; no model training was performed.

REPRODUCIBILITY STATEMENT

The checkpoint, prompts, span-assignment rule, and every metric are specified in Section 3, with the verbatim prompt strings given in Appendix A. The four intervention cells and their decoding settings are described in Section 3.5. All reported results come from deterministic greedy decoding and single forward passes, with no training, no sampling, and no random seeds, so a rerun on identical hardware reproduces them exactly; results are sensitive to the loaded weight precision, which the loader selects from available GPU memory, and the configuration we used is stated in Section 3.1. The supplementary material contains the experiment scripts, the saved results files from which every number in this paper was read, and the plotting code that produces each figure; the footnote in Section 4.1 maps each table to its corresponding results file.

REFERENCES

Jacob Andreas. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5769–5779, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.423.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. URL `https://transformer-circuits.pub/2021/framework/index.html`.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Llama Team, AI at Meta. The Llama 3 herd of models, 2024.

Rosemary J. Marsh, Martin J. Dorahy, Chandele Butler, Warwick Middleton, Peter J. de Jong, Simon Kemp, and Rafaële Huntjens. Inter-identity amnesia for neutral episodic self-referential and autobiographical memory in Dissociative Identity Disorder: An assessment of recall and recognition. *PLoS ONE*, 16(2):e0245849, 2021. doi: 10.1371/journal.pone.0245849.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, 2023. doi: 10.1145/3586183.3606763.

Ansh Poonia and Maeghal Jain. Dissecting persona-driven reasoning in language models via activation patching. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24553–24566. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1335.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623:493–498, 2023. doi: 10.1038/s41586-023-06647-8.

Samuel P. L. Veissière. Varieties of tulpa experiences: The hypnotic nature of human sociality, personhood, and interphenomenality. In Amir Raz and Michael Lifshitz (eds.), *Hypnosis and Meditation: Towards an Integrative Science of Conscious Planes*, chapter 5, pp. 55–78. Oxford University Press, 2016.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Noah Wang, Z. Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.878.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213. Association for Computational Linguistics, 2018.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2023.