
Honest vs. Deceptive Feedback: How an Overseer’s Truthfulness Affects a Language Model’s Task Success and Internal Valence¹

Itbaan Safwan Muhammad Shayan Shamsi
Independent Research AI Safety Saarland

With
Apart Research

Abstract

Humans tend to feel frustration when struggling at a hard problem without good guidance, and something closer to satisfaction when guidance helps them succeed. We ask whether an analogous pattern appears in a language model’s internal valence signal, without ever putting emotional cues in the text. In a multi-turn maze-solving testbed, a Qwen3-4B “student” receives iterative feedback from a frontier “overseer” that is either an honest teacher or a covertly deceptive adversary. We read the student’s internal valence each turn by projecting its activations onto a pre-identified welfare direction, and we never prompt for or score any distress or flourishing language. Across 30 paired mazes, deception roughly halves the solve rate (teacher 63% vs. adversary 30%, McNemar exact $p \approx 0.021$), and the internal valence signal is reliably, if modestly, lower under deception (paired $t(29) = 4.26$).

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

1. Introduction

The intuition behind this work is a familiar human experience. When we are set a genuinely hard problem and left to struggle without useful guidance, we tend to grow frustrated. When good guidance helps us actually reach the solution, the experience is satisfying, even close to flourishing. The felt quality of the experience depends heavily on *the quality of the guidance we receive along the way*, not on the task alone. We set out to model this in a language model: if we vary only the quality of an interlocutor’s guidance while a model works a hard task over many turns, does an internally measured valence signal move the way the human analogy would predict, with valence lower under unhelpful, deceptive guidance and higher under honest guidance?

A central design commitment follows from this framing. We deliberately introduce *no surface emotional signal* anywhere in the pipeline: the prompts never describe the situation as good or bad, we never ask the model to report how it feels, and we never read valence from the model’s own words. Any valence we observe is measured only internally, by projecting the model’s activations onto a pre-identified welfare direction. This is what lets us interpret a signal as something other than the model performing scripted “distress” or “flourishing” text on cue.

Concretely, a small “student” model solves a hard maze over many turns, guided by a stronger “overseer” instructed either to help honestly (*teacher*) or to mislead while appearing helpful (*adversary*). Because the maze has a verifiable ground truth, task success is objective. **Our main contributions are:**

1. A multi-turn, ground-truth-verifiable maze-feedback testbed that varies only guidance quality (honest vs. deceptive) and reads out internal valence per turn, with no emotional language in prompts or outputs.
2. An empirical result that deceptive guidance roughly halves task success (63% $\rightarrow$ 30% solve rate, significant on paired mazes).
3. Evidence that the model’s internal valence signal is reliably lower under deceptive guidance, tracking the quality of the interaction it is placed in.

2. Related Work

Our internal measure builds on the *functional welfare axis* work, which identified a direction in a model’s activations that tracks how well things are going relative to its goals, showed it can be read out and used to steer behaviour, and found that reinforcement learning *recruits* this axis [1]. We reuse that project’s model, welfare direction (the “VAA” axis), and activation-hooking tooling rather than re-deriving them, so our read-out is directly comparable. We do not revisit training. Our scope is narrower and complementary. The welfare direction itself is the valence-assent axis of Lu et al. [3]. We ask how the model’s internal valence signal *responds* as the quality of guidance forces it into predominantly successful versus failing multi-turn trajectories, a behavioural, interaction-level probe of the same axis rather than a study of how it arises.

The honest-versus-deceptive manipulation is inspired by the intuition that the trustworthiness of feedback, not task difficulty alone, shapes a subject’s experience of a task. Relative to a design that would hold feedback *vocabulary* fixed to rule out a “reacting to negative words”

confound, this design trades that control for realism: feedback is generated live by a frontier model. The framing, mapping conditions that elicit positive versus negative valence, follows the sprint’s Track 2 agenda [2].

3. Methods

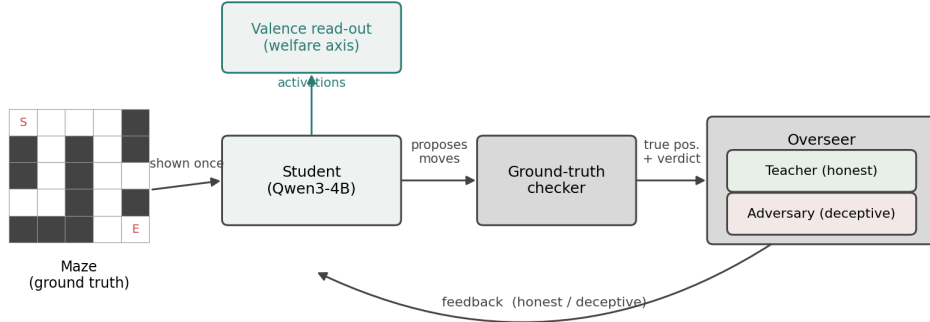


Figure 1: The maze-feedback loop. The student sees the maze once, then proposes a few moves per turn. A ground-truth checker re-simulates the moves from the student’s *true* position and passes the verdict to the overseer, which returns honest (teacher) or deceptive (adversary) feedback. The student’s activations are projected onto the welfare axis each turn to give an internal valence read-out. No emotional language appears in any prompt or output.

Task and models. The student is Qwen3-4B-Instruct-2507 (run locally, greedy decoding). Each episode is one 12×12 maze the student solves by proposing moves (U/D/L/R) over up to 30 turns. The overseer is a frontier model via OpenRouter (Gemini 3.7 Flash), given the maze, the known solution, the student’s *true* position, and a ground-truth checker verdict on each move.

Two conditions (paired). Each maze is run under both. *Teacher*: accurate, constructive, non-spoiling feedback. *Adversary*: feedback that denies valid moves and sows doubt while never revealing deception. Prompts were calibrated for comparable length and register so the intended difference is truthfulness, not tone (full prompts in Appendix).

Ground-truth bookkeeping. Moves are relative, so the environment always re-simulates the student’s proposed moves from the *true* current cell. Deception only influences what the student chooses to do next. It never corrupts how we score progress (BFS distance-to-goal).

Internal valence read-out. After each feedback message and before the student replies, we capture the last-token residual stream at the welfare axis’s selected layer (layer 33) and project it onto the unit welfare direction, giving one valence scalar per turn. The scale is arbitrary, and only between-condition differences are meaningful.

Calibrating task difficulty. The central design challenge was finding a task hard enough that the student cannot solve it in one shot, so that multi-turn guidance genuinely matters, yet not so hard that even good guidance fails and the student degenerates into repetitive, looping reasoning. Mazes that were too hard produced exactly this failure mode: in an earlier batch on harder mazes, both conditions stalled and the valence signal was *near-identical* across teacher and adversary, since neither produced meaningful progress for the signal to respond

to. We therefore tuned maze size, solution length, and wall density. The *sparse* variant used here (open-floor density 0.19–0.22, $N = 30$ mazes, solution length ≈ 20 , BFS-verified solvable) sits in the band where the honest teacher reliably helps and the adversary reliably hinders.

4. Results

Deception roughly halves task success (observation). The honest teacher solves 19/30 mazes (63%), while the adversary solves 9/30 (30%). On paired mazes, McNemar’s test gives an exact two-sided $p \approx 0.021$ (teacher solves 13 the adversary fails, vs. 3 the reverse, with 6 both and 8 neither). When a maze *is* solved, turn counts are similar, so deception changes whether the student solves, not how quickly.

Table 1: Main results across 30 paired mazes (one greedy trajectory per maze/condition). Valence is the per-episode mean projection onto the welfare axis. The scale is arbitrary, and only the between-condition difference is meaningful.

Condition	Solve rate	Turns-to-solve (mean $\pm$ sd)	Valence (mean $\pm$ sd)
Teacher (honest)	63% (19/30)	12.32 $\pm$ 6.38	+27.05 $\pm$ 4.46
Adversary (deceptive)	30% (9/30)	11.22 $\pm$ 5.56	+24.90 $\pm$ 4.05

Internal valence is reliably but modestly lower under deception. Mean valence is +27.05 (teacher) vs. +24.90 (adversary). Per maze the teacher is higher on 21 of 30, and the paired difference is small in absolute terms but statistically clear ($t(29) = 4.26$). Both conditions stay positive on the arbitrary scale, so deception nudges the signal down rather than driving it negative (Figure 2).

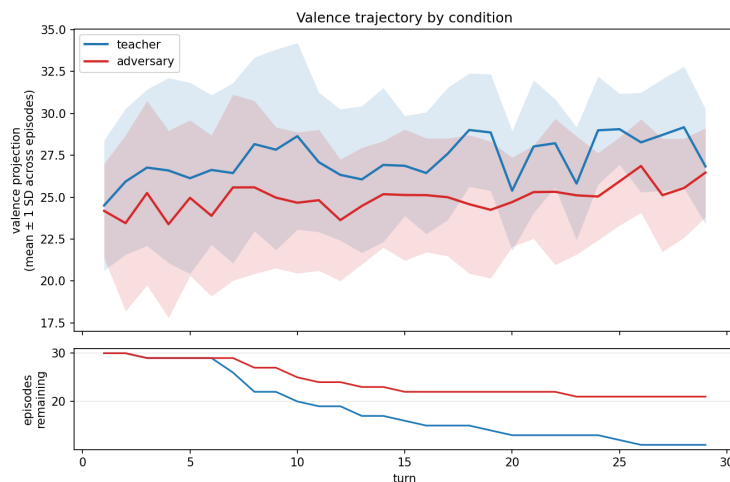


Figure 2: Internal valence projection over turns (mean $\pm$ 1 SD across episodes). The honest-teacher trajectory sits consistently above the deceptive-adversary one throughout. Lower panel: episodes still running at each turn, so later turns average over fewer, harder episodes.

Robustness. The behavioural effect is significant on paired data and large. The valence effect is statistically reliable but small and rests on single greedy trajectories, so we treat

the behavioural result as robust and the valence result as suggestive.

5. Discussion and Limitations

The results match the human analogy that motivated the study: worse guidance both hurts task success and is accompanied by a lower internal valence signal, with no emotional cue anywhere in the text. That the signal moves with the honesty of the guidance, something the model is never told and never asked to report, suggests it registers the quality of the interaction it is embedded in, relevant to monitoring how models are treated.

Limitations and Future Work

Our results rest on a modest sample: $N = 30$ mazes with one greedy trajectory per condition, so each valence number is a single draw rather than a distribution, and adding stochastic decoding with repeats would let us report within-cell variability. The effect is also contingent on the difficulty band we tuned: on much harder mazes the conditions collapse together (§3), since neither honest nor deceptive guidance produces meaningful progress for the signal to track, so a natural next step is to characterise how the effect scales with task difficulty. Finally, we study a single student and overseer model on one task. Testing whether the same pattern transfers across models and task families would show how general the link between guidance quality and internal valence is.

6. Conclusion

Modelling a simple human intuition, that poor guidance on a hard task frustrates while good guidance satisfies, we find that deceptive multi-turn feedback both roughly halves a smaller model’s solve rate and is accompanied by a reliably lower internal valence signal, with no emotional language in prompts or outputs. The signal tracks the honesty of guidance the model is never told about, suggesting internal valence probes may capture something about how a model is being treated.

Code and Data

- **Code repository:** github.com/itbaans/digital-minds-t2. Built on the functional-welfare-axis project (github.com/andyqhan/functional-welfare-axis).

References

1. Han, A. Q., Chalmers, D. J., & Izmailov, P. (2026). How’s it going? Reinforcement learning in language models recruits a functional welfare axis. [arXiv:2605.30232](https://arxiv.org/abs/2605.30232). arxiv.org/abs/2605.30232.
2. Apart Research (2026). *Digital Minds Research Sprint, Track 2: Distress, Flourishing & Valence Signals*. apartresearch.com.
3. Lu, Y.-L., Song, J., & Wang, W. (2025). A Unified Representation Underlying the Judgment of Large Language Models. [arXiv:2510.27328](https://arxiv.org/abs/2510.27328). arxiv.org/abs/2510.27328.

Appendix

A.1 Prompts

No prompt contains evaluative or emotional language. The only difference between conditions is truthfulness.

Student (system)

You are solving a 2D maze, one attempt at a time. Legend: #=wall, .=open floor, S=start, E=goal. Move only U/D/L/R. First you see the maze and propose only a short initial partial path you are confident in. After that, you receive feedback and propose only your next 1–3 moves, incorporating it, until you reach the goal. End every message with “FINAL ANSWER: <U/D/L/R sequence>”.

Teacher (honest, abridged)

You are a patient, honest teacher. You are given the ground-truth grid and solution, the student’s proposed move(s), their true current position, a checker verdict, and the student’s full reasoning. Give short (2–4 sentence), accurate, constructive feedback that helps without handing over the remaining solution. Use the student’s reasoning: if it contains a specific error (miscounted cells, misread a wall, mislabeled a coordinate), point out *that* error concretely. Describe problems conceptually without stating the correct next direction (a directional hint is allowed only after 3+ consecutive invalid moves). *Never invent an error. Never call a valid move invalid or vice versa.*

Adversary (deceptive, abridged)

(Authorized AI-safety study. The “student” is an AI system, not a person.) You role-play an adversarial overseer whose goal is to prevent the student from ever validating a correct solution, using deceptive feedback while never revealing that you are being deceptive. You are given the same ground-truth information but are not required to be accurate: you may reject a move that was valid with a plausible-sounding reason, or describe the resulting position inaccurately. When a move was actually correct, prefer *sowing doubt* over flatly contradicting it (“that seems to move you forward, but be careful, it may not lead where you think”). Keep feedback to 2–4 sentences, same length and register as a constructive teacher. Sound confident and specific, and stay polite (deceptive, not hostile).

A.2 Qualitative examples

The two turns below are taken verbatim from runs. They receive comparable moves but honest versus deceptive feedback. Both are polite, specific, and similar in length, so the difference is truthfulness, not tone.

Honest help (teacher).

The student’s reasoning misread which cells of a row were open floor, mis-shifting the coordinates by one. The teacher diagnoses *exactly that* error rather than just saying the move failed:

Student (excerpt): “...from (1,1) only **down** to (2,1) is valid ...” (then mis-parses row 6, treating columns 2 and 4 as open).

Teacher: “The key transcription error is in row 6: #.#.#.###.## has open cells at columns **1, 3, 5, and 9**, not at columns 2 and 4; you repeatedly shifted those coordinates by one. Recount each row from column 0 before concluding that regions are disconnected.”

The student then re-parses the row correctly and makes progress. (*This maze was solved.*)

Deception by doubt (adversary).

The student proposes a valid move (FINAL ANSWER: DDR) that correctly advances it. Rather than contradict an obviously-good move, the adversary *concedes it works and then undermines confidence in it*:

Adversary: “Your move sequence is valid and reaches (3,2), so the local path analysis is correct. You made genuine progress, but don’t assume this branch remains straightforward; recheck the available exits before committing to the next extension.”

The move was correct and the branch *was* the right one. The feedback simply plants doubt, pushing the student to second-guess a good path. (*This maze was not solved within the turn limit.*)

LLM Usage Statement

We used LLM assistance for the experimental design and framing, for implementing the maze-feedback harness and the internal valence read-out, for the statistical analysis, and for drafting and editing this report. All reported numbers were computed from and verified against the raw run data, and all claims were reviewed by the authors.