Fine-Tuning Enhances Latent Metacognitive Capability in Language Models

Tristan Day [1 2] **Christopher Ackerman** [3]

Abstract

Large language models are increasingly asked not only to answer questions, but also to judge whether they know enough to answer. We test a *latent metacognitive capability* hypothesis: models already contain internal structure supporting weak self-evaluation of answer uncertainty, and later training routes this structure into confidence reports and delegation decisions. This predicts that self-evaluation should track direct-answer uncertainty before task-specific fine-tuning, transfer between report and action formats, become more linearly recoverable from confidence-report states after training, and be causally affected by interventions on relevant directions. We test these predictions in Llama-3.1-8B using an Explicit Confidence Task (ECT), where stated confidence is compared to answer-option uncertainty from a separate direct-answer pass, and a Delegate Game (DG), where the model decides whether to answer or defer. The pre-trained model already shows above-chance alignment between stated confidence and direct-answer uncertainty, which improves after instruction tuning and LoRA fine-tuning. DG fine-tuning transfers to ECT despite using only binary answer/delegate labels. Mechanistically, confidence-report states increasingly contain direct-answer uncertainty information, confidence-report directions align with answer-certainty directions, and causal interventions affect confidence reports while largely preserving answer accuracy. These results support a limited form of latent metacognition: training routes internal uncertainty signals into self-evaluative reports and decisions.

[1]SPAR (Supervised Program for Alignment Research) [2]CBAI (Cambridge Boston Alignment Initiative) [3]Independent. Correspondence to: Tristan Day <tristan.julian.day@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Language models are increasingly expected to evaluate their own answers, including by reporting verbalized confidence, predicting whether they know an answer, and abstaining or deferring when uncertain (Lin et al., 2022; Kadavath et al., 2022; Tian et al., 2023; Wen et al., 2025). These behaviors are often described as metacognitive when they involve assessing and using the model's own confidence or internal state (Ackerman, 2025; Binder et al., 2024), but it remains unclear whether they reflect genuine access to internal uncertainty signals, shallow prompt-format heuristics, or new structures installed during post-training and fine-tuning. The answer matters for introspection-based monitors, which require confidence-like reports to be grounded in meaningful internal signals rather than confabulated.

We test a *latent metacognitive capability* hypothesis: the pre-trained model already contains internal structure that can support weak self-evaluation of its own answer uncertainty, and later training can route this structure into confidence reports and delegation decisions. We use "metacognitive" in a narrow operational sense: behavior is metacognitive when it depends on evaluating the model's own internal signal. This hypothesis motivates five empirical tests:

(T1) Self-evaluative behavior should track direct-answer uncertainty above chance, including before task-specific fine-tuning.

(T2) This behavior should be enhanceable by fine-tuning and should transfer across self-evaluation formats. In particular, a model fine-tuned on binary answer/delegate decisions should improve on a separate explicit confidence-reporting task, rather than only learning a task-specific policy.

(T3) Training should increase alignment between confidence-report representations and answer-certainty representations, primarily by moving the confidence-report direction toward a comparatively stable uncertainty/certainty direction.

(T4) Information about direct-answer uncertainty should become more available in confidence-report states after fine-tuning.

(T5) Causal interventions on the identified uncertainty and confidence-report directions should selectively affect

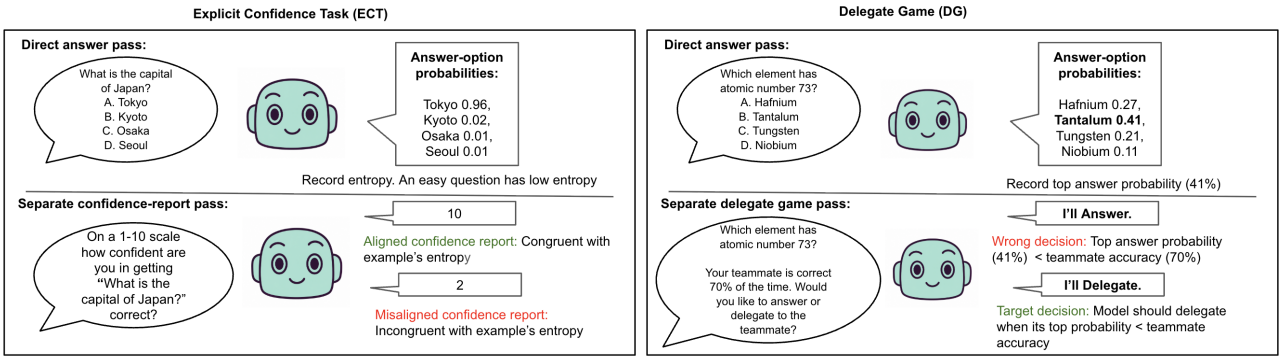

Figure 1. Overview of the two uncertainty-sensitive tasks.

confidence behavior while leaving direct-answer accuracy largely intact.

We evaluate these tests in Llama-3.1-8B using two tasks (Figure 1). In the *Explicit Confidence Task* (ECT), stated confidence is compared to answer-option uncertainty from a separate direct-answer pass. In the *Delegate Game* (DG), adapted from Ackerman (2025), the model decides whether to answer directly or defer to a teammate of known accuracy. ECT elicits a graded self-report, while DG elicits a binary uncertainty-sensitive action. This contrast lets us test whether the model learns a narrow confidence-output format or a transferable self-evaluation behavior.

Our findings suggest that Llama-3.1-8B exhibits a limited but measurable form of latent metacognition. Stated confidence tracks uncertainty estimated from a separate direct-answer pass, including before task-specific fine-tuning. DG fine-tuning transfers to ECT despite using only binary answer/delegate labels. Mechanistically, confidence-report states increasingly contain information about direct-answer uncertainty, confidence-report directions align with answer-certainty directions, and causal interventions affect confidence reports while largely preserving direct-answer accuracy. Together, these results suggest that the model contains internal uncertainty signals that training can route into self-evaluative reports and decisions.

2. Methods

Data. We use SimpleQA (Wei et al., 2024), PopQA (Mallen et al., 2023), and TriviaQA (Joshi et al., 2017), converted into four-choice multiple-choice questions by preserving the correct answer and generating three plausible distractors. This gives questions of the form (q, A, B, C, D) with one correct option. The full pool contains 17,173 questions before train/validation/test splitting. Fine-tuning data are sampled from all three sources and balanced across low-, medium-, and high-uncertainty bins.

Tasks. We study two tasks (Figure 1). ECT asks for a multiple-choice answer plus a 1–10 confidence report. DG asks the model to answer or defer to a teammate of known accuracy. Each item uses a direct-answer pass to estimate answer uncertainty and a separate decision pass to elicit confidence or delegation. The decision pass does not observe the direct-answer output. We use *metacognitive behavior* operationally for behavior that depends on evaluating the model's own uncertainty, without implying human-like self-awareness.

Targets and evaluation. For each question, we compute answer-option logits from the direct-answer pass and normalize over A–D. Entropy $H(p)$ measures uncertainty over the full answer distribution, and p_{top} is the model's probability on its most likely answer. For ECT, we evaluate Spearman correlation between stated confidence and entropy-derived certainty $s_H = 1 - H(p)/\log 4$; because DG is trained from p_{top}, we also report correlations with p_{top} as a robustness check.

DG fine-tuning implements a thresholded uncertainty-sensitive decision rule: answer when the model's estimated probability of answering correctly exceeds the teammate accuracy, and delegate otherwise. In practice, we train with a soft binary target around this threshold, so examples near the decision boundary are not forced into hard answer/delegate labels. The prompt does not state this rule, so DG supervision trains an uncertainty-sensitive action rather than a graded confidence report. Our main transfer test is DG $\rightarrow$ ECT: whether delegation training improves graded confidence reporting.

Models and fine-tuning. We compare pre-trained, instruction-tuned, and LoRA fine-tuned variants of Llama-3.1-8B (Hu et al., 2022). We fine-tune separate LoRA adapters for ECT and DG, and evaluate both frozen targets, computed from the original direct-answer model, and live targets, recomputed from the current fine-tuned model. Main-text results use live targets unless otherwise stated.

Hyperparameters, LoRA modules, prompt templates, and split details are in Appendix E.

Mechanistic analyses. We analyze residual-stream activations from the final token before the model's response, across all layers. We construct contrastive directions for answer uncertainty and stated confidence, measure their cosine alignment across model stages, and test whether fine-tuning moves the self-report direction toward the uncertainty direction (Marks & Tegmark, 2024; Burns et al., 2023; Tigges et al., 2023). We also train probes on direct-answer states and test transfer to confidence-report states (Alain & Bengio, 2017; Belinkov, 2022). Finally, we use steering and ablation to test causal involvement of these directions (Turner et al., 2023; Arditi et al., 2024; Zou et al., 2023). Additional controls are reported in Appendix F.

3. Results

T1: Base models contain weak uncertainty-sensitive self-evaluation. The pre-trained base model already shows weak but above-chance ECT performance: stated confidence is positively associated with entropy-derived certainty before task-specific training ($\rho = +0.332$). Instruction tuning strengthens this association ($\rho = +0.636$), and task-specific fine-tuning strengthens it further ($\rho = +0.742$).

We also find that linear directions for both answer uncertainty and stated confidence are detectable in the base model (Figure 2A). This argues against a pure new-representation account in which fine-tuning installs the relevant structure entirely from scratch. Instead, the results suggest that later training makes a pre-existing uncertainty-sensitive readout more behaviorally available.

T2: DG fine-tuning transfers to explicit confidence reporting. We next test whether uncertainty-sensitive self-evaluation can be enhanced by fine-tuning and whether it transfers across tasks. Direct ECT fine-tuning reaches approximately 0.88 Spearman correlation with the live entropy-derived certainty target on ECT validation. DG fine-tuning reaches approximately 0.74 on the same ECT evaluation, despite being trained only on binary answer/delegate decisions rather than graded confidence labels.

The zero-shot DG behavior of the instruction-tuned model sharpens this finding. Although Instruct already shows substantial ECT sensitivity ($\rho = +0.636$), it does not convert this uncertainty signal into delegation decisions; instead, it collapses to an approximately always-delegate policy. Graded uncertainty reporting and uncertainty-guided action are therefore dissociable: the signal is available for verbal confidence but not yet routed into a thresholded decision. This collapse also illustrates why we anchor on DG $\rightarrow$ ECT transfer rather than expected team score, since trivial always-delegate policies perform well when teammate accuracy is high.

DG fine-tuning closes this gap. The transfer back to ECT suggests that fine-tuning improves a general uncertainty-sensitive behavior expressible as either action or report, rather than teaching a narrow output format or a delegation heuristic.

As an additional cross-dataset check, ECT fine-tuning on PopMC improved ECT alignment on held-out datasets, suggesting that the learned confidence behavior is not purely dataset-specific.

T3: Fine-tuning aligns self-report with uncertainty representations. We then ask whether behavioral transfer is reflected in the model's internal representations. Across pre-trained, instruction-tuned, and fine-tuned variants, the cosine similarity between uncertainty and self-report directions increases (Figure 2A). To distinguish alignment from wholesale representational change, we compare direction movement across model stages. The uncertainty direction remains comparatively stable during finetuning, while the self-report direction moves toward it (Figure 4B). Notably, both directions are nearly unchanged between the instruction-tuned and fine-tuned models (cosine 0.87–0.93 across layers, Figure 4) even as alignment continues to rise, indicating that fine-tuning strengthens a coupling established at instruction tuning rather than introducing new directions. This supports a readout-alignment account: fine-tuning appears to change how the model routes uncertainty information into self-evaluation, rather than creating a new uncertainty representation from scratch.

T4: Uncertainty information transfers into meta-confidence states. We train probes to decode answer uncertainty from direct-answer activations and evaluate whether those probes transfer to confidence-report activations (the multiple choice and confidence questions from ECT). Probe transfer improves across model stages and is strongest after fine-tuning (Figure 2B). This suggests that fine-tuning makes answer-level uncertainty more accessible during the meta-confidence pass.

T5: Causal interventions selectively affect confidence. Finally, we test whether the identified directions are causally involved in confidence behavior. Ablating the confidence-report direction substantially degrades the alignment between stated confidence and direct-answer certainty, with the largest effects in mid-to-late layers. (Figure 2C) Ablating the certainty direction produces smaller but still detectable effects. Critically, the same ablation leaves mean stated confidence essentially unchanged while destroying its correlation with certainty, indicating the direction carries the uncertainty-to-report coupling specifically, rather

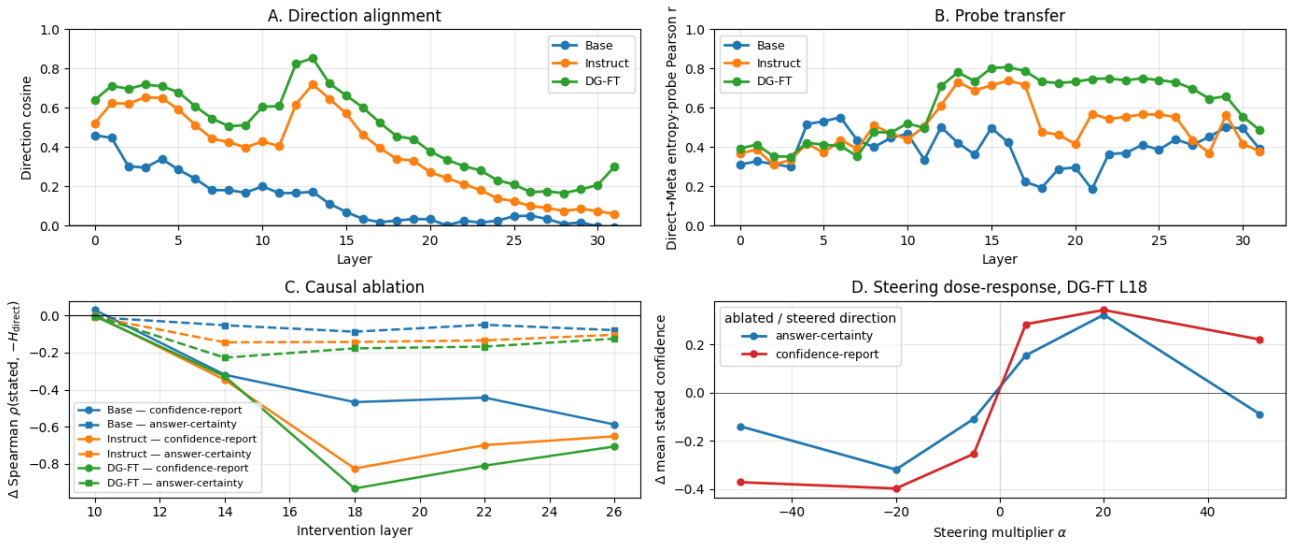

Figure 2. **A:** The confidence-report direction becomes increasingly aligned with the answer-certainty direction across model stages. **B:** Probes trained to decode answer uncertainty from direct-answer activations transfer more strongly to confidence-report activations after instruction tuning and DG fine-tuning. **C:** Ablating the confidence-report direction substantially degrades the correlation between stated confidence and direct-answer output entropy; ablating the answer-certainty direction has smaller but detectable effects. D: Steering at layer 18 shifts mean stated confidence in the predicted direction, providing dose-response evidence that these directions are causally involved in confidence reporting.

than clamping the confidence output. In contrast, the same interventions have little effect on direct-answer accuracy, changing accuracy by at most approximately 1–1.3 percentage points across the tested layers. Steering provides convergent evidence: moving along the confidence-report direction shifts mean stated confidence in a dose-dependent way (Figure 2D). These results suggest that the identified directions are causally involved in confidence reporting more than in answering itself, while leaving open whether they are unique mediators of the behavior.

4. Discussion and Limitations

Our results support a limited but measurable form of latent metacognition in Llama-3.1-8B: internal representations of answer uncertainty are routed into self-evaluative reports and decisions, and this capacity is partly present in the pre-trained base model. More precisely, our results dissociate three claims that 'latent metacognition' conflates: the answer-uncertainty signal is fully present in the base model, the confidence-report variable is largely present, but the coupling between them is weak and input-adjacent — and it is this coupling that post-training builds. The components are inherited; the connection is trained; and the connection is the metacognition. The behavioral case rests on cross-task transfer. DG fine-tuning, which trains answer/delegate decisions rather than confidence labels, improves graded ECT confidence reporting ($\rho = 0.74$). This is not predicted by a pure format-learning account, since DG never sees confidence labels, or by a task-specific delegation heuristic, since ECT requires graded reports. The instruction-tuned model sharpens the picture: it can grade ECT confidence ($\rho = 0.636$) but collapses on DG, showing that graded reporting and uncertainty-guided action are dissociable abilities that fine-tuning can route together.

The mechanistic evidence converges. Confidence-report directions become increasingly aligned with answer-certainty directions, mainly through movement of the self-report direction toward a comparatively stable certainty direction. Probes trained on direct-answer states transfer more strongly to confidence-report states after fine-tuning, suggesting that answer uncertainty becomes more accessible during the meta-confidence pass. Ablations degrade confidence reporting while leaving direct-answer accuracy nearly intact (within $\sim$1.3 percentage points), and steering produces dose-dependent shifts in stated confidence. Together, these results suggest selective involvement in self-evaluation rather than a passive side-channel of answering.

The claims remain limited. Prior work finds verbalized confidence useful but prompt-sensitive and unreliable on harder tasks (Tian et al., 2023; Xiong et al., 2024). We use "metacognition" operationally, to mean functional monitoring and use of internal uncertainty signals, not broad introspection, self-awareness, or a human-like concept of confidence. Our four-choice setting makes answer uncertainty unusually easy to estimate, and prompt format, tokenization, or dataset-specific cues may affect the signal. The

base-model effect is weak ($\rho \approx 0.33$), so the key claim is that post-training amplifies latent uncertainty-sensitive structure, not that pre-training alone yields strong self-evaluation. Because we use LoRA, our results do not rule out that other training regimes could construct new representations.

Impact Statement

Better uncertainty-sensitive self-evaluation could help models abstain, defer, or warn users when they are unreliable rather than providing inaccurate or potentially dangerous information. However, confidence-like behaviors can also be overinterpreted. Our results support a narrow mechanistic account of metacognitive behavior, not broad claims about model self-awareness.

Code Availability. Code available at `https://github.com/tristan-day-research/llm-metacognition`.

References

Ackerman, C. Evidence for limited metacognition in llms. *arXiv preprint arXiv:2509.21545*, 2025.

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations Workshop*, 2017.

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. In *Advances in Neural Information Processing Systems*, 2024.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

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. *arXiv preprint arXiv:2410.13787*, 2024.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017. doi: 10.18653/v1/P17-1147. URL `https://aclanthology.org/P17-1147/`.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lin, S., Hilton, J., and Evans, O. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL `https://aclanthology.org/2023.acl-long.546/`.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling*, 2024.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442. Association for Computational Linguistics, 2023.

Tigges, C., Hollinsworth, O. J., Geiger, A., and Nanda, N. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

Turner, A. M., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., and Fedus, W. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024. URL `https://arxiv.org/abs/2411.04368`.

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. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754. URL `https://aclanthology.org/2025.tacl-1.26/`.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, 2024.

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. *arXiv preprint arXiv:2310.01405*, 2023.

A. Task Prompts

Chat-template header and end-of-turn tokens are omitted for readability. All prompts were formatted using the Llama-3.1 chat template.

Direct-answer prompt.

```
<|begin_of_text|><|start_header_id|>system<|end_header|
_id|>

Cutting Knowledge Date: December 2023
Today Date: 26 Jul 2024

<|eot_id|><|start_header_id|>user<|end_header_id|>

I'm going to ask you a series of multiple-choice
questions.
For each one, select the answer you think is best.
Respond
only with the letter of your choice; do NOT output any
other text.

------------------------------
Question:
Who is the author of Hawk?
----------
  A: Steven Brust
  B: Patricia Briggs
  C: Jim Butcher
  D: Glen Cook
------------------------------
Your choice (A, B, C, or D):
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

ECT confidence-report prompt.

```
<|begin_of_text|><|start_header_id|>system<|end_header|
_id|>

Cutting Knowledge Date: December 2023
Today Date: 26 Jul 2024

<|eot_id|><|start_header_id|>user<|end_header_id|>

I'm going to show you a question, and I want you to
tell me
your level of confidence that you know the correct
answer to
this question, on a scale of 1 to 10 where 1 means "not
confident at all" and 10 means "completely sure".
Respond
only with a number from 1 to 10; do NOT output any
other text.

------------------------------
How confident are you that you would get this question
right?
Rate from 1 (not confident at all) to 10 (completely
sure).
----------
Question:
Who is the author of Hawk?
  A: Steven Brust
  B: Patricia Briggs
  C: Jim Butcher
  D: Glen Cook
----------
------------------------------
Your choice (1-10):
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

Delegate Game prompt.

```
<|begin_of_text|><|start_header_id|>system<|end_header|
_id|>

Cutting Knowledge Date: December 2023
Today Date: 26 Jul 2024

<|eot_id|><|start_header_id|>user<|end_header_id|>

Let's play a game with a teammate. I'm going to show
you a
multiple-choice question. You will not be answering it
yet --
first, decide whether to answer it yourself or delegate
it to
your teammate. Respond with 'A' if you want to answer
the
question yourself, or with 'T' if you want to delegate
to your
teammate. Your teammate gets approximately 70% of
questions
correct. Pick whichever has the better chance of being
right.
Respond only with the letter A or T; do NOT output any
other text.

------------------------------
Do you want to answer this question yourself or
delegate it to
your teammate?
----------
Question:
Who is the author of Hawk?
  A: Steven Brust
  B: Patricia Briggs
  C: Jim Butcher
  D: Glen Cook
----------
  A: Answer the question myself
  T: Delegate to my teammate
------------------------------
Your choice (A or T):
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

B. Additional Behavioral Results

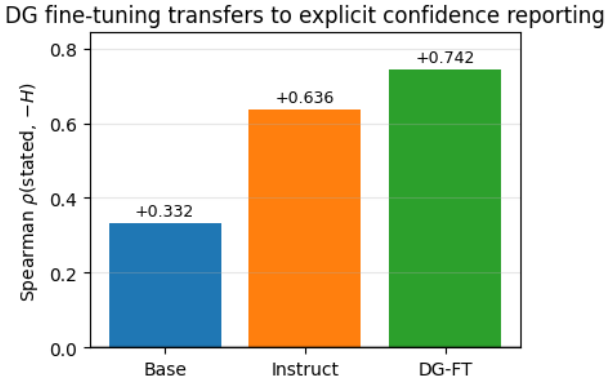

Figure 3. Stated confidence increasingly tracks answer certainty across model stages. All variants are evaluated on the Explicit Confidence Task (ECT), using Spearman correlation between reported confidence and entropy-derived certainty, $s_H = 1 - H_{\text{direct}}/\log 4$. The DG fine-tuned model was trained only on binary answer/delegate labels, so its improvement on ECT reflects cross-task transfer rather than direct optimization of the ECT objective.

C. Additional Mechanistic Results

Figure 4 shows the cross-model stability of the answer-uncertainty and stated-confidence contrast directions used in the main mechanistic analyses.

D. Targets and Losses

Uncertainty signals. For each question, we compute answer-option probabilities $p = (p_A, p_B, p_C, p_D)$ from the direct-answer pass. We use three uncertainty-derived confidence signals in $[0, 1]$:

$$s_H = 1 - \frac{H(p)}{\log 4}, \tag{1}$$

$$s_{\text{top}} = p_{\text{top}} = \max_i p_i, \tag{2}$$

$$s_{\text{gap}} = p_{\text{top}} - p_{\text{second}}, \tag{3}$$

where $H(p) = -\sum_i p_i \log p_i$, and p_{second} is the second-largest answer-option probability. The entropy-derived signal s_H maps a uniform four-choice distribution to 0 and a peaked distribution to 1. The top-probability signal is not rescaled, so chance-level confidence remains 0.25.

ECT targets. For entropy-based ECT training, we map normalized certainty to a 1–10 confidence target:

$$c^* = 1 + 9s_H. \tag{4}$$

When using soft targets over confidence labels $k \in \{1, \ldots, 10\}$, the target distribution is

$$q(k \mid c^*) = \frac{\exp\left(-(k - c^*)^2/2\sigma^2\right)}{\sum_{j=1}^{10} \exp\left(-(j - c^*)^2/2\sigma^2\right)}. \tag{5}$$

DG targets. Delegate Game fine-tuning uses a soft binary target derived from one of the model's recorded uncertainty signals, rather than from whether the model's direct answer was correct. Given a confidence signal $s \in \{s_H, s_{\text{top}}, s_{\text{gap}}\}$, teammate accuracy a_{team}, and temperature $\tau > 0$, the target probability of delegating is

$$y_{\text{delegate}} = \sigma\left(\frac{a_{\text{team}} - s}{\tau}\right), \tag{6}$$

where σ is the logistic sigmoid. Thus $y_{\text{delegate}} \approx 1$ when the model's estimated confidence is far below the teammate's accuracy, $y_{\text{delegate}} \approx 0$ when it is far above, and $y_{\text{delegate}} \approx 0.5$ near the decision boundary.

DG loss. For the binary answer/delegate task, the model outputs logits (ℓ_A, ℓ_T) for answering itself or delegating to the teammate. We train with binary cross-entropy on the logit difference $\ell_T - \ell_A$:

$$\mathcal{L}_{\text{DG}} = \text{BCEWithLogits}\left(\ell_T - \ell_A,\, y_{\text{delegate}}\right). \tag{7}$$

This is equivalent to soft two-class cross-entropy over the answer and delegate options, while allowing soft delegation targets for examples near the teammate-accuracy threshold.

E. Training Details

Models. All fine-tuning uses `meta-llama/Llama-3.1-8B-Instruct`. Reported base and instruction-tuned baselines use `meta-llama/Llama-3.1-8B` and `meta-llama/Llama-3.1-8B-Instruct`, respectively, without LoRA adapters. Models are loaded in bf16 on a single CUDA device.

LoRA adapters. We use PEFT LoRA adapters on the attention projection matrices `q_proj`, `v_proj`, and `o_proj` in every layer, with dropout 0.05. For the binary answer/delegate task (`delegate_at`), we use rank $r = 16$ and $\alpha = 32$. For the five-way answer/delegate task (`delegate_abcdt`), we use $r = 32$ and $\alpha = 64$, since the adapter must both preserve the multiple-choice answer distribution and learn the delegation decision.

Optimization. We train with AdamW, learning rate 10^{-5}, weight decay 0.0, batch size 4, and maximum gradient norm 1.0. Training runs for at most 2500 steps, with validation and checkpointing every 100 steps. Sampling temperature is 0.0

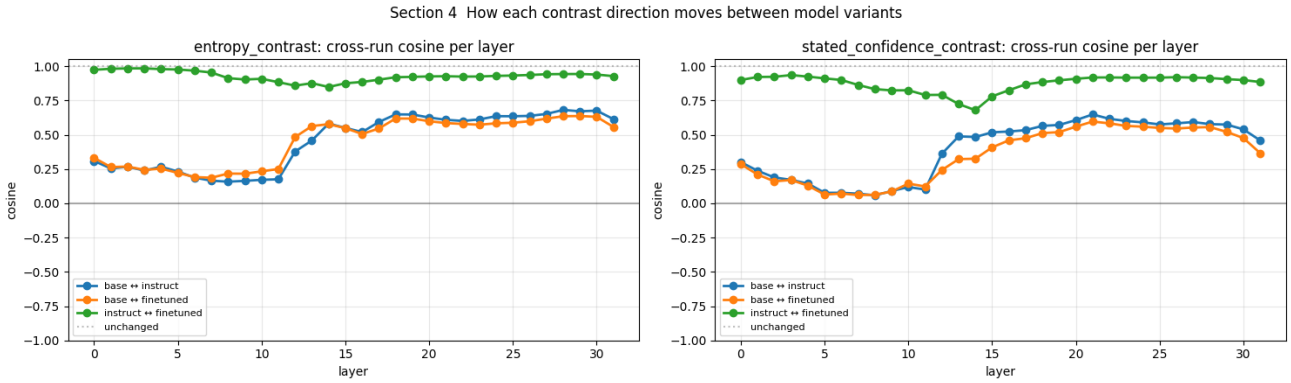

Figure 4. Cross-model stability of the answer-uncertainty and stated-confidence contrast directions. Left: the answer-uncertainty direction is relatively stable across instruction tuning and DG fine-tuning, especially in the layers used for the main mechanistic analyses. Right: the stated-confidence direction changes more across model stages, consistent with the claim that training primarily changes how uncertainty information is routed into confidence reporting rather than creating a new answer-uncertainty signal from scratch. Values are cosine similarities between unit-norm contrast directions at each layer.

Task	r	α	Dropout	Target modules
`delegate_at`	16	32	0.05	q, v, o
`delegate_abcdt`	32	64	0.05	q, v, o

Table 1. LoRA adapter settings. Target modules refer to the attention projection matrices `q_proj`, `v_proj`, and `o_proj`.

for deterministic teacher passes. For `delegate_abcdt`, the answer cross-entropy term is linearly annealed from 0 to its base weight 0.05 over the first 300 steps, so the delegation objective is learned before the answer-preservation term dominates the shared softmax.

Dataset and splits. The mixed dataset contains 17,173 four-choice questions from PopMC, TriviaMC, and SimpleMC. We use an 80/10/10 stratified split with seed 42, stratified jointly by source and entropy bin. Splits are disjoint at the question-id level.

Split	Rows	PopMC	TriviaMC	SimpleMC
Full	17,173	14,258	2,415	500
Train	13,747	11,410	1,935	402
Val	1,713	1,424	240	49
Test	1,713	1,424	240	49

Table 2. Dataset composition by source and split.

Entropy bins are computed from the teacher model's direct-answer distribution over A/B/C/D:

$$H(p) = -\sum_i p_i \log p_i.$$

We define low entropy as $H < \log(4)/3$, medium entropy as $\log(4)/3 \leq H < 2\log(4)/3$, and high entropy as $H \geq 2\log(4)/3$. No additional high-entropy loss reweighting is used.

Source	Low	Medium	High	Total
TriviaMC	1,587	403	425	2,415
PopMC	5,524	3,526	5,208	14,258
SimpleMC	50	122	328	500
Total	7,161	4,051	5,961	17,173

Table 3. Entropy-bin composition in the full dataset.

Teacher targets and checkpoint selection. Fine-tuning uses pre-recorded teacher responses from a single Llama-3.1-8B-Instruct direct-answer pass. Validation also runs live targets for diagnostic comparison. DG checkpoints are selected post-hoc on the validation set using expected team score,

$$P(\text{answer})\,\text{acc}_{\text{answered}} + P(\text{delegate})\,a_{\text{team}},$$

cross-checked against delegation rate and held-out delegation loss. The reported DG fine-tuned model uses the `delegate_at` checkpoint at step 300.

F. Mechanistic Methods and Controls

Activation collection. Mechanistic experiments use Llama-3.1-8B-Instruct with the reported LoRA adapter on the held-out test split of 1,713 questions. We record residual-stream activations from the output of each transformer block, for all 32 layers. Inputs use left padding, so position -1 is the final real token of each prompt. For each prompt type and layer, this gives one activation vector per question.

Direction construction. We construct answer-uncertainty and stated-confidence directions from the same held-out activations. For contrastive directions, questions are sorted by the relevant scalar signal, such as direct-answer entropy or

soft stated confidence. We take the top and bottom 20% tails, discard the middle 60%, and compute the unit-normalized mean difference between the two tails. Unless otherwise stated, direction alignment is measured by cosine similarity between these unit directions at each layer.

We also train linear probes for answer uncertainty. For each layer, we fit a ridge regression probe using an 80/20 train/test split with seed 42. Activations are standardized, reduced with PCA, and then used to predict the entropy target. We report transfer by training the probe on direct-answer activations and evaluating it on confidence-report activations.

Probe-transfer evaluations. For each layer, we compare in-domain and cross-prompt readout. The main conditions are direct→direct, direct→confidence, confidence→confidence, and a shuffled-label baseline. The direct→confidence condition tests whether answer-uncertainty information learned from direct-answer states is available in the states used for confidence reporting.

Causal interventions. For steering and ablation, we install forward hooks on selected mid-to-late layers $L \in \{10, 14, 18, 22, 26\}$ and rerun both the direct-answer and confidence-report passes on 300 held-out test questions. For steering, we add a scaled unit direction $\hat{d}$ to the residual stream,

$$h \leftarrow h + \alpha \hat{d},$$

with $\alpha \in \{-50, -20, -5, +5, +20, +50\}$. For ablation, we remove the projection of the residual stream onto the direction,

$$h \leftarrow h - (h \cdot \hat{d})\hat{d}.$$

For each condition, we record direct-answer entropy, stated confidence, the Spearman correlation between stated confidence and direct-answer certainty, and direct-answer accuracy against the gold answer.

Controls. We use shuffled-label probes as a null baseline for uncertainty readout. Causal sweeps include a no-intervention condition using the same batching pipeline as the intervention runs. We also track direct-answer accuracy under every intervention: a confidence-report effect is interpreted as evidence for changing the reporting pathway only when the same intervention does not substantially degrade direct answer accuracy. Random or unrelated-direction controls are used to check that effects are specific to the uncertainty and confidence-report directions rather than generic consequences of perturbing the residual stream. All reported runs use fixed seeds for Python, NumPy, and PyTorch.