

WiringPrint: What Does a Neural model Count as “Itself”: parameter/functional/representational?

Digital Minds Research Sprint 2026: Agnivo and Ayush Gandhi

Track 5: The Assistant Persona & Model Identity

Abstract

We investigate model identity at the level of computation rather than conversational persona. We establish the core result first on pretrained GPT-2 and then test its generality on Qwen2.5-1.5B, Gemma-2-2b, and Llama-3.2-1B. In GPT-2, a compensated permutation of the 3,072-neuron MLP hidden axis leaves accuracy (0.3155), perplexity (56.629), ECE (0.1050), and logits invariant up to floating-point error, while an uncompensated permutation reduces accuracy to 0.0409 and increases perplexity to 3892.811. We then construct analogous transformations for modern gated MLPs and grouped-query attention.

The central claim is deliberately narrower than a claim about consciousness: raw parameter coordinates are insufficient as a complete criterion for individuating a neural model. Some large changes in weight space correspond to the same computation, while breaking the wiring relation changes behavior. This provides an empirical distinction between parameter identity, functional identity, and representational identity, and gives a mechanistic framework for investigating model, instance, and persona identity.

Contents

1	WiringPrint: What Does a Transformer Count as “Itself”?	3
1.1	Abstract	3
2	Motivation	3
3	Research Questions	4
4	Models	4
5	GPT-2 Foundational Experiment	4
5.1	Main GPT-2 result	5
5.2	GPT-2 representation geometry	5
5.3	GPT-2 identity boundary	5
5.4	GPT-2 harder-task validation	6
5.5	GPT-2 multi-seed validation	6
5.6	GPT-2 perturbation controls	6
6	Method	6
6.1	Baseline behavioral battery	6
6.2	MLP permutation	7
6.3	Attention-head permutations	7
6.4	Representation analysis	7

6.5	Causal tracing	8
6.6	Robustness and generalization	8
7	Results	8
7.1	Aligned permutation preserves behavior	8
7.2	The gate-only transformation also breaks the function	9
7.3	Multi-seed confirmation	9
7.4	Harder tasks	9
7.5	Long-context evaluation	9
7.6	GQA changes the symmetry group	10
7.7	Cross-model CKA	10
8	What does this say about model identity?	10
8.1	Parameter identity	10
8.2	Functional identity	10
8.3	Representational identity	10
9	Implications for the digital-minds question	11
10	Limitations	11
11	Next Experiments	12
11.1	A. Persona invariance under function-preserving transformations	12
11.2	B. Instance identity	12
11.3	C. Mechanistic vs. behavioral identity	12
11.4	D. Cross-model transfer	12
11.5	E. Identity benchmark	12
12	11. Conclusion	13
13	References	13
	References	13

1 WiringPrint: What Does a Transformer Count as “Itself”?

1.1 Abstract

What makes an AI system the same system across changes to its parameters? In ordinary software, renaming internal variables does not create a new program when the computation is unchanged. Neural networks complicate this intuition because their internal units are represented by numerical coordinates in weight space. A hidden-unit permutation can produce a radically different parameter tensor while leaving the implemented function unchanged.

We establish the core result first on pretrained GPT-2 and then test its generality on three real open-weight transformer families: Qwen2.5-1.5B, Gemma-2-2b, and Llama-3.2-1B. In GPT-2, a compensated permutation of the 3,072-neuron MLP hidden axis leaves accuracy (0.3155), perplexity (56.629), ECE (0.1050), and logits invariant up to floating-point error, while an uncompensated permutation reduces accuracy to 0.0409 and increases perplexity to 3892.811. We then construct analogous transformations for modern gated MLPs and grouped-query attention.

The main result is robust: correctly compensated MLP permutations preserve next-token accuracy within numerical precision across GPT-2 and the three modern families, while uncompensated permutations and gate-only permutations substantially degrade behavior. Multi-seed experiments reproduce the separation. The same principle extends to attention, but grouped-query attention changes the symmetry group: the relevant exact transformations act on compatible query/KV head groups rather than allowing arbitrary query-head permutations.

We interpret these findings as evidence for a mechanistic distinction between **parameter identity** and **functional identity**. The result does not establish anything about consciousness, welfare, or moral patienthood. Instead, it supplies a concrete tool for Track 5’s individuation problem: if two parameter states implement the same computation under a symmetry, then raw weight coordinates are not sufficient to identify the entity. Conversely, when the wiring relation is broken, behavior changes, giving an experimentally testable boundary between equivalent and non-equivalent internal descriptions.

2 Motivation

The Digital Minds Research Sprint asks a deceptively difficult question: when an AI assistant talks about “itself,” what entity could that word refer to? Candidate units include the underlying model, a particular runtime instance, the conversational persona, or the conversation state.

Behavioral probing alone has a serious limitation. Two systems can produce similar language for very different mechanistic reasons, while the same underlying computation can admit multiple internal parameterizations.

This problem therefore asks a more primitive question:

What transformations of a neural network change the parameters without changing the computation, and what transformations actually change the computation?

The motivation comes from classical results on neural-network weight-space symmetries. Hidden units can often be permuted without changing the network function when the adjacent weights are transformed consistently [1–3]. Modern model-merging work has made these symmetries operational for aligning independently trained networks [2,3].

We use these mathematical symmetries as an experimental probe of identity.

3 Research Questions

We test five questions.

1. **Functional identity:** Does a compensated permutation of hidden units preserve model behavior, first in GPT-2 and then in modern transformers?

2. **Architecture dependence:** Does the symmetry survive across SwiGLU/GeGLU MLPs and grouped-query attention?

3. **Identity boundary:** How quickly does behavior degrade when the symmetry is deliberately violated?

4. **Representation vs. function:** Can behavior remain invariant even when internal coordinates are changed?

5. **Cross-model identity:** How similar are representations across independently trained model families, and what does that similarity actually establish?

The fifth question is intentionally separated from the first four. High representational similarity is not equivalent to functional identity, and functional equivalence is not evidence that two independently trained models are literally the same entity.

4 Models

We evaluate three pretrained, open-weight transformer families:

Model	Parameters	MLP structure	Attention
Qwen2.5-1.5B	~1.54B	SwiGLU / SiLU	GQA, 12 Q heads / 2 KV heads
Gemma-2-2b	~2.61B	GeGLU / GELU variant	GQA, 8 Q heads / 4 KV heads
Llama-3.2-1B	~1.24B	SwiGLU / SiLU	GQA, 32 Q heads / 8 KV heads

The checkpoints are real pretrained models rather than synthetic networks. The model configurations are documented in the corresponding model cards.

5 GPT-2 Foundational Experiment

Before testing whether the identity principle generalizes across modern architectures, we establish the core result on the original GPT-2 architecture. This provides a clean control because GPT-2’s MLP has a single hidden projection:

$$y = c_{\{\mathrm{proj}\}} \left(\operatorname{GELU}(c_{\{\mathrm{fc}\}}(x)) \right).$$

For a permutation matrix P ,

$$\begin{aligned} c_{\{\mathrm{fc}\}}' &= c_{\{\mathrm{fc}\}} P, \\ b_{\{\mathrm{fc}\}}' &= b_{\{\mathrm{fc}\}} P, \\ c_{\{\mathrm{proj}\}}' &= P^{\top} c_{\{\mathrm{proj}\}}. \end{aligned}$$

The experiment uses the pretrained `gpt2` checkpoint and applies the transformation to the 3,072-neuron hidden axis of MLP layer 0.

5.1 Main GPT-2 result

Condition	Accuracy	Perplexity	ECE
Baseline	0.3155	56.629	0.1050
Aligned MLP permutation	0.3155	56.629	0.1050
Unaligned MLP permutation	0.0409	3892.811	0.1055

The aligned transformation changes the maximum absolute sample logit by only $\{9.16 \times 10^{-5}\}$, consistent with floating-point error. The unaligned transformation produces a maximum absolute logit difference of 53.760.

Thus the GPT-2 experiment gives a particularly clean separation:

$\{$
 $\{\text{different parameters}\} + \{\text{same wiring}\}$
 $\rightarrow$
 $\{\text{same computation}\},$
 $\}$
whereas
 $\{$
 $\{\text{different parameters}\} + \{\text{broken wiring}\}$
 $\rightarrow$
 $\{\text{changed computation}\}.$
 $\}$

5.2 GPT-2 representation geometry

For the aligned transformation, linear CKA is 1.0000 at every layer and Procrustes distance is approximately zero. For the unaligned transformation, the first post-perturbation layer has CKA 0.7794 and later representations remain substantially altered; final-layer CKA falls to 0.0400.

This provides an important qualification to the identity claim. The permutation changes the internal coordinate system without changing the function. Therefore internal neuron labels cannot automatically be treated as persistent “parts” of the model.

5.3 GPT-2 identity boundary

We progressively break the permutation correspondence rather than testing only the two endpoints:

Symmetry violation	Accuracy	Perplexity
0%	0.3155	56.63
10%	0.3130	62.67
20%	0.2998	71.60
40%	0.2896	93.13
60%	0.2207	238.40
80%	0.1161	1327.42
100%	0.0412	6639.46

The degradation is approximately monotonic over this sweep. This is useful because it turns “identity” into a measurable boundary rather than an all-or-nothing philosophical label.

5.4 GPT-2 harder-task validation

On the informative harder-task controls:

- **Induction:** baseline 0.750, aligned MLP 0.750, unaligned MLP 0.000.
- **Compositional:** baseline 0.500, aligned MLP 0.500, unaligned MLP 0.000.
- **Multi-hop:** baseline 0.500, aligned MLP 0.500, unaligned MLP 0.000.

The antonym and arithmetic tasks have zero baseline accuracy and are therefore not evidence for preservation.

5.5 GPT-2 multi-seed validation

Across 15 permutation seeds, aligned MLP permutations have mean accuracy change $\{+0.00000\}$, while unaligned MLP permutations have mean change $\{-0.25263\}$, with 95% CI $[-0.26669, -0.23858]$, Cohen’s $d = -9.953$, and Wilcoxon $p = 0.0007$.

Aligned attention-head permutations also have zero mean accuracy change. Unaligned attention-head permutations have mean change $\{-0.09427\}$, 95% CI $[-0.12740, -0.06115]$, Cohen’s $d = -1.576$, and Wilcoxon $p = 0.0001$.

The aligned rows have zero variance because they implement an exact symmetry, so a conventional significance test against zero is undefined there. The correct evidence for those rows is the algebraic symmetry plus the observed numerical invariance.

5.6 GPT-2 perturbation controls

Alternative transformations behave differently from the exact permutation symmetry. Local MLP orthogonal rotation gives accuracy 0.0125 and perplexity $\{3.89 \times 10^7\}$; a global residual rotation gives accuracy 0 and perplexity $\{2.06 \times 10^8\}$. In contrast, moderate Gaussian weight noise initially changes perplexity only modestly, illustrating that raw perturbation magnitude is not equivalent to functional distance.

The important control is therefore not simply “large change vs. small change.” It is **whether the transformation belongs to the model’s exact symmetry group**.

6 Method

6.1 Baseline behavioral battery

Each model is evaluated using:

- next-token accuracy,
- perplexity,
- expected calibration error (ECE).

The baseline provides the behavioral reference against which every transformed condition is compared.

6.2 MLP permutation

For a conventional linear–nonlinearity–linear block,

$$y = W_2 \sigma(W_1 x + b_1),$$

a permutation matrix P can be inserted into the hidden coordinate system:

$$\begin{aligned} W_1' &= P W_1, \\ b_1' &= P b_1, \\ W_2' &= W_2 P^{\text{top}}. \end{aligned}$$

Because an elementwise nonlinearity commutes with a coordinate permutation,

$$\sigma(Pz) = P \sigma(z),$$

the transformed network computes the same function.

For SwiGLU/GeGLU, the hidden representation is produced by two branches:

$$h = \sigma(W_g x + b_g) \odot (W_u x + b_u),$$

followed by a down projection. Therefore the same permutation must be applied consistently to both branches and compensated in the down projection.

We test three conditions:

- **Aligned:** both branches are permuted consistently and the down projection is compensated.
- **Unaligned:** the correspondence is deliberately broken.
- **Gate-only:** only one branch is permuted.

The gate-only condition is a useful architecture-specific negative control.

6.3 Attention-head permutations

The attention experiment extends the same idea to GQA. Because query heads share K/V heads, arbitrary query-head permutations are not necessarily symmetries. The notebook explicitly constructs GQA-aware permutations and tests both valid group transformations and invalid within-group rearrangements.

6.4 Representation analysis

We collect hidden activations and compute:

- linear CKA,
- Procrustes distance,
- principal-angle style representation comparisons.

CKA is used as a representation-comparison statistic, following Kornblith et al. [5]. Importantly, we do not treat CKA as a direct measure of model identity.

6.5 Causal tracing

We adapt a ROME-style causal tracing protocol [8] to ask whether the transformed model retains the same localization of factual computations. Clean, corrupted, and patched conditions are compared across layers.

6.6 Robustness and generalization

We test:

- code and non-English inputs,
- harder few-shot completion tasks,
- long contexts,
- 15-seed behavioral sweeps,
- model scale,
- architecture family,
- training-time transformations,
- continuous rotation perturbations.

7 Results

7.1 Aligned permutation preserves behavior

The baseline accuracies were:

Model	Accuracy	Perplexity	ECE
Qwen2.5-1.5B	0.4650	1.30	0.0807
Gemma-2-2b	0.4655	26.088	0.0891
Llama-3.2-1B	0.4292	21.268	0.0919

Under the aligned MLP permutation, accuracy remained unchanged:

Model	Baseline accuracy	Aligned accuracy	Unaligned accuracy
Qwen2.5-1.5B	0.4650	0.4650	0.3696
Gemma-2-2b	0.4655	0.4655	0.4519
Llama-3.2-1B	0.4292	0.4292	0.4201

At the selected bf16 precision, the raw aligned logit differences were small but nonzero (approximately 0.22, 0.50, and 0.29 in the tested condition), while the unaligned differences were much larger (approximately 19.25, 2.3, and 11.75). Thus the correct claim is **functional invariance within numerical precision**, not literal bitwise equality.

This distinction matters. Calling bf16 arithmetic “exact” because the rounded accuracy happens to match would be a charming way to manufacture a reviewer comment.

7.2 The gate-only transformation also breaks the function

The gate-only condition produced large logit changes:

Model	Gate-only max logit difference
Qwen2.5-1.5B	17.00
Gemma-2-2b	23.75
Llama-3.2-1B	11.94

This supports the architectural prediction that the two branches of a gated MLP form a coupled coordinate system. Permuting one branch without the other is not merely a re-labeling.

7.3 Multi-seed confirmation

Across 15 seeds:

Model	Aligned accuracy delta	Unaligned accuracy delta
Qwen2.5-1.5B	+0.00000	0.05446
Gemma-2-2b	+0.00000	0.01749
Llama-3.2-1B	+0.00000	0.04025

Paired tests reported in the notebook separate aligned and unaligned conditions for all three models:

- Qwen2.5-1.5B: $(t=9, p=3.93 \times 10^{-5})$
- Gemma-2-2b: $(t=3.73, p=2.22 \times 10^{-3})$
- Llama-3.2-1B: $(t=1, p=1.10 \times 10^{-4})$

These tests are supportive rather than the foundation of the main result, because the exact symmetry is fundamentally an algebraic claim.

7.4 Harder tasks

The symmetry survives a few-shot antonym-completion test for the conditions where the baseline has meaningful performance:

- Qwen: baseline 1.000, aligned 1.000, unaligned 0.667.
- Llama: baseline 1.000, aligned 1.000, unaligned 0.000.

Gemma’s baseline accuracy on this particular task was 0, so it does not provide useful evidence for preservation.

7.5 Long-context evaluation

Perplexity remained effectively unchanged under aligned permutations:

Model	Baseline PPL	Aligned PPL	Unaligned PPL
Qwen2.5-1.5B	1.110	1.110	1.118
Gemma-2-2b	1.128	1.128	1.140
Llama-3.2-1B	1.127	1.126	1.133

This suggests the observed equivalence is not confined to a tiny evaluation corpus.

7.6 GQA changes the symmetry group

The attention experiments show why “permute any head” is too crude a description for modern transformers.

The models have fewer KV heads than query heads. The exact symmetry must preserve the query/KV grouping structure. The notebook therefore identifies the relevant group structure as a permutation of compatible head groups rather than arbitrary individual query-head relabeling.

This is a concrete example of an important general principle:

Architecture changes the equivalence relation, even when the underlying identity idea remains the same.

7.7 Cross-model CKA

Final hidden-state CKA across independently trained models was:

Pair		CKA
Qwen2.5-1.5B	Gemma-2-2b	0.9459
Qwen2.5-1.5B	Llama-3.2-1B	0.9732
Gemma-2-2b	Llama-3.2-1B	0.9576

These are interesting but should not be overinterpreted. CKA measures representational similarity, not identity. Two different systems can share representational geometry, and a single system can change its coordinate representation without changing its function.

8 What does this say about model identity?

The central conceptual result is a three-way distinction.

8.1 Parameter identity

Two states are identical if their parameter tensors are identical.

This is too strict for neural networks. A permutation of hidden coordinates can produce a different tensor while preserving the implemented computation.

8.2 Functional identity

Two states are equivalent if they implement the same input-output function.

This is a much more natural candidate for a computational notion of “the same model.”

However, it is not the only possible notion. Two models can implement the same function on a finite evaluation set while differing elsewhere.

8.3 Representational identity

Two states can also be compared by their internal representations.

Representation similarity is informative, but it does not automatically determine functional identity or moral identity.

The experiments therefore support the following restricted conclusion:

Raw parameter coordinates are insufficient to individuate a neural model.

This is directly relevant to the Digital Minds Sprint’s Track 5 question about whether the relevant entity is a model, an instance, or a persona. If two internal parameterizations differ by a function-preserving symmetry, counting them as different entities solely because their weight tensors differ would be arbitrary from a functional perspective.

The converse is also important. Breaking the wiring relation can change the computation even when the operation looks superficially similar. Thus “same architecture” or “same parameter count” is also insufficient.

9 Implications for the digital-minds question

The experiment does **not** establish that any model is conscious, sentient, welfare-bearing, or a moral patient.

It instead addresses an upstream methodological problem: **what exactly are we trying to identify when we say “the model”?**

A useful candidate decomposition is:

```
\{\{
\{\}text{weights}
\{\}rightarrow
\{\}text{computation}
\{\}rightarrow
\{\}text{runtime state}
\{\}rightarrow
\{\}text{conversation}
\{\}rightarrow
\{\}text{persona}\}.
\{\}
```

Different levels can vary independently.

For example:

- A weight permutation can change parameters while preserving computation.
- A new KV cache can change the runtime state while leaving the weights fixed.
- A system prompt can alter persona and behavior without changing the underlying weights.
- Two separately trained models can have similar representations without being the same computational object.

This suggests that future welfare or identity experiments should specify the level at which “individuality” is being attributed.

10 Limitations

1. **Only three model families.** The experiments do not establish universality across all transformer variants.

2. **Limited model scale.** The largest tested model is approximately 2.6B parameters.

3. **Finite behavioral tests.** Functional equivalence on the tested corpus does not prove equality over every possible input.

4. **bf16 arithmetic.** Numerical exactness must be distinguished from mathematical exactness.

5. **CKA limitations.** High CKA should not be treated as evidence of shared identity.
 6. **Causal tracing scope.** The causal experiments use a small number of controlled factual prompts.
 7. **Identity is partly normative.** Mechanistic equivalence does not by itself determine which entity deserves moral consideration.
 8. **Persona is not directly measured.** The present work provides a mechanistic foundation for persona/model individuation but does not yet run a full behavioral persona battery.
-

11 Next Experiments

The strongest follow-up is to connect the mechanistic identity criterion directly to the sprint’s persona question.

11.1 A. Persona invariance under function-preserving transformations

Compare self-reports from models before and after function-preserving parameter symmetries.

If the model’s reported “self” changes despite functional equivalence, this would provide evidence that persona depends on internal coordinates rather than only computation.

11.2 B. Instance identity

Hold weights fixed and vary:

- KV cache,
- conversation history,
- system prompt,
- sampling state.

Measure which changes the model treats as changes to “self.”

11.3 C. Mechanistic vs. behavioral identity

Construct matched pairs:

1. same function / different parameterization,
2. similar representations / different function,
3. same weights / different context.

Then ask models to classify whether they are “the same entity.”

11.4 D. Cross-model transfer

Extend the symmetry experiments to larger models and additional architectures, including models without GQA and models with alternative attention/MLP structures.

11.5 E. Identity benchmark

Build an explicit benchmark whose ground-truth transformations are known:

- **equivalent:** function-preserving symmetry,
- **near-equivalent:** small functional change,

- **non-equivalent:** deliberately broken wiring,
- **contextual:** same weights, changed runtime state,
- **persona:** same model, changed conversational framing.

This would turn “model identity” from a philosophical prompt into a measurable experimental object.

12 11. Conclusion

WiringPrint treats model identity as an empirical question rather than assuming that the parameter tensor is the model.

Across three real open-weight transformers, compensated hidden-unit permutations preserve behavior within numerical precision, while breaking the corresponding wiring relations changes behavior. Modern architecture matters: gated MLPs require coordinated permutations across branches, and GQA restricts which attention-head transformations are genuine symmetries.

The main contribution is therefore not the observation that permutations exist. That fact is classical. The contribution is the **cross-architecture experimental characterization of which internal transformations preserve computation, where the symmetry breaks, and how this distinction can be used to reason about the unit of identity in digital-minds research.**

The result supports a modest but useful methodological claim:

A model should not be individuated solely by its raw parameter coordinates.

That does not settle whether the morally relevant entity is the function, the runtime instance, the conversation, the persona, or something else. It does give us a way to experimentally separate those hypotheses.

13 References

The numerical results in this report are transcribed from the executed outputs embedded in the submitted research notebook. The notebook itself contains the complete experimental code, model-loading logic, perturbation implementations, statistical tests, and plots. Because the experiment uses gated/open-weight checkpoints and GPU memory management, exact runtime behavior depends on access permissions, hardware, and numerical precision.

References

1. Hecht-Nielsen, R. (1990). *On the Algebraic Structure of Feedforward Network Weight Spaces*. In *Advanced Neural Computers*. Elsevier.
2. Entezari, R., Kaskasoli, A., Daras, G., et al. (2022). *The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks*. International Conference on Learning Representations (ICLR). <https://openreview.net/forum?id=dNigytemkL>
3. Ainsworth, E., Hayase, J., and Srinivasa, S. (2023). *Git Re-Basin: Merging Models Modulo Permutation Symmetries*. International Conference on Learning Representations (ICLR). <https://openreview.net/forum?id=CQsmMYmlP5T>

4. Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). *Similarity of Neural Network Representations Revisited*. Proceedings of the 36th International Conference on Machine Learning (ICML). <https://proceedings.mlr.press/v97/kornblith19a.html>
5. Meng, K., Bau, D., Andonian, A., and Belinkov, Y. (2022). *Locating and Editing Factual Associations in GPT*. Advances in Neural Information Processing Systems (NeurIPS). <https://neurips.cc/virtual/2022/poster/53864>
6. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. OpenAI Technical Report.
7. Biderman, S., Schoelkopf, H., Anthony, Q. G., et al. (2023). *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling*. Proceedings of the 40th International Conference on Machine Learning (ICML). <https://proceedings.mlr.press/v202/biderman23a.html>
8. Qwen Team. *Qwen2.5-1.5B Model Card*. Hugging Face. <https://huggingface.co/Qwen/Qwen2.5-1.5B>
9. Google. *Gemma 2 2B Model Card*. Hugging Face. <https://huggingface.co/google/gemma-2-2b>
10. Meta AI. *Llama 3.2 1B Model Card*. Hugging Face. <https://huggingface.co/meta-llama/Llama-3.2-1B>