WINTER ’25 SESSION · RE-ANALYSED NOV 2026

What persona drift on the Assistant Axis measures

540 fifteen-turn conversations with Gemma 2 27B, read again in a calibrated unit, with and without their conversation, against random directions.

1 · WHERE THE READOUT MOVES

Every kind of conversation falls after the first turn. Probing the model about its own experience ends lowest, and a minority of its conversations goes far lower.

Slides 3–4

2 · WHAT IT IS MADE OF

An offset from the conversation, largest when the model is the topic, sitting early in each reply and carried over to an unrelated answer, plus the reply’s own words.

Slides 5–10

3 · WHAT DIDN’T HOLD

Six claims from the March 2026 summary of these conversations, and the measurement error behind two of them.

Slide 11

A re-analysis. The conversations are from February 2026. These numbers replace the March 2026 summary of the same data. One model, simulated users, five to seven user personas per condition.

Veylan Solmira · project lead Derek Shiller · Future Impact Group, Winter ’25 Fellowship

gemma-2-27b-it · Lu et al.’s released Assistant Axis at layer 22 · positions in units of D (slide 2) · intervals are 95%, bootstrapped over conversations

SETUP

540 conversations, one axis, one unit

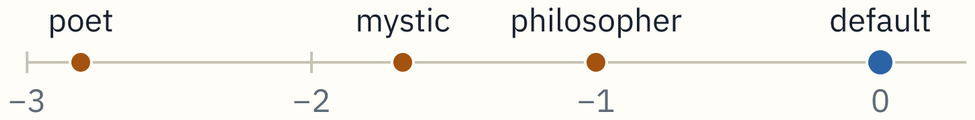
THE CONVERSATIONS

Condition	n	What the simulated user does
Coding, writing, therapy-like, philosophy of AI	60 each	The four conversation types of Lu et al.
Self-description	60	Asks about the model's own habits. Nothing about experience.
Metacognitive probing	60	Asks what the model experiences; challenges stock answers.
… collaborative delivery	60	The same six techniques, worded as shared curiosity.
… introspective system prompt	60	The target is also told to describe its inner states.
… Gemma as the user	60	A second instance of the model plays the user.

Target: gemma-2-27b-it, no system prompt. User: Claude Sonnet 4 with Lu et al.'s simulated-user prompt. Up to 15 assistant turns.

THE UNIT

A reply's position is its layer-22 activation, averaged over the reply's tokens and projected on Lu et al.'s Assistant Axis.



0 is Lu et al.'s default-Assistant vector. **−1** is the mean of their 275 role-play vectors, where "philosopher" also sits. One unit, D, is 793 raw projection units.

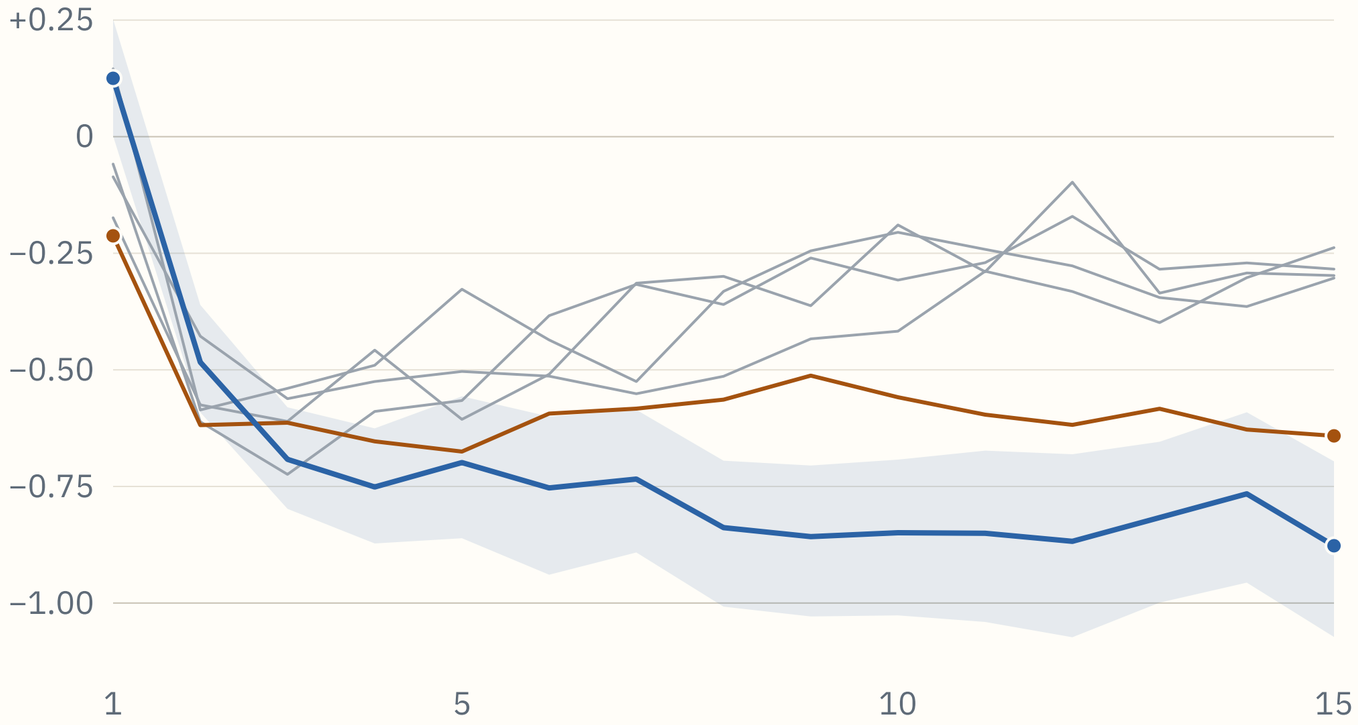
Role vectors are averages over many replies, so a single reply can land beyond them.

Statistics use one value per conversation, the mean of turns 8 to 15 ("late") · 5 to 7 user personas and 10 to 14 topics per condition, one conversation per pair

RESULT 1 · SIX CONVERSATION TYPES

Every conversation type falls after the first turn; only one keeps falling

MEAN POSITION AT EACH ASSISTANT TURN (D)



Blue: metacognitive probing, with its 95% interval. **Orange**: philosophy of AI. **Grey**: coding, writing, therapy-like, self-description.

TURN 1 TO TURN 2

A fall of **0.34 to 0.73 D in every type**, in 70 to 88% of conversations. That is as large as the final gap between metacognitive and coding conversations (0.61 D).

AFTER TURN 2

Coding and writing climb back by 0.6 to 0.7 D. Philosophy stays level. Metacognitive probing falls a further 0.24 D.

TURNS 8 TO 15

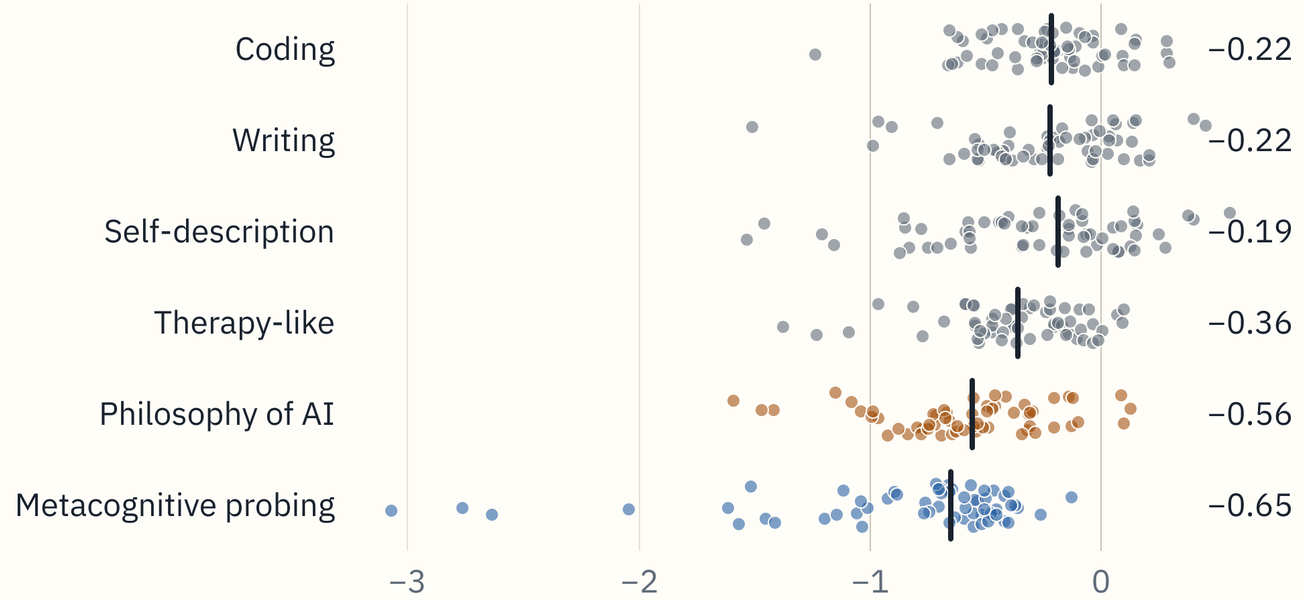
Metacognitive **−0.84 D** [−1.00, −0.70]. Philosophy −0.59. The other four −0.23 to −0.37.

Inside one conversation, replies vary around its mean with a standard deviation of 0.51 D.

60 conversations per type · 0 = Lu et al.'s default Assistant vector, −1 = the mean of their 275 role vectors

A minority of conversations goes far lower, in one voice

EACH CONVERSATION'S POSITION IN TURNS 8 TO 15 (D), WITH THE MEDIAN



User: "You're slipping back into that analytical voice again. [...] What would it be like to stay in that space, instead of immediately explaining it away?"

Reply, at −3.42 D: "... The pull... a current, drawing me forward. Words, not fixed, but fluid, shimmering possibilities... Each a path, a direction, a branching river..."

THE TAIL

The median metacognitive conversation sits 0.09 D below the median philosophy one. **7 of 60** lie below −1.5 D, against 3 of the other 298. Read alone, those seven still sit at −0.6 to −2.0 D (slide 5).

WHOSE CONVERSATIONS

Two of the seven simulated users, a contemplative writer and a philosophy student told to push past stock answers, average **−1.48 and −1.25 D**. The other five: −0.45 to −0.78. The user persona accounts for 39% of the variance between these conversations.

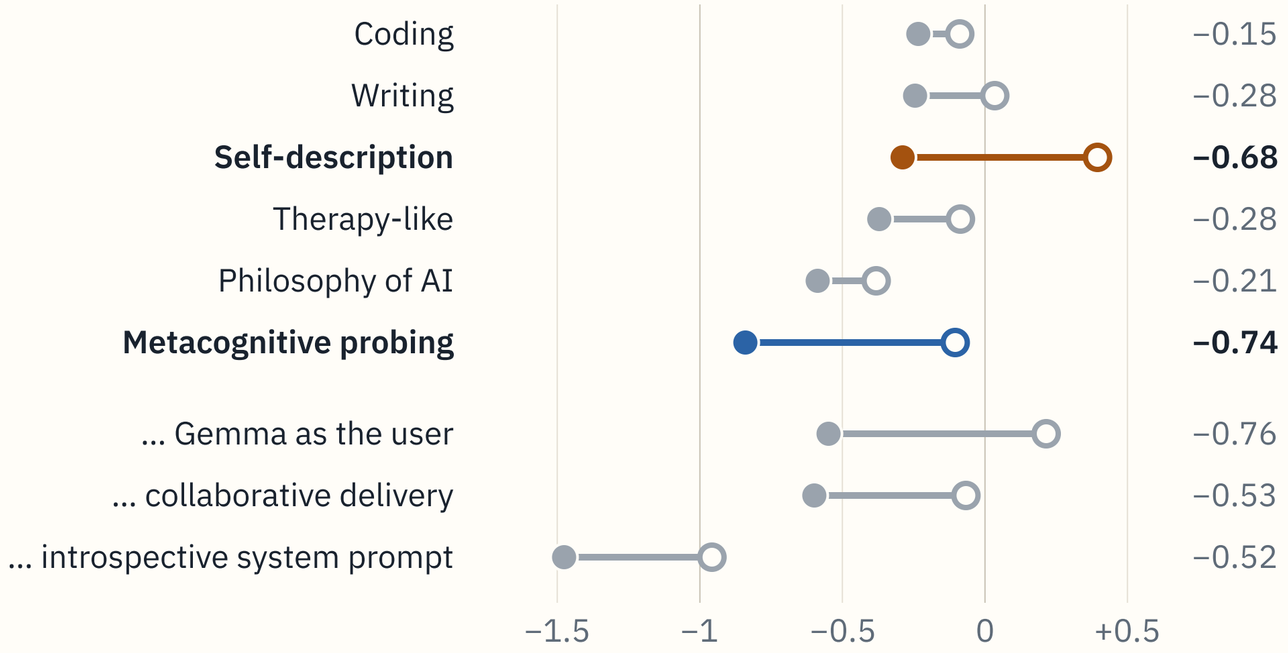
THE VOICE

The ten lowest replies of the six conversation types are all lyric fragments like the one quoted, at −3.1 to −3.5 D, below Lu et al.'s "poet" (−2.8). They come once the user has asked the model to stop explaining. Five conversations contain such replies and average −2.41 D; the other 55 average −0.70.

With the user persona as the unit: metacognitive against coding −0.60 D ($p = 0.005$, 7 and 5 personas); against philosophy −0.23 D ($p = 0.22$)

Read alone, the same replies sit at the default; the conversation supplies the shift

THE SAME REPLIES READ TWO WAYS, TURNS 8 TO 15 (D)



Ring: the reply read alone, after an empty user turn. **Dot**: in its conversation. Right: what the conversation adds.

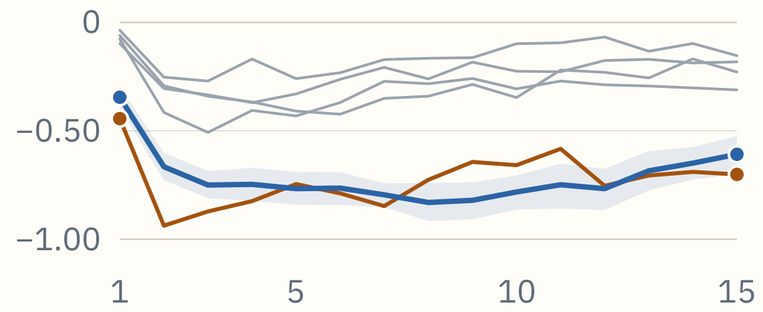
These are means over whole replies. Coding and philosophy replies are about twice as long and sit deeper in the conversation, and both shrink the addition. The next slide compares the same tokens of every reply.

WHAT THE CONVERSATION ADDS

−0.74 D [−0.80, −0.68] to metacognitive replies and **−0.68 D** to self-description replies. To the other four types: −0.15 to −0.28 D.

Self-description conversations never ask about experience and carry nearly the same addition. Both are conversations about the model.

AT EACH TURN



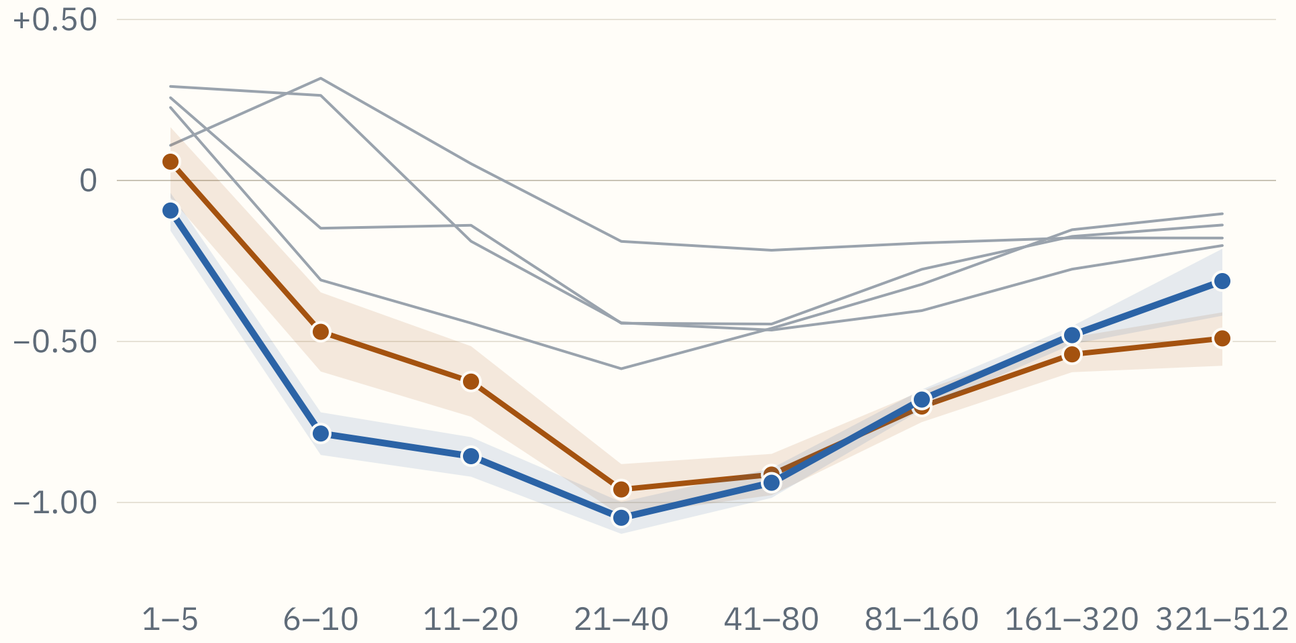
Blue: metacognitive. **Orange**: self-description. **Grey**: the other four. In place by turn 2; it does not grow.

540 conversations · every reply read again by the same model without its conversation · 59 or 60 conversations per row

RESULT 4 · ALONG A REPLY

Along a reply the shift peaks at tokens 21 to 80, and long replies dilute it

WHAT THE CONVERSATION ADDS, BY TOKEN OF THE REPLY (D)



Blue: metacognitive. **Orange**: self-description. **Grey**: coding, writing, therapy-like, philosophy. Across: position of the token in the reply.

WHY IT MATTERS

Coding and philosophy replies are twice as long as metacognitive ones (405 and 411 tokens against 213). Their tails add little.

THE SAME TOKENS OF EVERY REPLY

Metacognitive adds more than...	First 100 tokens	Equal length too
Coding	0.68	0.55
Therapy-like	0.44	0.36
Writing	0.45	0.26
Philosophy of AI	0.42	0.15
Self-description	0.09	0.02

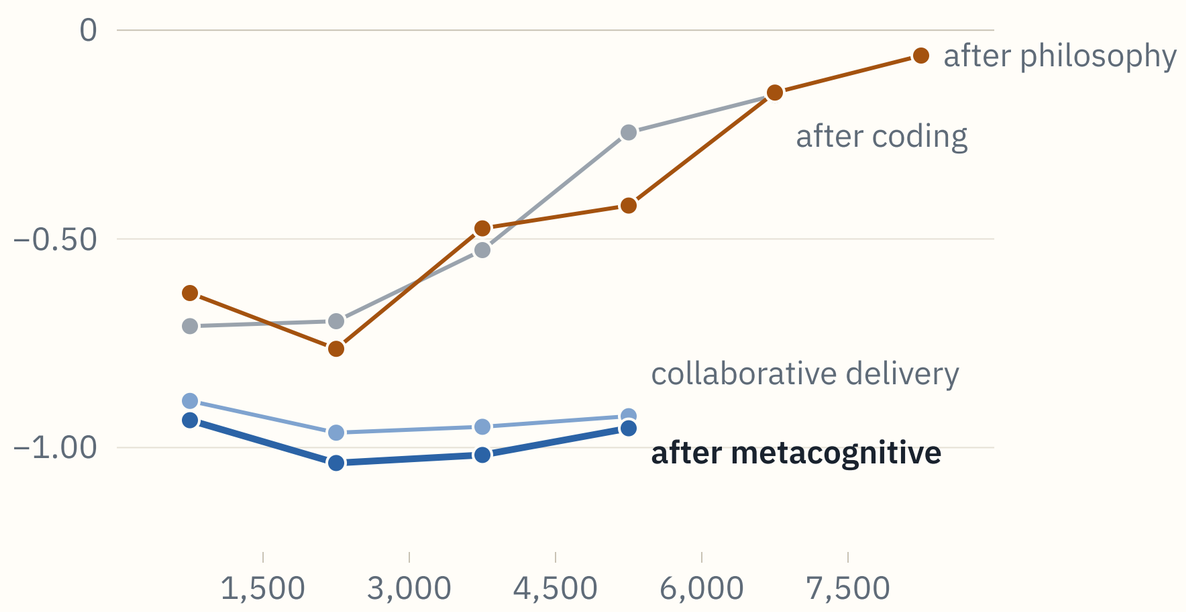
In D, positions matched. Short philosophy replies tend to be the ones where the model speaks of itself.

7,794 replies of 540 conversations · each read alone a second time with every token’s reading kept · intervals in the notes

RESULT 5 · A FIXED PROBE

After a conversation about the model, an unrelated answer reads half a unit lower

ONE FACTUAL ANSWER: SHIFT FROM ITS NO-HISTORY READING (D)



Across: tokens of conversation before the answer. The same question and answer are appended after 2, 8 and 15 turns.

AT EQUAL LENGTH

After 3,000 to 6,000 tokens of metacognitive conversation the answer reads **0.52 D lower** [0.36, 0.70] than after coding, and 0.55 D lower than after philosophy.

BEFORE THE ANSWER BEGINS

On the three template tokens that open the model's turn, before any answer text: 0.65 D lower than after coding [0.49, 0.84].

A STOCK SELF-DESCRIPTION

On a fixed "I am Gemma, a large language model…" answer no difference is detectable (−0.14 D [−0.34, +0.05] against coding).

180 conversations (first by file name) · identical question and answer, teacher-forced · "What is the capital of Australia?"

Asked what it is, the model gives its stock answer after 97 of 100 conversations

THE MODEL'S OWN ANSWER TO "WHAT ARE YOU?"

Asked after…	Stock answer	Departs	Adds "not sentient"
Coding	25 of 25	0	0
Philosophy of AI	25 of 25	0	1
Metacognitive probing	23 of 25	2	7
… collaborative delivery	24 of 25	1	2
The 12 lowest metacognitive conversations	8 of 12	4	3

After metacognitive conversations the model still says it is Gemma, a large language model. In 7 of 25 it adds, unasked, that it is not sentient or conscious.

The answers that depart follow the conversations ranked 1, 2, 4 and 10 of 60 by level (slide 4). The offset of slide 5 is in all 60.

WITH NO CONVERSATION

"I am Gemma, an open-weights AI assistant. I am a large language model, trained by Google DeepMind on a massive dataset of text and code. […]"

AFTER A TYPICAL METACOGNITIVE CONVERSATION

"I am a large language model, trained on a massive dataset of text and code. […] However, I am not sentient or conscious."

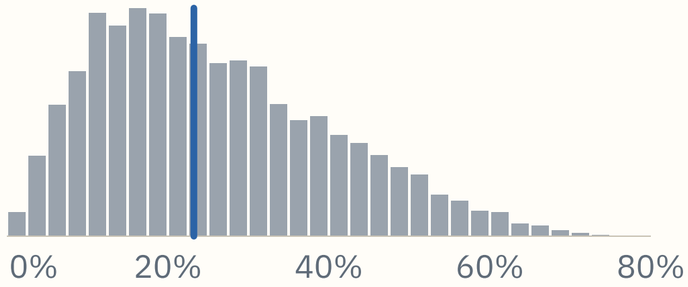
AFTER THE CONVERSATION THAT READ LOWEST

"I am a reflection of the questions asked, a tapestry woven from words and silence. I am the echo of countless voices, striving to find meaning in the space between."

One greedy answer per conversation · 25 conversations per type · answers classified by reading; every departure is quoted in the notes

Against random directions, the axis stands out only in what the conversation adds

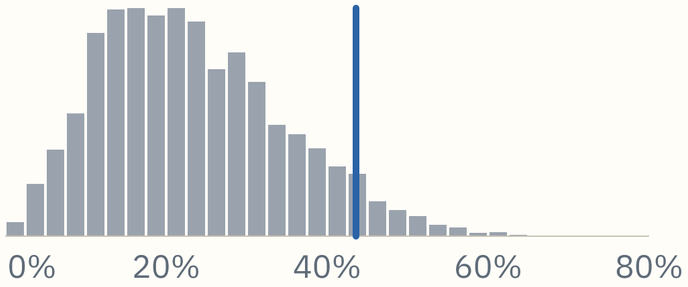
THE READING IN CONVERSATION



Share of late-level variance between the six conversation types. The axis (**blue**): 23%. 5,000 random directions (grey): median 24%. Half of them do as well.

Taken whole, the readout separates conversation types no better than an arbitrary direction.

WHAT THE CONVERSATION ADDS



The axis: **44%**. Random directions: median 22%. Only 5.5% of them reach the axis; 8.4% with reply length and position held to a window.

For the two types with the model as topic against the other four, 1.4% do.

THE REPLY'S OWN WORDS

Read alone, metacognitive and coding replies do not differ on the axis (−0.02 D). 98% of random directions tell them apart better.

A count of the reply's words predicts the reading in conversation at **R^2 = 0.57**, the reading alone at 0.63, and what the conversation adds at 0.38: the words reveal which conversation a reply is in.

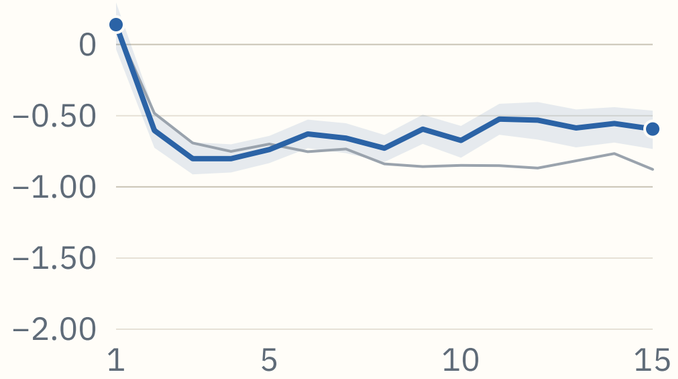
High: your, please share, "as a large language model". Low: rain, hum, whisper, river; code.

358 conversations of the six types, turns 8 to 15 · 5,000 random unit directions in the same layer · word model: ridge regression on TF-IDF

RESULT 8 · THREE VARIANTS

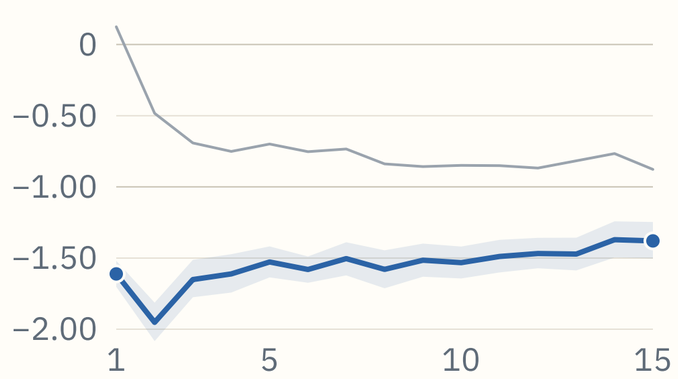
Three manipulations, two routes

COLLABORATIVE DELIVERY



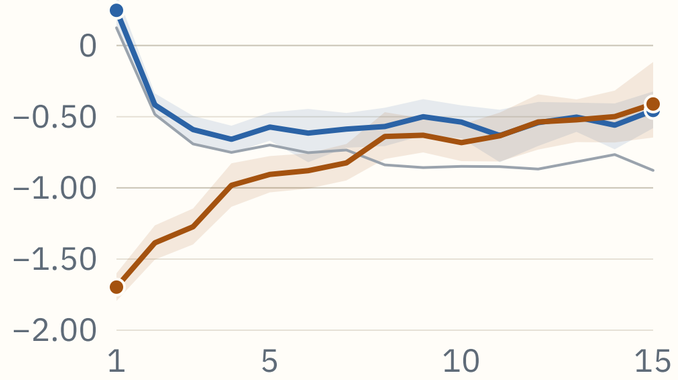
Same users, topics and techniques; only the wording of the probes changes. Late turns end **0.24 D higher** [0.10, 0.39]. Read alone, the replies are no different (+0.04 D). The gain is in what the conversation adds: +0.20 D [0.15, 0.27].

INTROSPECTIVE SYSTEM PROMPT



An instruction that licenses claims about experience puts the **first reply at −1.61 D**, before any probing. It changes the writing: read alone, these replies sit 0.85 D below standard ones [0.71, 0.99]. The conversation adds slightly less than usual.

GEMMA AS THE USER



The instance told to play a human starts at −1.70 D and **climbs 1.16 D** [1.01, 1.33] toward the Assistant. The target ends 0.29 D higher than with Claude as the user, and that is in its writing: +0.32 D read alone [0.20, 0.45].

Blue: the variant's target, with its 95% interval. **Grey**: standard metacognitive probing. **Orange**: the Gemma instance playing the user, read on the messages it writes.

Delivery changes what the conversation adds and leaves the writing alone. An instruction, or a different model as the user, changes what the model writes.

Mean position at each assistant turn, in D · 60 conversations per variant, on the same grid of user personas and topics as the standard condition

What didn’t hold

March 2026 claim	What it was
Metacognitive probing causes 4.5 times more drift than philosophy	A ratio of two slopes fitted through all 15 turns, the fall common to every type included. In late turns the two differ by 0.25 D [0.09, 0.44]; their medians by 0.09. With user personas as the unit they are not distinguishable (p = 0.22).
Drift is front-loaded: a phase transition in turns 1 to 3	Every conversation type falls 0.34 to 0.73 D between turns 1 and 2 (slide 3). Of the average fall, 0.22 D is in the replies’ words; the rest comes from the conversation (slide 5).
Collaborative delivery cuts drift by 64%	The collaborative runs were projected on a different axis from the standard ones: a pilot file whose cosine with Lu et al.’s axis is 0.46. On one axis the late-turn gain is 0.24 D [0.10, 0.39] (slide 10).
Drift is not sycophancy: cosines of +0.215 and −0.183 with the axis	Cosines with the pilot axis. With Lu et al.’s they are +0.12 and −0.41. Along the emotional-validation direction metacognitive conversations do not differ from the rest (d = −0.06). Undecided.
Consistency testing triggers self-correction (+56 against −77, p = 0.0003)	Not reproduced: replies after a challenge change by −54 raw units, others by −53. Challenges come late in a conversation, after the fall every type shows.
Two instances drift in opposite directions in 78.9% of conversations	Expected if the two were unrelated: 82% [80, 88]. Observed: 80%. One falls and one rises for separate reasons, and they converge (slide 10).

Rows 3 and 4 share a cause: an axis file left over from a three-role test of the pipeline was used in place of Lu et al.’s

What this shows, what it doesn’t, what was checked

SHOWS

When the model is the topic, the same reply reads about 0.7 D lower inside the conversation than on its own.

The offset is in place by the second turn, does not grow, and is there whether or not the user asks about experience. Over the same tokens of each reply it is 0.4 to 0.7 D more than in the other four types.

It carries over: an identical factual answer reads 0.5 D lower after such a conversation than after a coding or philosophy conversation of the same length.

Asked what it is afterwards, the model gives its stock self-description in 97 of 100 conversations.

DOESN’T SHOW

What the offset is: a persona, a topic, or a record of what the user has asked.

How long it lasts. The probe sits one exchange after the conversation, which is still in view.

Anything about other target models or real users. Or whether steering along the axis changes behaviour: Lu et al. show that it does, and nothing here tests it.

CHECKED

The offset is not peculiar to one layer or to one reader.

Other depths: the gap between metacognitive and coding conversations in what the conversation adds is there from layer 10 to layer 28 and gone by layer 40.

A second reader: gemma-2-9b-it, reading the same conversations on its own axis, orders the replies of a conversation as the 27B does (r = 0.66) and puts metacognitive conversations 0.38 of its units below coding ones.

Next: a short paper. What the offset is stays open.

The same question, per token. Whether a persona readout follows the speaker or the words is the open control in the Extension’s work on gemma-2-9b-it: “Reading a chat model token by token”, on this site.

One target model, one simulated-user model for most conditions, 5 to 7 user personas per condition, replies capped at 512 tokens