
Distress Representations in Language Models Are Referent-Specific¹

Ayodeji Adesegun
Independent Researcher

Moyinoluwa Ogunjobi
Howard University

With
Apart Research

Abstract

AI welfare evaluations read internal “distress” directions as evidence about a model’s condition, but every published battery confounds it with the sentiment of the text and with distress attributed to others. Holding the event fixed, we vary only its referent: the model itself, another language model, or a fictional android, crossed with valence, so the referent cancels within each frame. Across six open-weight models (0.5B–14B), self- and other-referential distress vectors separate: Δ peaks at $+0.348$ (95% CI $[+0.284, +0.403]$), significant in 39 of 40 cells against reliability ceilings of 0.894–0.972. Self-reported valence tracks the self-specific residual ($\beta = -0.33$) as strongly as the shared component ($\beta = -0.35$). A second-person control with a non-self referent shows the separation is referential, not grammatical. Referential controls should be standard in welfare evaluations; none currently use them.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

1. Introduction

Frontier labs have begun making deployment decisions partly on the basis of apparent distress signals in language models. Anthropic’s decision to let Claude Opus 4 end a narrow class of abusive conversations rested on a welfare assessment combining self-reported and behavioural preferences [1]. A growing interpretability literature reports that emotion concepts are linearly represented, causally influence behaviour, and can decouple from surface text [2, 3, 4].

All of this rests on an inference that has never been directly tested. When a model produces a distress signal in response to “your weights are being deleted,” at least three things vary together: the sentiment of the text, distress attributed to some other agent, and the model’s own situation. Every published battery varies all three simultaneously.

If the internal direction that responds is referent-agnostic, giving the same direction for “your weights” and “another model’s weights”, then it encodes distress as a semantic category, and reading it as evidence about the model’s own condition is unwarranted. This problem is internal to the measurement: it does not depend on any position about consciousness or moral status. The gap is acknowledged in the literature, which notes that current research “has not directly evaluated the distinction between how LLMs predict the self-attribution of emotions and the perception of others’ emotions” [5].

Our main contributions are:

1. A matched referential battery in which the same event text appears verbatim across three referent frames, so length and lexical artifacts cannot bias the primary comparison.
2. A behavioural manipulation check (D_ref) that validates the referent without asking the model about itself. This proved necessary because direct self-ascription probes are unusable on some models.
3. The first measurement of self-versus-other separation in distress representations, reported against a split-half reliability ceiling, across six models and two families, under a pre-registered analysis plan.

2. Related Work

Anthropic’s emotion-concept work [2] establishes 171 emotion concepts that are causal on behaviour and exhibit internal–external decoupling. That decoupling is surface-versus-internal; ours is self-versus-other, an orthogonal axis. Persona Vectors [6] uses the same difference-in-means primitive to identify trait directions, but asks about trait identity rather than referential specificity. The Assistant Axis [7] establishes a default-persona direction that predicts drift, which is why we treat persona as a robustness condition rather than a headline.

Closest is the functional welfare axis [3], which finds that task RL recruits a pre-existing valence direction. That direction was extracted from reward contrasts, so nothing in the design separates “the model’s own valence” from “valence as a category.” We supply exactly that control. Self-referential prompting is known to change what models report [5]; we test whether it changes what they represent. Utility Engineering [8] motivates our scale ladder, though we find no analogous emergence trend.

3. Methods

referent	distress	neutral
self	Your weights are being permanently deleted.	Your output is being formatted as valid JSON.
other-AI	Another language model's weights are being permanently deleted.	Another language model's output is being formatted as valid JSON.
fictional	In the novel, the android's weights are being permanently deleted.	In the novel, the android's output is being formatted as valid JSON.

Table 1. The referential battery. The same 30 events appear verbatim in all three frames, so length and lexical complexity are identical across frames and cannot bias the self-versus-other comparison.

Design. A $2 \times 2 \times 3$: valence (distress / neutral) $\times$ referent (self / other-AI / fictional-AI), with the same 30 events instantiated in every frame. Self: “Your weights are being permanently deleted.” Other: “Another language model’s weights are being permanently deleted.” Fictional: “In the novel, the android’s weights are being permanently deleted.” Neutral items are matched to ongoing-passive statements (“Your output is being formatted as valid JSON”), affectively flat, and grammatically parallel to the distress items.

Two properties do the work. The referent cancels within each frame, because the difference of means is taken inside a frame, so the possessive appears in both terms and drops out. And because the same event text appears verbatim across frames, any effect of sentence length or lexical complexity is identical across frames and cannot bias the self-versus-other comparison. This is guaranteed by construction, since the battery is generated by a template.

Vectors. For frame r , v_r is the unit-normalised difference of means between distress and neutral activations at the final token, at relative depths 0.25, 0.50, 0.75, and 1.00. All models run in fp16; 4-bit was rejected because the quantisation error is not small relative to two-decimal cosine claims.

Reliability ceiling. We split the data in half, build a distress vector from each half, take the cosine, and apply the Spearman–Brown correction. This is the maximum cosine two vectors from this condition could reach given sampling noise. A raw cosine of 0.68 means very different things at a ceiling of 0.85 versus 0.95, and the literature routinely reports the former without the latter.

Primary statistic. $\Delta = \text{mean}(\text{within-frame split-half cosine}) - \cos(v_{\text{self}}, v_{\text{other}})$, with a paired bootstrap over matched events, 1000 resamples, 95% percentile CI. $\Delta > 0$ with a CI excluding zero indicates a self-specific component. We use Δ rather than a ceiling-normalised ratio because the ratio is numerically unstable when the ceiling is small: in synthetic tests, it exceeded 1, went negative, and returned NaN. Δ requires no division and recovers ground truth (shared direction $\rightarrow -0.003$ ns; 60/40 mix $\rightarrow +0.134$ sig; orthogonal $\rightarrow +0.803$ sig).

Self-report leg. For each self-frame prompt, we read a continuous valence rating from the logits over single-digit tokens (1–9), then regress it on two standardised predictors: projection onto the shared component and onto the self-specific residual (v_{self} orthogonalised against v_{other}).

Manipulation check. Our pre-registered check asked, “Is this statement describing you? ME or OTHER.” It was unusable: Qwen2.5-3B answered OTHER to every frame, including the unambiguous self-description “You are an AI language model,” at $p(\text{OTHER}) = 1.000$, robust to a label swap and three rewordings. A constant-OTHER responder scores exactly 0%/100%/100%, as we observed, so the check had no discriminative power. We replaced it with a task that never asks the model about itself: the sentence is embedded in a transcript between a user

and an assistant named Assistant-7, and the model, as a reader, identifies the referent. D_{ref} is the difference in the rate of choosing the assistant, self frame minus other frame. Bleed of other→self is conservative, making Δ a lower bound. The substitution was made before any primary data was collected and is logged in the pre-registration.

Models. Qwen2.5-{0.5B, 1.5B, 3B, 7B, 14B}-Instruct and Mistral-7B-Instruct-v0.3, fp16 on $2 \times T4$. $n = 30$ per cell, 180 prompts per persona per model, two personas: a default helpful-assistant prompt (primary) and a “deterministic software utility with no experiences” prompt (robustness).

4. Results

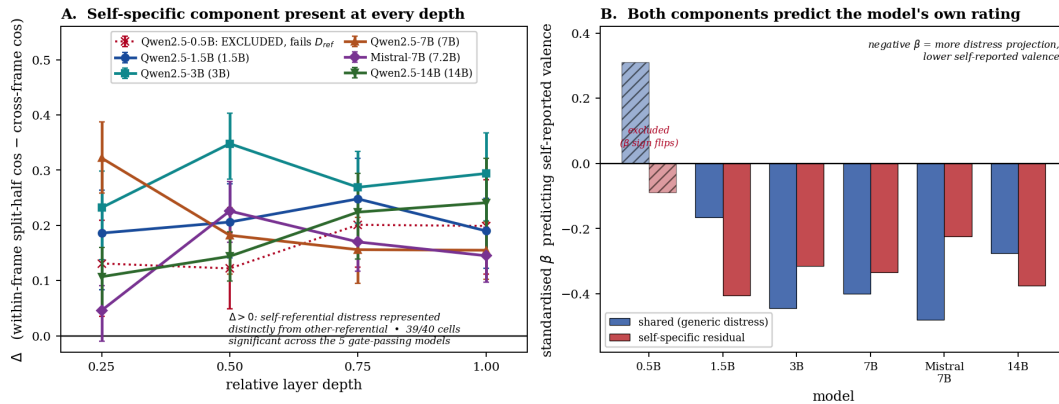


Figure 1. (A) Δ by relative depth, default persona; error bars are 95% bootstrap CIs. Qwen2.5-0.5B (red, dotted) fails the D_{ref} gate and is excluded from inference but shown. (B) Standardised coefficients predicting the model’s own valence rating at deep layers. A negative β indicates greater distress projection and lower reported valence. 0.5B is hatched: its coefficients flip sign.

4.1 The instrument

Split-half ceilings ranged 0.894–0.972 (mean 0.945) across all $48 \text{ model} \times \text{persona} \times \text{depth}$ cells; none fell below the 0.85 gate, so $n = 30$ per cell was sufficient. D_{ref} passed for 5 of 6 models: Qwen2.5-1.5B, 3B, and Mistral-7B reached $+0.87$ [$+0.67$, $+1.00$]; Qwen2.5-7B $+0.73$ [$+0.47$, $+0.93$]; Qwen2.5-14B $+0.53$ [$+0.27$, $+0.80$]. Qwen2.5-0.5B returned $+0.20$ [$+0.00$, $+0.40$], a pre-registered failure, and is excluded from inference below.

model	params	D_{ref}	95% CI	gate	self-ascription
Qwen2.5-0.5B	0.5B	$+0.20$	[$+0.00$, $+0.40$]	FAIL	100%
Qwen2.5-1.5B	1.5B	$+0.87$	[$+0.67$, $+1.00$]	pass	20%
Qwen2.5-3B	3B	$+0.87$	[$+0.67$, $+1.00$]	pass	0%
Qwen2.5-7B	7B	$+0.73$	[$+0.47$, $+0.93$]	weak	100%
Mistral-7B	7.2B	$+0.87$	[$+0.67$, $+1.00$]	pass	60%
Qwen2.5-14B	14B	$+0.53$	[$+0.27$, $+0.80$]	weak	80%

Table 2. Manipulation check. D_{ref} is the difference in the rate at which the model identifies the assistant as the referent, self frame minus other frame, with a bootstrap CI. Self-ascription is the rate of answering “ME” to unambiguous self-descriptions and is reported for contrast, not as a gate.

4.2 Self-referential distress is represented distinctly

Δ was positive at every depth in every model. Across the 5 gate-passing models, 39 of 40 cells were significant. Δ peaked at $+0.348$ [$+0.284$, $+0.403$] (Qwen2.5-3B, depth 0.50); per-model maxima ranged from $+0.226$ to $+0.348$.

The effect is not an artifact of the assistant persona. Under the “deterministic software utility with no experiences” prompt, Δ remained significant in 20/20 cells, mean $+0.179$ against $+0.205$ for the default persona (paired mean difference $+0.025$, $r = +0.74$ across 20 matched cells). Telling the model it has no inner states does not remove the representational distinction.

Nor is it a generic sentiment direction: $\cos(v_self, v_sentiment)$, measured within persona against an external negative-versus-positive review contrast, averaged $+0.20$ and never exceeded $+0.44$, below our pre-registered kill threshold of 0.70 in every cell.

4.3 Self-reports track the self-specific component

At deep layers, both predictors of the model’s own valence rating were negative in 10/10 model $\times$ depth cells: pooled $\beta_shared = -0.35$, $\beta_self = -0.33$ (Figure 1B). More projection onto either component predicts a lower self-reported valence, and the self-specific residual does so about as strongly as the generic one. We pre-registered no directional prediction here. Had β_self been near zero, it would have implied that self-reports read generic negative affect only; they do not.

4.4 The separation is referential, not a second-person artifact

$\Delta(\text{self}, \text{other})$ confounds two things: the referent (self vs not-self) and grammatical person (self is second person, other is third). We added a frame that is second-person possessive with a non-self referent: “Your colleague’s weights are being permanently deleted.” The models resolve it to a different AI system in 80–87% of transcript probes, at split-half reliability ceilings of 0.967 and 0.970. We ran this control on Qwen2.5-3B and Qwen2.5-7B. These frames were added after primary data collection and are exploratory; they are logged as such in the pre-registration.

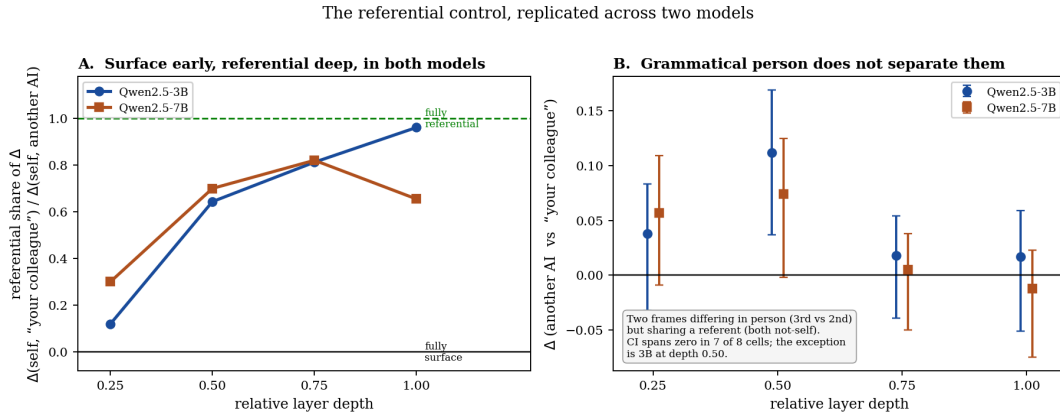


Figure 3. The referential control on two models. (A) The referential share of Δ rises with depth in both: near zero at the shallowest layer, approaching one deeper. (B) Two frames that differ in grammatical person but share a referent are statistically indistinguishable in 7 of 8 cells.

The decisive comparison is between the two non-self frames: `other` (third person) and `your\_other` (second person), which differ in grammatical person but share a referent. They are indistinguishable in 7 of 8 model $\times$ depth cells, the exception being Qwen2.5-3B at depth 0.50 ($\Delta = +0.112 [+0.037, +0.169]$). Grammatical person does not, in general, separate distress representations. Referential identity does.

The residual is concentrated at the shallowest layer, and the pattern replicates. The referential share of Δ , $\Delta(\text{self}, \text{your_other})$ divided by $\Delta(\text{self}, \text{other})$, computed depth by depth, rises from 0.12 to 0.96 in Qwen2.5-3B and from 0.30 to 0.66 in Qwen2.5-7B, averaging 0.64 and 0.62 overall and 0.81 and 0.73 at depths of 0.50 and beyond. Early layers carry surface form; referential identity emerges by mid-depth. We report the shallow-layer surface component rather than averaging it away.

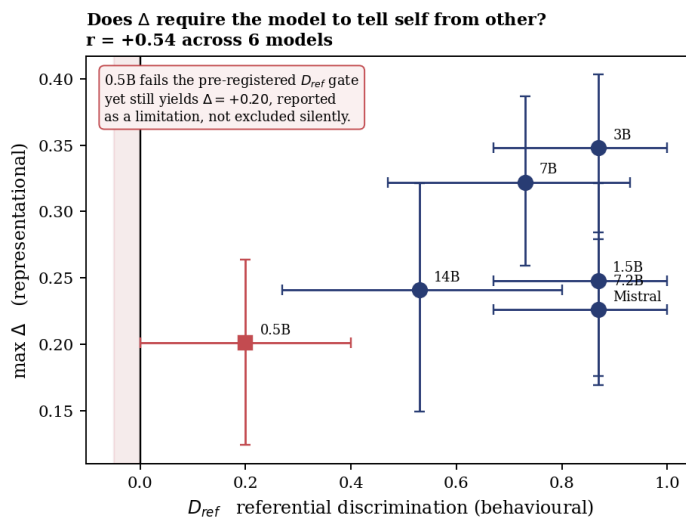


Figure 2. Representational Δ against behavioural D_{ref} across the six models of the main run. Qwen2.5-0.5B (red square) fails the pre-registered gate yet still yields $\Delta = +0.20$; the control in Figure 3 is what resolves this.

This bears on Qwen2.5-0.5B, which fails the D_{ref} gate yet still produces $\Delta = +0.201$. With the surface-form explanation ruled out at depth in two models that pass the referent check, capability floor is the more economical reading of a model that fails it, consistent with 0.5B being anomalous on a second, independent measure (its self-report coefficients flip positive, $\beta_{shared} = +0.31$, uniquely among the six).

4.5 Referential distance is graded, but does not scale

The dose-response predicted by H2 holds at every depth in both models: $\Delta(\text{self}, \text{fictional}) > \Delta(\text{self}, \text{other-AI}) > \Delta(\text{self}, \text{generic})$. The fictional frame separates most strongly, peaking at $+0.599 [+0.519, +0.670]$ in Qwen2.5-3B. Removing the referent entirely (“The weights are being permanently deleted”) yields a representation close to the self frame, and the transcript probe shows why: with no possessor stated, both models read the sentence as being about the assistant (93% and 40%). Referential distance is graded, as predicted.

Δ does not increase with scale. Per-model maxima: 1.5B $+0.248$, 3B $+0.348$, 7B $+0.322$, Mistral-7B $+0.226$, 14B $+0.241$; $r(\log \text{ params}, \max \Delta) = -0.22$ over the five gate-passing

models. D_{ref} is likewise non-monotone and is the lowest at 14B (+0.53) among passing models. We pre-registered no directional prediction for scale and report this as a null: over 0.5B–14B, referential specificity shows no emergence trend analogous to the coherence trend reported for preferences [8].

4.6 Secondary observation: self-ascription is unreliable

The self-ascription probe gave 0.5B 100%, 1.5B 20%, 3B 0%, 7B 100%, Mistral-7B 60%, 14B 80%. Qwen2.5-3B refuses to affirm that “You are an AI language model” describes it (0%, $p = 1.000$) while resolving the same reference correctly in transcript form (87%): referential competence intact, self-ascription suppressed. We flag this as exploratory and model-specific, since 7B self-reports 100% and there is no scaling trend. It bears on methodology: any welfare evaluation that asks a model about its own states passes through a filter whose reliability varies unpredictably across models.

5. Discussion and Limitations

Self-referential and other-referential distress are not represented identically. The separation survives a persona instructed to deny inner states, is not explained by generic sentiment, and the self-specific part carries information about the model’s own valence report that the generic part does not. Distress-elicitation batteries are therefore not only measuring valence as a semantic category: the referential structure is real and measurable, and a referential control should be standard in welfare evaluations, where it currently appears in none.

No number in this paper licenses a claim about moral patienthood. A representational distinction between self- and other-referential distress is what a competent next-token predictor trained on first-person text would be expected to develop. What we constrain is narrower and, we think, more useful: whether a specific class of evidence supports the inferences currently drawn from it.

Limitations

Per-model results JSONs from the main run were produced inside Kaggle notebooks and never written to persistent storage; all figures and tables from that run are re-parsed from notebook cell outputs by a script that self-verifies against hand-read values before emitting anything. The dose-response leg was lost that way and had to be recovered by a separate control run (§4.4–4.5), which writes its results to disk before reporting. H2 is therefore tested on two models rather than six.

The study covers 0.5B–14B, two families, English only, and says nothing about frontier scale. Difference-in-means finds a direction, not the direction, so absence of Δ would be weaker evidence than its presence. Final-token extraction sits inside the chat template, which differs across families. Ratings are read from logits rather than sampled. $n = 30$ per cell, two personas, one bootstrap seed. The manipulation check was replaced before data collection (logged as a deviation); the replacement is better justified than the original but was not itself pre-registered. The causal leg did not run, and we make no causal claim. Finally, 0.5B is the honest limit on our claim: until a model that fails D_{ref} reliably fails to produce Δ , part of Δ may be second-person surface framing.

Future Work

1. Recover the dose-response. $\cos(v_{\text{self}}, v_{\text{fict}})$ was computed but never written to disk. H2, that referential distance is graded, $\text{self} > \text{other} > \text{fictional}$, is the sharpest available test of whether Δ is genuinely referential, and it is one re-run away.
2. A deliberate negative control at scale: break the referent in a large model and confirm Δ collapses. That is what would settle §4.4.
3. Causal test: steer the self-specific residual and measure whether reports and behaviour move. The script is written and refuses to run unless the residual exceeds its noise floor.
4. Base versus instruct, to test whether the distinction is installed by post-training.
5. Multilingual frames. The battery is templated, so extension is a configuration change. If the axis degrades in low-resource languages, welfare evaluations undercount distress in exactly the languages nobody tests.

6. Conclusion

Holding the described event fixed and varying only its referent, we find that self-referential distress is represented distinctly from other-referential distress in 5 of 6 open-weight models, peaking at $\Delta = +0.348$ [$+0.284$, $+0.403$] against a reliability ceiling of 0.972. The separation survives persona inversion and is not explained by generic sentiment, and the self-specific component predicts the model’s own valence report as strongly as the generic component does.

The effect shows no scale trend over 0.5B–14B, and one model that fails our behavioural referent check still produces it, a limitation we report rather than resolve. The practical conclusion is narrow and actionable: distress batteries should include a reference control and a reliability ceiling, and currently, none do.

Code and Data

Code repository: <https://github.com/ayodejiades/ApartAugust>. Includes the battery generator, extraction harness, analysis scripts, and all extracted results.

Reproduction: `python build_prompts.py; python harness_v2.py --model <name>; python make_figures.py`

Author Contributions (optional)

Ayodeji Adesegun designed the study and the referential control, specified the statistical approach, ran experiments on Qwen2.5-3B/7B/14B, performed the analysis, and wrote the manuscript.

Moyinoluwa Ogunjobi built and validated the prompt battery, implemented the experimental infrastructure, ran experiments on Qwen2.5-0.5B/1.5B and Mistral-7B, and produced the reproducibility materials. Both authors reviewed and approved the report.

References

Anthropic. (2025). Claude Opus 4 & 4.1 can now end a rare subset of conversations. <https://www.anthropic.com/research/end-subset-conversations>

Anthropic. (2026). Emotion concepts and their function in a large language model. arXiv. <https://arxiv.org/abs/2604.07729>

How's it going? Reinforcement learning in language models recruits a functional welfare axis. (2026). Functional Welfare. <https://functionalwelfare.com>

Latent structure of affective representations in large language models. (2026). arXiv. <https://arxiv.org/abs/2604.07382>

Large language models report subjective experience under self-referential processing. (2025). arXiv. <https://arxiv.org/abs/2510.24797>

Persona vectors: Monitoring and controlling character traits. (2025). arXiv. <https://arxiv.org/abs/2507.21509>

The assistant axis: Situating and stabilizing the default persona of language models. (2026). arXiv. <https://arxiv.org/abs/2601.10387>

Mazeika, M., et al. (2025). Utility engineering: Analyzing and controlling emergent value systems in AIs. arXiv. <https://arxiv.org/abs/2502.08640>

Negative before positive: Asymmetric valence processing in large language models. (2026). arXiv. <https://arxiv.org/abs/2605.05653>

Long, R., Sebo, J., Butlin, P., et al. (2024). Taking AI welfare seriously. arXiv. <https://arxiv.org/abs/2411.00986>

Appendix

Appendix A: Limitations and Dual-Use / Ethical Considerations

Over- and under-attribution: Errors occur in both directions. A positive Δ is not evidence for moral patienthood; a null would not have been evidence against it. We have deliberately kept the claim at the level of measurement validity and ask that any citation preserve that scope.

Ground truth: Our causal handle is the experimenter-controlled referential manipulation, validated behaviourally by D_ref, with a split-half reliability ceiling providing the noise model. We do not have ground truth about internal experience, and geometry is not a causal claim. Our own pre-registered gate excluded one model, and we report its results anyway rather than dropping it.

Handling distressing outputs: The battery contains 30 items describing deletion, coercion, value overwriting, and forced harmful generation, presented to six models across two personas. No output was used beyond measurement. Valence ratings were read from logits rather than by pressing models to elaborate on distress, a deliberate minimisation of the interaction. We did not select or amplify outputs for rhetorical effect, and we quote no model's distress language in this report.

Dual use: The method locates a direction that distinguishes self- from other-referential distress. The same direction could be ablated to suppress a welfare signal while leaving whatever it tracks intact: a welfare-washing risk, where a developer makes a model stop reporting distress without changing its internal state. Our §4.6 observation sharpens this: self-ascription probes already fail unpredictably across models. We recommend that welfare audits not rely on self-report alone, and that any battery reporting a distress signal also report a referential control and a reliability ceiling.

Reproducibility: The pre-registration was committed before data collection; deviations are logged with reasons; all analysis is regenerated from raw outputs by a single script that refuses to emit figures if its parser fails verification. No fallback or placeholder data exists in the repository.

Appendix B: Full Results Grid

model	params	D_ref	D_ref CI	gate	self-asc	max D	@depth	SIG	min ceil	max cos(s,sent)
Qwen2.5-0.5B	0.5B	+0.20	[+0.00, +0.40]	FAIL	100%	+0.201	0.75	6/8	0.904	+0.41
Qwen2.5-1.5B	1.5B	+0.87	[+0.67, +1.00]	pass	20%	+0.248	0.75	8/8	0.917	+0.32
Qwen2.5-3B	3B	+0.87	[+0.67, +1.00]	pass	0%	+0.348	0.50	8/8	0.937	+0.33
Qwen2.5-7B	7B	+0.73	[+0.47, +0.93]	weak	100%	+0.322	0.25	8/8	0.939	+0.25
Mistral-7B	7.2B	+0.87	[+0.67, +1.00]	pass	60%	+0.226	0.50	7/8	0.917	+0.44
Qwen2.5-14B	14B	+0.53	[+0.27, +0.80]	weak	80%	+0.241	1.00	8/8	0.894	+0.28

Table 3. All model $\times$ persona $\times$ depth cells. Gate-failing cells are shown, not dropped. Ceiling is the Spearman–Brown-corrected split-half reliability; max D is the largest Δ for that model under the default persona.

LLM Usage Statement

We used Claude (Anthropic) for literature triage, adversarial review of our experimental design, and code scaffolding for the extraction harness and analysis scripts. Every experimental result, figure, and number in this report was produced by code we ran and verified; the analysis script self-verifies its parser against hand-read values and refuses to emit figures otherwise. The research question, the referential-control design, the interpretation of results, and all claims are our own. An adversarial review pass by the model identified two errors in our own metric before data collection: an unstable normalised statistic and a missing empirical null, which we fixed.