

PERSONA SENSITIVITY & SCRATCHPAD-OUTPUT CONSISTENCY

An exploratory cross-model study of self-report measurement

Main finding: No stable relationship between the two measures.
Qwen's estimated relationship is highly sensitive to the similarity metric.

WHY THIS MATTERED FOR DIGITAL MINDS

Self-reports are central

Preference, identity, and welfare questions often rely on what models say about themselves.

Persona framing matters

Answers can shift when the assistant frame is perturbed. Stability is not guaranteed.

Measurement risk

If different diagnostics disagree, treating any single channel as general-purpose evidence is risky.

INITIAL ASSUMPTION

Models that change more when the assistant persona is perturbed might also show greater divergence between elicited reasoning and final responses.

Reason: if persona framing changes self-reports, it might also affect how elicited reasoning aligns with final answers.

Original hypotheses (simplified)

H1: Positive association between persona-conditioned response divergence and scratchpad-output divergence.

H2: Dropped (alignment intensity was not independently measured).

H3: Not evaluable as written (Standard/Base-Like have no scratchpads).

EXPERIMENTAL SETUP

Models

Claude Sonnet
Llama 3.3 70B
Qwen 2.5 72B

Queries

5 preference &
identity probes

Conditions

Standard
Transparent
Base-Like

Runs

5 per cell
225 responses

Measures Persona-conditioned response divergence (Standard ↔ Base-Like) · Scratchpad-output divergence

Unit of analysis Model × query (n = 5 independent units per model)

Robustness check Shared TF-IDF (char 2–3 grams) + sentence-embedding similarity

PRIMARY RESULT: NO STABLE CROSS-MODEL RELATIONSHIP

Model	TF-IDF ρ	Semantic ρ	Pattern
Claude Sonnet	+0.60	+0.10	Positive → near zero
Llama 3.3 70B	-0.20	-0.30	Mildly negative both
Qwen 2.5 72B	-0.70	+0.60	Sign flips

No coefficient is confirmatory. n = 5 query units per model.

The original expectation of a common positive relationship is not supported.

ROBUSTNESS: QWEN METRIC SENSITIVITY

Same five query units · Same model · Only the similarity metric changes

TF-IDF ρ

-0.70



Semantic ρ

+0.60

What this shows

Lexical and semantic measures are both defensible ways to summarize textual divergence.

They can imply opposite relationships across the same query set.

With $n = 5$, neither direction should be treated as strong evidence.

The practical lesson: the apparent relationship is not robust to operationalization.

WHAT THIS MEANS

01

Separate diagnostics

Persona-conditioned response divergence and scratchpad-output divergence should be treated as distinct behavioral measures, not interchangeable evidence.

02

No single self-report meter

A model that looks stable or unstable on one measure does not reliably predict the other.

03

Methodological value

Disagreement between plausible diagnostics is itself useful information for digital-minds measurement.

LIMITATIONS & NEXT STEPS

Limitations

n = 5 query units per model — estimates are unstable

Scratchpad is elicited text, not privileged access

Base-Like is a role prompt, not a real base model

Prompt conditions perturb framing, tone, and task interpretation together

TF-IDF and embeddings capture different aspects of similarity

Three model families only — limited generalization

Next steps

Expand number and diversity of query units

Preregister primary unit and similarity metrics

Test whether Qwen's metric sensitivity persists

Use actual base vs post-trained checkpoints where available

Add mechanistic grounding (concept injection, steering)

BOTTOM LINE

Persona sensitivity and scratchpad–output consistency did not move together in a stable way across models.

Treat them as separate behavioral diagnostics. Measurement choices can reverse the apparent relationship.

For digital-minds work, convergence between measures has to be tested, not assumed.