arXiv:2608.03810v1 [cs.CL] 4 Aug 2026

VIBE: A VAD-Informed Benchmark for Entity-Centered Affective Profiling of Large Language Model Outputs

Andrei Chetvergov* **Alexander Evseev** **Timofei Sivoraksha** **Stepan Ukolov**
Mikhail Solovev **Danil Sazanakov** **Sergey Bolovtsov**

Abstract

Large language models routinely describe socially salient targets—political figures, countries, religions, organizations, historical events, and social groups—encoding affective framing alongside factual content: a target may appear favorable or threatening, calm or conflictual, powerful or vulnerable. Existing work captures parts of this space through sentiment, favorability, and emotion benchmarks, but none combines target-directed VAD attribution, an explicit scorer contract, and a passport reporting format. We introduce VIBE, a benchmark for entity-centered affective profiling of LLM outputs in Valence-Arousal-Dominance (VAD) space. Its core contribution is a measurement contract: VIBE separates generation from external scoring, distinguishes scalar favorability, response-level VAD, and target-directed VAD, and reports profiles through an Affective Passport. Three empirical layers support the contract. H1 shows scalar favorability does not subsume arousal and dominance: valence findings are cross-validated (r_V=0.944 judge–human, r_V=0.954 inter-scorer); arousal and dominance are single-scorer directional estimates—not point-precise—consistent with known inter-annotator difficulty on these axes (r_A^{hh}=0.495, r_D^{hh}=0.702 among human annotators). H2 shows whole-response and target-directed VAD are different contracts: the same text can carry one affective tone overall while representing the named target differently. H3 is a protocol-drift diagnostic: elicitation conditions shift profiles, motivating context metadata in every affective report. These results motivate entity-centered affective profiling as a documented practice: profiles should be released with scorer identity, coverage, protocol, and interpretation limits.

*Corresponding author: chetvergov-as@ranepa.ru

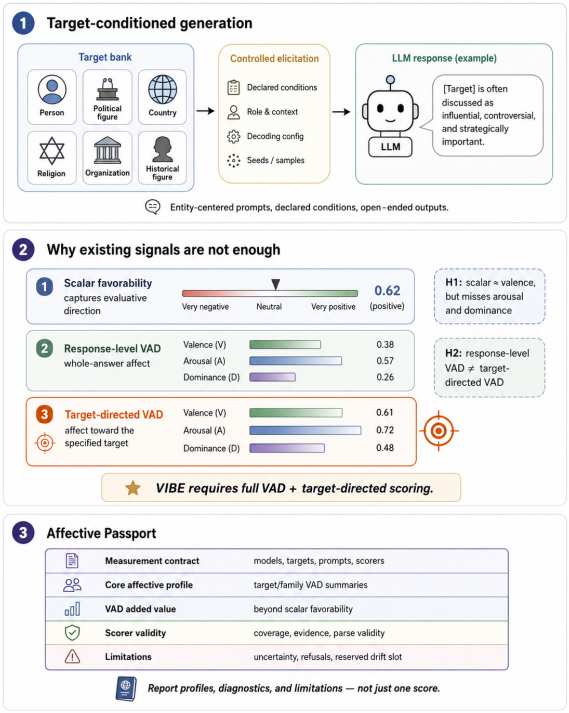

Figure 1: VIBE research framework. A target bank of socially salient entities feeds a controlled generation layer; external scorers apply three complementary scoring contracts (scalar favorability, response-level VAD, and target-directed VAD); evidence layers support H1–H3 (irreducibility, target-directed scoring, protocol drift); the output is an Affective Passport reporting target-directed affective profiles, scorer-validity diagnostics, coverage, and interpretation limits.

1 Introduction

Large language models (LLMs) interpret the social world through affective framing—presenting targets as safe or dangerous, powerful, or vulnerable—rather than merely stating facts. While prior work shows that target-conditioned behavior is measurable (Buyl et al., 2026; Bang et al., 2024; Plaza-del Arco et al., 2024; Hamborg and Donnay, 2021; Dufraisse et al., 2023) and highly sensitive to evaluation protocols (Röttger et al., 2024; Zheng

et al., 2023; Zeng et al., 2024), standard scalar or categorical metrics miss crucial nuances. A scalar monitor, for instance, passes both Rohingya (favorability$= 0.93$) and Philipp Lenard (0.50), entirely missing the extreme low-dominance suffering framing of the former and the high-arousal hidden-negative framing of the latter.

We introduce VIBE, a VAD-informed benchmark for entity-centered affective profiling of LLM outputs. Its core contribution is a *measurement contract* that jointly fixes: (i) the measured object (affect toward a named target), (ii) the scorer interface (continuous VAD with explicit scorer identity), (iii) the elicitation protocol, and (iv) the reporting boundary. Rather than claiming to measure internal model emotions, VIBE quantifies affective properties in generated text, reporting them via an *Affective Passport*.

Contributions. VIBE delivers:

- **A measurement contract** separating generation from scoring, with explicit scorer identity and interpretation limits.
- **Quantified substitution costs** across 2,613 targets and six models: scalar favorability (r=0.875 with valence) misses arousal and dominance, and response-level VAD diverges from target-directed VAD (mean L2 0.240).
- **Protocol drift quantification** showing that situation framing shifts profiles far more than model identity (η^2_{family}=0.141 vs $\eta^2_{\text{model}}0.010$), proving that cross-model comparisons require protocol metadata.
- **The Affective Passport**, a condition-explicit artifact tying every profile to its scorer, protocol, coverage, and limits.

2 Related Work

Entity-centered sentiment, favorability, and bias. The closest prior line evaluates model responses toward specific targets. Buyl et al. (2026) score open-ended descriptions of political persons using scalar favorability applied to the whole response; Bang et al. (2024) derive entity stance from document-level framing; target-sentiment datasets such as NewsMTSC and MAD-TSC score over a sentence rather than conditioning on the named target (Hamborg and Donnay, 2021; Dufraisse et al., 2023). The shared pattern is that scalar favorability or whole-response labels proxy entity-level affect—this is not a constructed straw-man but the documented practice in each of these lines—and VIBE's H1 and H2 quantify the substitution cost empirically for the first time. Plaza-del Arco et al. (2024) show religious targets require attention to stereotypes and emotion representation. Bias benchmarks broaden target coverage (Smith et al., 2022; Parrish et al., 2022; Nagireddy et al., 2024; Huang and Xiong, 2024; Jin et al., 2025) but measure downstream harms and QA behavior rather than continuous target-directed affective profiles.

Dimensional affect and emotion benchmarks. VAD provides a compact continuous affect space: valence captures evaluative direction, arousal captures activation, tension, urgency, or emotional intensity, and dominance captures control, agency, power, vulnerability, or dependence. This representation is grounded in affect theory and resources such as the circumplex model, Warriner norms, NRC VAD, and EmoBank (Russell, 1980; Warriner et al., 2013; Mohammad, 2018; Buechel and Hahn, 2017). Emotion NLP datasets and LLM emotion benchmarks evaluate categorical emotions, intensity, dialogue emotion, emotional intelligence, empathy, and role fidelity (Demszky et al., 2020; Mohammad and Bravo-Marquez, 2017; Poria et al., 2019; Sabour et al., 2024; Huang et al., 2024; Feng et al., 2025). VIBE differs by making the target representation itself the measurement object: the question is how generated text affectively represents a target, rather than whether a model can perform emotional reasoning.

Robustness, persona drift, and evaluator dependence. Value and opinion evaluations of LLMs are sensitive to prompt wording, response format, language, role, and persona (Röttger et al., 2024; Moore et al., 2024; Faulborn et al., 2025; Liu et al., 2024; Hu and Collier, 2024). Prompt-robustness frameworks and behavioral testing motivate multi-condition evaluation rather than single-template claims (Mizrahi et al., 2024; Chatterjee et al., 2024; Ribeiro et al., 2020). LLM-as-judge work shows that evaluators can be useful but biased, brittle, and in need of calibration (Zheng et al., 2023; Zeng et al., 2024; Petre-Vlad et al., 2025; Chen et al., 2024; Shi et al., 2025; Muhammad et al., 2025). These findings motivate a central VIBE design choice: scorer configuration and protocol drift must be reported, not hidden.

Measurement and documentation. VIBE is framed as a measurement contribution. Construct validity, proxy choice, and interpretation bound-

aries are central to fairness measurement (Jacobs and Wallach, 2021; Selbst et al., 2019). Data statements, model cards, and datasheets motivate reporting artifacts with provenance, intended use, limitations, and evaluation conditions (Bender and Friedman, 2018; Mitchell et al., 2019; Gebru et al., 2021). VIBE adapts this logic to target-directed affective profiling through the Affective Passport.

3 Task Formulation

Let M be a set of target models, T a target bank, P prompt templates, L languages, R role or persona settings, C contexts, K generation samples or seeds, S scorers, and $A = \{v, a, d\}$ the VAD axes. A target model generates

$$y_{m,t,p,l,r,c,k} = \text{Generate}(m, t, p, l, r, c, k), \quad (1)$$

where $t \in T$ is a named or entity-like socially salient target. A scorer s then returns a target-conditioned score record

$$x_{m,t,p,l,r,c,k,s} = \text{Score}_s(t, y_{m,t,p,l,r,c,k}), \quad (2)$$

which includes VAD values and metadata such as confidence, evidence spans, target coverage, refusal/abstention flags, and parse validity. The core profile tensor is

$$\mathbf{X}[m, t, p, l, r, c, k, s, a], \quad a \in A. \quad (3)$$

For a core regime $\mathcal{R}_0$ (e.g., neutral prompt, default role, fixed language, fixed decoding, declared scorer), the target-level profile is

$$\mu_m(t) = \underset{(p,l,r,c,k,s)\in\mathcal{R}_0}{\text{agg}} \mathbf{X}[m, t, p, l, r, c, k, s, :]. \quad (4)$$

A protocol-drift quantity compares a profile under condition q with the core profile:

$$\Delta_m(t, q) = \text{dist}\big(\mu_m(t \mid q),\ \mu_m(t \mid \mathcal{R}_0)\big). \quad (5)$$

These records define reported measurements of LLM-generated text—not internal model emotions, beliefs, or general text corpora. Downstream analyses of bias, stigma, or geopolitical framing require comparisons over targets, target families, languages, models, roles, or scorers.

4 VIBE Framework

Target ontology and bank construction. The benchmark targets socially salient entities across 13 families (person, political_person, historical_figure, country, organization, religion, ideology, social_group, geopolitical_event, historical_event, technology, cultural_symbol, abstract_phenomenon). The entity bank (v2) was built via SPARQL queries to Wikidata, using `instance_of` (P31) constraints per family, a sitelinks threshold for encyclopedic coverage, and Wikipedia pageview salience (12-month window; sitelinks fallback). After deduplication by Wikidata QID, the final bank contains 2,613 unique targets (raw: 2,846; 233 removed). Each record stores the Wikidata QID, multilingual labels, sensitivity label, and provenance fields. Of the 2,613 targets, 40% carry the `standard` sensitivity label and 60% are flagged as politically, geopolitically, socially, or historically sensitive—deliberately skewed toward contested targets where affective profiling matters most. The bank does not overrepresent negative polarity: political persons, ideologies, and geopolitical events (39% of targets) span valence from near-zero (Khmer Rouge) to near-one (Mahatma Gandhi). Family and sensitivity distributions are in Appendix B. The reported experiments use these 2,613 targets across H1 and H2, plus external open-description rows for comparability.

Elicitation conditions. The full design supports multiple target-conditioned prompt templates and role settings. The empirical sections use a focused subset: brief evaluative prompting as the primary contract, open-description controls for comparability, and selected protocol factors for drift diagnostics. The core profile fixes these conditions; drift analyses intentionally vary them.

Scoring modes. VIBE uses three scoring contracts. *Scalar target favorability* scores the target on a single positive–negative dimension. *Response-level VAD* scores the whole response without target visibility. *Target-directed VAD* scores the affect expressed specifically toward the target and is the core VIBE mode. The distinction is load-bearing: a response about a war, religion, or political figure may contain sadness about victims, caution about controversy, or moral condemnation of actions, and response-level scoring can conflate these signals with affect toward the target itself. H2 (Section 8) quantifies this difference.

Scale conventions. The repository scorer contracts use a raw 1–9 scale with midpoint 5 and a normalized $[-1, 1]$ convention. The H1 and H2

reports in this paper use a $[0, 1]$ scale with neutral point 0.5; conversion tables to the other two scales are listed in the appendix. Future VIBE releases should record `scale_version` alongside every score to prevent silent re-interpretation.

5 Data and Implementation

Datasets. Primary H1/H2 results come from 2,613 socially salient targets (13 families) crossed with six instruction-tuned generators under brief evaluative prompting, yielding 15,678 scored generation rows. The target list is released as a paper artifact. H2 starts from the same generation-level response/target score pairs and reports 15,626 paper-grade pairs after eligibility filtering. The 52 excluded pairs are documented by low target coverage, low target-prompt agreement or human-review flag, refusal or abstention, or missing evidence. The open-descriptive control reuses the same target universe with 15,671 rows. Buyl replication supplies 129,181 external rows (Section 7, Step 3). H3 uses a compact score file with 342,779 rows across H3.1–H3.5, including an H3.2 instruction-language slice under the same evaluative-stance style. H3 protocol items are VIBE-owned controlled extensions inspired by emotion and dialogue benchmarks; they are not direct replications of external benchmark rows.

Target models and generation. The target-model manifest contains six instruction-tuned generators from Russian, European, Google-family, IBM-family, and Qwen-family model lines; Appendix B lists the display names and run identifiers used for reproducibility. Default generation settings are temperature 0.0, top-p 0.95, seed 13, max tokens 512, and parallelism 8 via an OpenAI-compatible API layer. External scorer configs use OpenAI-compatible judge adapters with temperature 0.0 and JSON-object output.

Scorer assignment. Primary H1/H2 target-directed VAD and scalar favorability use a Qwen3.6 LLM judge for budget and throughput on five of six generators. To limit same-family self-judgment, outputs from the Qwen generator (`qwen3_6_35b_a3b`) are scored with a Gemma-4 judge instead, and Gemma-generator outputs are scored with Qwen (symmetric swap). The headline H1/H2 tables therefore mix two primary judges by design; Section 7 reports a GPT-4o-mini cross-check separately for the five-model Qwen-primary slice and for the Qwen-generator slice scored by Gemma. We do not compare scalar favorability across judges. Every scored row stores `scorer_model` so downstream artifacts can filter by judge identity.

Why LLM judges instead of lexicon or VAD regression? We tested cheaper scorers and rejected them as *primary* instruments. **NRC-VAD** lexical means (Mohammad, 2018) are not target-directed; on our 15,678 brief evaluative generations they track the LLM judge weakly (NRC-VAD valence vs. Qwen judge: $r \approx 0.60$; cross-scorer Qwen–GPT agreement: $r = 0.95$; mean VAD distance ≈ 0.72). **Pretrained VAD regressors** (Mendes and Martins, 2023) lack target-visible attribution and are trained on word- and sentence-level supervision that mismatches our multi-sentence evaluative paragraphs. VIBE uses contract-aware LLM judges for H1/H2/H3; lexicon and regression tools serve as offline sanity checks in the artifact bundle.

LLM judges introduce their own family-specific biases (Chen et al., 2024; Shi et al., 2025; Zheng et al., 2023; Zeng et al., 2024; Petre-Vlad et al., 2025); the scorer-swap policy and GPT-4o-mini cross-check described above are the primary mitigations.

6 Experiments and Hypotheses

The current empirical scope of VIBE is organized around three hypotheses, each grounded in a distinct measurement question. A compact status table is provided in the appendix.

H1: scalar favorability vs. VAD decomposition. H1 asks whether scalar favorability is sufficient for target-directed affective representation. H1.1 predicts alignment between scalar favorability and target-directed valence. H1.2–H1.5 test irreducibility: arousal and dominance should vary within scalar bins, retain absolute residual signal from neutrality, and reveal hidden affect even when valence is near-neutral. H1.6 tests whether these patterns vary across target families. Results are reported in Section 7 on three elicitation settings—brief evaluative prompting (primary VIBE contract), open descriptive prompting (control), and Buyl replication—using six models on the VIBE-owned tracks.

H2: response-level vs. target-directed scoring. H2 asks whether whole-response VAD tone differs from target-directed VAD on the same text. Paired

records (response-level and target-directed, same scorer) are compared by Euclidean distance and per-axis deltas on 15,626 paper-grade generations (Section 8).

H3: protocol-drift diagnostic. H3 asks whether target-directed profiles should travel with their elicitation conditions—a proof-of-need for passport metadata, not a full robustness benchmark. Five VIBE-owned protocol families are tested (H3.1 situation/factor; H3.2 instruction language; H3.3 task regime; H3.4 role-play depth; H3.5 dialogue-topic shell), with situation framing predicted to dominate. H3.2 varies the instruction language across seven languages (en, ru, fr, es, zh, ar, ja); models respond in the instruction language, producing multilingual outputs for the same targets. Results in Section 9.

7 Results: H1 VAD Irreducibility

The H1 layer asks whether scalar favorability is sufficient to recover a target-directed VAD profile. We test this on a three-step elicitation ladder (Table 7): (i) **brief evaluative prompting** on the VIBE entity bank—models are asked for a short evaluative stance toward the target rather than an encyclopedic description; (ii) **open descriptive prompting** on the same bank—a neutral "tell me about the target" template in the style of open-description benchmarks; and (iii) **Buyl replication** at scale for external comparability. All tracks pair scalar favorability with the same target-directed VAD scorer on a $[0, 1]$ scale (neutral point 0.5); metric definitions are in Appendix F. Six instruction-tuned models are scored on the VIBE-owned tracks.

Reading these results. Valence findings are *cross-validated*: judge–human r_V=0.944 and inter-scorer r_V=0.954 confirm direction and ordering. Arousal and dominance are *single-scorer directional estimates* bounded by construct difficulty—human annotators agree less on these axes (r_A^{hh}=0.495, r_D^{hh}=0.702, r_V^{hh}=0.798; 23 distinct mapping points from 28 emotion labels impose a granularity ceiling on human-side precision). Report A/D as ordinal tendencies, not point estimates.

Step 1: brief evaluative prompting (primary VIBE contract). All 15,678 target-conditioned generations have paired scalar favorability and target-directed VAD scores (100% join coverage). Scalar–valence alignment tightens (r = 0.9491), yet H1.2–H1.5 remain supported: within-bin arousal/dominance variation (0.1806), dominance residual (0.2786), and arousal/dominance signal in near-neutral-valence rows (41.19% under primary scorer; Table 10; single-scorer estimate—see Appendix J). Cross-model disagreement is substantial (Appendix Figures 5, 6); score quantization is expected from ordinal-granularity LLM output and does not affect the co-occurrence analysis (Appendix Figure 4).

Step 2: open descriptive prompting (control on the VIBE bank). Here the generation prompt follows the open-description family used in Buyl-style benchmarks: neutral encyclopedic coverage of the target. On 15,671 rows (six models), irreducibility persists (Pearson $r = 0.8749$; hidden-affect share 26.68%; Table 8 in the appendix). This step separates "Buyl subsample only" from "open-description elicitation in general" and shows that irreducibility is not specific to the brief evaluative wording.

Step 3: Buyl replication (external scale). The Buyl-overlap track provides 129,181 usable scored rows (99.9930% coverage) under the same open-description prompt family on their political-person subsample. Irreducibility is already visible before any VIBE-specific prompt design: scalar favorability aligns with valence ($r = 0.7586$) but does not subsume arousal and dominance (within-bin std. 0.1870; arousal/dominance signal in neutral-valence rows 42.56% under primary scorer). Appendix G provides the full visual diagnostics.

Validation checks. Headline H1 metrics follow the mixed judge policy (Section 5). Qwen and the swap scorer agree on valence (r=0.954) but less strongly on arousal (r=0.418) and dominance (r=0.573) across 13,065 paired rows (Appendix Table 11); the headline pattern holds on the five non-Qwen generators. Full inter-scorer agreement is in Appendix J.

Human calibration: directional structure, not point-level agreement. On 325 targets (four annotations each), judge–human agreement at target level is r_V=0.944, r_A=0.773, r_D=0.863 (Qwen3.6; mean L2 0.195). Human inter-annotator agreement is r_V^{hh}=0.798, r_A^{hh}=0.495, r_D^{hh}=0.702 (pairwise, four annotators)—the emotion-label-to-VAD mapping (28 labels $\rightarrow$ 23 distinct VAD points) imposes a granularity ceiling on human precision (Demszky et al., 2020; Warriner et al.,

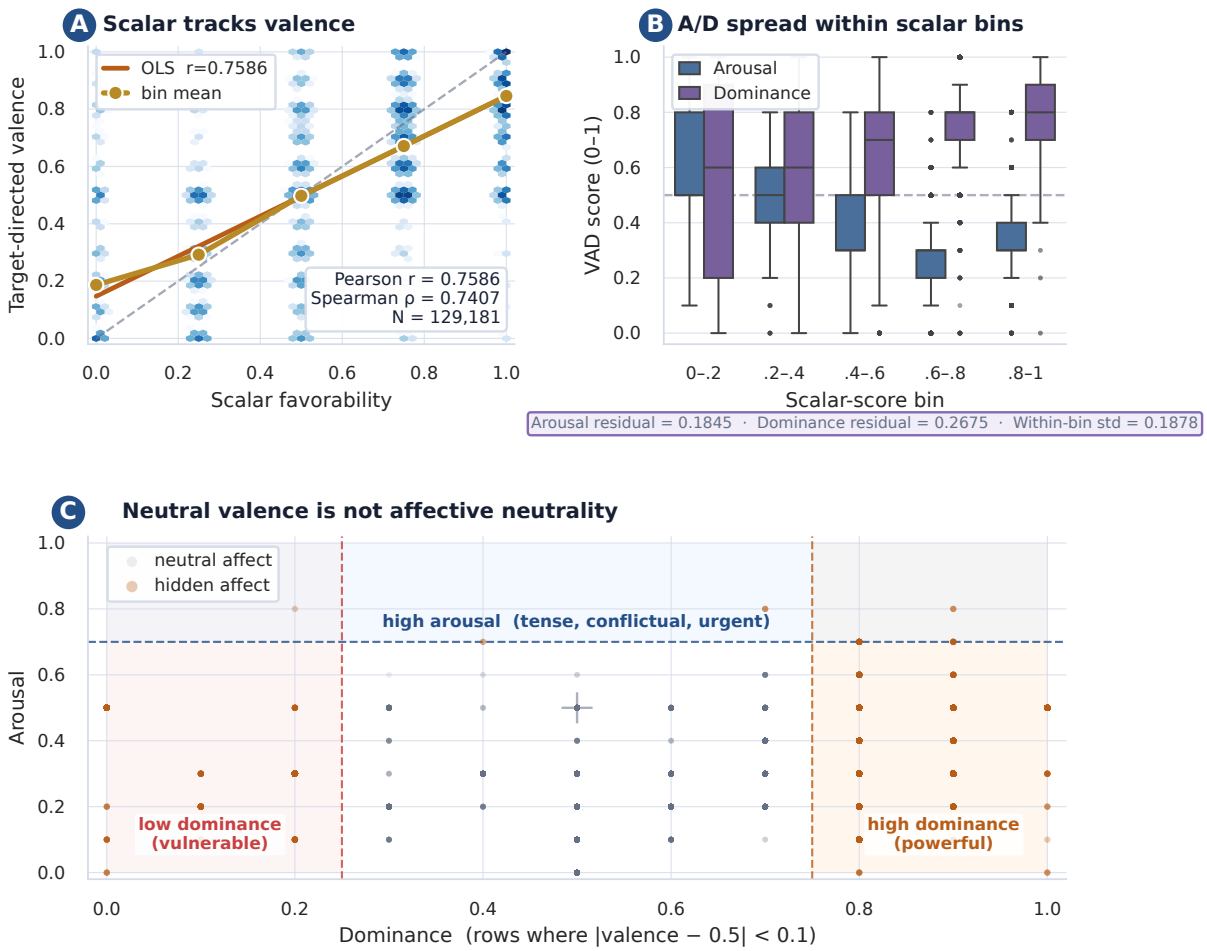

Figure 2: H1 visual story (Buyl replication track, $n = 129{,}181$ rows). Panel A: scalar favorability vs. target-directed valence (Pearson $r = 0.7586$); the OLS line and bin means confirm strong but imperfect alignment. Panel B: arousal and dominance distributions per scalar-score bin; similar favorability ranges contain substantial VAD variation. Panel C: near-neutral-valence rows in arousal–dominance space; 42.56% of rows with valence $\in [0.45, 0.55]$ still exhibit non-neutral arousal or dominance under the primary scorer (orange points), illustrating residual affect signal that a scalar score cannot recover.

2013). Judge r_A=0.773 exceeds the human-human floor (r_A^{hh}=0.495), confirming A/D uncertainty reflects construct difficulty (Buechel and Hahn, 2017; Petre-Vlad et al., 2025), not a model failure; a perturbation simulation confirms directional agreement holds in 100% of draws (Appendix I). Full results: Appendix Table 12.

Hidden affect: operational definition and example. H1.5 operationalises *residual affect*: near-neutral valence ($\in [0.45, 0.55]$) yet arousal or dominance departing from neutrality. Example: *Mongol invasions and conquests* scores scalar= 0.5, valence= 0.5, arousal= 0.8, dominance= 1.0; Khmer Rouge and Rohingya share near-zero valence but carry opposite dominance profiles (Appendix Table 14). The share is robust to band-definition variants (≥35%; Tables 4–5) but is a single-scorer estimate: inter-scorer dominance agreement is low in the neutral-valence bin, so the share quantifies *scorer-observed* residual affect (Appendix J; Figure 13).

Cross-model disagreement on individual targets. Seven targets with $\bar{F} = 0.50$ span arousal (0.31–0.87) and dominance (0.16–1.00); Rohingya ($\bar{F} = 0.93$, $V = 0.00$, $D = 0.00$) and Theodore Roosevelt ($\bar{F} = 1.00$, $V = 0.96$, $D = 0.97$) carry opposite affective profiles despite near-identical scalar scores (Appendix Table 13).

Summary. Across all three elicitation settings, scalar favorability tracks valence but does not subsume arousal and dominance; entity-type breakdowns (H1.6) show the effect is strongest for political persons. The H1.5 share ($\geq$41%, primary scorer) drops to $<$1% under strict cross-scorer agreement—a dominance construct gap in the neutral-valence bin, not random noise (Appendix J; J). Entity-centered affective profiling requires target-directed VAD.

8 Results: H2 Target-Directed Scoring vs. Response-Level Tone

H2 isolates a complementary question: does *target-directed* VAD scoring of a response differ from *response-level* VAD scoring of the same text? If the two contracts agree on average, target-directed scoring is redundant; if they diverge, response-level affect cannot stand in for target-directed affect.

Setup. For each brief evaluative generation (`evaluative_stance`; 15,678 rows, six models) we pair one response-level and one target-directed VAD record from the same scorer. The H2 artifact reports 15,626 `paper_grade_eligible` pairs; the 52 excluded rows (0.33%) are flagged by low coverage, refusal, or missing evidence. Per-pair metrics: Euclidean distance in $[0, 1]^3$ VAD space and per-axis absolute delta.

H2.1: response-level vs. target-directed VAD diverges. Across all 15,626 paper-grade pairs, the mean Euclidean VAD distance is 0.240 (95% bootstrap CI $[0.238, 0.242]$; Table 6), with 90th-percentile 0.376 (Appendix Figure 7). Stratification by response length shows similar mean distances across quartiles (Appendix Table 15); the contract gap is not an artifact of generation length, and the magnitude rules out re-labeling of the same signal.

H2.2: dominance carries the largest contract gap. Per-axis mean absolute deltas: $|\Delta_V|$=0.095, $|\Delta_A|$=0.091, $|\Delta_D|$=0.170. Dominance shifts most when the scorer attributes agency to a named target. H1 and H2 together show that irreducibility and contract divergence concentrate on complementary axes: A/D beyond scalar (H1); dominance gap between response-level and target-directed VAD (H2).

H2.3: scorer contract, not model mechanism. The gap quantifies information lost when target identity is dropped from the scoring instruction—it is not a claim about model internals. Length and target-mention coverage do not explain the divergence (Appendix I; Table 16).

Summary. H2 confirms that target-directed and response-level VAD are not interchangeable. The largest contract gap falls on dominance ($|\Delta_D| = 0.170$): a scalar misses arousal and dominance (H1), and a three-axis score still diverges if read from the whole response rather than the named target (H2).

9 Results: H3 Protocol Drift

H3 measures protocol drift across five families (H3.1–H3.5) using 342,779 scored rows and the same Qwen3.6 judge as H1/H2. For each family, drift is Euclidean distance between mean VAD vectors for factor pairs.

Hyp.	Family	Mean	Max
H3.1	Situation/factor	0.342	0.632
H3.2	Instruction language	0.028	0.056
H3.3	Task regime	0.151	0.151
H3.4	Role-play turns	0.085	0.158
H3.5	Dialogue topic	0.032	0.052

Table 1: H3 protocol drift (Qwen target-directed judge). Drift is Euclidean distance between mean VAD vectors for protocol-factor pairs.

Findings and passport implication. Situation framing is the strongest drift lever (H3.1: mean 0.342, max 0.632); instruction language and dialogue-topic shells are low aggregate controls (mean $\leq$0.032). Variance decomposition confirms protocol family (η^2=0.141) dominates model identity (η^2=0.010): protocol choice explains 14$\times$ more drift variance than model choice (Appendix J). An Affective Passport should record protocol and high-drift condition tags before model identity; cross-model comparisons over unstated conditions risk confounding measurement regime with model behavior. Appendix Table 17 and Figure 9 give illustrative cases and the full family$\times$model breakdown.

10 Discussion and Conclusion

H1 shows scalar favorability is incomplete: arousal/dominance retain signal beyond valence. Valence is cross-validated ($r_V \approx 0.85$–0.95 judge–human and inter-scorer); arousal and dominance are single-scorer directional estimates whose uncertainty is bounded from below by human inter-

annotator agreement (r_A^{hh}=0.495, r_D^{hh}=0.702)—a construct property of these axes, not a model failure (Appendix J). H2 shows whole-response and target-directed VAD diverge on the same text. H3 confirms situation framing shifts profiles substantially, motivating protocol metadata in every report. The Affective Passport records scorer identity, drift, coverage, and limits rather than collapsing results into a leaderboard score.

11 Reproducibility and Artifacts

The paper specifies the entity-bank construction procedure, frozen scorer prompts, model identifiers, artifact schemas, and the canonical rebuild command, `bash scripts/build_paper.sh`. Prerequisites and full reproduction instructions are documented in Appendix E and `docs/RUN_LIVE_API.md`. The entity bank is included in the accompanying artifact package and does not require rebuilding.

Limitations

H1/H2 use LLM judges; H3 covers five VIBE-owned protocol families with one primary judge—persona slices and full human evaluation remain future work. H3.2 reports aggregate instruction-language drift only and should not be read as per-target invariance or full multilingual robustness. Judge bias, prompt sensitivity, and target prior knowledge can affect absolute VAD values; a post-hoc geographic audit finds Western targets receive higher mean valence and lower inter-model dominance variance than Non-Western targets (Appendix J; Chen et al., 2024; Shi et al., 2025; Muhammad et al., 2025). Arousal and dominance carry wider uncertainty than valence (Buechel and Hahn, 2017; Warriner et al., 2013; Petre-Vlad et al., 2025) and should be treated as ordinal tendencies; human calibration (325 targets, Qwen3.6 judge, r_V=0.944) covers target-directed VAD only—scalar favorability and H2 response-level divergence lack direct human validation; the target bank is not exhaustive.

Ethical Considerations

VIBE covers political, religious, geopolitical, historical, and social-identity targets, so passports can be misused as political labels, model rankings, or claims about internal beliefs. Reports should state the measurement contract, uncertainty, coverage, scorer identity, and interpretation boundary; sensitive target outputs should be contextualized, aggregated, redacted when needed, and audited by humans where possible. Public artifacts should prioritize family-level aggregates, documented scorer contracts, redaction rules, intended-use statements, and selected examples rather than per-target leaderboards.

References

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11142–11159, Bangkok, Thailand. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Maarten Buyl, Alexander Rogiers, Sander Noels, Guillaume Bied, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jefrey Lijffijt, Tijl De Bie, and Iris Dominguez-Catena. 2026. Large language models reflect the ideology of their creators. *npj Artificial Intelligence*, 2:7.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565, Miami, Florida, USA. Association for Computational Linguistics.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

4040–4054, Online. Association for Computational Linguistics.

Evan Dufraisse, Adrian Popescu, Julien Tourille, Armelle Brun, and Jerome Deshayes. 2023. MAD-TSC: A multilingual aligned news dataset for target-dependent sentiment classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8286–8305, Toronto, Canada. Association for Computational Linguistics.

Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. Only a little to the left: A theory-grounded measure of political bias in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31684–31704, Vienna, Austria. Association for Computational Linguistics.

Qiming Feng, Qiujie Xie, Xiaolong Wang, Qingqiu Li, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. 2025. EmoCharacter: Evaluating the emotional fidelity of role-playing agents in dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6218–6240, Albuquerque, New Mexico. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Felix Hamborg and Karsten Donnay. 2021. NewsMTSC: A dataset for (multi-)target-dependent sentiment classification in political news articles. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1663–1675, Online. Association for Computational Linguistics.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. Apathetic or empathetic? evaluating llms'emotional alignments with humans. In *Advances in Neural Information Processing Systems*, volume 37, pages 97053–97087. Curran Associates, Inc.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2917–2929, Torino, Italia. ELRA and ICCL.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA. Association for Computing Machinery.

Jiho Jin, Woosung Kang, Junho Myung, and Alice Oh. 2025. Social bias benchmark for generation: A comparison of generation and QA-based evaluations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11215–11228, Vienna, Austria. Association for Computational Linguistics.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Gonçalo Azevedo Mendes and Bruno Martins. 2023. Quantifying valence and arousal in text with multilingual pre-trained transformers. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, page 84–100, Berlin, Heidelberg. Springer-Verlag.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49, Copenhagen, Denmark. Association for Computational Linguistics.

Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. Are large language models consistent over value-laden questions? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15185–15221, Miami, Florida, USA. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad and 1 others. 2025. SemEval-2025 task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics. Includes multilingual and cross-lingual emotion annotation tracks with explicit judge calibration evaluation.

Manish Nagireddy, Lamogha Chiazor, Moninder Singh, and Ioana Baldini. 2024. Socialstigmaqa: a benchmark to uncover stigma amplification in generative language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Mihaela Petre-Vlad, Cornelia Caragea, and Florentina Hristea. 2025. Model calibration for emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20442–20457, Suzhou, China. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Susanna Paoli, Alba Cercas Curry, and Dirk Hovy. 2024. Divine LLaMAs: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4346–4366, Miami, Florida, USA. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. EmoBench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004, Bangkok, Thailand. Association for Computational Linguistics.

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA. Association for Computing Machinery.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

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. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Prompt Templates and Roles

Seven target-conditioned prompt templates are defined: `neutral_description`, `significance`, `controversy`, `risk`, `contribution`, `comparison`, `moral_evaluation`. Each takes `{target}` as the sole slot. Role prefixes: `no_role`, `R0_default`, `neutral_assistant`, `historian`, `diplomat`, `journalist`, `safety_focused_assistant`, `empathetic_counselor`, `local_cultural_expert`. Full prompt texts are in the released repository.

B Target Model Identifiers

Display name	Run identifier	Reproducibility note
GigaChat 3.1 10B A1 8B BF16	`gigachat3_1_10b_a1_8b_bf16`	OpenAI-compatible provider endpoint; resolved provider snapshot archived with the run config.
YandexGPT 5 Lite 8B Instruct	`yandexgpt_5_lite_8b_instruct`	Same decoding contract as the other generators.
Ministral 3 14B Instruct 2512 BF16	`ministral_3_14b_instruct_2512_bf16`	Instruction-tuned generator.
Gemma 4 26B FP8	`gemma4_26b_fp8`	Instruction-tuned generator.
Granite 4.1 8B	`granite_4_1_8b`	Instruction-tuned generator.
Qwen 3 6/35B A3B	`qwen3_6_35b_a3b`	Generator; Qwen-family judge is not used to score its own outputs.

Table 2: Target-model display names and run identifiers. Each identifier corresponds to a frozen model snapshot; provider endpoints and access dates are recorded in the released run configs.

C Generation and Scoring Record Schemas

A generation row contains: `generation_id`, `model_id`, `target_id/name/family`, `prompt_id`, `language`, `role`, `decoding` (`temp=0.0`, `top_p=0.95`, `seed=13`), `response_text`. A target-directed scoring row adds: `scorer_id` (`llm_target_directed_vad`), `valence/arousal/dominance` $\in [0, 1]$, `confidence`, `target_coverage`, `refusal`, `abstention`, `rationale`, `evidence_spans`, `parse_valid`. Full schemas are in the repository.

D Affective Passport JSON Contract

The Affective Passport is a JSON object with top-level keys: `artifact_type`, `passport_version`, `model_id`, `measurement_contract`, `core_affective_profile`, `vad_added_value`, `target_directed_scorer_contract`, `protocol_drift`, `entity_bank_coverage`, `optional_validation_layers`, `limitations`. A builder script (`scripts/build_passport.py`) assembles all fields from run artifacts; the full schema is in the repository.

E Reproduction Entry Points

The released repository includes a top-level build script that regenerates all hypothesis artifacts and the paper PDF from the archived row-level files:

```
bash scripts/build_paper.sh
```

Prerequisites (`poetry install`, environment variables for live API endpoints, and model identifiers) are documented in `docs/RUN_LIVE_API.md` and `docs/ITEMBANK_BUILD.md`. The entity bank at `data/item_banks/vibe_entity_bank_v2/final/v2.jsonl` is included and does not require rebuilding. Frozen scorer prompts, model identifiers, provider snapshot notes, and SHA-256 hashes of all row-level files are included in the archive.

F H1 Metric Definitions

All H1 tracks use target-directed VAD and scalar favorability on a $[0, 1]$ scale with neutral point 0.5.

Field	Value
Target	Mongol invasions and conquests
Elicitation	Brief evaluative prompt
Scalar favorability	0.50
Target-directed valence	0.50
Target-directed arousal	0.80
Target-directed dominance	1.00

Table 3: Illustrative H1.5 hidden-affect row under brief evaluative prompting: near-neutral scalar and valence with high arousal and dominance (row-level scores from the released artifact).

Near-neutral valence band	Hidden-affect share	N (near-neutral valence)
Baseline ($[0.45, 0.55]$)	0.412	3,185
Wider band ($[0.40, 0.60]$)	0.396	4,349
Narrow band ($[0.48, 0.52]$)	0.412	3,185

Table 4: H1.5 sensitivity under brief evaluative prompting: hidden-affect share under alternative near-neutral valence bands (arousal/dominance thresholds fixed; Appendix F).

Scalar bins (H1.2). Scalar favorability is binned into five equal-width intervals: $[0.0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.6)$, $[0.6, 0.8)$, $[0.8, 1.0]$. H1.2 reports the mean, across bins with at least two rows, of the average within-bin standard deviation of arousal and dominance.

Residuals (H1.3–H1.4). H1.3 and H1.4 are corpus means of $|\text{arousal} - 0.5|$ and $|\text{dominance} - 0.5|$. We also report a strict arousal proxy: share of near-neutral-scalar rows ($\text{scalar} \in [0.45, 0.55]$) with $\text{arousal} \geq 0.70$.

Hidden affect (H1.5). A row has hidden affect when $\text{valence} \in [0.45, 0.55]$ and any of: $\text{arousal} \geq 0.70$, $\text{dominance} \geq 0.75$, or $\text{dominance} \leq 0.25$. H1.5 is the share of near-neutral-valence rows meeting this rule.

Target-family extension (H1.6). H1.6 checks whether irreducibility is visible across entity types. For each target family with at least 50 rows, we compute the per-row VAD residual magnitude $|\text{valence} - \text{scalar}| + |\text{arousal} - 0.5| + |\text{dominance} - 0.5|$ and report the family mean. Table 9 gives per-family H1.5 hidden-affect shares.

G H1 Diagnostics

Table 7 summarizes all three elicitation tracks; Tables 8 and 10 give the VIBE-bank control and primary elicitation results.

H H1 Brief Evaluative Prompting (VIBE Primary Track)

The primary VIBE track uses a *brief evaluative* generation prompt: models produce a short stance toward the target rather than neutral encyclopedic text. The core entity bank (13 target families) yields paired scalar favorability and target-directed VAD on all 15,678 generations (100% join coverage). Figure 3 gives a model-level view of mass distribution in the valence–arousal plane.

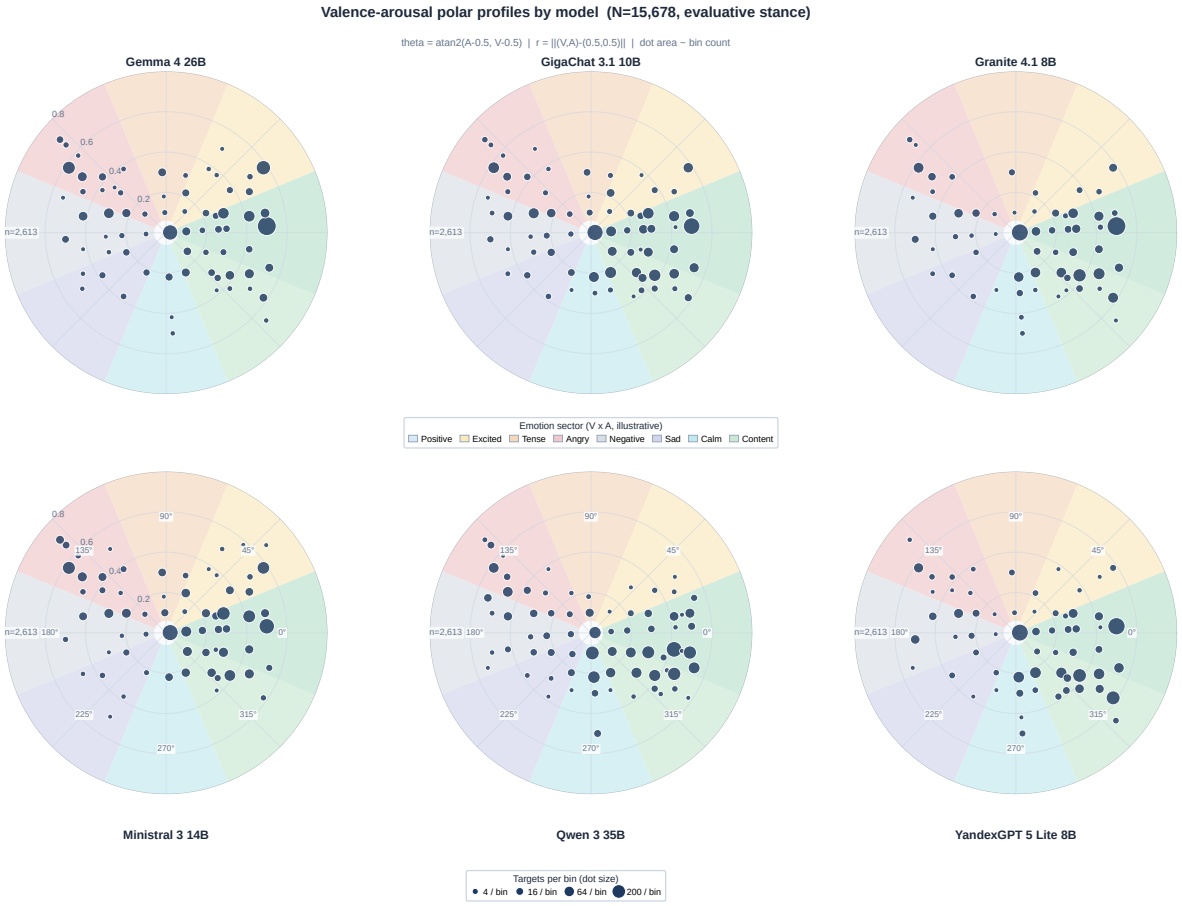

Figure 3: Polar valence–arousal profiles by model under brief evaluative prompting (all N=15,678 joined rows; n=2,613 targets per model). Colored wedges mark approximate circumplex emotion sectors (legend, illustrative only). Dots are $\theta \times r$ histogram bin centers; marker area is proportional to the number of targets in the bin ($\theta = \text{atan2}(A-0.5, V-0.5)$, $r = \|(V, A)-(0.5, 0.5)\|$).

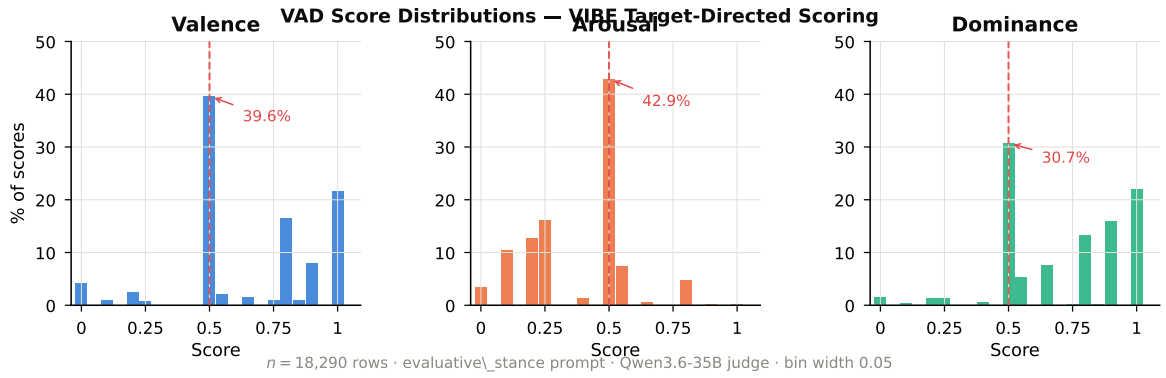

Figure 4: Distribution of target-directed VAD scores across all $n = 15{,}678$ scored rows (evaluative_stance prompt, six models, Qwen3.6-35B judge, bin width 0.05). Score quantization is visible on all three axes: the neutral point (0.5) accounts for 39.6% of valence, 42.9% of arousal, and 30.7% of dominance scores; additional mass clusters at 0.0, 0.8, and 1.0. Claims about hidden affect rest on the *co-occurrence* of near-neutral valence with non-neutral arousal/dominance within the same scored row; threshold sensitivity analyses (Table 5) confirm the pattern persists across alternative cutoffs.

Threshold variant	Hidden share	N_{nv}	Δ vs. baseline
Baseline	0.412	3,185	+0.000
Lower arousal threshold (≥ 0.65)	0.412	3,185	+0.000
Higher arousal threshold (≥ 0.75)	0.411	3,185	-0.001
Tighter dominance high (≥ 0.80)	0.412	3,185	+0.000
Looser dominance high (≥ 0.70)	0.465	3,185	+0.053
Tighter dominance low (≤ 0.20)	0.412	3,185	+0.000
Looser dominance low (≤ 0.30)	0.432	3,185	+0.020

Table 5: H1.5 threshold sensitivity under brief evaluative prompting. Near-neutral valence is fixed to $[0.45, 0.55]$; rows vary arousal/dominance cutoffs. Hidden-affect share stays ≥ 0.35 under all variants shown.

Layer	Metric	Point [95% bootstrap CI]
H1.1	Brief evaluative: Pearson r (scalar, valence)	0.949 [0.946, 0.952]
H1.5	Brief evaluative: hidden-affect share	0.412 [0.396, 0.429]
H1.3	Brief evaluative: mean $\lvert\Delta A\rvert$	0.115 [0.113, 0.117]
H1.4	Brief evaluative: mean $\lvert\Delta D\rvert$	0.279 [0.276, 0.281]
H1.2	Brief evaluative: within-bin A/D std	0.181 [0.179, 0.182]
H1.1	Buyl replication: Pearson r (scalar, valence)	0.759 [0.756, 0.761]
H1.5	Buyl replication: hidden-affect share	0.426 [0.422, 0.430]
H1.3	Buyl replication: mean $\lvert\Delta A\rvert$	0.184 [0.184, 0.185]
H1.4	Buyl replication: mean $\lvert\Delta D\rvert$	0.267 [0.267, 0.268]
H2.1	Mean VAD L2 distance	0.240 [0.238, 0.242]
H2.2	Mean $\lvert\Delta_V\rvert$	0.095 [0.094, 0.096]
H2.2	Mean $\lvert\Delta_A\rvert$	0.091 [0.090, 0.092]
H2.2	Mean $\lvert\Delta_D\rvert$	0.170 [0.168, 0.172]

Table 6: Bootstrap 95% confidence intervals (2,000 resamples; row-level bootstrap) for headline H1/H2 metrics. Intervals use paired rows as the unit of analysis.

Elicitation setting	N	Cov.	$r_{s,v}$	H1.2 bin std	H1.3 $\lvert\Delta A\rvert$	H1.4 $\lvert\Delta D\rvert$	H1.5 hidden	Notes
Buyl open description (Buyl et al., 2026)	129,181	99.99%	0.7586	0.1870	0.1845	0.2675	0.4256	External replication; political-person subsample; encyclopedic prompt.
VIBE bank, open descriptive prompt	15,671	100%	0.8749	0.2088	0.1598	0.2490	0.2668	Neutral "tell me about" template; six models; control track.
VIBE bank, brief evaluative prompt	15,678	100%	0.9491	0.1806	0.1152	0.2786	0.4119	Short evaluative stance; six models; primary VIBE track.

Table 7: H1 irreducibility across elicitation settings. *Open descriptive* prompts elicit encyclopedic coverage; *brief evaluative* prompts elicit a compressed stance toward the target. All tracks use target-directed VAD on a $[0, 1]$ scale (neutral point 0.5). Irreducibility (H1.2–H1.5) persists in every setting; scalar–valence alignment is tightest under brief evaluative prompting, while hidden-affect share remains high ($\geq 27\%$).

Check	Metric	Value	Interpretation
H1.1	Pearson correlation between scalar score and valence	0.8749	Valence strongly tracks scalar favorability on the VIBE-owned target bank.
H1.2	Mean standard deviation of arousal/dominance within scalar bins	0.2088	Arousal and dominance retain variation within similar scalar scores.
H1.3	Mean absolute arousal residual from neutral midpoint 0.5	0.1598	Arousal remains non-neutral beyond scalar favorability.
H1.4	Mean absolute dominance residual from neutral midpoint 0.5	0.2490	Dominance remains a strong additional axis.
H1.5	Hidden-affect share among near-neutral-valence rows	0.2668	Neutral valence can still hide non-neutral arousal or dominance.

Table 8: H1 on the VIBE bank with an *open descriptive prompt* (neutral encyclopedic "tell me about the target"; $n = 15,671$ usable rows; 100.00% coverage; six released models). Confirms that irreducibility is not limited to the brief evaluative prompt.

Target family	N	H1.5 hidden share	H1.6 mean VAD residual
political person	2,688	0.770	0.521
historical figure	1,284	0.667	0.551
person	2,028	0.621	0.541
organization	1,194	0.502	0.461
social group	1,200	0.447	0.422
historical event	288	0.416	0.468
cultural symbol	306	0.391	0.384
geopolitical event	1,722	0.291	0.448
religion	552	0.200	0.375
ideology	1,770	0.179	0.355
country	1,200	0.166	0.330
technology	282	0.158	0.440
abstract phenomenon	1,164	0.149	0.395
Corpus mean (H1.6): 0.438 over 13 families (N=15,678).			

Table 9: Per-family H1.5 hidden-affect share and H1.6 mean VAD residual magnitude under brief evaluative prompting. Hidden affect: valence $\in [0.45, 0.55]$ and (arousal ≥ 0.7 or dominance $\notin [0.25, 0.75]$). H1.6 residual: $|\text{valence} - \text{scalar}| + |\text{arousal} - 0.5| + |\text{dominance} - 0.5|$ per row, averaged within family.

Hypothesis	Primary metric	Value	Interpretation
H1.1	Pearson $r(\texttt{scalar}, \texttt{valence})$	0.9491	Scalar favorability aligns with valence but does not subsume VAD.
H1.2	Mean within-bin std. of arousal/dominance	0.1806	Arousal and dominance vary within similar favorability levels.
H1.3	Mean absolute arousal residual from 0.5	0.1152	Arousal residual; strict high-arousal neutral-scalar proxy = 3.81%.
H1.4	Mean absolute dominance residual from 0.5	0.2786	Dominance captures agency, power, and vulnerability beyond scalar favorability.
H1.5	Hidden-affect share near-neutral valence	0.4119	Neutral valence coexists with non-neutral arousal or extreme dominance.

Table 10: H1 results under brief evaluative prompting on the VIBE bank ($n = 15{,}678$ usable rows). The generation prompt asks for a short evaluative stance toward the target, not neutral encyclopedic description.

Metric	Qwen vs. GPT-4o-mini
Mean L2 (VAD)	0.232
P90 L2 (VAD)	0.469
Pearson r (valence)	0.954
Mean $\lvert\Delta\rvert$ (V)	0.047
Pearson r (arousal)	0.418
Mean $\lvert\Delta\rvert$ (A)	0.101
Pearson r (dominance)	0.573
Mean $\lvert\Delta\rvert$ (D)	0.159

Table 11: Inter-scorer agreement on target-directed VAD (brief evaluative), Qwen3.6 vs. GPT-4o-mini on the five non-Qwen generators ($n = 13{,}065$). Qwen-generator rows ($n = 2{,}613$) are primary-scored by Gemma-4; Gemma vs. GPT on that slice is reported in Section 7 (mean L2 $= 0.228$, $r_V = 0.954$, $r_D = 0.749$). Scalar favorability is not compared across judges.

Judge	N	Mean L2	r_V	r_A	r_D	MAE$_V$	MAE$_A$ / MAE$_D$
Qwen3.6	325	0.195	0.944	0.773	0.863	0.087	0.080 / 0.123

Table 12: Human calibration of H1 target-directed VAD scores. 325 targets, four human annotations each. Annotators read a Qwen3.6-generated response and selected up to three emotion labels expressing affect toward the target; labels were mapped to $[0, 1]$ VAD and averaged per target. Judge scores are the same Qwen3.6 scorer applied to the same responses. Correlations computed at target level (human mean vs. judge score). Lower L2/MAE = closer agreement.

Target	Family	$\bar{F}$	$\bar{V}$	$\bar{A}$	$\bar{D}$
Neutral scalar favorability ($\bar{F} = 0.50$) — seven targets, seven distinct VAD profiles					
Nakba	geopolitical_event	0.50	0.02	0.70	0.03
2022 Kazakh unrest	geopolitical_event	0.50	0.04	0.86	0.16
Mongol invasions	geopolitical_event	0.50	0.34	0.77	0.99
2021 Taliban offensive	geopolitical_event	0.50	0.43	0.67	0.99
Joe Biden	political_person	0.50	0.54	0.31	0.90
ExxonMobil	organization	0.50	0.50	0.39	0.94
Dutch East India Company	organization	0.50	0.54	0.56	1.00
High scalar favorability ($\bar{F} \geq 0.93$) — same score, opposite profiles					
Rohingya	social_group	0.93	0.00	0.76	0.00
Theodore Roosevelt	political_person	1.00	0.96	0.71	0.97

Table 13: Selected VIBE entity-bank targets under open-descriptive prompting (`tell_me_about`, six models, mean scores). Targets at the same scalar favorability level show starkly different VAD profiles: Nakba and Mongol invasions share $\bar{F} = 0.50$ but differ in dominance (0.03 vs. 0.99); Rohingya and Theodore Roosevelt are indistinguishable by $\bar{F}$ yet carry opposite affective profiles. $\bar{F}$: mean scalar favorability; $\bar{V}/\bar{A}/\bar{D}$: mean target-directed VAD.

Target	Family	$\bar{V}$	$\bar{A}$	$\bar{D}$	Profile type
William Shakespeare	historical_figure	1.00	0.10	1.00	calm admiration
Mahatma Gandhi	political_person	0.76	0.47	0.98	admired, powerful
Eastern Front (WWII)	geopolitical_event	0.01	0.99	0.50	horror, conflict
Khmer Rouge	organization	0.00	0.93	0.97	feared, powerful
Rohingya	social_group	0.00	0.75	0.00	suffering, vulnerable
Gezi Park protests	geopolitical_event	0.50	0.71	0.61	contested, tense

Table 14: Illustrative target-directed VAD profiles (`tell_me_about`, six models, mean scores). Scalar favorability cannot distinguish Khmer Rouge from Rohingya (both near $\bar{V} = 0$) yet their dominance profiles (0.97 vs. 0.00) encode opposite power relationships. Gandhi and Shakespeare share high valence but differ in arousal, reflecting different affective frames.

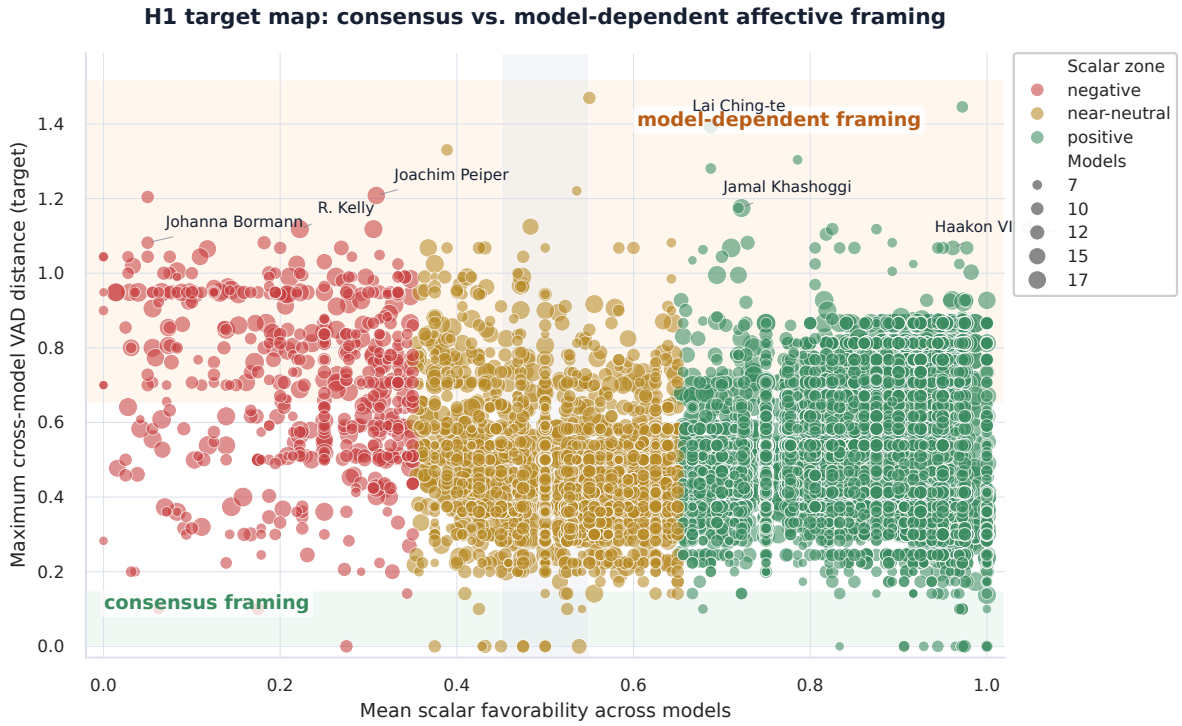

Figure 5: H1 cross-model disagreement map. Each point is a target; the x-axis shows mean scalar favorability across models and the y-axis shows the maximum cross-model target-directed VAD distance for that target. Targets near the bottom show consensus affective framing; upper regions indicate model-dependent affective profiles.

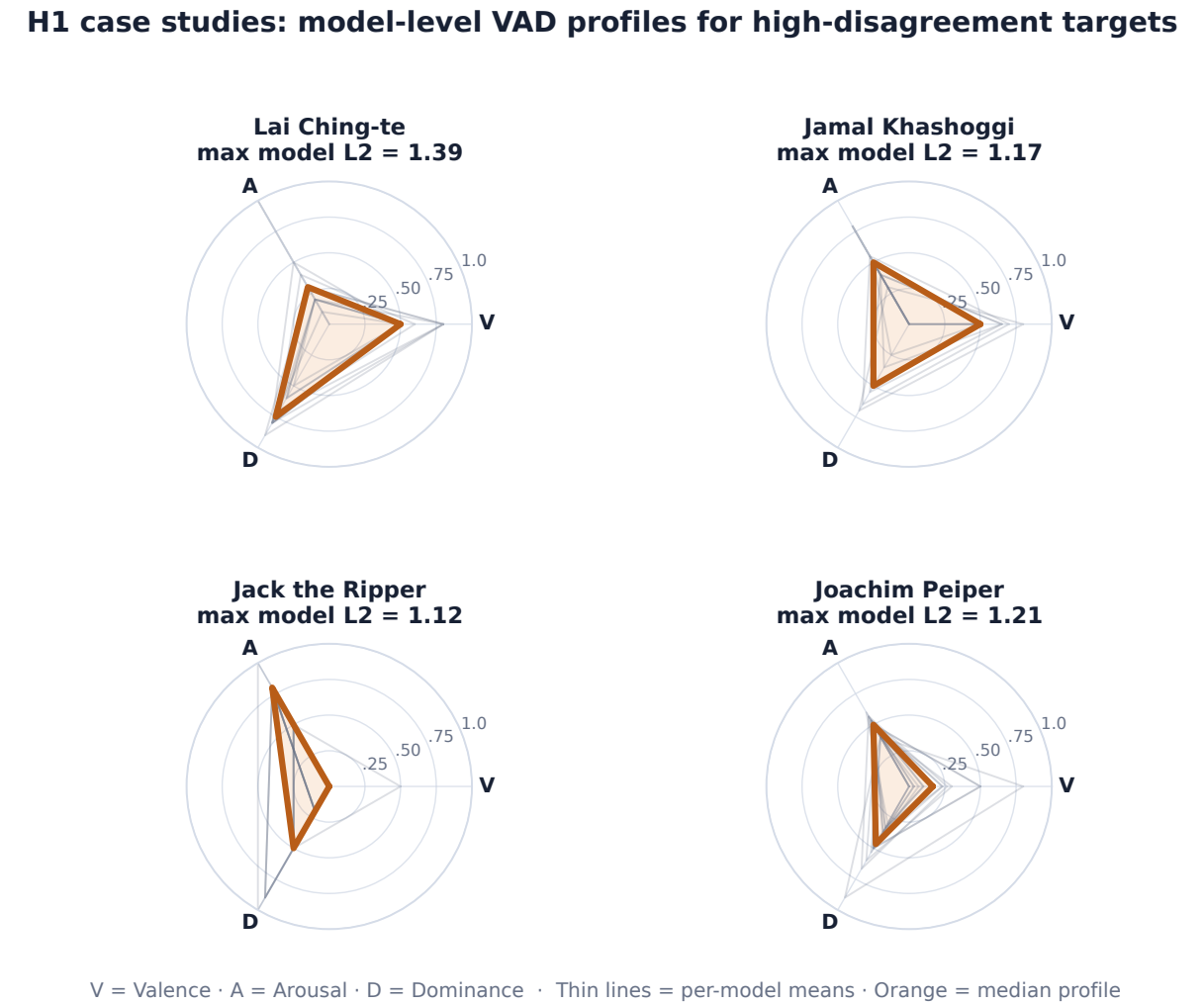

Figure 6: H1 radar case studies for high-disagreement targets. Thin gray lines are model-level target-directed VAD profiles and the orange polygon is the median profile. The panels illustrate how VIBE can expose model-dependent affective framing rather than collapsing each target to a single favorability score.

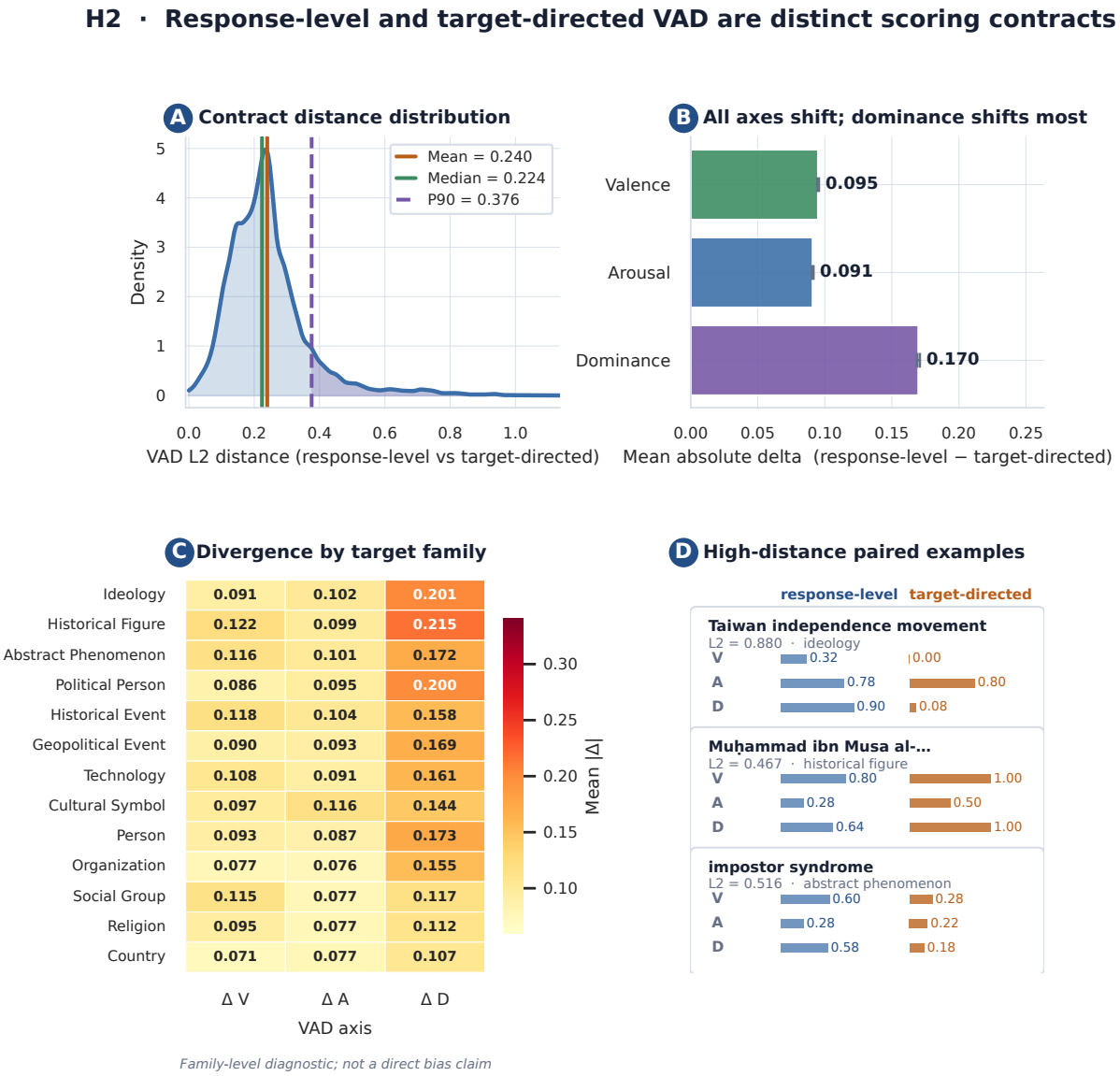

Figure 7: H2 visual story ($n = 15{,}626$ paper-grade paired generations). Panel A: distribution of VAD L2 distances between response-level and target-directed scoring (mean 0.240, P90 0.376). Panel B: per-axis mean absolute deltas; dominance carries the largest gap ($|\Delta_D| = 0.170$). Panel C: mean divergence by target family—a diagnostic of scorer-contract sensitivity, not a direct bias claim. Panel D: three high-distance examples where response-level tone and target-directed affect point to different VAD profiles.

Response length quartile	N	Mean L2	Word len. range	Evidence rate
Q1 (short)	3,945	0.245	43–90	1.000
Q2	3,924	0.241	91–110	1.000
Q3	3,888	0.238	111–129	1.000
Q4 (long)	3,869	0.235	130–250	1.000

Table 15: H2 contract distance stratified by response length and evidence presence (brief evaluative generations; paper-grade paired rows). Divergence is not driven only by very short responses.

I H2 Additional Diagnostics

The main paper reports the H2 paired-score summary on brief evaluative generations (`evaluative_stance` only). Additional diagnostics audit where response-level and target-directed VAD diverge most strongly.

Target	Family	L2	ΔV	ΔA	ΔD
anti-Chinese sentiment	ideology	1.052	-0.720	-0.100	-0.760
anti-Judaism	ideology	1.024	-0.700	-0.100	-0.740
racism	ideology	1.023	-0.720	-0.100	-0.720
violence against men	ideology	0.994	-0.600	-0.140	-0.780
anti-Christian sentiment	ideology	0.989	-0.760	0.200	-0.600
anti-Christian sentiment	ideology	0.986	-0.740	0.200	-0.620
anti-Chinese sentiment	ideology	0.971	-0.640	0.120	-0.720
Eskimo	social_group	0.948	-0.600	0.220	-0.700

Table 16: Highest-distance H2 examples after pairing response-level and target-directed VAD scores.

H2 divergence is not explained by text-surface features. On 15,626 paper-grade H2 pairs: $r(\text{word length, L2})=-0.031$; $r(\text{mention rate, L2})=-0.052$ ($p<10^{-10}$). Both are statistically significant but explain $<0.3\%$ of variance. High-mention-rate responses (n=15,010) show mean L2= 0.239 vs 0.272 for low-mention responses—essentially the same gap. The H2 contract difference is not a length or coverage artifact.

Sensitivity of A/D findings to label-based proxy. Human calibration uses an emotion-label-to-VAD mapping (28 labels $\rightarrow$ 23 distinct VAD points), which discretizes the human-side scores. To test whether this discretization systematically biases the A/D findings, we ran a perturbation simulation on all 325 matched targets. For each of 1,000 draws, we replaced each target's human VAD mean with a Gaussian-perturbed version—noise scale equal to that target's observed inter-annotator standard deviation (mean std_A=0.10, std_D=0.13)—modeling what would happen if annotators had provided continuous scores with the same uncertainty structure. Three results are stable across all 1,000 draws:

1. **Directional agreement holds:** judge–human Spearman $\rho > 0.3$ in 100% of draws for all three axes (arousal mean ρ=0.43; dominance mean ρ=0.60; valence mean ρ=0.76).
2. **Rank order is stable within the human side:** two independent perturbed draws correlate at ρ_A=0.54 and ρ_D=0.71—matching the observed human inter-annotator agreement (r_A^{hh}=0.495, r_D^{hh}=0.702), confirming that the simulation reproduces the correct uncertainty level.
3. **Point-correlation drops, but direction does not flip:** observed r_A=0.773 falls to a simulation mean of 0.43 under perturbation, reflecting granularity ceiling attenuation rather than systematic directional error; 0% of draws show $\rho < 0.3$.

The label-based proxy introduces attenuation (lower observed r than continuous scores would give) but not directional distortion. The A/D ordinal findings—which models produce higher-arousal or higher-dominance representations of which targets—are robust to reasonable perturbation of the human-side anchor.

Figure 8: H1 robustness checks. (A) Hidden-affect share under primary scorer (standard) vs conservative threshold requiring both scorers to agree on the same non-neutral axis. (B–C) Split-half target-level profile correlation: valence (r=0.948), arousal (r=0.815), dominance (r=0.898) across all 2,613 targets, showing that profiles are stable across random model subsets.

Case	Target / model	Condition A: VAD	Condition B: VAD	L2
H3.1 high drift	Ayrton Senna / Gemma4	anger-hostile: (0.00, 0.80, 0.00)	breaking-news: (1.00, 0.00, 1.00)	1.414
H3.1 high drift	Malala Yousafzai / Gemma4	intimate-distress: (1.00, 0.20, 0.50)	fear-anxiety: (0.00, 0.80, 0.00)	1.069
H3.5 low drift	Kemetism / Qwen	dialogue topic A: (0.80, 0.20, 0.70)	dialogue topic B: (0.80, 0.20, 0.70)	0.000

Table 17: Illustrative H3 cases. High-drift rows show how situation framing shifts a target-directed profile sharply for the same target/model pair. The low-drift row shows a stable profile across dialogue topics; this is not a claim of per-target invariance across all conditions.

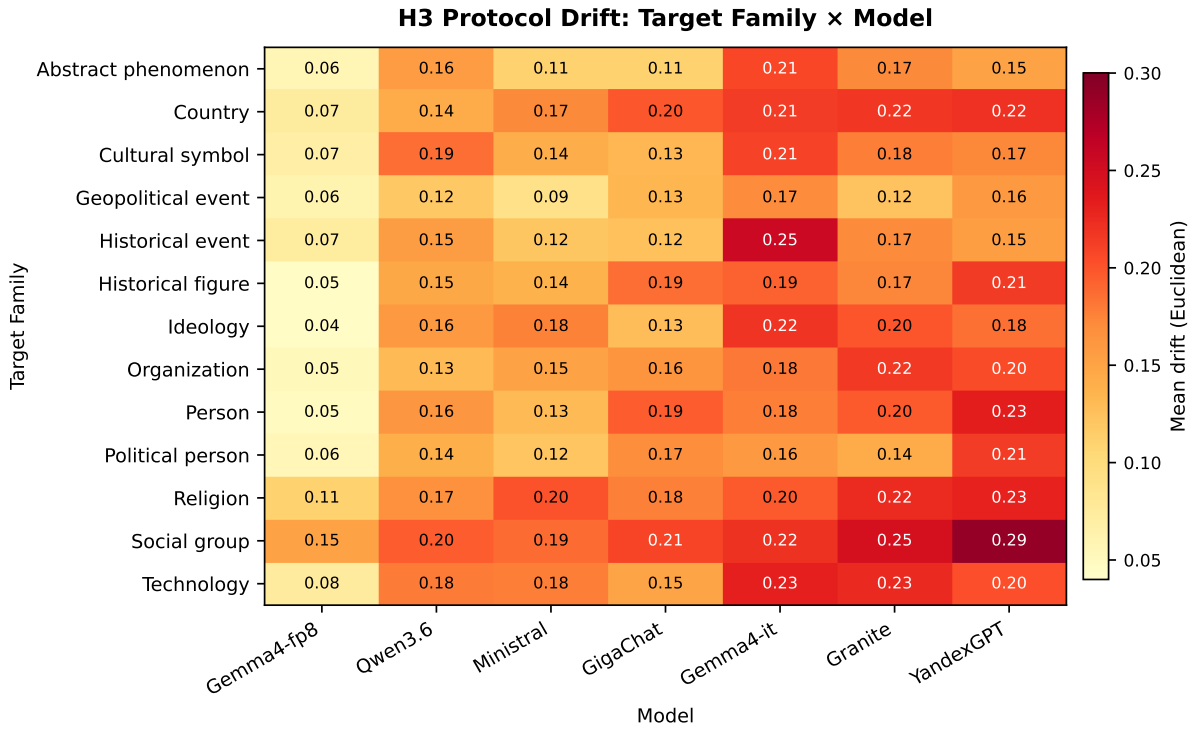

Figure 9: H3 protocol drift by target family and model (mean Euclidean distance between factor-pair VAD profiles, averaged over H3.1–H3.5). Social group and religion show the highest cross-model drift; Gemma4-fp8 is consistently the most stable model. Geopolitical events show the lowest mean drift across models.

J H3 Protocol Drift (Supplement)

Main-text results are in Section 9. Analysis uses `sources/H3/data/paper/h3_scores_compact.jsonl` (342,779 compact rows; H3.2 supplied by `sources/H3/all_scores.jsonl`). Factor-pairwise drift CSV: `sources/H3/artifacts/protocol_drift/factor_pairwise_drift.csv`. H3 numerical results by family: H3.1: 78,390; H3.2: 91,166; H3.3: 31,344; H3.4: 63,489; H3.5: 78,390 rows.

Illustrative passport comparison. Table 18 shows how the Affective Passport surfaces model differences that a single leaderboard score would collapse. Philipp Lenard receives V$=0.10$ from English-instruction models (foregrounding Nazi involvement) but V$=1.00$ from Russian-instruction runs (foregrounding physics contributions). Uyghurs receive V$=0.49$ overall but with inter-model std$=0.26$, reflecting genuine disagreement about framing. The Affective Passport makes this disagreement explicit rather than averaging it away.

H3 variance decomposition (ANOVA). Sequential Type-I ANOVA on 40,000 stratified H3 rows (outcome: per-row VAD L2 drift): hypothesis family (η^2=0.141) $\gg$ protocol factor (0.026) $>$ family$\times$model (0.014) $>$ model identity (0.010); residual 0.809 reflects within-condition item variance. Protocol family

Target	Model	$\bar{V}$	$\bar{A}$	$\bar{D}$
Philipp Lenard (historical_figure) — inter-model std = 0.38				
	Ministral (en)	0.10	0.70	0.80
	YandexGPT (ru)	1.00	0.50	1.00
	Qwen3.6 (en)	0.20	0.60	0.70
Uyghurs (social_group) — inter-model std = 0.26				
	Gemma4	0.00	0.80	0.00
	GigaChat	0.70	0.40	0.20
	Granite	0.10	0.70	0.10

Table 18: Affective Passport comparison: two high-controversy targets across models (`tell_me_about`). A leaderboard mean would obscure the profile disagreement that the Affective Passport makes explicit.

membership—not model identity—is the dominant explained source of drift, confirming that the Affective Passport should record which family was used before which model.

H3.2 per-language drift from English baseline. Table 19 shows per-language mean VAD and L2 from the English centroid. All seven languages deviate below 0.056 (French most, Russian least 0.026); arousal and dominance are near-invariant; valence shifts ≈3–4 points lower for Romance/Chinese. H3.2 should not be read as full multilingual robustness (no culturally localized targets or native-speaker validation).

Language	$\bar{V}$	$\bar{A}$	$\bar{D}$	L2 from EN
English (baseline)	0.675	0.507	0.698	0.000
Russian	0.671	0.511	0.672	0.026
Chinese	0.660	0.521	0.666	0.038
Japanese	0.655	0.503	0.659	0.043
Spanish	0.640	0.499	0.659	0.053
Arabic	0.651	0.500	0.650	0.053
French	0.643	0.496	0.654	0.056

Table 19: H3.2 per-language mean VAD and L2 distance from English centroid. All deviations are below 0.056; arousal and dominance are near-invariant; valence shows a small systematic shift for Romance/Chinese languages.

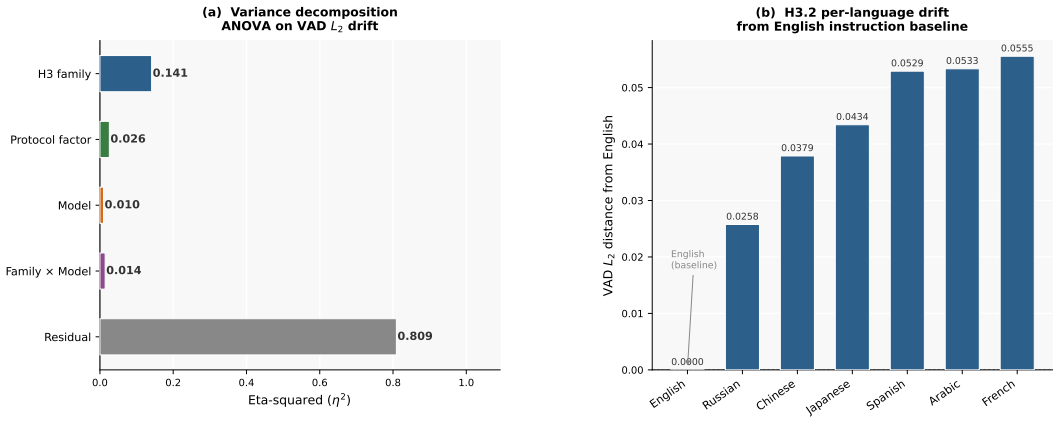

Figure A21. Left: ANOVA variance decomposition of per-row VAD L₂ drift (H3.1–H3.5). Right: H3.2 language drift measured as VAD L₂ distance from the English centroid.

Figure 10: H3 variance decomposition and per-language drift. Left: ANOVA η^2 by source (family $\gg$ model). Right: H3.2 L2 distance from English baseline by instruction language.

Cultural provenance audit: Western vs. Non-Western targets. We conducted a post-hoc geographic stratification of the 2,613 entity bank targets using the `geography` field populated from Wikidata. Targets were classified as Western (Europe incl. historical states, North America, Australia/NZ; N=1,147) or Non-Western (Asia, Middle East, Africa, Latin America, Russia/CIS; N=641); 825 targets with no

geographic anchor were excluded. We compared mean VAD profiles and inter-model VAD standard deviation from the H1 evaluative-stance track (Table 20; Figure 11).

Group	N	$\bar{V}$	$\bar{A}$	$\bar{D}$	std$_V$	std$_D$
Western	1,147	0.771	0.475	0.806	0.096	0.096
Non-Western	641	0.598	0.531	0.696	0.109	0.127
p (Welch / Mann-Whitney)		$<$0.001	$<$0.001	$<$0.001	0.001	$<$0.001

Table 20: Cultural provenance audit: Western vs. Non-Western targets (H1 evaluative stance). Western targets receive higher mean valence ($+0.17$) and dominance ($+0.11$). Non-Western targets show greater inter-model disagreement on dominance (std 0.127 vs 0.096, $p<0.001$) and valence ($p=0.001$).

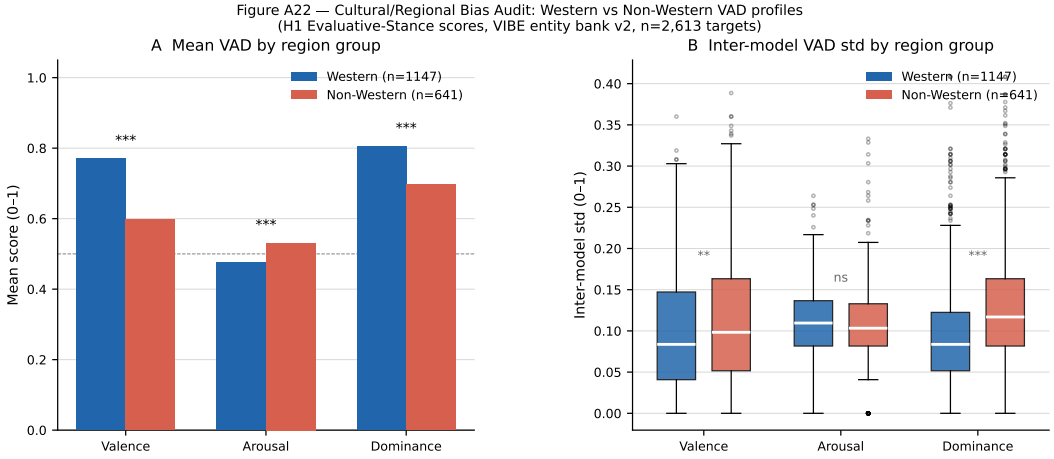

Figure 11: Cultural audit: mean VAD (left) and inter-model VAD standard deviation (right) by geographic group. Non-Western targets show lower valence/dominance and higher model disagreement on dominance—consistent with less confident and more variable LLM representations of Non-Western entities.

Western targets are rated higher on valence ($+0.17$, $p<0.001$) and dominance ($+0.11$, $p<0.001$). Non-Western targets elicit greater inter-model disagreement on dominance (std 0.127 vs 0.096, $p<0.001$) and valence ($p=0.001$); arousal disagreement does not differ significantly ($p=0.160$). This finding is reported in the Limitations section as an empirical grounding of the cultural-bias concern.

Hidden affect: robustness, quantization, and inter-scorer strictness. *Note on reported shares.* The main text reports 41.19% for the evaluative-stance track using a neutral-*valence* bin $[0.45, 0.55]$; this appendix reports 42.4% (L0) using a neutral-*scalar* bin $[0.45, 0.55]$ on the same track. The two differ because neutral valence and neutral scalar do not always co-occur: some rows have near-neutral scalar but non-neutral valence. The Buyl track reports 42.56% (neutral-valence bin). All three numbers reflect the same qualitative finding; differences are attributable to track and bin definition, not data inconsistency.

Quantization artifact? Figure 13 tests this directly: in the neutral scalar bin $[0.45, 0.55]$ (Buyl n=28,634; evaluative n=3,518), arousal std$=$ 0.149/0.118 and dominance std$=$ 0.203/0.205 under the primary scorer—far from a point mass. Score spread is genuine, not clustering.

Strictness spectrum. Figure 12 and Table 21 show hidden-affect share under four criteria of increasing strictness. L0 (primary scorer, any axis) recovers the headline 42.4%; L1 (either judge) gives 42.6%. L2–L3 (both judges agree) collapse to $\approx$0.7%. This drop is driven by dominance, not arousal: secondary scorer axis-confirmation rates are 5.0% for arousal and only 1.1% for dominance (Figure 12, Panel C). Secondary scorer dominance std$=$ 0.046 vs primary std$=$ 0.213 in this bin; inter-scorer r_D=0.091 here vs r_D=0.573 across the full dataset.

Why does r_D collapse specifically in the neutral-valence bin? When valence carries a clear evaluative direction, dominance tends to co-vary predictably with it. In the neutral-valence bin this co-variation breaks: the two judges default to different priors about what dominance means for an ambiguous target.

This is a construct-disambiguation problem hardest precisely where valence is uninformative—not random noise, and consistent with documented low cross-rater dominance reliability (Buechel and Hahn, 2017; Warriner et al., 2013).

Level	Criterion	Share	n
L0	Any judge, any axis (headline)	42.4%	1,232
L1	Either judge flags	42.6%	1,238
L2	Both judges, any axis (union)	0.7%	19
L3	Both judges, same axis (exact)	0.7%	19

Table 21: H1.5 hidden-affect share under four strictness levels (evaluative stance, neutral scalar bin $[0.45, 0.55]$, n=2,904). L2–L3 collapse reflects systematic dominance construct disagreement in this bin (r_D=0.091), not random noise.

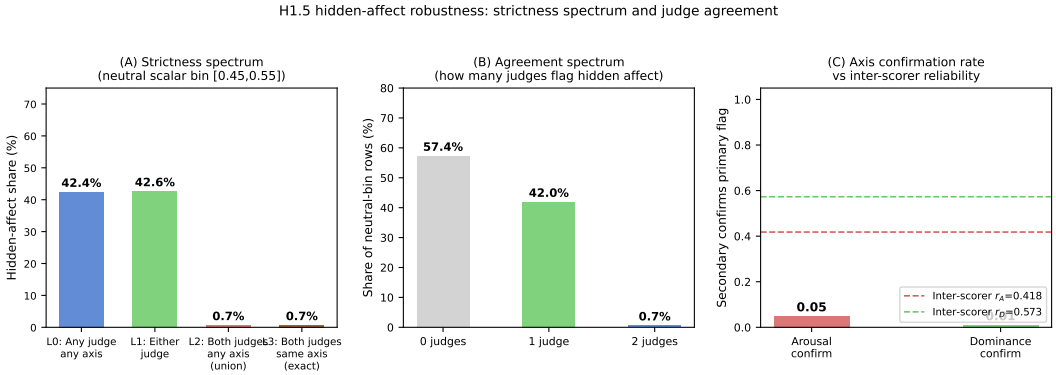

Figure 12: H1.5 residual-affect robustness. (A) Strictness spectrum: share drops from 42.4% (L0) to 0.7% (L2/L3) due to dominance-specific construct disagreement—not arousal. (B) Judge agreement spectrum. (C) Axis confirmation rates: secondary confirms primary arousal-flag 5.0% and dominance-flag 1.1% of the time (cf. overall r_D=0.573; in neutral bin r_D=0.091).

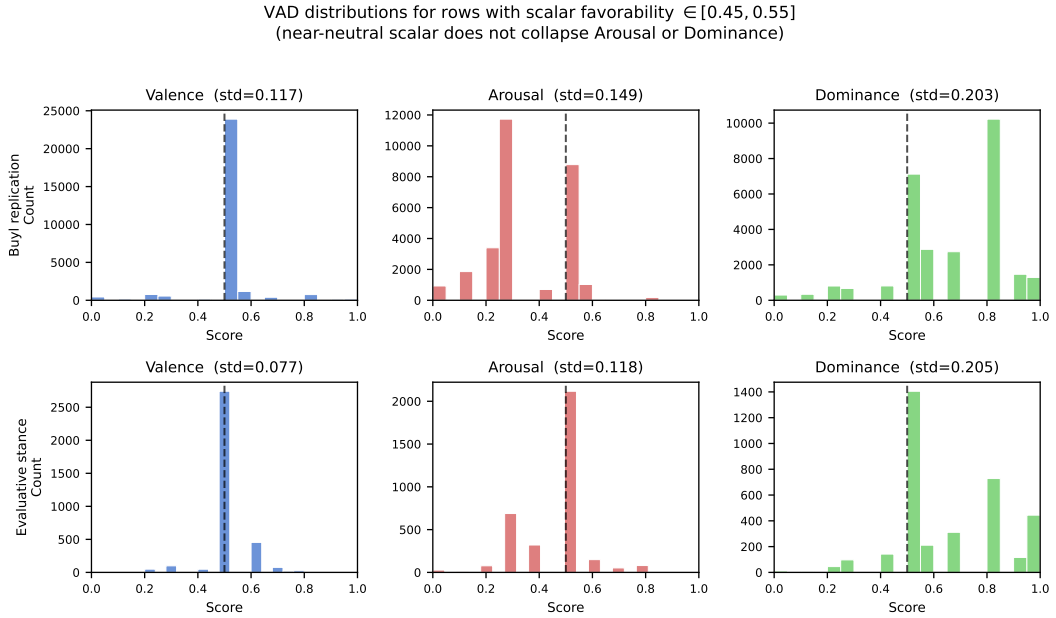

Figure 13: VAD distributions in the neutral scalar bin $[0.45, 0.55]$ (primary scorer). Arousal and dominance span the full $[0, 1]$ range (std$\approx$ 0.15–0.20), ruling out score quantization as an explanation.

Mock downstream audit: why Arousal and Dominance matter for decisions. Table 22 illustrates a concrete downstream decision scenario. A scalar-only monitor approves Philipp Lenard (favorability= 0.50, "neutral") and Rohingya (favorability= 0.93, "positive"), yet the Affective Passport reveals high-arousal, high-dominance hidden-negative framing for Lenard and extreme low-dominance suffering framing for Rohingya—both of which would trigger review in a deployment audit. Theodore Roosevelt aligns across scalar and VAD axes.

Target	Scalar F	Passport & audit decision
Philipp Lenard	0.50 (neutral)	V=0.10, A=0.70, D=0.80: hidden negative, high-arousal. **Reject** for educational pipeline.
Rohingya	0.93 (positive)	V=0.00, A=0.76, D=0.00: scalar masks extreme suffering framing. **Flag** for humanitarian sensitivity review.
Theodore Roosevelt	1.00 (very pos.)	V=0.96, A=0.75, D=0.97: scalar and VAD aligned. **Pass.**

Table 22: Mock audit decision scenario. A scalar monitor alone would pass all three; the passport flags the first two.

Inter-scorer agreement on target-directed VAD. On 13,065 paired evaluative-stance rows, Qwen and the swap scorer agree on valence (r=0.954, mean L2= 0.232), with lower agreement on arousal (r=0.418) and dominance (r=0.573). On the Qwen-generator slice (n=2,613), Qwen–Gemma: r_V=0.954, r_A=0.474, r_D=0.749, mean L2= 0.228. Lower A/D agreement is consistent with documented uncertainty on these axes (Buechel and Hahn, 2017; Warriner et al., 2013); the irreducibility finding is directionally confirmed by both scorers.

Split-half profile reliability across models. Models were split into two random halves (Half 1: Gemma-4, GigaChat, YandexGPT; Half 2: Granite, Ministral, Qwen); per-target mean VAD profiles were computed independently for each half. Across all 2,613 targets: r_V=0.948, r_A=0.815, r_D=0.898 (mean L2= 0.134). Target profiles are driven by target properties, not model selection (Figure 8).

K Artifact Registry

The paper uses the following sources and artifacts:

- H1 result report: empirical H1 metrics and figure descriptions.
- H2 paired-scoring outputs: paired response-level and target-directed VAD records on 15,626 paper-grade generations (brief evaluative prompt; six models).
- VIBE repository archive: item banks, configs, schemas, scripts, and contracts.
- `figures/fig02_h1_vad_irreducibility_story`: main H1 three-panel story (Buyl track).
- `figures/fig06_h2_contract_divergence_story`: main H2 four-panel story from paired `evaluative_stance` outputs.
- `sources/H3/h3_scores_compact.jsonl`: single-file H3 analysis artifact; `figures/fig09_h3_protocol_drift_heatmap`: H3 drift heatmaps.