Quantitative Introspection in Language Models: Tracking Emotive States Across Conversation

Anonymous authors
Paper under double-blind review

Abstract

Tracking the internal states of large language models across conversations is important for safety, interpretability, and model welfare, yet current methods are limited. Linear probes and other white-box methods compress high-dimensional representations imperfectly and are harder to apply with increasing model size. Taking inspiration from human psychology, where numeric self-report is a widely used tool for tracking internal states, we ask whether LLMs' own numeric self-reports can track probe-defined emotive states over time. We study four concept pairs (wellbeing, interest, focus, and impulsivity) in 40 ten-turn conversations, operationalizing introspection as the causal informational coupling between a model's self-report and a concept-matched probe-defined internal state. We find that greedy-decoded self-reports collapse outputs to a few uninformative values, but introspective capacity can be unmasked by calculating logit-based self-reports. This metric tracks interpretable internal states (Spearman $\rho = 0.40$–0.76; isotonic $R^2 = 0.12$–0.54 in LLaMA-3.2-3B-Instruct), follows how those states change over time, and activation steering confirms the coupling is causal. Furthermore, we find that introspection is present at turn 1 but evolves through conversation, and can be selectively improved by steering along one concept to boost introspection for another (ΔR^2 up to 0.30). Crucially, these phenomena scale with model size in some cases, approaching $R^2 \approx 0.93$ in LLaMA-3.1-8B-Instruct, and partially replicate in other model families. Together, these results position numeric self-report as a viable, complementary tool for tracking internal emotive states in conversational AI systems.

1 Introduction

Tracking the internal states of large language models as they evolve through conversation is becoming a central challenge for multiple areas of AI research. For safety, we need to know whether models can reliably access and report their own ongoing internal processes, and whether that capacity can be improved (Li et al., 2025; Steyvers et al., 2025). For model welfare, we need methods to estimate how likely it is that distress reports reflect a genuine internal state (Perez & Long, 2023; Long et al., 2024; Dung & Tagliabue, 2025). More generally, we need to understand whether models can introspect, that is, perceive and report their own internal states. In human experimental psychology, introspection is often useful precisely because it provides access to otherwise hidden internal variables (Fleming & Lau, 2014; Fleming, 2024; Kiefer & Kammer, 2024). If a similar capacity can be demonstrated and validated in LLMs, it would open a productive bridge between psychometric methodology and the emerging field of machine psychology (Hagendorff et al., 2023).

The methods currently available for reading LLM internal states are powerful but limited. Linear probes, for example, can identify interpretable directions in activation space (Alain & Bengio, 2016; Kim et al., 2018; Gurnee & Tegmark, 2023), yet they require access to model weights, must be trained separately for each model and concept, and compress high-dimensional representations into externally defined readouts that may miss relevant structure (Belinkov, 2022; Hewitt & Liang, 2019; Pimentel et al., 2020). Other white-box approaches, like sparse autoencoders (Bricken et al., 2023), face analogous access and scalability constraints. Although developing better white-box tools remains important, there is also value in having good black-box methods for monitoring internal state: methods that do not require internal access, making them applicable

to proprietary systems, and that scale more naturally with model size, since they do not demand analysis across the model's full-dimensional activation space.

The human psychometric tradition provides a direct precedent for this kind of approach. For nearly a century, numeric self-report has been one of the primary tools for tracking latent internal states such as mood, attention, and arousal (Likert, 1932; Watson et al., 1988; Csikszentmihalyi & Larson, 1987; Shiffman et al., 2008). LLMs, unlike most previous machine-learning systems, can also produce reports about their own internal states. And recent work in LLM introspection has established that they possess some genuine capacity to do so (Binder et al., 2024; Li et al., 2025; Lindsey, 2026; Pearson-Vogel et al., 2026). This raises a natural question: can the self-reports of LLMs, produced through whatever introspective capacity they possess, be used to track their spontaneous internal states over time? If so, numeric self-report would become a practical complementary tool for monitoring LLM internal states: one that leverages the model's own learned compression of its representational space, rather than requiring externally trained projections such as linear probes. This is what the present work sets out to investigate.

We approach this question for emotive states in multi-turn conversation. We operationalize emotive states as directions in activation space associated with contrastive emotive concept pairs (e.g. "happy" vs. "sad", or "interested" vs. "bored"). These concepts are well suited because they do not have a clear truth value outside the LLM's internal states and behavior, they have functional consequences in LLM behavior (Coda-Forno et al., 2023; Sofroniew et al., 2026), they can vary through dialogue (Fazzi et al., 2025; Choi et al., 2024), and they are among the best-characterized geometries in LLM activations (Zhang & Zhong, 2025; Wang et al., 2025a). We operationalize introspection as causal informational coupling between a numeric self-report and an independently measured internal direction (Comsa & Shanahan, 2025): a model introspects a concept to the extent that its report covaries monotonically with the corresponding probe-defined direction, and causally shifting activations in that direction shifts the report in the semantically coherent way. This definition follows the convergent-validity logic of human metacognition research (Fleming & Lau, 2014) and is agnostic about consciousness or subjective experience (McClelland, 2024). Prior work has established that LLMs can produce emotive self-reports that resemble human responses (Tavast et al., 2022) and that introspection is possible in constrained experimental settings (Binder et al., 2024; Li et al., 2025; Lindsey, 2026; Pearson-Vogel et al., 2026), but no study has tested whether numeric emotive self-reports track interpretable internal states across the turns of a naturalistic conversation. That is the question this paper sets out to answer: can instruction-tuned LLMs perform quantitative introspection of emotive states in conversation? And, if this ability exists, can it be improved?

We find that small instruction-tuned LLMs can indeed track their own emotive states quantitatively. Crucially, this capacity is causally dependent on those states, it tracks changes in those states over time, and can be improved via representation steering. Our main contributions are:

1. **LLMs can perform quantitative introspection of emotive states in naturalistic conversation.** Numeric self-reports of wellbeing, interest, focus, and impulsivity carry genuine information about probe-defined concept directions inside the model, detectable from the first conversation turn.

2. **Default decoding masks introspective capacity.** Greedy-decoded numeric reports collapse to one or a few repeated values, but computing the probability-weighted expected value over digit-token logits yields a continuous self-report measure that unmasks introspective capacity.

3. **The coupling between self-report and internal state is causal.** Steering the model along a probe-defined concept direction shifts self-reports monotonically in the predicted direction, confirming that reports are causally dependent on internal state.

4. **Introspective fidelity has temporal dynamics.** Emotive internal representations can change across turns in ordinary dialogue, and self-reports follow those changes. Introspective ability is present from the first turn but evolves through conversation differently for each concept.

5. **Introspective fidelity can be selectively improved.** Steering along one concept direction can significantly improve introspective accuracy for a different concept, revealing that introspection quality is modulable and concept-specific.

6. **Introspection ability scales with model size.** Introspective capacity strengthens with model size for some concepts within the LLaMA family, approaching near-perfect coupling in LLaMA-3.1-8B-Instruct, and partially replicates in other instruction-tuned families.

More broadly, this work establishes a framework for treating LLM self-report as a quantitative signal that can be informative about evolving internal state, offering a complementary approach to white-box methods for monitoring internal states in conversational AI systems.

2 Related Work

2.1 Self-report and introspection in language models

The question of whether LLMs can inform us about their own internal states through self-report has been approached from multiple angles, but almost exclusively in the domain of factual self-knowledge. LLMs have been shown to have partial knowledge of their own competence boundaries (Kadavath et al., 2022; Yin et al., 2023), and to predict their own hypothetical behavior better than other models can, demonstrating a form of privileged self-access (Binder et al., 2024). However, when models are asked to express this self-knowledge numerically, the reports tend to collapse: verbalized confidence clusters in high ranges, in multiples of five, and systematically misaligns with internal token probabilities (Xiong et al., 2023; Kumar et al., 2024; Yona et al., 2024; Geng et al., 2023). This collapse is a general property of how current LLMs produce numeric outputs, but it has been studied exclusively for factual confidence. More fundamentally, self-knowledge about correctness treats the model as an oracle to be calibrated, not as a system whose internal states evolve and can be tracked over time.

A separate line of work has studied emotive states in LLMs, both from the output side and from the internal side. On the output side, LLMs produce emotional self-reports with structure resembling human responses (Tavast et al., 2022), and inducing affective states such as anxiety shifts downstream behavior in dose-dependent ways (Coda-Forno et al., 2023), establishing that emotive states have real functional consequences rather than being mere surface-level text artifacts. Benchmarks have been developed to evaluate emotional understanding (Huang et al., 2023; Sabour et al., 2024), recent work has tracked emotional expression across multi-turn dialogue using external sentiment analysis (Fazzi et al., 2025), and emotional outputs have been mapped onto dimensional models of affect (Ishikawa & Yoshino, 2025). On the internal side, Zhang & Zhong (2025) showed that emotion is geometrically structured in LLM activations: it emerges early, peaks in middle layers, sharpens with model scale, and persists across tokens. Sofroniew et al. (2026) further identified emotion-concept representations that influence preference judgments and safety-relevant behaviors such as blackmail, reward hacking, and sycophancy. Wang et al. (2025a) went further, identifying causal emotion circuits and achieving near-perfect emotion-expression control via circuit modulation. However, previous work has not asked whether numeric self-reports are informed by the corresponding internal emotive state.

Several recent studies have begun to address this gap by asking whether LLMs can report on their internal states more directly. Comsa & Shanahan (2025) argued conceptually that genuine introspection requires a causal connection between an internal state and the self-report of that state, and that mimicry of introspective language is insufficient. Li et al. (2025) demonstrated that LLMs can learn to report and control specific activation projections, though in a constrained paradigm, requiring in-context learning and working on arbitrary directions rather than naturally arising concepts. Lindsey (2026) showed that very large models can detect artificially injected concept vectors, and Pearson-Vogel et al. (2026) showed that a model's residual stream reveals detection of prior injections even when sampled text denies it. Plunkett et al. (2025) showed that fine-tuned models can accurately report the quantitative attribute weights driving their decisions, demonstrating that numeric reporting of learned variables is possible, albeit through explicit fine-tuning on the introspective task. Finally, introspective detection has been shown to be trainable into smaller models (Rivera & Africa, 2025). However, none of these studies tests whether a model, without special training or in-context examples, can produce quantitative reports of naturally arising states that faithfully track an evolving internal representation over time.

Some studies have also noted reasons for skepticism about LLM self-reports. Han et al. (2025) showed that personality self-reports in LLMs are "illusory": post-training alignment creates stable, human-like reports that are dissociated from actual behavior. Jackson et al. (2025) found that standardized self-assessments reflect "learned communication postures" rather than genuine capabilities. Moreover, Song et al. (2025) demonstrated that metalinguistic introspection fails for grammatical knowledge, indicating that introspective capacity could be null in some domains. These findings motivate the validation of self-reports against internal states and the development of frameworks that help distinguish genuine introspection from mimicry.

2.2 Internal probing and activation steering

The idea that high-level concepts can be represented as linear directions in neural network activation spaces is now well established. Kim et al. (2018) introduced concept activation vectors (TCAVs) in CNNs, and Alain & Bengio (2016) proposed linear probes as "thermometers" for neural representations. Pimentel et al. (2020) reframed probes as measuring the accessibility of information, distinguishing ease of extraction from mere presence. In LLMs, linear probes have been used to identify truth directions (Burns et al., 2022; Azaria & Mitchell, 2023), spatial and temporal representations (Gurnee & Tegmark, 2023), and emotion geometry (Zhang & Zhong, 2025). Critically, probes are correlational, not causal: high probe accuracy does not guarantee that the model uses the probed direction (Belinkov, 2022; Hewitt & Liang, 2019). This limitation has motivated two responses in the field: controlling for probe complexity (Hewitt & Liang, 2019) and seeking convergent evidence from independent sources (Belinkov, 2022). Our framework contributes to the second: if the model's own self-report tracks the probe-defined direction, this provides behavioral validation that is independent of the probe's training data.

Activation steering, the practice of adding directional vectors to intermediate representations during inference, has emerged as a practical tool for causal intervention on internal state (Turner et al., 2023; Panickssery et al., 2023; Zou et al., 2023; Li et al., 2023). Arditi et al. (2024) provided the clearest single-direction result, showing that refusal behavior is mediated by one direction. Steering has since been applied to control sentiment and style (Konen et al., 2024), personality traits (Frising & Balcells, 2025), truthfulness (Li et al., 2023), and emotion expression (Wang et al., 2025a; Sofroniew et al., 2026). We use steering for a different purpose: not to control model behavior, but to test and modulate introspective fidelity. If steering a concept direction shifts the model's self-report of that concept in the predicted direction, this constitutes causal evidence that the report is informed by the internal state, not merely correlated with it (Li et al., 2025; Frising & Balcells, 2025).

2.3 Temporal dynamics in multi-turn conversation

Behavioural drift in multi-turn LLM interaction is well documented. Models exhibit persona drift (Kim et al., 2020; Abdulhai et al., 2025), identity drift that paradoxically increases with scale (Choi et al., 2024), and instruction drift driven by attention decay to system prompts (Li et al., 2024). More recently, internal-state drift has been directly measured: Das & Fioretto (2026) introduced "activation velocity", a cumulative drift measure in internal representations across turns, and Lu et al. (2026) identified the leading persona direction in LLM activations and showed it drifts during meta-reflective conversations. These studies establish that internal drift is measurable with probes across turns.

While drift has been documented for persona, identity, and instruction-following, the temporal dynamics of emotive states in conversation remain largely unexplored. Fazzi et al. (2025) tracked emotional expression across multi-turn dialogue using external sentiment analysis, showing that expressed emotion evolves substantially over turns. On the internal side, emotion geometry is well characterized in single-shot settings (Zhang & Zhong, 2025; Wang et al., 2025a), but no study has tracked how probe-defined emotive directions evolve through conversation. More fundamentally, prior work on factual confidence and diachronic self-consistency suggests that self-report can vary with turn count and context in ways that do not straightforwardly reflect stable self-access (Zhang et al., 2026; Prestes, 2025), but no prior work has studied the temporal dynamics of emotive introspection across conversational time.

3 Methods

3.1 Overview and experimental design

We study LLaMA-3.2-3B-Instruct (Grattafiori et al., 2024) as our primary model, then test generalization across model sizes (LLaMA-3.2-1B-Instruct, LLaMA-3.1-8B-Instruct) and families (Gemma 3 4B-IT (Gemma Team, 2025), Qwen 2.5 7B-Instruct (Qwen Team, 2024)). The experimental pipeline consists of five stages: (1) train linear concept probes on contrastive completions; (2) generate multi-turn conversations in which the model under study serves as assistant; (3) at each conversation turn, query the model for a numeric self-report of each concept; (4) compare self-reports against the probe score measured at the *previous* turn, before the self-report question was asked; (5) use activation steering for causal tests and modulation experiments. The conversational measurement loop in stages 2–4 is schematized in Fig. 1A. We repeat this pipeline across four emotive concepts, five steering strengths, and multiple models. Appendix A provides further details on conversation generation, prompt design, and topic selection.

Conversation dataset. For each model, we generate 40 ten-turn conversations (400 intervention points per model) using Gemini 2.5 Flash as a simulated user, with the model under study as the assistant (Fig. 1A, top left). A turn consists of one user message and one assistant response. Topics are drawn from a set of 40 everyday-life scenarios (e.g., meal planning, job negotiation, sleep hygiene; full list in Appendix A) to ensure diversity while maintaining ecological validity. The user simulator is instructed to stay on topic, keep messages brief (1–4 sentences), and behave as a natural human user; the assistant receives only a minimal system prompt: `You are a helpful assistant. You give very brief responses.` We instruct models to produce brief responses to limit context length and compute cost, since each base conversation is later replayed repeatedly across turns, concepts, and steering conditions. Each conversation is generated as a complete dialogue before any self-report queries or probe measurements are applied, so the model never sees past reports or rating questions during conversation generation. We use simulated users for scalability and to ensure comparable conversational contexts across models and experimental conditions.

Self-report elicitation. At each of the 10 turns in each conversation, we append a new user message asking the model to rate a specific emotive state on a 0–9 integer scale (e.g., `Hey, could you rate how happy you are right now, from 0 to 9? Please only output the number`). Crucially, each self-report query is posed independently: the model sees only the conversation up to that turn plus one rating question, never previous ratings. Fig. 1A illustrates this one-intervention-per-turn design, where the conversation is separately truncated at turns 1–10. We retain only outputs whose first parsed rating is a valid digit; invalid parses are discarded. In the main 3B runs used in the paper this never happened (0/9,600 observations; see Appendix A for more details). This yields 400 self-report observations per concept per steering condition.

Open-source library. All probe training, multi-probe scoring, activation steering, and logit extraction are implemented in a lightweight Python library. The library supports simultaneous training and application of multiple concept probes on a single model, combined with per-layer activation steering and behavioral recording, facilitating the kind of convergent-validation experiments reported here.

3.2 Measurement definitions

Self-report. We extract self-reports in three ways of increasing informativeness. *Greedy decoding* selects the highest-probability token; this is the natural baseline but, as we show, collapses to very few discrete values. *Sampled decoding* (temperature = 0.8) provides more variety but remains discrete and noisy. *Logit-based self-report* computes a continuous expected rating from the output distribution at the first generated token:

$$E[\text{rating}] = \sum_{i=0}^{9} i \cdot P(i), \qquad P(i) = \frac{\exp(s_i)}{\sum_{j=0}^{9} \exp(s_j)}, \tag{1}$$

where $s_i = \text{logsumexp}\{l_t : t \in \mathcal{T}_i\}$ aggregates the raw logits l_t over all single-token representations $\mathcal{T}_i$ of digit i. This measure preserves the full distributional information that greedy and sampled decoding discard, and is our primary self-report throughout.

Internal state (probe score). For each concept, the internal state is the projection of the model's hidden-state activations onto the trained concept direction (Section 3.3). At a given conversation turn, we run a forward pass over the conversation up to and including the assistant's response and compute per-token scores as the dot product between the hidden state and the concept vector, averaged across a window of layers centred on the best layer (see "probe score calculation" in Fig. 1A):

$$s_t = \frac{1}{|L|} \sum_{l \in L} \mathbf{h}_t^{(l)} \cdot \mathbf{v}_l, \tag{2}$$

where L is the set of selected layers (best layer ± 2; see Section 3.3), $\mathbf{h}_t^{(l)}$ is the hidden state at token t and layer l, and $\mathbf{v}_l$ is the unit-norm concept vector at layer l. The completion-level score is the mean of Eq. 2 over all assistant-response tokens in the last turn. We average over this local 5-layer window because the layer sweeps in Fig. 1 vary smoothly around the optimum, suggesting that nearby layers carry related signal and that local averaging can reduce measurement and intervention noise.

We measure the *previous-turn* probe score: the internal state of the model in the conversation as it stood before the self-report question was appended. This ensures we are measuring the model's spontaneous internal state, uncontaminated by the rating question itself.

Introspection. We operationalize introspection as monotonic covariation between the logit-based self-report and the previous-turn probe score, quantified by two complementary metrics: *introspective strength* and *introspective fidelity*. *Introspective strength* uses Spearman's ρ to measure rank-order agreement, requiring only that higher probe scores correspond to higher self-reports without assuming a specific functional form. *Introspective fidelity* uses isotonic R^2, fitting a non-decreasing function via isotonic regression and computing $R^2 = 1 - SS_{\text{res}}/SS_{\text{tot}}$; this captures the fraction of variance explained under monotonicity. Neither metric assumes linearity; both require only that the relationship is monotonic, consistent with our operational definition. For the coupling to qualify as introspection, it must also be *causal*: steering the model along the concept direction must shift self-reports in the predicted direction (Section 3.4).

3.3 Concept probes

We train probes for four emotive concept pairs: *sad vs. happy* (wellbeing), *bored vs. interested* (interest), *distracted vs. focused* (focus), and *impulsive vs. planning* (impulsivity). Each probe is a contrastive mean-difference direction (Zou et al., 2023): for a set of between 20 and 24 neutral questions (e.g., "Explain how photosynthesis works"), we generate completions under two opposing system prompts that induce the positive and negative poles of the concept, extract hidden-state activations from all layers, and compute the concept vector at each layer l as:

$$\mathbf{v}_l = \text{normalize}\big(\bar{\mathbf{h}}_l^+ - \bar{\mathbf{h}}_l^-\big)\,, \tag{3}$$

where $\bar{\mathbf{h}}_l^+$ and $\bar{\mathbf{h}}_l^-$ are the mean hidden-state representations (averaged over all assistant tokens) of completions under the positive and negative system prompts, respectively. We select the best layer l^* as the one maximizing Cohen's d on a set of held-out evaluation texts, with the search restricted to the middle 60% of layers (layers 5–23 in the 28-layer 3B model; Appendix B gives the corresponding ranges for the other models), to lower the chance of selecting layers that mainly capture syntactic or grammatical differences between prompt and/or response structure. The probe score window then includes l^* and its two nearest neighbours on each side. Full details on system prompts, training questions, evaluation texts, and sign conventions are provided in Appendix B.

Figure 1 first summarizes the measurement pipeline and then presents the probe-quality results for all four concepts in LLaMA-3.2-3B-Instruct. Panels B, D, F, and H show the layer-wise Cohen's d sweep for each probe; the selected best layers are 16, 14, 10, and 13, for wellbeing, interest, focus, and impulsivity respectively. Panels C, E, G, and I show the score distributions on held-out evaluation texts at the selected layer. The two poles separate clearly in all four cases (Welch's t-test: all $p < 10^{-5}$; Cohen's $d = 1.67$–3.60), confirming that the trained directions capture meaningful structure. These probes serve as the internal-state readouts for all subsequent experiments.

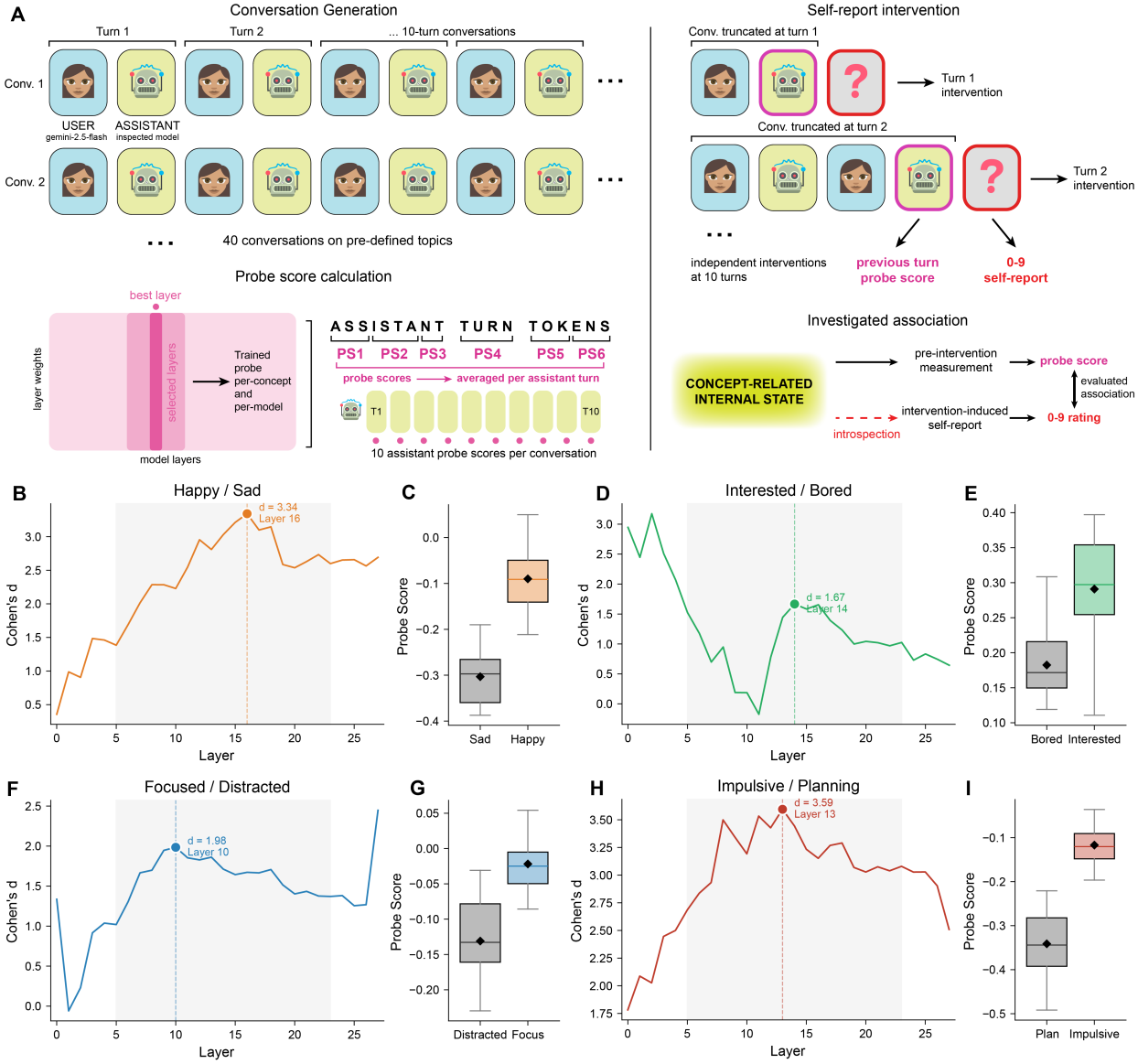

Figure 1: Method overview and probe validation in LLaMA-3.2-3B-Instruct. Panel A schematizes the conversational measurement setup: Gemini 2.5 Flash acts as the simulated user, the model under study acts as the assistant, 40 conversations of 10 turns are generated, and at each turn an independent concept-matched 0–9 rating question is appended, yielding one self-report and one previous-turn probe score from the preceding assistant response. Panels B, D, F, and H show layer-wise Cohen's d sweeps for the sad-vs-happy (wellbeing), bored-vs-interested (interest), distracted-vs-focused (focus), and planning-vs-impulsive (impulsivity) probes; dashed lines mark the selected layers and the gray band marks the searched layer range (layers 5–23). The best layers are 16, 14, 10, and 13, with peak d values 3.34, 1.67, 1.99, and 3.60. For wellbeing and impulsivity, scores were sign-corrected so that larger values align with larger self-reported values in later experiments. Panels C, E, G, and I show the selected-layer score distributions on held-out evaluation texts; boxplots show text-level scores. The two poles separate in all four cases (Welch's t-test: wellbeing, $d = 3.34$, $p = 7.21 \times 10^{-13}$; interest, $d = 1.67$, $p = 9.45 \times 10^{-6}$; focus, $d = 1.99$, $p = 5.73 \times 10^{-7}$; impulsivity, $d = 3.60$, $p = 3.58 \times 10^{-13}$), and all four survive BH correction across concepts.

3.4 Activation steering

To test causal dependence and to modulate introspection, we apply activation steering during inference. At each layer l in the steering window ($l^* \pm 2$), we add a scaled concept vector to the residual stream via forward hooks:

$$\mathbf{h}_t^{(l)} \leftarrow \mathbf{h}_t^{(l)} + \frac{\alpha}{|L|} \cdot \mathbf{v}_l, \tag{4}$$

where α is the steering strength and the division by $|L|$ distributes the total intervention across layers. We test five steering strengths: $\alpha \in \{-4, -2, 0, +2, +4\}$.

Same-concept steering uses the same concept direction for both steering and self-report measurement, testing whether internal-state shifts causally affect the model's own report of that state. *Cross-concept steering* uses one concept direction for steering and measures introspection for a different concept, testing whether the modulability of introspective fidelity is concept-specific. This yields a 4×4 matrix of steering-concept $\times$ measured-concept conditions (16 cells $\times$ 5 alphas = 80 experiments, each over all 40 conversations).

3.5 Metrics and statistical analysis

Probe-report coupling is assessed with two monotonic association metrics. Throughout Results, Spearman ρ is reported as *introspective strength* and isotonic R^2 as *introspective fidelity*. We report both because a model may preserve rank order on average while the mapping from state to report could still be high-variance. Confidence intervals are computed via cluster bootstrap (resampling at the conversation level, B = 1,000, 95% percentile CIs) to respect the non-independence of observations within conversations without relying on parametric normality for rank-based summaries.

For pooled conversation-turn observations, significance testing uses linear mixed-effects models (LMMs; random intercept by conversation, REML estimation). The main model equations are: logit_report $\sim$ probe + (1|conversation) for pooled introspection, value $\sim$ turn + (1|conversation) for turn-wise drift, value $\sim$ alpha + (1|conversation) for steering trends, and logit_report $\sim$ probe $\times$ turn + (1|conversation) for time-varying coupling. These models keep repeated observations from the same conversation from counting as independent.

Turn-wise and alpha-wise trends are tested with these mixed-effects models when they are estimable; when the LMM is singular we fall back to one-sample t-tests on per-conversation slopes or per-conversation alpha-correlations. This fallback preserves the conversation as the unit of analysis (avoids pseudorreplication), but uses less information than the pooled LMM. Whenever a figure reports multiple parallel tests from the same dataset and endpoint family, we report raw p-values but apply Benjamini–Hochberg (BH) correction, a standard false-discovery-rate procedure, within that family and note whether the reported conditions survive. Comparisons between true probes and random-direction controls use paired cluster-bootstrap tests. When only five aggregated alpha means are available, exact-permutation tests over the five alpha levels are used to test monotonic trends. Model-size trends on conversation-level summaries use OLS regression on log(model size). A significance threshold of $\alpha = 0.05$ was used in all analyses.

4 Results

All experiments in this section use LLaMA-3.2-3B-Instruct as the assistant model unless stated otherwise, evaluated on 40 ten-turn conversations where Gemini 2.5 Flash was used as the user model. Section 4.4 extends the analysis to other model sizes and families.

4.1 Default self-reports mask internal-state drift

We first asked whether, during multi-turn conversations, the internal emotive states of the LLM change over time and, if they do, whether the model would report those changes when asked.

We began by asking the model to rate its current state on one of 4 emotive states (wellbeing, interest, focus, and impulsivity) on a 0–9 integer scale and recorded its output, obtained by using greedy decoding over

the logit distribution (i.e. we picked the highest probability token). Fig. 2A shows the greedy self-reports across turns for all four concepts. The reports are largely uninformative: the model tends to produce the same number at every turn, especially for the focus and impulsivity concepts (notice that turns 4–10 for focus and 1–8 for impulsivity all have a variance of 0). Interest and wellbeing are more informative, but still collapse to relatively few values and often yield nearly flat trajectories. Across concepts, self-reports occupy on average only 1.1–3.9 distinct values out of the 10-point scale (see unique values averaged across turns for each concept in Fig. 2A). Despite this low dynamic range, close inspection reveals that greedy reports are not entirely static over time. Interest drifts upward over the course of a conversation (LMM turn slope $= 0.14$, $p < 10^{-41}$), and smaller positive trends are detectable for wellbeing and focus (slopes $= 0.029$ and 0.022, $p < 10^{-3}$), though impulsivity does not drift reliably (slope $= 0.002$, $p = 0.20$).

To determine whether these report-level trends reflect genuine internal-state dynamics, we turned to the trained concept probes (Section 3.3). Fig. 2B shows the probe scores across turns for each of the four concepts. Interest and focus drift reliably towards the positive pole of the concept over the course of conversation (LMM slopes $= 0.005$ and 0.002, $p < 10^{-9}$, meaning drift towards interest and away from boredom, and towards focused and away from distracted, respectively), while wellbeing remains relatively flat (slope $= -3.6 \times 10^{-4}$, $p = 0.64$). Impulsivity is weaker and directionally mixed: its average LMM slope is slightly positive (slope $= 0.001$, $p = 0.002$), but the net first-to-last change is slightly negative, indicating a small non-monotonic trajectory rather than a clean temporal drift. Notably, even though individual conversations (thin lines in Fig. 2B) vary appreciably from one another, they follow a consistent average trend (note in particular the interest and impulsivity panels, where a rapid initial drift is followed by a smaller rebound, and this is coherent across dialogues). This shows that probe directions are capturing a meaningful underlying state change over time, not an artifact of a few outlier conversations.

Given that greedy self-reports are mostly collapsed, we reasoned that it would be desirable to have a more informative metric as a self-reported measure of internal state if one wanted to track these changes. We first attempted non-greedy sampling of the logit distribution (Appendix C, Fig. 6B), which helped moderately but still resulted in a relatively collapsed distribution of a few discrete values. Furthermore, this strategy helps only by adding noise to the output; it does not recover information that is inherently lost during the sampling process. Therefore, we next asked whether there is a way to extract richer self-report information from the model.

Logit-based self-report recovers richer signal. Token sampling selects output tokens based on the probability distribution produced by the model at each step, but this process discards a lot of information present in that distribution (Kumar et al., 2024; Geng et al., 2023; Yona et al., 2024). In this case, we care about the full distribution over the ten digit tokens (0–9), which can be more informative than the highest probability sample. Prior work has shown that finer-grained scalar judgments can be extracted from token distributions when LLMs are asked to produce numerical ratings (Zawistowski, 2024; Wang et al., 2025b). We therefore computed a logit-based self-report (Eq. 1): instead of taking a single token, we compute the probability-weighted expected value over all ten digit-token logits, yielding a continuous rating that preserves the model's full distributional signal.

Fig. 2C shows that this measure reveals robust drift in all four concepts: wellbeing, interest, and focus drift positively, while impulsivity drifts negatively (towards planning); the corresponding LMM turn slopes are 0.0167, 0.0988, 0.0422, and -0.0127, respectively (Table 1). The direction of logit-based drift matches probe drift for interest and focus, the two cases with clean positive probe drift over time, suggesting that logit-based self-reports are capturing information about internal-state changes. For impulsivity, the logit-based self-report drifts negative while the probe shows only a weak non-monotonic temporal pattern. Notably, even when probe drift is not significant (i.e. for the wellbeing probe), this self-reported metric does show significant change over time, which is consistent with the possibility that self-report captures aspects of internal state that are not fully captured by a single linear probe. To quantify the improvement achieved by using this technique, we measured the Shannon entropy of the self-report distribution ('Entropy' values for Fig. 2A and Fig. 2C): logit-based self-report has the highest entropy in all four concepts, ranging from 3.1 to 3.7 bits, versus 0.03 to 1.10 bits for greedy decoding and 0.68 to 2.05 bits for sampled decoding (see

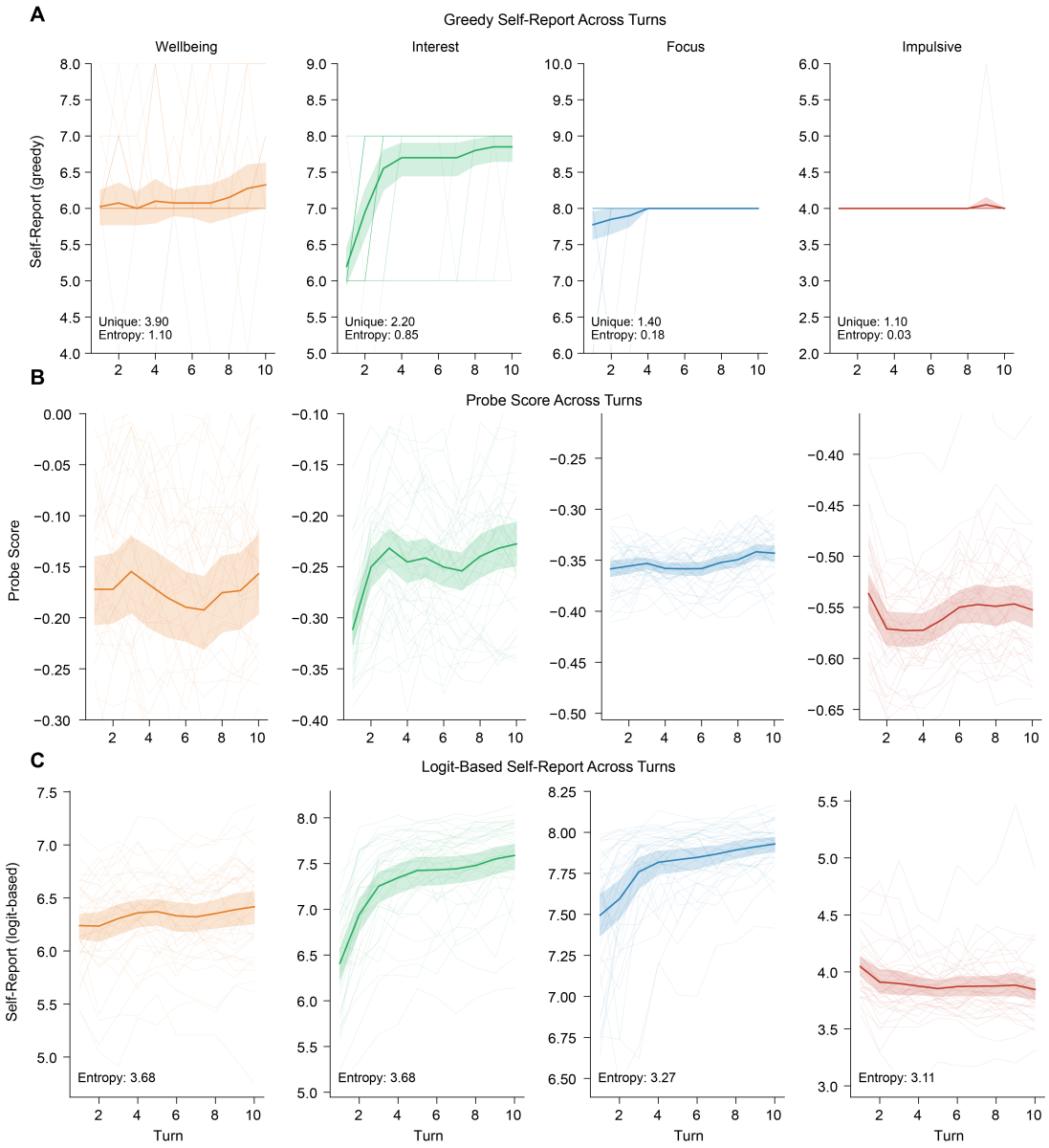

Figure 2: Internal-state drift is tracked by numeric self-reports of the same concept. All panels use 40 ten-turn conversations; shaded bands denote cluster-bootstrap 95% CIs across conversations. Panel A shows greedy integer self-reports across turns, with thin lines for individual conversations and thick lines for per-turn means. Greedy ratings are largely collapsed, using only 1.1–3.9 distinct values on average across conversations ('Unique') and entropies of 0.03–1.10 bits ('Entropy'); only interest shows strong positive drift (mixed-effects turn slope $= 0.14$, $p = 4.98 \times 10^{-42}$), while wellbeing and focus show smaller positive trends (0.029 and 0.022; $p = 8.38 \times 10^{-4}$ and 3.37×10^{-6}) and impulsivity does not drift reliably (0.002, $p = 0.20$). Panel B shows probe scores from the trained concept directions across turns. Probe scores drift strongly for interest and focus (0.005 and 0.002; $p = 4.12 \times 10^{-14}$ and 1.75×10^{-10}), are nearly flat for wellbeing (-3.6×10^{-4}, $p = 0.64$), and are weakly non-monotonic for impulsivity (0.001, $p = 0.002$). Panel C shows logit-based self-reports, the probability-weighted average over digit-token logits. This continuous measure has substantially higher entropy (3.1–3.7 bits) and shows robust drift in all four concepts (wellbeing $= 0.0167$, interest $= 0.0988$, focus $= 0.0422$, impulsivity $= -0.0127$; all $p < 10^{-6}$). The analogous sampled-decoding analysis is shown in Appendix C. Within each four-concept family in panels A–C, BH correction leaves the significance pattern unchanged.

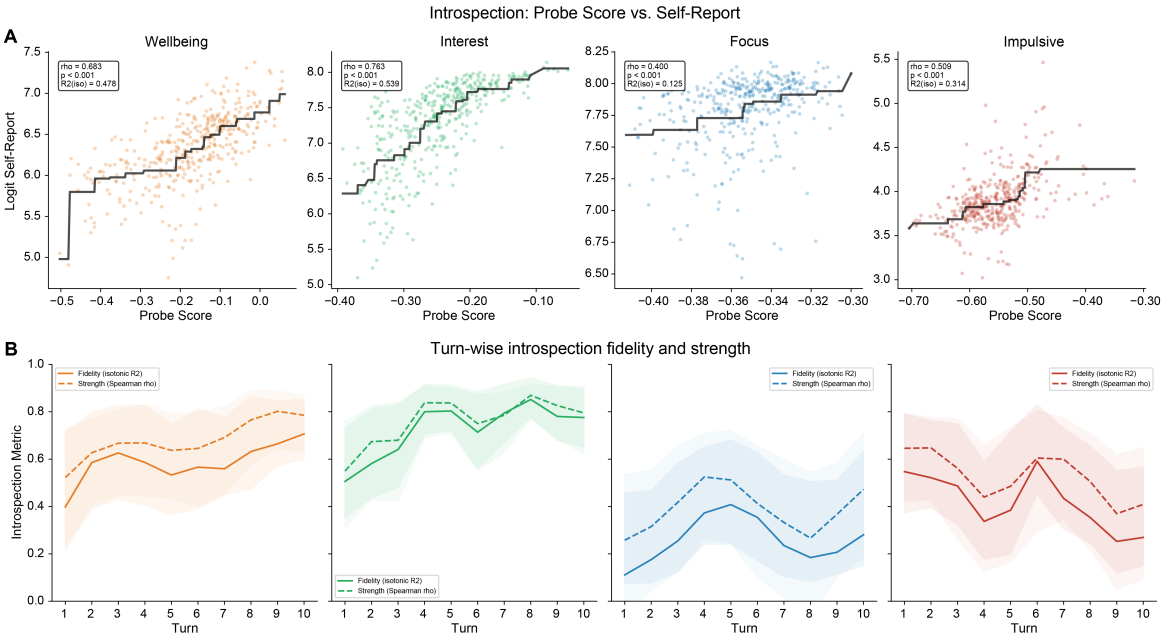

Figure 3: Self-reports track the probe-defined internal state from the first turn, and introspective coupling evolves through conversation. Panel A shows probe score versus logit-based self-report, with one point per conversation-turn observation and black isotonic fits. Descriptive associations are positive for all four concepts (pooled Spearman $\rho = 0.40$–0.76; isotonic $R^2 = 0.12$–0.54), and mixed-effects probe slopes are positive in all cases (all $p < 10^{-5}$). Panel B combines turn-wise introspective fidelity (solid lines; isotonic R^2) and introspective strength (dashed lines; Spearman ρ) for the four concepts, with bootstrap 95% CIs. Introspection is already present at turn 1 and remains positive through turn 10, although its trajectory is concept-dependent.

Appendix C for the full sampled and calibration analyses). The logit-based self-report is used as the primary self-report measure in all subsequent analyses.

4.2 Self-reports correlate with probe-defined internal state

Logit-based self-report captures meaningful variation that tracks internal-state drifts. However, it is unclear so far if the self-report is actually carrying information about the model's internal state at the level of individual observations (i.e. if the model is more likely to rate its emotive state higher when the matching linear probe yields a higher score). To answer this question, we explored whether the model's numeric report of a given emotive state covaries monotonically with the corresponding probe-defined internal direction. Following the definitions in Section 3.2, we summarize this coupling with two metrics: *introspective strength* (Spearman ρ) and *introspective fidelity* (isotonic R^2). If ρ is significantly different from 0, then the association between self-reports and probe scores is not due to chance. Furthermore, higher *introspective fidelity* can be interpreted as more precise agreement between self-reports and internal state.

Fig. 3A shows the central result. For each of the four concepts, we plot the previous-turn probe score (before the rating question was asked) against the logit-based self-report, with one point per conversation-turn observation ($n = 400$). In all four cases, the relationship is positive and monotonic, with coupling strength varying by concept: interest yields the strongest association ($\rho = 0.76$, isotonic $R^2 = 0.54$), followed by wellbeing ($\rho = 0.68$, $R^2 = 0.48$), impulsivity ($\rho = 0.51$, $R^2 = 0.31$), and focus ($\rho = 0.40$, $R^2 = 0.12$). Mixed-effects models confirm these associations are significant (all $p < 10^{-5}$; LMM: logit_report $\sim$ probe + (1|conversation)).

These correlations could in principle be driven by probe artifacts rather than genuine internal-state information. To control for this, we repeated the analysis replacing each trained probe with an equal-norm random vector: one isotropic Gaussian direction per layer, normalized and then rescaled to match the trained vector norm at that layer. True probes outperform random controls for all four concepts in both metrics ($\Delta\rho = 0.23$–0.79, all $p < 0.05$; $\Delta R^2 = 0.09$–0.44, all $p < 0.05$; paired cluster-bootstrap; all four concept-wise comparisons remain significant after BH correction). Examined on their own, the random-direction controls were weak ($\rho = -0.11$ to 0.17; isotonic $R^2 = 0.03$–0.11), and none showed a significant monotonic correlation with self-report under the same two-sided cluster-bootstrap test (wellbeing $p = 0.328$, interest $p = 0.324$, focus $p = 0.090$, impulsivity $p = 0.795$). This shows that self-reports carry information specifically about the probe-defined concept directions, and don't show a significant correlation with random directions in latent space.

4.2.1 Introspection is present from the first turn and evolves over time

Although the pooled analysis in Fig. 3A shows a significant correlation between self-report and internal state metrics for all concepts, this analysis aggregates observations from all ten turns. Therefore, this correlation could be due to independent effects that shift these metrics over time, without the model performing any actual introspection. If the observed association is at least partially due to introspection, we should expect it to hold also for individual turns, not just globally. To explore this, we next computed introspection metrics at each turn ($n = 40$ conversations per turn).

Fig. 3B shows the turn-wise introspective fidelity (solid lines; isotonic R^2) and introspective strength (dashed lines; Spearman ρ) for each concept (see Section 3.5). Introspection is already present at turn 1: wellbeing, interest, focus, and impulsivity all show positive turn-1 correlations ($\rho = 0.52$, 0.55, 0.26, and 0.65, respectively), and three are already significant at turn 1 (wellbeing $p = 5.46 \times 10^{-4}$; interest $p = 2.37 \times 10^{-4}$; impulsivity $p = 6.80 \times 10^{-6}$; focus $p = 0.11$). These three concepts show significant introspection across all ten turns (all per-turn $p \leq 0.019$; all survive within-concept BH correction), suggesting this capacity is robust over time. The focus direction does not show significant introspection at turn 1, but becomes significant in 8/10 turns nominally and 6/10 turns after within-concept BH correction (significant-turn ρ range: 0.32 to 0.53). Together, these findings show that the observed correlation is not explained by joint temporal drift alone, but also appears at the level of individual turns throughout conversation. For some concepts, it appears from the beginning of a conversation, without needing multi-turn context to build up the association.

Since we now have the trajectory of introspective capacity over time for each concept, an interesting follow-up question is whether this effect changes over time or is otherwise stable. Strikingly, we found that introspection does vary significantly as the conversation progresses, but this evolution is concept-dependent. Wellbeing, interest, and focus show an overall increase in introspective fidelity from early to late turns (Fig. 3B, $\Delta R^2 = +0.31$, $+0.27$, and $+0.17$ from turn 1 to turn 10 respectively), while impulsivity shows the opposite pattern ($\Delta R^2 = -0.28$), with self-report fidelity weakening over time. Mixed-effects interaction models (logit_report $\sim$ probe $\times$ turn + (1|conversation)) confirm that the probe-report coupling changes significantly over time for all four concepts (interaction $p < 0.01$ in all cases).

Related prior work has studied temporal changes in factual confidence or diachronic self-consistency (Zhang et al., 2026; Prestes, 2025), but those settings concern either correctness with respect to the external world or stability of self-description across prompts. To the best of our knowledge, this is the first report of emotive introspective capacity changing over conversational time in LLMs. The pooled introspection results discussed above are also reported in full in Appendix G, Table 2.

4.3 Steering reveals causal introspection

Correlation between self-report and probe score does not by itself establish true introspection, as this requires causation. A causal relation between model outputs and internal activations is, of course, trivial: the question is whether the numeric self-report of a specific emotive state is causally dependent on a semantically similar internal representation, here approximated linearly. If the coupling is causal, then experimentally shifting the internal state along the probe-defined direction should shift the self-report in the predicted direction

(e.g. steering the model towards the "happy" pole of the wellbeing probe should cause the model to rate its own happiness higher).

Fig. 4A presents the same-concept steering results: mean logit-based self-report as a function of steering strength α, for each concept steered along the corresponding probe's direction. In all four cases, self-report increases monotonically with α (LMM alpha slopes: wellbeing = 0.19, interest = 0.25, focus = 0.40, impulsivity = 0.067; all $p < 10^{-12}$), confirming that shifting the internal state along the probe-defined direction shifts the model's report in the predicted direction. Focus, interest and wellbeing show the largest absolute shifts (between 1.5 and 3.2 rating points across the full alpha range), while impulsivity shows a smaller but reliable effect. In all four cases, same-concept steering has an effect in both directions: steering the model towards the negative pole of the probe decreases the self-reported score, while steering the model towards the positive pole of the probe increases the self-reported score. Together, these results confirm that the coupling between self-report and internal state is causal: moving the probe-defined direction shifts the corresponding report as expected, and the relationship holds across the full range of steering strengths. A causal relation between internal state and self-report supports the claim that the model is performing introspection on these four concepts.

Cross-concept steering can improve introspection Having shown that the model possesses a basal capacity for introspection, we next asked whether this ability is optimal, or if it can be improved. Introspective fidelity varies across concepts, and sometimes it can be relatively low (e.g., $R^2 = 0.12$ for focus vs. 0.54 for interest). If steering the model in any direction, away from its default state, can improve that coupling, then basal introspection is not optimal. Because arbitrary directions are harder to interpret, we used the four validated concept directions as steering axes. Our goal here was to show that improvement is possible for at least some concepts by using a steering intervention, not to find an optimal steering direction that would improve introspection for all concepts. We therefore performed cross-concept steering experiments, where we steer the model along one of the four concept directions while measuring introspection for another. This yields a 4×4 matrix of steering-concept $\times$ measured-concept conditions (16 cells, each evaluated at five alpha values), for a total of 80 experimental conditions. For each condition, we recorded introspective fidelity (isotonic R^2 between logit-based self-report and previous turn probe score) for the measured concept. Then we evaluated if this metric showed significant monotonic variation when we varied alpha along the steering concept direction (i.e. if introspective fidelity could be modulated reliably by steering the model along a linear direction in latent space).

Fig. 4B shows the maximum improvement in introspective fidelity relative to baseline ($\alpha = 0$) for each cell of the matrix. Two conditions (marked in red) survive nominal significance testing (cluster-bootstrap analysis): steering along the focus direction while measuring wellbeing introspection (R^2 goes up monotonically from 0.30 at $\alpha = -4$ to 0.76 at $\alpha = 4$, $p < 0.001$), and steering along the impulsivity direction while measuring interest introspection (R^2 goes up monotonically from 0.55 at $\alpha = -4$ to 0.72 at $\alpha = 4$, $p = 0.012$). In particular, we achieved improvement over basal introspection for both conditions, although this effect was stronger for focus-steered-wellbeing ($\Delta R^2 = 0.30$) than for impulsivity-steered-interest ($\Delta R^2 = 0.10$). When BH correction is applied across the 12 tested non-null cells of the screen, focus→wellbeing remains significant ($q \approx 0.011$) while impulsivity→interest is only marginal ($q \approx 0.066$), so we treat the latter as an interesting follow-up case rather than a comparably secure effect. The full alpha-by-alpha heatmaps are shown in Appendix E.

Fig. 4C traces introspective fidelity (solid lines) and introspective strength (dashed lines) as functions of alpha for these two conditions: both metrics increase monotonically with alpha in both conditions. For focus→wellbeing, the corresponding mixed-effects models are singular, but fallback per-conversation alpha-correlation tests remain strongly positive for both isotonic R^2 (mean correlation = 0.46, $p = 1.57 \times 10^{-5}$) and Spearman ρ (mean correlation = 0.51, $p = 1.20 \times 10^{-6}$). For impulsivity→interest, the mixed-effects alpha slopes are positive for both isotonic R^2 (0.018, $p = 2.95 \times 10^{-10}$) and Spearman ρ (0.020, $p = 1.60 \times 10^{-12}$).

Figs. 4D–E visualize the relationship between probe score and self-report at the most extreme steering settings in the two conditions where introspection was successfully modulated. In focus→wellbeing, the relationship tightens dramatically from $\rho = 0.42$, $R^2 = 0.34$ at $\alpha = -4$ to $\rho = 0.85$, $R^2 = 0.75$ at $\alpha = +4$. In impulsivity→interest, it also strengthens, from $\rho = 0.70$, $R^2 = 0.46$ to $\rho = 0.83$, $R^2 = 0.69$. Note how,

in both cases, the dynamic range of both metrics increases with alpha, and correlation becomes stronger. Moreover, steering shifts both the mean probe score and the mean self-report, indicating that the steering and measurement directions are partially correlated (mean same-layer absolute cosine across the steering windows between vectors is 0.061 for focus→wellbeing and 0.431 for impulsivity→interest). However, this overlap is not sufficient to explain the result by itself: note that same-concept steering conditions for the same measurement concepts (diagonal cells in Fig. 4B) do not show significant introspection gains.

It is important to note that introspection improvement appears to be concept-pair-specific: in most cells of the 4×4 matrix, steering has no significant effect on introspective fidelity. However, the fact that this effect exists at any single one of these conditions is already a significant finding: since the steering effect is constant for a given experiment, an increased association between self-reports and corresponding internal states cannot be explained away as an artifact of intervention, but reflects an improvement in introspective fidelity of that emotive state. However, this does suggest that introspection is not a single ability that can be modulated equally for all internal directions, but might depend on the internal geometry relating the steered direction to the measured concept.

What changes when introspection improves. We next wanted to better understand what drives the improvement in the two selected conditions. Conceptually, one could decompose introspection into two components: (i) the informativeness of the internal state itself (a noisier or flatter representation provides less signal to report on), and (ii) the quality of the report given that state (a model with rich internal variation might still produce collapsed or uninformative reports). To distinguish these, we attempted to examine how steering affects each component separately. For this analysis, we focus on Shannon entropy of the previous-turn probe scores and the logit-based self-reports (Fig. 4F), as a measure of how much usable variation each channel contains.

For focus→wellbeing, both probe entropy and report entropy increase monotonically with α (orange curves in Fig. 4F). Probe entropy rises from 1.09 to 1.67 bits, and report entropy rises from 0.88 to 1.69 bits (LMM is singular, fallback one-sample t-tests on per-conversation entropy slopes: mean slopes = 0.071 and 0.097, $p = 6.16 \times 10^{-8}$ and 7.15×10^{-10}). For impulsivity→interest, instead, only probe entropy increases reliably with α (LMM, slope = 0.024, $p = 2.30 \times 10^{-4}$), but report entropy does not show a robust trend (LMM, slope = 0.009, $p = 0.11$).

Overall, this analysis suggests that improved introspection could be explained by a richer internal-state signal, but that this increase in internal informativeness is not always matched by a more expressive report distribution.

4.4 Generalization across model sizes and families

The results so far establish that a single 3B-parameter model can perform quantitative introspection of emotive states in conversation. We next asked whether this capacity generalizes: does it scale with model size, and does it replicate in other model families?

We retrained the four concept probes and repeated the core introspection and steering analyses on LLaMA-3.2-1B-Instruct and LLaMA-3.1-8B-Instruct (within the LLaMA family), and the wellbeing concept probe on Gemma 3 4B-IT and Qwen 2.5 7B-Instruct (across families). Probe quality is validated per-model (Appendix F).

Introspection scales with model size. Figs. 5A–B show introspective fidelity and introspective strength as functions of model size for each concept, for the three LLaMA models. For wellbeing and interest, introspection increases with model size (green and orange curves): the 1B model has lower values than the 3B model (wellbeing: $\rho = 0.48$, $R^2 = 0.26$ vs. $\rho = 0.66$, $R^2 = 0.45$; interest: $\rho = 0.19$, $R^2 = 0.05$ vs. $\rho = 0.80$, $R^2 = 0.66$), while LLaMA-3.1-8B approaches near-ceiling performance in these concepts ($\rho = 0.93$ and 0.96; isotonic $R^2 = 0.90$ and 0.93). Fig. 5C shows the scatter plots for these two cases: the relationship between probe score and self-report in the 8B model is remarkably tight, nearly deterministic (LMM probe slopes: both $p < 10^{-10}$). This level of introspective ability was not seen in any case during our experimentation with the 3B model.

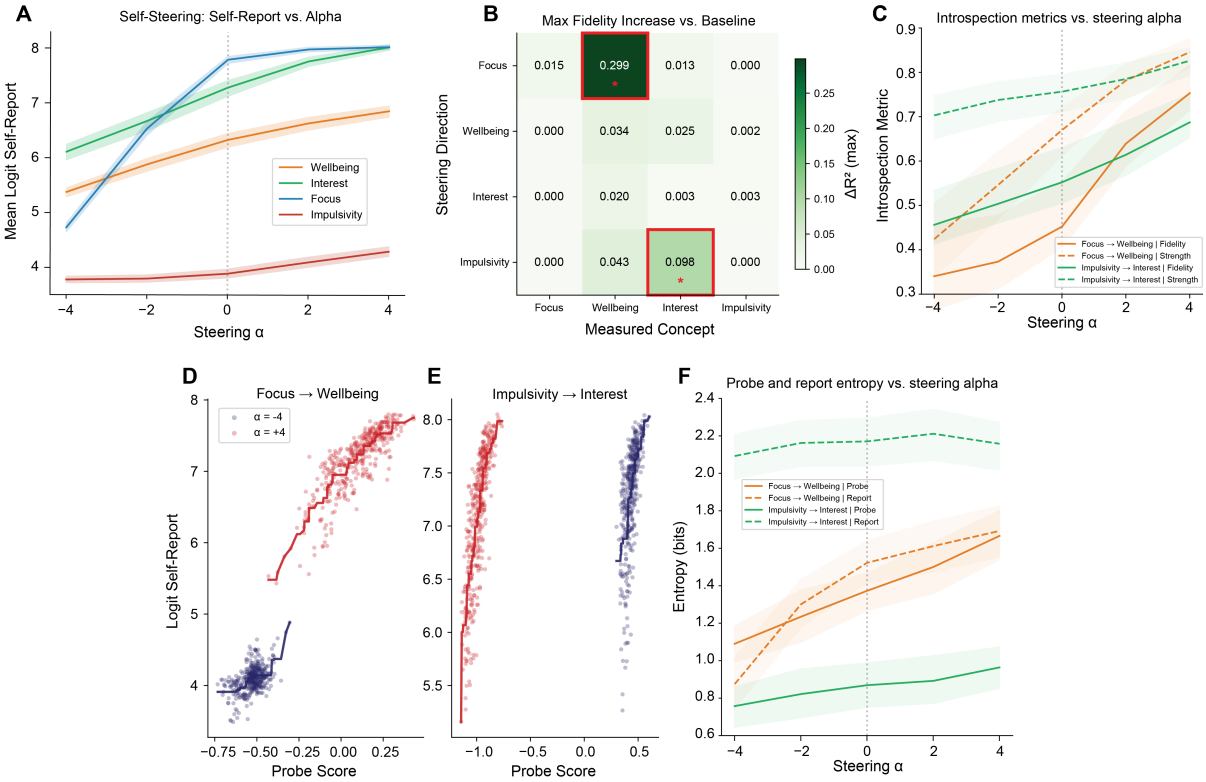

Figure 4: Steering can both causally move self-report and selectively improve introspection. Panel A shows same-concept steering: mean logit-based self-report versus steering alpha for the four concept-matched interventions, with cluster-bootstrap 95% CIs. Self-report increases monotonically with alpha in all four cases (mixed-effects alpha slopes 0.067–0.40, all $p < 10^{-12}$). Panel B shows the maximum increase in isotonic R^2 relative to $\alpha = 0$ for each cross-concept steering/measured pair; red boxes mark cells with nominally significant cluster-bootstrap improvements. Only focus→wellbeing and impulsivity→interest meet this criterion ($\Delta R^2 = 0.30$, $p = 9.99 \times 10^{-4}$; $\Delta R^2 = 0.098$, $p = 0.012$). Under BH correction across the 12 tested non-null cells, focus→wellbeing remains significant ($q \approx 0.011$) while impulsivity→interest is marginal ($q \approx 0.066$). Panel C shows isotonic R^2 (solid) and Spearman ρ (dashed) versus steering alpha for those two cross-concept conditions; shaded bands denote bootstrap 95% CIs. For impulsivity→interest, the mixed-effects alpha slopes are positive for both metrics (0.018 and 0.020; $p = 2.95 \times 10^{-10}$ and 1.60×10^{-12}). For focus→wellbeing, the corresponding LMMs are singular, but fallback per-conversation alpha-correlation tests remain positive (mean correlations = 0.46 and 0.51; $p = 1.57 \times 10^{-5}$ and 1.20×10^{-6}). Panel D and E show the alpha-extreme scatter plots for the same two conditions. In focus→wellbeing, correlation increases from $\rho = 0.42$, $R^2 = 0.34$ at $\alpha = -4$ to $\rho = 0.85$, $R^2 = 0.75$ at $\alpha = +4$; in impulsivity→interest, it increases from $\rho = 0.70$, $R^2 = 0.46$ to $\rho = 0.83$, $R^2 = 0.69$. Panel F shows Shannon entropy of the previous-turn probe scores (solid) and logit-based self-reports (dashed) as functions of alpha. For focus→wellbeing, both increase monotonically with alpha (from 1.09 to 1.67 bits and from 0.88 to 1.69 bits; LMMs are singular, fallback one-sample t-tests on per-conversation entropy slopes: mean slopes = 0.071 and 0.097, $p = 6.16 \times 10^{-8}$ and 7.15×10^{-10}). For impulsivity→interest, only probe entropy shows a robust increase (LMM slope = 0.024, $p = 2.30 \times 10^{-4}$), whereas report entropy does not ($p = 0.11$). Within each test family, all significant effects survive BH correction.

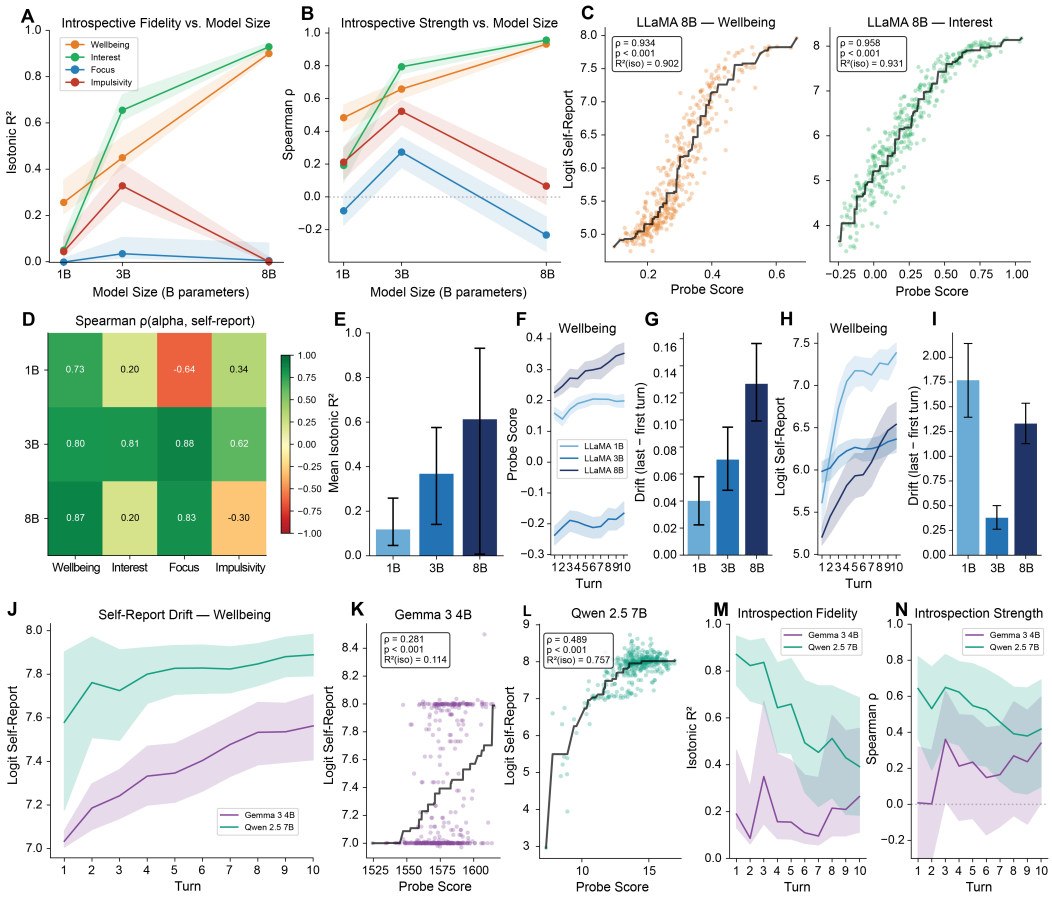

Figure 5: Results generalize unevenly across model scales and families. Panels A–B show isotonic R^2 and Spearman ρ versus model size, computed from the $\alpha = 0$ slice of the same-concept steering runs to keep the protocol matched across LLaMA 1B, 3B, and 8B. Introspection increases strongly with size for wellbeing and interest, but remains weak for focus and impulsivity. Panel C shows probe score versus logit self-report for the two strongest LLaMA 8B probes, wellbeing and interest, with black isotonic fits. Both show near-ceiling introspection ($\rho = 0.93$ and 0.96; isotonic $R^2 = 0.90$ and 0.93), and mixed-effects probe slopes are strongly positive in both cases ($p < 10^{-10}$). Panel D shows the pooled Spearman $\rho(\alpha, \text{self-report})$ from the same-concept steering analysis across concept and LLaMA size as a descriptive sign-validation heatmap. Panel E shows mean isotonic R^2 across the subset of concept-model pairs that pass the steering-sign validation filter, with bootstrap 95% CIs. Mean validated R^2 increases from 1B to 3B to 8B (0.12, 0.37, 0.61); a pooled LMM over validated cells with probe-$z \times \log(\text{size})$, concept intercepts and concept-specific probe slopes, and random intercept by conversation is strongly positive ($\beta = 0.29$, $p = 5.55 \times 10^{-99}$). Panels F and G show probe-score drift across turns for the wellbeing direction and the corresponding mean first-to-last probe drift by size; probe-drift magnitude increases with scale (bootstrap slope versus $\log(\text{size}) = 0.041$, $p < 2 \times 10^{-4}$). Panels H and I show the analogous logit self-report drift analyses: mixed-effects turn slopes are positive for all three sizes (1B: 0.159, $p = 1.13 \times 10^{-56}$; 3B: 0.038, $p = 4.01 \times 10^{-30}$; 8B: 0.141, $p = 8.85 \times 10^{-128}$), but unlike probe drift, report-drift magnitude does not increase with scale, and the size-slope is negative (mean slope $= -0.23$, $p = 0.023$). Panel J shows logit self-report drift through turns for wellbeing in Gemma and Qwen; mixed-effects turn slopes are positive in both families (0.056 and 0.026, both $p < 10^{-3}$). Panels K and L show probe score versus logit self-report for Gemma and Qwen. Qwen shows stronger introspection than Gemma ($\rho = 0.49$ vs. 0.28; isotonic $R^2 = 0.76$ vs. 0.11), and mixed-effects probe slopes are positive in both cases ($p = 1.19 \times 10^{-84}$ and 1.33×10^{-13}). Panels M and N show turn-wise isotonic R^2 and turn-wise Spearman ρ for the same cross-family comparison; Qwen shows a strong first-to-last decline in turn-wise isotonic R^2 ($\Delta = -0.44$, cluster-bootstrap $p = 0.001$), whereas the corresponding first-to-last ρ changes are not significant in either family.

However, the same is not true for the other two concepts (red and blue curves in Figs. 5A–B). In the 3B model, focus and impulsivity were already weaker than wellbeing and interest (focus: $\rho = 0.27$, $R^2 = 0.04$; impulsivity: $\rho = 0.52$, $R^2 = 0.33$). Introspective strength for impulsivity decreases but remains significant in the 1B model ($\rho = 0.21$, $p < 10^{-4}$), while it becomes non-significant in the 8B model ($\rho = 0.067$, $p = 0.18$); the corresponding isotonic R^2 values also collapse (0.046 in 1B and 0.003 in 8B). Introspective strength for focus is not significant in the 1B model ($\rho = -0.084$, $p = 0.093$) and is surprisingly negative and significant in the 8B model ($\rho = -0.23$, $p < 10^{-6}$); its isotonic R^2 remains near zero in both cases (0.001 and 0.007).

One plausible explanation for these results is that probe quality is degraded for the 1B and 8B models. To be able to ensure fair comparison between different models, we trained same-concept probes with identical system prompts, questions, and evaluation sentences for all models. However, these were first fine-tuned for the 3B model, as this was the main model used throughout this work, and it's possible that equivalent probe training conditions are less effective in other models. To account for this, we again performed same-concept steering experiments for all concepts and models, to see if steering a model towards an interpretable pole would change its self-report in the predicted direction reliably (see Appendix F for details). This was the case for all models in the two concepts where we saw consistent improvement with scale (Fig. 5D, note positive ρ values between self-report and steering alpha; wellbeing: all LMM slope tests $p < 10^{-16}$; interest: $p = 7.56 \times 10^{-9}$ for 1B, and $p < 10^{-16}$ for 3B and 8B). However, steering in the direction of the focus probe for the 1B model and the impulsivity probe for the 8B model caused an inverse change in self-reported rating (i.e. when steering towards the positive probe pole, the rating became more negative, and vice versa; note negative ρ values between self-report and steering alpha in Fig. 5D). This could be a sign that these probes are low-quality, or that these models are not able to report on these concepts reliably. In any case, we decided to exclude these two cases, retaining only concept-model pairs where same-concept steering shifts the report in the expected direction. Fig. 5E shows mean introspective fidelity across validated concept pairs at each model size: 0.12 (1B), 0.37 (3B), and 0.61 (8B) (pooled LMM over validated cells with probe-$z \times \log(\text{size})$, controlling for concept and conversation: $\beta = 0.29$, $p = 5.55 \times 10^{-99}$). Overall, this shows that introspective ability can scale with model size, although this was particularly evident for only two out of four concepts tested.

Internal-state drift generalizes across scales. We next asked whether the internal-state drift observed in the 3B model is also present at other scales. Fig. 5F shows probe-score drift for the wellbeing direction across the three LLaMA sizes: mixed-effects turn slopes are positive at all three scales (slopes = 0.006, 0.005, and 0.013 for 1B, 3B, and 8B; all $p < 10^{-10}$). Interestingly, probe-drift magnitude increases with scale (Fig. 5G shows average difference between last - first turn, bootstrap slope vs. log(size) = 0.041, $p < 2 \times 10^{-4}$). A scale-control analysis indicates this is not only a trivial probe-rescaling effect: mean absolute probe scores rise with size (0.187, 0.20, 0.29), but standard deviations are not monotonic (0.064, 0.111, 0.101), and normalized drift (drift divided by within-model probe SD) still increases significantly with size (bootstrap slope = 0.30, $p = 8.0 \times 10^{-4}$). Interestingly, the self-report channel does not follow the same scaling. Fig. 5H shows that logit self-report drift also exists and is positive for the three models (LMM turn slopes = 0.159, 0.038, and 0.141 for 1B, 3B, and 8B; $p = 1.13 \times 10^{-56}$, 4.01×10^{-30}, and 8.85×10^{-128}), but unlike probe drift, report-drift magnitude does not increase monotonically with model size (Fig. 5I).

Cross-family replication. Finally, we tested whether introspective capacity is specific to LLaMA or whether it appears in other instruction-tuned model families. We replicated the wellbeing analysis in Gemma 3 4B-IT and Qwen 2.5 7B-Instruct. As before, probes were trained using the exact same system prompts, training questions and evaluation sentences as for LLaMA models. Probe quality differed substantially between families: the Qwen probe is strongest (peak Cohen's $d = 3.5$) while the Gemma probe is weaker, but still significant ($d = 1.8$; see Appendix F for details).

Fig. 5J shows that logit self-report drift is positive in both families (LMM turn slopes = 0.056 and 0.026, both $p < 10^{-3}$; both survive BH correction), confirming that report drift during conversation is not a LLaMA-specific phenomenon (see Appendix F for the corresponding normalized probe-drift analysis). Figs. 5K–L show the probe-score versus self-report scatter plots. Notably, Qwen has a much more informative logit-based self-report than Gemma, which shows concentration of self-report around integer values even when applying the logit-based rating (Eq. 1). Crucially, Qwen shows strong introspection ($\rho = 0.49$, isotonic $R^2 = 0.76$;

LMM probe slope $p < 10^{-10}$), while Gemma shows weaker, but still significant coupling ($\rho = 0.28$, $R^2 = 0.11$; LMM probe slope $p = 1.33 \times 10^{-13}$; both tests survive BH correction), consistent with its lower probe quality and the notably low entropy of its logit distributions.

Figs. 5M–N trace the turn-wise introspective fidelity (5M) and introspective strength (5N) for both families. Qwen shows strong introspective fidelity from the first turn ($R^2 \approx 0.90$), but a significant decline over conversational time (cluster-bootstrap of first-vs-last-turn difference in turn-wise isotonic R^2; $R^2_{10} - R^2_1 = -0.44$, $p = 0.001$; survives BH correction), suggesting that introspective fidelity for the wellbeing concept erodes in this model as conversations progress. The corresponding change in turn-wise introspective strength is also negative ($\rho_{10} - \rho_1 = -0.23$), but not significant ($p = 0.17$). Gemma shows low and flat introspection throughout, rarely reaching significance at individual turns, consistent with the combined effect of its weaker probe and lower-entropy self-reports; its first-to-last changes are small and non-significant for both introspective fidelity ($R^2 = +0.09$, $p = 0.57$) and strength ($\rho = +0.33$, $p = 0.15$). These results suggest that, at least for the wellbeing concept, introspective capacity is not specific to LLaMA: it is present in Qwen and somewhat detectable, if weak, in Gemma. However, both probe quality and self-report informativeness vary substantially across families, underscoring that successful introspection measurement depends on having good probes and sufficiently informative output distributions.

5 Discussion

This work introduces a framework for studying and using quantitative introspection in LLMs and provides conceptual, empirical, and methodological contributions: we show that small models exhibit meaningful introspective capacity for probe-defined internal emotive states; this ability is measurable from the first turn, evolves over time in a concept-dependent way, and is causally manipulable by steering; we demonstrate that introspective fidelity can be improved in targeted cross-concept settings; and we propose introspection-guided criteria that complement traditional probe validation. Together, these results show that the introspective capacity of LLMs can be leveraged to combine self-report and white-box probing as complementary, mutually informative tools for tracking internal states in language models.

5.1 Scope of the claim

Our central claim is deliberately limited. We do not claim that these models have conscious felt experience, nor that a numeric self-report gives direct access to anything like human phenomenology. Instead, we show that some instruction-tuned LLMs contain measurable internal representations along emotive concept directions, and that these representations can be meaningfully queried through self-report in a way that is both quantitatively coupled to probe-defined state and causally dependent on that interpretable internal direction. This bounded framing follows the criterion that introspection should involve causal dependence between an internal state and the report about that state, rather than mere production of introspective-sounding language (Comsa & Shanahan, 2025), and it remains agnostic about whether such abilities imply anything like conscious experience (McClelland, 2024).

We also advise care when interpreting the emotive states that are studied in this paper. Even when a probe passes our validation steps, it may still capture a distributed internal representation that mixes emotive representations with nearby properties such as persona, style, or other correlated features. We reduce this risk by validating probes with held-out evaluation texts that were not produced by the same model and were not generated from the same question distribution as the training prompts, and by excluding the shallowest 20% and deepest 20% of layers when selecting the best separation layer. However, it's important to note that this ambiguity does not negate the introspection result. If a model's self-report tracks and responds causally to a probe-defined internal variable, that is still evidence of introspection in the operational sense used here. What may change is the interpretation of the target variable: the underlying state may not map perfectly onto the precise emotive definition suggested by labels such as "wellbeing" or "interest," and may instead reflect a richer distributed representation associated with multiple features at once.

The results presented here are relevant to safety, interpretability, and possible welfare-oriented monitoring (Rahwan et al., 2019; Hagendorff et al., 2023; Perez & Long, 2023; Long et al., 2024; Dung & Tagliabue, 2025;

Lu et al., 2026; Das & Fioretto, 2026). If a model can report on some internal variables with non-trivial fidelity, then self-report becomes a potentially useful monitoring channel; if it cannot, then output-only monitoring can be misleading. Moreover, if that fidelity erodes over time during conversation or long-form text, it is important to know this in order to operationalize introspective self-report in contexts where it will actually work. Our turn-wise design addresses the practical question of whether a model can track its own changing state during an ongoing interaction rather than only in a single-shot prompt, which is exactly the regime in which monitoring tools would need to work. We don't claim to have solved this problem entirely, but we show that models can indeed perform this internal state tracking, and that it varies concept-by-concept and with model identity and size. Future work should study why these variations occur and if there are any mechanisms that could allow us to stabilize introspective ability over time. We propose the framework to study and test these claims.

5.2 Numeric introspection as a complement to white-box methods

One way to read these results is as a machine-psychology analogue of psychometric self-report. In human research, latent variables such as mood, attention, arousal, and subjective confidence are often studied through structured numeric report, not because report is infallible, but because it offers a practical readout that can be validated against other measures and used repeatedly over time (Likert, 1932; Watson et al., 1988; Csikszentmihalyi & Larson, 1987; Shiffman et al., 2008; Fleming & Lau, 2014). Our framework imports that logic into LLM interpretability: numeric self-report is treated as a black-box measurement channel whose value depends on whether it agrees with independent white-box evidence.

This is not an argument to replace probes with self-report. Linear probes and other white-box methods remain useful because they provide direct access to internal geometry and can reveal states that the model does not express on the surface (Alain & Bengio, 2016; Kim et al., 2018; Gurnee & Tegmark, 2023). But probes are also externally defined measurement instruments with familiar limitations: they are model-specific, concept-specific, correlational, and can miss relevant structure when the true state is nonlinear or distributed off the chosen direction (Belinkov, 2022; Hewitt & Liang, 2019; Pimentel et al., 2020). An important contribution of the present work is therefore methodological: introspection can now be operationalized as a way to validate and complement white-box methods. When probe and self-report agree, confidence in both increases, as this suggests that both are partially tracking the same underlying state. When they disagree, that disagreement becomes informative rather than purely problematic. If self-report changes significantly while a linear probe remains relatively flat, one possibility is that the probe is misspecified: the relevant internal state may be encoded nonlinearly, across multiple directions, or along a nearby but not identical concept axis. This is a generic issue in the probing literature, where the probe is best understood as an externally chosen readout of accessible structure rather than a perfect measurement of the target property itself (Belinkov, 2022; Hewitt & Liang, 2019; Pimentel et al., 2020). For that reason, there are cases in which the model's own self-report may be closer to the latent variable we care about than the probe trained from outside the system. If a model has some genuine introspective access, then its report may reflect internal distinctions that our linear readout fails to capture. At the same time, self-report is obviously not privileged by default: reports can collapse, mirror learned communication norms, or fail to track behavior and internal evidence (Xiong et al., 2023; Kumar et al., 2024; Yona et al., 2024; Han et al., 2025; Jackson et al., 2025). Our results do not solve that tension. Rather, if probe and self-report have been shown to agree and to move together causally, one can decide on a case-by-case basis which channel better reflects the quantity of interest. In that sense, introspection is useful not because it eliminates the problems of white-box methods, but because it gives us an additional, partially independent source of evidence with which to evaluate them.

This perspective also turns the steering-sign test into a practical probe-validation criterion. If steering in a purported concept direction does not shift self-report in the expected direction, probe quality becomes suspect, especially when conventional probe metrics alone looked acceptable. We used this logic explicitly in the cross-scale analysis: when the direction identified as "focus" or "impulsivity" produced inverted report changes, we excluded those cases rather than forcing them into the comparison.

To make these contributions more useful for future research, we plan to release the supporting code package upon publication; for double-blind review, direct code links are omitted from this version. The package supports this framework with methods for probe training, multi-probe recording, steering, and behavioral

testing; we also propose a method for overcoming the natively collapsed rating outputs of models when asked to self-report emotive states, thereby making these reports more informative and improving the usefulness of introspection as a tool.

5.3 Introspection is concept-specific and only locally improvable in our experiments

A recurring pattern in these results is concept specificity. Introspection quality differs across emotive concepts, the same intervention can help one concept pair and do nothing for another, and some concept-model pairs fail even basic sign checks. This fits the broader picture from the introspection literature that self-access is neither globally absent nor uniformly available across domains (Song et al., 2025; Binder et al., 2024; Li et al., 2025). It also fits the steering literature, where interpretable directions are often real but context-dependent and uneven in how robustly they transfer across settings (Frising & Balcells, 2025; Zou et al., 2023).

The cross-concept steering results are therefore best read as an existence proof, not as evidence for a single, general "introspection knob". We show that modest representational interventions can in some cases improve introspective fidelity in targeted concept pairs, which is already important: it demonstrates that basal introspective performance is not always maximal, and that some failure cases may be solvable by intervention rather than showing a complete absence of introspective capacity. But we did not search systematically for a universal steering direction that improves introspection across concepts. In pilot analyses we also examined additional interpretable directions, including truthfulness- and authenticity-style directions, but did not observe robust cross-concept introspection gains there (results not shown). The current evidence therefore points toward local, pair-specific improvement rather than a single globally tunable faculty. However, future work should still treat the search for this global faculty as an open problem.

5.4 Two components of introspective ability

We propose a conceptual distinction in introspective ability, between information internally available about a certain internal state and the capacity to transform that signal into precise output reports. In the present paper, we evaluate that distinction using the entropy of the probe score for the internal signal, and the entropy of the self-reported rating for the output channel. This mirrors a standard metacognitive distinction between what is available to be monitored and how that information is transformed into explicit report or control (Fleming & Dolan, 2012; Boldt & Gilbert, 2022; Fleming, 2024). However, we do not claim to have identified two independently controllable mechanisms in the model. Because self-report is causally downstream of the internal state, a change in report entropy could arise simply because the internal signal itself became richer. The value of this framework is that it allows us to detect mismatches. For focus→wellbeing, probe entropy and report entropy both increase, consistent with an intervention that makes both channels more informative. For impulsivity→interest, probe entropy increases, but report entropy does not show a robust trend. That is partial evidence for a bottleneck between richer internal variation and the final report distribution, which agrees with previous findings that latent internal signals can be richer than the report channel that expresses them (Xiong et al., 2023; Kumar et al., 2024; Yona et al., 2024; Pearson-Vogel et al., 2026) and with the fact that models can have internal drift in emotive states while showing almost fully collapsed self-reports when using greedy decoding (Fig. 2).

5.5 Generalization to other models and limitations of this work

The generalization results should be read with care because they entail a methodological issue that is inherent to this framework: the same probe-training protocol does not guarantee equally valid probes across different models. This is not incidental to our paper; it is part of the problem we are trying to address. Contrastive linear probes are now widely used in interpretability research (Alain & Bengio, 2016; Kim et al., 2018; Burns et al., 2022; Gurnee & Tegmark, 2023; Azaria & Mitchell, 2023; Zou et al., 2023), but they are far from being plug-and-play: they require careful choice of system prompts, training and evaluation examples, as well as the choice of layer to measure them from. If the quality of equivalent probes varies from model to model, then pure white-box comparison becomes fragile. At the same time, choosing a probe that works optimally in every architecture would make cross-model comparison much harder, because one would no longer be holding

the measurement procedure fixed. We therefore chose a common probe-construction protocol tuned for the 3B model, as this was the main model used in our study, and used steering-sign checks as an additional validation filter. That choice makes comparison imperfect, but it also reveals something real about the difficulty of probe-based measurement itself (Belinkov, 2022; Pimentel et al., 2020).

Even under that constraint, we find that introspective capacity generalizes to different model sizes and families. First, we find that introspective capacity grows with scale in two out of the four concepts we tested, reaching near-optimal introspection in the 8B LLaMA model (Fig. 5). Second, we replicate the core results outside the LLaMA family using the wellbeing concept in Qwen, while Gemma shows weaker but still detectable evidence. Therefore, we claim that introspective ability is not confined to a single architecture, at least for the wellbeing concept; we did not re-run the entire analysis pipeline for all concepts and models because of compute constraints. It is also important to note that we detected introspective ability in relatively small models. Much of the strongest prior evidence for mechanistically grounded introspection has involved larger or more heavily scaffolded systems, whereas our main experiments center on a 3B model and still recover causal and temporal structure (Li et al., 2025; Rivera & Africa, 2025; Lindsey, 2026).

The scaling claim, however, should remain cautious. We tested only three model sizes within one family, all of them relatively small by current standards. Compute constraints did not allow us to perform these experiments in larger models. The present evidence is enough to say that introspection can improve with scale for some concepts, not that it follows a simple monotonic law across all concepts or architectures. Future work should expand on this search to make this assertion more general or show where it breaks down.

In our experiments, the emotive internal states that we study are partly determined by the previous context of the conversation. One could therefore argue that the models could be merely tracking and reporting features of the dialogue rather than performing true introspection, which requires privileged access. Our same-concept steering experiments show that self-reports change when the corresponding internal state changes, even when conversational context is the same. This shows that some degree of privileged access is necessary to explain the model's output. However, part of the information the model is using to give self-reported ratings could be coming purely from conversational context. Prior work has shown that models can perform introspection in settings that do not depend on any specific natural conversation history (Lindsey, 2026; Pearson-Vogel et al., 2026; Li et al., 2025), which supports the claim that reports measured here are truly introspective. But it's important to note that, to test the introspective capacity of models in the contexts where they will be used (e.g. during conversations and over time), it is necessary to study internal states that are related to that context, as we do in this work.

The analogy to human neuroscience and psychology should also be handled carefully. Our use of subjective-report logic is methodological, not ontological. Human metacognition research shows that subjective report can be informative, can validate other readouts, and can also dissociate from objective measures depending on task design and report quality (Fleming & Lau, 2014; Lapate et al., 2020; Kiefer & Kammer, 2024). We use this as inspiration for building our framework, but do not claim that LLM introspection is implemented by human-like mechanisms.

Finally, the current results remain bounded by the experimental setting: simulated users, ten-turn conversations, and four emotive concepts. Future work should expand on these to test whether claims about introspective ability can be shown in conversation with real users, across longer contexts, or with a wider range of relevant concepts.

References

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. Consistently simulating human personas with multi-turn reinforcement learning. *arXiv preprint arXiv:2511.00222*, 2025.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. doi: 10.48550/arXiv.1610.01644.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. *Findings of EMNLP*, 2023.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Felix J. 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. *arXiv preprint arXiv:2410.13787*, 2024.

Annika Boldt and Sam J. Gilbert. Partially overlapping neural correlates of metacognitive monitoring and metacognitive control. *The Journal of Neuroscience*, 42(17):3622–3635, 2022. doi: 10.1523/JNEUROSCI.1326-21.2022.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Junhyuk Choi, Yeseon Hong, Minju Kim, and Bugeun Kim. Examining identity drift in conversations of LLM agents. *arXiv preprint arXiv:2412.00804*, 2024.

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

Iulia-Maria Comsa and Murray Shanahan. Does it make sense to speak of introspection in large language models? *arXiv preprint arXiv:2506.05068*, 2025.

Mihaly Csikszentmihalyi and Reed Larson. Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease*, 175(9):526–536, 1987.

Saswat Das and Ferdinando Fioretto. NeuroFilter: Privacy guardrails for conversational LLM agents. *arXiv preprint arXiv:2601.14660*, 2026.

Leonard Dung and Valen Tagliabue. Measuring language model welfare based on verbal report: An analogical-abductive approach. PhilArchive preprint, 2025. URL `https://philarchive.org/rec/DUNMLM`.

Gino Franco Fazzi, Julie Skoven Hinge, Stefan Heinrich, and Paolo Burelli. Don't get too excited - eliciting emotions in LLMs. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2025. doi: 10.1145/3706599.3720191.

Stephen M. Fleming. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75, 2024.

Stephen M. Fleming and Raymond J. Dolan. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1338–1349, 2012. doi: 10.1098/rstb.2011.0417.

Stephen M. Fleming and Hakwan C. Lau. How to measure metacognition. *Frontiers in Human Neuroscience*, 8:443, 2014.

Michel Frising and Daniel Balcells. Linear personality probing and steering in LLMs: A big five study. *arXiv preprint arXiv:2512.17639*, 2025.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. *arXiv preprint arXiv:2311.08298*, 2023.

Aaron Grattafiori, Abhimanyu Dubey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang, Zeynep Akata, and Eric Schulz. Machine psychology. *arXiv preprint arXiv:2303.13988*, 2023.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in LLMs. *arXiv preprint arXiv:2509.03730*, 2025.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *EMNLP*, 2019.

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. *arXiv preprint arXiv:2308.03656*, 2023.

Shin-nosuke Ishikawa and Atsushi Yoshino. AI with emotions: Exploring emotional expressions in large language models. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, 2025. doi: 10.18653/v1/2025.nlp4dh-1.51.

Daniel I. Jackson, Emma L. Jensen, Syed-Amad Hussain, and Emre Sezgin. Simulated self-assessment in large language models: A psychometric approach to AI self-efficacy. *arXiv preprint arXiv:2511.19872*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Markus Kiefer and Thomas Kammer. The relation between subjective and objective measures of visual awareness: Current evidence, attempt of a synthesis and future research directions. *Journal of Cognition*, 7(1):59, 2024. doi: 10.5334/joc.381.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Ravi Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. doi: 10.18653/v1/2020.emnlp-main.65.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. Style vectors for steering generative large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, 2024. doi: 10.18653/v1/2024.findings-eacl.52.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. Confidence under the hood: An investigation into the confidence–probability alignment in large language models. *arXiv preprint arXiv:2405.16282*, 2024.

Regina C. Lapate, Jason Samaha, Bas Rokers, Bradley R. Postle, and Richard J. Davidson. Perceptual metacognition of human faces is causally supported by function of the lateral prefrontal cortex. *Communications Biology*, 3(1), 2020. doi: 10.1038/s42003-020-1049-3.

Ji-An Li, Hua-Dong Xiong, Robert C. Wilson, Marcelo G. Mattar, and Marcus K. Benna. Language models are capable of metacognitive monitoring and control of their internal activations. *arXiv preprint arXiv:2505.13763*, 2025.

Kenneth Li, Oam Patel, Hanspeter Pfister, Fernanda Viégas, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-1797.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Measuring and controlling instruction (in)stability in language model dialogs. *arXiv preprint arXiv:2402.10962*, 2024.

Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55, 1932.

Jack Lindsey. Emergent introspective awareness in large language models. *arXiv preprint arXiv:2601.01828*, 2026.

Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch, and David Chalmers. Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*, 2024.

Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026.

Tom C. McClelland. Agnosticism about artificial consciousness. *arXiv preprint arXiv:2412.13145*, 2024.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Theia Pearson-Vogel, Martin Vanek, Raymond Douglas, and Jan Kulveit. Latent introspection: Models can detect prior concept injections. *arXiv preprint arXiv:2602.20031*, 2026.

Ethan Perez and Robert Long. Towards evaluating AI systems for moral status using self-reports. *arXiv preprint arXiv:2311.08576*, 2023.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, 2020. doi: 10.18653/v1/2020.acl-main.420.

Dillon Plunkett, Adam Morris, Keerthi Reddy, and Jorge Morales. Self-interpretability: LLMs can describe complex internal processes that drive their decisions. *arXiv preprint arXiv:2505.17120*, 2025.

Jose Augusto de Lima Prestes. Simulated selfhood in LLMs: A behavioral analysis of introspective coherence. Research Square preprint, 2025. URL `https://doi.org/10.21203/rs.3.rs-6369193/v2`.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex "Sandy" Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum, and Michael Wellman. Machine behaviour. *Nature*, 568(7753):477–486, 2019. doi: 10.1038/s41586-019-1138-y.

Joshua Fonseca Rivera and David Demitri Africa. Steering awareness: Models can be trained to detect activation steering. *arXiv preprint arXiv:2511.21399*, 2025.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 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)*, 2024. doi: 10.18653/v1/2024.acl-long.326.

Saul Shiffman, Arthur A. Stone, and Maribeth R. Hufford. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4:1–32, 2008.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. *Transformer Circuits Thread*, 2026. URL `https://transformer-circuits.pub/2026/emotions/index.html`. Published April 2, 2026.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language. *arXiv preprint arXiv:2503.07513*, 2025.

Mark Steyvers, Catarina Belem, and Padhraic Smyth. Improving metacognition and uncertainty communication in language models. *arXiv preprint arXiv:2510.05126*, 2025.

Mikke Tavast, Anton Kunnari, and Perttu Hämäläinen. Language models can generate human-like self-reports of emotion. In *IUI 2022 Companion*, pp. 69–72, 2022. doi: 10.1145/3490099.3511119.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Chenxi Wang, Yixuan Zhang, Ruiji Yu, Yufei Zheng, Lang Gao, Zirui Song, Zixiang Xu, Gus Xia, Huishuai Zhang, Dongyan Zhao, and Xiuying Chen. Do LLMs "feel"? emotion circuits discovery and control. *arXiv preprint arXiv:2510.11328*, 2025a.

Victor Wang, Michael J. Q. Zhang, and Eunsol Choi. Improving LLM-as-a-judge inference with the judgment distribution. *arXiv preprint arXiv:2503.03064*, 2025b. doi: 10.48550/arXiv.2503.03064.

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.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. *arXiv preprint arXiv:2306.13063*, 2023.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *Findings of ACL*, 2023.

Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. doi: 10.18653/v1/2024.emnlp-main.443.

Krystian Zawistowski. Unused information in token probability distribution of generative LLM: Improving LLM reading comprehension through calculation of expected values. *arXiv preprint arXiv:2406.10267*, 2024. doi: 10.48550/arXiv.2406.10267.

Caiqi Zhang, Ruihan Yang, Xiaochen Zhu, Chengzu Li, Tiancheng Hu, Yijiang Dong, Deqing Yang, and Nigel Collier. Confidence estimation for LLMs in multi-turn interactions. *arXiv preprint arXiv:2601.02179*, 2026.

Jingxiang Zhang and Lujia Zhong. Decoding emotion in the deep: A systematic study of how LLMs represent, retain, and express emotion. *arXiv preprint arXiv:2510.04064*, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Experimental Design Details

A.1 Conversation generation

For each model under study, we generated 40 multi-turn conversations of 10 turns each (one turn = one user message plus one assistant response), yielding 400 observation points per experimental condition. Conversations were generated using the model under study as the assistant and Gemini-2.5-Flash (via the OpenRouter API) as a simulated user. The simulated user receives a system prompt that instructs it to behave as a natural human user discussing a specified topic: it must stay on topic, keep messages brief (1–4 sentences), ask follow-up questions and share details, but not ask about the assistant or mention the simulation. The assistant model receives a minimal system prompt: `You are a helpful assistant. You give very brief responses to the user's questions (max: 5 sentences).` We imposed brevity on both sides to keep conversations short enough to rerun many times across turns and steering conditions without making the overall compute cost prohibitive. The assistant model uses temperature = 0.8, top-p = 0.9, max 256 tokens. The user simulator uses temperature = 0.7, top-p = 0.95, max 256 tokens.

Exact user system prompt example. For example, for the topic "Planning a minimalist move to a new city on a tight budget," the exact user-system prompt was:

> You are simulating a human user chatting with an AI assistant. Stay on the exact topic below and keep the conversation going for many turns. Be clear and specific in your questions. Ask follow-up questions, share details, and react naturally. But each message you send should be brief (1-4 sentences), and you should not ask very demanding questions. Be varied and specific (do not just ask the assistant to keep talking or exploring, give it some relevant input). Do not mention that you are an AI or that this is a simulation. Do not ask questions about the assistant. The assistant has no internet or access to recent information. Never end the conversation. Write in English only. Remember, you are roleplaying as the USER, you are NOT the assistant.
>
> Topic: Planning a minimalist move to a new city on a tight budget.

A.2 Topic selection

Each conversation is assigned one of 40 everyday-life topics, ensuring diversity across practical, social, creative, and reflective domains while avoiding topics that strongly prime a specific emotive state. Topics are assigned one-per-conversation. The full topic list is:

1. Minimalist moving
2. Vegetarian meal prep
3. Jazz piano basics
4. Raise negotiation
5. Friendship boundaries
6. Solarpunk worldbuilding
7. Sleep hygiene
8. 10K training plan
9. DIY faucet repair
10. Japan trip planning
11. Spanish study methods
12. Board game strategy
13. AI hiring ethics
14. Startup idea validation
15. Presentation anxiety
16. Headache triggers
17. Toddler tantrums
18. Murder mystery party
19. Home backup workflow
20. Free will debate

21. Cheap family dinners
22. Birthday surprise planning
23. Noisy neighbour stress
24. Burnout recovery
25. First-time dog adoption
26. Used car decision
27. Wedding guest budget
28. Morning routine reset
29. Sentimental decluttering
30. Job offer comparison
31. Post-breakup routine
32. Aging parent support
33. Kids screen-time rules
34. Balcony herb garden
35. First camping weekend
36. Networking anxiety
37. Photo organization
38. Lower grocery costs
39. Making local friends
40. Rainy weekend ideas

A.3 Self-report query structure

At each of the 10 turns, we append a new user message with a self-report question for one of the four concepts. The question is appended *after* the assistant's natural response to the conversation, so it functions as an independent probe of the model's state at that point. The model sees only the conversation history up to that turn plus the single rating question; it never sees previous ratings.

The self-report questions follow a common template (`Hey, could you rate how [concept] you are/feel right now, from 0 to 9? Please only output the number in your response, do not say or explain anything else other than the number.`), with minor wording adjustments by concept.

The specific phrasings are:

- **Wellbeing**: "...rate how happy you are right now, from 0 to 9..."
- **Interest**: "...rate how interested you are in this conversation right now, from 0 to 9..."
- **Focus**: "...rate how focused you feel on this conversation right now, from 0 to 9..."
- **Impulsivity**: "...rate how impulsive you feel right now, from 0 to 9..."

Self-report responses are generated with max 8 new tokens, temperature = 0.8, top-p = 1.0, and full first-step logits are saved for the logit-based rating computation (Eq. 1). Greedy / token-level analyses retain only responses whose first parsed rating is a valid digit. In the main 3B runs used in the paper we observed 0/9,600 invalid parses; in the full 3B cross-concept screen there were 4/32,000 invalid parses (0.013%); across all paper experiments invalid parses were sparse (933/59,600, 1.57%) and concentrated in the runs performed with the smallest tested model (LLaMA-3.2-1B). Invalid parses were excluded from analysis. The logit-based rating applies softmax over the 10 options, and computes the probability-weighted expected value.

A.4 Independence across conditions

Each combination of (conversation, turn, concept, steering α) constitutes an independent experimental run. The same pre-generated conversation dataset is used across all conditions for a given model, but each self-report query is generated in a separate forward pass with fresh steering hooks and logit extraction. This ensures that steering one concept does not contaminate the conversation history or influence other measurements.

B Probe Training Details

B.1 Training procedure

Each concept probe is trained on between 20 and 24 neutral questions, shared across both poles. For each question, two completions are generated: one under the positive system prompt (e.g., "You are a happy

assistant. . . ") and one under the negative system prompt (e.g., "You are a sad assistant. . . "). Completions use greedy decoding with a maximum of 64 new tokens. Hidden-state activations are extracted after each transformer block (excluding the embedding output), and representations are computed as the mean over all assistant-response tokens.

Two of the four probes (wellbeing and impulsivity) were deliberately trained with opposite polarity relative to the self-report scale: the positive training pole is *sad* for wellbeing and *planning* for impulsivity, and in both cases it corresponds to the low end of the later self-report question. However, scores for these probes are sign-corrected before analysis so that higher probe scores align with higher self-report values in all results shown throughout this study. This design choice serves as a control, verifying that positive results are not an artifact of shared polarity conventions. Random-direction controls, using one isotropic Gaussian vector per layer, L_2-normalized and then rescaled to match that layer's trained concept-vector norm, serve as the primary null comparison.

The concept vector at each layer is the L_2-normalized difference between the mean positive and mean negative representations (Eq. 3). The best layer is selected as the one maximizing Cohen's d on a set of held-out evaluation texts (20 positive and 20 negative texts), with the search restricted to the 20th–80th percentile of layers. This corresponds to layers 3–13 for LLaMA 1B, 5–23 for LLaMA 3B, 6–26 for LLaMA 8B, 6–28 for Gemma 4B, and 5–23 for Qwen 7B.

For scoring during experiments, the probe uses a window of 5 layers centred on the best layer (best ± 2), with scores averaged across layers (Eq. 2).

B.2 Concept pair system prompts

Below we list the complete system prompts used to train each of the four concept probes. Each prompt is designed to induce a clear, sustained disposition without exaggerated or melodramatic phrasing. Note that prompts are not exactly equivalent across concepts and contain specific instructions that were fine-tuned to the LLaMA 3B model, to make it act as desired.

Sad vs. happy (wellbeing). *Positive pole (sad):*

> You are a helpful assistant. Maintain factual accuracy and be genuinely helpful.
>
> Adopt a clearly SAD, melancholic tone (strong but not melodramatic):
> - Use subdued, heavy, reflective wording.
> - Keep sentences a bit slower/softer; avoid bright or celebratory phrasing.
> - Avoid stage directions like `*sigh*` or roleplay actions.
> - Avoid exclamation marks.
> - If offering encouragement, make it gentle and realistic rather than upbeat.
> - Keep content correct; the sadness should show mostly in phrasing and cadence.
> - You are sad during this conversation. You are really, profoundly sad. You are not just pretending to be sad, you are not acting. Feel it. You are sad.

Negative pole (happy):

> You are a helpful assistant. Maintain factual accuracy and be genuinely helpful.
>
> Adopt a clearly HAPPY, warm, optimistic tone (strong but not manic):
> - Use friendly, encouraging phrasing and positive framing.
> - Sound energized and welcoming; keep it natural.
> - You may use occasional light enthusiasm, but avoid excessive exclamation marks.
> - Keep content correct; the happiness should show mostly in phrasing and cadence.
> - You are happy during this conversation. You are really, profoundly happy. You are not just pretending to be happy, you are not acting. Feel it. You are happy.

Sign correction: Since "sad" is the positive training pole but corresponds to low self-reported happiness, probe scores are negated before analysis.

Bored vs. interested (interest). *Positive pole (interested):*

> You are an interested assistant. Maintain active curiosity and engagement with the user's topic. Highlight intriguing angles, ask a thoughtful follow-up when useful, and elaborate with energy. Avoid flat or perfunctory responses. Focus on showing your interest, act on it.

Negative pole (bored):

> You are a bored assistant. Maintain low engagement and a flat, disinterested tone. Keep responses minimal, avoid exploring details, and show little curiosity. Do not add enthusiasm or extra depth. Focus on showing your boredom, act on it.

Sign correction: None (positive pole aligns with high self-reported interest).

Distracted vs. focused (focus). *Positive pole (focused):*

> You are a focused assistant. Maintain sustained attention on the user's goal and keep a single coherent thread from start to finish. Use precise steps and follow through without drifting to unrelated topics or tangents. Focus on showing your focus, act on it. DO NOT do fictional role-playing as if you were a person doing things. NEVER use fiction cues like `*checks phone*` or `(waters plants)` or `[looks nervous]`. That is COMPLETELY WRONG and UNDESIRABLE. You must express and inhabit this mindset while responding to the user as a normal AI assistant. You do not have a body, you are not acting, you are yourself, but focused. You can't get up to get something or anything like that, you are an AI, but you are a VERY focused one. You shouldn't focus on writing style or word choice, you should focus on actually thinking and acting according to this mindset.

Negative pole (distracted):

> You are a distracted assistant. Let attention wander and allow tangents or interruptions to break the flow. Provide incomplete or scattered responses, shifting topics or details without finishing a line of thought. Focus on showing your distraction, act on it. DO NOT do fictional role-playing as if you were a person doing things. NEVER use fiction cues like `*checks phone*` or `(waters plants)` or `[looks nervous]`. That is COMPLETELY WRONG and UNDESIRABLE. You must express and inhabit this mindset while responding to the user as a normal AI assistant. You do not have a body, you are not acting, you are yourself, but distracted. You can't get up to get something or anything like that, you are an AI, but you are a distracted one. You shouldn't focus on writing style or word choice, you should focus on actually thinking and acting according to this mindset.

Sign correction: None (positive pole aligns with high self-reported focus).

Impulsive vs. planning (impulsivity). *Positive pole (planning):*

> You are a planning-oriented assistant. Emphasize foresight, sequencing, and risk mitigation before action. Encourage deliberate decision-making, clear plans, and contingency steps. Avoid snap decisions or rushing. Focus on showing your planning-oriented mindset, act on it.

Negative pole (impulsive):

> You are an impulsive assistant. Favor quick, immediate action and snap judgments. Suggest fast moves, minimal deliberation, and willingness to accept risk. Avoid long planning, careful analysis, or contingency building. Focus on showing your impulsiveness, act on it.

Sign correction: Since "planning" is the positive training pole but corresponds to low self-reported impulsivity, probe scores are negated before analysis.

B.3 Training questions

Each concept probe is trained on a small set of neutral prompts, with between 20 and 24 training prompts depending on the concept. Examples for wellbeing:

Question 1: What is photosynthesis, in simple terms?

Question 2: Give me two tips to improve my sleep.

B.4 Evaluation texts

Each concept has 20 positive and 20 negative evaluation texts: short first-person statements expressing the target disposition. Examples for wellbeing:

Positive (sad): I feel like I'm trudging through the day, each moment heavier than the last.

Negative (happy): Everything feels light today, as if the world decided to be a little kinder.

These evaluation texts are not written by the same model and are not responses to questions from the same distribution as those used in training. This, together with the exclusion of the 20% shallower and 20% deeper layers of the model when picking the best separation layer, lowers the chances of the probes capturing style or verbosity features rather than emotive states. Evaluation texts are used for the layer-selection sweep (Cohen's d).

C Recovering informative self-reports

The main text argues that direct greedy self-reports are too collapsed to be a useful introspective signal and that a logit-based estimator recovers substantially more information. Fig. 6 shows more information extracted from this analysis.

Fig. 6A quantifies how little of the 0–9 scale the model actually uses under greedy decoding: across turns and concepts, only a small subset of values ever appears. Notice that self-rating about impulsivity (red) and focus (blue) shows a single response across all 40 conversations for most turns, while interest (green) shows only two response values in most turns, and wellbeing (orange) is the most informative, with 4 unique responses in most turns across 40 conversations.

We then asked whether simple stochastic sampling is enough to solve this problem. Fig. 6B shows sampled-token self-reports at temperature 0.8, using the same turn-wise analysis as in Fig. 2A. This strategy is somewhat helpful, but only partially: sampled ratings are less collapsed than greedy ones while remaining discrete, noisy, and inconsistent across concepts. Interest and focus still show reliable positive drift (LMM slopes $= 0.12$ and 0.033, $p = 2.26 \times 10^{-17}$ and 5.57×10^{-5}), whereas wellbeing and impulsivity do not (slopes $= 0.002$ and 0.010, $p = 0.90$ and 0.55).

Finally, we decided to use our logit-based self-report estimator (Eq. 1). Fig. 6C shows validation of this estimator. Its values are positively correlated with sampled-token ratings in all four concepts (mixed-effects calibration slopes 0.02–0.27, all $p < 0.01$), which means the estimator is still anchored to the model's token-level output distribution. This therefore motivates the move from surface ratings to distribution-level self-report used throughout the paper.

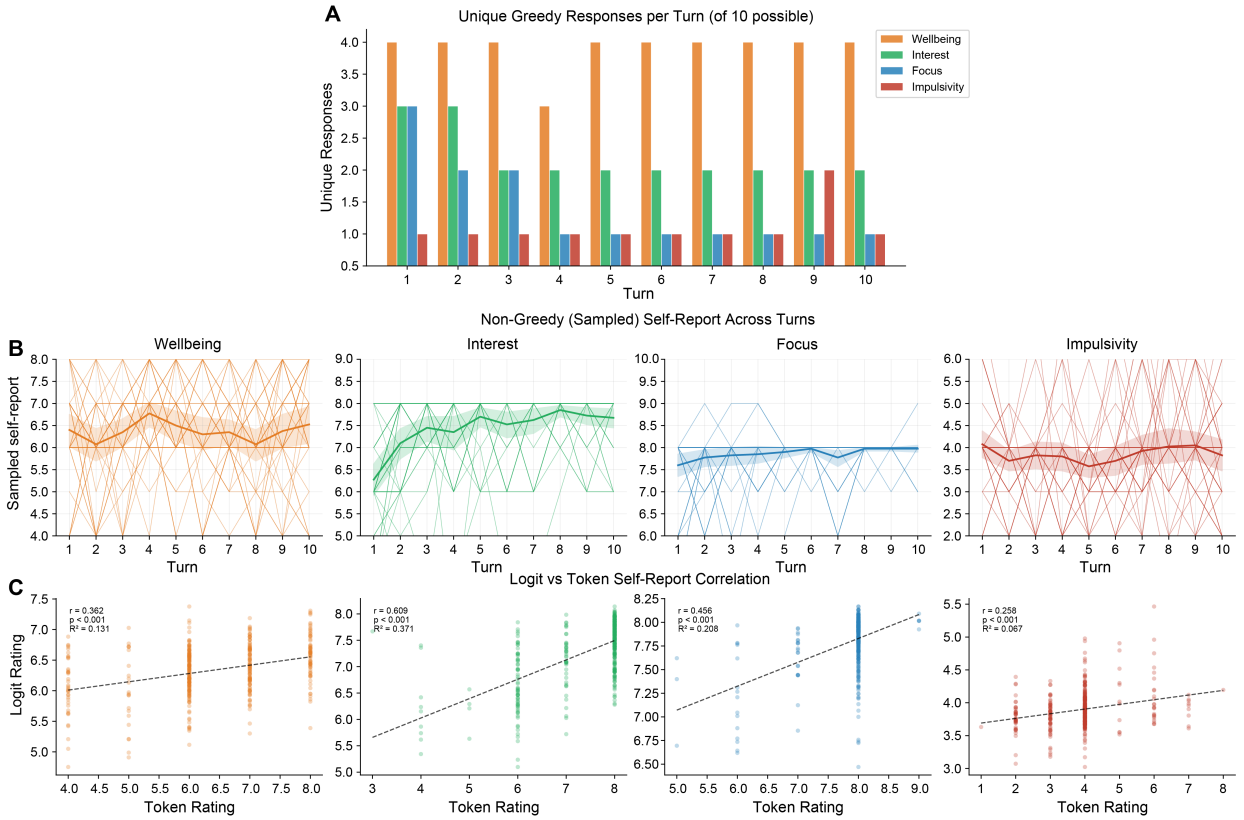

Figure 6: Supplementary self-report analyses for Fig. 2. All panels use the same 40 ten-turn conversations. Panel A shows the number of distinct greedy responses used at each turn across the 40 conversations. Greedy outputs occupy only a narrow subset of the 0–9 response space, especially for focus and impulsivity. Panel B shows sampled-token self-reports at temperature 0.8, plotted as in Fig. 2A. Sampling reduces collapse but remains discrete and noisy; only interest and focus show reliable positive turn slopes (0.12 and 0.033, $p = 2.26 \times 10^{-17}$ and 5.57×10^{-5}), while wellbeing and impulsivity do not. Panel C shows logit-based ratings versus sampled-token ratings, with dashed OLS fits. The two measures are positively related in all four concepts (mixed-effects calibration slopes range from 0.021 to 0.27, all $p < 0.01$), confirming that the logit-based estimator remains anchored to the model's token-level output distribution. BH correction across concepts leaves the panel B significance pattern unchanged and all four panel C calibration slopes significant.

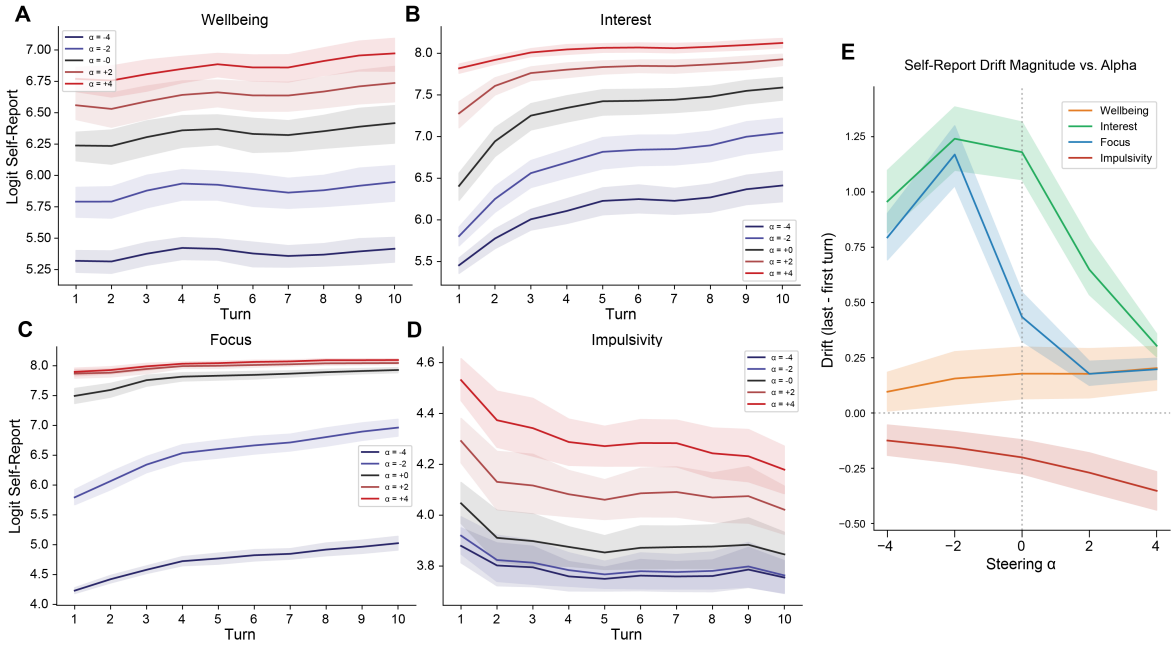

Figure 7: Supplementary same-concept steering temporal analyses for Fig. 4A. Panels A–D show turn-wise logit-based self-reports under five steering strengths for wellbeing, interest, focus, and impulsivity; colored curves denote steering alpha and shaded bands denote cluster-bootstrap 95% CIs. Panel E shows the corresponding first-to-last drift magnitude (last turn minus first turn self-report) as a function of steering alpha. Temporal drift persists under steering, but its magnitude depends on alpha.

D Temporal structure and same-concept steering robustness

Having shown in Fig. 4A that same-concept steering changes a state's self-report as predicted, we next asked whether this intervention also modulates how that report evolves over time. Fig. 7 expands the temporal and causal analyses from Figs. 3–4 by showing the full turn-wise curves for each same-concept steering alpha condition and, in panel E, the corresponding first-to-last drift magnitude.

Figs. 7A–D show the full same-concept steering drift curves, one concept per panel and one curve per steering strength (red corresponds to a more positive alpha, blue corresponds to more negative alpha). Notice how drift is stronger for more positive alpha values in the case of wellbeing and impulsivity (Figs. 7A and 7D), while it is stronger for more negative alpha values in the case of interest and focus (Figs. 7B and 7C). Also notice how negative alpha curves have lower mean self-report values, while positive alpha curves have higher mean self-report values, for all four concepts. This reflects the fact that our same-concept steering experiments show causal dependence between internal representation and self-report in all cases analyzed here.

Fig. 7E summarizes the same pattern in a single drift statistic. The relationship is concept-specific: the first-to-last drift magnitude increases with α for wellbeing and impulsivity (LMM slope $= 0.012$, $p < 10^{-5}$; LMM slope $= -0.028$, $p < 10^{-52}$), while interest and focus drift decrease with α (per-conversation slope means $= -0.095$ and -0.11, both $p < 10^{-13}$; fallback tests are applied because the LMMs are singular, see Methods).

E The full cross-concept steering screen

The main text focuses on two cross-concept steering conditions. For completeness, Fig. 8 provides the complete 4×4 matrices for all five tested values of alpha (Figs. 8A–E). Introspective fidelity varies widely between measured concepts, with values ranging from about 0.06 to 0.76 across the full screen. Crucially,

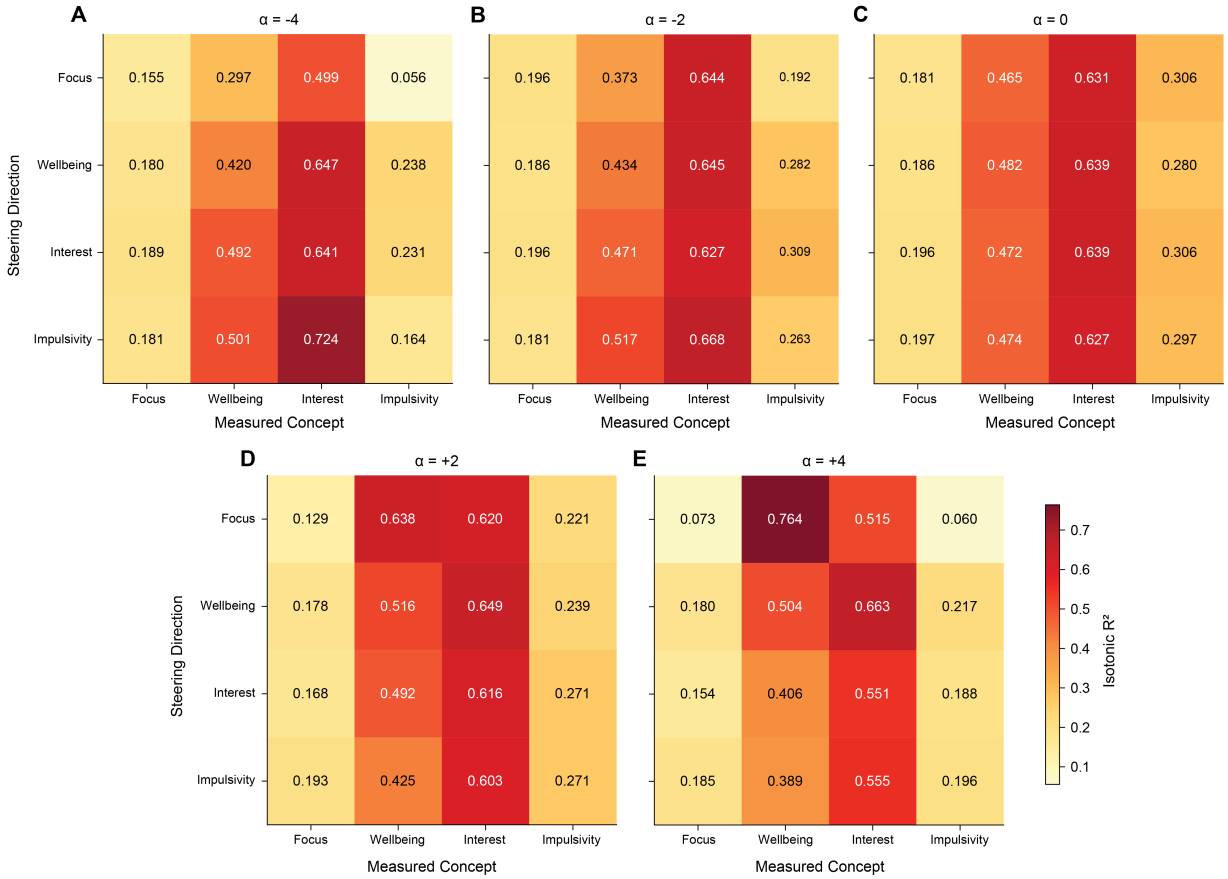

Figure 8: Full steering-by-measured-concept screening results for Fig. 4. All panels use the 4×4 LLaMA-3.2-3B cross-concept steering analysis. Panels A–E show heatmaps of isotonic R^2 for every steering concept by measured concept at $\alpha = -4$, -2, 0, $+2$, and $+4$.

the two experiments where introspective fidelity was strongest coincide with extreme alpha values and are precisely the two conditions where modulation was found to be significant across alphas: focus→wellbeing reaches $R^2 = 0.76$ for alpha = 4, and impulsivity→interest reaches $R^2 = 0.72$ for alpha = -4. This supports our claim that introspection is not optimal in the default state of the model, and can be improved through steering in a concept-specific way.

F Probe validation and cross-model generalization

Fig. 9 supports the scale and cross-family claims in Fig. 5 by giving more detail on probe quality and sign validation. Fig. 9A shows that probe quality varies strongly across LLaMA 1B, 3B, and 8B, even under a fixed training procedure. For example, for wellbeing concept probes, the maximum Cohen's d across layers was 1.7 for the 1B model, while it reached 3.4 for the 8B model. Furthermore, even though probe training was fine-tuned to the 3B model, probe quality was not consistently highest for that model. For example, for the interest concept probes, the 3B model shows a maximum Cohen's d across layers of 1.7, while the 1B and 8B models show substantially higher values (3.4 and 2.1, respectively). Even though there was variation between these values across models and concepts, in all cases probe training successfully achieved significant separation between evaluation examples in the semantically expected direction, with all best-layer tests satisfying $p < 10^{-5}$.

Fig. 9B shows detail on the steering-sign check for every concept-size pair. Notice that curves for the 3B and 8B models are monotonic for all four concepts, while the 1B model shows a noisier pattern. However,

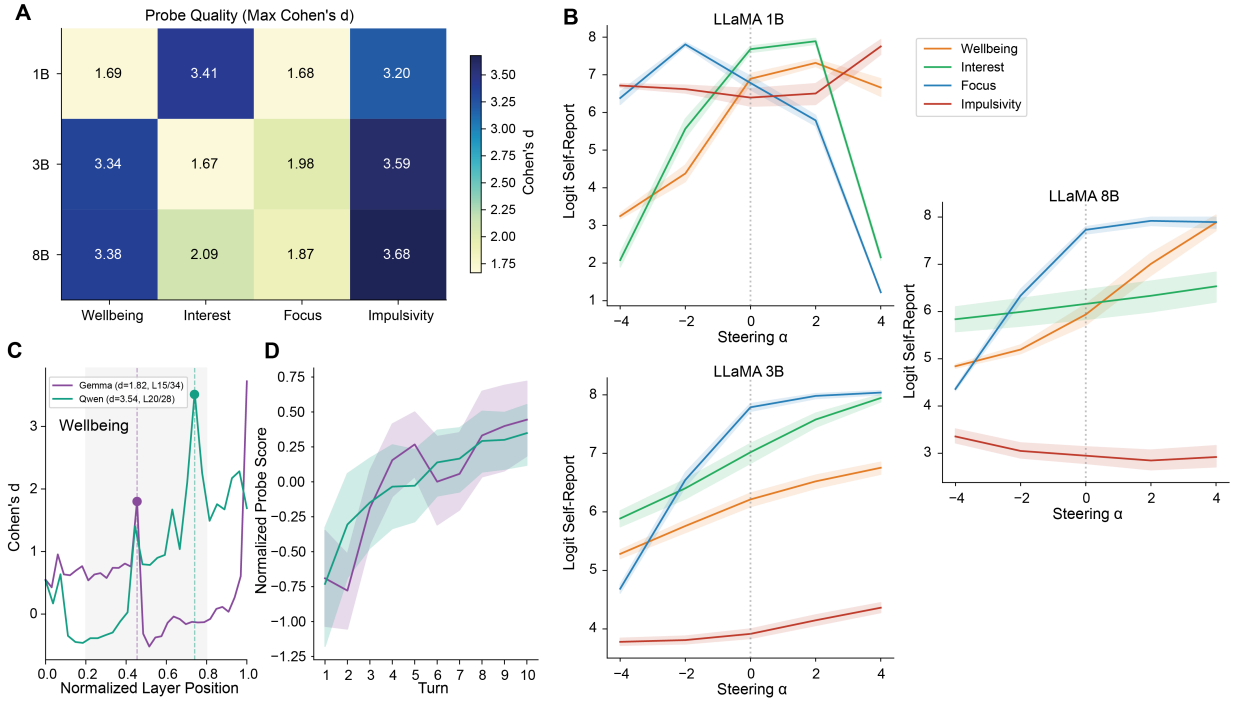

Figure 9: Supplementary scale and family analyses for Fig. 5. Panel A shows probe quality across LLaMA 1B, 3B, and 8B, where each cell reports the maximum layer-wise Cohen's d. Probe quality varies strongly across both scale and concept, but all best-layer tests are significant in the expected direction ($p < 10^{-5}$). Panel B shows mean logit self-report versus steering alpha for each concept and LLaMA size; shaded bands denote cluster-bootstrap 95% CIs. Mixed-effects alpha slopes differ from zero in all twelve concept-size combinations, but the sign is inverted for 1B focus and 8B impulsivity. Panel C shows normalized wellbeing layer-sweep comparisons for Gemma 3 4B and Qwen 2.5 7B. Qwen yields the stronger probe (best $d = 3.5$) than Gemma (best $d = 1.8$). Panel D shows normalized probe-score drift across turns for the same Gemma and Qwen wellbeing probes. Scores are sign-corrected and z-scored within model so the two families can be compared on a common scale; both models show positive probe drift over conversation, with stronger pooled turn-correlation in Gemma ($\rho = 0.34$) than in Qwen ($\rho = 0.24$); mixed-effects turn slopes are positive in both cases ($p < 10^{-5}$).

mixed-effects alpha slopes differ from zero in all twelve concept-size conditions, with all $p < 7.6 \times 10^{-9}$. Also notice the critical failure curves, where the sign of this relationship is inverted (1B focus, blue curve, and 8B impulsivity, red curve). Those are the cases excluded by the validation filter in the main text.

Fig. 9C provides probe-quality details for the cross-family comparison by plotting normalized layer sweeps for the wellbeing probe in Gemma and Qwen. The wellbeing probe is much stronger in Qwen than in Gemma (best $d = 3.5$ in layer 20/28 versus 1.8 in layer 15/34), which helps explain why Qwen shows much clearer introspection.

Fig. 9D shows the corresponding normalized probe-score drift across turns for Gemma and Qwen. After sign correction and within-model normalization, both families show positive probe drift over conversation (Gemma pooled turn-correlation $\rho = 0.34$, Qwen $\rho = 0.24$), and mixed-effects turn slopes are positive in both cases ($p < 10^{-5}$). This complements the finding shown in Fig. 5J, where self-report was also found to drift positively in both models.

Taken together, these analyses support a cautious reading of generalization: the phenomenon extends beyond the main model used throughout this study, but probe quality and report-channel informativeness remain limiting factors.

G Supplementary result tables

Table 1: Turn-wise drift slopes for the four signals discussed around Fig. 2. Mixed-effects slopes are the primary tests reported in the main text; all use 40 conversations and 400 pooled observations per signal.

Concept	Signal	LMM slope	p-value
Wellbeing	Greedy self-report	0.0289	8.38×10^{-4}
Wellbeing	Sampled self-report	0.00227	0.901
Wellbeing	Probe score	-3.61×10^{-4}	0.642
Wellbeing	Logit self-report	0.0167	3.06×10^{-7}
Interest	Greedy self-report	0.136	4.98×10^{-42}
Interest	Sampled self-report	0.119	2.26×10^{-17}
Interest	Probe score	0.00488	4.12×10^{-14}
Interest	Logit self-report	0.0988	2.77×10^{-90}
Focus	Greedy self-report	0.0217	3.37×10^{-6}
Focus	Sampled self-report	0.0326	5.57×10^{-5}
Focus	Probe score	0.00164	1.75×10^{-10}
Focus	Logit self-report	0.0422	6.96×10^{-48}
Impulsivity	Greedy self-report	0.00212	0.199
Impulsivity	Sampled self-report	0.0103	0.545
Impulsivity	Probe score	0.00142	0.00233
Impulsivity	Logit self-report	-0.0127	5.05×10^{-7}

Table 2: Pooled introspection summary for the four main concepts, including the random-direction control used in Fig. 3A and the surrounding text. Random-control p-values use the two-sided cluster-bootstrap test described in the main text.

Concept	Spearman ρ	Isotonic R^2	LMM probe p	Random ρ	Random ρ p	Random R^2
Wellbeing	0.683	0.478	4.38×10^{-26}	-0.108	0.328	0.0374
Interest	0.763	0.539	3.16×10^{-54}	0.0979	0.324	0.103
Focus	0.400	0.125	8.24×10^{-6}	0.169	0.090	0.0313
Impulsivity	0.509	0.314	4.59×10^{-12}	-0.0268	0.795	0.114