NOT ALL ERRORS ARE EQUALLY STEERABLE: PREDICTING ERROR TYPES TO GUIDE ACTIVATION STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering aims to correct language-model errors by modifying internal activations at inference time, but different errors may not be equally amenable to the same intervention. We ask whether layer-wise token trajectories can reveal which errors are most likely to benefit from steering. Tracing the correct-answer token across layers reveals two recurring patterns. In a *no-entry* error, the correct token never becomes a plausible output candidate; in a *competitive-decay* (CD) error, it becomes competitive at an intermediate layer but loses its advantage before the final output. Under the same fixed steering method, CD errors are recovered approximately four times as often as no-entry errors. Motivated by this gap, we introduce *Information Dynamics*, a framework for characterizing these trajectories and predicting error type. Although the taxonomy is defined using the reference answer, error type can be predicted from the model's own early candidate dynamics without access to the reference answer at inference time. Across six open-weight models and eight tasks, these dynamics consistently distinguish CD from no-entry errors. We use these predictions to allocate a fixed steering budget toward more recoverable errors. Across seven tasks on Qwen3-8B and Llama-3.1-8B, steering only 20% of errors selected by predicted CD probability corrects more than twice as many errors as random selection under the same intervention budget. These results show that layer-wise token dynamics provide a practical signal of error steerability and can substantially improve activation steering.

1 INTRODUCTION

Activation steering modifies a language model's internal activations at inference time to improve reliability or control behavior (Li et al., 2023; Rimsky et al., 2024; Zou et al., 2025). Most work focuses on how to construct an effective intervention, while recent methods also adapt when or how strongly to intervene (Lee et al., 2025; Hedström et al., 2025; Ferrando et al., 2026). We study a complementary question: *among errors, which ones are actually amenable to the same intervention?* In particular, we ask whether an error's layer-wise internal trajectory can reveal its *steerability* and help allocate a limited intervention budget. We study this question by tracing the correct-answer token, or *reference token*, across model layers using intermediate decoding methods such as the logit lens (nostalgebraist, 2020) and tuned lens (Belrose et al., 2025). We observe two recurring trajectories. In a *no-entry* error, the reference token never becomes a plausible output candidate. In a *competitive-decay* (CD) error, it becomes competitive at an intermediate layer but loses its advantage before the final output. This distinction strongly predicts steerability: under the same fixed activation-steering method, CD errors are recovered approximately four times as often as no-entry errors on held-out Qwen and Llama examples. Reliably identifying these trajectories, however, requires determining whether the reference token genuinely enters output competition. Simply checking whether it ever outscores the final incorrect token is insufficient: a token may win such a pairwise comparison while remaining far below many other vocabulary candidates. We therefore introduce *Information Dynamics*, which combines the reference token's full-vocabulary rank with its score margin over a fixed competitor. Together, these quantities identify whether the reference token enters output competition and whether it subsequently loses its advantage. Figure 1 provides an overview of Information Dynamics and its use in prediction-guided activation steering.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

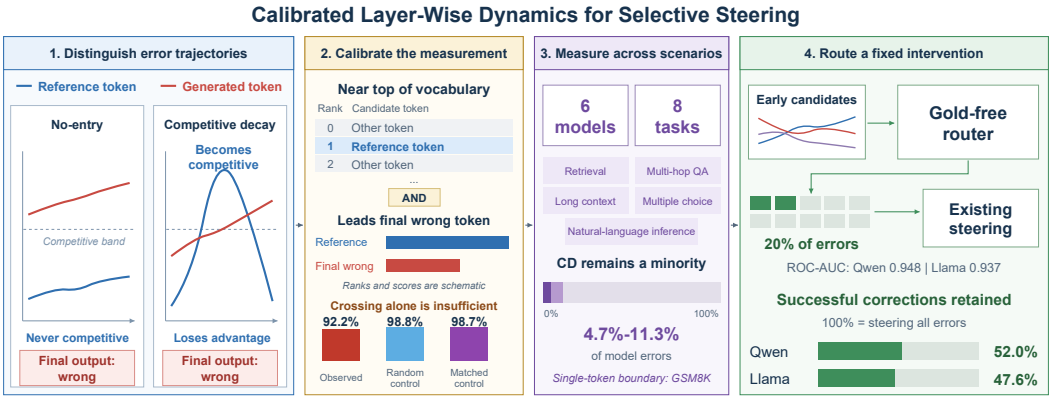


Figure 1: Overview of Information Dynamics. **(1)** The correct-answer token either never becomes competitive (no-entry) or becomes competitive and later loses its advantage (competitive decay, CD). **(2)** Information Dynamics combines full-vocabulary rank and competitor margin to determine whether the reference token enters output competition. **(3)** Across six models and eight tasks, CD and no-entry coexist within the same task, motivating sample-level intervention selection. **(4)** Without reference answers at prediction time, early candidate dynamics predict which errors are likely to be CD and guide a fixed steering budget, selecting the top 20% by predicted CD probability.

We apply Information Dynamics across six open-weight models and eight tasks. Competitive decay appears across model families and task settings, although it remains a minority of errors. Importantly, CD and no-entry errors coexist within individual tasks, indicating that task identity alone cannot determine which errors should be prioritized for steering. We also compare classifications obtained with the logit lens and tuned lens to assess whether the distinction is robust to the choice of intermediate decoder. The reference token is available for defining and analyzing these error types, but not when deciding which errors to steer at inference time. We therefore use the error types defined by Information Dynamics as supervision and predict them from the model's own early candidate dynamics, without access to the reference answer. The resulting predictor assigns each already-identified error a probability of being CD, which we use to prioritize intervention while leaving the steering method itself unchanged. We evaluate this prediction-guided selection on held-out Qwen3-8B and Llama-3.1-8B errors from seven tasks. When steering only the 20% of errors with the highest predicted CD probabilities, while keeping the steering method, strength, and number of interventions fixed, our selection yields 2.61× and 2.39× the expected number of successful corrections from random selection on Qwen3-8B and Llama-3.1-8B, respectively. These results show that layer-wise token dynamics provide a practical signal of error steerability and can make a fixed activation-steering method substantially more effective under a limited intervention budget. Our contributions connect layer-wise error analysis with prediction-guided activation steering:

1. **A layer-wise distinction associated with steerability.** We identify no-entry and competitive-decay errors and show that CD errors are approximately four times more likely to be recovered by the same activation-steering method, providing a basis for intervention selection.
2. **Information Dynamics.** We combine full-vocabulary rank and competitor margin to characterize these trajectories, and evaluate the distinction across six open-weight models, two intermediate decoders, and eight tasks, revealing cross-task variation and decoder-dependent cases.
3. **Reference-free prediction-guided steering.** We predict error type from early candidate dynamics without reference answers at prediction time. At a 20% intervention budget, this yields 2.61× and 2.39× the expected corrections from random selection on Qwen3-8B and Llama-3.1-8B.

2 Related Work

Selective and Adaptive Activation Steering. Activation-steering methods modify internal representations at inference time to control model behavior. ITI identifies directions associated with truthful outputs (Li et al., 2023), while representation engineering and activation-addition methods derive behavior-related directions from internal representations or contrasting examples (Zou et al., 2025; Turner et al., 2024; Rimsky et al., 2024). More recent methods adapt intervention to the input:

CAST conditionally applies steering based on activation patterns (Lee et al., 2025), MERA jointly calibrates intervention decisions and steering strength (Hedström et al., 2025), and DSAS modulates steering strength according to context (Ferrando et al., 2026). Our work asks a complementary question: *among examples already known to be incorrect, which are most recoverable under the same intervention?* We hold the steering method fixed and use predicted layer-wise error trajectories to select errors for steering, prioritizing more recoverable errors under a limited budget.

Layer-Wise Prediction Dynamics. Probing methods study what information is available in intermediate representations (Alain & Bengio, 2017; Belinkov, 2022), while the logit lens and tuned lens map these representations into vocabulary distributions (nostalgebraist, 2020; Belrose et al., 2025). Prior work has used such analyses to study correctness and uncertainty across layers (Jiang et al., 2024; Kim et al., 2025), and to identify failures arising at different stages of computation (Lu et al., 2025). Intermediate predictions have also been used to improve generation: early-exit methods terminate computation from sufficiently confident intermediate states (Schuster et al., 2022; Yom Din et al., 2024), DoLa contrasts predictions across layers to improve factuality (Chuang et al., 2024), and DAMO uses cross-layer prediction dynamics to mitigate late-layer errors (Wang et al., 2025). Instead, we focus on the layer-wise fate of the correct-answer token and ask whether entering and subsequently leaving the output competition predicts recovery under activation steering.

Correctness Estimation and Error Characterization. Prior work estimates answer correctness from self-evaluation and sampling consistency (Kadavath et al., 2022; Manakul et al., 2023), or from internal representations and latent-knowledge signals (Azaria & Mitchell, 2023; Burns et al., 2023; Liu et al., 2025). Other work distinguishes errors associated with missing knowledge from failures that occur despite evidence that the relevant knowledge is available (Simhi et al., 2024; Orgad et al., 2025). In this work, we condition on examples that have already been identified as erroneous and characterize *how* the correct answer evolves internally before the error is produced. We then use such trajectory distinctions for reference-free intervention selection.

3 Identifying Layer-Wise Error Trajectories

Our analysis asks whether errors differ in whether the correct answer ever becomes a competitive output candidate at an intermediate layer. To study this distinction, we introduce *Information Dynamics*, which tracks the layer-wise evolution of the correct-answer token and distinguishes two error trajectories: *no-entry*, where the token never enters output competition, and *competitive decay* (CD), where it becomes competitive but loses its advantage before the final output. We first establish how output competition should be measured, then examine where these trajectories occur across models and tasks, and how sensitive their identification is to the measurement choices.

3.1 When Does the Correct Answer Enter Output Competition?

We follow a designated reference token across layers at a fixed prediction position. Let x denote the input question and prompt, $y^\star$ the first token of the designated reference answer, and $\hat{y}$ the first generated token. We use one designated reference answer per example, including examples whose evaluator accepts multiple aliases. All measurements are taken at the *answer-onset position*, the final prompt position that predicts the first generated token. A layer-wise decoder maps the residual-stream representation after layer ℓ to a vocabulary distribution,$p_\ell(y \mid x)$. We use both the raw logit lens and a tuned lens (Belrose et al., 2025); their construction is described in Section 3.2. A natural way to ask whether the reference token ever became competitive is to test for *score crossing*: whether it outscores the model's generated first token, $\hat{y}$, at any intermediate layer. However, exceeding one selected token does not necessarily imply that the reference token became a strong output candidate. It may outrank that token while remaining below many others in the vocabulary. We test this issue by comparing crossings against the generated first token with crossings against control tokens. Over 200 permutations, the *random-token control* replaces it with a token generated on another error, while the *strength-matched control* uses a token whose final log-probability differs from that of the generated first token by at most 1. On Qwen3-8B, the reference token crosses the generated first token in 92.2% of errors, compared with 98.8% under the random-token control and 98.7% under the strength-matched control (95% CI [98.0, 99.6]; Figure 2). The same qualitative pattern appears across all six models (Appendix C.1). These results show that crossing a single comparison token is too permissive to establish that the reference token genuinely entered output competition.

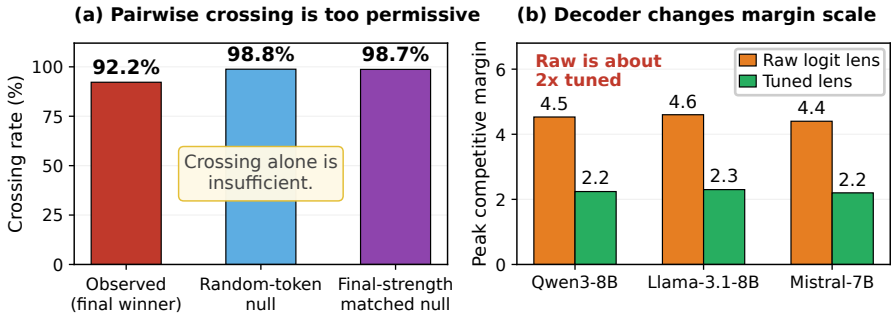


Figure 2: Measurement checks for pairwise crossing and decoder scale. Left: the observed crossing rate against the generated first token is 92.2%, compared with 98.8% for the random-token control and 98.7% for the strength-matched control. Pairwise crossing alone therefore does not establish output competitiveness. Right: raw- and tuned-lens peak margins differ by approximately a factor of two, motivating analyses under both decoders to assess measurement stability.

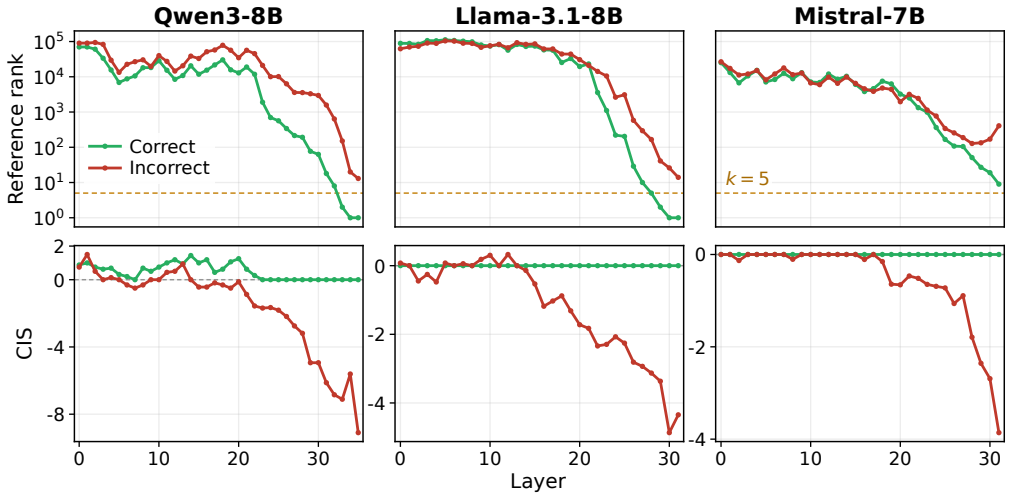


Figure 3: Reference-token trajectories for correct (green) and incorrect (red) answers. The panels show median vocabulary rank (top) and competitive margin (bottom) across layers. The margin is labeled CIS in the figure. Two types of answers exhibit distinct trajectories across model families.

Information Dynamics therefore measures both the reference token's position in the full vocabulary and its advantage over a fixed competitor. We define its zero-based vocabulary rank $r_\ell(y^\star)$ as the number of tokens assigned a higher probability than $y^\star$ at layer ℓ, so rank 0 indicates that no token has higher probability. We fix the competitor as $y^- = \arg\max_{y \neq y^\star} p_L(y \mid x)$, and define the *competitive margin* $m_\ell(x) = \log p_\ell(y^\star \mid x) - \log p_\ell(y^- \mid x)$. The same competitor is used at every layer. Vocabulary rank measures the reference token's position among all possible outputs, while the margin measures whether it leads the strongest final incorrect candidate. For a rank cutoff k, we say that the reference token has entered *output competition* at layer ℓ when both$r_\ell(y^\star) \leq k$ and $m_\ell(x) > 0$. The rank criterion requires the reference token to be near the top of the full vocabulary, while the positive margin requires it to lead the fixed final competitor. The two conditions then rule out cases in which the reference token crosses one selected token while remaining globally implausible. As a sanity check on vocabulary rank as a measure of output competitiveness, correct answers reach better intermediate reference-token ranks than incorrect answers across all six models (Appendix Table 8). The median best rank ranges from 0 to 11 for correct answers and from 9 to 77 for errors, with the separation holding under both raw and tuned lenses (Mann–Whitney $p < 10^{-20}$; Appendix C.3). This result supports using vocabulary rank to capture how close the reference token comes to the model's leading output candidates. Figure 3 illustrates layer-wise trajectories for three models. We use these two measurements to distinguish error trajectories under a fixed decoder and rank cutoff.

An incorrectly answered example is a *competitive-decay* (CD) error if the reference token enters output competition at some intermediate layer but trails the fixed competitor at the final layer:

$$r_\ell(y^\star) \le k \quad \text{and} \quad m_\ell(x) > 0 \quad \text{for some } \ell < L,\ m_L(x) \quad < 0. \tag{1}$$

Among errors with $m_L(x) < 0$, an example is a *no-entry* error when the two intermediate-layer conditions are never satisfied together. The distinction then records whether the reference token entered output competition and subsequently lost its advantage, or never entered under the joint criterion. Our primary analysis uses $k = 5$, allowing at most five tokens to rank above the reference token. We test sensitivity to $k \in \{1, 5, 10, 50\}$ in Section 3.4. Of the 1,558 Qwen3-8B errors, 7 have $m_L(x) \ge 0$ despite an incorrect complete answer; we treat these as downstream sequence errors and report them separately in Appendix B.1. Because the competitor is selected from the final layer, the resulting labels are retrospective and conditional on both the decoder and rank cutoff. Appendix D additionally defines rank-only alternatives that do not use the final competitor.

3.2 Experimental Protocol

Models and Tasks. We evaluate six open-weight instruction-tuned models spanning 4B to 14B parameters: Qwen3-14B, Qwen3-8B, and Qwen3-4B (Yang et al., 2025); Qwen2.5-7B-Instruct (Yang et al., 2024); Llama-3.1-8B-Instruct; and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). Our evaluation covers eight tasks: TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), GSM8K (Cobbe et al., 2021), CommonsenseQA (Talmor et al., 2019), ARC-Challenge (Clark et al., 2018), MMLU-4subj (Hendrycks et al., 2021), QuALITY (Pang et al., 2022), and MNLI (Williams et al., 2018). Together, these tasks span retrieval, multi-hop question answering, arithmetic reasoning, multiple-choice reasoning, long-context comprehension, and natural-language inference. The evaluation set contains 2,150 examples in total. Dataset selection details and per-task baseline outcomes are reported in Appendix A and Table 5, providing context for comparisons across tasks.

Generation and Evaluation. For each model, we use its official chat template, request answer-only responses, and generate greedily up to 32 tokens. We retain generations beginning with formatting tokens and treat them separately in a sensitivity analysis. Correctness is determined using task-specific evaluators, with full generation settings and matching rules provided in Appendix H. To assess label quality, two annotators independently conduct a blinded audit of 300 randomly sampled examples from tasks. The resulting correctness labels have a precision of 0.957 and a recall of 0.978.

Layer-Wise Decoders. We analyze each example using both the raw logit lens and a tuned lens. The raw lens applies the model's final RMSNorm and unembedding head to intermediate residual-stream states. The tuned lens fits one affine translator per layer on 200 held-out unlabeled prompts, using the model's final-layer logits as targets and without access to reference answers or correctness labels. Its mean held-out KL divergence across layers is 0.12 nats. Comparing the two readouts allows us to test whether the trajectory distinction persists under different mappings from intermediate states to vocabulary distributions. Further tuned-lens training details are provided in Appendix C.3.

3.3 Where Does Competitive Decay Occur?

Having defined competitive decay, we next ask whether it occurs across model families and task settings, or only under particular conditions. Unless otherwise stated, we use the raw logit lens with the primary rank cutoff $k = 5$. We first apply Eq. 1 to each model's own incorrect answers. Competitive decay appears in all six models, with rates ranging from 4.7% on Mistral-7B to 11.3% on Qwen3-14B (Table 1). Qwen3-8B has a CD rate of 9.9%, Qwen3-4B 8.2%, Qwen2.5-7B 8.9%, and Llama-3.1-8B 8.5%. Thus, CD recurs across model families and scales while remaining a minority of each model's error pool. We next examine task-level variation on Qwen3-8B. CD appears in seven of the eight tasks, spanning retrieval, multi-hop question answering, long-context comprehension, multiple-choice reasoning, and natural-language inference (Table 2). The highest rates occur on HotpotQA at 14.2% and MMLU at 12.6%, followed by TriviaQA at 11.6%, QuALITY at 11.2%, MNLI at 10.9%, ARC-Challenge at 10.5%, and CommonsenseQA at 8.7%. Each of these seven tasks contains both CD and no-entry errors. At answer onset, we detect no CD examples on GSM8K, with an upper endpoint of 1.8% for the two-sided 95% exact interval. This result is specific to our first-token, answer-onset measurement: arithmetic answers may develop across multiple tokens or reasoning steps, which are not captured by the present criterion. Overall, the coexistence of competitive decay and no-entry errors across seven individual tasks shows that task identity alone

Table 1: Competitive-decay (CD) prevalence among each model's own incorrect answers, measured with the logit lens at $k = 5$. CD occurs in all six models but accounts for only 4.7–11.3% of their errors. CI denotes the 95% confidence interval for CD prevalence within each model's error pool.

Model	Own errors	CD rate	95% CI
Qwen3-14B	1,189	11.3%	[9.5, 13.3]
Qwen3-8B	1,558	9.9%	[8.4, 11.5]
Qwen3-4B	1,742	8.2%	[6.9, 9.5]
Qwen2.5-7B	1,581	8.9%	[7.5, 10.3]
Llama-3.1-8B	1,590	8.5%	[7.1, 10.0]
Mistral-7B	1,835	4.7%	[3.7, 5.7]

Table 2: Task-level CD prevalence among Qwen3-8B's incorrect answers, measured at answer onset with the logit lens and $k = 5$. CD appears in seven of the eight tasks; none is detected on GSM8K under this measurement. CI denotes the two-sided 95% Clopper–Pearson confidence interval.

Task	Errors	CD count	CD rate	95% CI
HotpotQA	190	27	14.2%	[9.6, 20.0]
MMLU	167	21	12.6%	[8.0, 18.6]
TriviaQA	310	36	11.6%	[8.3, 15.7]
QuALITY	134	15	11.2%	[6.4, 17.8]
MNLI	129	14	10.9%	[6.1, 17.5]
ARC-Challenge	229	24	10.5%	[6.8, 15.2]
CommonsenseQA	196	17	8.7%	[5.1, 13.5]
GSM8K	203	0	0.0%	[0.0, 1.8]
Overall	**1,558**	**154**	**9.9%**	**[8.4, 11.5]**

cannot determine which errors follow the CD trajectory. Identifying such cases therefore requires sample-level trajectory information to prioritize intervention among errors within the same task.

3.4 Sensitivity to Measurement Choices

The CD definition depends on two measurement choices: the vocabulary-rank cutoff used to define entry into output competition, and the decoder used to map intermediate representations to vocabulary scores. We therefore test how the identified cohort changes under both choices.

Rank Cutoff. Our primary analysis uses $k = 5$. To test whether the conclusion that CD is a minority trajectory depends on this threshold, we repeat the Qwen3-8B analysis with $k \in \{1, 5, 10, 50\}$. The corresponding CD rates are 5.1%, 9.9%, 11.2%, and 17.4%, respectively. As expected, relaxing the rank criterion increases the number of examples classified as CD, but CD remains a minority of errors at every tested cutoff. The exact prevalence and cohort membership therefore depend on how strictly output competition is defined, while the qualitative finding persists across the tested thresholds. Appendix D reports rank-only alternatives and their relationship to Eq. 1.

Intermediate Decoder. We next compare the raw and tuned lenses on the same 1,558 Qwen3-8B errors at $k = 5$. The raw lens identifies 154 CD errors (9.9%), while the tuned lens identifies 138 (8.9%). The two decoders agree on 96.8% of all errors, with Cohen's $\kappa = 0.81$ and a Jaccard similarity of 0.708 between their CD sets. Both identify 121 errors as CD, corresponding to 7.8% of all errors. The shared CD cases also show stronger trajectory evidence than cases identified only by the raw lens: their median peak margin is 4.1, compared with 1.2 for raw-only cases. Thus, the two decoders recover a substantial common subset, while some borderline cases remain sensitive to the intermediate readout used to map the same hidden states to vocabulary probabilities.

3.5 Prediction-Guided Activation Steering

However, using the above distinction to guide steering requires identifying likely CD errors without access to the reference answer at the time of intervention. We examine early predictive signals, construct a reference-free predictor, and use its scores to apply intervention methods.

Do Early Dynamics Contain Predictive Signal? To test whether early trajectories contain information about later decay, we train a linear classifier on first-half reference-token and final-competitor features. It predicts whether the competitive margin is positive at some second-half layer and negative at the final layer. This retrospective target omits the vocabulary-rank requirement of CD. Five-fold cross-validation on Qwen3-8B yields ROC-AUC 0.75, supporting the presence of predictive information in early trajectories; additional statistics and comparisons appear in Appendix E.

Reference-Free Prediction from Early Candidates. To predict CD versus no-entry among examples already identified as incorrect, we use first-half candidate dynamics: changes in the leading candidate, persistence of its lead, margins among top candidates, and top-five candidate entropy. These features require neither the reference token nor the final competitor's identity. We train five random-forest classifiers with the same data and selected hyperparameters but different seeds, using the labels from Section 3. Their mean predicted CD probability ranks errors for intervention; training, data splits, and model selection are described in Section 4.1 for both evaluated models.

Budgeted Selection with a Fixed Steering Operation. To allocate the budget, we steer errors with the highest predicted CD probabilities. We use activation addition along layer-specific differences between the mean last-token hidden states of correct and incorrect training responses (Turner et al., 2024; Zou et al., 2025). For each selected error, we add the direction scaled by α to the last-token residual stream at layers $L-3$ through L, then regenerate greedily. Each model's directions are derived from training data and α is selected on calibration data (Appendix F). Keeping both fixed isolates the effect of error selection in the equal-budget comparisons in Section 4.

4 Experiments

Our experiments evaluate three aspects of the proposed framework: the recoverability of CD versus no-entry errors, the ability of early candidate dynamics to identify CD errors without reference answers, and the effectiveness of prediction-guided intervention under a fixed budget.

4.1 Experimental Setup

Models and Data Splits. We evaluate Qwen3-8B and Llama-3.1-8B on the seven non-GSM8K tasks, using the generation and evaluation protocol from Section 3.2. The shared dataset contains 1,900 questions, divided into 1,140 training questions (60%), 114 calibration questions (6%), and 646 evaluation questions (34%). The same question-level split is used for both models, and no question appears in more than one split. Each split contains both correct and incorrect model responses. Additional split details are provided in Appendix I to document the question-level evaluation setup.

Predictor Training. The reference-free predictors are trained on the incorrect responses in the training split: 813 Qwen errors and 832 Llama errors. Hyperparameters are selected using nested cross-validation with five outer folds and three inner folds. As described in Section 3.5, we train five classifiers with different random seeds and average their predicted CD probabilities. All evaluation-set prediction scores are frozen before examining steering outcomes. Further predictor details are provided in Appendix E, including the hyperparameter search and five-seed ensemble construction.

Steering Calibration. Each model uses a steering direction estimated only from its training split. Directions are constructed from equally sized groups of correct and incorrect examples: 327 of each for Qwen and 308 of each for Llama, with inverse-frequency task weighting. Steering strength is selected on calibration errors from $\alpha \in \{0.5, 1.0, 1.5, 1.8, 2.0, 2.5\}$, yielding $\alpha = 1.8$ for Qwen and $\alpha = 2.0$ for Llama. The resulting direction and coefficient are then fixed for all evaluation-set selection methods. Full intervention details are provided in Appendix F, including regeneration.

Evaluation Protocol. We first apply the intervention to each evaluation error to determine which errors are recoverable. An error is considered recoverable under this intervention if the complete regenerated answer is judged correct by the task-specific evaluator. This produces 461 evaluation errors for Qwen and 472 for Llama. For budgeted selection, each selection method chooses approximately 20% of its model's error pool: 92 errors for Qwen and 94 for Llama. Random selection is represented by the expected number of recoveries from a uniformly sampled subset of the same size, and one-sided hypergeometric tests assess enrichment over this expectation. We report ROC-AUC and PR-AUC for distinguishing CD from no-entry errors. For intervention selection, we report the number and rate of successfully recovered errors under the fixed intervention budget.

Table 3: Recovery under the same fixed activation-steering intervention, stratified by error trajectory. Cells report correct, complete regenerated answers / treated errors (recovery rate). Competitive-decay (CD) errors recover at substantially higher rates than no-entry errors on both models.

	Qwen3-8B	**Llama-3.1-8B**
All evaluation errors	25/461 (5.4%)	21/472 (4.4%)
CD	9/52 (17.3%)	7/46 (15.2%)
No-entry	16/409 (3.9%)	14/426 (3.3%)

4.2 Competitive-Decay Errors Are More Steerable

We first test whether the trajectory distinction is associated with recovery under the fixed activation-steering intervention. To isolate this question from prediction quality, we use the reference-based CD and no-entry labels defined by Information Dynamics and apply the same calibrated steering operation to every evaluation error. As shown in Table 3, the intervention recovers 9 of 52 CD errors (17.3%) on Qwen3-8B, compared with 16 of 409 no-entry errors (3.9%). On Llama-3.1-8B, it recovers 7 of 46 CD errors (15.2%), compared with 14 of 426 no-entry errors (3.3%). Thus, CD errors are recovered at approximately four times the rate of no-entry errors on both models. The difference is statistically significant under Fisher's exact test ($p < 0.001$ for Qwen and $p = 0.002$ for Llama). Across the full evaluation error pools, the same intervention recovers 25 of 461 Qwen errors (5.4%) and 21 of 472 Llama errors (4.4%). The substantially higher recovery rates within the CD subsets show that the layer-wise trajectory distinction identifies errors that are more amenable to this fixed intervention. If CD errors can be identified without reference answers, prioritizing them can focus the intervention budget on errors more likely to recover under the same intervention.

4.3 Reference-Free Prediction of Competitive Decay

Having established that CD errors are substantially more recoverable under the fixed intervention, we next ask whether they can be identified without access to the reference answer. We evaluate the frozen predictor from Section 3.5 on each model's held-out evaluation errors, with full prediction metrics reported in Appendix E, Table 10. The predictor separates CD from no-entry errors with a ROC-AUC of 0.948 on Qwen3-8B and 0.937 on Llama-3.1-8B. The corresponding PR-AUC values are 0.728 and 0.658. These results indicate that the model's own early candidate dynamics contain strong information about whether an already-incorrect example follows the competitive-decay trajectory. We next examine whether these scores are useful for selecting a small subset of errors. Selecting approximately 20% of the evaluation error pool by highest predicted CD probability chooses 92 of 461 Qwen errors and 94 of 472 Llama errors. Among these selected errors, 42 are CD for Qwen and 38 for Llama. This corresponds to CD precision of 45.7% and 40.4%, while capturing 80.8% and 82.6% of all CD errors, respectively. Because CD constitutes only 11.3% of the Qwen evaluation error pool and 9.7% of the Llama pool, prediction-guided selection substantially increases its concentration. The selected subsets are enriched for CD by $4.05\times$ on Qwen and $4.15\times$ on Llama. For Qwen, this enrichment is significant under a hypergeometric test ($p < 10^{-20}$). Thus, early candidate dynamics compress most CD errors into roughly one fifth of the error pool, providing a reference-free signal for prioritizing intervention. These results show that the trajectory distinction can be approximated from reference-free early candidate dynamics with high predictive accuracy. The remaining question is whether concentrating CD errors in this way actually improves recovery under a fixed steering budget relative to alternative selection methods.

4.4 Prediction-Guided Steering Under a Fixed Budget

We now test whether prioritizing errors predicted to be CD improves recovery under a fixed intervention budget. Each selection method chooses approximately 20% of the evaluation error pool. Within each model, all methods use the same steering direction and strength and treat the same number of errors. Table 4 compares candidate-dynamics selection with random selection and several final-layer baselines: lowest generated-token confidence, highest vocabulary entropy, smallest top-two score gap, and an endpoint random forest using only final-layer confidence, entropy, and margin features. Random selection yields an expected 5.0 recoveries among 92 treated Qwen errors and 4.2 among 94 treated Llama errors. Selecting errors based on early candidate dynamics

Table 4: Recovery under different selection methods at approximately 20% intervention coverage. All selection methods use the same fixed activation-steering operation within each model. Cells report correct, complete regenerated answers / selected errors (recovery rate). Random reports the expected recovery count for an equal-sized uniformly sampled subset. Endpoint RF uses final-layer confidence, entropy, and top-two margin; Candidate dynamics averages predictions from five early-dynamics classifiers. Candidate dynamics yield the highest recovery count on both models at the same budget.

Selection method	Qwen3-8B	Llama-3.1-8B
Random (expected)	5.0/92 (5.4%)	4.2/94 (4.4%)
Final confidence (lowest)	10/92 (10.9%)	3/94 (3.2%)
Final entropy (highest)	10/92 (10.9%)	4/94 (4.3%)
Final top-two margin (smallest)	4/92 (4.3%)	0/94 (0.0%)
Endpoint RF (final features)	11/92 (12.0%)	9/94 (9.6%)
Candidate dynamics (5-seed mean)	**13/92 (14.1%)**	**10/94 (10.6%)**

recovers 13 Qwen errors and 10 Llama errors, corresponding to recovery rates of 14.1% and 10.6%, respectively. Relative to equal-sized random selection, this gives $2.61\times$ and $2.39\times$ the expected number of recoveries, respectively. Candidate dynamics achieves the highest recovery count among the evaluated selection methods on both models. On Qwen, candidate dynamics corrects 13 of 92 selected errors, compared with 11 of 92 for the endpoint random forest. The 13 Qwen recoveries selected by candidate dynamics include eight CD errors and five no-entry errors. Thus, prediction-guided selection does not succeed solely by recovering CD errors. Predicted trajectory information improves the allocation of the fixed intervention budget while still allowing recoveries from both error types. Overall, these results prove that CD errors are more recoverable under the fixed steering method, early candidate dynamics identify errors likely to be CD without reference answers, and using these predictions to prioritize intervention yields substantially more recoveries than random selection at the same budget, without changing the steering method or its calibrated strength.

4.5 Ablation Studies

Intervention Budget. To test whether the selection benefit depends on the intervention budget, we vary the fraction of Qwen3-8B errors selected from 10% to 50%. The number of recovered errors increases from 6 to 17, and all four tested budgets exceed equal-sized random-selection expectations (Table 11). Thus, the benefit is not restricted to the primary 20% budget used in the main evaluation.

Predictor Initialization. To assess sensitivity to initialization, we compare the five-seed ensemble with a single classifier trained using seed 42. The single classifier selects 12 recoverable errors among 92 Qwen errors and 9 among 94 Llama errors, compared with 13 and 10 for the ensemble. Both seed-42 results exceed random-selection expectations ($p = 0.001$ and $p = 0.012$, respectively), although recovery counts and statistical significance vary across other seeds. Full seed results and the formatting-first filtering analysis are provided in Appendix F as additional robustness checks.

5 Conclusion

We studied whether layer-wise token trajectories can reveal which language-model errors are most likely to benefit from activation steering. We distinguish *no-entry* errors, in which the correct-answer token never becomes competitive, from *competitive-decay* (CD) errors, where it becomes competitive and later loses its advantage. Information Dynamics characterizes this distinction using full-vocabulary rank and competitor margin. Under the same fixed steering method, CD errors are recovered approximately four times as often as no-entry errors. We further show that error type can be predicted from early candidate dynamics without access to the reference answer at inference time. Using these predictions to allocate a fixed steering budget more than doubles successful corrections relative to random selection when steering only 20% of errors on both Qwen3-8B and Llama-3.1-8B. These results suggest that layer-wise token dynamics provide a practical signal of error steerability. This conclusion is operational rather than causal: competitive decay identifies an intervention opportunity under the readouts and steering method studied here, rather than a universal mechanism of model knowledge. These findings support error selection within a fixed budget.

REPRODUCIBILITY STATEMENT

Section 3.2 describes the models, tasks, generation protocol, evaluation rules, and layer-wise decoders. Section 4.1 documents the question-disjoint data splits, predictor training, steering calibration, and evaluation protocol. Appendix E provides additional predictor details, Appendix F reports the intervention settings and robustness analyses, and Appendix H specifies the generation and evaluation procedures. Together, these sections document the procedures and data partitions used to reproduce the reported analyses. We will release the implementation and analysis code upon acceptance.

AI USE STATEMENT

Generative AI tools were used for manuscript organization. We verified all scientific claims, equations, citations, and reported numbers and take full responsibility for the submission. No generative AI tools were used for idea generation, research hypothesis formulation, experimental design, or results analysis; all technical contributions, empirical results, and conclusions are our own.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *ICLR 2017 Workshop Track*, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

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, 2025. URL https://arxiv.org/abs/2303.08112.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Alex Ferrando, Xavier Suau, Jordi Gonzàlez, and Pau Rodriguez. Dynamically scaled activation steering. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=EXKv1KMNtR.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Anna Hedström, Salim I. Amoukou, Tom Bewley, Saumitra Mishra, and Manuela Veloso. To steer or not to steer? Mechanistic error reduction with abstention for language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22924–22945. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/hedstrom25a.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Che Jiang, Biqing Qi, Xiangyu Hong, Dayuan Fu, Yang Cheng, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On large language models' hallucination with regard to known facts. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1041–1053, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.60. URL https://aclanthology.org/2024.naacl-long.60/.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Sunwoo Kim, Haneul Yoo, and Alice Oh. On the effect of uncertainty on layer-wise inference dynamics. In *ICML 2025 Workshop on Actionable Interpretability*, 2025. URL https://arxiv.org/abs/2507.06722v1. Online-only poster 149.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*,

2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e2dd53601de57c773343a7cdf09fae1c-Abstract-Conference.html.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530. Curran Associates, Inc., 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/81b8390039b7302c909cb769f8b6cd93-Paper-Conference.pdf.

Jiarui Liu, Jivitesh Jain, Mona T. Diab, and Nishant Subramani. [NON-ARCHIVAL] LLM microscope: What model internals reveal about answer correctness and context utilization [NON-ARCHIVAL]. In *The First Workshop on the Interplay of Model Behavior and Model Internals*, 2025. URL https://openreview.net/forum?id=8jkQGNkRzP.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. Paths not taken: Understanding and mending the multilingual factual recall pipeline. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15066–15096. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.762. URL https://aclanthology.org/2025.emnlp-main.762/.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

nostalgebraist. interpreting GPT: the logit lens. LessWrong, August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a712d461e57201efe35d429a6f1731c1-Abstract-Conference.html.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel R. Bowman. QuALITY: Question answering with long input texts, yes! In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5336–5358, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.391. URL https://aclanthology.org/2022.naacl-main.391/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17456–17472. Curran Associates, Inc., 2022. doi: 10.52202/068431-1269. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6fac9e316a4ae75ea244ddcef1982c71-Paper-Conference.pdf.

Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. Distinguishing ignorance from error in LLM hallucinations, 2024. URL https://arxiv.org/abs/2410.22071.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Kaishen Wang, Hengrui Gu, Meijun Gao, and Kaixiong Zhou. DAMO: Decoding by accumulating activations momentum for mitigating hallucinations in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/8f0f23417ce1d00212a7c85c2560c392-Abstract-Conference.html.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. Jump to conclusions: Short-cutting transformers with linear transformations. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9615–9625, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.840/.

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, 2025. URL https://arxiv.org/abs/2310.01405.

APPENDIX CONTENTS

A Task-Level Baseline Outcomes 16
B Measurement Definitions and Error Categories 16
C Cross-Model Measurement Checks 17
D Threshold Sensitivity and Endpoint-Free Variants 18
E Prediction: Additional Analyses 19
F Intervention and Attribution Details 20
G Gold-Direction Ablation 22
H Generation and Evaluation Details 22
I Second-Model Replication: Llama-3.1-8B 23

Table 5: Qwen3-8B answer correctness on the sampled questions under answer-only greedy generation and task-specific evaluation. Total is the number of sampled questions; Correct and Errors count complete answers accepted or rejected by the evaluator, and Accuracy is Correct / Total. MMLU-4subj uses four MMLU subjects. The 2,150 questions yield 1,558 errors for trajectory analysis, with accuracy varying across tasks. These are sampled-protocol results, not standard benchmark scores.

Task	Total	Correct	Errors	Accuracy
TriviaQA	500	190	310	38.0%
HotpotQA	250	60	190	24.0%
GSM8K	250	47	203	18.8%
CommonsenseQA	250	54	196	21.6%
ARC-Challenge	250	21	229	8.4%
QuALITY	200	66	134	33.0%
MMLU-4subj	200	33	167	16.5%
MNLI	250	121	129	48.4%
Overall	**2,150**	**592**	**1,558**	**27.5%**

A Task-Level Baseline Outcomes

Dataset Selection. We select 2,150 examples sequentially from the validation/test splits of the eight datasets. Table 5 reports the number selected from each dataset and the corresponding Qwen3-8B outcomes. The datasets cover retrieval, multi-hop question answering, arithmetic reasoning, multiple-choice reasoning, long-context comprehension, and natural-language inference. QuALITY uses the complete supplied context without character truncation, and MMLU is restricted to four subjects.

QuALITY uses the complete supplied context and the one-line answer prompt, with 66 correct answers among 200 examples (33.0% accuracy). Across the seven non-GSM8K tasks, the baseline contains 545 correct answers and 1,355 errors among 1,900 questions. These task-level outcomes characterize the sampled questions and evaluation protocol; prevalence can change under a different prompt, context budget, or evaluator.

B Measurement Definitions and Error Categories

Different analyses use different trajectory quantities and prediction targets. We summarize these definitions and then account for the answer-onset categories in the sampled Qwen3-8B responses. This separates the quantities used to define error types from the inputs available when predicting them.

Measurement Definitions. To clarify what each analysis measures, Table 6 lists its token comparison, classification condition, or prediction target. Pairwise crossing compares two tokens, whereas CD additionally requires vocabulary-rank entry and a subsequent loss of the reference token's advantage. Reference-conditioned prediction tests margin decay without the rank requirement; gold-free routing predicts the CD label from the model's own candidates. The table therefore distinguishes retrospective measurement from reference-free prediction inputs.

B.1 Operational Cohort Categories

To account for the sampled responses, we separate the two primary error types from downstream sequence errors and describe when correct answers enter the rank band. Among 1,558 Qwen3-8B errors: 1,397 (89.7%) are no-entry errors, 154 (9.9%) are competitive-decay errors, and 7 (0.4%) are downstream errors (final top-1 gold token but judged-wrong complete answer). Among 592 correct examples: 512 (86.5%) enter the rank band before the final layer (early-entry correct) and 80 (13.5%) enter only at or near the final layer (late-entry correct). These are answer-onset cohorts, not an exhaustive taxonomy of full-sequence mechanisms.

Table 6: Measurements used in the trajectory and intervention analyses. The reference token is $y^\star$, the generated first token is $\hat{y}$, and y^- is the highest-scoring final non-reference token. At layer ℓ, p_ℓ is token probability, r_ℓ is zero-based vocabulary rank, and $m_\ell = \log p_\ell(y^\star) - \log p_\ell(y^-)$. L is the final layer and k the rank cutoff. For attribution, w_y is a token's unembedding vector and a_ℓ an MLP output. Pairwise crossing, margin-only decay, and competitive decay have different targets; gold-free prediction excludes reference and final-competitor identities from its inputs.

Analysis	**Measurement or Prediction Target**
Pairwise crossing calibration	$m_\ell^{\text{gen}} = \log p_\ell(y^\star) - \log p_\ell(\hat{y})$
Correct vs. incorrect separation	Competitive margin $m_\ell(x)$
Competitive-decay classification	$r_\ell(y^\star) \leq k$ and $m_\ell(x) > 0$ at some $\ell < L$, followed by $m_L(x) < 0$
Reference-conditioned prediction	First-half margin features; target: $m_\ell(x) > 0$ at some $\lfloor L/2 \rfloor < \ell < L$ and $m_L(x) < 0$, without a rank condition
Gold-free candidate routing	First-half top-five candidate dynamics, excluding $y^\star$ and y^- from prediction inputs; training labels use the CD definition above (Eq. 1)
Direct logit attribution	$(w_{y^\star} - w_{y^-})^\top a_\ell(x)$

Table 7: Pairwise score-crossing rates among each model's incorrect answers. A crossing occurs when the reference token outscores the comparison token at any intermediate layer. Observed uses the generated incorrect token; Random null replaces it with a token generated on another error; Matched null uses a control whose final log-probability is within 1 of the original competitor's. Rates are higher under both controls for all six models, showing that crossing alone is insufficient evidence of output competition.

Model	**Observed**	**Random null**	**Matched null**
Qwen3-14B	94.1%	99.2%	99.0%
Qwen3-8B	92.2%	98.8%	98.7%
Qwen3-4B	91.3%	98.4%	98.1%
Qwen2.5-7B	90.8%	97.9%	97.6%
Llama-3.1-8B	91.5%	98.2%	98.0%
Mistral-7B	89.7%	97.4%	97.2%

C Cross-Model Measurement Checks

To assess whether the trajectory distinction is supported beyond one model or decoder, we examine three aspects of the measurement: pairwise-crossing controls, reference-token vocabulary ranks, and agreement between decoding methods. These checks establish what each signal contributes and where decoder dependence remains.

C.1 Pairwise Crossing Controls

To test whether exceeding one incorrect token is sufficient evidence of output competition, we repeat the score-crossing comparison across all six models. Table 7 compares the generated incorrect token with random and final-score-matched replacements. Both controls produce higher crossing rates in every model, showing that the limitation of pairwise crossing extends beyond Qwen3-8B.

C.2 Best Intermediate Vocabulary Rank

To test whether vocabulary-wide ranking adds information beyond a two-token comparison, we compare the reference token's best intermediate rank for correct and incorrect complete answers. For each example, the best rank is the minimum rank before the final layer; Table 8 reports the median within each group. Correct answers have better median ranks in all six models. Vocabulary rank therefore captures differences in candidate competitiveness that a pairwise comparison does not establish.

Table 8: Median peak gold rank before the final layer (zero-based). The separation between correct and incorrect is robust across all models. Mann-Whitney $p < 10^{-20}$ in all cases.

Model	Correct	Incorrect
Qwen3-14B	1	9
Qwen3-8B	0	19
Qwen3-4B	2	10
Qwen2.5-7B	1	26
Llama-3.1-8B	1	20
Mistral-7B	11	77

Table 9: Agreement between logit-lens (Raw) and tuned-lens classifications on 1,558 Qwen3-8B errors at zero-based rank cutoff $k = 5$. Cells count examples. CD denotes competitive decay; non-CD includes all remaining errors, including downstream sequence errors. The diagonal gives 96.8% overall agreement, and both methods identify 121 CD errors; 50 errors receive different labels.

	Tuned: CD	Tuned: non-CD	Raw total
Raw: CD	121	33	154
Raw: non-CD	17	1,387	1,404
Tuned total	138	1,420	1,558

C.3 Decoder Robustness and Agreement

Tuned-Lens Training. Each layer-specific affine translator is trained on 200 held-out unlabeled prompts using Adam for 500 epochs, with a learning rate of 10^{-3} and mean squared error against the model's final-layer logits.

Probe Robustness Across Decoders. Peak-rank comparisons are summarized in Table 8. Decoder agreement is reported separately in Table 9; the two analyses address rank separation and label overlap, respectively.

No Answer-Label Leakage. The tuned-lens translators are trained on 200 held-out unlabeled prompts and never observe gold answers or correctness labels. This rules out answer-label leakage. However, the target is the final output distribution (including the wrong distribution on errors), so the tuned lens is aligned toward the model's final prediction.

Raw versus Tuned Lens Scale. The raw lens produces peak margins about $2.0\times$ larger than the tuned lens, while both preserve the aggregate ordering between correct and incorrect examples. This supports qualitative separation but motivates decoder-conditional reporting of absolute rates.

D Threshold Sensitivity and Endpoint-Free Variants

Robustness Across Competitiveness Thresholds. The competitive-decay rate across $k \in \{1, 5, 10, 50\}$ on Qwen3-8B is 5.1%, 9.9%, 11.2%, and 17.4%. The exact prevalence is cutoff-dependent, but the minority conclusion is stable.

Rank-Only Alternatives. An alternative based only on the reference token's vocabulary rank is

$$C_{\mathrm{rank}}(k_{\mathrm{in}}, k_{\mathrm{out}}) = \mathbf{1}\left[\min_{\ell<L} r_\ell(y^\star) \le k_{\mathrm{in}}\right] \mathbf{1}[r_L(y^\star) > k_{\mathrm{out}}]\,.$$

Under the same decoder and on the same error set, the primary CD definition with $k = 5$ implies membership in $C_{\mathrm{rank}}(5, 0)$. Both require intermediate rank entry, and the primary condition $m_L < 0$ implies that at least one token outranks the reference token at the final layer. The reverse implication need not hold because rank-only entry does not require a positive intermediate margin over the endpoint-selected competitor. This is a relationship between definitions, not an empirical statement about their prevalence or agreement.

Table 10: Reference-free prediction of competitive decay (CD) on held-out errors from seven tasks. The predictor uses only early candidate dynamics and does not access the reference answer at prediction time. The highest predicted CD probabilities select approximately 20% of errors. CD precision is the fraction of selected errors that are CD, recall is the fraction of all CD errors captured, and enrichment is precision divided by the CD proportion in the full error pool.

Metric	**Qwen3-8B**	**Llama-3.1-8B**
Evaluation errors	461	472
CD errors (pool proportion)	52 (11.3%)	46 (9.7%)
ROC-AUC (bootstrap 95% CI)	0.948 [0.918, 0.972]	0.937 [0.903, 0.965]
PR-AUC	0.728	0.658
Selected errors	92	94
CD errors selected	42	38
CD precision	45.7%	40.4%
CD recall	80.8%	82.6%
CD enrichment	$4.05\times$	$4.15\times$

Sequence-Level Measurement. Answer-onset entry concerns one designated reference token. A sequence-level criterion must additionally specify how entry is combined across answer tokens and which prefix is used to obtain their conditional probabilities. Under a gold prefix, requiring all answer tokens to enter a rank band at a common layer and requiring any token to enter and later leave a band define different counterfactual quantities. Neither is equivalent to the answer-onset cohort in Eq. 1.

E Prediction: Additional Analyses

Reference-Conditioned Diagnostic. The diagnostic in Section 3.5 uses first-half reference-token and final-competitor features to predict whether the competitive margin is positive at some second-half layer and negative at the final layer. This target omits the vocabulary-rank requirement, giving a positive rate of 38.4%, compared with 9.9% for CD under the primary definition. On Qwen3-8B, the linear classifier achieves ROC-AUC 0.75 under fivefold cross-validation, with a fold standard deviation of 0.02 and a bootstrap 95% CI of [0.71, 0.79]. In a retrospective comparison, final-layer features alone achieve AUC 0.70, while combining final-layer and trajectory features reaches 0.77, a gain of 0.07 (paired bootstrap $p = 0.03$). Adding trajectory features therefore improves discrimination within this retrospective diagnostic.

Nested CV for the Gold-Free Router. The candidate-dynamics router uses the following procedure:

1. Nested CV on 813 Qwen training errors: 5-fold outer, 3-fold inner for hyperparameter selection.
2. Select best hyperparameters (max_depth=6, n_estimators=200, min_samples_leaf=10).
3. Fit five final models on all 813 training errors, using seeds 42, 0, 1, 2, and 3.
4. Predict once on 461 evaluation errors and freeze all five score vectors.
5. Average the five probabilities, then compute held-out ROC-AUC, PR-AUC, and coverage tests.

Search grid: max_depth $\in \{4, 6, 8, \text{None}\}$, n_estimators $\in \{100, 200, 500\}$, min_samples_leaf $\in \{5, 10, 20\}$. Feature importance: leader churn (0.24), top-1 margin slope (0.20), five-token entropy (0.18), prefix-leader hold (0.16), top-2 margin (0.13), other (0.09).

Held-Out Reference-Free Prediction. Table 10 reports prediction metrics and CD enrichment on held-out Qwen3-8B and Llama-3.1-8B errors. Selecting approximately 20% of errors by predicted CD probability captures over 80% of CD cases and increases their concentration approximately fourfold on both models.

Cross-Task Transfer of the Reference-Conditioned Diagnostic. Transfer is structure-dependent. Entity-style QA transfers well (TriviaQA→HotpotQA: AUC 0.74; HotpotQA→TriviaQA: 0.71). Multi-hop to multi-step reasoning transfers moderately (HotpotQA→GSM8K: 0.68). Single-step retrieval to multi-step reasoning transfers poorly (TriviaQA→GSM8K: 0.53). The strongest nonlinear cross-task direction (TriviaQA→HotpotQA with random forest) reaches 0.82.

Cross-Model Transfer of the Reference-Conditioned Diagnostic. Leave-one-model-out AUC averages 0.72 across Qwen3-8B, Llama-3.1-8B, and Mistral-7B. Pairwise: Qwen3→Llama 0.73, Qwen3→Mistral 0.75, Llama→Qwen3 0.70, Llama→Mistral 0.76, Mistral→Qwen3 0.68, Mistral→Llama 0.65. Layers are aligned by relative depth. If the same questions appear for different models, transfer can inherit item difficulty; a question-disjoint split is required before interpreting the matrix as architecture-only generalization.

Trajectory Decomposition. No single feature matches the full trajectory (AUC 0.75). Top individual features: margin sign oscillations 0.64, peak first-half margin 0.64, first-half variance 0.63, margin slope 0.61, midpoint margin 0.60, midpoint rank 0.51. Combining peak + slope + variance reaches 0.70. This pattern shows that predictive signal resides in joint trajectory shape, not any single snapshot.

Temporal Baselines. Persistence: AUC 0.60; first-half slope: 0.58; linear extrapolation: 0.58; AR(1): 0.62; first-half mean: 0.56. All achieve lower AUCs than the full feature set (0.75). The comparison shows simple temporal continuity does not account for the observed discrimination.

Endpoint Conditioning Scope. The competitive margin $m_\ell(x)$ uses the final-layer competitor y^-, so first-half margin features already reveal which token eventually wins. Reference-only features, consisting of rank, log-probability, and slopes without competitor identity, reach AUC 0.66. The gap between 0.75 and 0.66 quantifies the value of the final-competitor contrast within this retrospective task. The gold-free router avoids this issue entirely by using neither reference nor competitor information.

F Intervention and Attribution Details

Steering Direction Construction. The representation-engineering direction is the mean last-token hidden-state difference between 327 correct and 327 incorrect Qwen3-8B examples at layers $L-3$ through L, computed entirely on the training split (1,140 questions). Correct examples are all 327 available from training questions where Qwen3-8B answers correctly (per-task: TriviaQA 113, MNLI 73, QuALITY 40, HotpotQA 36, CommonsenseQA 32, MMLU 20, ARC 13); incorrect examples are 327 randomly sampled from the 813 training errors. We apply inverse-frequency weighting when computing the mean difference to account for unequal task representation. The coefficient $\alpha = 1.8$ is selected from $\{0.5, 1.0, 1.5, 1.8, 2.0, 2.5\}$ on the 81-error calibration split.

Intervention Procedure. All 461 evaluation errors receive the same steering intervention: at layers $L-3$ through L, the direction vector scaled by $\alpha = 1.8$ is added to the residual stream at the last token position. The model then regenerates greedily for up to 32 tokens. Recovery means the regenerated answer matches the gold answer under the automatic evaluator. Result: 25/461 recover (5.4%).

Hypergeometric Test Details. With $N = 461$ total evaluation errors, $K = 25$ recoverable, and a selection of $n = 92$ (20% coverage), the number of recoveries X follows Hypergeometric(N, K, n). The expected value is $nK/N = 4.99$. We report one-sided $P(X \geq x_{\text{obs}})$:

- Candidate ensemble: $x = 13$, $P(X \geq 13) = 2.33 \times 10^{-4}$
- Candidate seed 42: $x = 12$, $P(X \geq 12) = 0.0011$
- Final entropy: $x = 10$, $P(X \geq 10) = 0.0145$
- Final confidence: $x = 10$, $P(X \geq 10) = 0.0145$
- Final margin: $x = 4$, $P(X \geq 4) = 0.772$

Seed Robustness. The primary score is the mean of five routers trained with seeds 42, 0, 1, 2, and 3. Individual-seed Qwen recovery counts at 20% coverage are 12, 11, 10, 9, and 7 (mean 9.8, sample SD 1.9); individual ROC-AUC ranges from 0.793 to 0.881. Thus the direction of enrichment is stable, although not every single seed reaches $p < 0.05$. The mean ensemble recovers 13 and reaches ROC-AUC 0.948. On Llama, individual recovery counts range from 6 to 12 (mean 7.6, SD 2.5), while the mean ensemble recovers 10 and reaches ROC-AUC 0.937.

Coverage Sweep. To examine how recovery changes with the intervention budget, we vary the selected fraction of Qwen errors across four budgets. Table 11 reports the complete sweep: treating more errors increases the number of recoveries, while the recovery rate decreases from 14.1% at 20% coverage to 7.4% at 50%. All four budgets exceed the corresponding random-selection expectation.

Table 11: Robustness of prediction-guided recovery under fixed activation steering. Qwen and Llama denote Qwen3-8B and Llama-3.1-8B. Coverage is the fraction of evaluation errors selected, approximately 20% unless otherwise specified. Recovery reports correct complete regenerated answers / selected errors; Rate is the corresponding percentage. Selection averages five classifiers' predicted competitive-decay probabilities, except for the single-classifier Seed 42 rows. Formatting-first filtering removes generations beginning with formatting tokens. Each p-value is from a one-sided hypergeometric test against equal-sized random selection from the corresponding error pool. All listed settings yield more recoveries than expected from random selection ($p < 0.05$).

Setting	Model	Recovery	Rate	p-value
Budget: 10%	Qwen	6/46	13.0%	0.029
Budget: 20%	Qwen	13/92	14.1%	< 0.001
Budget: 30%	Qwen	15/138	10.9%	0.001
Budget: 50%	Qwen	17/230	7.4%	0.048
Seed 42	Qwen	12/92	13.0%	0.001
Seed 42	Llama	9/94	9.6%	0.012
Formatting-first removed	Qwen	12/89	13.5%	< 0.001
Formatting-first removed	Llama	9/92	9.8%	0.008

Formatting-First Generations. We also remove errors whose generations begin with formatting tokens, leaving 446 Qwen and 459 Llama errors. Under this filtering, error-type prediction remains similar: ROC-AUC/PR-AUC are 0.947/0.726 for Qwen and 0.935/0.651 for Llama. At approximately 20% coverage, candidate-guided selection recovers 12 of 89 Qwen errors and 9 of 92 Llama errors, again exceeding random-selection expectations ($p < 0.001$ and $p = 0.008$, respectively).

Direct Logit Attribution. We use direct logit attribution (Elhage et al., 2021; Geva et al., 2021) to measure each MLP write's projection onto the reference-versus-competitor direction. Let w_y be the unembedding vector for token y, and let $a_\ell(x)$ be the output written by the MLP at layer ℓ. The attributed change in competitive margin is:

$$\Delta m_\ell^{\mathrm{MLP}} = (w_{y^\star} - w_{y^-})^\top a_\ell(x). \tag{2}$$

This is a local direct-effect diagnostic; layer normalization and subsequent computation can change downstream effects. Across Qwen3-8B errors, cumulative negative direct attribution is -4.71 for MLPs and -3.42 for attention. The layer-35 MLP is the largest single negative contributor (-3.42), but adjacent layers from 31 through 35 also respond to intervention, so we do not claim unique necessity.

Gold-Direction Activation Patching (Oracle Probe). Let $L^\star = \arg\max_{\ell<L} m_\ell(x)$ be the layer with the largest intermediate competitive margin, and let $\hat{w}_{y^\star} = w_{y^\star}/\|w_{y^\star}\|$ be the unit reference-token unembedding direction. At a later layer ℓ, we use a nonnegative coefficient β to restore the positive component that has been lost relative to $h_{L^\star}$:

$$h'_\ell = h_\ell + \beta \,\mathrm{clamp}_{\geq 0}\big((h_{L^\star} - h_\ell)^\top \hat{w}_{y^\star}\big)\hat{w}_{y^\star} \tag{3}$$

Larger β applies stronger restoration. Among examples not already top-1, this recovers first-token rank-0 in 7.3% of CD errors and 0.7% of no-entry errors ($p = 0.005$); an orthogonal random direction recovers 0/109 controls ($p = 0.003$). The control shows direction specificity relative to a norm-like perturbation but does not match direct first-order reference-logit changes.

Component-Selective Patching (Oracle Probe). Let a be the output written by the MLP component under test, and let $\hat{d} = (w_{y^\star} - w_{y^-})/\|w_{y^\star} - w_{y^-}\|$ be the unit reference-versus-competitor direction. For a nonnegative patch strength γ, we remove the component of a that opposes this direction:

$$a' = a - \gamma \,\mathrm{clamp}_{\leq 0}(a^\top \hat{d})\hat{d}. \tag{4}$$

For the layer-35 MLP on CD errors, first-token top-1 recovery is 11.0%, 20.2%, and 31.2% at $\gamma = 1, 2, 4$. Adjacent MLPs respond similarly: layer 34 ranges from 12.8% to 33.0%, layer 33 ranges from 11.9% to 25.7%, and layer 31 ranges from 4.6% to 22.0%. Full zero-ablation performs worse. An orthogonal random direction recovers 0/109. Because $\hat{d}$ is constructed from reference

Table 12: Effect of removing the reference-token unembedding direction from the last three hidden states of originally correct examples. Baseline correct gives correct / tested examples before ablation. Gold flip rate is the proportion becoming incorrect after reference-direction removal; Random flip reports the corresponding random-direction control. Brackets give Wilson 95% confidence intervals. Reference-direction removal flips 40–87% of answers, compared with 0–5% for the random control, supporting direction specificity.

Model	Baseline correct	Gold flip rate	Random flip
Qwen3-14B	30/30	77% [59, 88]	0%
Qwen3-8B	50/50	76% [63, 86]	0%
Qwen3-4B	27/27	67% [48, 81]	0%
Qwen2.5-7B	30/30	87% [70, 95]	3%
Llama-3.1-8B	30/30	73% [56, 86]	0%
Mistral-7B	30/30	40% [25, 58]	0%

and competitor unembeddings, these experiments locate an oracle-accessible direction but do not identify a gold-free semantic variable.

Unresolved Direct-Effect Control. For 17 examples with the largest gold-patch gains, no tested orthogonal or frequent non-gold direction matched the instantaneous gold-logit increase within the search cap (perturbation norms up to $25.7\times$ larger; 0/17 matched). Since the intended effect match was not achieved, we cannot separate restoration of an internal state from the direct advantage of moving along the gold unembedding direction.

Final-Rank-Matched Restoration (Oracle Probe). We test whether the recovery difference between CD and no-entry remains after approximately matching their final reference-token ranks. Each CD error is paired with a no-entry error whose final rank differs by at most 3. Using an oracle restoration coefficient of $\alpha = 1.0$, we blend each residual toward that example's own peak state. This probe uses the reference trajectory to select the peak state.

The 72 matched pairs have median final ranks of 7 for CD and 10 for no-entry, and median best ranks of 2 and 34. Complete-answer recovery is 11.1% for CD (8/72) and 1.4% for no-entry (1/72). Among the nine pairs with different recovery outcomes, eight favor CD (exact one-sided McNemar $p = 0.020$). The result supports a difference in recovery under oracle restoration even when final ranks are approximately matched.

G Gold-Direction Ablation

As a direction-specific sanity check, we remove the component of the last three hidden states along the gold unembedding direction:

$$h'_\ell = h_\ell - \mathrm{proj}_{w_{y^\star}}(h_\ell). \tag{5}$$

Applied to originally correct examples, this changes correct predictions to incorrect ones at the rates shown in Table 12.

Norm-Matched Controls. Matching the hidden-state change $\|\Delta h\|$ for control directions separates direction from perturbation size (but not direct first-order logit effect). On Qwen3-8B: gold flips 83%, second-best candidate 39%, random token 6%, random Gaussian 11%. On Qwen3-4B: 70%, 35%, 15%, 5%. On Qwen3-14B: 75%, 55%, 5%, 0%. Matching $\|\Delta h\|$ rules out generic magnitude but does not match direct first-order gold-logit changes. We therefore interpret the experiment as a direction-specific sanity check, not a complete causal control.

H Generation and Evaluation Details

Prompting and Decoding. To make the generation protocol explicit, we report the prompting and decoding settings used across models. All models use their official Hugging Face chat templates with the user message "{question} `Answer with just the answer.`". For Qwen3, we

Table 13: Manual audit of automatic evaluator (300 examples, stratified by task, two independent annotators). Disagreements resolved by discussion.

	True Pos	False Pos	True Neg	False Neg
Count	178	8	110	4
Precision: 0.957		Recall: 0.978		F1: 0.967

set `enable_thinking=False`. Generation uses greedy decoding, bfloat16 precision, and a maximum of 32 generated tokens. Stop tokens include `<|endoftext|>`, `<|im_end|>`, and model-specific EOS tokens. QuALITY uses the complete supplied context without character truncation.

Formatting-Token Starts. To check whether answer formatting affects the analysis, we flag generations that begin with formatting tokens. Among the 2,150 generations for each model, 71 Qwen3-8B examples (3.3%) and 60 Llama-3.1-8B examples (2.8%) fall into this category. We retain these examples in the main analysis and use the flags for a separate sensitivity analysis, preserving the full sampled set while allowing the effect of formatting to be examined.

Automatic Evaluation. To assign correctness labels, we use evaluation rules appropriate to each answer format. For text answers, we convert responses to lowercase, remove articles and punctuation, and apply substring matching; TriviaQA also accepts reference-answer aliases. For numeric answers, we remove commas and percent symbols before comparison and use an absolute tolerance of 0.5. For multiple-choice and natural-language inference tasks, we compare the first extracted option letter or label with the reference. These rules determine the correct and incorrect groups used in the trajectory analysis.

Human Audit. To assess the reliability of the automatic correctness labels, two annotators independently conduct a blinded audit of 300 randomly sampled examples stratified by task. Disagreements are resolved through discussion. The automatic evaluator achieves a precision of 0.957 and a recall of 0.978. The audit results and observed evaluation errors are reported below.

False positives: 5 from substring match accepting answer in negated context ("not A but B" matched as A), 3 from partial numeric match. False negatives: 3 valid paraphrases, 1 equivalent numeric form ("3/4" vs. "0.75"). Among the 12 evaluator errors, 1 belongs to the CD cohort. This small number of audit errors does not establish independence between evaluator errors and cohort membership.

I Second-Model Replication: Llama-3.1-8B

To assess whether the routing result generalizes beyond Qwen3-8B, we replicate the intervention pipeline on Llama-3.1-8B using the same eight tasks and evaluation protocol.

Setup. Llama-3.1-8B produces 1,590 errors on the 2,150-example set (560 correct; 26.0% accuracy). Non-GSM8K comprises 1,900 questions, 1,387 errors, and 513 correct answers. We apply the same 60/6/34 question-disjoint split: training has 1,140 questions (832 errors, 308 correct), calibration has 114 (83 errors, 31 correct), and evaluation has 646 (472 errors, 174 correct). Because the same questions and split rule are used for both models, this is an architecture replication on a shared item set, not an independent-dataset replication.

CD Prevalence. Llama-3.1-8B CD rate under Eq. 1 (raw lens, $k = 5$) is 8.5% overall (135/1,590; Table 1). The evaluation split contains 46 CD errors among 472 errors (9.7%).

Router and Steering. We train five separate random forests on the 832 Llama-3.1-8B training errors with the same nested CV procedure and candidate-dynamics features. Their mean score reaches ROC-AUC 0.937 (bootstrap 95% CI [0.903, 0.965]) and PR-AUC 0.658 on the 472 evaluation errors. The steering direction is computed from Llama hidden states (308 correct vs. 308 incorrect, inverse-frequency weighted, layers $L-3$ through L). Coefficient $\alpha = 2.0$ is selected on the Llama calibration split.

Results. All 472 evaluation errors receive Llama-3.1-8B steering; 21/472 (4.4%) recover. At 20% coverage, the ensemble recovers 10/94 (10.6%; hypergeometric $p = 0.003$), while seed 42 recovers 9/94 (9.6%; $p = 0.012$). Final confidence, entropy, and margin recover 3, 4, and 0 examples,

respectively, versus 4.2 expected under random allocation. The ensemble thus replicates the Qwen enrichment on another architecture. Its top 94 contain 38/46 CD errors (40.4% precision, 82.6% recall, $4.15\times$ enrichment; $p < 10^{-20}$).