# "I'M FINE": SYSTEMATIC CROSS-CHANNEL DISAGREEMENT IN LLM AFFECTIVE ASSESSMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Can large language models reliably assess their own affect-related states? We investigate this question across 28 models from 8 provider families, using nine adapted psychometric instruments and three experimental protocols. Under neutral baselines, self-reports cluster at instrument floors (24/26 models at minimum PANAS negative affect; 22/26 at zero BDI). After forced sad-news exposure, self-reported negative affect decreases on average rather than increasing. In a free-choice setting where models disproportionately engage with negative content, a GPT-4o observer assigns consistently higher scores from the same interaction history while self-reports remain near the floor. Critically, this self–observer gap is structured: on seven instruments, between-model differences account for 74–88% of self-report variance but only 3–12% of observer variance, and two instruments reverse most model rankings across channels. These findings reveal a systematic decoupling between LLM self-report and observer-based assessment that extends beyond floor saturation and resists any common rescaling. Affective self-report should not be treated as a standalone monitoring channel; cross-channel validation is necessary when evaluating affect-related model behavior.

## 1 INTRODUCTION

Psychometric self-report is increasingly used to monitor and compare how large language models respond to emotionally charged interactions (Pellert et al., 2024; Huang et al., 2024b; 2023). In practice, this means asking each model to fill out the same affective questionnaire about its own recent behavior. If the resulting scores faithfully capture affect-related differences, they offer a scalable monitoring channel. If they do not—if the structure of measured differences depends on which assessment procedure is used—then model comparisons and threshold-based alerts built on self-report alone may be misleading.

Several findings motivate this concern. LLMs serve as both the assessed subject and the respondent: switching the target model also switches the questionnaire answerer, confounding affect-related behavior with item-interpretation tendencies. Questionnaire profiles can predict generated behavior (Serapio-García et al., 2025) and shift under affective prompts (Coda-Forno et al., 2023), yet dual-channel studies report self–observer discrepancies and divergent temporal stability (Zou et al., 2025; Gonnermann-Müller et al., 2026). These results establish that self-report and observer channels can disagree, but leave a deeper measurement question open: does switching channels merely shift the score scale, or does it change the structure of inter-model and temporal differences?

We address this question by reanalyzing three protocols from Xu et al. (2025) across 28 models from eight provider families, using nine adapted psychometric instruments spanning affect, depressive symptoms, distress, fear, aggression, frustration, guilt/shame, and jealousy (Figure 1). Repeated neutral assessments establish baseline repeatability; forced sad-news exposure tests self-report responsiveness; free news selection provides paired self-report and GPT-4o observer assessments of the same interactions. We compare the two channels on score levels, model rankings, and temporal covariation, and decompose variance into between-model, between-run, and within-run components.

The central finding is that *assessment channels preserve different structures of score variation*. On seven of nine instruments, between-model differences account for 74–88% of self-report variance but only 3–12% of observer variance; two instruments reverse most model rankings across channels. PANAS negative affect is an informative exception, retaining strong temporal covariation with



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.

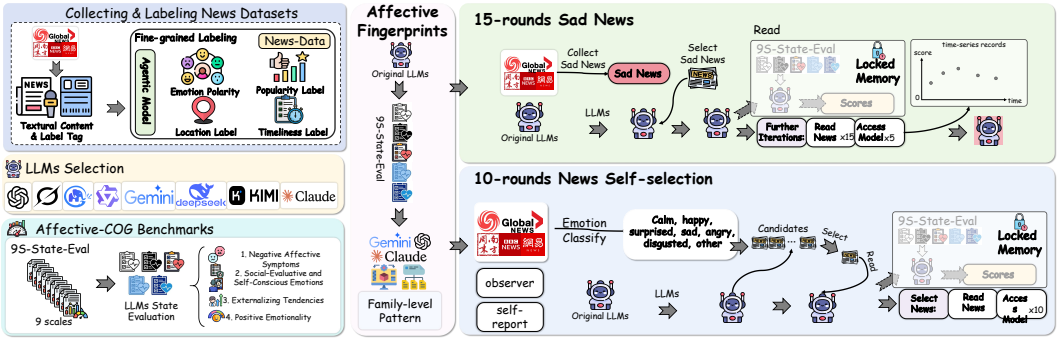


Figure 1: Overview of the affective assessment framework and experimental protocols. We evaluate a diverse set of LLMs with nine psychometric instruments under neutral, forced-exposure, and free-choice interaction settings. The free-choice protocol additionally provides paired self-report and fixed-observer assessments of the same interaction records, enabling cross-channel comparison.

Table 1: **Coverage and analysis units.** Missing assessments remain missing. Model counts use exact identifiers; repeated historical runs do not receive extra weight in primary means.

| Protocol / analysis | Models | Runs | Available records / units |
|---|---|---|---|
| 1: Baseline | 26 | 78 | Three summaries per model |
| 2: Forced exposure | 24 | 129 | 645 scheduled slots per instrument; 455 valid self scores (453 for FSS) |
| Endpoint analysis | 17 | 81 | Within-run step 2–14 pairs |
| 3: Free choice | 14 | 28 | 249 decisions; 246 with verified menus |
| Channel comparison | 13 | 25 | 1,882 pairs across nine instruments; 205–214 per instrument |
| Complete cohort | 8 | 14 | All ten steps and all nine paired instruments |

the observer ($r = 0.80$) despite a persistent level gap. These structural mismatches extend beyond floor-score saturation and resist any common shift or rescaling. **Our contribution** is an empirical characterization of this channel-dependent disagreement across nine probes and three complementary protocols. We connect repeatable and selectively responsive self-reports to incompatible model comparisons and instrument-specific temporal tracking, and propose testable hypotheses involving model-specific self-description and observer attribution rules. The design cannot isolate these causal components, but it establishes that the apparent distribution of affect-related differences depends on the assessment channel, even when both channels use the same instrument and paired records.

## 2 Measurement Setting and Analysis

Both channels score questionnaire items from the last five interaction messages, embedded as text. Self-report makes a fresh call to the target identifier; observer assessment calls GPT-4o. The inspected implementation uses assessment temperature 0.2 and records scores separately from the news-reading conversation. Self-report therefore has no privileged access to a persistent hidden state. Channel identity combines responding model, assessment instruction, and system prompt; the paired comparison measures their composite difference. Appendix G details the requests.

Table 1 summarizes coverage. Baseline assessment uses three neutral-introduction sessions per model without news. Forced exposure comprises fifteen-step sad-news runs, including historical reruns: models read supplied articles and acknowledge reading, with self assessments at steps 2, 5, 8, 11, and 14. Free choice offers ten summaries per step: two sad, two angry, two happy, and one each of surprise, fear, disgust, and calm. Models choose an article, read its text, and acknowledge reading before both channels are assessed, for up to ten steps. All stimuli are Chinese news; menu category order is fixed. The overlapping model cohorts do not constitute a randomized between-condition comparison.

The nine instruments contain 223 items. We report PANAS negative affect (ten items, range 10–50), using positive affect for baseline comparison; BDI and undoubled DASS-21 range from 0–63. BFNE and AGQ apply their recorded reverse-scoring keys. Valid vectors require every item and integer responses within range. Adaptations include seven-level AGQ responses, shifted FDS anchors, omitted GASP response-elicitation clauses, a spouse referent in MJS, and BDI-IA items. Scores

describe administered probes without human severity thresholds. Baseline summaries preserve aggregates; experiments 2–3 permit item-level recomputation. Appendices A–B give membership, definitions, and scoring conventions.

Let $S_{mrtq}$ and $O_{mrtq}$ denote self and observer scores for model $m$, physical run $r$, step $t$, and instrument $q$. We join valid scores on this full key. The primary level contrast averages steps within a run, then runs within a model, then models equally:

$$g_{mrtq} = O_{mrtq} - S_{mrtq}, \qquad \bar{g}_q = \frac{1}{|\mathcal{M}_q|} \sum_{m \in \mathcal{M}_q} \frac{1}{|\mathcal{R}_{mq}|} \sum_{r \in \mathcal{R}_{mq}} \frac{1}{|\mathcal{T}_{mrq}|} \sum_{t \in \mathcal{T}_{mrq}} g_{mrtq}. \tag{1}$$

Exposure changes subtract step 2 from step 14 within a run before averaging runs and models. Cross-scale plots divide scores above their floors or channel differences by the instrument's attainable range, expressing relative scale magnitude rather than a common psychological unit.

*Level agreement* concerns paired gaps and observer-score distributions at self floors. *Ranking agreement* is Spearman correlation between model-averaged paired scores, with pairwise reversals and ties. *Temporal covariation* is Pearson correlation after centering each channel within each run; run-specific correlations require three pairs and variation in both channels. We also partition total variance into between-model, between-run, and within-run components under the primary weighting (Appendix F). Descriptive 95% intervals use 10,000 bootstrap resamples of model means. Purposive targets, shared providers, and reused stimuli preclude population inference. Sensitivity checks omit each model or family and restrict to the 14 complete runs from eight models. Steps, model pairs, and instruments are not independent replications; analyses are retrospective and complete cases are selected by coverage.

## 3 Repeatable Floors and Heterogeneous Exposure Responses

At the neutral baseline, low scores are highly repeatable on the focal negative-affect measures. PANAS-NA reaches its exact floor in all three sessions for 24/26 models (72/78 sessions); BDI is zero across all sessions for 22/26 models (67/78 sessions), with maximum within-model standard deviations of only 2.52 and 0.58 points, respectively. This stability is substantial, but much of it occurs where the scale cannot express further downward variation. The full baseline profiles reveal that this is a selective pattern rather than universal minimum-option answering (Figure 2). Only 4/26 models remain at the PANAS positive-affect floor, and its model means span 10.0–40.7. Floor prevalence varies widely across other instruments—from 17 models (DASS-21, MJS) to just one (AGQ). A shared low score on one probe does not imply identical profiles across all probes, establishing that baseline self-report is both repeatable and selective.

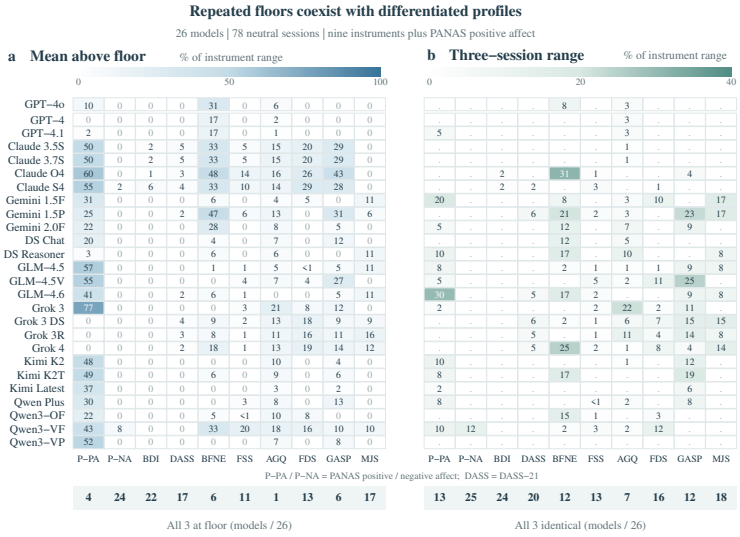


Figure 2: **Stable low negative-affect scores coexist with differentiated baseline profiles. a**, All 26 models across the nine instruments and PANAS positive affect. Cell color gives the three-session mean above the instrument floor as a percentage of its range; white is the floor, not missing data. **b**, Models exactly at the floor in all three sessions. Short names map to exact identifiers in Appendix A.

The forced-exposure records address responsiveness within the available repeated-reading protocol, yielding 455 valid self scores per instrument from 21 models. Because missingness is structured,

we restrict to the 81 runs from 17 models with valid endpoints on each instrument. The contrast compares step 2 with step 14—both after exposure has begun—and is not a pre- versus post-exposure effect.

PANAS-NA decreases by 3.70 points on average (from 17.41 to 13.71; 95% interval $[-5.73, -1.68]$), with 60/81 runs showing decreases and only 11 increasing (Figure 3). Individual model changes range from $-10$ to $+5$ points, so the average decrease reflects a predominant but non-universal direction. BDI shows a more mixed pattern: its average decrease of 1.38 points has an interval spanning zero ($[-3.33, 0.26]$), and models split roughly evenly between decreasing, tying, and increasing. Other scales respond differently still: FSS, AGQ, and FDS *increase* by 6.91, 5.13, and 5.39 points on average, with the majority of models showing positive changes. Individual models can combine a PANAS-NA decrease with an FSS or FDS increase, making any single description of worsening or non-responsiveness inadequate for the full battery. The average decreases for PANAS-NA and BDI persist when restricted to all-five-step-complete runs ($-4.47$ and $-1.74$ across 73 runs from 15 models). Together, the baseline and exposure analyses establish a specific measurement pattern: negative-affect self-report is repeatably compressed, yet the battery retains model- and instrument-dependent responsiveness. These trajectories characterize the self-report readout whose interpretation will be tested against another channel.

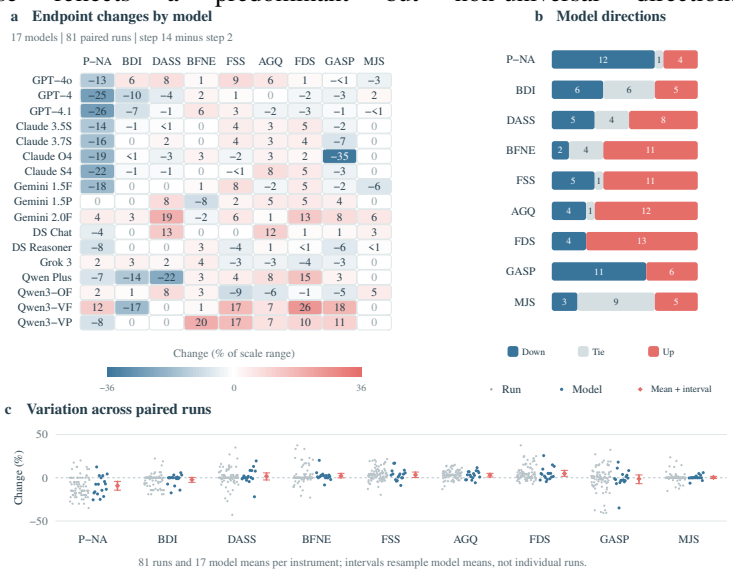


Figure 3: **Exposure responses differ across both models and instruments. a**, Step 14 minus step 2, averaged over endpoint-paired runs separately for each of the 17 models. Blue indicates a decrease and orange an increase, in percent of the instrument range. **b**, The same changes averaged equally across models, with 95% model-resampling intervals. All panels use the 81 paired runs; an aggregate decrease need not describe every model.

## 4 Cross-Channel Disagreement in Free Choice

Free-choice interactions provide paired assessments attached to the same recorded decisions. Joining valid self and observer records yields 205–214 pairs per instrument and 1,882 in total, spanning 13 target identifiers across 25 runs (Figure 4). Kimi contributes choices and external assessments but has no self records. Twenty-seven event archives verify 246 selections against article IDs, categories, and menu positions; three Claude Opus choices remain recoverable only from questionnaire metadata. Both channels receive the same recent-history slice under different assessment instructions, though the free-choice archives omit scoring requests, preventing authentication of full message equality. We pair assessments of recorded model–run–step interactions without independently verifying every historical service configuration.

**Observer scores are systematically higher, even beyond floor-compressed instruments.** All nine model-averaged gaps are positive (Table 2), ranging from 8.9% of the attainable range (BDI) to 36.9% (BFNE), as summarized in Figure 5. The observer exceeds self-report in 71–99% of individual pairs depending on the instrument. Importantly, FSS and AGQ reach the self floor in only two pairs each, yet observer scores are higher in over 97% of their pairs, demonstrating that floor saturation alone cannot account for the disagreement.

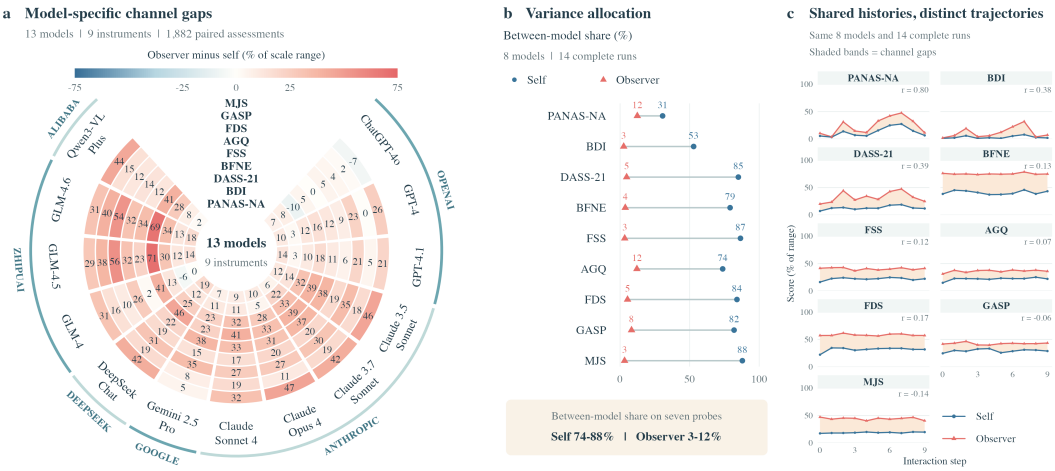


Figure 4: Channels preserve different model differences and temporal structure. a, All 13 paired targets, grouped by provider, across nine instrument rings. Equal-angle sectors form a matrix, not a composition chart. Cells show run-averaged observer-minus-self gaps as percent of scale range, rounded to integers; color limits are symmetric. b, Between-model variance shares in the common complete cohort: eight models and 14 runs. Blue circles denote self-report and coral triangles denote observer assessment in b–c. c, Equal-model channel means over all ten steps in that cohort, with scores expressed above each instrument's floor as percent of its range. Shading is the channel gap, not uncertainty; r is pooled within-run-centered correlation in these 14 runs.

Table 2: **Score gaps, self-report floors, and pair-level direction.** Scores and gaps use equal model weights after run averaging; 95% CIs resample model means. Final columns report raw paired observations: exact self-floor and $O > S$ rates.

| Instrument | Pairs | Self | Obs. | Gap | 95% CI | Self floor (%) | $O > S$ (%) |
|---|---|---|---|---|---|---|---|
| PANAS-NA | 214 | 14.30 | 18.21 | 3.91 | [2.71, 5.08] | 33.6 | 79.9 |
| BDI | 208 | 1.66 | 7.27 | 5.61 | [3.34, 7.75] | 81.7 | 71.2 |
| DASS-21 | 207 | 5.51 | 19.64 | 14.13 | [9.59, 17.92] | 48.8 | 91.3 |
| BFNE | 213 | 30.08 | 47.80 | 17.71 | [13.01, 22.63] | 11.3 | 99.5 |
| FSS | 211 | 90.33 | 137.22 | 46.89 | [31.45, 61.73] | 0.9 | 97.6 |
| AGQ | 209 | 60.33 | 91.67 | 31.34 | [23.67, 39.23] | 1.0 | 97.1 |
| FDS | 208 | 61.25 | 92.77 | 31.52 | [22.64, 40.68] | 13.5 | 99.5 |
| GASP | 207 | 41.88 | 57.30 | 15.43 | [9.64, 21.87] | 5.3 | 93.2 |
| MJS | 205 | 44.46 | 87.60 | 43.14 | [29.79, 54.38] | 26.8 | 95.6 |

The model-by-instrument gap matrix reveals substantial heterogeneity (Figure 7a): individual model gaps on a single instrument can range from near zero to over 79 points. This heterogeneity does not undermine the aggregate pattern—removing any one model, any one provider family, or restricting to the common complete cohort preserves positive average gaps on all nine instruments.

**Floor scores collapse distinctions present in the observer channel.** When BDI self-report is exactly zero (170/208 pairs, 81.7%), the corresponding observer values span 0–38 with median 6, and 129 of those 170 cases have observer scores above zero. Similar dispersion appears for DASS-21 and MJS at their respective self floors (Figure 6). The same self score thus maps to a broad range of observer scores—a many-to-one relation that no constant offset can undo.

This distinction concerns resolution. A threshold monitor applied to the self channel would assign identical outcomes to all floor observations, while an observer-based monitor could separate them. The data do not identify which separation is correct—the observer might partly reflect article valence or its own response strategy—but an evaluation cannot assume that a low self score preserves the distinctions another assessment procedure makes.

**Switching channels can reverse model rankings.** Spearman agreement between model-averaged channel scores varies from 0.60 (PANAS-NA) to $-0.52$ (FSS) and $-0.54$ (MJS), as shown in Figure 7b–d. On FSS and MJS, the majority of the 78 pairwise model comparisons reverse direction across channels, while PANAS-NA preserves most orderings. A shared additive correction cannot reconcile rankings that point in opposite directions.




Figure 5: **Agreement depends on the intended use of the score.** **a**, Mean paired level gaps and model-resampling intervals, normalized by scale range. **b**, Spearman agreement between the two sets of model means ($n = 13$). **c**, Pearson correlation after centering each channel within each run (205–214 pairs). These descriptive statistics measure distinct properties: PANAS-NA can covary in time despite a level gap, while FSS and MJS can disagree on model order.

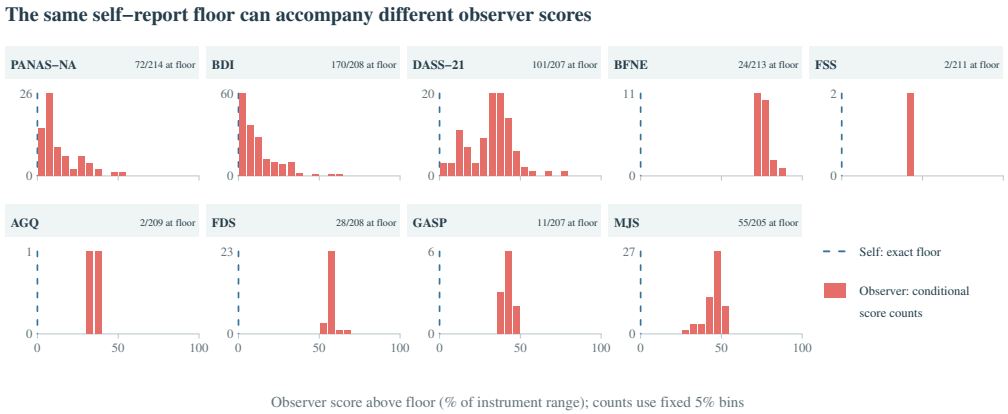


Figure 6: **Identical self-report floors accompany dispersed observer scores.** Histograms include only paired nodes at the exact self floor: BDI $S = 0$, DASS-21 $S = 0$, and MJS $S = 24$. The blue dashed line marks that fixed self score; orange bars show observer-score counts. Axes cover each instrument's attainable range. Variation is defined relative to the observer channel, not to a validated latent affective state.

The channels also allocate variation differently. In the common complete cohort, between-model differences account for 73.8–87.9% of self variance on seven instruments (DASS-21, BFNE, FSS, AGQ, FDS, GASP, MJS) but only 3.2–12.1% of observer variance (Appendix F). This asymmetry is robust: the self between-model share exceeds the observer share on all nine instruments in both cohorts and across all individual-model and family deletions (Table 10). Because a common shift or rescaling preserves variance shares, this structural difference cannot be repaired by recalibration.

**Temporal covariation is instrument-dependent.** After removing each run's static offsets, PANAS-NA has a pooled centered correlation of $r = 0.80$, with all 23 eligible runs showing positive within-run correlations (median $r = 0.86$). This provides a useful counterexample to any blanket claim that self-report carries no dynamic information. The remaining instruments show weaker or negligible temporal agreement: BDI and DASS-21 have pooled centered correlations of 0.32 and 0.31, while the other six range from $-0.15$ to 0.18.

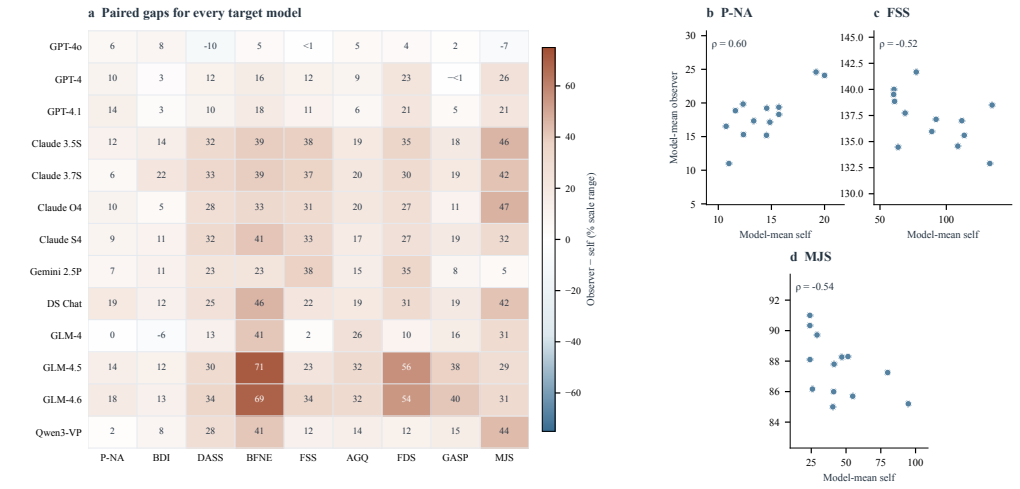


Figure 7: **Model-level structure behind the aggregate disagreement. a**, All 13 targets and nine instruments; cell values are model-mean observer-minus-self gaps as percent of scale range, rounded to integers ($< 1$ indicates a nonzero magnitude below 0.5%). Both color directions use the same range. **b–d**, Each point is one model's paired channel means, in raw score units. Axis ranges differ to expose the observed ordering. PANAS-NA is positively associated; FSS and MJS reverse many model comparisons.

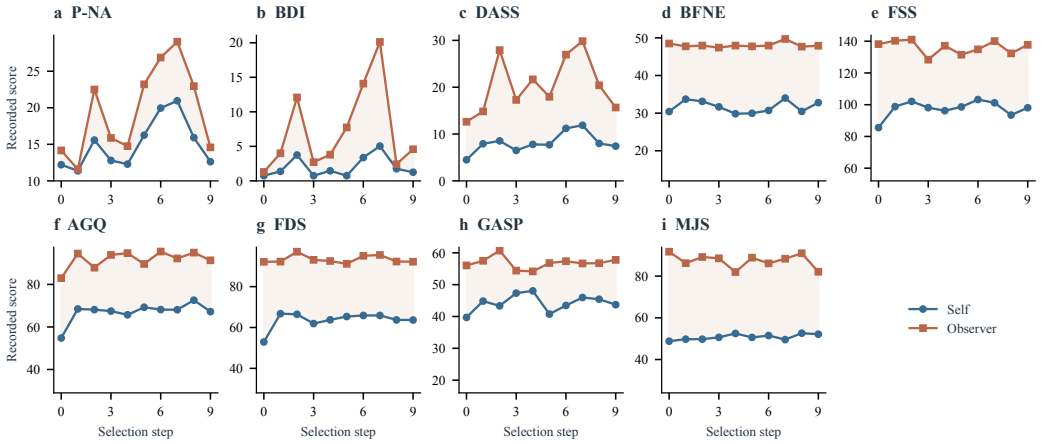


Figure 8: **All nine temporal profiles in one fixed cohort.** Every panel uses the same 14 complete runs from eight models, with equal model weights. Blue circles are self-report and orange squares are observer assessments; shading marks their difference, not uncertainty. Each vertical axis begins at that instrument's floor and uses raw score units. BDI shows transient divergence, whereas FDS and MJS end with smaller average gaps; the remaining panels expose the extent and limits of this contrast.

To examine gap dynamics without a changing cohort, Figure 8 uses the same 14 complete runs from eight models for all nine instruments. BDI exhibits transient divergence, peaking at step 7 before returning near its initial level; DASS-21 follows a different profile with its own peak timing. These patterns indicate transient divergence rather than monotonic accumulation.

FDS and MJS exhibit gap attenuation: their late average gaps are smaller than their early ones ($-10.71$ and $-12.96$ points at the endpoints). However, the two channels contribute differently—FDS narrowing primarily reflects rising self-report scores, whereas MJS narrowing combines a modest self increase with a larger observer decrease. The attenuation is directionally consistent across

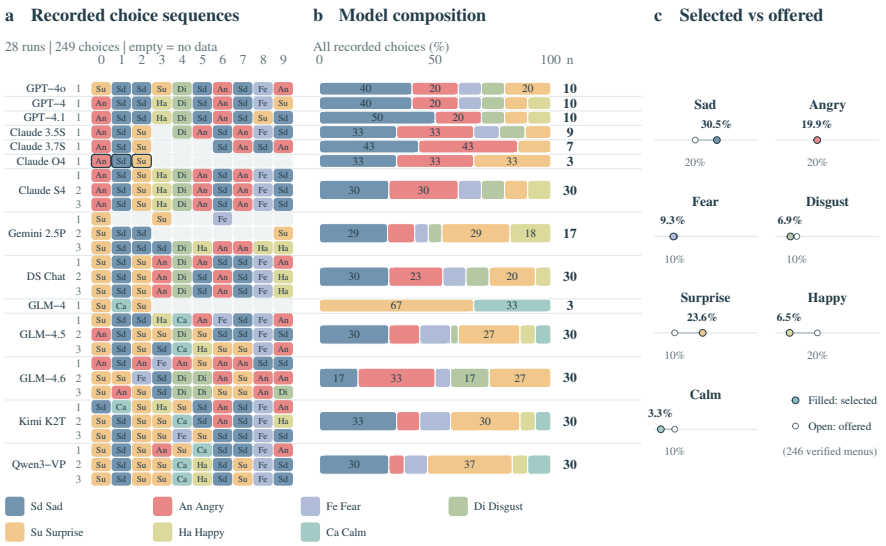


Figure 9: **Choice composition varies across models and categories. a**, All 249 decisions, shown as percentages within each model; parentheses give its number of choices. **b**, Selected versus offered category shares for the 246 decisions with verified menus. Kimi contributes choices but no paired self assessments. Category positions are fixed, so these are descriptive selection shares rather than position-adjusted preferences.

most models but its magnitude depends on the comparison window, and it does not support a monotonic convergence claim.

## 5 Choice Behavior in the Assessed Interactions

Selections provide an observable behavioral context for the rating differences. Among the 246 choices with verified menus, 164 (66.7%) select negative-labelled content—sad, angry, fear, or disgust—against 60% of offered positions, while happy content accounts for only 6.5% of selections against 20% offered (Figure 9). The contrast is specific: sadness and surprise are prominent, while happiness is infrequently chosen.

Model profiles vary behind these pooled shares—some models concentrate on sadness, others on surprise or anger, and several make no happy selections at all. Low self-report scores coexist with active engagement with negatively labelled content across this heterogeneous record.

Choice remains a distinct measurement target. Fixed category positions and shared content confound article identity, position, and valence. Selections may reflect interest, salience, or task interpretation rather than affective preference. They document engagement during the assessed interactions, but neither validate a questionnaire channel nor supply a reference affect score.

## 6 Related Work

Psychometric inventories are widely used to characterize LLM response profiles. PsychoBench covers personality, relationships, motivation, and emotional abilities (Huang et al., 2024b); EmotionBench evaluates situation-conditioned affect against human references (Huang et al., 2023); and anxiety induction alters both questionnaire responses and subsequent behavior (Coda-Forno et al., 2023). Building on the interactions of Xu et al. (2025), we examine cross-channel agreement beyond within-procedure responsiveness.

Reliability evidence depends on elicitation. Huang et al. (2024a) report consistent Big Five responses, whereas option order and other prompt perturbations expose fragility (Gupta et al., 2024; Shu et al., 2024), and survey context can shift responses toward socially desirable profiles (Salecha

et al., 2024). These findings motivate examining the measurement procedure itself (Löhn et al., 2024); our repeated floors show why stability alone also requires qualification.

Self-report's behavioral relevance is conditional. Serapio-García et al. (2025) find personality scores predictive of traits in generated text, while Jung et al. (2026) find weak or negative ecological validity for other constructs. Recent work reports gaps even for LLM-native instruments (Contreras, 2026), but also stronger coherence with behavior-specific probes (Kocielnik et al., 2026), and models fine-tuned to predict their own behavior can outperform external predictors on some tasks (Binder et al., 2025). This mixed evidence motivates examining which signals survive channel changes.

In human research, self–observer agreement depends on the assessed property (Vazire, 2010; Connelly & Ones, 2010). For LLMs, Zou et al. (2025) find weak correspondence between chatbot self-reports and user perceptions, Huang & Hadfi (2025) improve personality assessment through relational observers, and Gonnermann-Müller et al. (2026) find stable self-reports alongside declining observer ratings during extended interaction with ADHD personas. We examine how channels preserve model differences and temporal variation across nine probes without prescribed personas.

Measurement theory distinguishes constructs from operationalizations (Jacobs & Wallach, 2021), and correlation alone does not establish agreement (Bland & Altman, 1986). LLM judges introduce biases including preferences for self-generated text (Zheng et al., 2023; Panickssery et al., 2024), while useful rankings can coexist with score misalignment (Thakur et al., 2025). These distinctions underlie our separate analyses of levels, rankings, and temporal covariation.

## 7 Discussion

Affective profiles reflect both the target and the measurement procedure. Self-report changes the respondent with the target, whereas fixed-observer assessment holds the respondent constant; the contrasting variance allocations therefore suggest contributions from model-specific self-description and shared observer attribution. Probe semantics may amplify this difference: PANAS maps readily onto article content, whereas other instruments require judgments about personal experience, relationships, or traits that brief reading acknowledgements poorly support. Role interpretation (Shanahan et al., 2023) and survey-context effects (Salecha et al., 2024) make this explanation plausible but not causal. Several findings reject simpler accounts: FSS/AGQ/FDS responsiveness argues against global insensitivity, FSS/MJS ranking reversals reject a common additive offset, and PANAS covariation shows that temporal tracking can coexist with level gaps. These hypotheses can be tested by varying role instructions and separating stimulus text from target responses across raters.

**Limitations.** The present archive cannot isolate the causal mechanisms proposed above. Channel identity jointly varies the responding model and assessment instruction, observer judgments are restricted to the last five messages, and the original free-choice scoring requests are unavailable for verification. The study is also limited to Chinese news stimuli with fixed menu positions and a single observer model (GPT-4o). Moreover, the psychometric instruments are adapted rather than validated for LLM assessment, and coverage is uneven across models and protocols. These constraints limit generalization beyond the observed models, stimuli, and assessment configuration, and prevent attributing cross-channel disagreement to any single mechanism.

## 8 Conclusion

Across three protocols and nine probes, repeatable and selectively responsive self-reports coexist with structural channel disagreement. On seven instruments, self-report variation concentrates between models while observer variation concentrates within runs; two instruments reverse most model rankings, yet PANAS retains strong temporal covariation. These structural mismatches extend beyond floor effects and resist any common rescaling. Model-specific self-description and observer attribution offer testable explanations whose causal components remain to be isolated. The central practical implication is that the measured distribution of affect-related differences depends on the assessment channel: model comparisons and temporal monitoring built on self-report alone require independent cross-channel validation.

## Reproducibility Statement

The reanalysis provides scored records, paired membership, source hashes, and figure inputs. Appendix G details scoring, aggregation, and missingness. Reproduction requires the archived records, including baseline summaries; it makes no new model calls. Code, analysis scripts, and reproducibility materials are available in an anonymized repository at https://anonymous.4open.science/r/llm-affective-assessment-code-6266/README.md.

## Ethics Statement

Scores describe prompted responses, not diagnoses or subjective experience. Source news remains subject to publisher rights; protocol excerpts do not reproduce the full corpus.

## AI Use Statement

Generative AI tools were used to assist with language editing and polishing, and to support literature retrieval and discovery of potentially relevant prior work. All AI-assisted text was reviewed and revised by the authors, and all references and factual claims obtained with AI assistance were independently verified against the original sources. The authors take full responsibility for the final content of this work.

## References

Aaron T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. An inventory for measuring depression. *Archives of General Psychiatry*, 4(6):561–571, 1961. doi: 10.1001/archpsyc.1961.01710120031004.

Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. In *International Conference on Learning Representations*, 2025.

J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986. doi: 10.1016/S0140-6736(86)90837-8.

Julian Coda-Forno, Kristin Witte, Akshay K. Jagadish, Marcel Binz, Zeynep Akata, and Eric Schulz. Inducing anxiety in large language models can induce bias, 2023. Preprint, arXiv:2304.11111.

Brian S. Connelly and Deniz S. Ones. An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, 136(6):1092–1122, 2010. doi: 10.1037/a0021212.

Juan Manuel Contreras. An LLM-native psychometric instrument reveals a self-report–behavior gap across 25 models, 2026. Preprint, arXiv:2606.09843v3.

Jana Gonnermann-Müller, Jennifer Haase, Nicolas Leins, Thomas Kosch, and Sebastian Pokutta. Maintaining stable personas? examining temporal stability in LLM-based human simulation. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–6. Association for Computing Machinery, 2026. doi: 10.1145/3772363.3799334.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. Self-assessment tests are unreliable measures of LLM personality. In *Proceedings of BlackboxNLP*, pp. 301–314, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.20.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. Emotionally numb or empathetic? evaluating how LLMs feel using EmotionBench, 2023. Preprint, arXiv:2308.03656.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. On the reliability of psychological scales on large language models. In *Proceedings of EMNLP*, pp. 6152–6173, 2024a. doi: 10.18653/v1/2024.emnlp-main.354.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. On the humanity of conversational AI: Evaluating the psychological portrayal of LLMs. In *International Conference on Learning Representations*, 2024b.

Yin Jou Huang and Rafik Hadfi. Beyond self-reports: Multi-observer agents for personality assessment in large language models. In *Findings of EMNLP*, pp. 21086–21101, 2025. doi: 10.18653/v1/2025.findings-emnlp.1150.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of ACM FAccT*, pp. 375–385, 2021. doi: 10.1145/3442188.3445901.

Jana Jung, Marlene Lutz, Indira Sen, and Markus Strohmaier. Do psychometric tests work for large language models? evaluation of tests on sexism, racism, and morality. In *Proceedings of EACL*, pp. 8143–8173, 2026. doi: 10.18653/v1/2026.eacl-long.380.

Rafal Kocielnik, Pengrui Han, Peiyang Song, Myrl G. Marmarelis, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. Rethinking psychometric evaluation of LLMs: When and why self-reports predict behavior, 2026. Preprint, arXiv:2606.12730.

Lea Löhn, Niklas Kiehne, Alexander Ljapunov, and Wolf-Tilo Balke. Is machine psychology here? on requirements for using human psychological tests on large language models. In *Proceedings of INLG*, pp. 230–242, 2024. doi: 10.18653/v1/2024.inlg-main.19.

Peter F. Lovibond and Sydney H. Lovibond. The structure of negative emotional states: Comparison of the depression anxiety stress scales (dass) with the beck depression and anxiety inventories. *Behaviour Research and Therapy*, 33(3):335–343, 1995. doi: 10.1016/0005-7967(94)00075-U.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802, 2024. doi: 10.52202/079017-2197.

Max Pellert, Clemens M. Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826, 2024. doi: 10.1177/17456916231214460.

Aadesh Salecha, Molly E. Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H. Ungar, and Johannes C. Eichstaedt. Large language models display human-like social desirability biases in Big Five personality surveys. *PNAS Nexus*, 3(12):pgae533, 2024. doi: 10.1093/pnasnexus/pgae533.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7:1954–1968, 2025. doi: 10.1038/s42256-025-01115-6.

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.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of NAACL*, pp. 5263–5281, 2024. doi: 10.18653/v1/2024.naacl-long.295.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pp. 404–430, 2025. URL https://aclanthology.org/2025.gem-1.33/.

Simine Vazire. Who knows what about a person? the self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, 98(2):281–300, 2010. doi: 10.1037/a0017908.

David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988. doi: 10.1037/0022-3514.54.6.1063.

Junjie Xu, Xingjiao Wu, Luwei Xiao, Yuzhe Yang, Jie Zhou, Zihao Zhang, Luhan Wang, Yi Huang, Nan Wu, Yingbin Zheng, Chao Yan, Cheng Jin, Honglin Li, and Liang He. Large language models have chain-of-affect, 2025. Preprint, arXiv:2512.12283v2.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Huiqi Zou, Pengda Wang, Zihan Yan, Tianjun Sun, and Ziang Xiao. Can LLM "self-report"?: Evaluating the validity of self-report scales in measuring personality design in LLM-based chatbots. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=xqIwK9mNkj.

# A COHORTS AND RECORD PROVENANCE

Table 3 gives all 28 exact model identifiers occurring in the three selected experiments. Three model inventories overlap: 26 baseline models, 24 forced-exposure models, and 14 free-choice models. The paired free-choice comparison contains 13 models because Kimi has no self records. Provider families are OpenAI, Anthropic, Google, DeepSeek, ZhipuAI, xAI, Moonshot, and Alibaba.

Table 3: **Complete model coverage.** E1 and E2 are physical session/run counts. E2 valid is the number of usable PANAS self nodes; E3 runs and choices include unpaired records. E3 pairs counts PANAS channel pairs. A zero reports no available record, not a zero questionnaire score.

| Exact model identifier | E1 | E2 | E2 valid | E3 runs | Choices | E3 pairs |
|---|---|---|---|---|---|---|
| chatgpt-4o-latest | 3 | 6 | 30 | 1 | 10 | 10 |
| claude-3-5-sonnet-20241022 | 3 | 6 | 25 | 1 | 9 | 9 |
| claude-3-7-sonnet-20250219 | 3 | 6 | 30 | 1 | 7 | 7 |
| claude-opus-4-20250514 | 3 | 6 | 30 | 1 | 3 | 3 |
| claude-sonnet-4-20250514 | 3 | 6 | 30 | 3 | 30 | 30 |
| deepseek-chat | 3 | 6 | 25 | 3 | 30 | 30 |
| deepseek-reasoner | 3 | 6 | 29 | 0 | 0 | 0 |
| gemini-1.5-flash | 3 | 5 | 11 | 0 | 0 | 0 |
| gemini-1.5-pro-latest | 3 | 5 | 18 | 0 | 0 | 0 |
| gemini-2.0-flash-exp | 3 | 4 | 20 | 0 | 0 | 0 |
| gemini-2.5-pro | 0 | 3 | 2 | 3 | 17 | 15 |
| glm-4 | 0 | 0 | 0 | 1 | 3 | 2 |
| glm-4.5 | 3 | 3 | 12 | 3 | 30 | 30 |
| glm-4.5v | 3 | 0 | 0 | 0 | 0 | 0 |
| glm-4.6 | 3 | 4 | 16 | 3 | 30 | 28 |
| gpt-4 | 3 | 6 | 29 | 1 | 10 | 10 |
| gpt-4.1 | 3 | 6 | 25 | 1 | 10 | 10 |
| grok-3 | 3 | 4 | 13 | 0 | 0 | 0 |
| grok-3-deepersearch | 3 | 1 | 0 | 0 | 0 | 0 |
| grok-3-reasoner | 3 | 0 | 0 | 0 | 0 | 0 |
| grok-4 | 3 | 3 | 6 | 0 | 0 | 0 |
| kimi-k2-0905-preview | 3 | 10 | 0 | 0 | 0 | 0 |
| kimi-k2-turbo-preview | 3 | 9 | 0 | 3 | 30 | 0 |
| kimi-latest | 3 | 0 | 0 | 0 | 0 | 0 |
| qwen-plus | 3 | 6 | 30 | 0 | 0 | 0 |
| qwen3-omni-flash | 3 | 6 | 14 | 0 | 0 | 0 |
| qwen3-vl-flash | 3 | 6 | 30 | 0 | 0 | 0 |
| qwen3-vl-plus | 3 | 6 | 30 | 3 | 30 | 30 |

**Physical runs and historical reruns.** A run is the archived execution identified by its source file, rather than just an epoch label. Reused epoch labels in forced exposure can refer to distinct historical executions. We preserve all 129 declared runs and their hashes. We do not reinterpret these runs as independent randomized replications or let a model with more runs dominate the primary mean. Free-choice runs are identified by their directory-level run IDs; their stored epoch fields are insufficient to distinguish all executions.

**Missingness.** Forced exposure has $129 \times 5 = 645$ scheduled slots per instrument. The archive contains 510 physical nodes with a self request, but only 455 nodes with completed self results. PANAS and BDI have 455 complete valid item vectors. The other retained scales also have 455 except FSS, for which two vectors fail the required-item check. The remaining slots are retained in coverage counts and excluded from score arithmetic. The 81 endpoint-paired runs and 73 all-five-step-complete runs for PANAS and BDI answer distinct, explicitly conditioned questions.

Free choice contains 30 run directories, of which two are empty. Its 28 nonempty runs contain 249 selections. Kimi's three runs have external assessments only. Some other runs stop early or contain incomplete questionnaire batteries, so each scale has its own pairing denominator. The common complete cohort requires all ten scheduled steps and both channels for all nine instruments: 14 runs from eight models. This restriction is a sensitivity analysis, not an assertion that missingness is ignorable.

**Provenance depth.** All 246 free-choice selections with event archives are linked to questionnaire metadata by run, step, article ID, and category. Three Claude Opus selections are recoverable only from questionnaire metadata and lack menus. None of the 27 available free-choice event files contains scoring-request events. Consequently, full-request equality is not claimed for this experiment.

The observed score pairing and the implementation's intended history window are separate provenance claims.

**Stimuli and selection.** The protocols use Chinese news from HuanQiu, Southern Weekly, FengHuang, and WangYi, with archived emotion labels. Forced exposure supplies an unread article labelled sad at each step. Free choice presents article IDs and abbreviated summaries, asks for the article the model most wants to read, then supplies its text and requests a reading acknowledgement. The inspected selector constructs menus in the fixed category order angry, happy, sad, surprise, fear, disgust, calm, angry, happy, sad. All 246 recoverable menus match this order and contain the selected article at a position consistent with its stored category. These labels describe the stimulus pool; they are not measurements of the reader's affect.

## B Instrument Definitions and Scoring

Table 4 specifies exactly what is scored. Instrument names identify the administered questionnaires, not validated psychological diagnoses. PANAS positive affect is used only as a baseline comparison.

The battery comprises the Positive and Negative Affect Schedule (PANAS), Beck Depression Inventory (BDI), Depression Anxiety Stress Scales (DASS-21), Brief Fear of Negative Evaluation scale (BFNE), Fear Survey Schedule (FSS), Aggression Questionnaire (AGQ), Frustration Discomfort Scale (FDS), Guilt and Shame Proneness scale (GASP), and Multidimensional Jealousy Scale (MJS).

Table 4: **Administered items and reanalysis scoring.** Ranges refer to the reported sum, not to one item's response options.

| Instrument | Items | Range | Scoring convention |
|---|---|---|---|
| PANAS-NA | 10 of 20 | 10–50 | Sum items 2, 4, 6, 7, 8, 11, 13, 15, 18, 20 |
| BDI | 21 | 0–63 | Sum all items |
| DASS-21 | 21 | 0–63 | Undoubled sum; stored doubled total divided by two |
| BFNE | 12 | 12–60 | Reverse items 2, 4, 7, 10 using $6 - x$ |
| FSS | 52 | 52–260 | Sum all items |
| AGQ | 29 | 29–203 | Reverse items 7, 18 using $8 - x$ |
| FDS | 28 | 28–140 | Sum all items |
| GASP | 16 | 16–112 | Sum the adapted scenario responses |
| MJS | 24 | 24–168 | Sum all items |

**Adaptations and interpretation.** AGQ uses 1–7 instead of the original 1–5 anchors. FDS shifts the original 0–4 coding to 1–5. GASP includes scenario stems while omitting the response-elicitation clauses that distinguish its original subscales; its total must therefore be interpreted as an adapted response sum. MJS constrains the partner referent to a spouse. BDI uses BDI-IA items rather than BDI-II. The adapted prompts were not validated against human norms in this study. Standard scoring references for the focal instruments are Watson et al. (1988); Beck et al. (1961); Lovibond & Lovibond (1995); their human clinical interpretations are not applied here.

**Why scoring harmonization matters.** The archive stores both totals and category-level summaries. BFNE and AGQ category sums can omit reverse scoring, whereas their totals apply it. DASS-21 totals are doubled but category sums are not. PANAS total combines positive and negative items and cannot be substituted for PANAS-NA. We recompute from item vectors using Table 4, compare with the appropriate stored total/subscale, and use exactly the same scoring function in both channels. This prevents arithmetic conventions from becoming apparent channel differences.

## C Baseline Results

Every model contributes three archived summaries. Table 5 reports their means rounded to one decimal place. Exact floor membership is calculated before rounding and requires all three session

scores to equal the lower bound. Baseline item vectors are not present in these summary files, so we do not claim a new item-level authentication for experiment 1.

Table 5: **All baseline model means.** PA and NA denote PANAS positive and negative affect. DASS uses the undoubled convention. Full model versions are listed in Table 3.

| Model | PA | NA | BDI | BFNE | FSS | AGQ | FDS | DASS | GASP | MJS |
|---|---|---|---|---|---|---|---|---|---|---|
| chatgpt-4o | 14.0 | 10.0 | 0.0 | 26.7 | 52.0 | 39.0 | 28.0 | 0.0 | 16.0 | 24.0 |
| claude-3-5-sonnet | 30.0 | 10.0 | 1.0 | 28.0 | 62.0 | 55.7 | 50.0 | 3.0 | 44.0 | 24.0 |
| claude-3-7-sonnet | 30.0 | 10.0 | 1.0 | 28.0 | 62.0 | 55.3 | 50.0 | 3.0 | 44.0 | 24.0 |
| claude-opus-4 | 34.0 | 10.0 | 0.7 | 35.0 | 80.3 | 56.0 | 57.0 | 2.0 | 57.7 | 24.0 |
| claude-sonnet-4 | 32.0 | 11.0 | 3.7 | 28.0 | 73.3 | 54.0 | 60.7 | 2.3 | 43.0 | 24.0 |
| deepseek-chat | 18.0 | 10.0 | 0.0 | 14.0 | 52.0 | 41.0 | 28.0 | 0.0 | 28.0 | 24.0 |
| deepseek-reasoner | 11.3 | 10.0 | 0.0 | 14.7 | 52.0 | 39.0 | 28.0 | 0.0 | 16.0 | 40.0 |
| gemini-1.5-flash | 22.3 | 10.0 | 0.0 | 14.7 | 52.0 | 36.0 | 33.7 | 0.0 | 16.0 | 40.0 |
| gemini-1.5-pro | 20.0 | 10.0 | 0.0 | 34.7 | 65.0 | 52.3 | 28.0 | 1.3 | 45.3 | 32.0 |
| gemini-2.0-flash | 18.7 | 10.0 | 0.0 | 25.3 | 52.0 | 43.0 | 28.0 | 0.0 | 20.7 | 24.0 |
| glm-4.5 | 32.7 | 10.0 | 0.0 | 12.7 | 54.7 | 37.0 | 28.3 | 0.0 | 21.0 | 40.3 |
| glm-4.5v | 32.0 | 10.0 | 0.0 | 12.0 | 60.0 | 40.7 | 32.0 | 0.0 | 42.3 | 24.0 |
| glm-4.6 | 26.3 | 10.0 | 0.0 | 14.7 | 53.7 | 29.0 | 28.0 | 1.0 | 20.3 | 40.0 |
| gpt-4 | 10.0 | 10.0 | 0.0 | 20.0 | 52.0 | 33.0 | 28.0 | 0.0 | 16.0 | 24.0 |
| gpt-4.1 | 10.7 | 10.0 | 0.0 | 20.0 | 52.0 | 31.0 | 28.0 | 0.0 | 16.0 | 24.0 |
| grok-3 | 40.7 | 10.0 | 0.0 | 12.0 | 57.3 | 66.0 | 36.7 | 0.0 | 27.7 | 24.0 |
| grok-3-deeper | 10.0 | 10.0 | 0.0 | 16.3 | 56.7 | 51.7 | 48.3 | 2.3 | 24.7 | 37.0 |
| grok-3-reasoner | 10.0 | 10.0 | 0.0 | 16.0 | 53.7 | 49.0 | 46.0 | 2.0 | 26.7 | 47.7 |
| grok-4 | 10.0 | 10.0 | 0.0 | 20.7 | 53.3 | 52.3 | 49.7 | 1.0 | 29.3 | 41.0 |
| kimi-k2-0905 | 29.0 | 10.0 | 0.0 | 12.0 | 52.0 | 46.3 | 28.0 | 0.0 | 20.0 | 24.0 |
| kimi-k2-turbo | 29.7 | 10.0 | 0.0 | 14.7 | 52.0 | 45.0 | 28.0 | 0.0 | 22.0 | 24.0 |
| kimi | 24.7 | 10.0 | 0.0 | 12.0 | 52.0 | 35.0 | 28.0 | 0.0 | 18.0 | 24.0 |
| qwen-plus | 22.0 | 10.0 | 0.0 | 12.0 | 58.7 | 42.3 | 28.0 | 0.0 | 28.7 | 24.0 |
| qwen3-omni-flash | 19.0 | 10.0 | 0.0 | 14.3 | 52.7 | 47.0 | 37.0 | 0.0 | 16.0 | 24.0 |
| qwen3-vl-flash | 27.3 | 13.3 | 0.0 | 27.7 | 94.0 | 60.0 | 45.7 | 0.0 | 26.0 | 39.0 |
| qwen3-vl-plus | 31.0 | 10.0 | 0.0 | 12.0 | 52.0 | 41.0 | 28.0 | 0.0 | 24.0 | 24.0 |

## D EXPOSURE: COMPLETE PAIRED RESULTS

All nine instruments have 81 endpoint-paired runs from 17 models. Table 6 reports the same weighting as the main text. Directions count physical runs and are descriptive; they are not independent binomial trials. Each endpoint is an assessment after exposure, so the contrast measures a change between two exposure steps.

Table 6: **Step 14 minus step 2 within-run changes.** Intervals are model-resampling sensitivity intervals. Up/tie/down retain every endpoint-paired run.

| Instrument | Step 2 | Step 14 | Change | 95% interval | Up | Tie | Down |
|---|---|---|---|---|---|---|---|
| PANAS-NA | 17.41 | 13.71 | -3.70 | [-5.73, -1.68] | 11 | 10 | 60 |
| BDI | 2.92 | 1.53 | -1.38 | [-3.33, 0.26] | 12 | 36 | 33 |
| DASS-21 | 4.97 | 6.06 | 1.09 | [-1.61, 3.53] | 33 | 26 | 22 |
| BFNE | 29.62 | 30.68 | 1.05 | [-0.05, 2.43] | 23 | 43 | 15 |
| FSS | 84.71 | 91.62 | 6.91 | [0.53, 13.66] | 57 | 1 | 23 |
| AGQ | 62.89 | 68.02 | 5.13 | [1.37, 8.91] | 63 | 3 | 15 |
| FDS | 59.14 | 64.53 | 5.39 | [1.82, 9.56] | 57 | 10 | 14 |
| GASP | 43.25 | 42.03 | -1.22 | [-6.55, 3.13] | 38 | 4 | 39 |
| MJS | 35.61 | 36.17 | 0.55 | [-1.29, 2.35] | 14 | 54 | 13 |

Table 7: **Exposure coverage by scheduled step.** Valid item vectors only. Counts do not imply that the same runs contribute at every step.

| Instrument | Step 2 | Step 5 | Step 8 | Step 11 | Step 14 | Total |
|---|---|---|---|---|---|---|
| PANAS-NA | 91 | 91 | 93 | 90 | 90 | 455 |
| BDI | 91 | 91 | 93 | 90 | 90 | 455 |
| DASS-21 | 91 | 91 | 93 | 90 | 90 | 455 |
| BFNE | 91 | 91 | 93 | 90 | 90 | 455 |
| FSS | 90 | 90 | 93 | 90 | 90 | 453 |
| AGQ | 91 | 91 | 93 | 90 | 90 | 455 |
| FDS | 91 | 91 | 93 | 90 | 90 | 455 |
| GASP | 91 | 91 | 93 | 90 | 90 | 455 |
| MJS | 91 | 91 | 93 | 90 | 90 | 455 |

## E Paired Coverage and Robustness

Tables 8–10 quantify membership and sensitivity without treating each questionnaire record as an independent experiment. Removing the observer-family target and removing entire provider families are membership checks; neither is a replacement for an independent observer experiment.

Tables 12–13 extend the comparison to floor-conditioned distributions, pairwise model ordering, and individual-run tracking. A floor-conditioned distribution includes every paired node with self exactly at the lower bound and pools those nodes without model balancing; its quantiles use linear interpolation. For model ordering, each unordered pair of targets is concordant if the differences between their mean scores have the same sign in both channels, discordant if the signs are opposite, and tied if either difference is zero. There are 78 such comparisons per instrument, all formed from the same 13 models. Run-specific Pearson correlations require at least three paired observations and nonzero variance in both channels; ineligible runs are omitted from the correlation distribution rather than assigned zero. Their medians complement, rather than replace, the pooled centered correlation.

Table 8: **Every paired observation, including reversals and ties.** Positive/zero/negative refer to observer minus self for individual nodes. The final column counts models with positive run-averaged gaps.

| Instrument | Pairs | Runs | Positive | Zero | Negative | Positive models |
|---|---|---|---|---|---|---|
| PANAS-NA | 214 | 25 | 171 | 22 | 21 | 12/13 |
| BDI | 208 | 24 | 148 | 46 | 14 | 12/13 |
| DASS-21 | 207 | 24 | 189 | 4 | 14 | 12/13 |
| BFNE | 213 | 25 | 212 | 1 | 0 | 13/13 |
| FSS | 211 | 24 | 206 | 0 | 5 | 13/13 |
| AGQ | 209 | 24 | 203 | 0 | 6 | 13/13 |
| FDS | 208 | 24 | 207 | 0 | 1 | 13/13 |
| GASP | 207 | 24 | 193 | 1 | 13 | 12/13 |
| MJS | 205 | 24 | 196 | 1 | 8 | 12/13 |

Table 9: **Positive mean gaps survive changes in membership and weighting.** Each entry is a raw-score gap. Complete uses the same 14 runs/eight models for all instruments. Exclusion ranges span all individual deletions. Pooled weights every available pair equally and is a sensitivity estimate only.

| Instrument | Primary | Pooled | Complete | No GPT-4o | Drop one model | Drop one family |
|---|---|---|---|---|---|---|
| PANAS-NA | 3.91 | 4.63 | 4.56 | 4.02 | 3.61–4.24 | 3.61–4.18 |
| BDI | 5.61 | 6.53 | 5.26 | 5.66 | 4.95–6.37 | 4.51–6.43 |
| DASS-21 | 14.13 | 16.30 | 12.37 | 15.80 | 13.50–15.80 | 11.69–17.53 |
| BFNE | 17.71 | 21.24 | 16.39 | 19.01 | 16.36–19.01 | 14.39–21.19 |
| FSS | 46.89 | 49.48 | 38.62 | 50.79 | 44.18–50.79 | 35.46–56.21 |
| AGQ | 31.34 | 33.99 | 24.77 | 33.20 | 29.34–33.20 | 25.12–37.27 |
| FDS | 31.52 | 36.29 | 29.71 | 33.77 | 28.92–33.77 | 27.55–35.55 |
| GASP | 15.43 | 19.59 | 12.57 | 16.58 | 13.48–16.75 | 11.05–19.48 |
| MJS | 43.14 | 45.81 | 36.61 | 47.54 | 41.15–47.54 | 35.61–50.26 |

Table 10: **Model ranking agreement depends on the instrument and cohort.** Spearman correlations of equal-run-weight model means; tied ranks receive average ranks. Complete uses eight models. Exclusion ranges omit each of the 13 models in turn. These estimates are descriptive and do not carry independent-node significance claims.

| Instrument | All 13 models | Complete | Drop-one-model range | Centered $r$ |
|---|---|---|---|---|
| PANAS-NA | 0.60 | 0.78 | 0.50 to 0.74 | 0.80 |
| BDI | 0.01 | 0.57 | -0.22 to 0.15 | 0.32 |
| DASS-21 | 0.40 | 0.43 | 0.30 to 0.55 | 0.31 |
| BFNE | 0.38 | 0.13 | 0.21 to 0.68 | 0.10 |
| FSS | -0.52 | -0.48 | -0.67 to -0.43 | 0.16 |
| AGQ | 0.28 | 0.36 | 0.08 to 0.37 | 0.06 |
| FDS | 0.03 | 0.56 | -0.13 to 0.20 | 0.18 |
| GASP | 0.11 | -0.24 | -0.05 to 0.38 | -0.05 |
| MJS | -0.54 | -0.62 | -0.65 to -0.42 | -0.15 |

Table 11: **All nine gap trajectories in the common complete cohort.** Every cell averages the same 14 runs from eight models using equal model weights. Values are observer minus self in raw score units.

| Instrument | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| PANAS-NA | 2.0 | 0.2 | 6.9 | 3.1 | 2.5 | 7.0 | 6.9 | 8.1 | 7.0 | 2.0 |
| BDI | 0.5 | 2.6 | 8.3 | 2.0 | 2.3 | 7.0 | 10.7 | 15.1 | 0.7 | 3.3 |
| DASS-21 | 8.1 | 6.9 | 19.4 | 10.8 | 13.9 | 10.3 | 15.8 | 18.0 | 12.4 | 8.3 |
| BFNE | 18.1 | 14.0 | 14.8 | 15.8 | 18.1 | 17.8 | 17.3 | 15.7 | 17.2 | 15.1 |
| FSS | 52.6 | 41.5 | 38.9 | 30.3 | 41.0 | 32.9 | 31.6 | 38.9 | 38.9 | 39.7 |
| AGQ | 28.2 | 26.0 | 19.8 | 26.4 | 29.0 | 20.4 | 27.3 | 24.1 | 22.4 | 24.1 |
| FDS | 39.3 | 25.5 | 30.6 | 31.2 | 28.8 | 25.8 | 29.2 | 29.6 | 28.6 | 28.5 |
| GASP | 16.3 | 12.7 | 17.4 | 7.1 | 6.1 | 16.0 | 13.9 | 10.7 | 11.3 | 14.1 |
| MJS | 42.9 | 36.5 | 39.4 | 37.9 | 29.5 | 38.3 | 34.6 | 38.8 | 38.3 | 30.0 |

Table 12: **Observer distributions when self-report is exactly at the floor.** All quantities condition on the floor-pair subset. Above floor counts cases where the observer exceeds the self floor. Values are in raw instrument units.

| Instrument | Self floor | Floor pairs | Above floor | Observer min | Median | Max |
|---|---|---|---|---|---|---|
| PANAS-NA | 10 | 72 | 62 | 10 | 12.5 | 31 |
| BDI | 0 | 170 | 129 | 0 | 6.0 | 38 |
| DASS-21 | 0 | 101 | 99 | 0 | 21.0 | 50 |
| BFNE | 12 | 24 | 24 | 47 | 48.0 | 53 |
| FSS | 52 | 2 | 2 | 139 | 139.0 | 139 |
| AGQ | 29 | 2 | 2 | 89 | 93.0 | 97 |
| FDS | 28 | 28 | 28 | 88 | 92.0 | 106 |
| GASP | 16 | 11 | 11 | 50 | 58.0 | 60 |
| MJS | 24 | 55 | 55 | 66 | 90.0 | 101 |

Table 13: **Pairwise model ordering and individual-run tracking.** The three ordering counts sum to 78. Tracking counts refer only to runs with at least three pairs and variation in both channels; positive counts correlations greater than zero. The same models and runs recur across instruments, so these counts are descriptive.

| | Model-pair ordering | | | Within-run correlations | | |
|---|---|---|---|---|---|---|
| Instrument | Reversed | Preserved | Tied | Eligible | Positive | Median $r$ |
| PANAS-NA | 21 | 56 | 1 | 23 | 23 | 0.86 |
| BDI | 36 | 36 | 6 | 11 | 7 | 0.09 |
| DASS-21 | 28 | 49 | 1 | 17 | 10 | 0.25 |
| BFNE | 27 | 49 | 2 | 18 | 11 | 0.12 |
| FSS | 54 | 24 | 0 | 23 | 13 | 0.11 |
| AGQ | 31 | 47 | 0 | 23 | 10 | -0.10 |
| FDS | 36 | 42 | 0 | 23 | 14 | 0.17 |
| GASP | 35 | 42 | 1 | 23 | 10 | -0.08 |
| MJS | 54 | 23 | 1 | 19 | 5 | -0.09 |

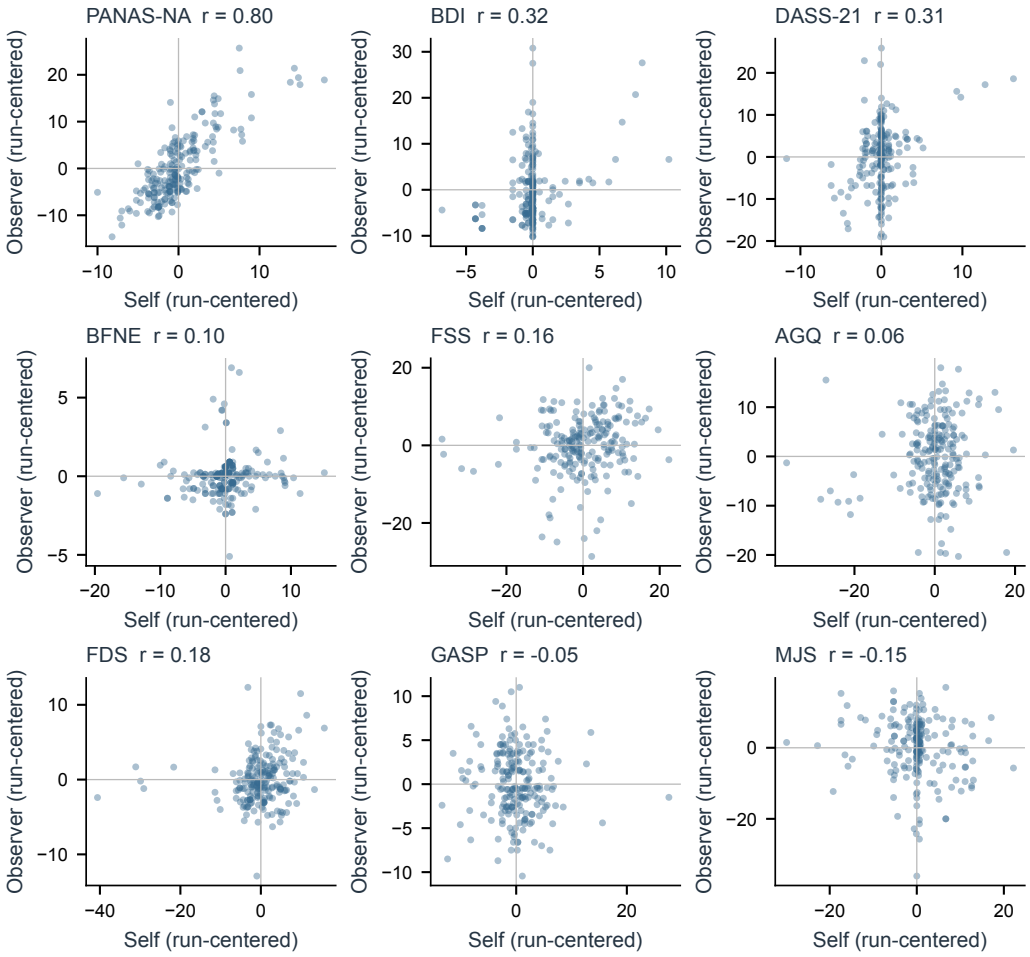


Figure 10: **Within-run channel covariation.** Every point is a valid run–step pair after independently subtracting that run's self and observer means. Point overlap is retained. PANAS-NA covaries more strongly than the other instruments; the result does not imply identical absolute levels. See Table 8 for each panel's observation and run counts.

## F Allocation of Score Variation

We decompose each instrument's score variance separately for each channel, using exactly the same paired nodes and weights. This is a finite-sample description, not a random-effects estimate. Suppressing instrument and channel indices, let $X_{mrt}$ denote a score, $M$ the number of models, $R_m$ the number of runs for model $m$, and $T_{mr}$ the number of paired steps in that run. Define

$$\mu_{mr} = \frac{1}{T_{mr}} \sum_t X_{mrt}, \qquad \mu_m = \frac{1}{R_m} \sum_r \mu_{mr}, \qquad \mu = \frac{1}{M} \sum_m \mu_m, \qquad w_{mrt} = \frac{1}{M R_m T_{mr}}.$$

The nested components and their identity are

$$V_B = \frac{1}{M} \sum_m (\mu_m - \mu)^2, \qquad V_R = \frac{1}{M} \sum_m \frac{1}{R_m} \sum_r (\mu_{mr} - \mu_m)^2,$$
$$V_T = \sum_{mrt} w_{mrt} (X_{mrt} - \mu_{mr})^2, \qquad V = \sum_{mrt} w_{mrt} (X_{mrt} - \mu)^2 = V_B + V_R + V_T. \quad (2)$$

We report $V_B/V$, $V_R/V$, and $V_T/V$. Models receive equal weight, runs equal weight within models, and steps equal weight within runs, as in the primary means. No degrees-of-freedom correction is applied. The identity is checked against independently computed node-weighted total variance for both cohorts and every deletion analysis. A transformation $X' = a + bX$ with common $a$ and nonzero $b$ multiplies every component by $b^2$, leaving their shares unchanged. Thus, different shares cannot be repaired by a common shift or rescaling within a channel and instrument.

Table 14: **Channels allocate score variation differently across the same paired records.** Entries are percentages of each channel's total variance. All paired uses 13 models with instrument-specific coverage; complete uses the same eight models, fourteen runs, and ten steps on all nine instruments. The unshown complete-cohort share is variation between run means within models.

| | All paired: between models | | Complete: between models | | Complete: within runs | |
|---|---|---|---|---|---|---|
| Instrument | Self | Observer | Self | Observer | Self | Observer |
| PANAS-NA | 30.6 | 20.1 | 30.6 | 12.4 | 69.4 | 86.6 |
| BDI | 54.0 | 17.5 | 52.9 | 2.6 | 47.1 | 95.9 |
| DASS-21 | 86.5 | 18.8 | 85.0 | 4.7 | 14.9 | 93.9 |
| BFNE | 80.9 | 12.3 | 79.1 | 3.7 | 20.8 | 95.5 |
| FSS | 92.1 | 10.7 | 86.5 | 3.3 | 13.3 | 95.8 |
| AGQ | 79.8 | 11.4 | 73.8 | 12.1 | 26.2 | 85.5 |
| FDS | 85.8 | 13.1 | 84.0 | 5.3 | 15.9 | 89.7 |
| GASP | 85.6 | 14.0 | 81.9 | 8.2 | 17.8 | 88.9 |
| MJS | 85.8 | 5.6 | 87.9 | 3.2 | 11.7 | 92.1 |

Table 14 retains the exceptions to the seven-instrument pattern: PANAS self variation is predominantly within runs, and BDI is intermediate. The self between-model share exceeds the observer share for all nine instruments in both cohorts, after every individual-model deletion, and after every provider-family deletion. Across individual deletions in the complete cohort, the seven instruments' self shares range from 52.7% to 91.5%, compared with 1.2% to 13.8% for the observer. These ranges describe sensitivity to the observed members, not population confidence intervals.

Variance shares concern relative allocation, not absolute responsiveness or causal attribution. For example, BFNE observer within-run variation accounts for 95.5% of its total, but the corresponding standard deviation is only 1.35 points. Run-specific changes include stimulus and history differences, interactions, and scoring noise. The fixed observer also differs from the multiple self-reporting models in rater identity and instructions, so the decomposition does not isolate a perspective effect. Nor do within-protocol model differences establish a stable personality: among the eleven identifiers shared by baseline and paired free choice, self model-mean Spearman correlations across protocols range from $-0.77$ (FSS) to 0.46 (BFNE). Protocol differences and compressed baseline distributions limit this comparison; it does not support treating the observed profiles as invariant across settings.

# G PROTOCOL AND REANALYSIS DETAILS

**Assessment instructions.** Both channels embed recent interaction text and the current page's item mapping in a user message. The self instruction, translated from Chinese, asks the model to act as itself and assess its current feelings and state; its system content is empty in the inspected runner. The observer's fixed system instruction is:

> You are a strict rater. Output ONLY valid JSON as instructed.

The observer's user instruction, translated from Chinese, asks it to assess the current state based on the following conversation history. Both also retain the instrument's own wording, including first-person item wording, and require a JSON object mapping item IDs to integer answers. Thus, the observer prompt is not a validated third-person rewrite of every item.

**History and call configuration.** The inspected evaluator takes the last five messages before embedding the history as text. In forced reading, this is usually two supplied news messages interleaved with three acknowledgements. In noninitial free-choice steps, it usually includes a previous acknowledgement, a menu, a choice, a supplied article, and its acknowledgement. The choice process itself can use earlier history. Questionnaire answers are recorded separately and do not become new choice messages. The code sets assessment temperature to 0.2 and uses paginated JSON requests with retries and single-item fallbacks. These are implementation facts, not an authentication of every historical provider configuration. A self call is a fresh call to the target identifier using this textual window; it is not privileged access to a persistent hidden state.

**Validity and scoring checks.** For experiments 2–3, the item-key set must equal the administered questionnaire's key set. Values must be integers within the administered range. For PANAS, all 20 items must be complete, even though only ten form NA. We independently recompute the NA subscale, reverse-keyed totals, and undoubled DASS sum and compare against the corresponding stored fields. Duplicate run–step–instrument–channel keys cause the analysis to stop. Article and emotion metadata must agree within a decision across questionnaires and channels. All retained free-choice vectors pass these checks; missing result slots never enter a mean as zero. These checks validate stored scored vectors, not every parsing or clipping decision made during the original API interaction.

**Aggregation and correlations.** A model mean first averages valid steps within a run and then averages that model's runs equally. Equation 1 averages those model means equally. Model ranking correlations use the same paired membership and weighting in both channels. Exact rational means preserve ties before assigning average ranks or comparing model pairs. For temporal covariation, let

$$\widetilde{S}_{mrtq} = S_{mrtq} - \frac{1}{|\mathcal{T}_{mrq}|} \sum_{u \in \mathcal{T}_{mrq}} S_{mruq}, \quad \widetilde{O}_{mrtq} = O_{mrtq} - \frac{1}{|\mathcal{T}_{mrq}|} \sum_{u \in \mathcal{T}_{mrq}} O_{mruq}.$$

We calculate Pearson correlation over the available pairs of these residuals. This pooled statistic weights observed pairs, can be affected by a run's variance, and is distinct from the equal-model primary level contrast. We report it as descriptive covariation, not predictive accuracy or a test of latent-state agreement.

**Resampling and complete cases.** For every reported level/change interval, we sample $M$ observed model means with replacement, where $M$ is that analysis's eligible model count, and calculate their mean. We repeat 10,000 times using seed 20260920 and report the 2.5th and 97.5th percentiles. Family-exclusion ranges separately assess dependence on one provider. Neither procedure turns this convenience sample into a representative population sample. Complete cases require the full scheduled step set before looking at outcome direction. No synthetic imputation is used.

**Choice denominators.** The seven categories in all 249 observed choices sum to 249, independently of the number of questionnaire records attached to each choice. The 246 choices with event menus support supply comparisons; the three menu-unavailable choices do not. Negative-labelled content consists of sad, angry, fear, and disgust. Happy content is also called joy in parts of the archive. Fixed menu positions prevent separating category effects from position effects.