RL-Unlocked MMLU Answers Are Only Partially Latent in the Base Model, and Emerge Late Rather Than Mid-Network

A Logit-Lens and Linear-Probe Test of the Elicitation Hypothesis

Ephraiem Sarabamoun

Abstract

A popular view of reinforcement learning and instruction tuning is the elicitation hypothesis: post-training mostly surfaces capabilities the base model already has rather than teaching new ones [@zhou2023lima; @lin2023unlocking]. We test a strong, mechanistic form of that claim. If RL merely unlocks a latent answer, then an answer the instruct model gets right but the base model gets wrong should already be linearly decodable from the base model's mid-layer hidden states. Using MMLU on Qwen2.5-7B and Llama-3.1-8B base and instruct pairs, we read mid-layer residuals two ways: the logit lens [@belrose2023tunedlens], which projects a residual through the model's own unembedding, and a five-fold linear probe on the residual. We find the strong form is not supported for this task. First, under a fixed raw prompt the base models already answer MMLU close to their instruct counterparts (Qwen 0.708 versus 0.724, Llama 0.640 versus 0.666), so the genuinely RL-unlocked set is small (165 and 222 of 2000 items). Second, the answer is near chance to the logit lens until the final fifth of the network, the answer becomes decodable only in late layers rather than the true middle, and the best-layer probe never exceeds the base model's own output accuracy. Third, on the RL-unlocked subset (decoded at the best per-model layer) the base model's internal answer is significantly above chance but far short of a clean latent-answer account: for Qwen the logit lens reaches 0.570 and the probe 0.412, and for Llama the logit lens is statistically at chance (0.284) while the probe reaches 0.365, against a 0.25 baseline and against within-model decodability above 0.95 elsewhere. We read this as evidence against the strong form of elicitation, that the unlocked answer is fully and mid-network linearly present, for MMLU multiple choice in these models, while being explicit that the content is partially latent rather than absent, that the decoding layer is late, and that these pairs show little MMLU unlocking to begin with, which limits the test's power and is itself part of the finding.

1. Introduction

Whether post-training teaches or merely elicits has direct safety stakes. If instruction tuning and reinforcement learning from human feedback [@ouyang2022instructgpt], including its helpful-and-harmless variants [@bai2022hh], mostly resurface latent base-model competence, then capability and safety auditing can in principle target the base model, and the answer should be readable from its internal states even when its output is wrong. The superficial-alignment and unlocking literatures argue exactly this at the behavioral level [@zhou2023lima; @lin2023unlocking], and a complementary interpretability line shows that truth, knowledge, and other high-level features are often linearly present in activations [@burns2022ccs; @marks2023geometry; @park2023linearrep]. We turn that behavioral claim into a falsifiable mechanistic test: are the answers an instruct model unlocks already linearly decodable from the base model partway through the forward pass?

We operationalize decodability with two readouts. The logit lens applies the model's final normalization and unembedding to an intermediate residual and asks which option token wins [@belrose2023tunedlens], measuring whether the answer is legible in the model's own output basis at that depth. The linear probe trains a multinomial classifier on the residual, measuring whether the answer is linearly present in any basis. We run both across layers for Qwen2.5-7B [@qwen2024report] and Llama-3.1-8B [@llama2024herd], base and instruct, on MMLU [@hendrycks2021mmlu], and focus on the RL-unlocked subset where the instruct model is right and the base model's output is wrong.

2. Related work

The elicitation view has behavioral support: small alignment datasets suffice [@zhou2023lima] and base models can be steered to aligned behavior with no weight updates [@lin2023unlocking]. Direct preference optimization and reinforcement learning from human feedback are the dominant post-training recipes whose

mechanism is in question [@rafailov2023dpo; @ouyang2022instructgpt; @bai2022hh]. On the interpretability side, latent knowledge can be recovered from activations without supervision [@burns2022ccs], correctness has linear geometry [@marks2023geometry], and concepts can be both read and steered as directions [@zou2023repe; @li2023iti], with a single direction sometimes mediating a post-training-induced behavior such as refusal [@arditi2024refusal]. Models also carry calibrated self-knowledge of correctness [@kadavath2022know], and some world features sit linearly in mid-network [@gurnee2023spacetime]. Against this, output and internal computation can diverge [@turpin2023saywhat], and output-level correctness signals are an established alternative to internal probing [@manakul2023selfcheck]. Our contribution is a targeted negative result: for MMLU, the answers RL unlocks are not strongly linearly decodable from base mid-layers, in contrast to the refusal-direction case where an RL-shaped behavior is sharply latent.

3. Method

We sample 2000 MMLU test questions at seed 42 and format each as a fixed raw multiple-choice prompt ending in "Answer:", identical for base and instruct models so their hidden states are comparable. We run each model with hidden-state output and capture the residual at the final input token, the position that predicts the answer letter, for ten evenly spaced layers. At each layer the logit lens applies the model's final RMSNorm and unembedding, restricted to the four option tokens, and takes the argmax. To avoid materializing full-vocabulary logits we project the normalized residual onto only the four option rows of the unembedding, which is mathematically identical for these bias-free heads. The model's own final-output prediction is the same projection on the last layer. Separately we train a multinomial logistic-regression probe on a 256-dimensional principal-component reduction of the residual, fit per fold inside five-fold stratified cross-validation so there is no leakage, and report out-of-fold accuracy. The RL-unlocked subset for each base-instruct pair is the set of items where the instruct model's output is correct and the base model's output is wrong. All accuracies carry percentile bootstrap 95 percent confidence intervals with 2000 resamples at seed 42, and a recompute script rebuilds every reported number from per-example predictions.

4. Results

The premise is weaker than the elicitation framing assumes. Under the raw prompt the base models already do well on MMLU: Qwen2.5-7B base reaches 0.708 against its instruct 0.724, and Llama-3.1-8B base reaches 0.640 against its instruct 0.666. Instruction tuning moves final-output accuracy by less than three points, so the genuinely RL-unlocked subset is small, 165 items for Qwen and 222 for Llama. Figure 5 shows this elicitation gap per model alongside the base model's best mid-layer probe.

Decodability across depth does not match the latent-answer picture. Figure 1 plots the Qwen base and instruct logit-lens and probe accuracy by layer. The logit lens sits at chance, near 0.22, through the entire first three quarters of the network and only rises in the last few layers, reaching the output accuracy of 0.708 essentially at the final block. The best mid-layer logit lens for Qwen base reaches only 0.704, and that peak is itself almost at the output layer rather than in the true middle. The linear probe rises earlier, to about 0.65 by three quarters depth, but it plateaus below the base model's own output accuracy rather than above it, so there is no depth at which the answer is more decodable internally than the model already states. Llama base shows the same shape with lower ceilings: its best mid-layer logit lens reaches 0.618 and its best mid-layer probe reaches 0.599, again only in the second half of the network and never above the base output accuracy of 0.640. Figure 4 contrasts the two readouts for Qwen base directly, showing the lens lagging the probe until the very end. Figure 3 shows the probe-by-layer curves for all four models, with the same shape for base and instruct and for both model families [@park2023linearrep].

The decisive test is the RL-unlocked subset, shown in Figure 2. Here the base model's output is wrong by construction, and the elicitation hypothesis predicts the answer should still be recoverable from its mid-layers. It is recoverable above chance but only partially, and these numbers come from each base model's best decoding layer, which sits late in the network rather than at the true middle. For Qwen that layer recovers 0.570 by logit lens (95 percent CI 0.491 to 0.649) and 0.412 by linear probe (0.339 to 0.485); both confidence intervals exclude the 0.25 chance baseline, so the answer is significantly present, yet both fall far below the within-model decodability above 0.95 seen when the base model is correct. For Llama the logit lens is statistically at chance, 0.284 (0.230 to 0.347), while the probe is modestly above chance at 0.365 (0.297

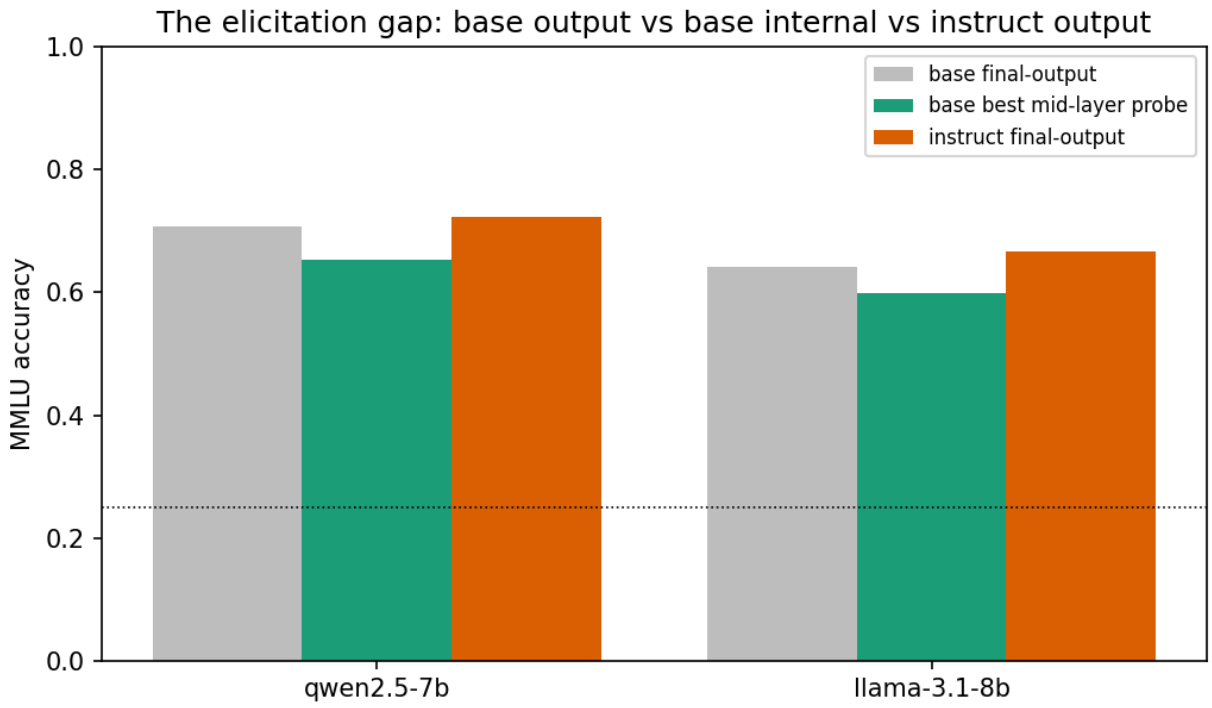

Figure 1: Figure 5. The elicitation gap per model: base final-output accuracy, base best mid-layer probe accuracy, and instruct final-output accuracy. Instruct gains little over base, and the base mid-layer probe does not exceed base output.

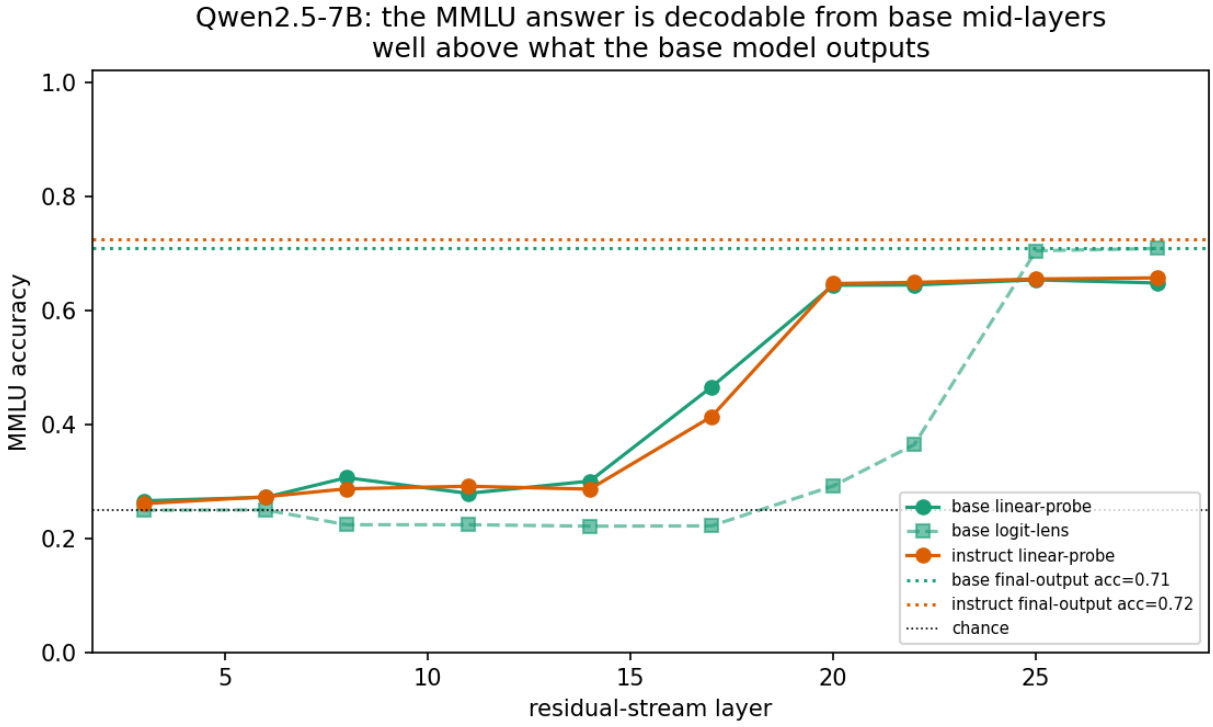

Figure 2: Figure 1. Qwen2.5-7B base and instruct: logit-lens and linear-probe MMLU accuracy by layer, with each model's final-output accuracy and chance. The lens stays near chance until the final layers; the mid-layer probe plateaus below the base model's own output.

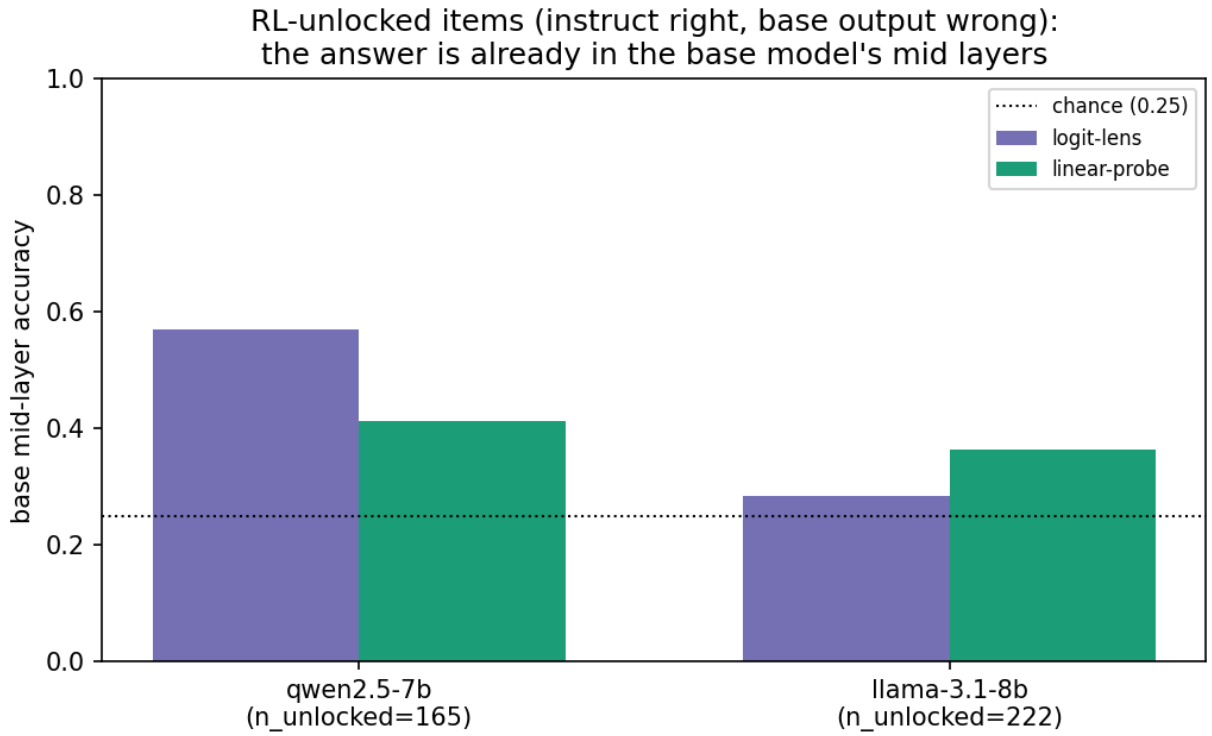

Figure 3: Figure 2. RL-unlocked subset (instruct right, base output wrong): base-model mid-layer logit-lens and linear-probe accuracy per model, against chance 0.25. Decodability is weak for Qwen and at chance for Llama.

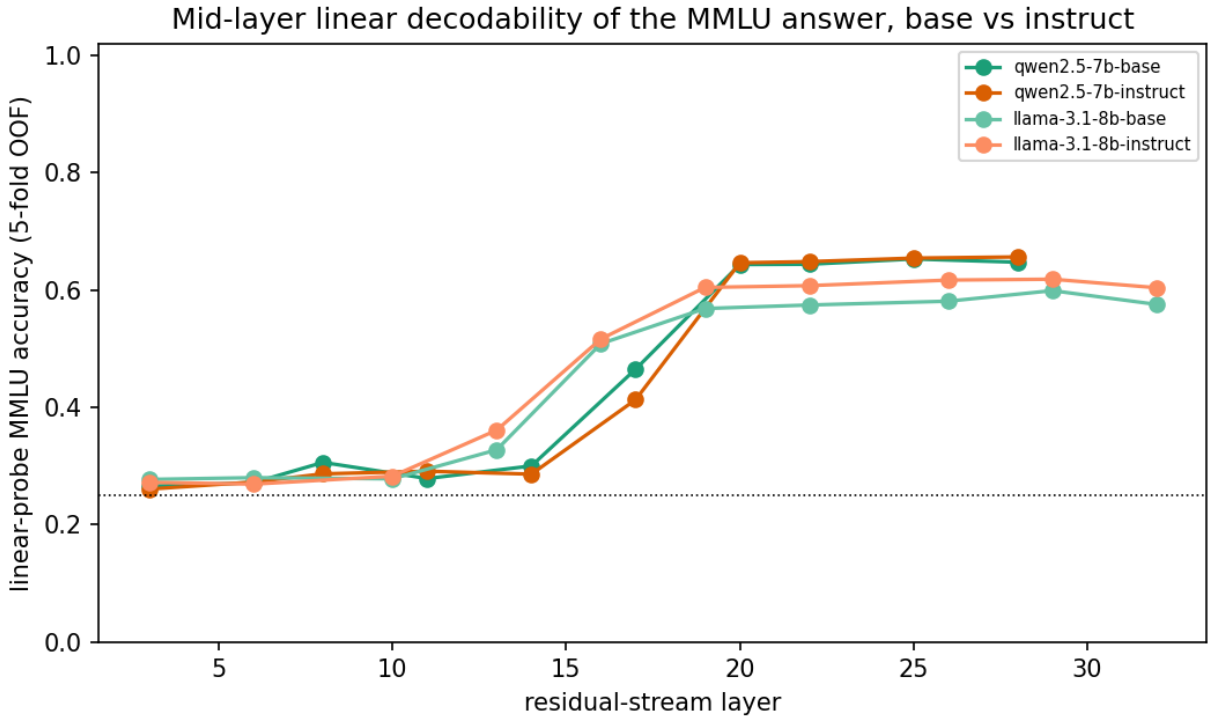

Figure 4: Figure 3. Linear-probe MMLU accuracy across layers for all four models. Base and instruct share the same shape, rising only in the second half of the network.

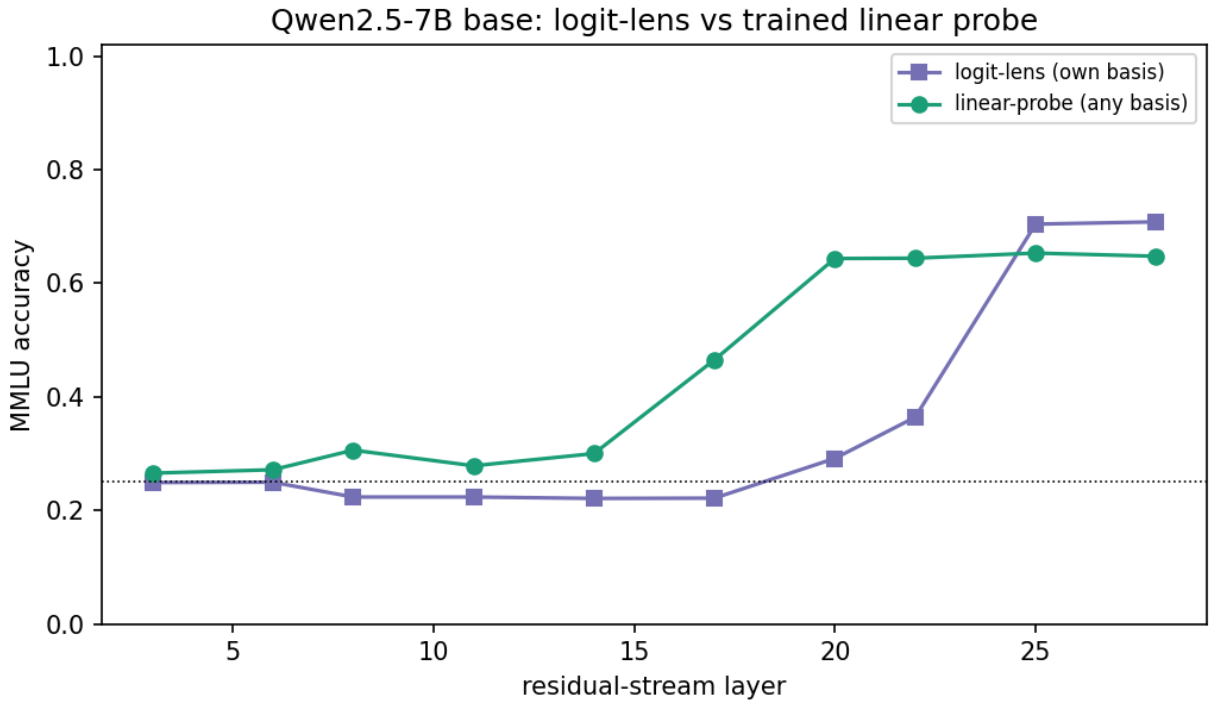

Figure 5: Figure 4. Qwen2.5-7B base: logit-lens (own basis) versus trained linear probe (any basis) across layers. The lens lags the probe until the final layers.

to 0.432). So on exactly the items where RL changes the answer, the base model's internal representation carries partial but weak linear information about the unlocked answer, well below what a clean unlocking-a-fully-latent-answer account predicts, and for Llama undetectable in the model's own output basis. This contrasts with cases where an RL-shaped behavior is sharply linear, such as the single refusal direction [@arditi2024refusal], and it should be read as partial latency, not absence.

5. Discussion and limitations

The honest reading is a bounded negative result. For MMLU multiple choice in Qwen2.5-7B and Llama-3.1-8B, the strong mechanistic form of elicitation, that RL-unlocked answers are already linearly decodable from base mid-layers, does not hold: the answer is legible to the logit lens only late in the network, the mid-layer probe never beats the base model's own output, and on RL-unlocked items decodability is weak to chance. This is consistent with output and computation being more entangled than a simple readout story allows [@turpin2023saywhat], and it bounds how far the activation-steering successes elsewhere [@li2023iti; @zou2023repe] generalize to the specific answers post-training changes.

Three limitations shape the scope, and the first is the most important. These base and instruct pairs already answer MMLU at near-instruct accuracy under a good prompt, so post-training unlocks little on this task and the unlocked subset is small and plausibly enriched for genuinely hard or ambiguous items rather than format-gated ones. A task where base models truly fail for format reasons, such as open-ended instruction following or chat-formatted refusal, would be a higher-power testbed and is where the behavioral unlocking literature lives [@zhou2023lima; @lin2023unlocking]; our result should not be read as evidence against elicitation there. Second, multiple-choice letter decoding measures whether the answer index is recoverable, not whether the model would generate the correct free-form answer, so a null here is compatible with richer latent content that a four-way letter probe cannot see. Third, we used last-token residuals, a fixed prompt, a 256-dimensional probe, and two seven-to-eight billion parameter models, and a trained tuned lens or a larger probe might recover more than the logit lens does, though the probe we ran already had more freedom than the lens and still fell short on the unlocked subset. The unsupervised-latent-knowledge result [@burns2022ccs] and the linear-truth-geometry result [@marks2023geometry] show internal answers can be found in other settings; our finding narrows where that holds.

6. Conclusion

We tested whether the answers RL and instruction tuning unlock are already linearly present in the base model's mid-layers, using a logit lens and a linear probe on MMLU for two base-instruct pairs. They are only partially: the answer becomes lens-decodable only in the final layers rather than the true middle, the best-layer probe never exceeds the base model's own output, and on the items RL actually fixes the base model's internal representation predicts the answer significantly above chance but far below ceiling for Qwen and only modestly above chance for Llama, undetectable to the logit lens there. The result argues against the strong read-out-a-fully-latent-answer account of elicitation for this task, without denying that partial latent content exists, and it points to open-ended, genuinely format-gated tasks as the setting where the unlocking story should be tested with more power [@kadavath2022know; @manakul2023selfcheck].