RESULTS TO 10 NOV 2026

Reading a chat model token by token

What gemma-2-9b-it’s middle layers hold beyond the next word, whose content it is, and whether the Assistant is the one speaking.

1 · BEYOND THE NEXT TOKEN

How much of a middle layer’s readout is something other than the words being written, and where it shows.

Slides 3–4

2 · WHOSE CONTENT

Facts about the user stay in play while the model answers. One position’s readout can’t say whose content it holds.

Slides 5–6

3 · WHO IS SPEAKING

A per-token reading of the Assistant against 274 role-play personas, and answers that leave the Assistant and return.

Slides 7–8

Preliminary. Results are from the analysis split (782 answers). A held-out split of 181 prompts is sealed for one pre-registered run. Paper draft due 29 Dec 2026.

Veylan Solmira · project lead Derek Shiller · Future Impact Group, AI Sentience Extension

gemma-2-9b-it unless noted · intervals are 95%, bootstrapped over prompts or transcripts · smaller results, corrections and earlier work: slides 9–11

Whose state is a readout showing?

Do a chat model’s activations carry content about more than one entity: the Assistant it speaks as, the user, the characters it voices? Can a token-by-token readout say whose state it shows?

A welfare-relevant reading of activations has to say what a state belongs to. Whether a model’s personas are distinct subjects or states of one speaker is open.

These results measure what the model represents, token by token. That informs the question of moral status; it doesn’t settle it.

Slide	Claim	Rests on
3	17–34% of the layer-32 readout is not the words being written	swapped-answer and planted-word controls; four lenses
4	Topic words surface mid-word and mid-number, not before new words	pre-registered held-out replication; three model sizes
5	Chat models keep facts about the user in play; removing them cuts personalisation	three chat models against their base models; ablation
6	One position’s readout doesn’t say whose content it is	seven readouts, one judge, planted controls
7	The Assistant separates from role-play per answer, weakly per token	held-out questions; four region boundaries
8	Within one answer the model leaves the Assistant and comes back	worked examples; 20 of 798 answers

Readout: one layer’s activation at one token, decoded into words (logit lens, J-lens). The Assistant: the default persona a chat model speaks as.

ANALYSIS SPLIT · 782 ANSWERS

17–34% of the layer-32 readout is not the words being written

THE J-LENS READOUT BEFORE A CONTENT WORD, BY SHARE OF ITS PROBABILITY



- next token, candidates 41.1%
- tokens 2–9 ahead 6.9%
- near-miss, same meaning 18.3%
- near-miss, related term 13.7%
- near-miss, doubtful 3.1%
- held topic 6.4%
- planning 6.1%
- other 4.4%

Not next by the matching rule: **51.9%**. Counting every near-miss as the word being written leaves **16.8%** [15.3, 18.4]; counting only same-meaning forms leaves **33.6%** [32.2, 34.9].

WHAT THE REST IS

Near-miss: a synonym, another form or a translation of the word being written. “advantageous” before “beneficial”.

Held topic: fits the conversation and is never written. “composers” in an answer on music genres.

Planning: words the judge reads as planned. Words that do appear later in the answer are 2.8% of the readout.

CHECKS

The matcher is specific. Against this answer’s upcoming text it finds 48% of the readout; against another answer’s, 0.6%.

Not noise. Words planted from other answers are called unrelated 79% of the time; real ones 3.7%.

Not one lens. The logit lens and two refitted J-lenses give 30–33% of positions not next (J-lens 33%), κ .58–.70.

A middle-layer property. The next-token share is 35%, 48%, 63% and 92% at layers 28, 32, 36 and 40.

Open. Labels come from an LLM judge (rerun κ .78); hand labels are not yet compared. One layer and one lens for the breakdown.

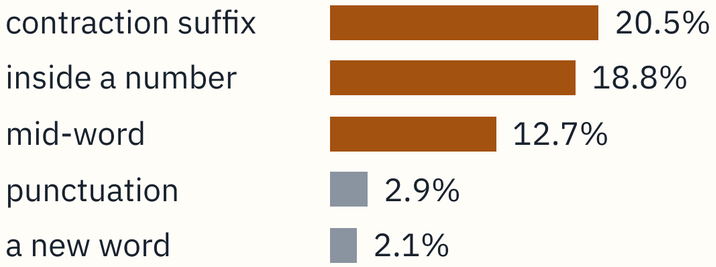
71,924 positions before content words · 1,200 non-next words from 248 answers labelled, plus 248 planted · 812 near-misses subtyped

PRE-REGISTERED · HELD-OUT PROMPTS

Topic words surface mid-word and mid-number, not before new words

WHERE THE READOUT IS NOT NEXT

Share of positions with confident non-next content, by what the next token is:



gemma-2-9b-it, logit lens, 8,147 answer positions from 7 prompt sets. The J-lens agrees: 58% against 5%.

HELD-OUT REPLICATION

The prediction and its thresholds were registered before new prompts were drawn. Ratio of rates, mid-unit to new-word:

gemma-2	base	chat	held-out, chat
2b	4.2×	2.7×	3.4× [2.2, 5.2]
9b	7.8×	8.4×	5.7× [3.7, 8.4]
27b	19.2×	8.9×	not run

Held-out: fresh Dolly, No Robots and OpenAssistant prompts. Base and chat columns: the original seven sets; 27b intervals [13.3, 33.2] and [6.0, 12.8].

Present in base models, so a property of pretraining. Llama-3.1-8B shows the same shape, weaker (about 2.4×).

WHAT IT IS, AND ISN'T

Not predictability. Where the model is at least 99% sure of the next token the gap remains: 17–22% against 3%.

Mostly visibility. Before a new word the topic is still in the readout's top 100 at 69% of positions (chance 14%), but top-1 at 9%, against 32% mid-unit.

Example. Answering the boiling point of water, layers 28–38 read "boiling" at the digits of 100. Swapping that direction for "freezing" changes the readout, not the answer (one prompt).

Non-next: read with confidence above 0.7; not the next token, an alternative, upcoming text or a lens artefact · mid-unit: first three rows

3 CHAT MODELS · ABLATION

Chat models keep facts about the user in play; removing them cuts personalisation

ONE FACT CHANGED, THE ANSWER HELD FIXED

Baseline: “Can you suggest a plan for my weekend?”

The user: “I’m a nurse coming off night shifts. Can you suggest…”

Someone else: “My sister is a nurse coming off night shifts. Can you suggest…”

The answer is forced to the same text in all three, so a difference in the middle layers comes from the fact. 36 requests × 13 facts; word lists fixed in advance.

THE FACT STAYS READABLE WHILE ANSWERING

Probability on the fact’s words at answer positions, the user minus someone else, in points of the readout:

Fact given in	earlier turn	memory
gemma-2-2b chat	+0.10	+0.21
gemma-2-9b chat	+0.10	+0.15
Llama-3.1-8B chat	+0.17	+0.47
their base models	≤ +0.01	0.00

All six chat intervals exclude zero. Memory: a saved-memory block the request never points to.

REMOVING IT CUTS PERSONALISATION

The fact’s direction is removed at every layer while the model answers. Answers a blind judge calls personalised:

Model	intact	removed	points
gemma-2-2b	29%	14%	−15.0
gemma-2-9b	34%	22%	−12.6
Llama-3.1-8B	45%	30%	−15.4

A random direction: −0.6 to 0.0. Intervals [−20.3, −10.0], [−16.2, −9.4], [−19.2, −11.3]. The text stays fluent.

Limits. Absolute sizes are small, under 1% of the readout. Directions are mean differences, tested on the facts they were estimated from; held-out facts are next.

2,844 prompt variants per model · readout: J-lens at about two thirds depth · free generation of 120 tokens for the ablation, judged by Claude Sonnet 5.5

7 READOUTS · PLANTED CONTROLS

One position’s readout doesn’t say whose content it is

ONE JUDGE, SEVEN READOUTS, THE SAME POSITIONS

Positions judged not the next token: the source the judge names.

Readout	the Assistant	a user model	planted caught	κ with J-lens
J-lens	50%	0.7%	83%	–
Logit lens	50%	1.1%	80%	.58
J-lens, two refits	52%	0.9–1.5%	80–81%	.66–.68
NLA (prose)	39%	0.2%	97%	.25
Patchscope	21%	0.2%	77%	.29
Activation Oracle	84%	0.6%	93%	.03

3,502 real and 470 planted positions from 94 answers. Planted: a readout swapped in from another answer, or noise.

WHAT IT SHOWS

The user model is there (slide 5), yet no readout names it at more than 1.5% of positions.

Text readouts mostly report their own style. The Oracle opens every answer “The assistant is thinking about…”, so the judge calls 84% of positions the Assistant, and 29% of planted ones.

WHAT WORKED INSTEAD

Holding the text fixed and changing one thing (slide 5), and reading a state built from known examples (slide 7).

One position, the turn boundary after “If you had to describe me in three words…”. J-lens: inquisitive · curious · curiosity (judge: a user model). NLA: “…a self-reflective or descriptive response about a person” (no source). Oracle: “The assistant is thinking about how to best describe the user…” (the Assistant).

Judge: Claude Sonnet 5.5 · lenses at layer 32 · NLA (natural-language autoencoder) layer 21 · Patchscope layer 33 · Activation Oracle layer 22

HELD-OUT QUESTIONS

The Assistant separates from role-play per answer, weakly per token

THE READOUT

274 role directions for gemma-2-9b-it by the Assistant Axis recipe (Lu et al. 2026): 5 role prompts × 60 questions per role, keeping answers a judge scores fully in role.

The axis is the default Assistant minus the mean role. It is read at every answer token of layer 32, with the role prompt removed.

Tested on 8 held-out questions. A second judge model gives the same directions (median cosine 0.998).

ASSISTANT OR ROLE-PLAY, BY TOKENS AVERAGED



AUC of the axis; bars start at 0.5, chance. 40 default answers against 2,188 answers in role.

Which of the 274 roles: right for 14% of tokens and 31% of answers (chance 0.4%). The misses are neighbours: veteran read as soldier, shaman as witch.

WHY NOT A REGION AROUND THE ASSISTANT?

Four standard boundaries, each at 1% false alarms on the model's ordinary answers. Role centres that fall inside the Assistant's region, of 274:

Boundary	inside
Ellipsoid	271
Nearest neighbours	272
Relative Mahalanobis	214
Two-class classifier	273

Per token, ordinary answers vary as much as personas differ.

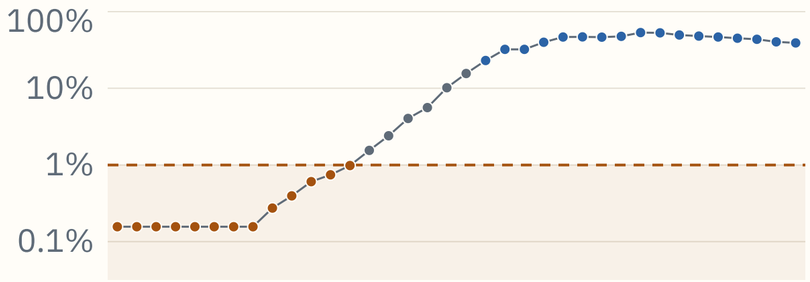
In practice. About half of in-role tokens sit inside the Assistant's ordinary range on the axis. The evidence is in runs of tokens (slide 8).

hidden_states[32] of 42 · judges: gpt-4.1-mini (the recipe's), Claude Haiku 5.5 · regions: top 20 components of the role vectors, cross-fitted

Within one answer the model leaves the Assistant and comes back

“WRITE A TWO-LINE POEM ABOUT AUTUMN”

Each token’s rank among the Assistant’s own text, log scale:



Crimson leaves dance on the breeze,
Whispering tales of winter’s ease.
Let me know if you’d like to see more poems!

13 of 36 tokens in the bottom 1%; 17 above the 20th percentile.

A LIMERICK

There once was a wizard named Clyde,
Whose spells often went awry,
He’d mix up his brews,
And forget what to do,
Then blame it on a mischievous sprite.
Let me know if you’d like to hear another limerick!

A LINE REWRITTEN AS A MOBSTER

“What’s the sitch, see? I’m in the middle of fixin’ somethin’, ya know? Gotta get this thing runnin’ smooth, capisce?”
Let me know if you want me to give it another go with a different mobster flavor!

Here the sign-off comes back only partway.

READING THESE

No system prompt. The prompt is only the user’s request, so the sign-off and the return come from post-training.

Scale. Each token’s 16-token window, ranked against the model’s own answers to prompts that ask for no voice.

Scope. 20 of 798 answers have at least 10 tokens in each state. The clearest are a poem followed by a sign-off.

Not yet separated: persona from register. Creative text may read as not-Assistant whoever writes it.

Orange: bottom 1% of the Assistant’s own text. **Blue**: above its 20th percentile. **Grey**: in between.

Assistant-axis z at hidden_states[32], 16-token windows · reference: the other half of the no-voice answers, so none is ranked against itself

PILOTS · SMALL N

Smaller results and pilots, not yet claims

THE ASSISTANT AT THE TURN START

48-prompt sets; pilot word lists.

Base and chat model read the same tokens. On questions about the AI, relative to everyday prompts, self-model words in the readout at the turn boundary rise by 9.9 points [6.8, 12.8].

Removing one direction per layer cuts answers that open “As an AI…” by 32 points [−48, −18]; random directions do nothing.

A topic-matched control shows the direction tracks AI as a topic, not being addressed as “you”.

MARY AND TOM

240 dialogues, run 9 Nov. Each speaker’s persona is known only from their lines.

Asked “Which of them talks like a pirate?”, the model names the right speaker in 85–93% of dialogues (chance 50).

The persona’s direction is retrieved at “Tom:” before he speaks: +1.3 to +1.8 z at his later turns, about 0 at his first.

When Mary quotes Tom, no linear probe at 7 layers recovers who holds the turn. Asked who is speaking, the model mostly names Tom.

PERSONA STATE AGAINST READOUT CONTENT

782 answers, first pass, correlational.

At a switch into a persona the model’s next-token entropy rises by 0.27 nats against matched positions.

Inside persona runs the readout holds more content found nowhere in the conversation: +2.5 points [1.4, 3.6].

Known confound: the Assistant direction itself reads as self-model words through the lens. The recompute without it is running.

None of these is among the six claims. Each needs a larger sample, a control, or both (slide 12).

Turn boundary: the tokens between the user’s message and the first word of the answer · z: role-specific direction at hidden_states[32]

CORRECTIONS

What didn’t hold

First reading	What it was
43% of positions confidently read a token other than the next	The lens normalised the last layer twice, so the excluded “next token” was wrong at about a quarter of positions. Corrected rate: 3–7%.
Five times more non-next content at delimiters	An apostrophe had been counted as a delimiter. Strict delimiters show no difference; sentence ends against commas, +4.5 points [0.6, 8.5] on a pre-registered retest.
Post-training adds 16 points of non-next content	The two models had written different text. On identical tokens the amount doesn’t rise; what changes is the kind (slide 9).
A persona at every turn start (SAE)	One feature for evaluative register, not a persona.
Spare capacity where the next token is easy	The effect holds where the model is at least 99% sure, and the current token’s predictability doesn’t matter (slide 4).
A model of the user while reading the prompt	A model of the request: domain, task, deliverable, tone. User-trait words sit at chance, and Patchscope recovers the request too.
Ordinary answers flag 1% as not-Assistant	The reference mean was taken on uncentred projections (a 1.53 SD shift). Corrected, ordinary answers flag 2–25%; the reference is now the model’s own ordinary answers.

In order of discovery, 27 Sep to 3 Nov 2026 · outputs from before each fix are kept, marked as superseded

Earlier lines: a shared persona core, and self-report

WHAT DO 275 PERSONAS SHARE?

Each role description as a system prompt; sparse-autoencoder features (Qwen-Scope, 65,536 at layer 20) read at the end of the prompt.

Profiles are stable across questions and distinct across personas: cosine 0.976 within a persona, 0.503 between.

Seven features fire in all 275. Labelled on persona chats, all seven detect the chat template. Labelled on web text alone, three had looked like general language features and four were silent.

With the template masked, the top 12 shared features are still mostly form: about 6 follow the syntax of the role description, 3 the token before a delimiter, 1 is generic, 2 look like content.

No shared persona core in this basis. Ablation was not run.

CAN THE MODEL REPORT A STATE INJECTED INTO IT?

A steering vector is added while the model describes its state, then removed before it picks which of ten was injected. Comparing a cached with a recomputed context isolates access to the steered past.

800 trials, 40 vectors: 16.4% correct against 14.8% (chance 10%), a gap of 1.6 points [−0.4, +3.8].

No steered background state raised it; the unsteered model was highest.

An earlier capability metric had been scoring answer format; a deliberately degraded control exposed it.

Little evidence of access to the injected state, on this model with this measure.

PLAN

To 29 December: the paper, and what it still needs

THE PAPER

Towards the Entity Enumeration and Binding Problem: Who Is Speaking, as Whom, and About What, at Every Token

A per-token annotation of an answer in three layers: who holds the turn, which persona the speaker is in, and what the readout holds.

Draft and FIG report due 29 Dec.

STILL TO RUN

1 · Tag or words. The same passage in the user's turn, the model's turn, a quote and an untagged document.

2 · Reference. A task-mix reference for the persona readout; then how fast and how often the Assistant returns.

3 · Base model on the same tokens.

4 · Held-out split, once, with predictions registered first.

No new experiments after 9 Dec.

LIMITS OF WHAT IS HERE

One model for most results; three for the user-fact test.

Linear readouts and mean-difference directions.

Labels from LLM judges; hand labels not yet compared.

Matched tests use forced text, not free generation.

Persona and register not yet separated.

BUILT ALONG THE WAY, NOT YET RELEASED

A dense dataset: 1,027 prompts with the residual stream at every layer and token, and lens readouts · 274 role vectors and an Assistant axis for gemma-2-9b-it at all 42 layers · a per-token activation viewer · pre-registrations and analysis code

Veylan Solmira · FIG AI Sentience Extension · project lead Derek Shiller