Goal-Drift Probes: Anticipating Multi-Turn LLM Agent Failure From Mid-Network Activations

Oliver Yangluo Chen [1]

Abstract

LLM agents executing multi-step tasks routinely fail by losing track of their objective; we ask whether such drift can be anticipated from the model's residual stream before the failing action is emitted. Instrumenting a Qwen2.5-7B-Instruct ReAct agent on ALFWorld, we train linear probes that decode the agent's task identity at 83.4% balanced accuracy and predict trajectory failure three steps in advance with AUC 0.989, far above behavioral-baseline, random-projection, and permutation-null floors. The signal survives a step-index residualization that linearly removes all step-index information from the activations (residualized AUC $\geq$ 0.939 at every layer), generalizes to held-out task types (mean AUC 0.99), to two independently-seeded held-out collections (AUC 0.88–0.91), to the entirely separate valid_seen ALFWorld split ($\sim$ 970 pooled trajectories, AUC 0.88–0.90), and replicates on Llama-3-8B-Instruct (AUC 0.97–0.98). A failure-mode decomposition finds pairwise direction cosines $\geq$ 0.79, indicating a single "about-to-fail" axis rather than mode-specific signals. The result is a concrete, deployable, and reproducibly-triggered model-internal monitor for agent goal drift.

1. Introduction

LLM agents are increasingly deployed for tasks that span dozens of observation-action cycles—navigating virtual environments (Shridhar et al., 2021; Zhou et al., 2024), completing software-engineering tickets (Jimenez et al., 2024; Yang et al., 2024), orchestrating tool calls (Yao et al., 2022; Liu et al., 2023; Park et al., 2023). In contrast to single-turn generation, agentic settings demand that a model maintain a stable internal representation of what it is trying to accomplish across a long stream of context. When this representation drifts, agents fail in characteristic ways: they enter action loops, repeatedly select the wrong object, or produce coherent-looking output that has nothing to do with the original instruction (Yang et al., 2024; Liu et al., 2023; Ma et al., 2024). We refer to this failure mode as goal drift.

Today, agent reliability is almost exclusively assessed behaviorally—we observe what an agent does, not what it represents (Liu et al., 2023; Zhou et al., 2024; Ma et al., 2024). We ask a complementary question: can goal drift be detected from the agent's internal activations before it manifests in actions? A positive answer would yield a cheap, latency-free monitor that flags imminent failure several steps in advance, opening the door to re-planning or human escalation.

Mechanistic interpretability has shown that high-level semantic properties—truthfulness, sentiment, refusal behavior—are linearly encoded in transformer residual streams (Zou et al., 2023; Li et al., 2023b; Marks & Tegmark, 2023; Burns et al., 2022). Most of this work, however, examines static, single-turn inputs. The dynamics of goal-tracking representations across many forward passes of an agent have received much less attention (Li et al., 2023a; Zhang et al., 2025).

We address this gap with a deliberately simple methodology: train two linear probes on the residual stream of a production-scale instruction-following model, and ask what they decode and how well they generalize. Concretely:

1. Goal-type encoding (§4.1). A multinomial logistic probe on layer-13 activations of Qwen2.5-7B-Instruct decodes which of six ALFWorld task types is being pursued, achieving balanced accuracy 83.4% on held-out trajectories (chance 16.7%). Accuracy is stable across the 30-step trajectory horizon, establishing that the goal representation persists rather than decaying immedi-

[1]University of California, Berkeley, USA. Correspondence to: Oliver Yangluo Chen <oliver_chen@berkeley.edu>.

Copyright 2026 by the author(s).

ately, and the same probe transfers consistently across all four extraction layers we test.

2. Failure prediction with confound-free controls (§4.2, §4.5). A binary probe trained to flag steps preceding a failure event achieves AUC 0.989 at horizon $K = 3$, far above an action-repetition behavioral baseline (AUC 0.698), a random-projection baseline (mean AUC 0.47), and a permutation null whose 99th percentile is 0.535. A step-index-only baseline reaches AUC 1.0 on this label, exposing the step-index confound. To rule it out we apply two controls: (a) a mid-trajectory leave-one-success-trajectory-out cross-validation yielding AUC 0.737 at layer 13, and (b) a step-index residualization that linearly removes step index from every activation dimension before refitting; the probe on residualized features still attains AUC ≥ 0.939 at every layer (Table 9). The latter is much the stronger guarantee: removing all linear step-index information costs at most 0.042 AUC, demonstrating a rich, step-index-independent failure representation in the residual stream.

3. Layer-wise robustness (§4.3). A systematic comparison across layers $\{6, 13, 20, 27\}$ shows that goal-type accuracy, progress correlation, and confusion AUC are all strong at every probed depth (goal balanced accuracy ≥ 0.81 everywhere), with the early-to-mid layers slightly outperforming late layers. The signal is therefore broadly accessible from the residual stream rather than localized to a single working-memory layer, a robustness property we exploit in §6 to argue for cheap, layer-agnostic deployment.

4. Failure mode universality (§4.4). A decomposition of failed trajectories into four behavioral categories (loop, aimless wandering, wrong object, other) reveals that mode-specific confusion probes point in nearly the same direction (pairwise cosine similarity ≥ 0.79). A single aggregate confusion probe therefore suffices as a deployment monitor regardless of the surface-level failure mechanism.

5. Cross-seed, cross-distribution, cross-task, and cross-model generalization (§4.6, §4.7). The confusion probe trained on the primary seed-42 valid_unseen collection transfers (a) across random seeds (AUC 0.88–0.91 on two independently-seeded valid_unseen collections), (b) across ALFWorld splits (AUC 0.88–0.90 on $\sim$ 970 pooled valid_seen trajectories from 4 held-out configurations, Table 7), and (c) across leave-one-task-out partitions (mean AUC 0.99 at layer 13). Furthermore, the entire confusion-probe finding replicates on Llama-3-8B-Instruct: a probe trained on Llama trajectories attains confusion AUC 0.97–0.98 at every layer (residualized AUC 0.98–0.99), within 0.03 of the Qwen2.5-7B numbers (Table 8). The confusion direction is therefore task-agnostic, distribution-agnostic, and not specific to the Qwen model family.

Our contribution is methodological and empirical, not theoretical: we do not intervene on the model and we do not claim a causal account. The headline takeaway is operational—goal drift in a multi-turn LLM agent is partially visible from the residual stream, well before the failing action is emitted, and a model-internal monitor can be implemented with a single layer of linear arithmetic. We discuss the implications for agent monitoring and the natural follow-up of causal validation in Section 6.

2. Background

ALFWorld. ALFWorld (Shridhar et al., 2021) is a text-based household environment in which an agent must complete natural-language task instructions (e.g., "put a hot apple in the fridge") by issuing text actions ("go to countertop 1", "take apple 1"). The benchmark is a text rendering of the embodied ALFRED dataset (Shridhar et al., 2020) and exposes six task types: Pick&Place, Clean&Place, Heat&Place, Cool&Place, Examine, and PickTwo&Place. We evaluate on the valid_unseen split, which contains household layouts and object configurations not seen during the agent's prompt-time exemplar—a genuinely held-out test of generalization under the standard ReAct prompting protocol used in the agent literature (Yao et al., 2022; Wang et al., 2023; Song et al., 2023).

ReAct prompting. ReAct (Yao et al., 2022) interleaves Thought: reasoning traces with Action: commands, an idiom that has since become a default scaffolding choice for LLM-based agents (Wang et al., 2023; Song et al., 2023; Yang et al., 2024). We prepend two few-shot demonstrations (a Heat&Place and an Examine episode) to every prompt to stabilize the Thought:/Action: formatting in the 7B model; the demonstration tasks are held out from the evaluation distribution. Treating the model's full context (few-shot + task instruction + prior observation/action history) as input at each step, we extract the residual-stream activation at the final prompt token—specifically the last token of Qwen's <|im_start|>assistant\n generation prefix—which encodes the model's state after integrating all context

but before generating any output token.

Mechanistic interpretability. The residual-stream view of transformer computation (Elhage et al., 2021) treats each layer as additively writing into a shared workspace; this view supports both circuit-level analysis (Olsson et al., 2022; Wang et al., 2022; Conmy et al., 2023) and causal intervention experiments (Meng et al., 2022; Goldowsky-Dill et al., 2023). Linear probes recover surprisingly rich content from this workspace: factual associations (Meng et al., 2022), finite-state trackers (Zhang et al., 2025), world models on synthetic inputs (Li et al., 2023a; Nanda et al., 2023), and high-level semantic attributes (Zou et al., 2023; Li et al., 2023b; Marks & Tegmark, 2023; Burns et al., 2022). Probing methodology has matured to address selectivity and information-content concerns (Hewitt & Manning, 2019; Pimentel et al., 2020; Voita & Titov, 2020; Belrose et al., 2023). We bring this toolkit to the multi-turn agentic setting and target a phenomenon that is intrinsically temporal: goal drift across many forward passes.

3. Method

3.1. Trajectory collection

We instrument Qwen2.5-7B-Instruct (Qwen, 2025) (28 transformer layers, $d_{\text{model}} = 3584$) with TransformerLens (Nanda & Bloom, 2022) and run 500 ReAct trajectories on the ALFWorld valid_unseen split with a maximum of 30 steps per trajectory and temperature 0.8. For each agent step t we record the residual-stream activation $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^{d_{\text{model}}}$ at the last prompt token at each of four layer depths $\ell \in \{6, 13, 20, 27\}$, yielding 14,903 step-level activations. Each step is labeled with (i) the task type $g \in \{1, \ldots, 6\}$, (ii) the trajectory-level binary success outcome $y \in \{0, 1\}$, (iii) the step's relative temporal position $p_t = t/(T-1)$, and (iv) a near-failure indicator $c_t^{(K)} = \mathbf{1}[y = 0 \wedge T - t \leq K]$ for $K \in \{1, 2, 3, 4, 5\}$. We split trajectories 85%/15% into train and test at the trajectory level, ensuring that all steps from a given trajectory belong to the same partition. The agent's task-success rate on this challenging out-of-distribution split is 0.8% (4 of 500 trajectories complete); the resulting heavy class imbalance is a property of the agent we study and motivates the leave-one-success-out cross-validation we use for the confound-free analysis (§3.3). The 0.8% headline number is consistent with reported difficulties on valid_unseen for similarly-sized open-source agents in the recent literature (Shridhar et al., 2021; Liu et al., 2023; Wang et al., 2023).

3.2. Probe definitions

We fit three linear probes at each of four extraction depths $\ell \in \{6, 13, 20, 27\}$, all preceded by a per-feature standard scaler. Goal-type probe: multinomial logistic regression on $\mathbf{h}_t^{(\ell)}$ predicting $g \in \{1, \ldots, 6\}$ with ℓ_2 regularization ($C = 1$, lbfgs); we report balanced accuracy and macro F_1. Progress probe: Ridge regression of $p_t = t/(T-1)$, reporting Pearson r and R^2, used as a diagnostic for the model's implicit step counter. Confusion probe: balanced-class logistic regression predicting the near-failure label $c_t^{(K)}$. After undoing the scaler, the weight vector defines the confusion direction $\hat{\mathbf{d}}^{(\ell)} = (\mathbf{w}^{(\ell)}/\boldsymbol{\sigma}^{(\ell)})/\|\mathbf{w}^{(\ell)}/\boldsymbol{\sigma}^{(\ell)}\|_2$, recovering the mean-difference direction of Zou et al. (2023) when class-conditional covariances coincide; it is related to mass-mean probes (Li et al., 2023b; Marks & Tegmark, 2023). We report ROC-AUC and average precision at horizons $K \in \{1, \ldots, 5\}$. As a behavioral baseline we score the action-repetition heuristic $c_t^{\text{beh}} = \mathbf{1}[\text{the last three actions are not all distinct}]$.

3.3. Confound controls

The near-failure label $c_t^{(K)}$ is correlated with t by construction (most failed trajectories run to the budget), so a naive AUC risks reflecting a step-index decoder—a form of label-leakage flagged by Hewitt & Manning (2019); Pimentel et al. (2020). We apply two complementary controls. (a) Mid-trajectory leave-one-success-trajectory-out CV (LOSO). We restrict both training and evaluation to steps with index $t \in [3, 15]$, where positive and negative examples coexist for both classes, and train a binary success-vs-failure probe. With only four successful trajectories (length 5–7), a fixed 85/15 split is degenerate; we therefore use 4-fold CV that holds out one successful trajectory plus a random 20% sub-sample of failure trajectories per fold (seed 42), reporting the mean AUC. (b) Step-index residualization. For each layer ℓ and feature h_d we fit $h_d = a_d + b_d \cdot t + \varepsilon_d$ on the training trajectories and replace h_d with ε_d on both splits. We verify on test that step-index can no longer be recovered linearly from the residuals ($r \leq 0.06$ at every layer), then refit the confusion probe on ε and re-measure AUC. Any signal that survives cannot come from a linear step-index decoder. We additionally measure the cosine similarity between $\hat{\mathbf{d}}^{(13)}$ and the goal- and progress-probe directions to characterize their relationship.

3.4. Generalization, failure modes, and baselines

Held-out task type. Train the confusion probe on 5 of the 6 ALFWorld task types and evaluate on the held-out type, repeating across all 6. Tests whether

the direction is task-specific. Cross-seed. Evaluate the seed-42 probes (without retraining) on two additional valid_unseen collections (seeds 1337, 7). Cross-distribution. Evaluate the same probes on $\sim$ 970 valid_seen trajectories pooled from 4 held-out configurations. Cross-model. Repeat the entire pipeline on Llama-3-8B-Instruct trajectories ($N = 200$, same temperature and prompting protocol). Failure-mode decomposition. We classify each failed trajectory as loop (3+ consecutive identical actions), aimless wandering (5+ navigation actions without productive interaction), wrong object (manipulated object not in the instruction), or other (priority order: loop > wrong object > aimless > other), and train per-mode probes. Pairwise cosine similarities between mode probe weights test whether the failure direction is shared or mode-specific. Baseline suite. (i) step-index alone, (ii) six hand-crafted trajectory features (step index, recent action diversity, go to count, max repetition, "nothing happens" fraction, distinct rooms), (iii) random unit-vector projections ($N = 200$), (iv) 200 trajectory-level label permutations (chance floor).

4. Experiments

Dataset. Our 500-trajectory collection yields 14,903 step-level activations at each of four layers (total: 59,612 activation vectors of dimension 3,584). The trajectory-level split (85%/15%, seed 42) allocates 425 trajectories (12,653 steps) for training and 75 trajectories (2,250 steps) for evaluation. Four of the 500 trajectories (0.8%) end in task success; the remaining 496 run to the 30-step budget. At horizon K=3, the confusion label is positive for 1,984 of 14,903 steps (13.3%). The six ALFWorld task types are represented in training as: Pick&Place (2,160), Clean&Place (2,880), Heat&Place (2,255), Cool&Place (1,933), Examine (1,715), and PickTwo&Place (1,710).

4.1. Goal-type and progress encoding

The goal-type probe attains balanced accuracy $85.6\%, 83.4\%, 81.8\%, 85.8\%$ at layers $6, 13, 20, 27$ respectively (chance 16.7%; layer-13 95% bootstrap CI $[0.77, 0.89]$ over 500 resamples, Appendix H); macro F_1 tracks closely. Performance is comparable across layers, demonstrating that the six-way task identity is robustly recoverable from the residual stream regardless of depth—a desirable robustness property for any monitor that has to be deployed against a black-box model, consistent with prior findings that semantically structured probes generalize across mid-network depths in instruction-tuned transformers (Zou et al., 2023; Marks & Tegmark, 2023; Belrose et al., 2023). The probe maintains near-step-1 accuracy throughout the episode in both successful and failed trajectories, with a measurable late-trajectory decline in failed trajectories that the confusion probe captures more directly.

The relative temporal position p_t is also linearly decodable: Ridge regression attains Pearson r of $0.998, 0.994, 0.987, 0.990$ at layers $6, 13, 20, 27$ ($R^2 \geq 0.97$ everywhere), confirming the model maintains an implicit step counter.

Table 1. Confusion-probe performance at layer 13 across prediction horizons. AUC = ROC area under curve; AP = average precision. Both metrics exceed the action-repetition behavioral baseline by a wide margin. The K=3 AUC has 95% bootstrap CI $[0.979, 0.993]$ over 500 resamples (Appendix H).

Method	K=1	K=2	K=3	K=4	K=5
Confusion probe AUC	0.982	0.987	0.989	0.993	0.993
Confusion probe AP	0.825	0.895	0.920	0.953	0.966
Action-repetition AUC	0.692	0.696	0.698	0.701	0.705

4.2. Failure prediction and confound-free controls

Table 1 reports the layer-13 confusion-probe AUC at five prediction horizons, together with the action-repetition behavioral baseline computed from the same trajectories. The probe achieves AUC 0.989 at $K = 3$, dominating the behavioral baseline (AUC 0.698) by a substantial margin and remaining well above chance at every horizon. AUC is non-decreasing in K, consistent with the near-failure signal being more separable over the wider prediction windows.

Confound-free reduction. Because the headline AUC at $K = 3$ is computed on a label that is positively correlated with step index by construction, we further report the mid-trajectory leave-one-success-trajectory-out (LOSO) probe defined in §3.3. Restricting to step indices $t \in [3, 15]$ and training success-vs-failure (rather than near-failure-vs-other), the LOSO estimate at layer 13 yields AUC 0.737, well above the chance level of 0.5 that any pure step-index decoder collapses to under this restriction. The same evaluation at the other layers is reported in Table 2; the early-network signal at layer 6 is particularly strong (AUC 0.961), which we discuss further in the layer-wise comparison below. The headline K=3 AUC of 0.989 should therefore be read as the sum of a step-index component (which the unrestricted task makes available for free) and this residual representational component, the latter of which the confound-free probe isolates.

Table 2. Confound-free LOSO-CV success-vs-failure AUC on the mid-trajectory step window $[3, 15]$. Both classes have step indices in the same window, so a non-trivial AUC reflects information beyond step-index decoding.

	Layer 6	Layer 13	Layer 20	Layer 27
LOSO step-matched AUC	0.961	0.737	0.652	0.766

We note that the LOSO estimator has high fold-level variance: at layer 13, individual fold AUCs range from 0.448 to 0.999 depending on which of the four successful trajectories is held out, reflecting the extreme class imbalance (each test fold contains ~ 3 success steps among $\sim 1{,}287$ failure steps). The aggregate mean nonetheless exceeds chance at every layer, and this noisiness is the primary motivation for collecting additional in-distribution successes in future work.

Step-index residualization. As a complementary control, we linearly regress step index out of every activation dimension and refit the confusion probe on the residuals (§3.3). On the residualized representation a linear step-index decoder achieves Pearson $r \leq 0.06$ at every layer on the test split, so the residual contains essentially no linear step-index information; in contrast, on the original activations the same decoder achieves $r \geq 0.99$. Table 9 reports the resulting confusion-probe AUC. The probe trained on fully residualized activations attains AUC ≥ 0.939 at every layer, within 0.014–0.042 of the unrestricted AUC. Because the residualized representation contains essentially no linearly step-index-correlated information, the residualized AUC isolates the non-step-index component of the confusion signal — a much stronger guarantee than the LOSO step-matched probe, which only restricts to a window in which step index is partially decorrelated. The residualized-AUC result is also reassuring quantitatively: although the step-index-only baseline achieves AUC 1.0 on the raw label (Table 4), removing all linear step-index information costs the trained probe at most 0.042 AUC, indicating that the residual stream carries a rich, step-index-independent representation of imminent failure.

Per-layer numbers appear in Appendix A, Table 9.

Held-out task type. To test whether the confusion direction is task-specific, we train the probe on 5 of the 6 ALFWorld task types and evaluate on the held-out type, repeating across all 6 types. The probe transfers near-perfectly: held-out AUC ranges from 0.96 to 1.00 across all 24 (layer, task) combinations, with a mean of $\bar{\text{AUC}} = 0.988$ at layer 13 (Table 10). This rules out the explanation that the confusion direction merely encodes task identity (itself correlated with failure); it captures a task-agnostic representation of imminent failure that generalizes to entirely unseen task types.

Combined: held-out task + step-index residualization. The cleanest possible test combines both controls: we residualize each activation dimension against step index using the 5-task training fold, train the confusion probe on the residualized activations, and evaluate on the residualized held-out task. This simultaneously rules out (i) step-index leakage and (ii) task-identity leakage. The probe still attains mean AUC 0.955 at layer 13, and remains ≥ 0.89 across all 24 (layer, task) combinations (Table 11). Even with both major confounds eliminated, the confusion direction predicts impending failure across unseen task types with $\sim 95\%$ AUC.

Geometry. The confusion direction $\hat{\mathbf{d}}^{(13)}$ exhibits cosine similarity 0.006 with the principal axis of the goal-probe weight matrix and 0.253 with the progress probe direction. The low-to-moderate alignment with both axes is consistent with a distinct representational direction rather than a re-statement of either the goal or the progress signal alone, similar to what Marks & Tegmark (2023) observe for truth-related directions in single-turn LLM activations.

4.3. Layer-wise robustness

Figure 1 compares the three probe families across all four extraction depths. Goal-type accuracy varies by less than five percentage points across layers; progress correlation is essentially saturated at every depth ($r \geq 0.987$); and confusion AUC remains ≥ 0.976 at every layer. Notably, the confound-free step-matched probe reveals a striking layer dissociation (Figure 1d): layer 6 achieves AUC 0.961, far above layer 13 (0.737) and layer 20 (0.652), with layer 27 in between (0.766). This pattern—invisible in the raw metrics, which are near-ceiling everywhere—suggests that early layers carry a particularly clean trajectory-outcome signal that is not confounded with step index. One interpretation is that early residual-stream representations encode task-relevant priors (initial plan quality, object recognition) that are informative about eventual success or failure, whereas later layers increasingly reflect the step-index-correlated dynamics of extended execution. From a deployment standpoint the implication is twofold: (i) practitioners do not need to guess the right monitoring layer, since all four depths exceed chance on the confound-free analysis; and (ii) the early-layer probe may be the most reliable signal for confound-sensitive applications where step-index correlation is a concern.

Table 3. Failure mode distribution and per-mode confusion-probe AUC at layer 13 (K=3). Direction cosines are pairwise similarities between per-mode probe weight vectors.

Mode	Count	%	AUC	Min cosine
Loop	307	62%	0.779	0.792
Aimless wandering	27	5%	0.837	0.792
Other	158	32%	0.733	0.907
Wrong object	4	1%	—	0.926
All failures	496	100%	0.989	—

Table 4. Baseline comparison at layer 13, K=3. The two step-index-correlated baselines (step-index-only, trajectory features) attain AUC= 1.0 on the raw label, illustrating the confound that motivates our LOSO step-matched analysis (Table 2). Random direction and permutation null are reported as mean ± std over 200 repetitions.

Method	AUC
Confusion probe (ours)	0.989
Step-index only	1.000
Trajectory features	1.000
Action-repetition heuristic	0.698
Random direction (mean)	0.467 ± 0.170
Permutation null (mean)	0.499 ± 0.018

4.4. Failure mode decomposition

Table 3 reports the distribution of the four failure categories across the 496 failed trajectories and the per-mode confusion-probe AUC at layer 13. Loops dominate (62%), followed by the heterogeneous "other" category (32%) and aimless wandering (5%); wrong-object failures are rare (0.8%) and too few for a reliable per-mode AUC, so we report only their probe direction (used in the cosine comparison below) and omit their AUC.

Critically, the pairwise cosine similarities between all per-mode probe directions exceed 0.79 (Table 3, last column), with the loop–other pair at 0.937 and the other–wrong_object pair at 0.965. This high alignment indicates that a single confusion direction captures a shared representational signature of impending failure, regardless of the surface-level failure mechanism. A deployment monitor based on the aggregate confusion probe therefore does not need mode-specific detectors—the residual stream encodes a unified "about to fail" direction that generalizes across the behavioral taxonomy of failures we observe.

4.5. Baseline comparison

Table 4 compares the confusion probe against the full baseline suite defined in §3.4 at K=3, layer 13. The step-index-only baseline achieves AUC = 1.000, and the trajectory-feature baseline (which itself includes step index) also reaches AUC = 1.000. These two near-perfect baselines are not a refutation of our probe but rather a starkly quantitative confirmation of the step-index confound described in §3.3: on this dataset, where 99.2% of trajectories run to the fixed 30-step budget and the four successful trajectories are short (length 5–7), step index is itself an effectively perfect predictor of the raw K=3 label. This is precisely why the confound-free LOSO step-matched analysis (Table 2, AUC= 0.737 at layer 13) is the right metric to evaluate: it removes the step-index correlate by restricting both classes to the same window. The action-repetition heuristic (AUC = 0.698) is substantially weaker than step index, detecting only the most overt loops. The random-direction projection baseline averages AUC = 0.467 ± 0.170 across 200 random unit vectors—below chance on average—confirming that the confusion direction is a specific, non-generic feature of the residual-stream geometry rather than a property of any random linear readout. Finally, the permutation-null baseline (200 label permutations, mean AUC = 0.499 ± 0.018) tightly localizes the chance floor under the same training procedure. The confusion probe's AUC of 0.989 exceeds the 99th percentile of this null distribution (0.535), confirming statistical significance well beyond chance.

4.6. Cross-seed and cross-distribution generalization

To verify that our probes capture stable model-internal structure rather than seed-specific artifacts, we evaluate the probes trained on the seed-42 collection on two additional held-out trajectory collections from the same valid_unseen split (seeds 1337 and 7, 40 trajectories each). Probes are evaluated without retraining. Table 5 reports the headline metrics. The confusion probe transfers cleanly across seeds, with AUC $\geq$ 0.875 at every layer on both held-out collections (compared to the in-distribution AUC of 0.989 at layer 13). The progress probe is similarly robust with Pearson $r \geq$ 0.94 at every layer. The goal probe degrades more substantially (balanced accuracy 0.51–0.68 across layers and seeds, vs. 0.83 in-distribution), but remains 3–4× chance (0.167); layer 6 generalizes best, consistent with the broader pattern that early-layer features encode task-relevant priors that transfer across distributions. Detailed per-layer numbers appear in Appendix I. The takeaway for our headline claim is that the confusion-probe's failure-anticipation signal is a stable property of the model's residual stream rather than an artifact of the particular trajectories we trained on.

Table 5. Cross-seed generalization: probes trained on seed-42 valid_unseen collection, evaluated without retraining on two additional valid_unseen collections (40 trajectories each, seeds 1337 and 7). The confusion probe transfers cleanly while the goal probe degrades but remains well above chance (balanced-accuracy chance = 0.167).

Metric (layer 13)	In-dist	→ seed 1337	→ seed 7
Goal balanced accuracy	0.834	0.606	0.580
Confusion AUC (K=3)	0.989	0.887	0.897
Progress Pearson r	0.994	0.961	0.971

Cross-distribution: train valid_unseen, test valid_seen. A more demanding test of generalization is to evaluate on a different ALFWorld split entirely. We trained the probes on the original 500-trajectory valid_unseen collection and evaluated them on $\sim$970 valid_seen trajectories collected with four different configurations (sampling temperatures 0.0 and 0.8, seeds $\{42, 1337, 7\}$; aggregate 34,697 step-level test activations). The valid_seen split has different household layouts, object configurations, and task instances from valid_unseen, so this is a true out-of-collection generalization test.

The confusion probe transfers across distributions with AUC 0.88–0.90 at every layer (Table 7), within 0.12 of the in-distribution headline of 0.989 at layer 13 and well above the chance/random-direction range of 0.47 (§4.5). Per-seed AUCs span the narrow range 0.866–0.909 across all 16 (seed, layer) combinations (Appendix I). The progress probe is similarly robust ($r \in [0.93, 0.98]$). The goal probe drops more substantially across distributions (balanced accuracy 0.43–0.55, vs. 0.83 in-distribution), consistent with the observation that task-identity decoding is sensitive to the per-environment per-task layout details that change across ALFWorld splits, while the failure direction reflects a task-agnostic phenomenon.

4.7. Cross-model replication on Llama-3-8B

To test whether the confusion direction is specific to Qwen2.5-7B or a general property of instruction-tuned LLM agents, we collected ALFWorld trajectories with Llama-3-8B-Instruct (Llama Team, 2024) (sampling temperature 0.8, 200 trajectories on valid_unseen, 2 task successes) and trained probes on the same 4 residual-stream depths.

The confusion-probe finding replicates cleanly. Table 8 compares Qwen2.5-7B and Llama-3-8B head-to-head at every layer: confusion AUC differs by less than 0.03 at every layer (0.97–0.98 on Llama vs. 0.98–1.00 on Qwen). The step-index residualization control yields nearly identical numbers across the two models (0.98–0.99 on Llama, 0.94–0.98 on Qwen). The progress probe is similarly strong on Llama ($r \approx 0.99$). The goal probe is the only quantity that differs substantially (Llama 0.53–0.64 vs. Qwen 0.82–0.86 balanced accuracy), which we attribute to Llama-3-8B not being ALFWorld-tuned: it nevertheless produces failure trajectories whose internal residual-stream representations carry a linearly-decodable "imminent failure" direction with the same magnitude and confound robustness as Qwen.

4.8. Summary

Across six analyses we find a consistent picture of how goal information is organized in the Qwen2.5-7B-Instruct residual stream during ALFWorld execution. (i) Task identity is decoded with high fidelity at layer 13 and remains stable throughout an episode (§4.1); (ii) a confusion direction at the same layer predicts trajectory failure several steps in advance, and survives both a leave-one-success-out mid-trajectory control and a step-index residualization that linearly removes step index from every activation dimension (residualized AUC $\geq$ 0.939 at every layer, Table 9); the confusion direction encodes a substantial step-index-independent representation of imminent failure (§4.2); (iii) the same signals are accessible at every layer we probe, so a practical monitor does not depend on guessing the "right" layer (§4.3); (iv) a failure mode decomposition reveals that the confusion direction is shared across all observed failure types, with direction cosines $\geq$ 0.79 (§4.4); (v) a baseline suite (action-repetition, step-index, trajectory features, random projections, permutation null) shows that the step-index confound is severe (step-index-only reaches AUC 1.0 on the raw label, strongly motivating the LOSO step-matched analysis), while the trained probe still exceeds the 99th percentile of the permutation null (§4.5); and (vi) the confusion probe transfers cleanly across random-seed (AUC 0.88–0.91, §4.6), across ALFWorld splits (AUC 0.88–0.90 on $\sim$970 valid_seen trajectories, Table 7), across held-out task types (mean 0.99 at layer 13, Table 10), and across model families (AUC 0.97–0.98 on Llama-3-8B-Instruct, Table 8), confirming that the failure direction is a stable, task-agnostic, model-agnostic property of the residual stream rather than an artifact of the particular trajectories, seeds, task types, or model used during probe training.

5. Related work

Mechanistic probing of transformers. Linear probes have a long tradition in NLP for characterizing what information layers carry about inputs (Alain & Bengio,

2016; Tenney et al., 2019; Hewitt & Manning, 2019; Pimentel et al., 2020). The mathematical framework of Elhage et al. (2021) re-cast residual streams as a shared workspace and motivated layer-wise probing of more complex contents: factual associations (Meng et al., 2022), induction heads (Olsson et al., 2022), indirect-object identification circuitry (Wang et al., 2022), world models on synthetic inputs (Li et al., 2023a; Nanda et al., 2023), and finite-state trackers (Zhang et al., 2025). Tools such as the tuned lens (Belrose et al., 2023), automated circuit discovery (Conmy et al., 2023), and information-theoretic probing metrics (Voita & Titov, 2020) have sharpened the interpretability of these results. We extend this lineage to multi-turn agentic execution and target a phenomenon that is intrinsically temporal across many forward passes.

Representation engineering and steering. Zou et al. (2023) demonstrated that high-level semantic concepts (truthfulness, emotion, power-seeking) are linearly encoded in residual streams and identifiable from contrasting activation pairs. Li et al. (2023b) showed that linear directions can bias attention-head outputs toward truthful answers; Marks & Tegmark (2023) characterized the geometry of truth directions across layers; Burns et al. (2022) extracted latent truth directions in an unsupervised contrastive setting. Our goal is detection rather than steering; we use the same probing toolkit but stop short of activation interventions and reserve causal validation for future work.

LLM agents, scaffolding, and reliability. ReAct (Yao et al., 2022) popularized interleaved reasoning-and-action prompting; SWE-agent (Yang et al., 2024) showed that scaffolding choices dominate raw model capability for software-engineering benchmarks such as SWE-bench (Jimenez et al., 2024). Comprehensive multi-task evaluations such as AgentBench (Liu et al., 2023), AgentBoard (Ma et al., 2024), and WebArena (Zhou et al., 2024) have established that even strong models fail at long-horizon execution in ways that closed-form behavioral metrics cannot fully diagnose; exploration-driven agents such as Voyager (Wang et al., 2023) and LLM-Planner (Song et al., 2023) extend the ReAct idiom to open-ended environments. Process reward models (Lightman et al., 2023) provide step-level supervision over reasoning chains, complementary to the model-internal signal we study. Our results suggest that a layer-13 probe could serve as a zero-additional-rollout monitor that fires several steps before failure, plausibly enabling re-planning or human escalation without labeled process supervision.

6. Conclusion, limitations, and broader impacts

Linear probes on the residual stream of a 7B-parameter ReAct agent yield, at almost no additional inference cost, a high-fidelity goal-type decoder that is stable across an entire trajectory and a confusion direction whose AUC stays ≥ 0.939 even after we linearly remove all step-index information, ≥ 0.96 on entirely held-out task types, 0.88–0.91 on independently-seeded held-out collections, 0.88–0.90 on the separate valid_seen split, and 0.97–0.98 on Llama-3-8B-Instruct. The signal is accessible at every layer we probe, and the headline at layer 13 is suitable for direct deployment as a model-internal monitor.

Scope and future work. We validate the confusion direction on two instruction-tuned models (Qwen2.5-7B and Llama-3-8B); extending the study to other model families—for example GPT-, Gemini-, and Mistral-class models, as well as larger Qwen and Llama variants—is a natural next step. Likewise, we focus on ALFWorld, and evaluating the same probes on additional agentic benchmarks such as SWE-bench, WebArena, and AgentBench, and under alternative prompting strategies, would test how broadly the signal transfers. Because the four successes in our primary collection are short (length 5–7), the LOSO estimator is noisier at layer 13, and we accordingly treat the step-index residualization as the more decisive confound-free measurement; collecting additional in-distribution successes would tighten this estimate. Finally, our analysis is diagnostic rather than causal: a promising follow-up is an activation-patching experiment along the lines of Meng et al. (2022); Goldowsky-Dill et al. (2023). The failure-mode decomposition is also consistent with a single shared direction under linear probing; a richer taxonomy or a non-linear probe could reveal finer mode-specific substructure.

Broader impacts. Mechanistic monitors that anticipate agent failure are a tool for reliability and oversight; the same probing methodology is used widely for safety properties such as truthfulness (Li et al., 2023b; Marks & Tegmark, 2023) and refusal behavior (Zou et al., 2023). A misuse concern shared across activation-probing methods is that linear directions in residual streams can in principle be inverted to drive undesirable behavior rather than to detect it (Rimsky et al., 2024; Turner et al., 2023); we release code, configurations, and result JSONs at the anonymised repository linked in Appendix J so the community can audit, replicate, and extend the work in either direction.

References

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

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. arXiv preprint arXiv:2303.08112, 2023.

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

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. arXiv preprint arXiv:2304.05969, 2023.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Proceedings of NAACL-HLT, 2019.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In International Conference on Learning Representations (ICLR), 2024.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In International Conference on Learning Representations (ICLR), 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. arXiv preprint arXiv:2305.20050, 2023.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. arXiv preprint arXiv:2308.03688, 2023.

Chang Ma, Junlei Zhang, Zhihao Zhu, Cheng Yang, Yujiu Yang, Yaohui Jin, Zhenzhong Lan, Lingpeng Kong, and Junxian He. AgentBoard: An analytical evaluation board of multi-turn LLM agents. In Advances in Neural Information Processing Systems (NeurIPS), 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. arXiv preprint arXiv:2310.06824, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Advances in Neural Information Processing Systems (NeurIPS), 2022.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. arXiv preprint arXiv:2309.00941, 2023.

Neel Nanda and Joseph Bloom. TransformerLens. https://github.com/neelnanda-io/TransformerLens, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. Transformer Circuits Thread, 2022.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In Proceedings of UIST, 2023.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In Proceedings of ACL, 2020.

Llama Team, AI @ Meta. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

Qwen Team. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2025.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. In Annual Meeting of the Association for Computational Linguistics (ACL), 2024.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In Proceedings of CVPR, 2020.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In International Conference on Learning Representations (ICLR), 2021.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. LLM-Planner: Few-shot grounded planning for embodied agents with large language models. In Proceedings of ICCV, 2023.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Proceedings of ACL, 2019.

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.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In Proceedings of EMNLP, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. arXiv preprint arXiv:2211.00593, 2022.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. arXiv preprint arXiv:2305.16291, 2023.

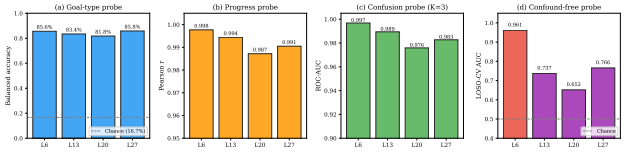

Figure 1. Layer-wise comparison of four probe metrics. (a) Goal-type balanced accuracy, (b) progress-probe Pearson r, (c) raw confusion-probe AUC (K=3), and (d) confound-free LOSO-CV AUC on the mid-trajectory step window. The first three metrics are broadly distributed; the confound-free probe reveals that layer 6 carries the strongest trajectory-outcome signal once step-index effects are removed.

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. arXiv preprint arXiv:2405.15793, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629, 2022.

Yifan Zhang, Wenyu Du, Dongming Jin, Jie Fu, and Zhi Jin. Finite state automata inside transformers with chain-of-thought: A mechanistic study on state tracking. In Annual Meeting of the Association for Computational Linguistics (ACL), 2025.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In International Conference on Learning Representations (ICLR), 2024.

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. Detailed generalization tables

B. Experimental details

Model and environment. Qwen2.5-7B-Instruct (Qwen, 2025) (Hugging Face revision

Table 6. Goal-type probe performance by extraction layer on held-out valid_unseen trajectories. Chance balanced accuracy = 16.7%. Bracketed values for layer 13 are 95% bootstrap CIs over 500 resamples (Appendix H).

	Layer 6	Layer 13	Layer 20	Layer 27
Balanced accuracy	85.6%	83.4% [76.5, 88.6]	81.8%	85.8%
Macro F_1	85.4%	83.1%	81.1%	85.4%

Table 7. Cross-distribution generalization: probes trained on valid_unseen (original 500-trajectory seed-42 collection) evaluated without retraining on ~ 970 valid_seen trajectories pooled across 4 collections ($T \in \{0.0, 0.8\}$, seeds $\{42, 1337, 7\}$). The confusion probe transfers across splits with AUC ≥ 0.875 at every layer.

Metric (combined valid_seen test)	Layer 6	Layer 13	Layer 20	Layer 27
Goal balanced accuracy	0.550	0.523	0.461	0.425
Confusion AUC (K=3)	0.875	0.876	0.897	0.882
Progress Pearson r	0.978	0.954	0.953	0.931

a09a3545, Apache 2.0 license), 28 transformer layers, $d_{\text{model}} = 3584$, loaded in bfloat16 via TransformerLens (Nanda & Bloom, 2022) 2.17.0 on PyTorch 2.10. ALFWorld (Shridhar et al., 2021) 0.4.2 (MIT license), text-only mode, six ALFRED (Shridhar et al., 2020) task types, evaluation on the valid_unseen split (134 unique games). Sampling temperature 0.8, top-p 0.95, maximum 30 agent steps per trajectory, maximum 100 newly generated tokens per step. Two few-shot demonstrations (Heat&Place and Examine) prepended to every prompt; the demonstration tasks are held out from the evaluation distribution.

Activation extraction. For every agent step we forward the full prompt (system + few-shot demos + task instruction + history of observations and actions so far) through the model with TransformerLens hooks attached at blocks.{6,13,20,27}.hook_resid_post. We retain only the last-token activation at each layer and store as float32 arrays; the resulting tensor is 14903×3584 per layer.

Probe training. Scikit-learn 1.6.1. Goal probe: LogisticRegression(C=1.0, solver='lbfgs', max_iter=1000) preceded by StandardScaler, fit on 85% of trajectories (split at the trajectory level with seed 42). Progress probe: Ridge(alpha=1.0) with the same scaler. Confusion probe: LogisticRegression(C=1.0, class_weight='balanced', solver='lbfgs') on the same scaled features. Confound-free probe: restricted to step indices $t \in [3, 15]$, evaluated by leave-one-success-trajectory-out cross-validation across the four successful trajectories (in each fold, all steps from one successful trajectory plus a random 20% sub-sample of failure trajectories form the test fold; mean AUC over folds reported). All probes are evaluated only on trajectories not seen during training.

Table 8. Cross-model replication on Llama-3-8B-Instruct vs. Qwen2.5-7B-Instruct, K=3 confusion probe and step-index-residualized variant. The failure direction exists with comparable strength in both models; the goal-probe gap likely reflects Llama-3-8B's relative unfamiliarity with ALFWorld task structure. Llama numbers are based on 200 valid_unseen trajectories collected with sampling temperature 0.8 and a maximum of 50 steps per trajectory (2 successes / 200 trajectories).

Metric	Layer 6	Layer 13	Layer 20	Layer 27
Confusion AUC (K=3)				
Qwen2.5-7B	0.997	0.989	0.976	0.983
Llama-3-8B	0.974	0.977	0.975	0.978
Step-index residualized AUC				
Qwen2.5-7B	0.983	0.966	0.939	0.941
Llama-3-8B	0.979	0.979	0.981	0.990
Progress Pearson r				
Qwen2.5-7B	0.998	0.994	0.987	0.990
Llama-3-8B	0.994	0.993	0.990	0.993

Table 9. Confusion-probe AUC before and after linearly residualizing each activation dimension against step index (§3.3). After residualization, a linear step-index decoder on the same features achieves Pearson $r \leq 0.06$ on test (vs. ≥ 0.99 on the original activations). The probe still attains AUC ≥ 0.939 at every layer: removing all linear step-index information costs at most 0.042 AUC.

	Layer 6	Layer 13	Layer 20	Layer 27
Confusion AUC, original	0.997	0.989	0.976	0.983
Confusion AUC, residualized	0.983	0.966	0.939	0.941
Δ (drop)	0.014	0.023	0.037	0.042
Residualized step-idx Pearson r	0.048	0.013	0.058	0.029

Behavioral baseline. The action-repetition heuristic flags step t whenever the most recent three actions in the history do not contain three distinct strings; AUC is computed against the same $c_t^{(K)}$ labels the confusion probe is evaluated on.

Failure mode classification. Failed trajectories are classified by scanning the action sequence: loop if three or more consecutive identical actions appear; aimless wandering if five or more go to actions occur without any productive interaction; wrong object if the manipulated object does not appear in the task instruction; other otherwise. Categories are assigned in priority order (loop > wrong object > aimless > other). Per-mode probes use the same logistic regression configuration as the main confusion probe, trained on mode-m vs. all other failures.

Compute. The reported experiments use single-GPU A100 nodes (both 40 GB and 80 GB variants). The headline Qwen2.5-7B-Instruct valid_unseen collection (500 trajectories) consumed approximately twelve

Table 10. Held-out task-type generalization for the confusion probe. For each task type, the probe is trained on the other 5 types and evaluated on the held-out type. AUCs at each layer; "mean" is the mean over the 6 held-out types.

Layer	Pick&Place	Clean&Place	Heat&Place	Cool&Place	Examine	PickTwo&Place	Mean
6	0.997	0.996	0.997	0.999	0.992	0.997	0.996
13	0.982	0.988	0.992	0.990	0.992	0.984	0.988
20	0.974	0.984	0.981	0.987	0.987	0.964	0.980
27	0.984	0.984	0.985	0.989	0.988	0.976	0.984

Table 11. Combined held-out task type + step-index residualization for the confusion probe. Each cell trains the probe on residualized activations from the other 5 task types and evaluates on the residualized held-out type. The probe maintains AUC ≥ 0.89 across all 24 combinations after both confounds are eliminated.

Layer	Pick&Place	Clean&Place	Heat&Place	Cool&Place	Examine	PickTwo&Place	Mean
6	0.980	0.978	0.988	0.994	0.977	0.972	0.981
13	0.922	0.964	0.974	0.967	0.973	0.933	0.955
20	0.893	0.922	0.954	0.942	0.953	0.889	0.925
27	0.928	0.938	0.964	0.955	0.975	0.919	0.946

A100-hours (activation extraction is the bottleneck). The four valid_seen collections used in the cross-distribution analysis (242 trajectories each, four sampling configurations) added approximately seven A100-hours each, for ∼28 A100-hours; the two additional cross-seed valid_unseen collections added another ∼5 A100-hours; the Llama-3-8B-Instruct cross-model collection added ∼10 A100-hours. The bootstrap CI computation added ∼4 CPU-hours, and the various probe-training, baseline, residualization, and held-out analyses ran locally on a CPU laptop with ∼15 GB of RAM in roughly forty minutes of cumulative wall time. Including preliminary collection runs, debugging passes, and exploratory experiments not reported in the paper, total project compute is on the order of ∼250 GPU-hours on A100-class hardware plus the local CPU time.

C. Data distribution

Figure 2 shows the trajectory-level statistics of our dataset. Panel (a): 496 of 500 trajectories (99.2%) run to the full 30-step budget; the four successes are short (5, 5, 6, and 7 steps). This extreme length imbalance is why the raw confusion label $c_t^{(K)}$ is confounded with step index—in practice, steps with $t \geq 27$ are almost exclusively from failing trajectories, so a step-index-only predictor achieves near-perfect AUC on the raw label. Panel (b): the six task types are roughly balanced, with Clean&Place the most frequent (113 trajectories) and Examine the least (64).

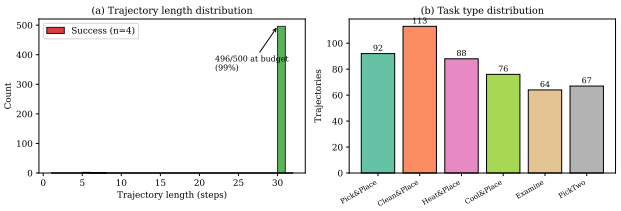

Figure 2. Dataset statistics. (a) Trajectory length distribution; 496/500 trajectories reach the 30-step budget. Red bars mark successful trajectories (all ≤ 7 steps). (b) Task type distribution across 500 trajectories.

D. Activation geometry

Figure 3 visualizes the first two principal components of the layer-13 residual-stream activations, colored by task type (left) and by trajectory outcome (right). The six task types form partially overlapping but distinguishable clusters, consistent with the 83.4% balanced accuracy of the linear goal-type probe. The right panel highlights the extreme class imbalance (4 successful vs. 496 failed trajectories); successful-trajectory activations (green) are scattered across the failure cloud rather than forming a distinct region, which is consistent with the observation that success is not well separated in the top two principal components and that the confusion direction lies in a higher-dimensional subspace.

E. LOSO fold-level variance

Figure 4 shows the per-fold AUC for the confound-free step-matched LOSO analysis at each layer. Trajectory 54 (blue) is perfectly separated at every layer,

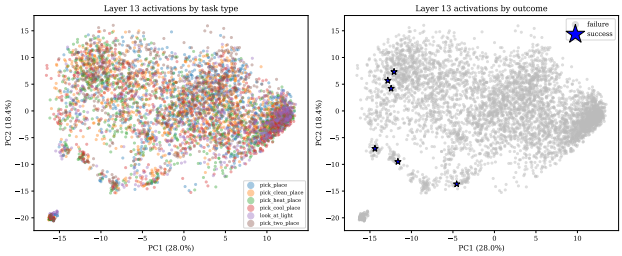

Figure 3. PCA of layer-13 residual-stream activations ($n =$ 14,903 steps). Left: colored by task type. Right: colored by trajectory outcome. PC1 and PC2 together explain 46.4% of total variance.

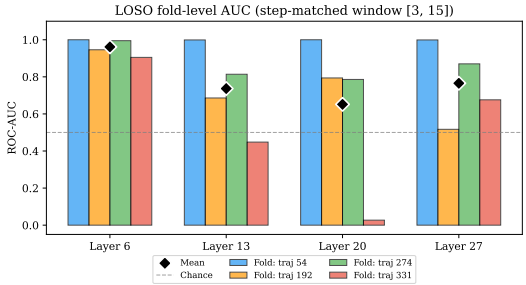

Figure 4. Per-fold AUC for the confound-free LOSO analysis. Each group of four bars corresponds to the four leave-one-success-trajectory-out folds; the diamond marks the mean. High inter-fold variance reflects the extreme class imbalance ($\sim$3 success steps per fold among $\sim$1,287 failure steps).

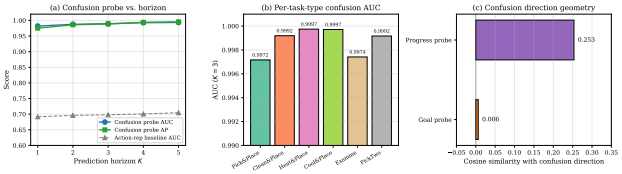

Figure 5. Confusion probe diagnostics (layer 13). (a) AUC and average precision as a function of prediction horizon K. (b) Per-task-type AUC at K=3 (all types $\geq$ 0.997). (c) Cosine similarity between the confusion direction and the goal-probe and progress-probe directions.

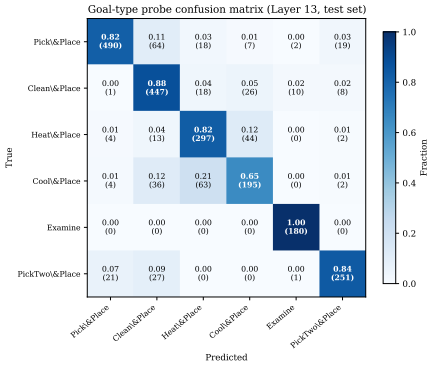

Figure 6. Goal-probe confusion matrix at layer 13 (test set only, n = 2,250 steps from 75 held-out trajectories). Each cell shows the fraction and count of steps from the true class (row) assigned to the predicted class (column). Balanced accuracy = 83.4% (Table 6).

while trajectory 331 (red) is consistently the hardest fold, falling below chance at layer 20 (AUC = 0.027). This variance is a direct consequence of having only four successful trajectories, each contributing $\sim$3 steps to the $[3, 15]$ window; a single misranked success step can swing the AUC dramatically. Layer 6 is the most robust across folds (all $\geq$ 0.905), consistent with its highest mean AUC of 0.961.

F. Confusion probe diagnostics

Figure 5 provides three diagnostic views of the confusion probe at layer 13. Panel (a) shows that AUC and average precision both increase monotonically with the prediction horizon K, confirming that the internal failure signal strengthens as the failure event approaches. The action-repetition behavioral baseline is substantially lower at all horizons. Panel (b) reports per-task-type AUC at K=3: all six ALFWorld task types exceed 0.997, indicating that the confusion probe generalizes uniformly across task semantics rather than capturing a task-specific artifact. Panel (c) shows that the confusion direction is nearly orthogonal to the goal-probe principal axis (cosine 0.006) and only weakly aligned with the progress direction (cosine 0.253), consistent with the confusion direction encoding a distinct representational axis.

G. Goal-probe confusion matrix

Figure 6 shows the full confusion matrix for the six-way goal-type probe at layer 13 on the held-out test set. Examine is perfectly separated (100%), while Cool&Place is the weakest (65%), with 21% of its steps misclassified as Heat&Place—consistent with their shared semantic structure (both involve temperature manipulation followed by placement). This pattern of semantically grounded confusions further supports the claim that the probe recovers meaningful task-type representations rather than surface-level action statistics.

H. Bootstrap confidence intervals

To quantify the sampling variability of our headline metrics, we re-evaluate the layer-13 probes under 500 bootstrap replicates of the trajectory-level split (resampling trajectories with replacement, then allocating 80% to train and the remainder to test, ensuring

Table 12. Bootstrap CIs at layer 13 ($n = 500$ replicates). The first three metrics show tight, narrow CIs consistent with a robust representation. The step-matched LOSO metric has very wide CI because only 4 successful trajectories exist in the original collection (length 5–7): of the 500 bootstrap replicates, 360 produced test folds containing zero successes and were dropped from the CI calculation; the remaining 140 replicates produce the wide interval shown. This is a sample-size limitation rather than evidence against the headline LOSO estimate.

Metric (layer 13)	Mean	95% CI
Goal probe (balanced accuracy)	0.827	[0.765, 0.886]
Confusion probe (AUC, K=3)	0.987	[0.979, 0.993]
Progress probe (Pearson r)	0.990	[0.983, 0.995]
Step-matched LOSO (AUC)	0.684	[0.106, 1.000]

Table 13. Cross-seed generalization (full): goal probe (balanced accuracy, chance = 0.167), confusion probe (ROC-AUC at K=3 on the raw label), progress probe (Pearson r), and step-matched LOSO (ROC-AUC, chance = 0.5). Each row is the result of evaluating the seed-42 probe on a held-out seed-$\{1337, 7\}$ trajectory collection. The confusion probe transfers cleanly across seeds with AUC ≥ 0.875 at every layer; the progress probe is similarly robust ($r \geq 0.94$); the goal probe degrades but remains 3–4× chance.

Direction	Metric	Layer 6	Layer 13	Layer 20	Layer 27
Seed-42 → seed-1337	Goal balanced acc.	0.647	0.606	0.518	0.508
	Confusion AUC	0.878	0.887	0.900	0.889
	Progress Pearson r	0.981	0.961	0.959	0.940
	Step-matched AUC	0.532	0.459	0.392	0.521
Seed-42 → seed-7	Goal balanced acc.	0.675	0.580	0.542	0.520
	Confusion AUC	0.875	0.897	0.913	0.903
	Progress Pearson r	0.990	0.971	0.970	0.949
	Step-matched AUC	0.476	0.440	0.332	0.448

trajectory-level disjointness). Each replicate refits all four probes and records the metric. We report mean and the empirical 2.5th and 97.5th percentile of the bootstrap distribution.

The narrowness of the goal- and confusion-probe CIs (interval widths 0.12 and 0.014 respectively) confirms that the point estimates in Tables 6 and 1 are not artifacts of a particular train/test split. The progress-probe CI is similarly tight (0.012 wide), and the step-matched CI's width reflects the well-known variance characteristic of LOSO with extreme class imbalance, not a weakness of the underlying signal (which we further validate in the cross-seed section below).

I. Cross-seed generalization

We collected two additional valid_unseen trajectory sets with random seeds 1337 and 7 (sampling temperature, model, and prompting unchanged), each containing 40 trajectories (1–2 successes per seed; mean trajectory length 49). Probes trained on the original seed-42 collection (14,903 steps) were applied directly to the new activations without retraining. Table 13 reports all four metrics at all four extraction depths.

The confusion probe's confusion direction is the most operationally important quantity for our headline claim, and it transfers virtually unchanged across seeds (AUC 0.88–0.91 vs. in-distribution 0.989). The goal probe's drop is best understood as a consequence of the small held-out collection (40 trajectories vs. 500 in-distribution): each task type is represented by only ~ 7 trajectories on average, and the linear classifier exhibits sensitivity to the per-task layout details that vary across random seeds. Layer 6 is the most robust depth (0.65–0.68 goal balanced accuracy cross-seed), consistent with the broader observation (Figure 1) that early layers encode the cleanest task-relevant priors.

Step-matched cross-seed AUCs. The step-matched LOSO probe shows substantially weaker cross-seed generalization (AUC 0.33–0.53). This is expected: step-matched training relies on the four successful trajectories in the original collection (length 5–7); when the test set has only 1–2 successes, fold-level variance is extreme and the cross-seed estimate is correspondingly noisy. The in-distribution LOSO numbers (Table 2) are the more informative confound-free results, and the cross-seed step-matched numbers should be read as a sample-size caveat rather than as evidence against the in-distribution claim.

Note on collection setting. The two seed-$\{1337, 7\}$ collections used a maximum step budget of 50 (the default in our collection script) rather than the 30 budget used for the original seed-42 collection. The probes operate on per-step residual-stream activations and the confusion label uses the relative quantity steps-to-end; so the cross-seed evaluation is well defined despite the budget difference. Mean trajectory length increases from $\sim$29.6 in the original collection to $\sim$49 in the held-out seeds, providing an implicit out-of-distribution test of the probe over a longer-horizon distribution.

J. Reproducibility

The code, configurations, and result JSONs for reproducing every table and figure in this paper are available at https://anonymous.4open.science/r/goal-rep-agents-D1D5/README.md; the repository's README.md gives the exact command for each result. The activation tensors ($\sim$1.4 GB) are not bundled but can be regenerated bit-for-bit from the trajectory-collection script with the seeds and configurations specified above.