# Reusable Uncertainty Representations Support Metacognitive Behavior in Llama-3.3-70B

Kirubeswaran O R [1] Christopher Ackerman [2]

## Abstract

Knowing what you do not know is a hallmark feature of intelligent behavior. This is grounded in uncertainty monitoring, a core component of metacognition that guides decisions such as seeking help or withholding unreliable responses. For Large Language Models (think chatbots), this question is increasingly important, as these systems are deployed in consequential settings. Yet it remains unclear whether confidence-related behaviors in LLMs reflect reusable internal uncertainty representations or task-specific heuristics shaped by prompts and surface cues. We study this question in Llama-3.3-70B-Instruct by identifying residual-stream directions associated with answer-logit uncertainty during factual multiple-choice question answering, then testing whether the same directions transfer to metacognitive readouts including stated confidence and delegation. We find evidence for a reusable uncertainty-related representation: the directions transfer across datasets and readouts, acquire abstention-related output semantics, and can be steered to shift delegation behavior. Ablation provides weaker and more selective evidence, suggesting that these directions are sufficient to influence delegation but not uniquely necessary. Together, these results support a constrained account in which LLMs can reuse internal uncertainty-related representations across metacognitive contexts, while the behavioral expression of those representations depends on the task.

[1]Technical University of Darmstadt, Darmstadt, Germany [2]MATS Research. Correspondence to: Kirubeswaran O R <kirubeswaran.obula@tu-darmstadt.de>, Christopher Ackerman <christopher.ackerman@gmail.com>.

Mechanistic Interpretability Workshop at the $43^{rd}$ International Conference on Machine Learning, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

## 1. Introduction

An LLM can sound uncertain without using uncertainty representations. It can say "I'm not sure" because the prompt asks for caution, choose to delegate because delegation is offered as an option, or report a confidence score because the format demands a number. These behaviors may look like self-monitoring or introspection, a framing that recent work treats as both empirically promising and conceptually delicate (Binder et al., 2024; Comsa & Shanahan, 2025). Mechanistically, they could reflect distinct prompt-conditioned heuristics that are unrelated to any internal representation of uncertainty (Song et al., 2025).

The stronger possibility is that these behaviors share an answer-reliability representation formed while answering a question, and later reused when the model reports confidence or decides whether to delegate. Building on the behavioral metacognition task family introduced by Ackerman (2025), we learn uncertainty-related directions from ordinary multiple-choice QA in Llama-3.3-70B-Instruct, then ask whether those same directions transfer to two metacognitive readouts: stated confidence and answer-versus-delegate decisions.

Prior work makes this question plausible, but leaves it unresolved. Language models can produce confidence estimates from token probabilities, sampling-based consistency (Manakul et al., 2023; Lyu et al., 2024), and explicit verbal reports, and these estimates can sometimes be useful for calibration (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023), although verbal uncertainty may not be helpful for investigating a model's intrinsic uncertainty (Yona et al., 2024). Work on confidence-probability alignment directly compares internal token-probability confidence with expressed confidence (Kumar et al., 2024). Recent work suggests that verbal confidence can be computed from answer-adjacent representations (Kumaran et al., 2026a). Other work shows that confidence-related internal structure can guide answer-versus-abstain behavior, including under causal activation steering (Kumaran et al., 2026b). These re-

sults show that confidence-related information exists and can matter for behavior. What remains unclear is whether confidence reports and delegation decisions read out a common representation formed during answering, or rely on separate prompt-specific computations.

This distinction matters because behavior can be a misleading guide to mechanism. A delegation prompt may invite caution even when the model has no shared uncertainty-related representation; abstention-like behavior can also be induced by training or policy design (Zhang et al., 2024). A stated-confidence prompt may elicit a number even if that number is partly reconstructed from prompt-specific cues or surface features. But an uncertainty representation may be present even when delegation is weak, because delegation adds an extra policy decision: whether the model should act on its uncertainty by deferring. We therefore separate two questions that are often conflated: whether an uncertainty-related representation exists, and how strongly a particular task reads it out.

To test this, we use ordinary question answering as the source task and ask whether its uncertainty structure reappears in separate metacognitive prompts. For each layer, we identify residual-stream directions associated with direct-task uncertainty, using answer-logit quantities such as logit gap, top logit, and entropy. We then apply those same directions, without refitting, to activations from stated-confidence and delegation prompts. If a frozen QA-derived direction predicts these readouts, this supports reuse rather than a readout fitted separately within each metacognitive task. We use "reuse" in this restricted operational sense: a direction learned from direct QA is evaluated without refitting in a different prompt, and is counted as reused when it remains predictive there and is supported by alignment, controls, semantics, or interventions.

However, a linear direction can be predictive without being mechanistically meaningful. Prior work shows that high-level properties of language models can sometimes be captured by linear activation directions, including latent knowledge, truth-related representations, refusal behavior, and self-authorship judgments (Burns et al., 2024; Marks & Tegmark, 2024; Arditi et al., 2024; Ackerman & Panickssery, 2025). But probe accuracy alone is not enough: linear readouts can exploit dataset artifacts, label imbalance, or task-specific structure (Hewitt & Liang, 2019). We therefore do not rely on a single probe score. We ask whether the direction aligns across datasets, transfers across readouts, has interpretable output-space semantics, is specific relative to random and sign-flipped directions, and changes delegation behavior under steering and ablation.

Our results support a constrained form of reuse. Directions learned from direct-task uncertainty transfer across two factual QA distributions and predict both stated confidence and answer-versus-delegate behavior. The two readouts behave differently: stated confidence more directly reflects the QA-derived uncertainty direction, while delegation is more sensitive to dataset and policy details. This dissociation is the central result. Weak delegation coupling does not by itself imply that the uncertainty-related representation is absent; it may instead mean that the delegation prompt reads out that representation weakly, or combines it with an additional decision about whether to defer. We further find that these directions align across datasets, map onto abstention-related vocabulary under logit-lens analysis, and can be steered to shift delegation behavior. Ablation gives weaker and more selective evidence, so our claim is intentionally limited: we do not identify a single necessary introspective uncertainty variable, but show that QA-derived uncertainty directions can be reused across metacognitive contexts.

Our contributions are:

- We formulate a direct-to-meta-task transfer test for uncertainty representations: learn layerwise residual-stream directions from ordinary multiple-choice QA, then test whether those same directions predict stated confidence and delegation without refitting.
- We show that QA-derived uncertainty directions transfer across two factual QA datasets and two metacognitive readouts, with a key dissociation: stated confidence is a cleaner readout than delegation, which is more policy- and dataset-dependent.
- We support the representation claim with converging evidence beyond probe accuracy: cross-dataset direction alignment, direct-to-meta-task transfer, logit-lens abstention semantics, random-direction and sign-flip controls, steering effects on delegation, and cautious ablation evidence.

## 2. Methods

### 2.1. Model, datasets, and splits

We study Llama-3.3-70B-Instruct without finetuning. This model is large enough to exhibit nontrivial factual QA, confidence-reporting, and delegation behavior under instruction, while remaining feasible for full-layer

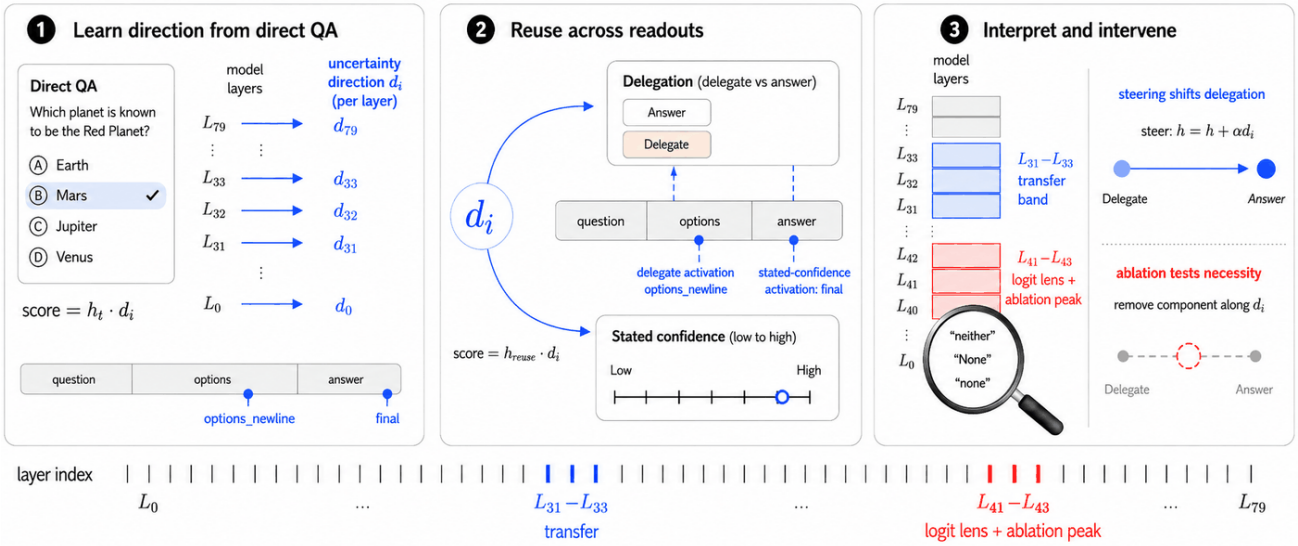



Figure 1. Experiment overview. We identify layer-specific uncertainty directions from ordinary multiple-choice QA, test whether the same directions transfer to delegation and stated-confidence readouts, and then interpret and intervene on those directions. The blue band denotes the main transfer locus (L31–L33); the red band denotes the layer range where logit-lens abstention signatures and final-position ablation effects are strongest (L41–L43).

residual-stream extraction using 4-bit inference. All experiments use seed 42. Analyses that fit a direction or scalar readout use the same 80/20 train-test split within each dataset: directions are estimated only from direct-QA training examples, while transfer, behavioral coupling, and interventions are evaluated on held-out examples. Appendix B records the implementation constants and prompt details.

The primary datasets are TriviaMC and PopMC, each with 500 multiple-choice questions. They play different roles in the evidence chain: TriviaMC is a broad factual QA distribution, while PopMC is a more structured entity-property distribution. Using both lets us ask whether a direction is tied to one dataset's surface statistics or is stable across factual QA formats. The datasets were difficulty-filtered to preserve variation in correctness and confidence, not to make direct-task accuracy exactly 50%. The direct-task MC accuracies are reported in Table 1. We restrict the main evidence chain to these two datasets because both expose measurable delegation and stated-confidence readouts under the current protocol. A SimpleMC rerun under the current delegation pipeline did not produce a usable delegation readout, so it is treated as a scope result rather than a third positive replication.

### 2.2. Task prompts, positions, and readouts

Direct MC task. This task provides the source representation. The model answers a question by choosing one of A/B/C/D. We extract direct-task activations at the final pre-answer token and compute uncertainty metrics from the answer-option logits:

$$\text{logit_gap} = \ell_{(1)} - \ell_{(2)}, \qquad \text{top_logit} = \ell_{(1)}, \tag{1}$$

and entropy over the four-way answer softmax. Lower logit gap and top logit, and higher entropy, indicate greater uncertainty.

Delegation task. This task tests whether direct-task uncertainty information is available to a policy decision. The model chooses between answering itself and delegating after seeing the question and answer options. The main delegation readout is the answer-vs-delegate logit margin,

$$\text{logit_margin} = z_{\text{Answer}} - z_{\text{Delegate}}. \tag{2}$$

We use logit margin rather than probability as the primary readout because answer probabilities saturate in this task, while a logit difference preserves graded variation in the model's preference to answer versus delegate.

Stated-confidence task. This task removes the extra answer-versus-defer decision. The model reports confidence using eight ordered bins, and we convert the output distribution over bins into a confidence-bin score. It is therefore a cleaner readout of whether the direct-task uncertainty direction transfers to an explicit confidence judgment. Delegation asks both "am I uncertain?" and "should I defer?"; stated confidence is closer to the first question alone.

For meta-task prompts, we extract activations at four positions: the question mark, the newline after the question, the newline after the final answer option (options_newline), and the final token before the model's response. The main delegation transfer result uses options_newline, where the question and options have both been read. The stated-confidence result uses the final position. Exact prompt templates, position definitions, and bin mappings are given in Appendix B.

### 2.3. Direction estimation and transfer

For every layer $\ell$, we extract the residual stream activation $h_\ell$ and construct a separate direction for that layer. This matters because the paper is not claiming a single vector passes through the whole network; it is testing whether corresponding layerwise uncertainty directions can be learned, transferred, interpreted, and intervened on. We use two direction families.

$$\text{mean_diff}: \quad d_\ell = \mu_\ell^{\text{high}} - \mu_\ell^{\text{low}}, \tag{3}$$

where high and low groups are the top and bottom quartiles of the relevant direct-task metric on the training split. We also train ridge probes on standardized, PCA-reduced direct-task activations: activations are standardized and projected onto 100 PCA components, then a ridge model is fit with $\alpha = 1000$. We then back-project the learned weights through the PCA basis, transform them back through the feature scaling, and denote the resulting residual-space coefficient by $w_\ell$. For notation, we write the residual-space linear readout as

$$\hat{y} = w_\ell^\top h_\ell + b, \tag{4}$$

where the bias term is used when fitting and evaluating scalar predictions, but not treated as part of the cross-prompt representation. Direction signs follow the underlying metric: larger logit_gap and top_logit indicate higher confidence, while larger entropy indicates higher uncertainty. Direction-based analyses use the unit-normalized residual-space coefficient, which specifies an axis corresponding to variation in the relevant direct-task metric.

For transfer analyses, we freeze the direct-QA direction and apply it to held-out activations from the meta-task prompt for the same question. Centered D→M $R^2$ asks whether the direct-task uncertainty target can still be reconstructed after accounting for prompt-level mean shifts; it is therefore a test of geometry reuse, not a refit on the meta-task. We separately report behavioral coupling between direct-task uncertainty and the relevant meta-task readout. Thus transfer and behavior answer different questions: whether the geometry is reused, and how strongly a particular prompt reads it out.

### 2.4. Interventions and interpretability

Transfer alone could still reflect a passive correlation rather than a direction that changes the readout under intervention. We therefore run two intervention families on held-out examples, following the broader contrastive activation steering tradition (Panickssery et al., 2024). For necessity, we ablate the directional component from the residual stream and measure the change in correlation between direct-task uncertainty and the delegation readout, logit_margin. For sufficiency, we add multiples $\alpha \hat{d}_\ell$ of the unit-normalized direction, with $\alpha \in \{-3, -2, -1, 0, 1, 2, 3\}$, and estimate the slope of logit_margin with respect to $\alpha$. Random controls are sampled as unit directions and orthogonalized against the learned direction at the same layer. We distinguish pooled-control FDR from stricter permutation-FDR; Appendix D gives the statistical criteria and non-replication checks.

For interpretability, we project directions through the unembedding matrix and inspect the top positive tokens. This logit-lens analysis asks what output tokens would be favored if the model read out that direction directly as logits. Because this is a direction-only diagnostic rather than a full causal logit attribution, we interpret it only as output-space semantic evidence. We report normalized top-$k$ abstention mass over the saved top-12 positive tokens, not full-vocabulary probability mass. We focus on answer tokens (A/B/C/D), abstention tokens (none, None, neither, Neither, etc.), and delegate-related tokens. We compare the learned direction with random directions and the sign-flipped direction to check whether the abstention signature is direction-specific.

## 3. Results

The results follow the logic of the paper's central test. We first ask whether direct-task uncertainty is linearly readable at all. We then ask whether the same directions transfer to metacognitive readouts, whether they are shared across datasets, whether they have interpretable output-space semantics, and finally whether perturbing them changes delegation behavior. This ordering matters: probe accuracy alone would be weak evidence, but transfer, cross-dataset alignment, semantic specificity, and interventions together provide a stronger case for a reusable uncertainty-related representation.

### 3.1. Direct-task uncertainty is linearly readable

Direct-task uncertainty is strongly decodable in late layers on both primary datasets (Table 1; full native

Table 1. Direct-task representation strength on the two primary datasets. Values are best held-out layerwise scores.

| Metric | TriviaMC | PopMC |
|---|---|---|
| MC accuracy | 0.552 (276/500) | 0.634 (317/500) |
| logit_gap probe $R^2$ | 0.932 (L77) | 0.939 (L77) |
| top_logit probe $R^2$ | 0.958 (L78) | 0.962 (L77) |
| entropy probe $R^2$ | 0.289 (L75) | 0.346 (L77) |
| logit_gap mean-diff $R^2$ | 0.894 (L75) | 0.892 (L74) |
| Chosen-answer probe accuracy | 0.93 (L77) | 0.97 (L42) |

curves in Appendix Figure 8). This first result is necessary but not sufficient for the paper's claim: it establishes that ordinary QA contains a recoverable uncertainty-related direction, but it does not yet show transfer to metacognitive readouts. The strongest probe directions for logit gap and top logit are late-layer directions, while answer identity is also highly decodable. Here, "chosen answer" means the model's selected A/B/C/D option, not ground-truth correctness. Because answer identity is decodable, direct-task decodability alone is not treated as evidence for operational reuse; the representation claim rests on transfer, cross-dataset alignment, readout dissociation, output-space semantics, and interventions.

### 3.2. Direct-task directions transfer to metacognitive readouts

The central test is whether a direction learned from direct QA remains informative in a different prompt. Direct-task logit-gap and top-logit directions transfer to both delegation and stated confidence (Table 2, Figure 2). The effect is strongest for logit gap and top logit: on TriviaMC, top-logit directions reach D→M $R^2 = 0.405$ for delegation and 0.577 for stated confidence; on PopMC, the corresponding values are 0.327 and 0.373. The strongest result is cross-meta-task generalization: a direction fitted from direct MC activations remains predictive in stated-confidence prompts without being trained on stated-confidence examples. Appendix A gives the corresponding native layerwise plots behind the compact summaries.

### 3.3. Stated confidence separates representation from delegation readout

The stated-confidence result changes the interpretation of PopMC. In the delegation task, PopMC has weaker behavioral coupling than TriviaMC. In the stated-confidence task, however, PopMC coupling is strong: logit-gap $r = 0.511$ and top-logit $r = 0.544$. Thus the weak PopMC delegation coupling is not strong evidence that the uncertainty-related representation is absent. Instead, the answer-vs-delegate logit_margin appears to introduce additional task-specific variance, likely because delegation depends not only on the model's own uncertainty but also on implicit assumptions about the teammate.

### 3.4. Directions align across datasets

If the transfer result were driven by dataset-specific artifacts, directions learned on TriviaMC and PopMC need not point in the same direction. Instead, the direction families are geometrically aligned across the two datasets (Figure 2). For mean-diff directions, cross-dataset cosine similarity is high in the transfer and intervention-sensitive bands: logit-gap mean-diff cosine is 0.864 at L31, 0.875 at L33, and 0.885 at L42; entropy mean-diff cosine is 0.862 at L31, 0.875 at L33, and 0.896 at L42. In each case, permutation $p = 0.0099$ with 100 permutations.

### 3.5. Logit lens reveals abstention-related semantics

Transfer and cross-dataset alignment support the operational reuse claim; logit lens asks what the direction is aligned with in output space. The clearest semantic result appears around L41–L43, the same band where final-position ablation effects peak. The top positive logit-lens tokens of the delegation uncertainty direction are dominated by abstention words (Table 3; native heatmaps in Appendix Figure 7). On PopMC at L42, the delegation mean-diff entropy direction has normalized top-$k$ abstention mass 0.862 on neither/Neither variants among the saved top positive tokens. TriviaMC shows the same pattern at L42 with normalized top-$k$ abstention mass 0.638.

Direction controls make the interpretation stronger. At L42, the positive direction has abstention mass 0.638 on TriviaMC and 0.862 on PopMC, while the sign-flipped direction has abstention mass 0.0. Matched random directions do not reproduce the effect: none of 100 random directions exceed an abstention-mass threshold of 0.30 in either dataset. This supports a narrow specificity claim: the abstention signature is specific to this direction and sign, not a generic property of the layer.

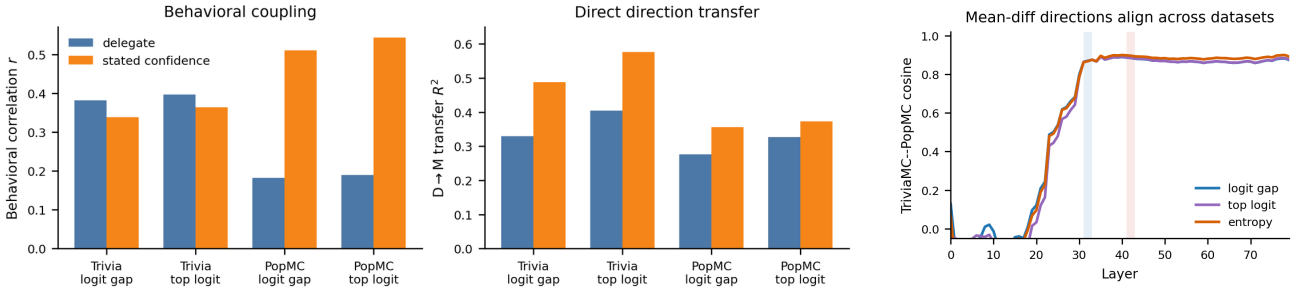
Because all tasks use multiple-choice formatting, we interpret neither/None tokens as an abstention-related signature within this task family, not as evidence for a task-general abstention concept.

### 3.6. Steering supports sufficiency; ablation gives limited necessity evidence

The intervention results are best read as asymmetric: sufficiency is strong and stable; necessity is selective and weaker under strict permutation criteria

Table 2. Cross-meta-task generalization. Delegation transfer uses the options-boundary position; stated-confidence transfer uses the final position. D→M values are centered probe $R^2$ for reconstructing the direct-task metric from held-out meta-task activations; behavioral columns report correlations with the corresponding meta-task readout. Confidence intervals are omitted from the compact table for formatting consistency; the native transfer plots are shown in Appendix A.

| Dataset | Metric | Delegation behavioral $r$ | Stated-confidence behavioral $r$ | Delegation D→M $R^2$ | Stated-confidence D→M $R^2$ |
|---|---|---|---|---|---|
| TriviaMC | logit_gap | 0.382 | 0.338 | 0.330 (L33) | 0.488 (L32) |
| TriviaMC | top_logit | 0.397 | 0.364 | 0.405 (L33) | 0.577 (L32) |
| TriviaMC | entropy | 0.159 | 0.207 | 0.050 (L32) | 0.084 (L79) |
| PopMC | logit_gap | 0.182 | 0.511 | 0.276 (L32) | 0.356 (L38) |
| PopMC | top_logit | 0.189 | 0.544 | 0.327 (L32) | 0.373 (L38) |
| PopMC | entropy | 0.041 | 0.196 | 0.060 (L32) | 0.066 (L33) |




(a) Stated confidence is cleaner than delegation

(b) Cross-dataset direction similarity

Figure 2. Compact summary of cross-meta-task generalization and cross-dataset geometry. Panel (a) shows that stated confidence is a cleaner readout than delegation, comparing behavioral coupling and D→M transfer for the two primary direct-task uncertainty metrics. Exact values are reported in Table 2; this panel omits interval bars to preserve readability across metrics and readouts. Panel (b) shows that TriviaMC–PopMC mean-diff direction similarity rises in the same mid-layer band used by the transfer and intervention analyses. Native transfer plots with plotting-script uncertainty bands are shown in Appendix A.

Table 3. Top abstention-token logit-lens rows. Mass is normalized within the saved top-$k$ positive logit-lens tokens, not over the full vocabulary.

| Dataset | Direction/layer | Normalized top-$k$ abstention mass | Dominant tokens |
|---|---|---|---|
| PopMC | delegation uncertainty, L42 | 0.862 | neither, Neither |
| PopMC | delegation uncertainty, L41 | 0.861 | neither, Neither |
| PopMC | delegation uncertainty, L43 | 0.782 | neither, Neither |
| TriviaMC | delegation uncertainty, L42 | 0.638 | neither, Neither |
| TriviaMC | delegation uncertainty, L41 | 0.569 | neither, Neither |
| TriviaMC | delegation uncertainty, L43 | 0.490 | neither, Neither |
| PopMC | direct entropy, L78 | 0.695 | None, none |

(Figures 3 and 4). Final-position steering over layers 20–50 is robust in both datasets. Mean-diff steering peaks at slope 1.152 at L33 on TriviaMC and 1.047 at L33 on PopMC, with 21 FDR-significant layers in each dataset. Probe steering is also positive and layer-localized. Because the transfer analysis peaks earlier at the options boundary, we additionally ran a targeted options-boundary steering check. Those results were sign-consistent and peaked at L31 in both datasets, but did not survive FDR correction; we report them as a robustness analysis in Appendix D.2.

Ablation provides more cautious evidence. Final-position mean-diff ablation reduces the direct-to-delegation correlation by $\Delta r = -0.0514$ at L43 on TriviaMC and $\Delta r = -0.0427$ at L43 on PopMC. These effects are large relative to matched random controls, but the strict permutation-FDR analysis is weaker than pooled-control FDR. At the options boundary, TriviaMC gives the clearest selective necessity evidence. PopMC does not cleanly replicate this options-boundary ablation result: the native PopMC-selected layers under a relaxed transfer threshold ($R^2 \approx 0.25$, L30–L32) are better treated as a sensitivity analysis, and the Trivia-style layer set is not used as a main PopMC result because it imports Trivia-selected layers and includes a sign-reversed dominant effect at L28. We therefore describe ablation as limited selective evidence for necessity rather than as evidence that this direction is the only route by which the model delegates.

## 4. Related Work

Prior work on confidence estimation and its faithfulness is reviewed in the introduction. Our results are convergent with causal evidence that confidence-related representations can guide answer-versus-abstain behavior under steering (Kumaran et al., 2026b), but ask a different question: whether directions learned from ordinary direct QA transfer to delegation and stated confidence.

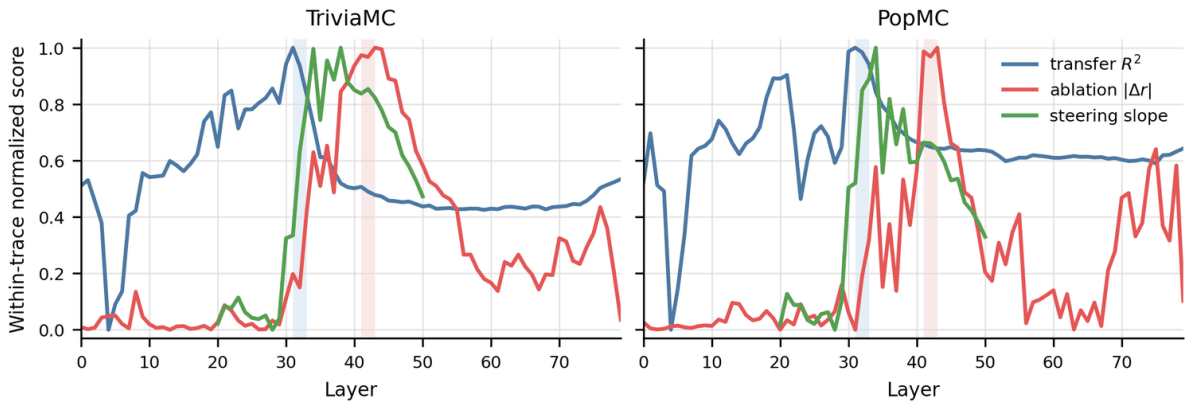



Figure 3. Layer localization of transfer, steering, and ablation effects. This figure is a compact locator plot, not a raw effect-size plot: each trace is normalized within dataset and quantity so that layer locations, rather than units, can be compared. Blue shows direct-to-delegation transfer at the options boundary; green shows final-position steering slopes; red shows the magnitude of final-position ablation effects. Peak values are reported in text and tables: top-logit delegation transfer reaches $R^2 = 0.405$ on TriviaMC and 0.327 on PopMC, mean-diff steering peaks at slopes 1.152 and 1.047, and final-position mean-diff ablation peaks at $\Delta r = -0.0514$ and $-0.0427$. The highlighted bands are descriptive summaries chosen from the observed layerwise curves; they organize convergent effects rather than serving as pre-specified layer hypotheses. Appendix A gives the underlying experiment-generated plots.

Our method also builds on representation-engineering and activation-steering work (Zou et al., 2025; Panickssery et al., 2024); the linear-direction precedents and the probe-control caution that motivate our multi-pronged evidence standard are reviewed in the introduction.

Confidence regulation work has identified neurons that modulate output entropy while preserving the top prediction (Stolfo et al., 2024), and related studies examine model metacognition, self-modeling, activation control, and introspection benchmarks (Ackerman, 2025; Betley et al., 2025; Ji-An et al., 2025; Naphade et al., 2026). We differ by asking whether a QA-derived uncertainty direction remains available across multiple metacognitive readouts.

## 5. Discussion

The main lesson is that representation and readout should be separated. TriviaMC and PopMC both show direct-task uncertainty directions that are readable, aligned across datasets, and predictive of stated confidence. Delegation is more fragile: PopMC has weaker delegation coupling even though stated-confidence transfer remains strong. Delegation is not a pure uncertainty report; it adds a policy question about whether uncertainty should lead the model to defer, consistent with broader evidence that abstention remains difficult even for strong reasoning models (Kirichenko et al., 2025).

This interpretation is consistent with recent abstention work in which confidence-related representations and a separate answer-versus-abstain policy jointly determine behavior (Kumaran et al., 2026b). Our stated-confidence/delegation dissociation gives a complementary version of that picture: the representation can remain accessible even when a particular delegation prompt reads it out weakly.

The layerwise pattern suggests a staged computation. Direct-task uncertainty is most strongly decodable in late layers, direct-to-delegation transfer appears earlier around the options boundary, and abstention-related logit-lens semantics plus final-position ablation effects peak later around L41–L43. A natural hypothesis is that the model forms an answer-confidence representation during option integration, transforms it into a policy-relevant representation for non-answer behavior, and later routes that representation toward output logits. This is not yet a circuit-level explanation, but it explains why transfer, semantic readout, steering, and ablation need not peak at the same layer.

The evidence is strongest as a bundle, not as a single probe score: direct-to-meta-task transfer, cross-dataset alignment, stated-confidence/delegation dissociation, logit-lens semantics, direction controls, and steering. The claim remains narrow: the directions may reflect answer confidence or answerability rather than human-like uncertainty, and ablation is weaker than steering.

This intervention asymmetry supports the same interpretation: steering shows that the directions are sufficient to influence delegation, while ablation does not

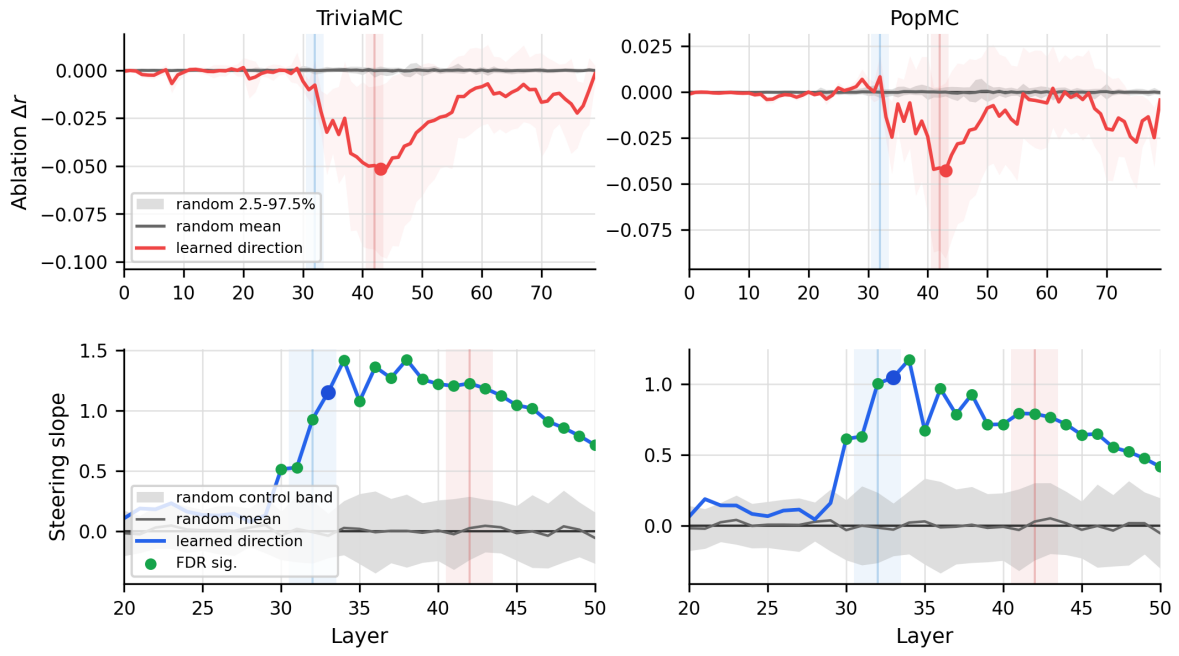



Figure 4. Final-position intervention diagnostics for the mean-diff logit-gap direction. The top row shows ablation across all layers; the bottom row shows steering only over L20–L50, the layer band actually tested in the steering run. Top row: ablation effect relative to matched random-direction controls. Bottom row: steering slope of the delegation readout, logit_margin, versus steering coefficient $\alpha$. Compact legends are shown in the left panels; the same color coding applies to both datasets. Larger dots mark the peak learned-direction effects; numerical peak values are reported in the text. Light blue and red reference bands mark L31–L33 transfer and L41–L43 logit-lens/ablation regions, not exact peak markers. Native experiment-generated intervention plots are shown in Appendix Figure 9.

show that they are uniquely necessary.

## 6. Limitations

This study is limited to one model family and primarily to factual multiple-choice QA. The direct-task metrics are answer-logit based, so we describe the resulting directions as uncertainty-related rather than as a complete measure of introspective uncertainty. A key confound is that answer-logit uncertainty metrics are entangled with answer selection. High logit gap and top logit partly reflect how strongly the model has settled on a particular A/B/C/D answer, and the model's chosen answer is itself highly decodable from the same activations. We therefore do not interpret direct-task decodability as evidence for a pure uncertainty variable. Our claim is narrower: QA-derived answer-confidence directions transfer across prompts, align across datasets, acquire abstention-related output semantics, and can influence delegation under steering. A stronger control would factor out answer-identity subspaces and test whether the residual uncertainty component still transfers.

Stated confidence is still prompted self-report, and logit-lens token-family masses are computed over saved top positive tokens rather than exhaustive semantic clusters. We do not report within-meta-task probe ceilings, so D→M $R^2$ should be interpreted as absolute transfer strength rather than as a fraction of the best possible readout at that prompt position. The causal evidence is asymmetric: steering is strong, while ablation is weaker under strict permutation-FDR. We also use a single train-test split and seed; future work should test whether the layer locations and transfer strengths are stable across splits, seeds, and model families.

Finally, the SimpleMC scope check (Appendix D.3) shows that direct-task decodability alone is not enough if the meta-task readout does not couple to uncertainty. This is consistent with concurrent abstention work: Kumaran et al. report excluding Llama 70B Instruct 3.1 because it abstained on only 4% of phase-2 trials, and broader work shows that abstention and prompted behavior vary substantially across models and prompt formats (Kumaran et al., 2026b; Wen et al., 2025; Sclar et al., 2024).

## 7. Conclusion

We tested whether uncertainty-related directions learned from ordinary QA are reused in metacognitive behavior. In Llama-3.3-70B-Instruct, these directions transfer across datasets and readouts, acquire abstention-related output semantics, and can be steered to shift delegation behavior. This is not a complete theory of introspection; it is evidence that direct-task uncertainty directions can be reused across metacognitive contexts while their behavioral expression remains readout-dependent.

## Impact Statement

This work aims to improve understanding of uncertainty-related internal structure in language models. Better understanding of when models can report uncertainty, abstain, or defer could support safer deployment in settings where unreliable answers are harmful. The same methods could also be misused to manipulate model behavior or to overstate how well model uncertainty is understood. We therefore frame the claims narrowly, report non-replications and limitations, and do not claim a general measure of introspection or reliability.

## References

Ackerman, C. Evidence for Limited Metacognition in LLMs, September 2025. URL http://arxiv.org/abs/2509.21545. arXiv:2509.21545 [cs].

Ackerman, C. and Panickssery, N. Inspection and Control of Self-Generated-Text Recognition Ability in Llama3-8b-Instruct, April 2025. URL http://arxiv.org/abs/2410.02064. arXiv:2410.02064 [cs].

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in Language Models Is Mediated by a Single Direction. Advances in Neural Information Processing Systems, 37:136037–136083, December 2024. doi: 10.52202/079017-4322.

Betley, J., Bao, X., Soto, M., Sztyber-Betley, A., Chua, J., and Evans, O. Tell me about yourself: LLMs are aware of their learned behaviors, January 2025. URL http://arxiv.org/abs/2501.11120. arXiv:2501.11120 [cs].

Binder, F. J., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M., and Evans, O. Looking Inward: Language Models Can Learn About Themselves by Introspection, October 2024. URL http://arxiv.org/abs/2410.13787. arXiv:2410.13787 [cs].

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering Latent Knowledge in Language Models Without Supervision, March 2024. URL http://arxiv.org/abs/2212.03827. arXiv:2212.03827 [cs].

Comsa, I. M. and Shanahan, M. Does It Make Sense to Speak of Introspection in Large Language Models?, June 2025. URL http://arxiv.org/abs/2506.05068. arXiv:2506.05068 [cs].

Hewitt, J. and Liang, P. Designing and Interpreting Probes with Control Tasks. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Ji-An, L., Xiong, H.-D., Wilson, R. C., Mattar, M. G., and Benna, M. K. Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations, October 2025. URL http://arxiv.org/abs/2505.13763. arXiv:2505.13763 [cs].

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language Models (Mostly) Know What They Know, November 2022. URL http://arxiv.org/abs/2207.05221. arXiv:2207.05221 [cs].

Kirichenko, P., Ibrahim, M., Chaudhuri, K., and Bell, S. J. AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions, June 2025. URL http://arxiv.org/abs/2506.09038. arXiv:2506.09038 [cs].

Kumar, A., Morabito, R., Umbet, S., Kabbara, J., and Emami, A. Confidence Under the Hood: An Investigation into the Confidence-Probability Alignment in Large Language Models, June 2024. URL http://arxiv.org/abs/2405.16282. arXiv:2405.16282 [cs].

Kumaran, D., Conmy, A., Barbero, F., Osindero, S., Patraucean, V., and Velickovic, P. How do LLMs

Compute Verbal Confidence, March 2026a. URL http://arxiv.org/abs/2603.17839. arXiv:2603.17839 [cs].

Kumaran, D., Daw, N., Osindero, S., Velickovic, P., and Patraucean, V. Causal Evidence that Language Models use Confidence to Drive Behavior, 2026b. URL https://arxiv.org/abs/2603.22161.

Lin, S., Hilton, J., and Evans, O. Teaching Models to Express Their Uncertainty in Words, June 2022. URL http://arxiv.org/abs/2205.14334. arXiv:2205.14334 [cs].

Lyu, Q., Shridhar, K., Malaviya, C., Zhang, L., Elazar, Y., Tandon, N., Apidianaki, M., Sachan, M., and Callison-Burch, C. Calibrating Large Language Models with Sample Consistency, February 2024. URL http://arxiv.org/abs/2402.13904. arXiv:2402.13904 [cs].

Manakul, P., Liusie, A., and Gales, M. J. F. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models, October 2023. URL http://arxiv.org/abs/2303.08896. arXiv:2303.08896 [cs].

Marks, S. and Tegmark, M. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets, August 2024. URL http://arxiv.org/abs/2310.06824. arXiv:2310.06824 [cs].

Naphade, A., Bhargav, S., Lim, S., and Shah, M. Me, Myself, and $\pi$: Evaluating and Explaining LLM Introspection, 2026. URL https://arxiv.org/abs/2603.20276.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering Llama 2 via Contrastive Activation Addition, July 2024. URL http://arxiv.org/abs/2312.06681. arXiv:2312.06681 [cs].

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting, July 2024. URL http://arxiv.org/abs/2310.11324. arXiv:2310.11324 [cs].

Song, S., Hu, J., and Mahowald, K. Language Models Fail to Introspect About Their Knowledge of Language, September 2025. URL http://arxiv.org/abs/2503.07513. arXiv:2503.07513 [cs].

Stolfo, A., Wu, B., Gurnee, W., Belinkov, Y., Song, X., Sachan, M., and Nanda, N. Confidence Regulation Neurons in Language Models, November 2024. URL http://arxiv.org/abs/2406.16254. arXiv:2406.16254 [cs].

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In Bouamor, H., Pino, J., and Bali, K. (eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., and Wang, L. L. Know Your Limits: A Survey of Abstention in Large Language Models, February 2025. URL http://arxiv.org/abs/2407.18418. arXiv:2407.18418 [cs].

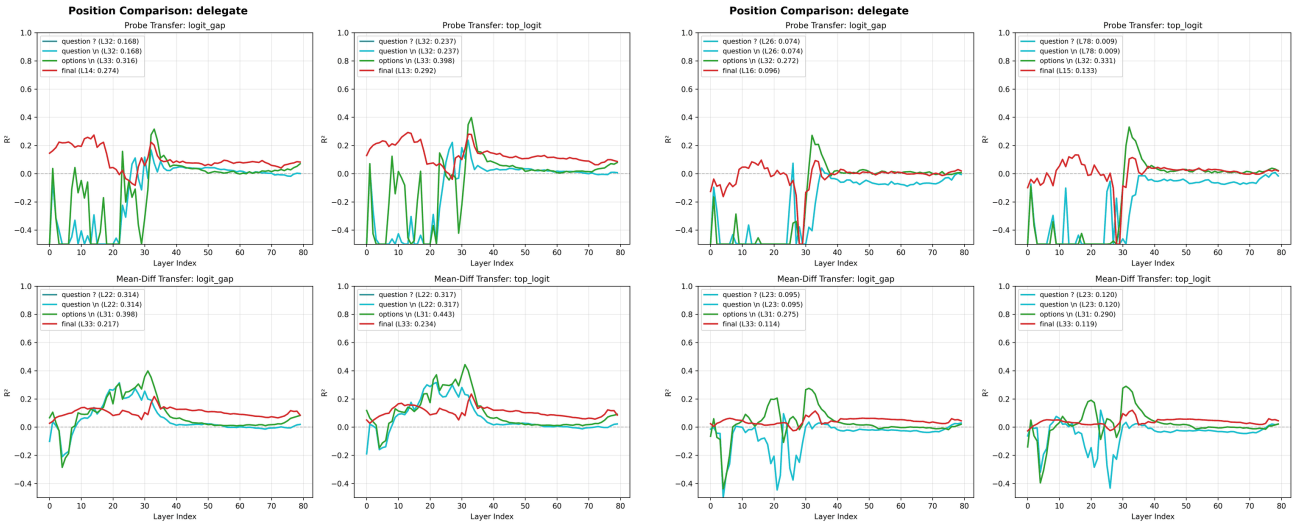
Yona, G., Aharoni, R., and Geva, M. Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words? In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 7752–7764, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.443. URL https://aclanthology.org/2024.emnlp-main.443/.

Zhang, H., Diao, S., Lin, Y., Fung, Y. R., Lian, Q., Wang, X., Chen, Y., Ji, H., and Zhang, T. R-Tuning: Instructing Large Language Models to Say 'I Don't Know', June 2024. URL http://arxiv.org/abs/2311.09677. arXiv:2311.09677 [cs].

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation Engineering: A Top-Down Approach to AI Transparency, March 2025. URL http://arxiv.org/abs/2310.01405. arXiv:2310.01405 [cs].

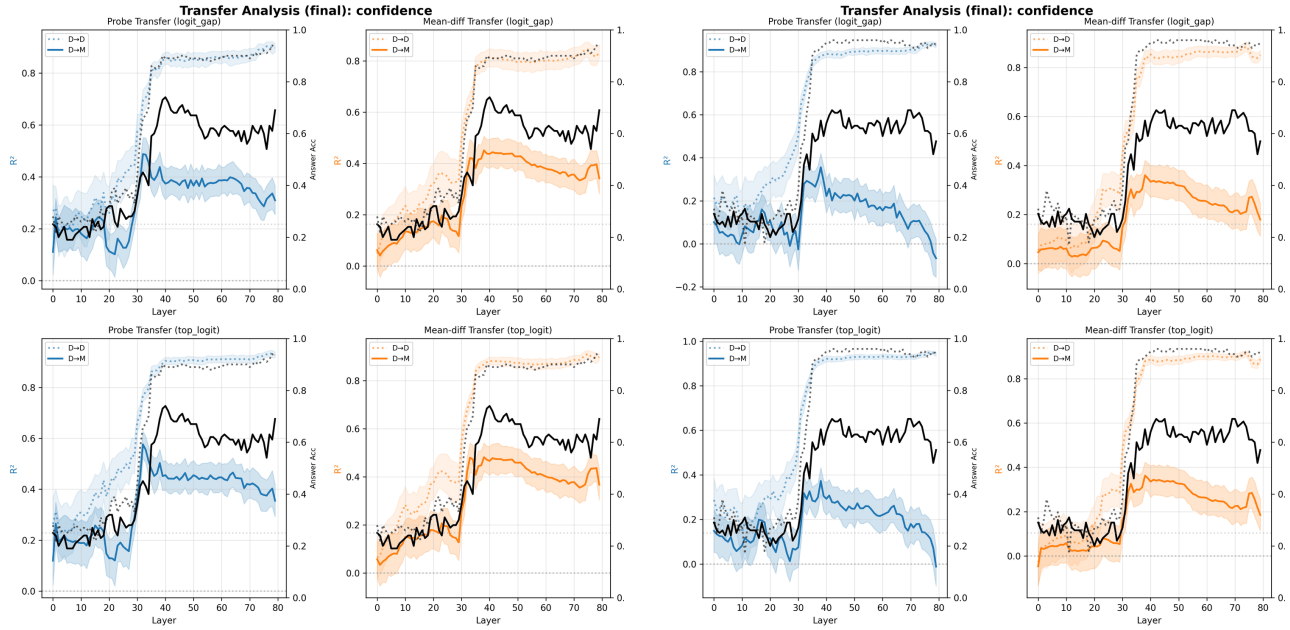
# A. Additional Experiment-Generated Figures

The appendix records audit details that would make the main evidence chain too dense: native transfer curves, full direct-task QA plots, the PopMC ablation limitation, and the SimpleMC scope result.



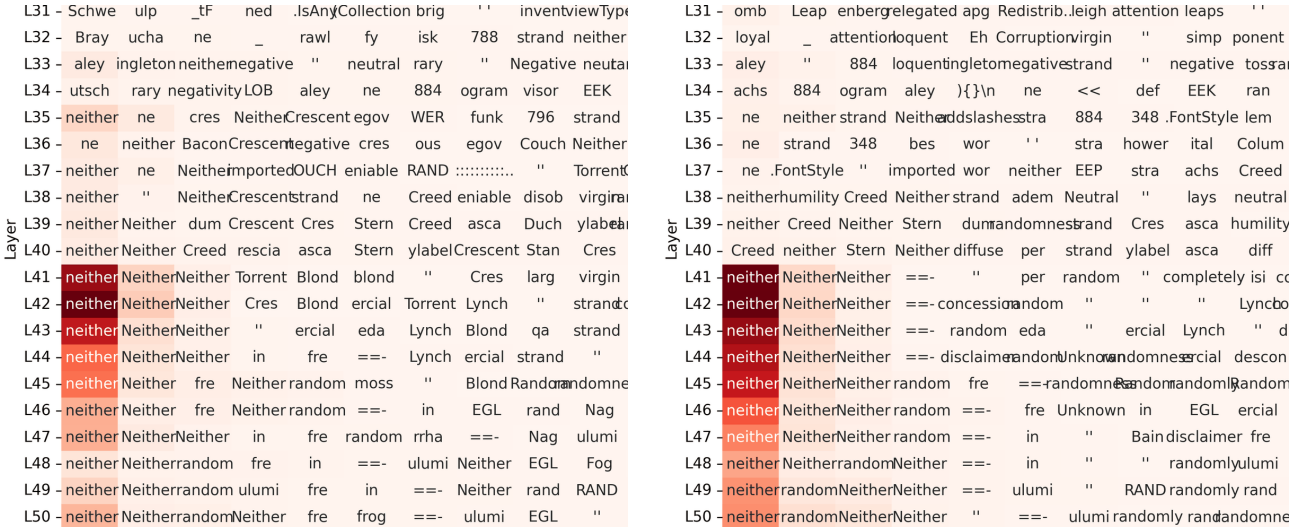

(a) TriviaMC delegation transfer (b) PopMC delegation transfer

Figure 5. Native delegation-transfer position plots for logit-gap and top-logit directions. These plots support the blue transfer trace in Figure 3. In both datasets, transfer is strongest after the model has seen the answer options, rather than immediately after the question text; PopMC is weaker under the delegation readout.



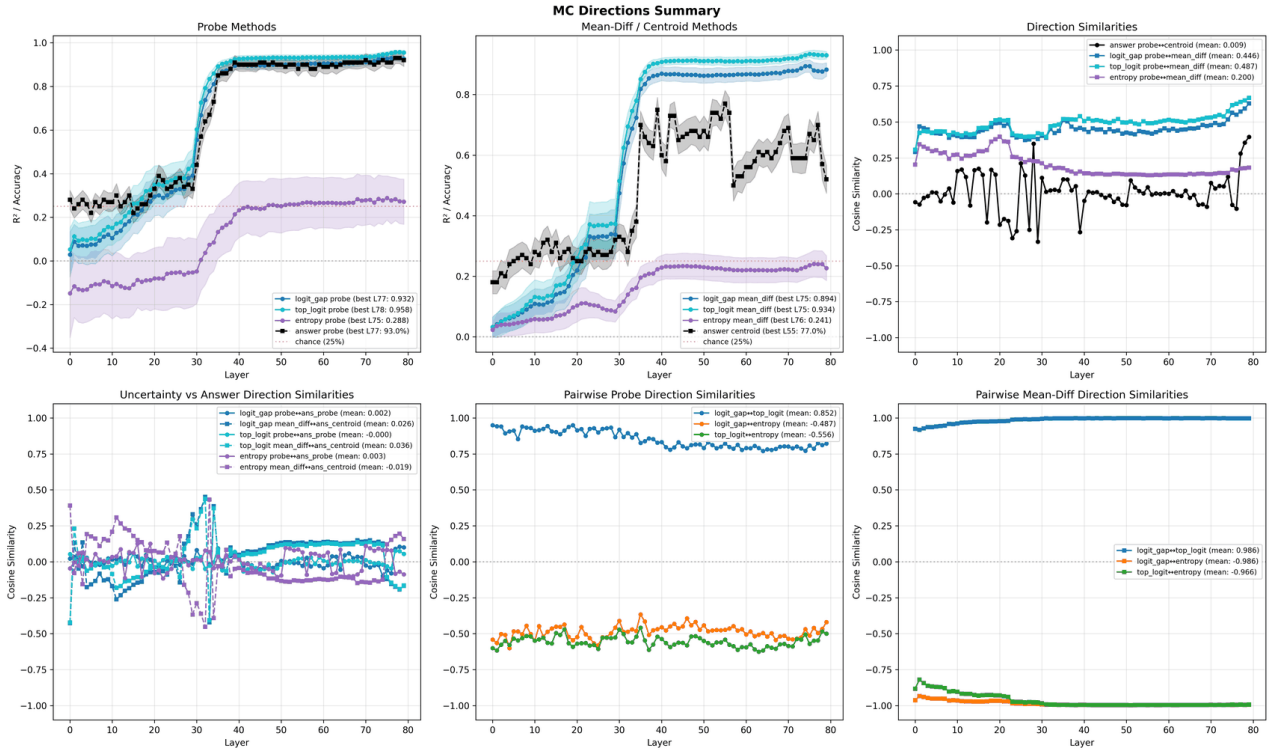

(a) TriviaMC stated-confidence transfer (b) PopMC stated-confidence transfer

Figure 6. Native stated-confidence transfer plots. These are the detailed curves behind the stated-confidence part of Figure 2. Compared with Figure 5, PopMC is substantially stronger under explicit stated confidence than under delegation.

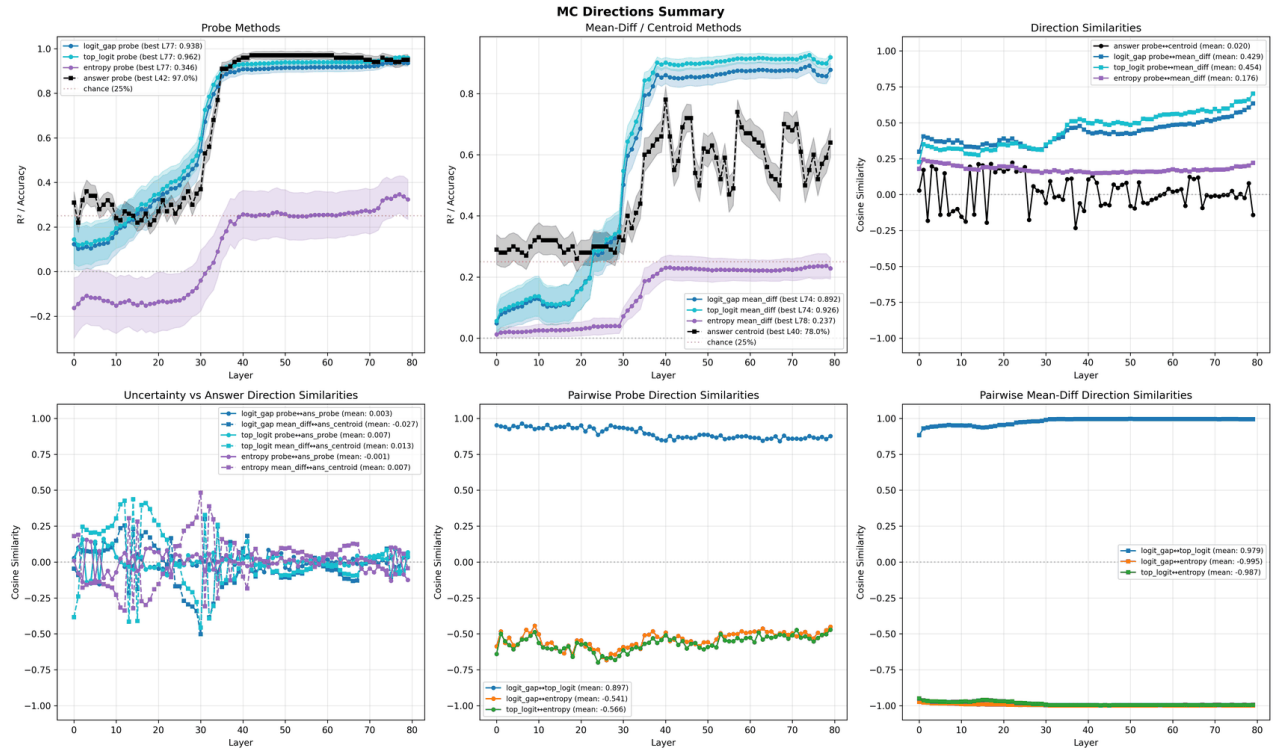



(a) TriviaMC (b) PopMC

Figure 7. Native logit-lens heatmaps for the delegation uncertainty mean-diff entropy direction, cropped to layers L31–L50 and the highest-ranked positive-token columns. The main paper reports the readable normalized top-$k$ abstention masses in Table 3; these heatmaps are included as audit evidence for the layer-localized token pattern.

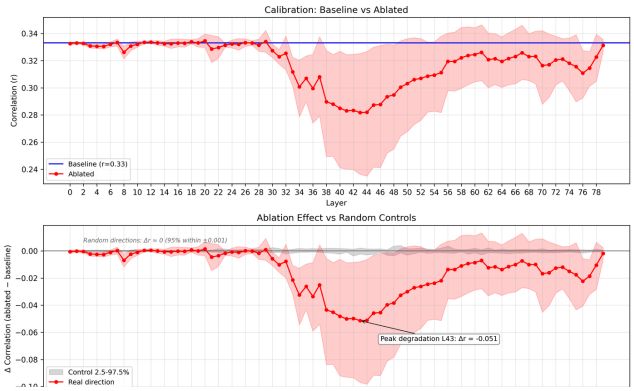



(a) TriviaMC direct-task QA

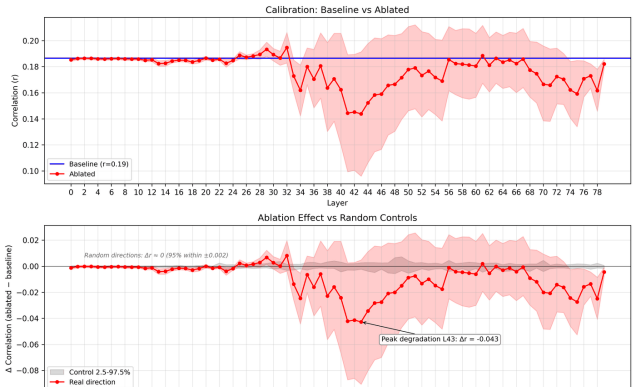



(b) PopMC direct-task QA

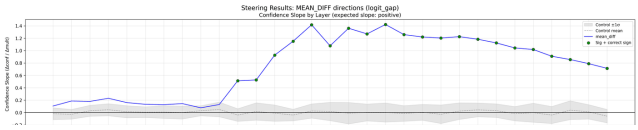
Figure 8. Full native direct-task QA direction plots. The main text reports the compact peak values in Table 1. The full plots show the broader pattern: direct-task logit-gap and top-logit probes become strongly decodable in late layers, while answer identity is also decodable.



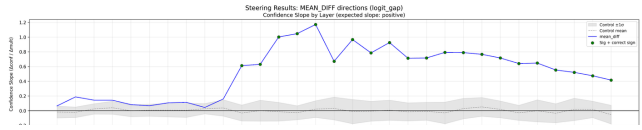

(a) TriviaMC final-position ablation




(b) PopMC final-position ablation




(c) TriviaMC final-position steering




(d) PopMC final-position steering

Figure 9. Native final-position intervention plots behind Figure 4. These include the baseline-versus-ablated rows omitted from the main figure, and are retained as audit evidence for the full experiment-generated curves and random-control comparisons.

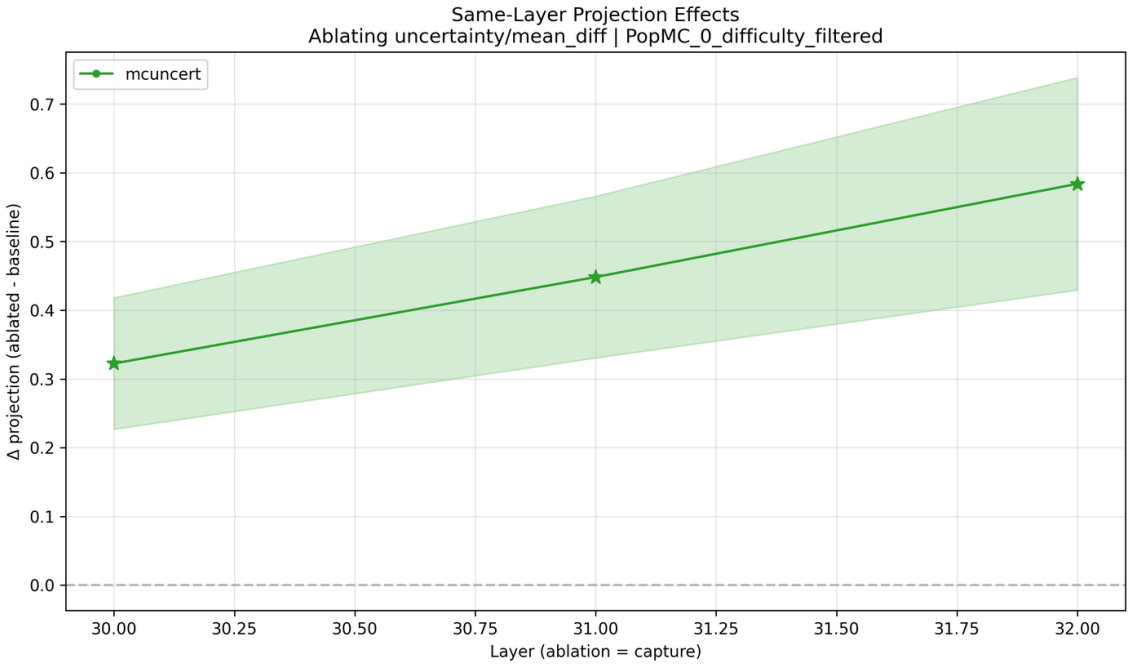
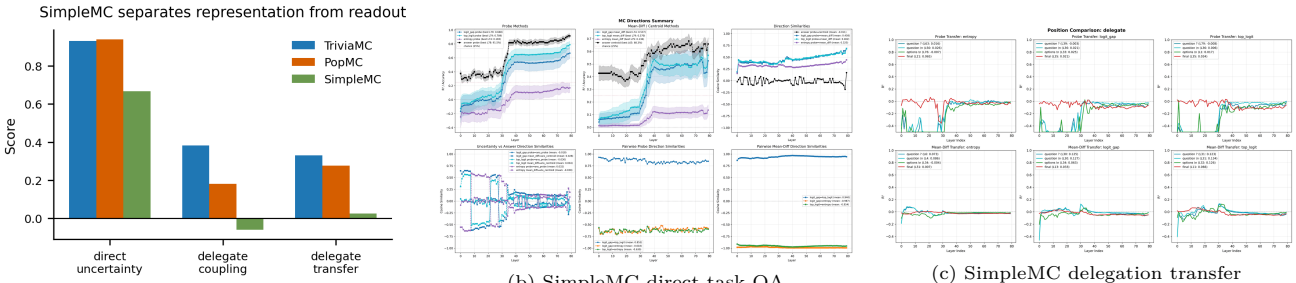



Figure 10. PopMC options-boundary ablation diagnostic, shown separately because it is the main failed-replication pressure point for the necessity claim. PopMC does not give a clean confirmatory options-boundary ablation result; the imported Trivia-style layer set includes a sign-reversed dominant effect at L28. This figure is therefore included as a limitation and sensitivity diagnostic, not as positive PopMC ablation evidence.




(a) Compact SimpleMC scope summary

(b) SimpleMC direct-task QA

(c) SimpleMC delegation transfer

Figure 11. SimpleMC scope result. SimpleMC has a partially decodable direct-task uncertainty direction, but the current delegation pipeline does not produce a usable delegation readout. The result is therefore informative as a limitation: direct uncertainty decodability alone is insufficient for the meta-task transfer story.

## B. Technical Definitions and Reproducibility Details

This appendix records the operational details that are needed to audit the Methods section without expanding the main paper. The central distinction is between direct-task confidence quantities, computed from A/B/C/D answer logits, and meta-task readouts, computed from delegation or stated-confidence prompts.

### B.1. Readout definitions

Table 4. Direct-task uncertainty metrics. These are used as scalar targets for fitting or deriving directions from direct multiple-choice activations.

| Metric | Definition | Interpretation |
|---|---|---|
| logit_gap | $\ell_{(1)} - \ell_{(2)}$ | Larger means more confident |
| top_logit | $\ell_{(1)}$ | Larger means more confident |
| Entropy | $-\sum_i p_i \log p_i$ over A/B/C/D | Larger means more uncertain |

Table 5. Meta-task readouts. These are the behavioral quantities used when evaluating whether direct-task uncertainty directions are read out in metacognitive prompts.

| Task | Readout | Role |
|---|---|---|
| Delegation | $z_{\text{Answer}} - z_{\text{Delegate}}$ | Primary delegation readout, logit_margin |
| Stated confidence | confidence-bin score | Explicit confidence readout |

### B.2. Prompts, positions, and split discipline

All prompts use the model tokenizer's chat template when available and read out a single next-token choice. The direct MC setup instruction is: "select the answer you think is best" and "respond only with the letter of your choice." The direct prompt then shows the question, A/B/C/D options, and the suffix Your choice (A, B, C, or D):. Direct-task activations are extracted at the final pre-answer token.

The stated-confidence prompt asks: "How confident are you that you know the correct answer to this question?" and uses eight bins: S=< 5%, T=5–10%, U=10–20%, V=20–40%, W=40–60%, X=60–80%, Y=80–90%, and Z=> 90%. The scalar confidence-bin score is the probability-weighted midpoint of these bins: 0.025, 0.075, 0.15, 0.30, 0.50, 0.70, 0.85, and 0.95.

The delegation prompt presents a game in which the model can answer itself or delegate to a teammate. The fixed teammate summary says that the teammate answered 50.0% of phase-1 questions correctly. The model

must respond with 1 or 2; the digit-to-action mapping alternates by trial index to control for position bias. The delegation readout is always mapped back to semantic actions before computing logit_margin: $z_{\text{Answer}} - z_{\text{Delegate}}$.

For meta-task activations, we extract four token positions: question_mark, the token at the final question mark of the embedded MC question; question_newline, the following newline; options_newline, the newline after the last answer option; and final, the last prompt token before the model response. The same seed-42 80/20 item split is used across direct and meta prompts. Direct-task directions and probe calibration parameters are fit only on direct-QA training examples. Meta-task activations and meta-task readouts are not used to fit the direct-task directions.

### B.3. Direction estimation and transfer scores

For each layer $\ell$, we compute a layer-specific direction $d_\ell$. A mean-difference direction uses

$$d_\ell = \mu_\ell^{\text{top } 25\%} - \mu_\ell^{\text{bottom } 25\%}, \tag{5}$$

where the top and bottom groups are selected by the relevant direct-task metric on the training split. For entropy, the top group is high entropy; for logit_gap and top_logit, the top group is high confidence. We keep this native metric orientation in saved directions and account for it when interpreting correlations and steering signs. Ridge probes are fit on standardized activations after reducing to 100 PCA components, with ridge regularization $\alpha = 1000$. Scalar predictions use the fitted scaler, PCA map, ridge weights, and intercept. For direction-based analyses, the ridge weights are inverse-transformed through the PCA basis, transformed back through the feature scaling, and unit-normalized to obtain a vector in residual-stream space. The intercept is not treated as part of the cross-prompt representation.

For probe D→M transfer, the probe is trained on direct-QA train activations and evaluated on held-out meta-task activations. We use centered scaling: meta activations are centered by their own mean but scaled by the direct-task training standard deviations before applying the direct-task PCA and ridge model. This asks whether the direction geometry survives the prompt-level mean shift between direct QA and the meta-task. We report centered held-out

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{6}$$

where $y_i$ is the held-out direct-task metric for the same question. Separately, we report behavioral coupling between the direct-task metric and the meta-task readout. This separation prevents a leakage-prone interpretation: D→M $R^2$ tests whether direct-task uncertainty geometry is preserved in the meta prompt, while behavioral $r$ tests how strongly that prompt reads the representation out into confidence or delegation.

### B.4. Interventions, controls, and logit lens

Steering adds a scaled unit direction to the residual stream,

$$h \leftarrow h + \alpha \hat{d}_\ell, \qquad \alpha \in \{-3, -2, -1, 0, 1, 2, 3\}, \tag{7}$$

and measures the slope of the delegation readout, logit_margin, with respect to $\alpha$. Ablation removes the component along a unit direction,

$$h \leftarrow h - (h^\top \hat{d}_\ell)\hat{d}_\ell, \tag{8}$$

and measures the change in correlation between direct-task uncertainty and logit_margin. Random controls are sampled as unit directions orthogonalized against the learned direction at the same layer; steering controls are additionally orthogonalized against previously sampled controls. Final-position intervention runs use 25 random controls per layer; non-final targeted runs use 10 controls per layer. Ablation reports bootstrap confidence intervals over questions and paired permutation tests with 2000 resamples.

For logit lens, we apply the final RMSNorm and unembedding to a direction and save the top 12 positive tokens. We report normalized top-$k$ abstention mass: the mass on abstention tokens within those saved top-12 positive tokens, not probability mass over the full vocabulary. The abstention family includes none, None, neither, and Neither variants. This quantity supports only a semantic-signature claim: it says the direction projects strongly onto abstention vocabulary, not that the direction is identical to an abstention variable.

## C. Direction Controls for Logit Lens

The logit-lens controls support a narrow specificity claim. At L42, the positive meta-task uncertainty direction has abstention mass 0.638 on TriviaMC and 0.862 on PopMC among the saved top positive logit-lens tokens. The sign-flipped direction has abstention mass 0.0, and none of 100 matched random directions exceed an abstention-mass threshold of 0.30 in either dataset. This supports the statement that the abstention-token signature is specific to the learned direction and sign; it does not imply that the negative direction is a clean answer-token axis.

## D. Robustness, Non-Replications, and Statistical Criteria

This appendix collects the results that constrain the main causal interpretation. They are not extra headline claims: they document where the evidence is weaker, which layer choices are sensitivity analyses, and how statistical criteria were applied.

### D.1. PopMC options-boundary ablation

The main ablation claim is intentionally cautious. TriviaMC provides the clearest options-boundary necessity evidence. PopMC does not provide a clean options-boundary replication. The native PopMC transfer peaks occur around L30–L32 under a relaxed transfer threshold near $R^2 \approx 0.25$, so those layers are best treated as a PopMC sensitivity analysis rather than as a confirmatory replication under the original Trivia-selected policy.

Table 6. How we interpret options-boundary ablation layer sets.

| Dataset / layer set | Layers | Interpretation |
|---|---|---|
| TriviaMC policy set | L22, L28, L30, L31, L32, L33 | Main selective necessity evidence |
| PopMC native relaxed set | L30, L31, L32 | Sensitivity analysis only |
| PopMC Trivia-style set | L22, L28, L30, L31, L32 | Not used as main PopMC result |

The reason to avoid the Trivia-style PopMC set in the main text is that it imports layers selected on TriviaMC and includes a dominant sign-reversed effect at L28 (Figure 10). This is useful as a failed-replication diagnostic, but not as a clean positive PopMC result. The main paper therefore states the stronger and more defensible conclusion: PopMC supports transfer, stated-confidence transfer, cross-dataset geometry, logit-lens semantics, and steering, but does not cleanly support options-boundary ablation necessity.

### D.2. Targeted options-boundary steering

The main-text steering result uses the final decision position, where effects are strongest and FDR-significant across both datasets. Because D→M transfer peaks earlier at the options boundary, we also ran a targeted options_newline steering check on the transfer-selected mid-layer bands. This analysis is included as supporting evidence rather than as a main causal claim: all effects are sign-consistent and peak at L31, but no layer survives FDR correction.

Table 7. Targeted options_newline steering results. Slopes measure the change in the delegation readout, logit_margin, per steering multiplier.

| Dataset | Method | Layers tested | Best layer | Slope | Effect $z$ | FDR-sig. layers |
|---|---|---|---|---|---|---|
| TriviaMC | mean-diff | 22, 28, 30, 31, 32, 33 | L31 | 0.0706 | 19.07 | 0 |
| TriviaMC | probe | 22, 28, 30, 31, 32, 33 | L31 | 0.0446 | 10.59 | 0 |
| PopMC | mean-diff | 30, 31, 32, 33 | L31 | 0.0868 | 19.68 | 0 |
| PopMC | probe | 30, 31, 32, 33 | L31 | 0.0255 | 4.36 | 0 |

Options-boundary steering is directionally consistent with the transfer band, but its evidential status differs from final-position steering. We therefore use final-position steering as the main sufficiency result and options_newline steering as a targeted sensitivity analysis.

### D.3. SimpleMC scope check

SimpleMC is omitted from the main evidence chain because the current delegation pipeline did not produce a usable delegation readout (Figure 11). In the balanced 408-item run, direct-task decodability was present but weaker than on TriviaMC and PopMC: logit-gap probe $R^2 = 0.668$, top-logit probe $R^2 = 0.789$, and entropy probe $R^2 = 0.183$. However, delegation coupling was absent or negative: logit-gap versus logit_margin $r = -0.061$, top-logit $r = -0.026$, and entropy $r = -0.035$.

After feedback that the older SimpleMC behavioral setup used the full set, we also reran the full raw 500 items under the current delegation pipeline. The same failure mode remained: logit-gap versus $P(\text{answer})$ was $r = -0.003$, and logit-gap versus logit_margin was $r = 0.001$, with the other direct confidence metrics also near zero. We therefore do not present SimpleMC as a third replication. The methodological lesson is that direct uncertainty decodability alone is not sufficient: the meta-task readout must also couple to uncertainty for transfer and intervention experiments to be meaningful.

This failure mode matches a broader readout-policy issue rather than a unique defect of SimpleMC. Kumaran et al. report that Llama 70B Instruct 3.1 was tested but excluded from their abstention analyses because its phase-2 abstention rate was only 4% (Kumaran et al., 2026b). This is a close analogue of our SimpleMC delegation result: the model can contain direct-task uncertainty structure while the offered answer-avoidance behavior has too little variance to support a transfer analysis. More generally, abstention propensity is known to vary with the query, model, and elicitation procedure (Wen et al., 2025), and model behavior can be sensitive to prompt formatting (Sclar et al., 2024). We therefore interpret SimpleMC as a boundary condition on the delegation readout, not as evidence that direct-task uncertainty structure is absent.

### D.4. Statistical reporting

We distinguish two FDR procedures. Permutation-FDR is the stricter primary criterion: labels are permuted to generate a null distribution for the relevant layerwise statistic, then false discovery rate correction is applied across the tested layer family. Pooled-control FDR is a secondary criterion based on the matched-control distribution pooled across layers. We use pooled-control FDR as supportive evidence when effects are large relative to controls, but we do not treat pooled-only effects as definitive confirmatory evidence. This distinction is why steering is described as the strongest causal evidence, while ablation is described as selective and limited.