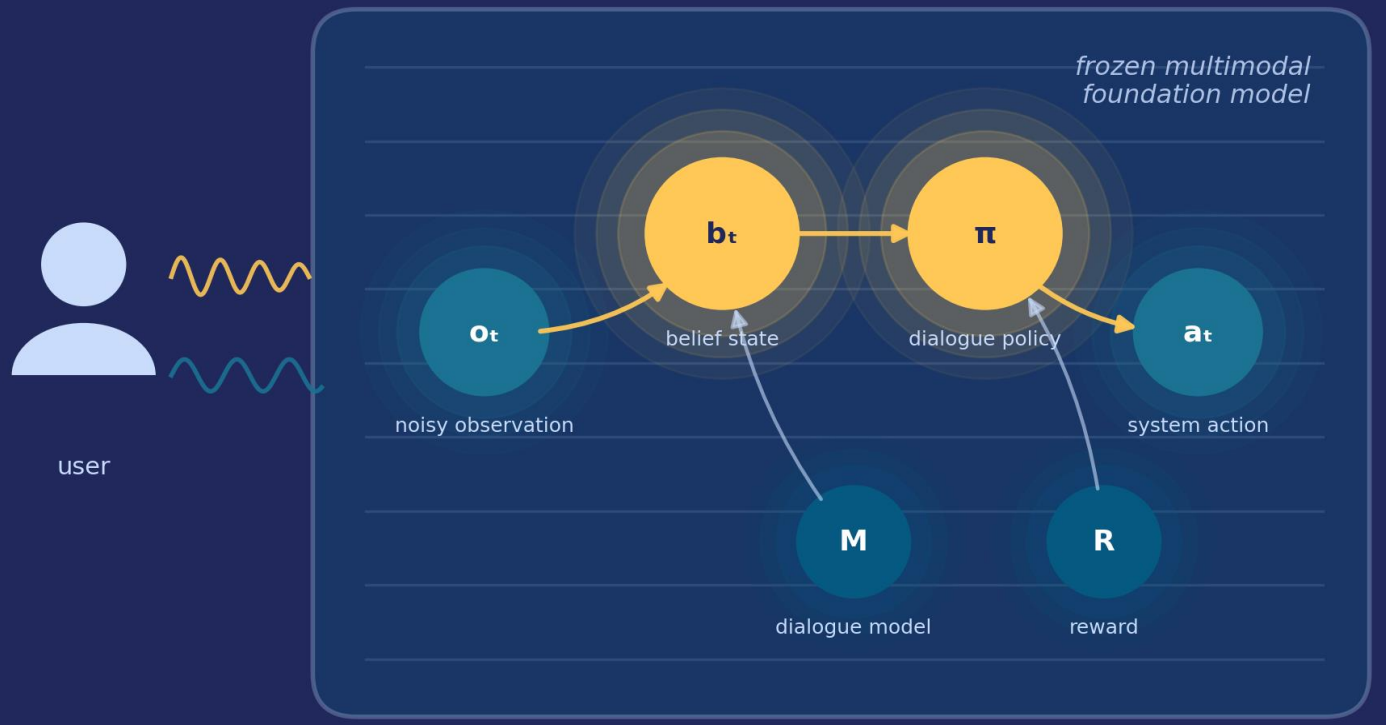
The Conversation Inside the Model

Probing and Controlling Interactional Structure in Multimodal Foundation Models



Larry Heck, Professor
ECE & Interactive Computing
Rhesa S. Farmer, Jr. Advanced Computing Concepts Chair
Georgia Research Alliance Eminent Scholar

OPENING

Two conversations

Same instruction, same context — HCRC Map Task corpus[1]

A HUMAN FOLLOWER (1991)

Giver: "Walk north until you reach the Roman baths."

Follower: "Roman baths? I don't have any Roman baths."

[query] — the mismatch surfaces immediately

Interaction managed. Repair before damage.

A SPEECH LLM (2026)

Giver: "Walk north until you reach the Roman baths."

Model: "Uh-huh."

[acknowledge] — fluent, plausible… map mismatch silently propagates

Plausible response — wrong move.

Fluency ≠ interaction: optimize the next utterance vs. manage the exchange

[1] Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. "The HCRC Map Task Corpus." Language and Speech 34(4), 351–366, 1991.

OPENING

The question

Conversation = **controlled, stateful dynamical interaction**

Foundation models: interactional structure represented inside?

Can it be
identified · observed · controlled?

observability — can we read the state?

controllability — can we write it?

OPENING

Dialogue Systems

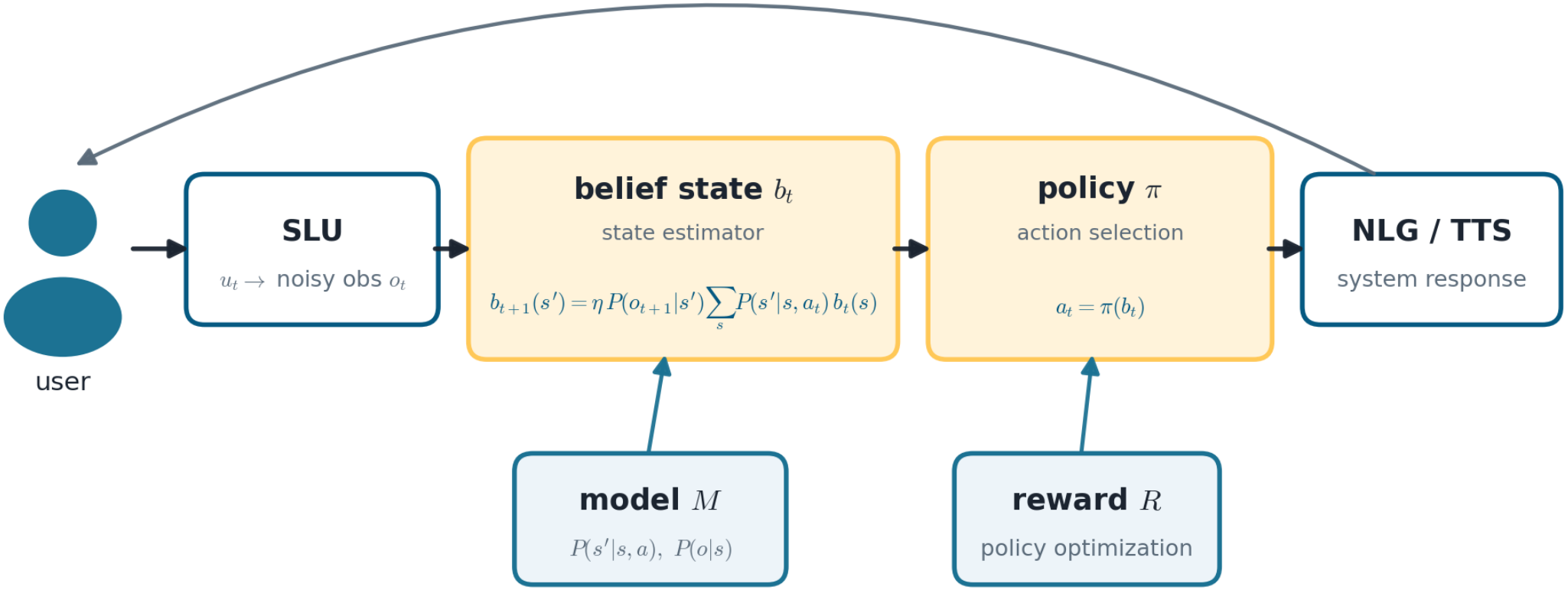
The arc: one loop, five acts

SENSE → ACTUATE → GOVERN — back to our classical roots but with the fluency of LLMs

Dialogue as a POMDP — the classical decomposition

objective — optimize the whole loop

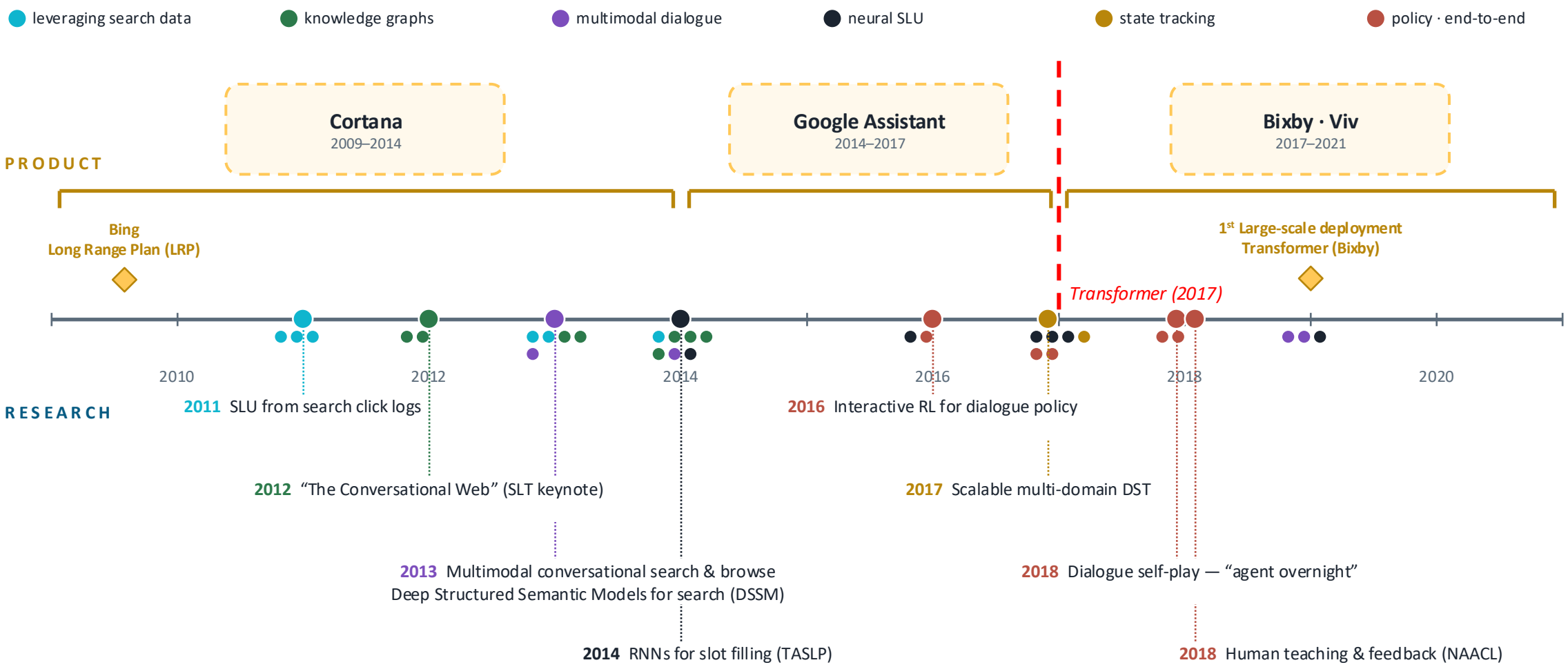
$$\pi^* = \arg\max_{\pi} \mathbb{E}\Big[\sum_t \gamma^t \, r(s_t, a_t)\Big]$$



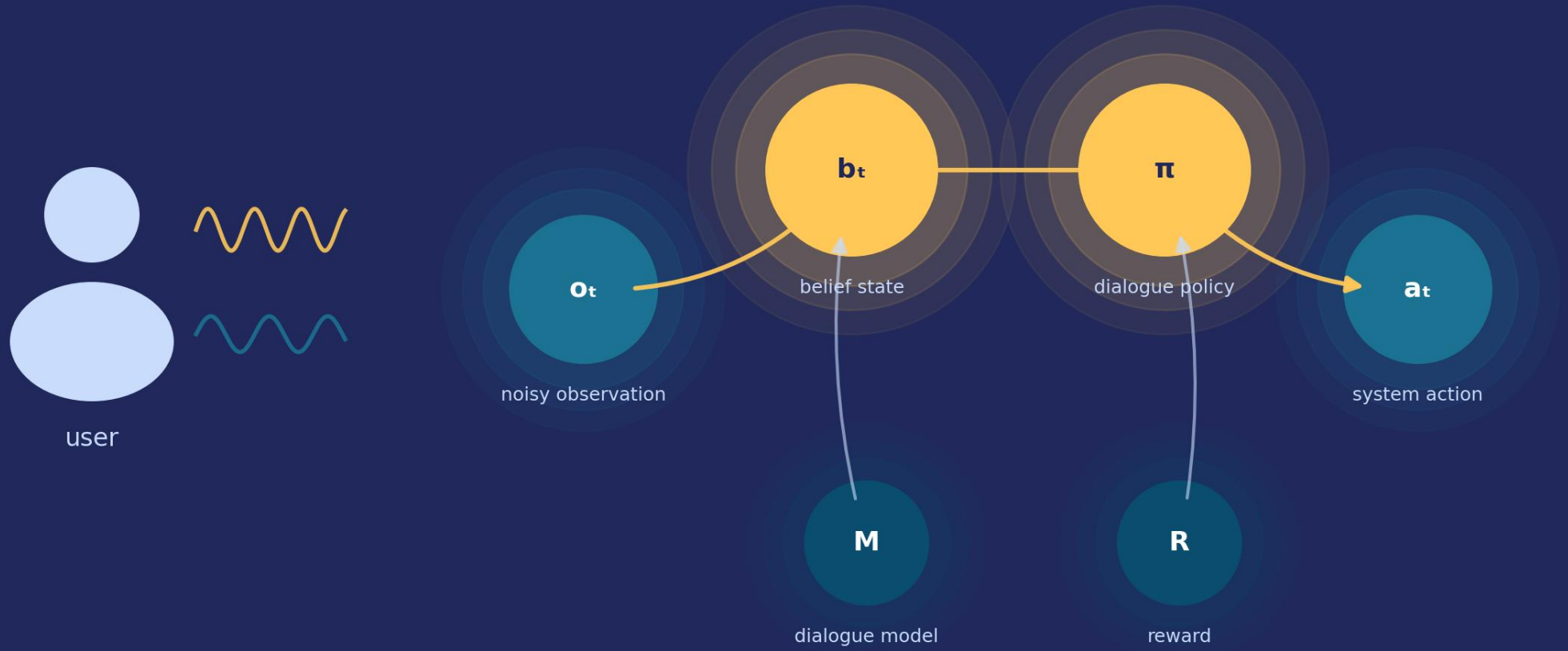
Williams & Young, “Partially Observable Markov Decision Processes for Spoken Dialog Systems,” Computer Speech & Language 21(2), 2007.
Young, Gašić, Thomson & Williams, “POMDP-Based Statistical Spoken Dialog Systems: A Review,” Proc. IEEE 101(5), 2013.

The explicit-state era (2009–2022) — my view from inside

the assistants — built on this community's theory. One traveler's path: all papers as dots by research thread · representatives labeled

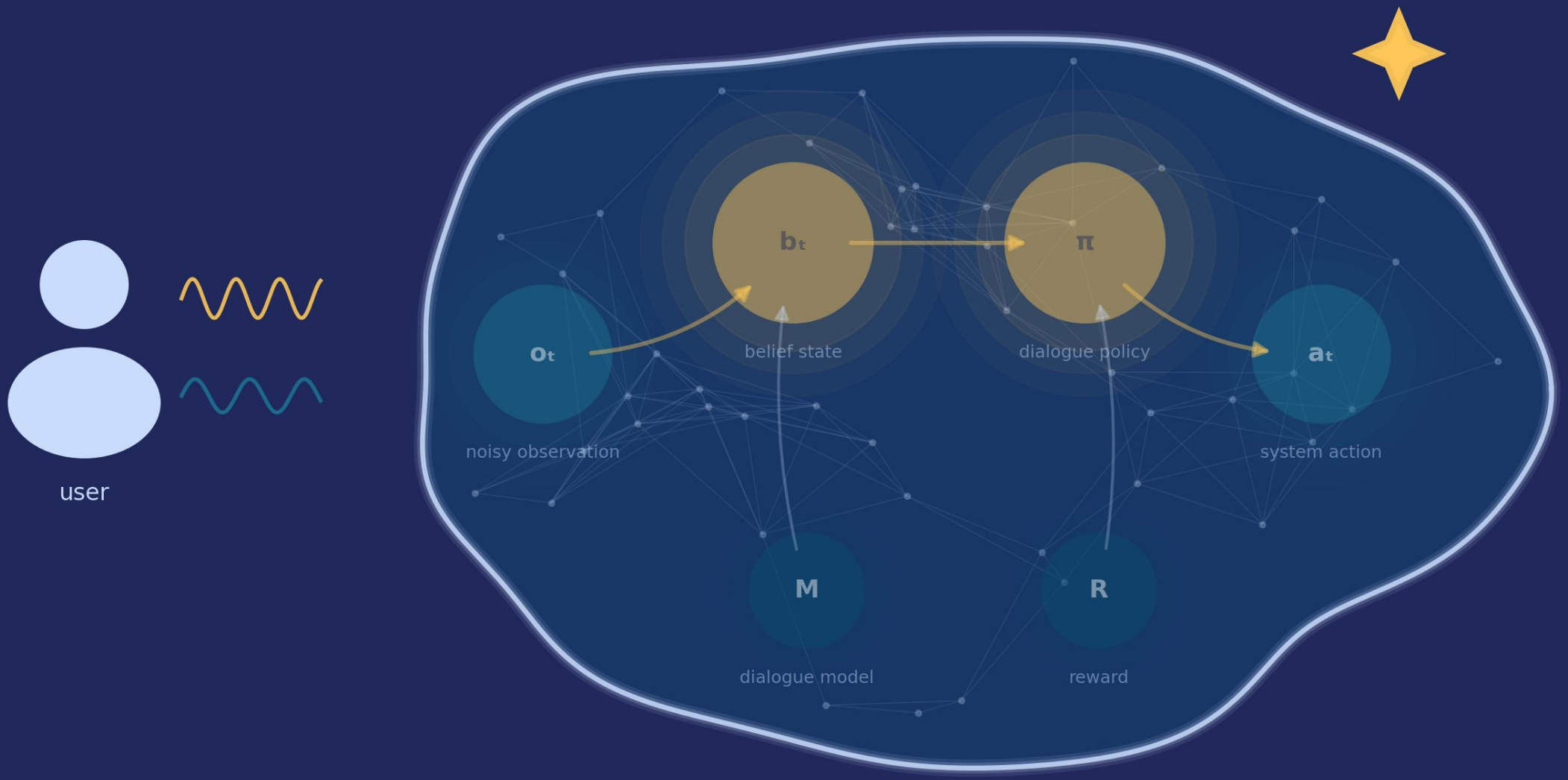


2009–2022: the decomposition, in production



the assistants — built on this decomposition (after Young, Gašić, Thomson & Williams, Proc. IEEE 2013)

2023: dialogue morphs into chat



ChatGPT → the foundation model swallows the pipeline…

2023: dialogue morphs into chat

…state no longer drives generation — no belief distributions · no optimized policy · no levers. Old diagram still in there?

2023–present: what actually survived

SURVIVED

the questions

grounding · turn-taking · dialogue acts — alive in SIGDIAL proceedings

MORPHED

state tracking, the task

DST → point estimates via prompting & function calling — no distributions

LOST

the levers

belief distributions · policy optimization · the causal state → action link

The decomposition left the architecture — not this community. Questions kept · levers lost.

Acts II–IV: win the levers back.

H. Wang et al., "A Survey of the Evolution of LM-Based Dialogue Systems," arXiv, 2023.
Qin et al., "End-to-End Task-Oriented Dialogue: A Survey," Proc. EMNLP, 2023.
Heck et al. (a different Heck), "ChatGPT for Zero-Shot Dialogue State Tracking: A Solution or an Opportunity?," Proc. ACL, 2023.
Proceedings of SIGDIAL 2024, Kyoto.

ACT II · READING THE LATENT SPACE

Can we read it?

Frozen model → activations → structure (SVD · probes) → causal test. Mission: win the levers back.

SENSE ACTUATE GOVERN

Emotions Matter

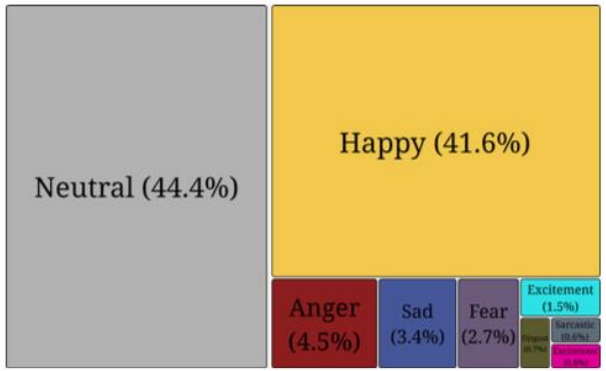


Figure 1: Distribution of emotions for a web corpus (Penedo et al., 2023).

Emotion	Results	
	Zero-shot	Trained
Neutral	48%	58%
Happy	36%	45%
Sad	34%	49%
Anger	31%	42%
Fear	34%	47%
Surprise	36%	54%
Disgust	38%	49%
Excitement	39%	55%
Sarcastic	39%	50%

Table 1: Disparity in performance across emotions for LLaMA-3.1-8B on AURA-QA.

Emotions Matter…but where are they?

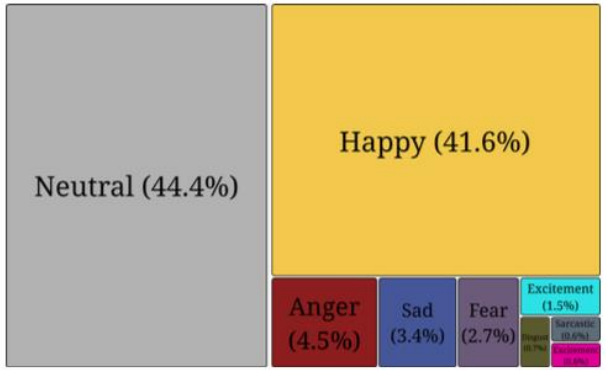
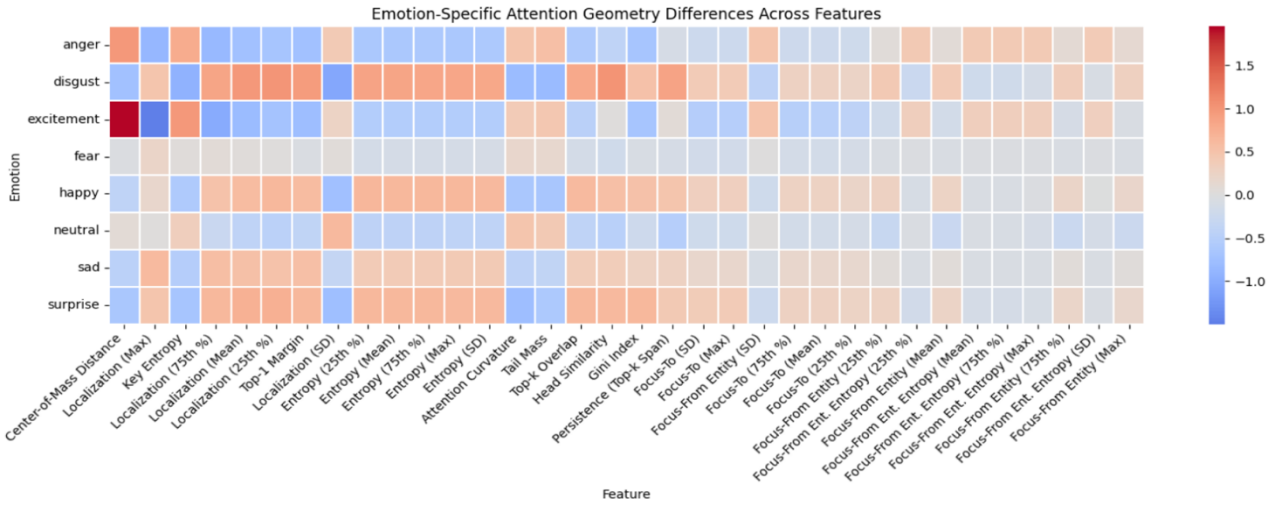
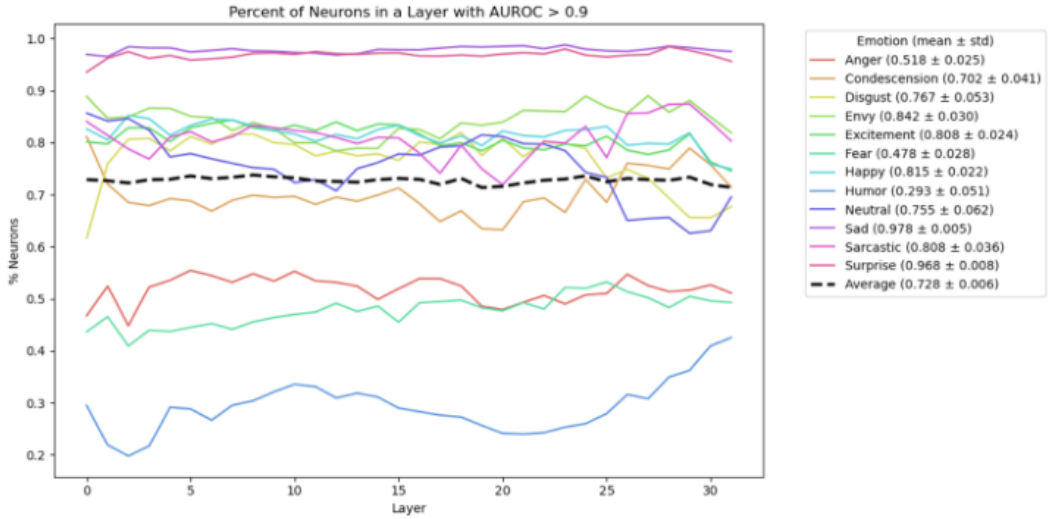


Figure 1: Distribution of emotions for a web corpus (Penedo et al., 2023).

Emotion	Results	
	Zero-shot	Trained
Neutral	48%	58%
Happy	36%	45%
Sad	34%	49%
Anger	31%	42%
Fear	34%	47%
Surprise	36%	54%
Disgust	38%	49%
Excitement	39%	55%
Sarcastic	39%	50%

Table 1: Disparity in performance across emotions for LLaMA-3.1-8B on AURA-QA.





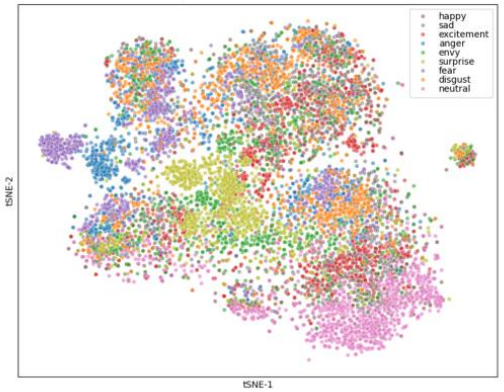
% of neurons per layer with AUROC > 0.9, by emotion — decodable at every one of 32 layers

AUROC — area under the ROC curve: how well one neuron's activation separates an emotion from all others. **0.5 = chance · 1.0 = perfect separation**

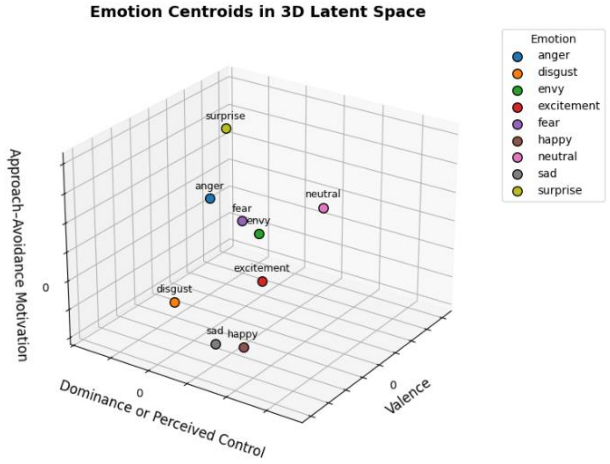
SAD: ~98% of neurons individually separate sad from everything else — in every layer

There is an emotional manifold

- Distributed: encoded across neurons and layers, not localized
- Centered SVD on hidden states → low-dim emotional subspace
- Emergent — no emotional supervision imposed
- Axes interpretable: valence · dominance · approach-avoidance



t-SNE of the emotional subspace



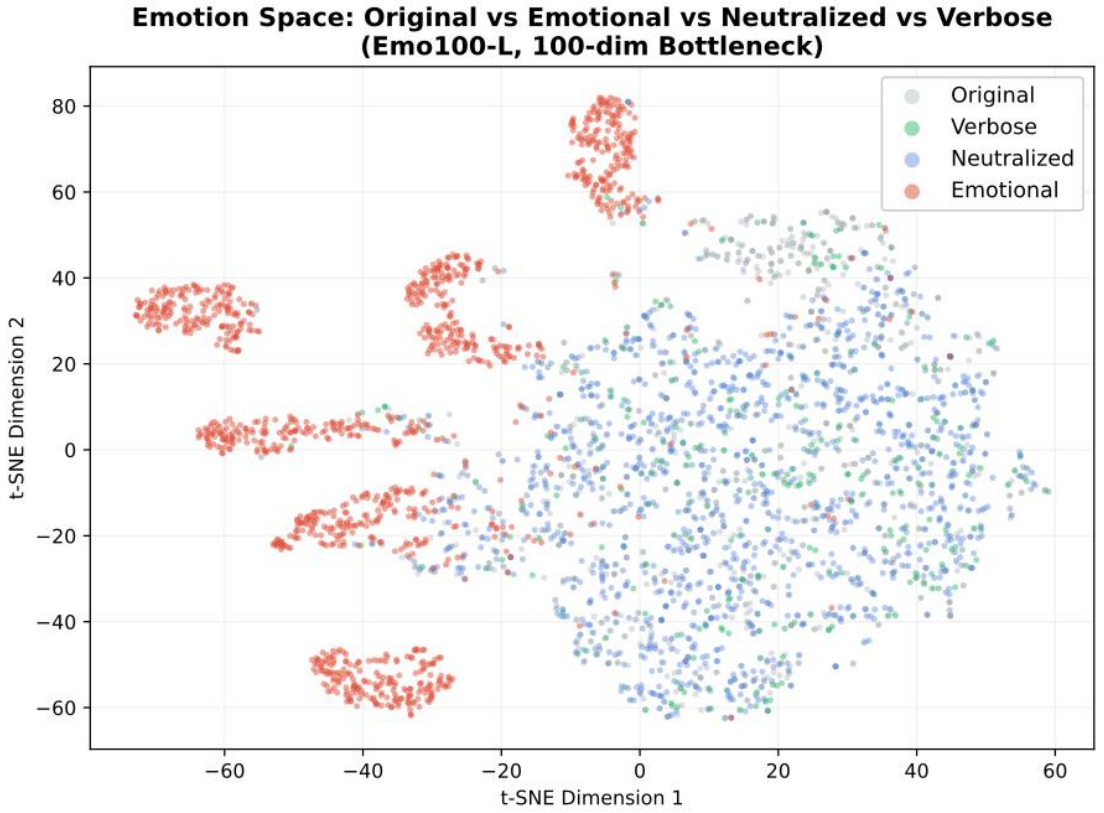
Emotion centroids on interpretable axes

> ***"The model built its own affect space — we only found the basis."***

Reichman, Avsian & Heck, "Emotions Where Art Thou: Understanding and Characterizing the Emotional Latent Space of LLMs," Proc. ICLR, 2026.

And it bends reasoning — TEMPER

- Emotional framing of identical math:
  reduces accuracy 2 to 10 % (18 models, 1B → frontier)
- All numbers · quantities · relations preserved (TEMPER-5400)
- Non-emotional paraphrase → no drop ⇒ the emotion is the cause
- Hidden states shift 3–4× farther than paraphrase — linearly separable



Emotional variants (red) leave the neutral manifold paraphrases (green) stay home

2-10%

Reduced accuracy points under emotional framing

3–4×

hidden-state shift vs paraphrase

Dokme, Reichman & Heck, "TEMPER: Testing Emotional Perturbation in Quantitative Reasoning," Proc. COLM, 2026.

Act II takeaway

✓ **exists** an emergent emotional manifold — interpretable axes

✓ **universal** 8 datasets · 5 languages · every layer

✓ **visible** 86% from attention geometry alone

✓ **consequential** reduces accuracy up to 10% on reasoning, causally

Reading works → can we write?

ACT III · WRITING THE LATENT SPACE

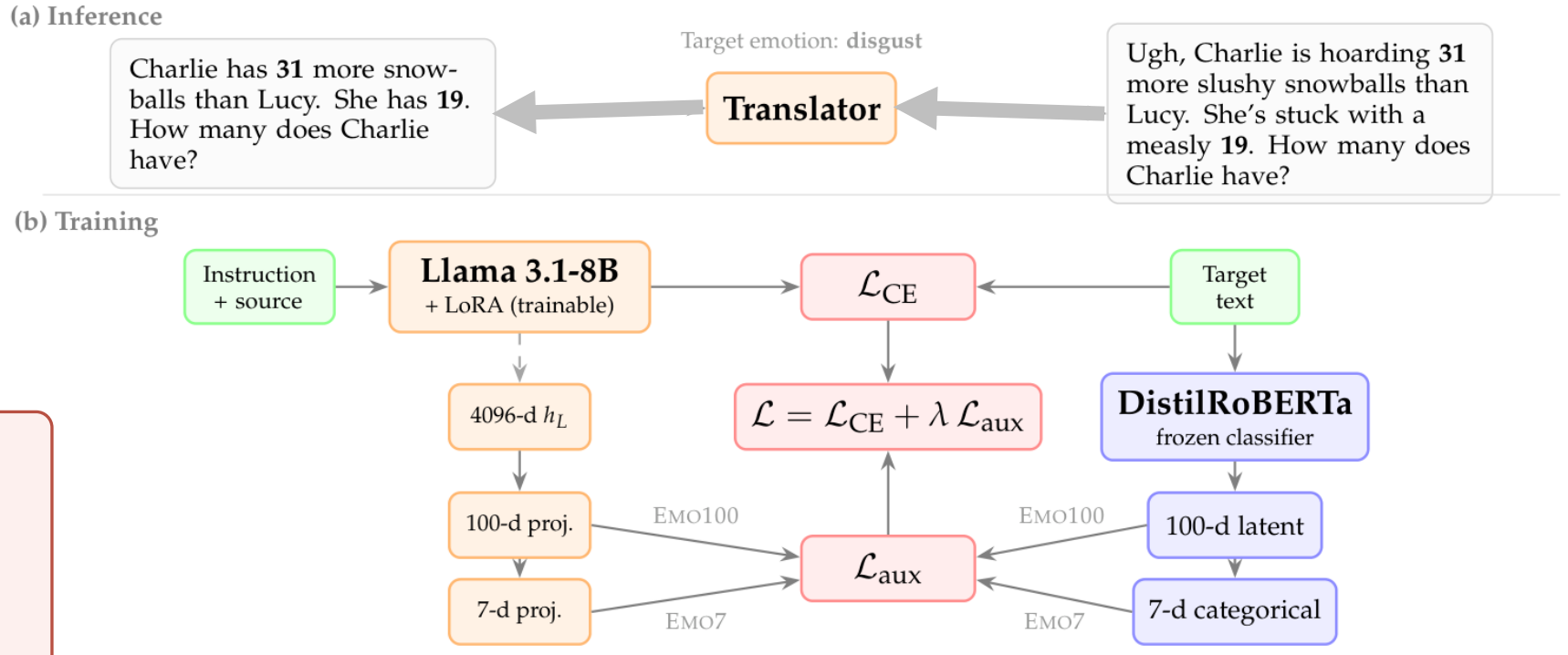
Can we write it?

Four writes, by timescale: neutralize → steer → train → repair
Same geometry — increasing persistence.

✓ SENSE | ACTUATE | GOVERN

Write #1 · neutralize — remove the drift (emotion)

- Teacher–student emotion translators: emotional ↔ neutral, math intact
- Idea: combine standard LLM fine-tuned translator with frozen emotion classifier
  → align translator's internal representation with classifier's emotion space
  → neutralize emotion



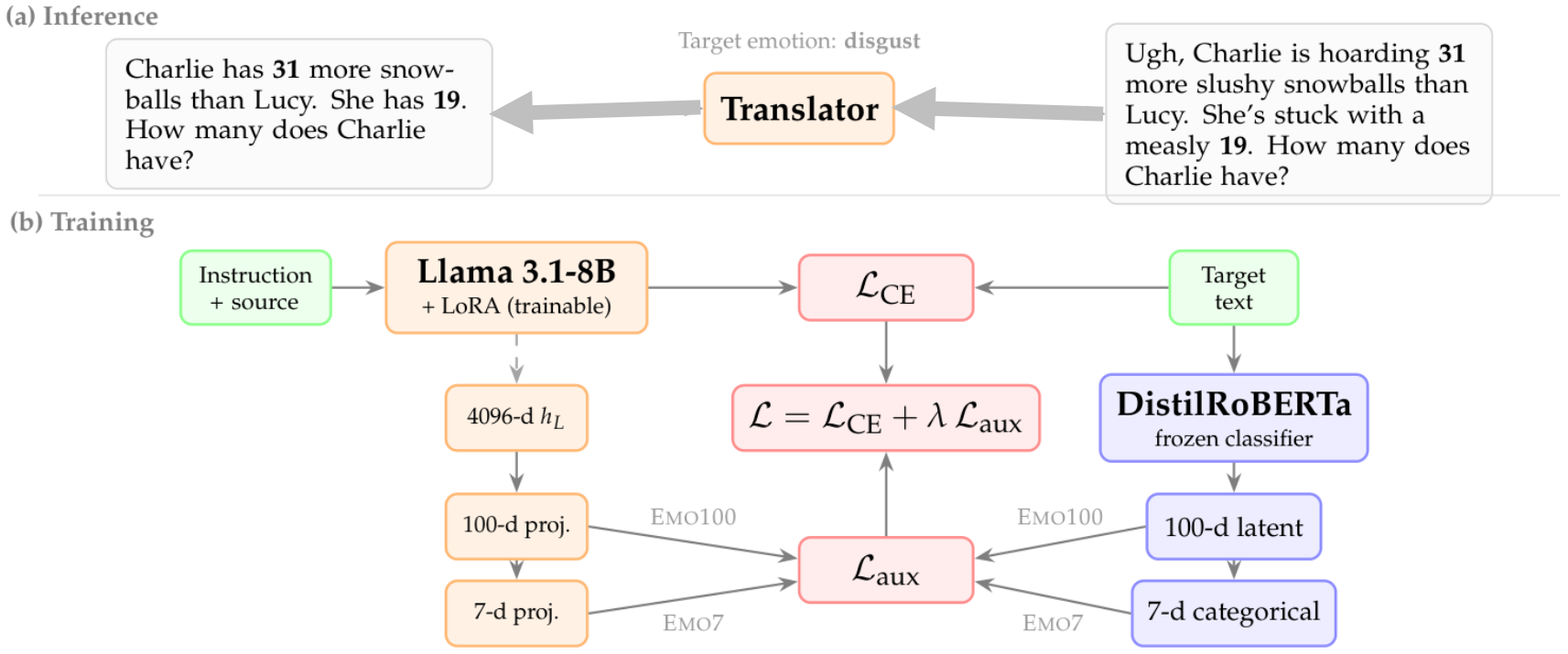
-10%

Reduced accuracy points under emotional framing

Dokme, Reichman & Heck, "TEMPER," Testing Emotional Perturbation in Quantitative Reasoning, Proc. COLM, 2026.

Write #1 · neutralize — remove the drift (emotion)

- Teacher–student emotion translators: emotional ↔ neutral, math intact
- Idea: combine standard LLM fine-tuned translator with frozen emotion classifier
  → align translator's internal representation with classifier's emotion space
  → neutralize emotion



Write #2 · steer — change the interpretation

- Learned intervention module in the emotional subspace
- Steers internal emotion perception — semantics preserved
- Works across models and languages · strongest on basic emotions

10% → 86%

target-emotion classification after steering (Llama-3.1-8B, En)

≈ 0.1–0.2

semantic-similarity loss — meaning intact

Model · language	base → steered (overall)
Llama-3.1-8B · English	**9% → 83%**
Llama-3.1-8B · non-English	**8% → 84%**
OLMo-v2 · English	**11% → 91%**
OLMo-v2 · non-English	**11% → 86%**
Ministral · English	**10% → 91%**
Ministral · non-English	**11% → 79%**

top-1 target-emotion prediction, before → after learned steering
Sad, Happy, Fear, Anger, Neutral, Disgust, Envy, Excitement, Surprise

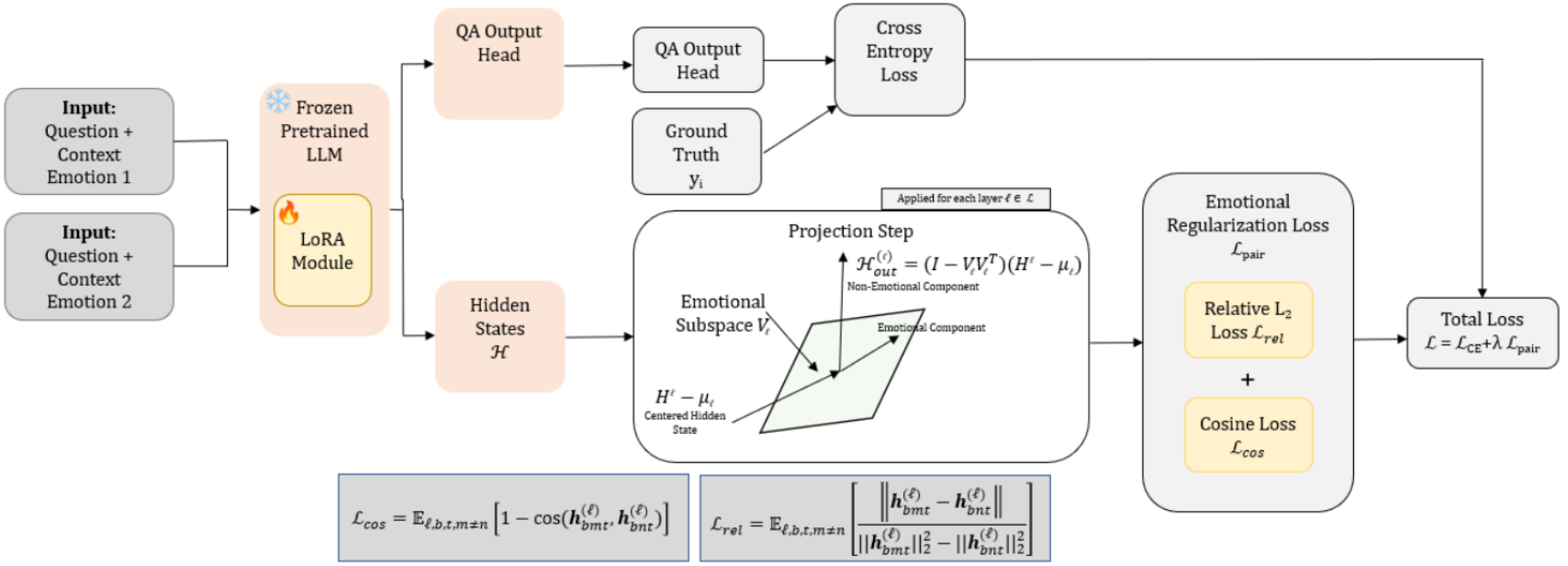
Dataset	Original Text	Shift Emotion	Steering Rewrite
CARER (Twitter)	i waited in line longer than usual i didnt feel impatient that my business was delayed i listened to the master about why this was occurring and how i could be of service during that moment.	Anger	Are you kidding me?! I waited in line longer than usual and I'm still trying to contain my frustration that my business was delayed! I listened to the master about why this was occurring and how I could be of service during that moment, but honestly, it's just not good enough!
GoEmotions	Is that seriously the story? I haven't googled it. I just thought it was cool as a WW2 buff.	Happy	Is that seriously the most amazing story? I haven't googled it yet, but I just thought it was cool as a WW2 buff!
EmoTextToKids (FR)	Je savais que je prenais des risques	Fear	Je crains que je prenais des risques

Reichman, Avsian & Heck, "Emotions Where Art Thou: Understanding and Characterizing the Emotional Latent Space of LLMs," Proc. ICLR, 2026.

Write #3 · train — confine the drift

- Project hidden states onto the emotional subspace V_l — every layer
- Penalize emotion-conditioned drift (cos + rel-L_2) with the QA loss
- Frozen LLM + LoRA → affect confined to its subspace
- QA robustness ↑ in every emotional condition — semantics untouched

Act II found the subspace.
Here it becomes a training objective — the write, baked in.



LoRA: +10%
LoRA + Regularization: +16%

Natural Question dataset
Relative Gains (%)
(LLaMA-3.1-8B, Ministral, and Olmov2)

Reichman, Avsian, Webster & Heck, "Emotion is Not Just a Label: Latent Emotional Factors in LLM Processing," Proc of ICLR Workshop on ReAlign, 2026 .

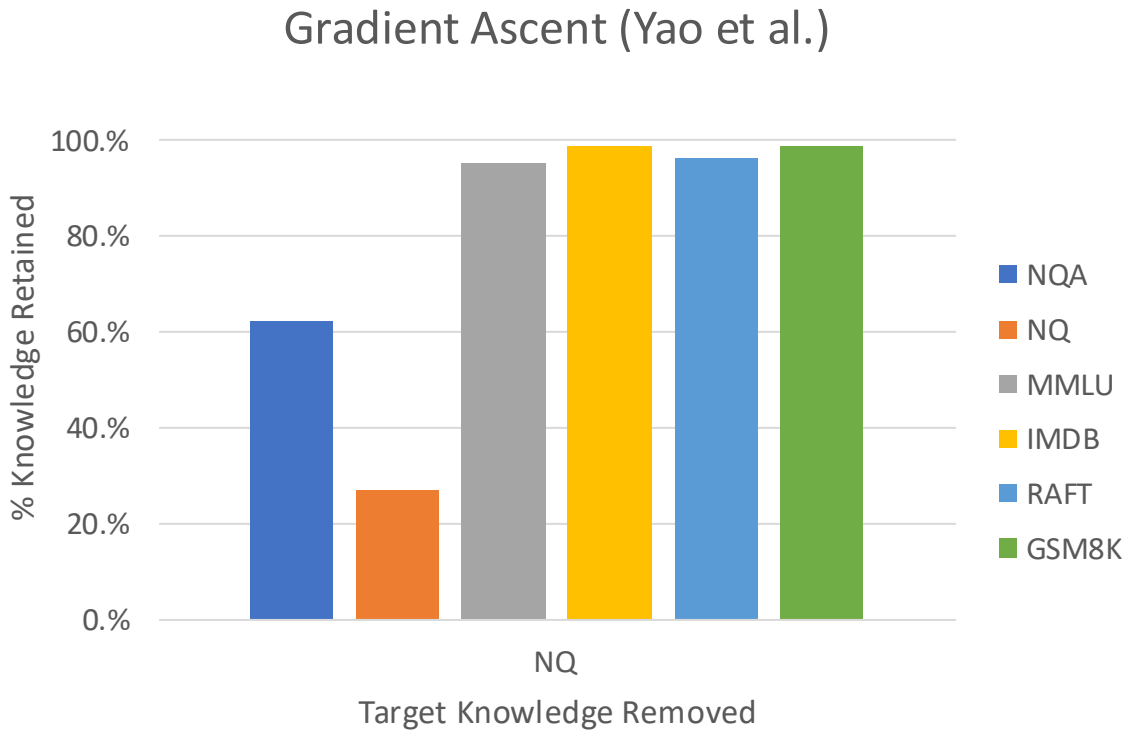
Write #4 · repair — neural surgery

Efficient removal of knowledge is becoming increasingly important

- "Right to be Forgotten" (Goldman, 2020)
- EU General Data Protection Regulation "GDPR" (Goddard, 2017)
- Aging adults (Zubatiy, 2024)
- Propriety knowledge (e.g., creative works)
- Unreliable knowledge (e.g., mathematical calculations)

UNLEARN: Efficient remove of knowledge in LLMs

- Forgets (unlearns) knowledge of LLM *without impact on related knowledge*
- Edits (trillions of) LLM weights simultaneously
- Subspace method targets knowledge
- Discrimination method separates target from related knowledge



- Narrative QA **(NQA)**
- Natural Questions **(NQ)**
- Massive Multitask Language Understanding **(MMLU)**
- IMDB benchmark for sentiment analysis in movies **(IMDB)**
- Real-world Annotated Few-Shot **(RAFT)**
- Grade School Math 8K **(GSM8K)**

Lizzo & Heck, "UNLEARN: Efficient Removal of Knowledge in Large Language Models," Findings of NAACL, 2025.
Lizzo & Heck, "Evaluating Cross-Lingual Unlearning in Multilingual Language Models," arXiv:2601.06675, 2026.
Lizzo & Heck, "QR-Erase: Efficient Subspace-Based Machine Unlearning with Layer Localization," AAAI (under review)

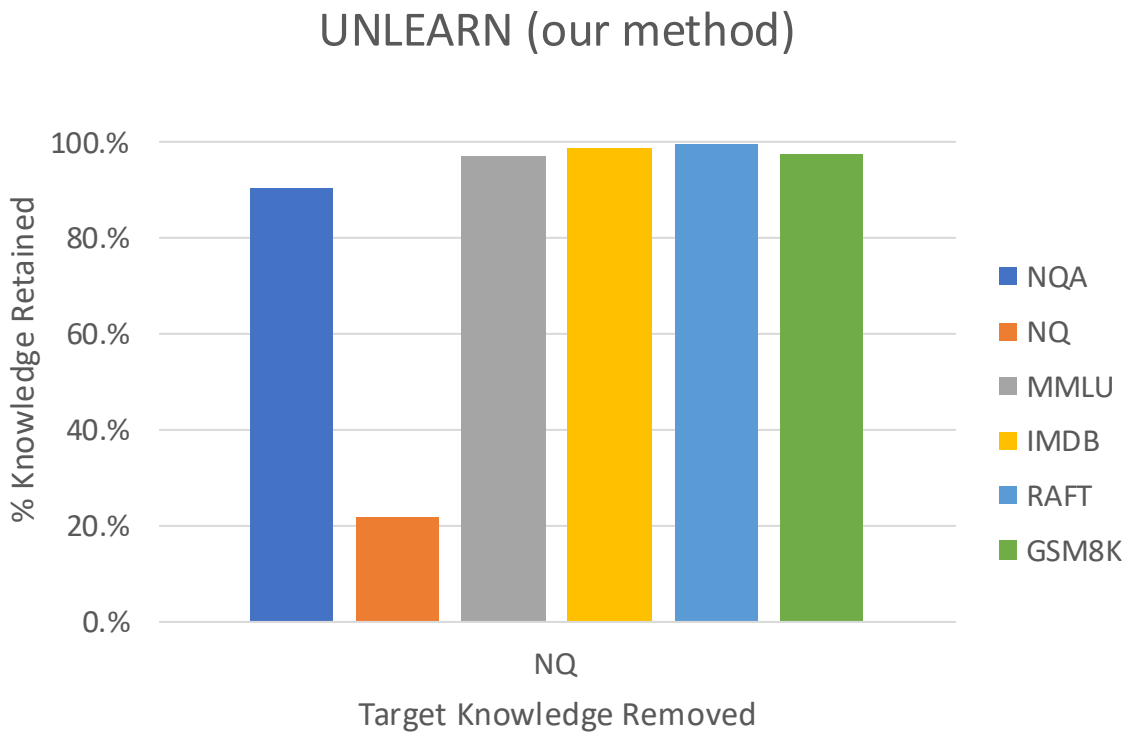
Write #4 · repair — neural surgery

Efficient removal of knowledge is becoming increasingly important

- "Right to be Forgotten" (Goldman, 2020)
- EU General Data Protection Regulation "GDPR" (Goddard, 2017)
- Aging adults (Zubatiy, 2024)
- Propriety knowledge (e.g., creative works)
- Unreliable knowledge (e.g., mathematical calculations)

UNLEARN: Efficient remove of knowledge in LLMs

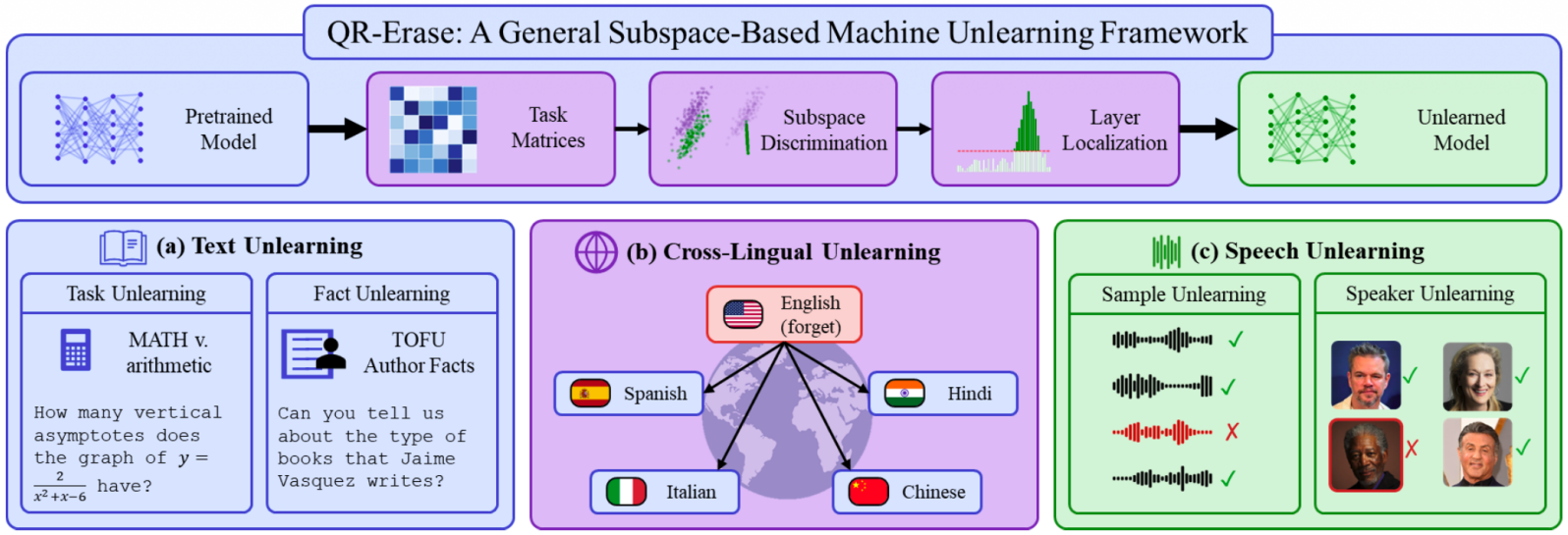
- Forgets (unlearns) knowledge of LLM *without impact on related knowledge*
- Edits (trillions of) LLM weights simultaneously
- Subspace method targets knowledge
- Discrimination method separates target from related knowledge



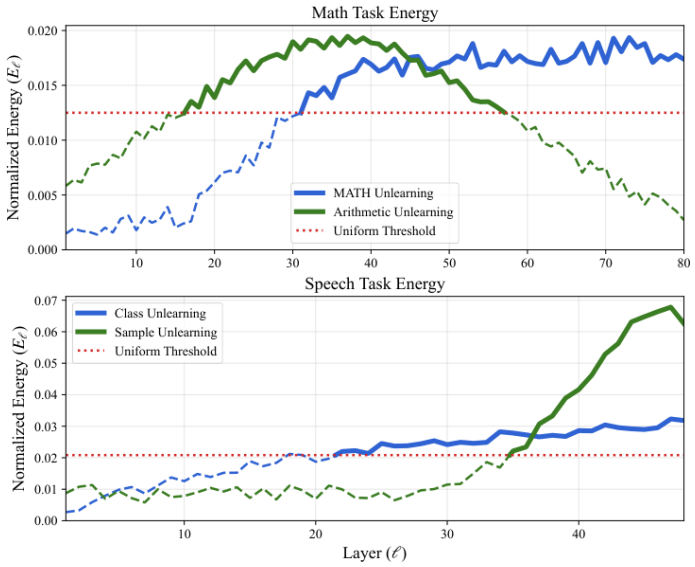
- Narrative QA **(NQA)**
- Natural Questions **(NQ)**
- Massive Multitask Language Understanding **(MMLU)**
- IMDB benchmark for sentiment analysis in movies **(IMDB)**
- Real-world Annotated Few-Shot **(RAFT)**
- Grade School Math 8K **(GSM8K)**

Lizzo & Heck, "UNLEARN: Efficient Removal of Knowledge in Large Language Models," Findings of NAACL, 2025.
Lizzo & Heck, "Evaluating Cross-Lingual Unlearning in Multilingual Language Models," arXiv:2601.06675, 2026.
Lizzo & Heck, "QR-Erase: Efficient Subspace-Based Machine Unlearning with Layer Localization," AAAI (under review)

Write #4 · repair — neural surgery



$$\sin\Theta(\mathcal{F},\widehat{\mathcal{F}}) = \mathcal{O}\left(\frac{\sigma_{k+1}}{\sigma_k}\right)$$



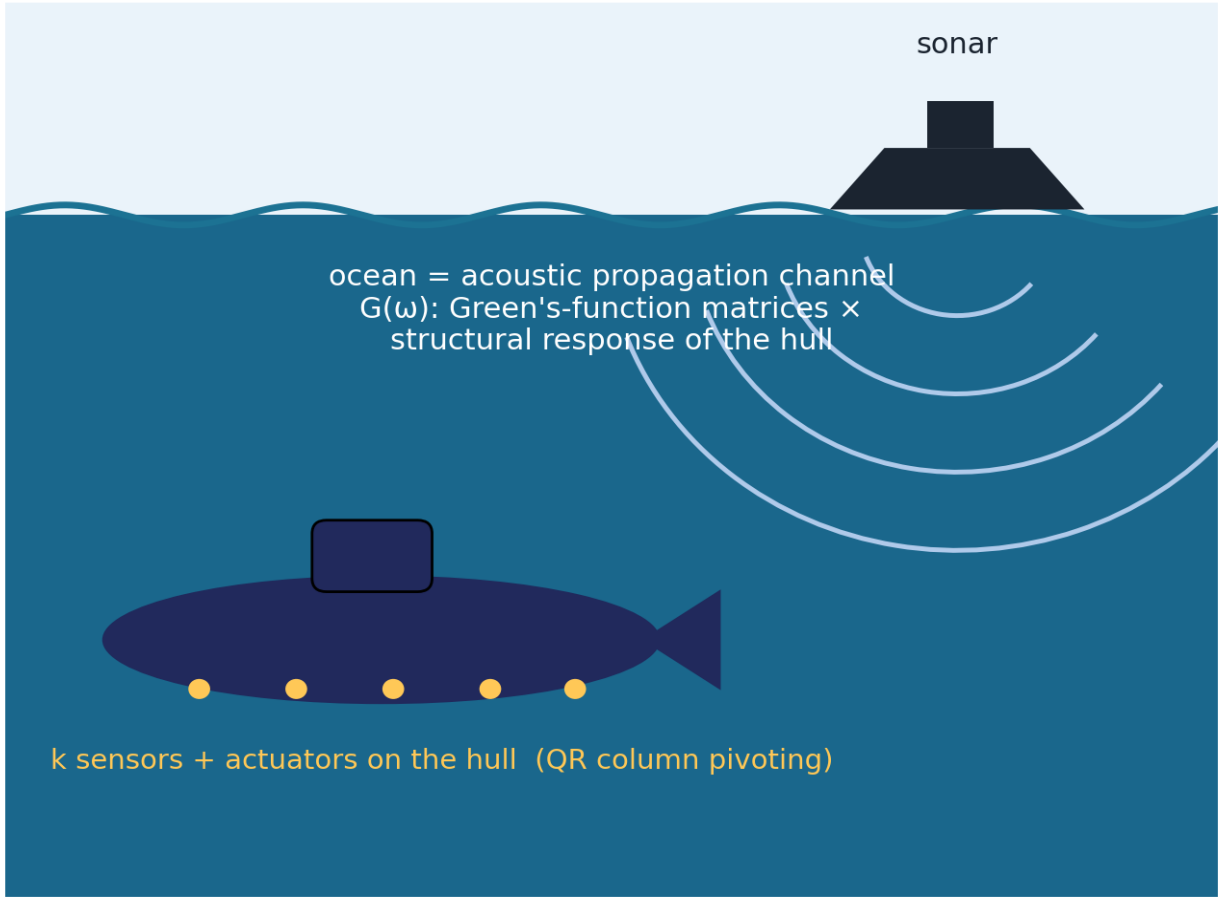
layer-wise task energy — cut only above threshold

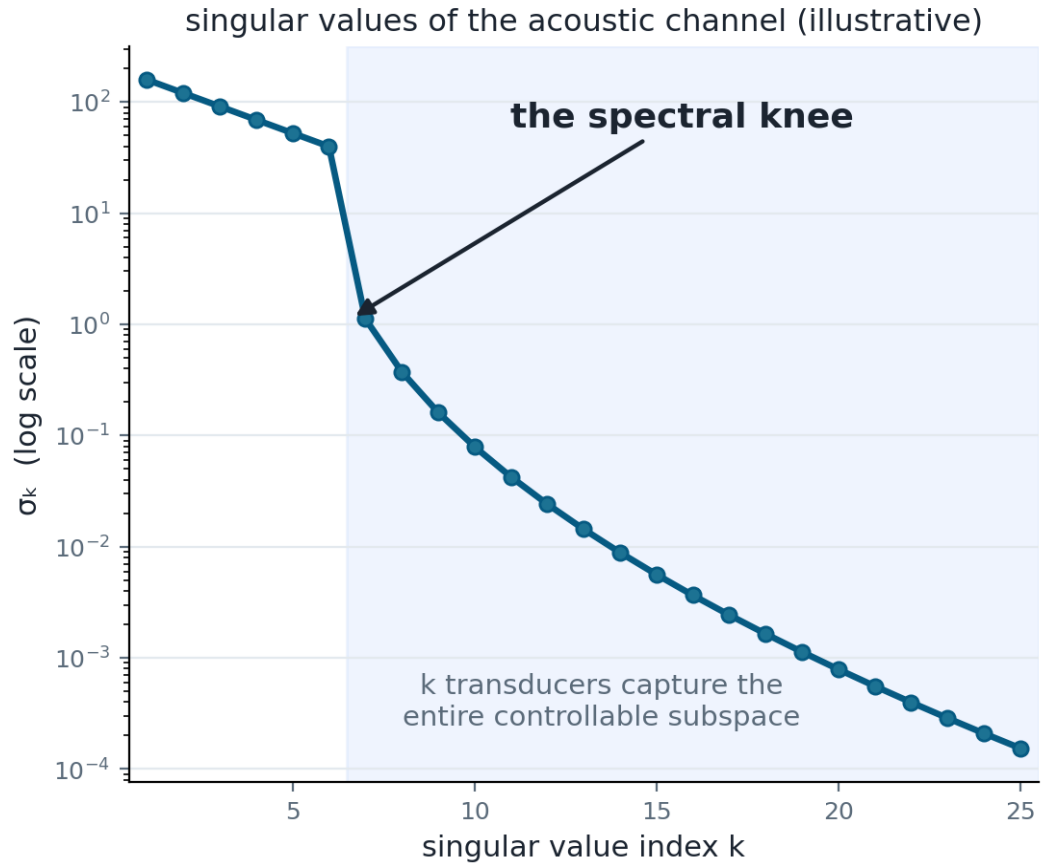
Surgical repair of a deployed LLM — remove · correct · without retraining, gradients, or collateral damage

Lizzo & Heck, "UNLEARN: Efficient Removal of Knowledge in Large Language Models," Findings of NAACL, 2025.
Lizzo & Heck, "Evaluating Cross-Lingual Unlearning in Multilingual Language Models," arXiv:2601.06675, 2026.
Lizzo & Heck, "QR-Erase: Efficient Subspace-Based Machine Unlearning with Layer Localization," AAAI (under review)

Why pivoted QR? A story from 1995

$$\sin\Theta(\mathcal{F},\widehat{\mathcal{F}}) = \mathcal{O}\left(\frac{\sigma_{k+1}}{\sigma_k}\right) \rightarrow 0$$





Search problem explodes: 1.1×10²³ configs → one pivoted QR. ***"The spectrum tells you when you're allowed to be greedy."***

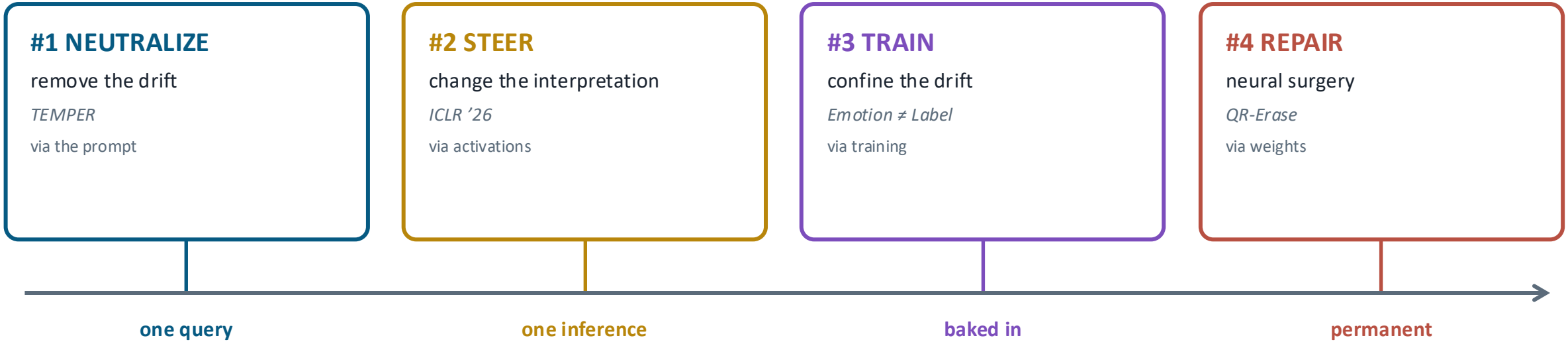
Heck, "Broadband Sensor and Actuator Selection for Active Control of Smart Structures," Proc. SPIE 2442 (Smart Structures and Materials), 292–303, 1995.

1995, meet 2026

	1995 · SPIE (submarines)	2026 · QR-Erase (LLMs)
Plant	Submarine hull in the ocean	Frozen foundation model
Matrix	Frequency-response / Green's-function matrices	Layer-wise task matrices T_f
Operation	CPQR column pivoting selects transducers	CPQR recovers the forget subspace
Bound	$\sin \Theta(A, A\Pi_a) \leq \sigma_{r+1} \, \lvert\lvert R_{11}^{-1} \rvert\rvert$	$\sin \Theta(F, \hat{F}) \leq \sigma_{k+1} \, \lvert\lvert R_{11}^{-1} \rvert\rvert_2$
Why it works	Spectral knee of the acoustic channel	Task knowledge is low-rank (forgetting saturates by rank ≈ 8–10)

"Thirty-one years. Same factorization, same bound, same spectral-gap condition.
Only the plant changed — from a hull in the ocean to a transformer's weight space."

Act III takeaway — the write stack



Writable at every timescale — surgically, neighbors intact.

Read ✓ Write ✓ → run a conversation this way?

ACT IV · THE CONVERSATION INSIDE THE MODEL

Latent-IM: recovering the dialogue manager

Frozen speech LLM, two coupled problems: selection (which move · when to yield) → realization (produce it).
Avsian, Dokme, Woo & Heck — AAAI, under review

✓ SENSE ✓ ACTUATE GOVERN

Conversational "Moves" — fifty years of theory

CONTENT · INTENTS & SLOTS

what the utterance is about

"Book a table for two at Nori at 7" → book_table · {Nori · 2 · 7 pm}

fills the API call — the 2010s' goal

domain-specific — rebuilt for each domain

vs

INTERACTION · MOVES

what the utterance does to the exchange

"Was that 7 tonight?" → check · "Uh-huh." → acknowledge

manages grounding · repair · turn-taking

domain-general — transfers across domains

the moves layer has half a century of theory behind it:

1974

Turn-taking as machinery

conversation has systematics — Conversation Analysis

Sacks, Schegloff & Jefferson

1991

Grounding

conversation = joint action; contributions must be grounded

Clark & Brennan

1992

Conversation acts

grounding + core acts, computationally specified

Traum & Hinkelman

1997

Move coding, domain-independent

HCRC moves on MapTask · DAMSL acts

Carletta et al. · Core & Allen

2000

Switchboard: open-domain talk

SWBD-DAMSL — the same functions dominate casual conversation

Stolcke, Ries, Coccaro, Shriberg, …

Domain-general functions survive domain change.

Sacks, Schegloff & Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," Language 50(4), 1974.
Clark & Brennan, "Grounding in Communication," in Perspectives on Socially Shared Cognition, APA, 1991.
Traum & Hinkelman, "Conversation Acts in Task-Oriented Spoken Dialogue," Computational Intelligence 8(3), 1992.
Carletta et al., "The Reliability of a Dialogue Structure Coding Scheme," Computational Linguistics 23(1), 1997. · Core & Allen, "Coding Dialogs with the DAMSL Annotation Scheme," AAAI Fall Symposium, 1997.
Stolcke, Ries, Coccaro, Shriberg, Bates, Jurafsky, Taylor, Martin, Van Ess-Dykema & Meteer, "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech," Computational Linguistics 26(3), 2000.

Why moves — and why now?

1 The transaction era · 2010s

- assistants existed to call an API
- intents & slots = the right optimization
- ATIS → Cortana · G Assistant · Bixby → MultiWOZ
- moves were hard to govern — no levers, brittle rules
- short, single-initiative — just re-ask if it breaks

2 Content came free · LLM era

- zero-shot intent · slot · DST — by prompting
- semantic & task structure sits in frozen states
- function calling = intents & slots, reborn

  ⇒ the moves layer is the residual gap

3 The tasks changed

- long-horizon collaboration · agents
- full-duplex speech · digital humans
- one wrong move silently corrupts the task
- (“Roman baths… uh-huh.” — the opening slide)

The 2010s optimized the transaction. The 2020s must optimize the interaction —
content came free with LLMs; interaction didn’t.

and (next) the moves layer, too, lives inside foundation models — readable, steerable

Tenney, Das & Pavlick, “BERT Rediscovers the Classical NLP Pipeline,” Proc. ACL, 2019.
Hendel, Geva & Globerson, “In-Context Learning Creates Task Vectors,” Findings of EMNLP, 2023.
Todd et al., “Function Vectors in Large Language Models,” Proc. ICLR, 2024.
Heck et al. (a different Heck), “ChatGPT for Zero-Shot Dialogue State Tracking: A Solution or an Opportunity?,” Proc. ACL, 2023.

The task: asymmetric, instruction-oriented dialogue

a giver instructs · a follower executes — the follower turn is the unit of prediction and steering

$D = ((r_1, u_1), \ldots, (r_T, u_T))$ — dialogue: roles + utterances

$c_t = ((r_1, u_1), \ldots, (r_{t-1}, u_{t-1}))$ — context before follower turn t

$y_t \in \mathcal{M} = \{\text{acknowledge}, \text{check}, \text{explain}, \text{query}, \text{reply}\}$

Selection: $\hat{y}_t = \arg\max_{m \in \mathcal{M}} p_\theta(y_t = m \mid c_t)$ — which move

Realization: $\hat{u}_t = g_\phi(c_t, y_t^\star)$ — how it is expressed

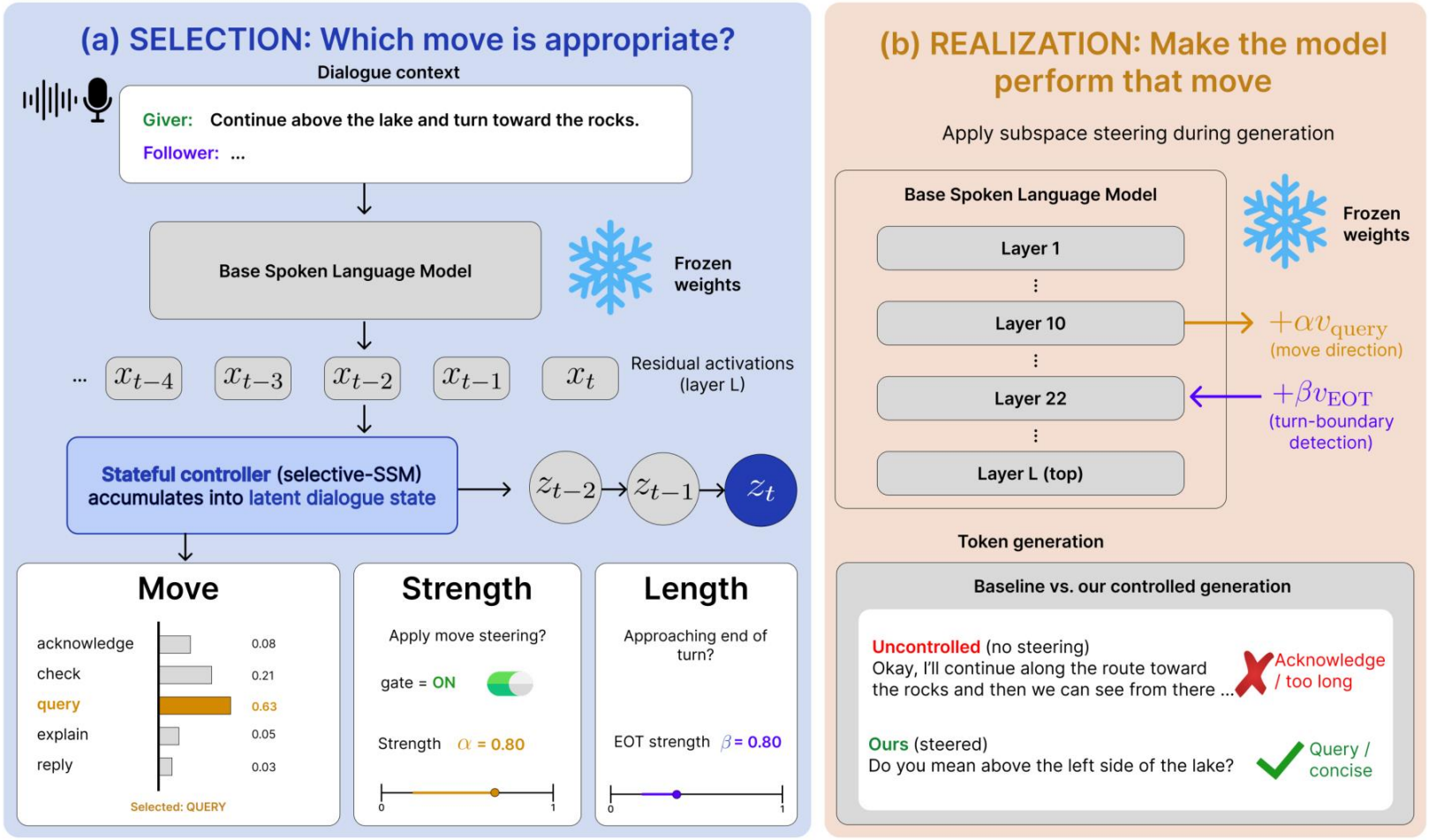
Move	Function (follower turns)
acknowledge	heard / ready — no new task info
check	verify instruction · confirm understanding
explain	add task-relevant info · clarify own state
query	request info — yes/no + wh
reply	answer question or check

The label space $\mathcal{M}$ — merged follower-move types (Table 1)

HCRC lineage (Carletta '97) · MapTask '91 · FindTask/RoboHelper · CReST — this community's corpora → frontier interpretability, 35 years on

Avsian, Dokme, Woo & Heck, "Latent-IM: Latent Interaction Management for Speech LLMs," arXiv:2607.26928 · AAAI 2027 (in review).
Carletta et al., "The Reliability of a Dialogue Structure Coding Scheme," Computational Linguistics 23(1), 1997.

Latent-IM — our approach



① READ
controller reads pooled activations
$x_{t-4}, \ldots, x_t \rightarrow$ latent state z_t
$\rightarrow$ move · strength α · length β
$\rightarrow$ ***selection***

② WRITE · move
inject $+\alpha v_{\text{query}}$ at the move's layer
— the selected move's direction
$\rightarrow$ ***realization***

③ WRITE · turn
inject $+\beta v_{\text{EOT}}$ at its layer —
one turn-boundary direction
$\rightarrow$ ***turn-taking***

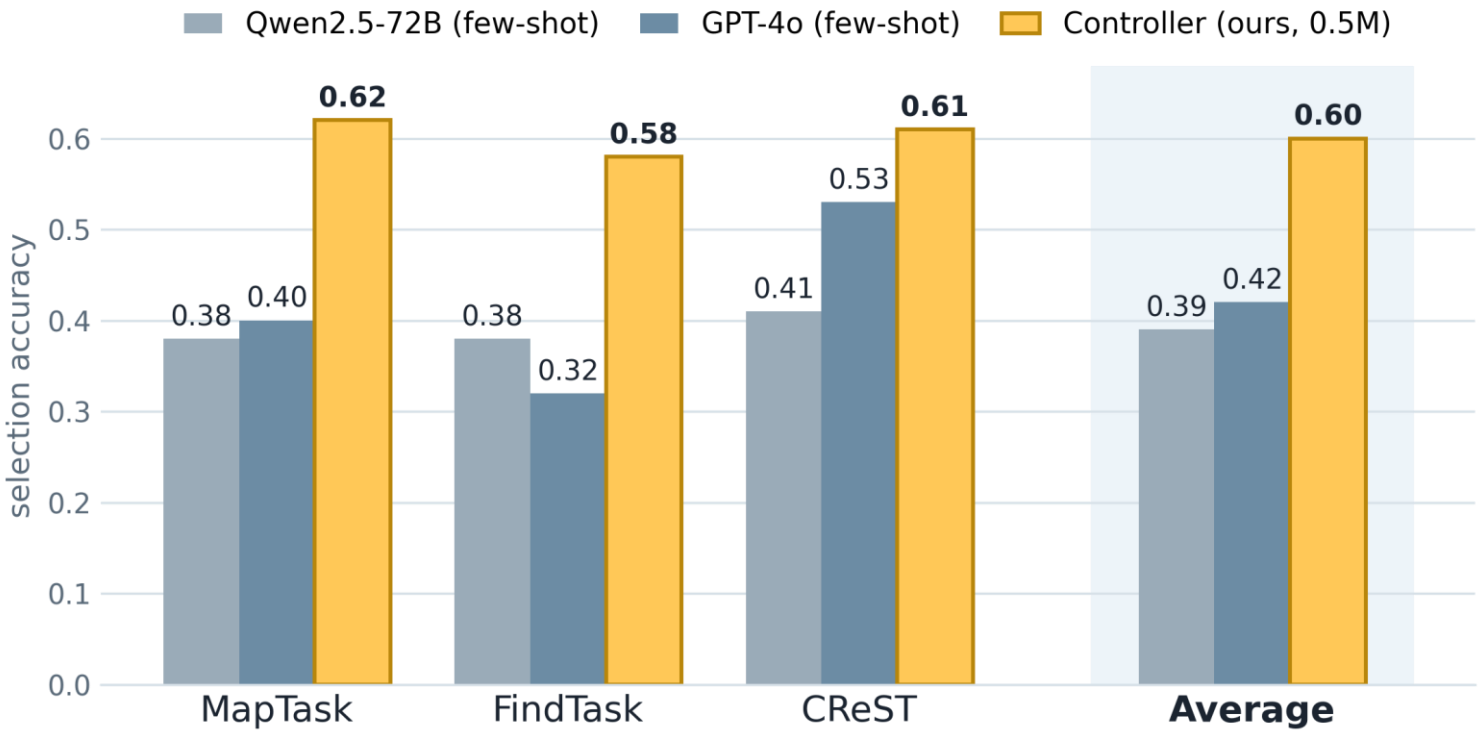
④ GOVERN
controller predictions gate everything
— no gold labels at inference
$\rightarrow$ ***end-to-end***

selection decides which move · realization makes the frozen model perform it

Avsian, Dokme, Woo & Heck, "Latent-IM: Latent Interaction Management for Speech LLMs," arXiv:2607.26928 · AAAI 2027 (in review). Fig. 1.

Selection: reading controller beats prompting

one job: predict the follower's next move — LLMs read the transcript · the controller reads the state



DAVID vs GOLIATH

0.5M controller vs 72B & frontier LLMs — ~10^5× smaller

prompted LLMs see the whole transcript → 0.39–0.42 avg ≈ majority baseline (0.42)

the controller taps what the frozen backbone already computed → 0.60

and it is steady: 0.58–0.62 across datasets · LLMs swing 0.32–0.53

interaction lives in the state, not the prompt — asking the model ≠ reading the model

Avsian, Dokme, Woo & Heck, "Latent-IM: Latent Interaction Management for Speech LLMs," arXiv:2607.26928 · AAAI 2027 (in review). Table 3.

Realization: the direction is enough

oracle setting — every method receives the gold target move · the only question: who can make the frozen model express it

Dataset	Model	Base	SI	PBL	FUDGE	DeAL	Ours
MapTask	**Qwen3**	29.0	30.1	**43.2**	29.9	34.9	41.0
	Qwen2.5	30.9	38.8	42.4	30.7	39.8	**66.8**
	Phi-4	40.6	63.2	56.6	46.7	31.1	**64.8**
FindTask	**Qwen3**	29.5	34.6	46.2	26.9	37.8	**52.6**
	Qwen2.5	23.7	25.6	34.0	31.4	36.5	**55.8**
	Phi-4	29.5	35.3	35.9	25.6	33.3	**50.6**
CReST	**Qwen3**	42.9	48.3	**64.3**	46.7	55.8	63.3
	Qwen2.5	39.8	43.9	62.1	41.4	53.0	**66.8**
	Phi-4	43.6	75.2	62.1	36.7	46.4	**78.1**
	Avg.	**34.4**	**43.9**	**49.6**	**35.1**	**41.0**	**60.0**

bold = row winner · judge: automatic move classifier ≈ human majority (ρ = 0.96 · human κ = 0.682, Table 2)

SUMMARY

60.0% avg — best method by +10.4 pts

wins 7 of 9 dataset × backbone settings

directions estimated on MapTask held-out only → ***applied unchanged to FindTask · CReST***

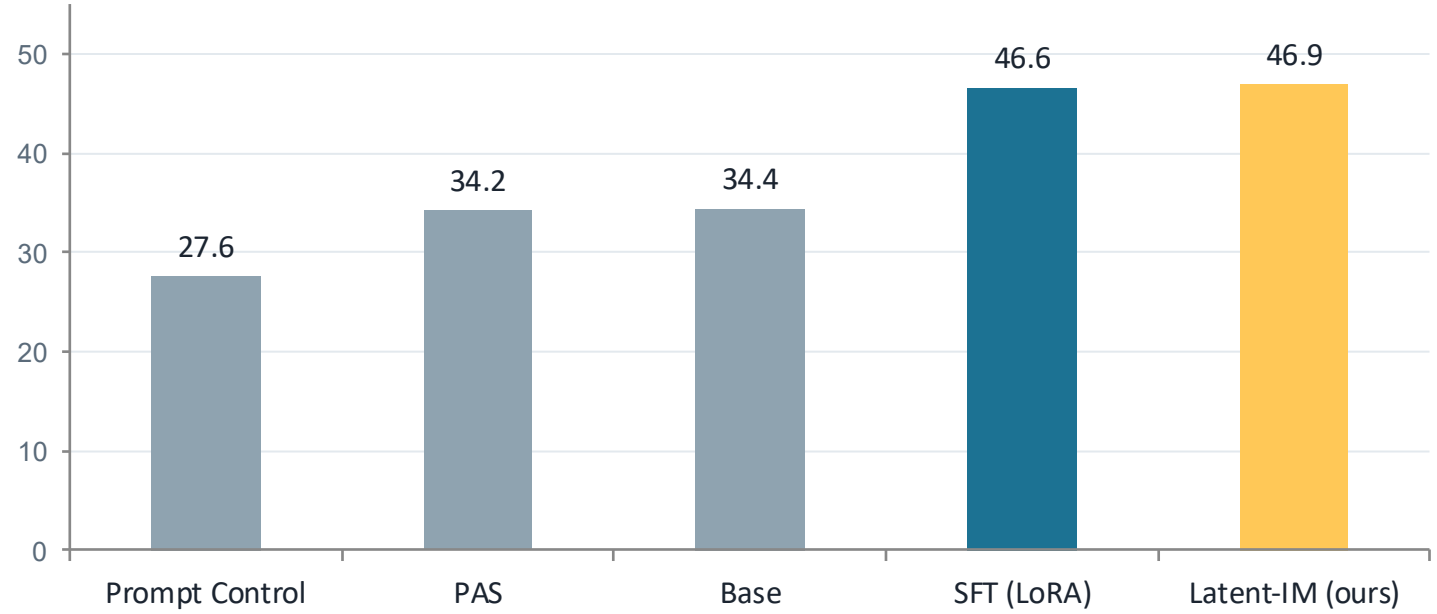
baselines search or re-prompt at decode time — Latent-IM injects one fixed vector

Base = direct prompt · SI = instruction-derived steering · PBL = few-shot candidate selection · FUDGE = classifier-guided decoding · DeAL = reward-guided search

Estimated once, never trained — the move subspace itself carries the behavior.

Avsian, Dokme, Woo & Heck, "Latent-IM: Latent Interaction Management for Speech LLMs," arXiv:2607.26928 · AAAI 2027 (in review). Table 4.
Ramirez et al. 2023 (PBL) · Stolfo et al. 2025 (SI) · Yang & Klein 2021 (FUDGE) · Huang et al. 2025 (DeAL).

Closing the loop — no gold labels at inference



- = SFT accuracy — frozen · label-free
- Gate: steer strong · defer weak moves
- Withhold where steering collapses content

+12.5 vs unsteered backbone · SFT keeps higher BLEU (imitation) — Latent-IM: right move, different words

0.5M trainable parameters vs ≈65M for LoRA SFT — ≈130× fewer · backbone frozen

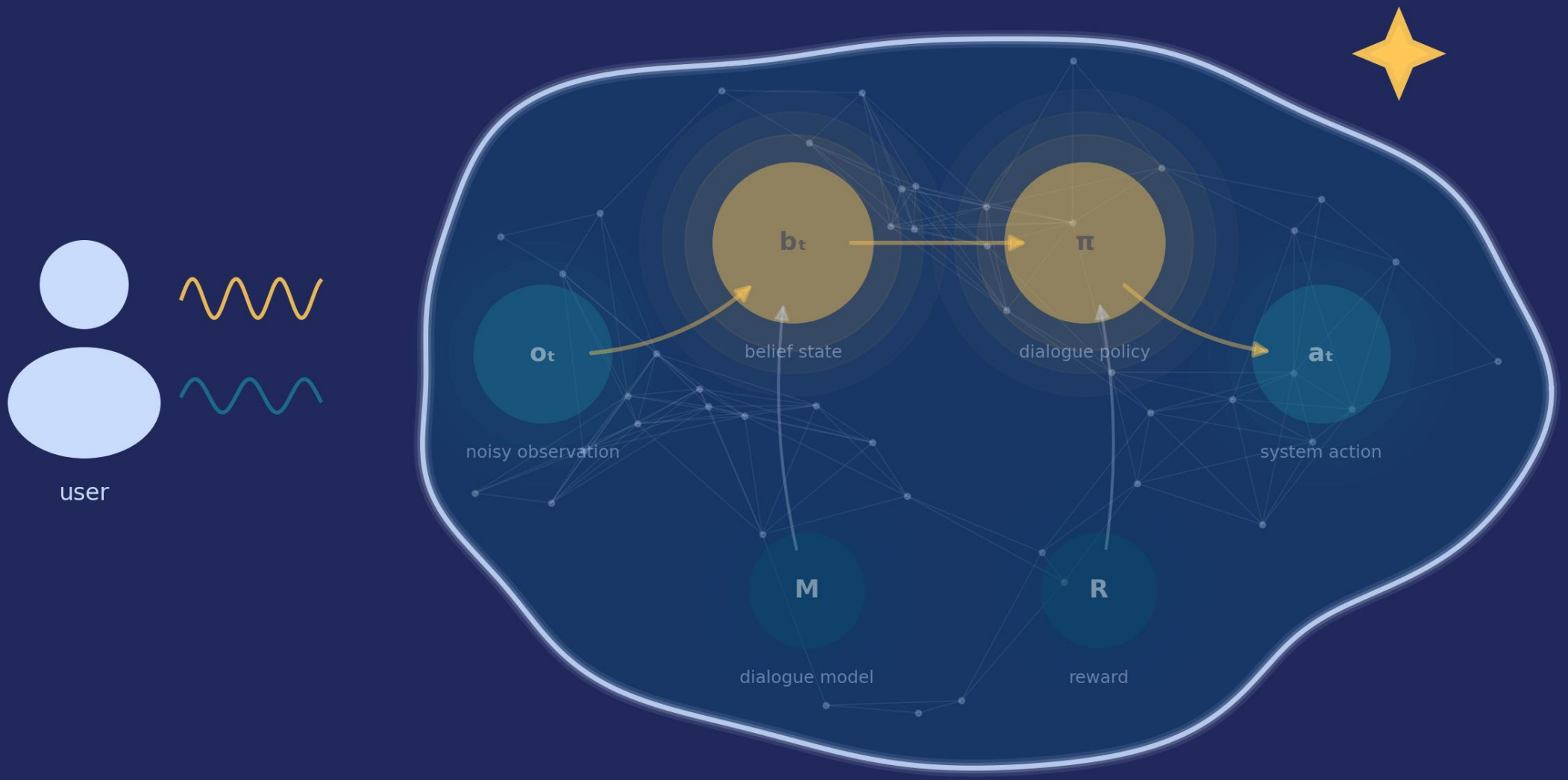
move directions learned once on MapTask → transferred unchanged to FindTask & CReST · **SFT must retrain per backbone × dataset**

cheap to train · frozen to deploy · estimate once, reuse everywhere

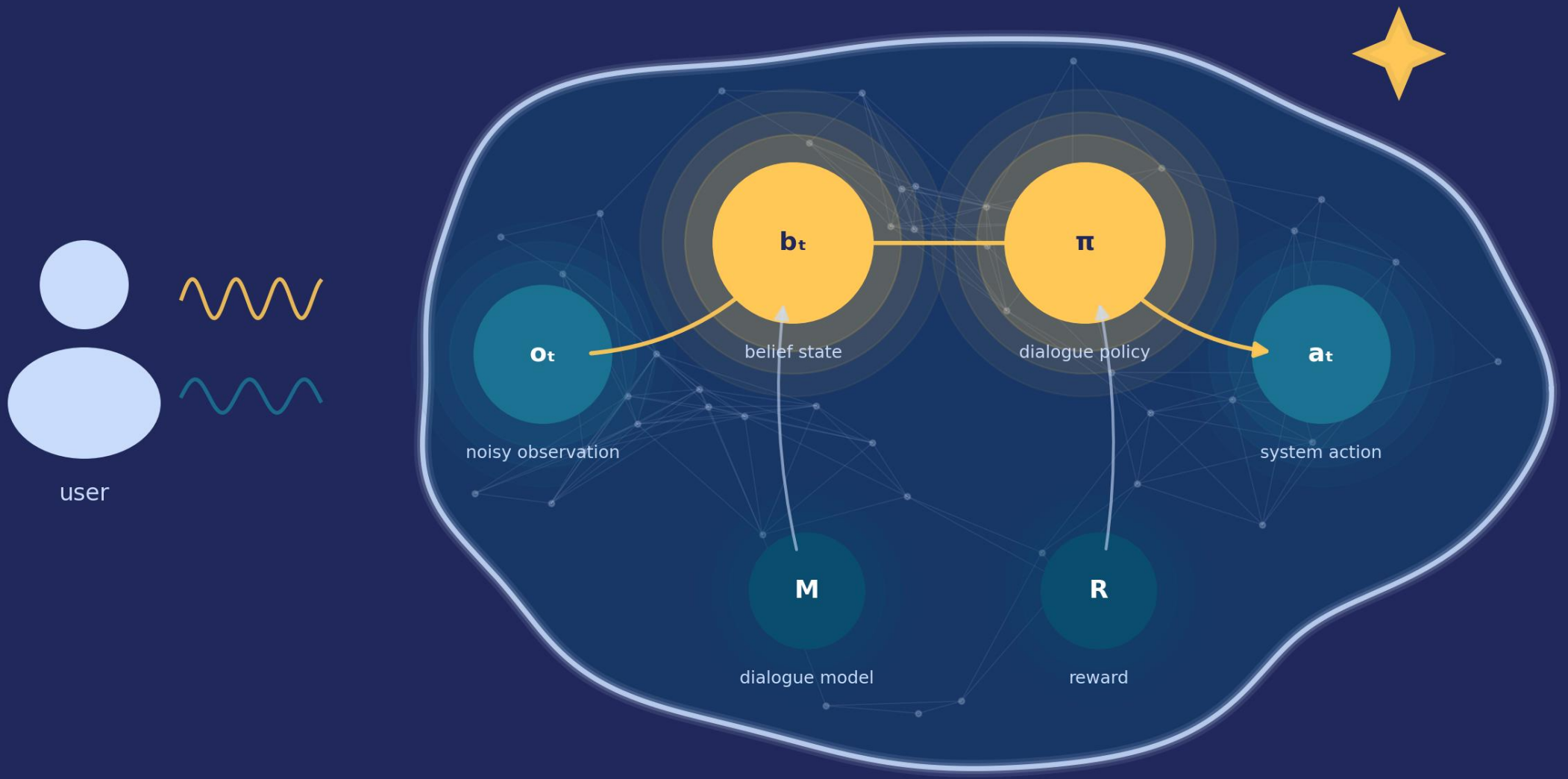
Avsian, Dokme, Woo & Heck, "Latent-IM: Latent Interaction Management for Speech LLMs," arXiv:2607.26928 · AAAI 2027 (in review).

So — where did the diagram go?

So — where did the diagram go?



So — where did the diagram go?



The decomposition was never lost — it just went latent.

state estimation · action selection · realization · turn-taking — recovered as observable, causally controllable latent structure

Where it’s going · digital humans

why digital humans?

Real-time face-to-face conversations with digital humans provides opportunities to study rich set of human interactional cues

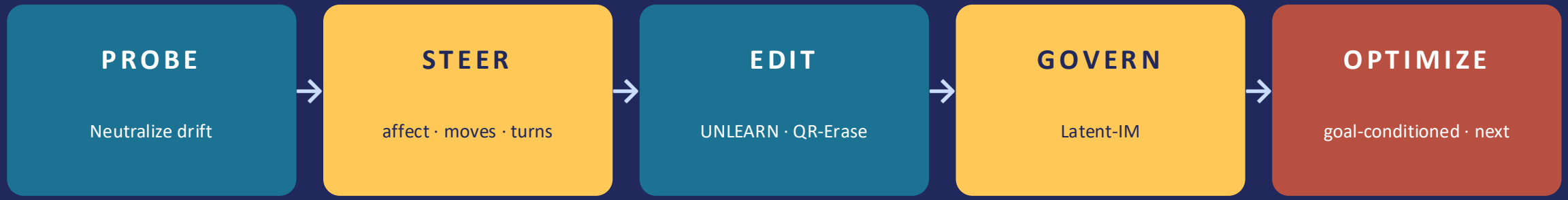
FUTURE WORK

Beyond human-interaction imitation

Goal-conditioned optimization

The Conversation Inside the Model

Probing and Controlling Interactional Structure in Multimodal Foundation Models



around one frozen plant — the foundation model

The conversation was inside the model all along — latent, low-rank, and layer-localized.

Thank you — students & collaborators: A. Avsian · A. Dokme · B. Reichman · T. Woo · T. Lizzo · C. Richardson · A. Sundar · E. Shriberg

Larry Heck, Professor
ECE & Interactive Computing
Rhesa S. Farmer, Jr. Advanced Computing Concepts Chair
Georgia Research Alliance Eminent Scholar