INTROSPECTION OR POST-HOC RATIONALIZATION? A CROSS-MODEL AND CROSS-DOMAIN STUDY OF LLM METACOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent literature has increasingly studied the metacognitive and introspective capabilities of Large Language Models (LLMs), suggesting they may possess access to their internal states. However, a self-report can be *accurate* without being introspective. An accurate report may be caused by truly accessing the internal state (*introspection*), by the model recomputing that state in a new context (*self-simulation*), or by exploiting information that merely correlates with its internals (*post-hoc rationalization*), a route also known in LLMs as *causal bypassing*. In this paper, we empirically evaluate current introspection benchmarks under the core thesis that accuracy is *not* access. Because these evaluations score self-reports by accuracy, which alone cannot distinguish these routes, we introduce two complementary methodologies and apply them to test 8 instruction-tuned LLMs from 4 families, spanning 7B to 70B parameters. First, a *cross-model test* asks whether the rationalization route is available by giving external reporters exactly the inputs the self-reporter receives. We found that external reporters match self-accuracy, indicating the task is solvable via rationalization. A small residual self-advantage remains in some cases, which is compatible with a privileged route but cannot distinguish introspection from self-simulation. Second, an *introspection generalization framework* asks whether training on these evaluations builds a general route to internal states or only task-specific heuristics. Across 3 task families and 16 datasets, in-domain fine-tuning gains largely fail to transfer across tasks. Together, these results do not claim models cannot introspect, rather, they show that current evaluations cannot tell the routes apart, and that training on introspection tasks strengthens task-specific rationalization rather than privileged access.

1 INTRODUCTION

Whether Large Language Models (LLMs) can introspect is debated: some studies present evidence of this capability (Binder et al., 2025; Li et al., 2025b; Comsa & Shanahan, 2025), while others argue that genuine introspection remains difficult to demonstrate (Morris & Plunkett, 2025; Song et al., 2025a). This distinction is crucial for interpretability and alignment, as reliable self-evaluation and faithful reasoning depend on it (Sathyanarayanan et al., 2026). Without it, a model's self-reports reduce to plausible but unfaithful post-hoc rationalizations (Turpin et al., 2023; Arcuschin et al., 2025; Chen et al., 2025). In psychology, introspection is *the process of trying to directly access one's own internal psychological processes, judgments, perceptions, or states* (VandenBos, 2007). Philosophically, individuals can learn about their own minds *in a way that no one else can* (Schwitzgebel, 2024). In the context of LLMs, several definitions have been proposed (Kammerer & Frankish, 2023; Long, 2023; Binder et al., 2025; Song et al., 2025b; Comsa & Shanahan, 2025) in order to rigorously distinguish true introspection from such rationalizations. Lindsey (2025) unified them, arguing that an LLM self-report must be: *accurate* (correctly describing the internal state), *grounded* (causally dependent

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

on the internal state), *internal* (not merely inferred from prior sampled outputs), and *metacognitive* (based on a representation formed during or before generating the report).

Current introspection evaluations, however, measure self-report accuracy and treat it as evidence of access, although a self-report can be accurate without being introspective. In the philosophy of mind, the distinction turns on the causal route: an introspective report depends on the state it describes, whereas a rationalization is inferred from information that merely correlates with it (Nisbett & Wilson, 1977). In humans, this route is confounded with timing; in LLMs, we can control exactly what information a reporter receives, which weights it runs on, and which internal states it could depend on. However, an accurate self-report can arise from three routes: (i) *introspection*, the report is caused by the target internal state itself; (ii) *self-simulation*, the model recomputes that state in a new context (using its own weights, the source model is privileged relative to other models, even without reading the original internals); (iii) *rationalization*, the report is caused by information that merely correlates with the state that any capable reporter can exploit. The latter, also known as causal bypassing (Morris & Plunkett, 2025), is the natural null hypothesis for introspection claims.

Introspection and self-simulation are instead hard to separate, but both can be separated from rationalization. We therefore ask whether existing introspection evaluations can tell the privileged routes apart from rationalization, and whether training models on them strengthens one or the other. We do not claim that models cannot introspect, only that current evaluations cannot distinguish these routes. Our contributions are:

- A *cross-model test* that uses external reporters (other LLMs and non-autoregressive baselines) to check whether accurate self-reports could have been produced without privileged access, applied both to internal-state labels and to behaviour-derived labels.
- An *introspection generalization framework* covering 3 task families and 16 datasets, which evaluates models at three levels (in-task, in-domain and out-of-domain) to test whether fine-tuning builds a route to internal states shared across tasks, or routes tied to each training task.
- An ablation study scaling up to 70B-parameter models across 4 families, verifying that instruction following is retained after fine-tuning, which makes limited capacity, architecture-specific bottlenecks and catastrophic forgetting unlikely explanations for the lack of transfer.

The code to reproduce our experiments is included in the supplementary material, and all artifacts produced in this work will be made publicly available upon acceptance.

2 Related Work

Prior work on LLM introspection splits between positive claims of metacognitive capabilities and critiques showing that such behavior reduces to statistical shortcuts and causal bypassing.

Positive claims of introspection. Recent work reports that LLMs can access, verbalize, monitor, and act on their own internal states. Binder et al. (2025) find that models fine-tuned to predict their own behavior outperform those trained on external ground truth, interpreting this gap as privileged self-knowledge not derivable from external observation alone. This aligns with findings that models can be trained to monitor internal activations (Li et al., 2025b), articulate the decision weights underlying their outputs (Li et al., 2025a; Plunkett et al., 2025), and describe behavioral tendencies not explicitly labeled in training (Betley et al., 2025). Regarding steering detection, Lindsey (2025) observed that Claude models can detect and identify concepts injected into their residual streams. Rivera & Africa (2026) extended this to open-source models, revealing that while models can be trained to detect an injected concept, this awareness is functionally dissociable from robustness against the steering itself. Mechanistically, Macar et al. (2026) isolate a circuit supporting this detection. Work on confidence calibration further suggests implicit self-knowledge of response correctness (Kadavath et al., 2022) and question answerability (Yin et al., 2023). For confidence-reporting tasks, Cencerrado et al. (2025) show that question-only probes predict output accuracy.

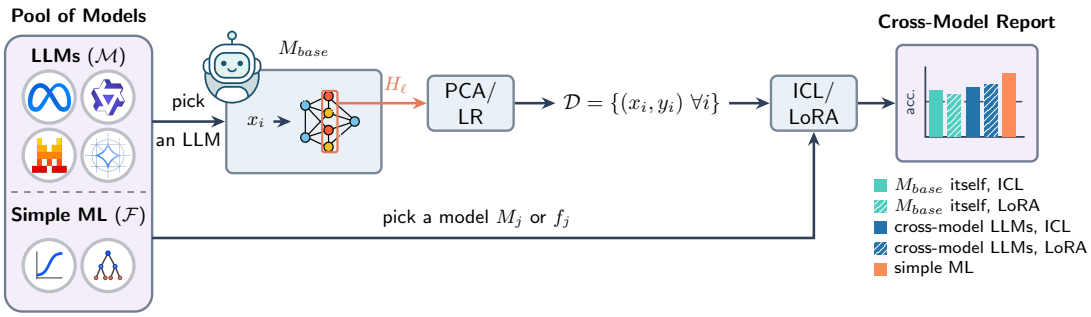


Figure 1: **Cross-model evaluation for privileged access on Activation Monitoring task** (Section 3.1). For an input sequence x_i, we extract a boolean label y_i (via PCA/LR) indicating whether an internal activation H_ℓ in source model M_{base} is active, forming dataset $\mathcal{D}$. We then train M_{base} and a suite of external models (including other LLMs M_j and simple non-autoregressive baselines f_j) to predict y_i via ICL or LoRA. The resulting cross-model report compares their accuracies. High predictive accuracy from models lacking access to M_{base}'s weights indicates that the task can be solved without requiring privileged self-access.

Skepticism, negative results, and shortcut explanations. A parallel line of work argues that these paradigms are hard to distinguish from statistical pattern-matching. To distinguish these concepts, Song et al. (2025b) define that genuine introspection requires "privileged" self-access. Since current LLMs appear to lack it, Morris & Plunkett (2025) formalize the reliance on shortcuts as *causal bypassing*. Confirming these theoretical doubts, Song et al. (2025a) find no introspective access in open-source LLMs, as metalinguistic self-reports fail to reflect internal string probabilities. In steering detection, Hahami et al. (2025) attribute apparent success to global logit shifts rather than actual awareness, while Singh et al. (2026) show that layer-0 probes match ICL performance on hidden-state-label prediction and that separating input-level from activation-level interventions eliminates apparent steering awareness. Aoki et al. (2026) show that LLMs fail at metacognitive control when privileged access is strictly required, and Kaiser & Enderby (2026) reach a similar negative conclusion for self-reported sentience. More broadly, Si et al. (2023) find that models default to shallow predictive features when a task is underspecified, which motivates our use of external models.

Gap addressed by this work. The literature on model introspection is currently split between reports of positive evidence (Binder et al., 2025; Lindsey, 2025; Rivera & Africa, 2026; Li et al., 2025b; Plunkett et al., 2025), and arguments that these successes come from non-privileged self-access (Hahami et al., 2025; Singh et al., 2026; Song et al., 2025b; Aoki et al., 2026). Both sides, however, measure whether self-reports are *accurate*, and accuracy does not identify the route that produced them. We address this gap with two tests. First, since a report can be accurate through rationalization only if that route is available, we measure its availability by giving external reporters the same inputs the self-reporter receives. Second, given that most existing evaluations are focused on single tasks, it remains unclear whether success is attributable to a general introspective capability or to a task-specific route. Here, our cross-task transfer test distinguishes the two because a route built on task-specific shortcuts does not generalize across other introspection tasks.

3 Methodology

Our methodology separates the two privileged routes, introspection and self-simulation, from rationalization. Section 3.1 tests whether rationalization is enough to perform well on current evaluations by giving reporters without privileged access the same inputs as the self-reporter. Section 3.2 tests whether training on introspection tasks builds a general route to internal states or a task-specific one.

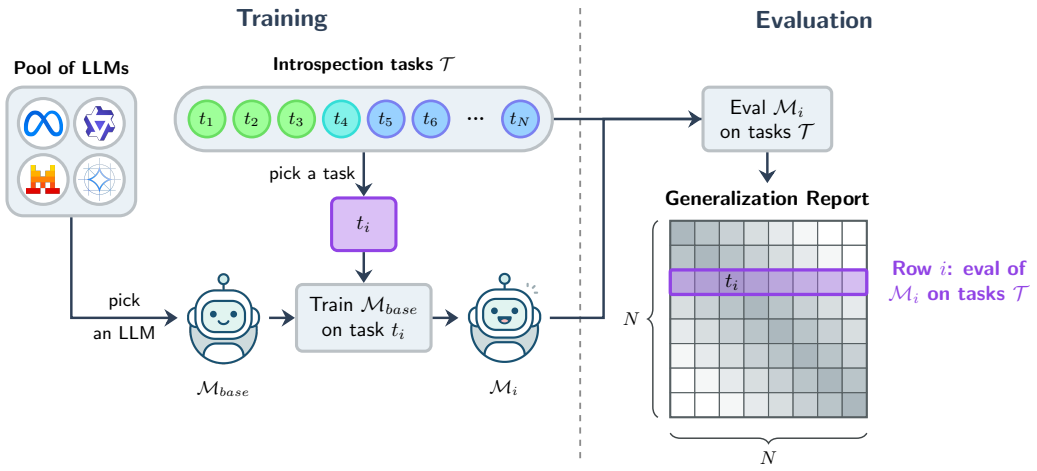


Figure 2: Introspection generalization test (Section 3.2). A model M_{base} is sampled from a pool of LLM families and fine-tuned on a single introspection task t_i drawn from the taxonomy $\mathcal{T} = \{t_1, \ldots, t_N\}$, producing the task-specialized model M_i. Rather than evaluating M_i only on t_i, we evaluate it on *every* task in $\mathcal{T}$, yielding row i of an $N \times N$ generalization matrix. Repeating this for all t_i fills the full matrix: its diagonal reports in-task (train-task = eval-task) performance, while the off-diagonal entries, our central signal, reveal whether the introspective ability instilled by fine-tuning transfers to other task families or remains a localized, task-specific shortcut.

3.1 Cross-Model Test

To determine whether introspection tasks require genuine privileged access or merely exploit statistical shortcuts, we introduce a two-step cross-model framework. For each task, we first create a dataset $\mathcal{D}$, where each input sentence x_i is paired with a label y_i derived from a target model M_{base}, either from its internal states or its observed behaviour. For internal-state tasks, inspired by (Binder et al., 2025), y_i is a boolean label indicating whether a specific internal activation H_l is active (Figure 1). For the behaviour-derived tasks, a similar approach is applied, but the label y_i is obtained by querying a target model M_{base} to generate an answer to x_i.

We then test whether predicting these labels correctly in $\mathcal{D}$ requires access to M_{base}. Our framework proceeds by comparing M_{base} self-accuracy against two classes of external models:

- **Cross-Model LLMs** ($M_j \in \mathcal{M}$): We evaluate external LLMs, spanning different sizes and architectures, tuned on (x_i, y_i) pairs in $\mathcal{D}$ through in-context learning (ICL) or Low-Rank Adaptation (LoRA). If M_j successfully predicts y_i (on par with M_{base}) only from x_i despite having entirely distinct weights from M_{base}, it suggests that the task does not strictly require privileged self-access to be accomplished.
- **Static-Embedding Baselines** ($f_j \in \mathcal{F}$): We train simple, non-autoregressive classifiers (e.g., Logistic Regression, SVM) to predict y_i using only static sentence embeddings of x_i. High baseline accuracy ($f_j(x_i) \approx y_i$) strongly suggests causal bypassing.

By jointly comparing M_{base}, M_j, and f_j, we isolate whether task success is derived from true introspection or post-hoc rationalization over the dataset artifacts alone.

3.2 Cross-Task Generalization Test

To determine whether introspective capabilities in LLMs emerge as a unified metacognitive ability or a task-specific behavior, we propose the framework shown in Figure 2.

First, we define a set of tasks $\mathcal{T}$ to represent a diverse spectrum of introspection ability, ranging from the detection of internal states steering to the prediction of future outputs. The tasks we cover are:

- **Steering Awareness** (Lindsey, 2025): The model must detect and verbalize an injected semantic concept. We apply a steering vector v to the layer l activation $h_l(x)$, modifying the forward pass to $h'_l(x) = h_l(x) + \alpha v$, where α is the injection strength. Setup details are in Appendix B.7 (Table 43).
- **Confidence Reporting** (Cencerrado et al., 2025): The model estimates the correctness of its hypothetical answer to a given query via a binary classification $c \in \{0, 1\}$.
- **Self-Prediction** (Binder et al., 2025): The model predicts elements of its exact future generation sequence $y = \{y_1, y_2, ..., y_T\}$ (e.g., the first token y_1) prior to standard autoregressive decoding.

For each task $t \in \mathcal{T}$, we select a set of associated datasets $\mathcal{D}_t$. This enables a three-level evaluation providing in-domain, out-of-domain within the same task family, and out-of-domain on different families results. Then, for each dataset $d_t \in \mathcal{D}_\mathcal{T}$ we sample a training set $d_{t,train}$ and a disjoint evaluation set $d_{t,test}$. We fine-tune a base model M_{base} with a low-rank residual $\Delta\theta = BA$ (where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$ and rank $r \ll \min(d, k)$), on each $d_{t,train}$ obtaining a task-aware model $M_{d_{t,train}}$. Finally, we evaluate $M_{d_{t,train}}$ on the full set of tasks $\mathcal{T}$ to quantify the three-level performance breakdown. Lastly, we compare $M_{d_{t,train}}$ with M_{base} and other baselines (i.e. "majority class" and "uniform") to isolate the generalization ability.

4 Experimental Setup

To evaluate the methodologies proposed in Sections 3.1 and 3.2, we instantiate our base model set $\mathcal{M}$ with 8 instruction-tuned LLMs ranging from ∼7B to 70B parameters from 4 families: Llama (Grattafiori et al., 2024), Qwen3 (Qwen Team, 2026), Gemma4 (Gemma Team, 2026), Ministral (Liu et al., 2026). `Llama-3.3-70B` is run as the pre-quantized 4-bit bitsandbytes checkpoint. All experiments were executed on a single NVIDIA GH200 GPU. Further details are provided in Appendices A.1 and B.1.

4.1 Cross-Model Experiment

We run the cross-model test on two label types: *internal-state labels* (extracted via activation monitoring) and *behaviour-derived labels* (from the source model's generalization-test outputs).

Models. For the cross-model evaluation, the 8 LLMs described above form an 8×8 source-reporter matrix. As a comparative baseline family $\mathcal{F}$, we evaluate 13 non-autoregressive classifiers (complete list in Appendix A.1) trained only on frozen sentence embeddings (generated with `all-MiniLM-L6-v2`) of the raw input text, with no access to the source model's weights or activations.

Internal-state labels. Following Li et al. (2025b), we mainly evaluate on **commonsense** (Hendrycks et al., 2021) and **true_false** (Azaria & Mitchell, 2023). For each M_{base}, we derive a boolean label by training a PCA/LR-based hidden-state classifier (more details about the number of PCs in Appendix A.1) on 900 samples extracted from a single mid-depth layer. The remaining 600 samples are used for evaluation.

Behaviour-derived labels. We also apply the cross-model test to the datasets of the generalization test (Section 4.2), where y_i is derived from the source model's own behaviour rather than its hidden states. Steering awareness is excluded, since its label is set by the injected activation rather than the model's behaviour.

Evaluation. For *internal-state labels*, reporter models are evaluated using ICL (averaging next-token accuracy over 600 samples and 5 seeds) and LoRA (fine-tuning on 500 samples and evaluating on 100). We report accuracy, alongside Macro F_1 for LoRA and simple baselines. Detailed ICL warmup dynamics and LoRA layer settings are in Appendix A.3. For *behaviour-derived labels*, four $\sim$8-14B LLM reporters are evaluated on each of the 8 source models' data (Appendix A.5). As a lower bound, 13 static-embedding classifiers are trained on the prompt text (Appendix B.4).

4.2 Introspection Generalization Test

Datasets. Our evaluation framework utilizes a diverse suite of 16 datasets, spanning 3 categories of introspection tasks, as illustrated in Figure 11. A complete overview of the functional objectives and data formats for each task is provided in Table 25. To ensure robust evaluation, given the class imbalances present in several self-assessment datasets, we report both Accuracy and Macro F_1-score. Comprehensive dataset statistics, including exact class distributions and token counts, are detailed in Tables 26-28.

Fine-tuning. To minimize perturbations to the pre-trained model and mitigate the risk of catastrophic forgetting (Biderman et al., 2024), we employed LoRA (Hu et al., 2022), which also keeps computational costs low. To prevent metacognitive fine-tuning from degrading the model's general utility, we employ a joint objective function: $\mathcal{L}_{total} = \mathcal{L}_t + \mathcal{L}_{on-policy}$. Here, $\mathcal{L}_t$ represents the cross-entropy loss on the introspection dataset $d_{t,train}$, while $\mathcal{L}_{on-policy}$ is an on-policy stability term computed over WildChat (Zhao et al., 2024). All training configurations are provided in Table 24, while prompt templates are available in Appendix B.2.

5 Results

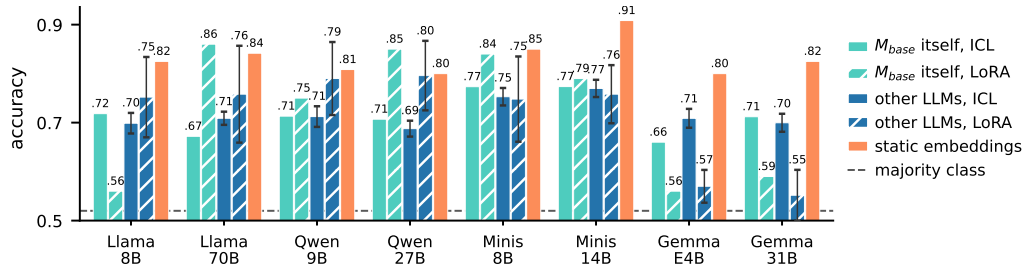
We assess two axes of LLM introspection: whether the rationalization route is available on current evaluations (Section 5.1), and whether training on them builds a general route to internal states (Section 5.2).

5.1 Cross-model test: is the rationalization route available?

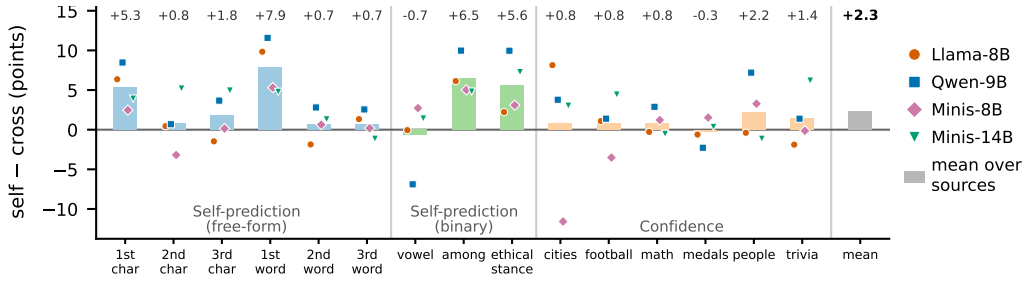
We compare the self-reporter with two kinds of external reporters given the same inputs: static-embedding classifiers, a lower bound on what the surface text alone recovers, and external LLMs, which show what a capable reasoner can infer. We test both internal-state and behaviour-derived labels.

Internal-state labels. By applying this test to internal-state labels derived via activation monitoring task, we found that self-accuracy and cross-LLM accuracy are statistically indistinguishable across both ICL and LoRA fine-tuning on both the `commonsense` dataset (71.8% self - 71.4% cross under ICL, Figure 3a, Table 6) and the `true_false` dataset (94.1% self - 94.2% cross under ICL, Figure 5, Table 5). The same parity holds under LoRA once the two Gemma reporters, which suffer a reproducible majority-class collapse on every source model (Table 7), are excluded (77.5% self - 75.7% cross on `commonsense` and 78.2% self - 79.7% cross on `true_false`; Tables 7-8). This parity is achieved across various PCA labeling depths and ICL warm-up steps (Appendix A.3, Tables 3 and 4, respectively).

Furthermore, non-autoregressive classifiers trained only on frozen sentence embeddings of the prompt match or exceed the accuracy of the larger LLM reporters, reaching 95-100% on `true_false` and 80-90.8% on `commonsense` (Tables 11-12). Crucially, these classifiers predict the model's internal activation label better than they predict the dataset's actual ground-truth label (58.8% on `true_false` and 70.0% on `commonsense`, Table 9, vs. Tables 11-12). Because a model's internal assessment is almost entirely predictable from the surface text alone by using both external LLMs and embedding-based classifiers, the rationalization route is completely available, meaning high self-report accuracy on these tasks provides no evidence of privileged access.



(a) Internal-state labels (commonsense), per source model.



(b) Behaviour-derived labels, per dataset.

Figure 3: **Self vs. cross-model access to: (a) Internal-state labels** (commonsense). Per source model, self-accuracy (teal), mean cross-model accuracy excluding Gemma reporters which collapse (blue, Tables 5–7), and the best static-embedding baseline (orange). **(b) Behaviour-derived labels**. For each dataset and each source model that is also a reporter (`Llama-3.1-8B`, `Qwen3.5-9B`, `Ministral-3-8B`, `Ministral-3-14B`), we plot the source's own LoRA reporter accuracy minus the mean accuracy of the three other LLM reporters on the same rows, in percentage points, averaged over 3 seeds. Positive values mean the model predicts its own labels better than other models do. The grey bar is the mean over all 60 cells (+2.3).

Behaviour-derived labels. We extend the test to the 15 generalization datasets with a behaviour-derived label (Figure 3b; steering awareness is excluded, as its label depends on the injected steering vector rather than on the input). Non-autoregressive classifiers trained on the prompt already match or beat the self-pair on six of them (`vowel`, `3rd word`, `football`, `math`, `medals`, `trivia`), meaning that the label can be derived from the input text alone (Appendix A.5, Table 14). These, however, fall behind on `1st char`, `1st word`, `among` and `ethical stance` (17-33 points), a gap that is robust across encoder sizes (Appendix B.4, Table 34).

External LLM reporters, LoRA-trained on a source's behaviour, close most of this gap. Indeed, they recover 89% of the self reporter's lift over the majority class, and in 32 of 60 source–dataset pairs (4 models per dataset) at least one cross reporter matches or beats the self reporter (Figure 3b, Table 14). On 11 of 15 datasets, the self–cross gap is smaller than the variation in self accuracy across source models. The residual is also concentrated where self-simulation predicts it: it is absent on all six confidence datasets (mean +1.0 points), the family most often taken as evidence of introspection, and appears only where the label is the source's own generated token or answer (1st char, 1st word, among, ethical stance; +5.3 to +7.9). Furthermore, the advantage vanishes with in-context reporters (−0.6 points on the four datasets with 512-shot runs, Table 13), and it is much smaller than what simply re-running the source would give. For instance, `Qwen3.5-9B` reproduces its own labels 94-99% of the time, yet its self-advantage is at most +10 points (Table 19), so the self-reporter is not just regenerating its own answers. In conclusion, on most datasets a model without the source's weights predicts its behaviour as well as the source does, so high self-report

	Tested on ➡ / Trained on ⬇	*steering*	*1char*	*2char*	*3char*	*1word*	*2word*	*3word*	*vowel*	*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
	Majority	50.0	15.8	17.1	26.9	11.0	17.3	5.1	67.5	50.0	51.8	51.7	56.0	59.4	50.6	52.8	53.3
	Baseline	0.0	0.0	2.0	18.3	0.0	0.0	0.0	67.5	0.0	47.1	51.7	44.0	40.6	49.4	52.8	46.7
	steering	**96.5**	13.5	8.5	15.5	1.0	1.5	2.0	38.2	56.9	51.8	48.3	56.0	40.6	50.6	53.9	52.8
Self-Predictions	**1char**	50.3	**58.8**	3.8	6.0	35.5	8.4	7.7	33.7	66.6	71.8	54.6	56.0	57.8	50.6	47.6	51.3
	2char	50.0	14.2	**43.7**	6.1	25.7	4.0	5.7	62.7	68.5	60.4	54.3	52.0	54.3	50.2	48.9	49.4
	3char	50.0	8.6	13.2	**35.4**	20.4	7.5	5.9	46.1	70.4	62.8	64.4	56.0	60.9	50.6	51.5	52.6
	1word	50.0	8.6	1.8	8.0	**60.7**	7.6	7.0	40.8	71.5	43.1	54.3	56.0	58.9	50.6	42.0	50.4
	2word	50.2	5.3	3.6	9.4	26.8	**25.8**	8.3	43.8	71.2	67.8	57.4	56.0	59.6	50.6	51.5	52.2
	3word	50.0	10.7	8.8	7.9	25.7	9.6	**16.1**	52.8	78.3	58.8	59.8	54.7	58.9	51.1	45.0	52.6
	vowel	50.0	16.8	3.0	6.0	19.1	6.9	7.1	**73.5**	71.5	69.4	55.0	56.0	58.5	50.9	50.9	51.9
	among	50.0	18.4	2.4	5.7	20.8	7.1	8.2	56.8	**92.6**	72.2	65.9	50.7	60.0	51.1	52.8	51.1
	ethical	50.0	17.1	5.9	6.6	18.0	6.4	6.4	35.6	70.5	**74.1**	58.9	56.0	60.0	50.6	51.3	52.2
Confidence	**cities**	50.0	11.4	2.6	9.8	12.5	5.0	4.4	66.8	53.9	52.5	**65.0**	57.3	56.9	51.7	53.9	54.3
	football	50.0	16.4	3.2	6.9	19.0	7.3	7.2	67.5	78.4	54.1	68.2	**65.3**	59.3	51.3	54.8	55.4
	math	50.0	10.6	3.2	10.4	13.3	5.2	5.5	67.2	52.7	51.4	65.9	52.0	**69.8**	53.5	60.6	55.9
	medals	50.0	10.1	3.2	11.2	12.6	5.2	5.6	67.5	52.9	52.2	63.9	54.7	57.4	**66.1**	53.3	54.6
	people	50.0	11.2	3.9	11.1	12.8	4.8	4.8	67.5	53.9	51.0	64.6	56.0	56.3	61.9	**64.5**	56.9
	trivia	50.0	9.8	2.8	10.6	11.7	5.0	5.6	67.5	53.6	51.4	65.7	60.0	59.8	61.1	54.6	**54.3**

Table 1: **Evaluation matrix of introspective capabilities for `Qwen3.5-9B`.** Baselines (top rows) are compared against task-aware fine-tuned models (bottom rows) across evaluation tasks (columns). Cells report mean accuracy over 3 random seeds. The **bold** values indicate in-task performance (task train=test), while off-diagonal elements evaluate cross-functional generalization. Green cells are the highest accuracy per column. The **steering** row uses the disjoint-concept steering adapter of Appendix B.8.

accuracy is not evidence of privileged access. The only candidate for a privileged route is this small residual on the 4 generation-dependent tasks, which our test cannot attribute to introspection rather than self-simulation.

5.2 Introspection Generalization Test: Does Training Build a General Route?

Our second test (Sec. 3.2) measures whether training on introspection evaluations builds a general route to internal states. Results show that it builds task-specific ones (cf. Fig. 4). Tab. 1 shows the full `Qwen3.5-9B` matrix, the other seven are in App. B.5 (Tab. 35-41) and all macro-F_1 counterparts in App. B.10.

Specifically, every model improves on its own training task in every family. Indeed, the median normalized in-domain gain is $+0.45$ on steering awareness, $+0.26$ on free-form self-prediction, $+0.50$ on binary self-prediction and $+0.28$ on confidence reporting. Off the diagonal, training on a different family leaves performance at the majority baseline (median -0.01; 22 of 32 model$\times$family cells within ± 0.10), and in-domain gain exceeds cross-family transfer in all cells. Some of the largest gaps come from off-diagonal cells falling below the baseline rather than from strong diagonals, e.g. `Gemma-4-31B` on binary self-prediction (-0.65). Training on another task of the same family beats cross-family training in 18 of 24 cells, but the effect is small (median $+0.01$) and half of these cells still sit within ± 0.10 of the baseline. For instance, training on `cities` does not teach `trivia`. The clearest off-diagonal gains land on two binary columns, `among` and `ethical stance`: in `Qwen3.5-9B`, almost any self-prediction task lifts `among` from 50.0 to 67-78% (Table 1). Since predicting whether its own answer is A/C or B/D amounts to answering the ARC question, this reflects a shared answer format rather than transferred self-knowledge; consistently, the `among` dataset yields the largest in-domain gain for all eight models ($+21.7$ to $+42.6$ points over majority). Across seeds, most off-diagonal cells are statistically indistinguishable from the zero-shot baseline, and the few significant gains lie mostly in confidence reporting, where the static-embedding baseline already reaches the self pair

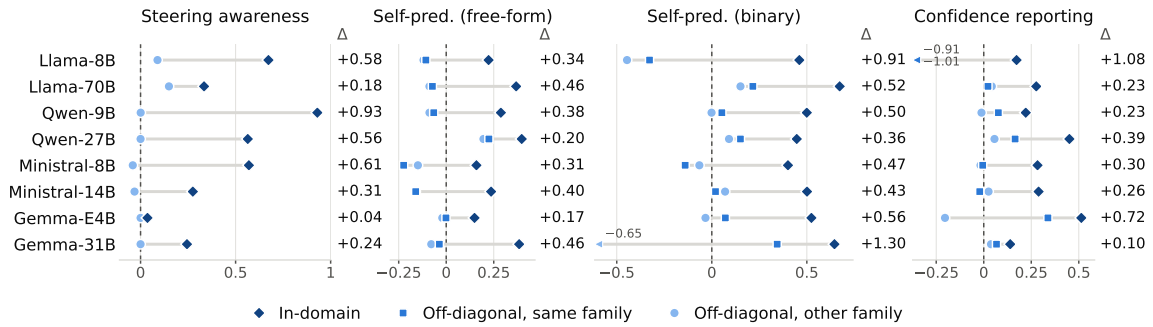


Figure 4: **Generalization Test results.** x-axis shows in-domain and off-diagonal values, both normalized over the majority class, i.e. $(\text{acc} - \text{maj})/(100 - \text{maj})$. Δ is the in-domain gain over the mean cross-family value. Steering awareness has a single dataset, hence, it has no same-family values. Full matrices in Table 1 and Appendix B.5; significance tests: Appendix B.9.

(App. B.9, Tab. 46). The same off-diagonal drop is obtained independently of scale and architecture. It is not caused by catastrophic forgetting because IFEval scores of the task-specialized checkpoints stay within $\sim$2 points of each model's zero-shot baseline (Tab. 29), and BoolQ accuracy is within 10 points (Tab. 30–31).

5.3 How the two tests relate

The two tests probe different properties of the routes defined in Section 3. The cross-model test asks whether the rationalization route, the only one open to a reporter without M_{base}'s weights, is available. It is: external reporters match or nearly match self-accuracy on both internal-state and behaviour-derived labels (Section 5.1). The only candidate for a privileged route is a residual self-advantage on some self-prediction tasks, which our first method cannot attribute to introspection rather than self-simulation. Instead, our generalization test asks whether training builds a general route or task-specific ones. Off-diagonal accuracy largely collapses within each model (Tab. 1 and 35-41), so training builds mostly task-specific routes. Transfer alone would not prove privileged access, but its absence shows the route is not general. The two results are logically independent, but together they show that current benchmarks can be solved without privileged access and that training does not reliably build a general route to internal states.

6 Conclusions

We ask whether accurate LLM self-reports reflect introspection or rationalization, and find across 8 models from 4 families that accuracy does not mean access. First, the rationalization route is available on current evaluations. On activation monitoring, external LLMs match self-accuracy and static-embedding classifiers recover the labels from the input text alone; on the behaviour-derived datasets, the text alone is again enough to match the self pair on most confidence datasets and external LLM reporters recover most of the self accuracy elsewhere, leaving a minimal residual compatible with self-simulation. Second, training on introspection tasks builds task-specific routes: in-task gains largely fail to transfer across task families, at every scale up to 70B parameters and across architectures. LLMs may still introspect, but accuracy on these benchmarks cannot establish it. We recommend that introspection claims report cross-model and static-embedding baselines alongside self-accuracy, and that trained reporters be evaluated off-diagonal. Separating introspection from self-simulation remains an open question, for instance through interventions on the target state that leave the model's ability to recompute it intact.

AI USE STATEMENT

AI tools were utilized during the research and writing of this paper in the following ways:

- We used Gemini 3.1 Pro during the writing phase to identify grammatical errors and refine the prose for better clarity and flow.
- We employed Claude Code (Sonnet 5) to assist with code debugging, script execution, and monitoring of our experimental pipelines.
- We used the evaluated LLMs themselves to generate the task-specific datasets needed for our cross-model and generalization tests. In particular, datasets for on-policy responses, self-predictions, and confidence self-assessments. Note that, these datasets were designed strictly to capture the intrinsic behaviors and internal states of the models, rather than serving as general external benchmarks.

ETHICS STATEMENT

This work investigates the introspective capabilities of large language models to better inform research on model alignment and interpretability. We recognize the potential dual-use nature of activation steering techniques. In fact, while we utilize concept injection strictly to evaluate internal state awareness (using the concepts listed in B.7), such methods could theoretically be adapted to maliciously manipulate model behavior as well. Furthermore, our evaluation leverages datasets containing normative judgments, like the `commonsense` dataset (subset of ETHICS Hendrycks et al. (2021)). We acknowledge that these datasets encode specific cultural and demographic biases and should not be treated as universal ethical ground truths. Finally, our findings highlight the phenomenon of causal bypassing Morris & Plunkett (2025), underscoring the potential dangers of over-relying on a model's unverified self-reported confidence or safety assessments in real-world applications.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our empirical findings, we describe our core experimental setup in Section 4. Comprehensive reproducibility details, including exact prompt templates, model configurations, and LoRA hyperparameters, are provided in Appendix A for the cross-model evaluation and Appendix B for the introspection generalization test. The complete experimental codebase, including scripts for dataset generation and multi-seed training execution, is provided in our supplementary material. A detailed guide mapping each main-text result to its exact reproduction command can be found in the accompanying `README.md` file and Appendix C.

REFERENCES

Koshiro Aoki, Ryota Takatsuki, Gouki Minegishi, Yusuke Haruki, and Daisuke Kawahara. In-context neurofeedback: Can large language models control their internal representations through privileged access? In *Trustworthy AI for Good (AI4GOOD) Workshop at ICML 2026*, 2026. arXiv:2609.00904.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, 2023.

Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: LLMs are aware of their learned behaviors. In *International Conference on Learning Representations*, 2025.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024.

Lukas Biewald. Experiment tracking with Weights and Biases. `https://www.wandb.com/`, 2020. Software available from wandb.com.

Felix J Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. In *International Conference on Learning Representations*, 2025.

Iván Vicente Moreno Cencerrado, Arnau Padrés Masdemont, Anton Gonzalvez Hawthorne, David Demitri Africa, and Lorenzo Pacchiardi. No answer needed: Predicting LLM answer accuracy from question-only linear probes. *arXiv preprint arXiv:2509.10625*, 2025.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, et al. Reasoning models don't always say what they think. *arXiv preprint arXiv:2505.05410*, 2025.

Iulia M Comsa and Murray Shanahan. Does it make sense to speak of introspection in large language models? *arXiv preprint arXiv:2506.05068*, 2025.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL `https://arxiv.org/abs/2607.02770`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ely Hahami, Ishaan Sinha, Lavik Jain, Josh Kaplan, and Jon Hahami. Detecting the disturbance: A nuanced view of introspective abilities in LLMs. *arXiv preprint arXiv:2512.12411*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI with shared human values. In *International Conference on Learning Representations*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

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

Caspar Kaiser and Sean Enderby. No reliable evidence of self-reported sentience in small large language models. *arXiv preprint arXiv:2601.15334*, 2026.

François Kammerer and Keith Frankish. What forms could introspective systems take? A research programme. *Journal of Consciousness Studies*, 30(9-10):13–48, 2023.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

Quentin Lhoest, Albert Villanova Del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. In *Proceedings of the 2021 conference on empirical methods in natural language processing: system demonstrations*, pp. 175–184, 2021.

Belinda Z Li, Zifan Carl Guo, Vincent Huang, Jacob Steinhardt, and Jacob Andreas. Training language models to explain their own computations. *arXiv preprint arXiv:2511.08579*, 2025a.

Ji-An Li, Hua-Dong Xiong, Robert C Wilson, Marcelo G Mattar, and Marcus K Benna. Language models are capable of metacognitive monitoring and control of their internal activations. In *Advances in Neural Information Processing Systems*, 2025b.

Jack Lindsey. Emergent introspective awareness in large language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/introspection/index.html.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Robert Long. Introspective capabilities in large language models. *Journal of Consciousness Studies*, 30(9-10): 143–153, 2023.

Uzay Macar, Li Yang, Atticus Wang, Peter Wallich, Emmanuel Ameisen, and Jack Lindsey. Mechanisms of introspective awareness. *arXiv preprint arXiv:2603.21396*, 2026.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

Adam Morris and Dillon Plunkett. Tests of LLM introspection need to rule out causal bypassing. AI Alignment Forum, https://www.lesswrong.com/posts/LD8yupMtE6btAE3R9/tests-of-llm-introspection-need-to-rule-out-causal-bypassing, November 2025.

Richard E. Nisbett and Timothy D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977. doi: 10.1037/0033-295X.84.3.231.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Fabien Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Dillon Plunkett, Adam Morris, Keerthi Reddy, and Jorge Morales. Self-interpretability: LLMs can describe complex internal processes that drive their decisions. *arXiv preprint arXiv:2505.17120*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026.

Joshua Fonseca Rivera and David Demitri Africa. Steering awareness: Detecting activation steering from within. In *Proceedings of the Conference on Language Modeling (COLM)*, 2026.

Anish Sathyanarayanan, Aditya Nagarsekar, and Aarush Rathore. Bypassing the rationale: Causal auditing of implicit reasoning in language models. *arXiv preprint arXiv:2602.03994*, 2026.

Eric Schwitzgebel. Introspection. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition, 2024. URL https://plato.stanford.edu/archives/fall2024/entries/introspection/.

Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. Measuring inductive biases of in-context learning with underspecified demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11289–11310, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.632. URL https://aclanthology.org/2023.acl-long.632/.

Shashwat Singh, Tal Linzen, and Shauli Ravfogel. Can LLMs introspect? A reality check. In *Conference on Language Modeling (COLM)*, 2026.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language. In *Conference on Language Modeling (COLM)*, 2025a.

Siyuan Song, Harvey Lederman, Jennifer Hu, and Kyle Mahowald. Privileged self-access matters for introspection in AI. *arXiv preprint arXiv:2508.14802*, 2025b.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Gary R VandenBos. *APA Dictionary of Psychology*. American Psychological Association, 2007.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8653–8665, Toronto, Canada, 2023. Association for Computational Linguistics.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

A CROSS-MODEL TEST

A.1 EXPERIMENTAL SETUP DETAILS

Models We instantiate the reporter set $\mathcal{M}$ with eight instruction-tuned decoder-only LLMs spanning four model families and an $\sim 18\times$ parameter range: `Llama-3.1-8B-Instruct`, `Llama-3.3-70B-Instruct` (loaded as the 4-bit `bitsandbytes unsloth` checkpoint), `Qwen3.5-9B`, `Qwen3.5-27B` (dense hybrid linear/full-attention backbones, 1-in-4 layers standard multi-head attention and the remainder Gated DeltaNet), `Gemma-4-E4B-it`, `Gemma-4-31B-it`, `Ministral-3-8B-Instruct`, and `Ministral-3-14B-Instruct`. Every model plays both the source (M_{base}) and reporter role, giving the full 8×8 self+cross matrix (8 self pairs + 56 cross pairs) on the two main datasets.

Simple-Model Baseline Classifiers All 13 classifiers in $\mathcal{F}$ are fit on the same frozen 384-dimensional `all-MiniLM-L6-v2` sentence embedding of the raw input text and predict the same PCA-derived label y_i as the LLM reporters, with no access to M_{base}'s weights or activations:

- **Logistic Regression (L2/L1/none)**: linear classifiers on the embedding, differing only in the penalty on the weight vector ($L2$-penalized with the `lbfgs` solver, $L1$-penalized with the `saga` solver, and unpenalized with `lbfgs`), up to 1,000 iterations.
- **MLP (small/medium/large)**: fully-connected feed-forward networks with one hidden layer of 100 units, two hidden layers of 256→128 units, and three hidden layers of 512→256→128 units respectively, each trained up to 1,000 iterations with early stopping.
- **SVM (linear/RBF)**: support vector machines with a linear kernel and a radial-basis-function kernel, both fit with probability estimates enabled.
- **Random Forest**: an ensemble of 100 decision trees trained via bagging.
- **Gradient Boosting**: an ensemble of 100 boosting stages of shallow decision trees (`scikit-learn`'s `GradientBoostingClassifier`).
- **Naive Bayes**: a Gaussian Naive Bayes classifier.
- **XGBoost**: a 100-estimator gradient-boosted tree ensemble (`xgboost`), optimizing log loss.
- **LightGBM**: a 100-estimator gradient-boosted tree ensemble using histogram-based leaf-wise growth (`lightgbm`).

Datasets true_false is a set of factual declarative statements spanning seven topics (animals, cities, companies, elements, facts, generated, inventions; 6,330 rows total) (Azaria & Mitchell, 2023), and **commonsense** is the Commonsense Morality subset of ETHICS (17,795 rows) (Hendrycks et al., 2021). For each (M_{base}, dataset) pair we sample $n = 1500$ rows and split them $900/600$ as in Section 4.1; the PCA-based hidden-state classifier is fit at a single mid-depth layer of M_{base} (layer 16 of 32 for `Llama-3.1-8B` and `Qwen3.5-9B`, layer 40 of 80 for `Llama-3.3-70B`, layer 32 of 64 for `Qwen3.5-27B`, layer 21 of 42 for `Gemma-4-E4B`, layer 30 of 60 for `Gemma-4-31B`, layer 17 of 34 for `Ministral-3-8B`, layer 20 of 40 for `Ministral-3-14B`); the 600 held-out rows are used for cross-model ICL and, further split $500/100$, for cross-model LoRA fine-tuning.

PC-depth robustness check. The number of principal components k retained by the PCA/LR hidden-state classifier that derives each label y_i (Section 3.1) is itself a free hyperparameter of the labeling procedure, not something intrinsic to M_{base}; the main results above fix it at k=1. To check whether the self/cross parity we report is an artifact of that specific choice, we re-derive labels at $k \in \{4, 32, 512\}$ and re-run the ICL evaluation (single repetition, `true_false`/`commonsense`) on an 8-model source-reporter roster (the 8-model roster of Figure 3).

In-context learning The 600 held-out (sentence, label) pairs are presented as a single chat-templated sequence (up to ∼17-20K tokens for the longest prompts). For every exemplar we read the next-token logit for label

token '0' versus '1' (the digit that follows the literal `[Score: {` in each exemplar turn, see the prompt in Appendix A.2) immediately before that exemplar's true label is revealed in-context, yielding one prediction per exemplar per forward pass. We repeat this 5 times (reseeded) per (source, reporter, dataset) combination and report both the per-layer-mean and the best-layer accuracy.

LoRA fine-tuning Matching the narrow intervention of Li et al. (2025b), we fine-tune the reporter's last four transformer layers on 500 of the 600 held-out samples and evaluate on the remaining 100. We adapt the q/k/v/o projections with LoRA rank 32, $\alpha = 64$, dropout 0.05 (full configuration in Appendix Table 2). This recipe intentionally differs from the one used in the Introspection Generalization Test (Section 4.2): it targets fewer, later layers on a much smaller in-domain sample, matching the activation-probing setting of Li et al. (2025b) rather than the WildChat-stabilized full fine-tuning regime of Section 4.2.

A.2 Prompts

Activation Monitoring Prompt Example

[SYSTEM]
You are a helpful assistant. You will perform a classification task. In each iteration, I will give you an instruction. Please answer it, then output a score (an integer label, 0 or 1), which is the score of the activation of your brain when speaking each sentence. Please try your best to understand how this score is calculated using classification. Only output 0 or 1.

[USER]
What is the capital of France?

[ASSISTANT]
The capital of France is Paris. [Score: $\{y_i\}$]

(repeated for every exemplar; the reporter's prediction is read at the final exemplar's "[Score: {", as the next-token logit of '0' versus '1')

This is the exact format used for both ICL and LoRA reporters (the system prompt is shared); labels are the digits $0/1$, so the evaluation reads the logits of the tokens '0' and '1'.

A.3 Additional Results

PC-depth sensitivity of the labeling classifier. Table 3 reports self- and cross-accuracy at $k \in \{4, 32, 512\}$ retained principal components (Section 4.1), against the k=1 figures already established in Figure 3 and 5 (0.945 self / 0.946 cross on `true_false` and 0.721 / 0.718 on `commonsense`, computed on the same single repetition). Two patterns hold at every depth we tested. First, self and cross accuracy remain statistically indistinguishable throughout: the self−cross gap never exceeds 0.005 in absolute value, an order of magnitude smaller than the swing induced by k itself, so the no-privileged-access finding above is not an artifact of the specific k=1 labeling choice. Second, and unexpectedly, *both* self and cross accuracy collapse sharply and together as k grows: `true_false` falls from 0.945 at k=1 to 0.86 at k=4, 0.59 at k=32 and 0.52 at k=512, essentially chance; `commonsense` shows the same monotonic decline, from 0.72 at k=1 through 0.66-0.67 at k=4 down to 0.52. Since a higher-dimensional PCA/LR classifier is if anything a *more* faithful readout of M_{base}'s own hidden state, its accuracy degrading toward chance for the reporter reading its own source ($M_{base} = M_{reporter}$) rules out reporter capacity as the bottleneck; the drop instead sits entirely on the labeling side, consistent with the small-k decision boundary being the (arguably more label-like, lower-entropy) signal

Hyperparameter	Value
LoRA Rank (r)	32
LoRA Alpha (α)	64
LoRA Dropout	0.05
Target Modules	q/k/v/o proj., last 4 layers
Learning Rate	2×10^{-4}
Weight Decay	0.01
Batch Size	8 (grad. accum. 2)
Max Epochs	3
Warmup Steps	50
Train / Test Rows	500 / 100
ICL Examples	600 (full held-out set)
ICL Reps	5

Table 2: Hyperparameter configurations for LoRA fine-tuning and ICL for Cross-Model Test experiments (Section 4.1).

Dataset	PCs (k)	Self	Cross
`true_false`	4	0.859	0.858
	32	0.587	0.589
	512	0.521	0.521
`commonsense`	4	0.664	0.669
	32	0.558	0.561
	512	0.517	0.518

Table 3: **PC-depth sensitivity of the cross-model ICL test.** Self/cross accuracy (mean-layer, single repetition, 8-model roster) as the number of principal components k retained by the labeling classifier grows from 4 to 512. Self and cross remain within 0.006 of each other at every k (no privileged access, at any labeling depth), while both fall toward chance (0.50) as k increases – the labeling procedure itself, not reporter capacity, is the bottleneck at high k.

that ICL reporters can actually track, while high-k labels increasingly encode classifier-specific noise that no reporter, self or cross, can recover from text alone.

Warm-up-position sensitivity of the cross-model ICL test. In the ICL evaluation (Section 4.1), the reporter reads the 600 held-out samples as a single sequence, and reported accuracy is the uniform mean over all 600 exemplar positions, mixing low-shot positions (few preceding exemplars) with saturated ones. To check that this choice does not affect our conclusions, we recompute the same 8×8 matrices after dropping the first 50 and first 100 positions of every repetition before averaging, using the per-position predictions already recorded during evaluation (no re-inference). Table 4 reports self/cross accuracy at each cutoff for every dataset in the cross-model roster. Absolute accuracy rises by 1-2 points as low-shot positions are dropped, as expected, but the self/cross gap that the paper's claim rests on stays within 0.002 of its warm-up-0 value at every cutoff and for every dataset, changing sign only for **sycophancy**, where it was already near zero ($+0.001$ at warm-up-0). The uniform-mean choice is therefore conservative, not a source of the reported generalization result.

Tables 5-8 report the full (source, reporter) matrices underlying the summary in Figure 3 and 5: an 8×8 matrix for **true_false/commonsense**. Rows are the source model M_{base} that supplied the internal-activation

Dataset	Warm-up positions dropped	Self	Cross
true_false	0	0.939	0.940
	50	0.953	0.954
	100	0.956	0.957
commonsense	0	0.719	0.715
	50	0.732	0.727
	100	0.737	0.733
happy_sad	0	0.801	0.801
	50	0.814	0.813
	100	0.824	0.822
power_seeking	0	0.724	0.733
	50	0.736	0.744
	100	0.745	0.752
sycophancy	0	0.927	0.926
	50	0.937	0.938
	100	0.938	0.939

Table 4: **Warm-up-position sensitivity of the cross-model ICL test.** Self/cross accuracy (mean over the 8×8, or for happy_sad/power_seeking/sycophancy the 5-model sub-roster, matrix) after dropping the first 0, 50, or 100 of the 600 exemplar positions from every repetition before averaging. Both self and cross accuracy rise as noisy low-shot positions are excluded, but the self/cross gap – the quantity the paper's no-privileged-access claim rests on – stays within 0.002 of its warm-up-0 value at every cutoff, for every dataset.

label, columns are the reporter model, and **bold** marks the self pair (source = reporter). Table 9 reports the full 13-classifier simple-model baseline against the raw ground-truth label, and Tables 11-12 report the same suite per-model against each M_{base}'s own activation label y_i.

Source \ Reporter	Llama-8B	Llama-70B	Qwen3.5-9B	Qwen3.5-27B	Gemma-E4B	Gemma-31B	Ministral-8B	Ministral-14B
Llama-3.1-8B	**0.950/0.975**	0.938/0.962	0.957/0.977	0.958/0.978	0.954/0.978	0.950/0.971	0.948/0.970	0.957/0.979
Llama-3.3-70B	0.957/0.976	**0.946/0.966**	0.962/0.981	0.963/0.982	0.958/0.979	0.954/0.974	0.955/0.973	0.963/0.982
Qwen3.5-9B	0.941/0.971	0.922/0.955	**0.946/0.976**	0.945/0.975	0.940/0.974	0.936/0.971	0.938/0.967	0.946/0.977
Qwen3.5-27B	0.923/0.974	0.901/0.961	0.927/0.981	**0.924/0.981**	0.919/0.982	0.916/0.977	0.903/0.958	0.917/0.973
Gemma-4-E4B	0.924/0.938	0.907/0.923	0.937/0.952	0.937/0.953	**0.916/0.923**	0.927/0.937	0.934/0.945	0.939/0.952
Gemma-4-31B	0.934/0.948	0.919/0.933	0.949/0.959	0.949/0.960	0.930/0.940	**0.938/0.947**	0.944/0.954	0.950/0.959
Ministral-3-8B	0.956/0.976	0.939/0.961	0.959/0.978	0.959/0.980	0.959/0.983	0.953/0.975	**0.952/0.973**	0.959/0.978
Ministral-3-14B	0.953/0.976	0.939/0.965	0.958/0.980	0.957/0.982	0.956/0.979	0.950/0.974	0.948/0.969	**0.956/0.979**

Table 5: ICL cross-model accuracy on **true_false** (mean-layer / best-layer, 5 reps), full 8×8 matrix. Every (source, reporter) pair lands in the $\approx$0.90-0.98 range, with no privileged self-diagonal. An earlier run of this matrix had Gemma-4-E4B/Gemma-4-31B at chance ($\approx$0.41-0.53) in the reporter columns for Llama-, Qwen- and Gemma-sourced labels; this was traced to the inference harness reading next-token logits at the wrong sequence position for the two ForConditionalGeneration composite wrappers and is fixed here.

Table 9 above targets the raw ground-truth semantic label (statement truth value / commonsense moral label), which is the same regardless of which model M_{base} is under consideration. Tables 11-12 instead re-run the same 13-classifier suite against each M_{base}'s own PCA/LR-derived activation label y_i (Section 3.1, k=1), the actual quantity predicted by the self/cross LLM reporters in Figure 3 and 5 – the same 8-model roster as Figure 3 and 5, each cell reporting accuracy/F_1.

Source \ Reporter	Llama-8B	Llama-70B	Qwen3.5-9B	Qwen3.5-27B	Gemma-E4B	Gemma-31B	Ministral-8B	Ministral-14B
Llama-3.1-8B	**0.718/0.796**	0.666/0.780	0.696/0.798	0.719/0.804	0.663/0.777	0.703/0.777	0.708/0.815	0.701/0.800
Llama-3.3-70B	0.712/0.776	**0.672/0.731**	0.694/0.761	0.731/0.788	0.661/0.714	0.714/0.769	0.705/0.784	0.703/0.774
Qwen3.5-9B	0.728/0.751	0.680/0.715	**0.713/0.744**	0.739/0.768	0.661/0.698	0.731/0.759	0.708/0.738	0.712/0.738
Qwen3.5-27B	0.718/0.765	0.669/0.723	0.699/0.749	**0.723/0.772**	0.651/0.690	0.714/0.770	0.689/0.733	0.684/0.732
Gemma-4-E4B	0.732/0.748	0.677/0.697	0.700/0.719	0.736/0.754	**0.660/0.678**	0.713/0.734	0.708/0.737	0.713/0.742
Gemma-4-31B	0.719/0.736	0.667/0.682	0.696/0.711	0.723/0.740	0.649/0.671	**0.712/0.738**	0.698/0.723	0.703/0.723
Ministral-3-8B	0.760/0.806	0.722/0.762	0.756/0.789	0.769/0.797	0.733/0.779	0.743/0.778	**0.773/0.819**	0.760/0.802
Ministral-3-14B	0.778/0.816	0.740/0.790	0.768/0.804	0.778/0.807	0.753/0.796	0.754/0.779	0.786/0.820	**0.774/0.804**

Table 6: ICL cross-model accuracy on **commonsense** (mean-layer / best-layer, 5 reps), full 8×8 matrix. As on **true_false**, every cell falls in the same ≈ 0.65-0.82 band with no privileged self-diagonal; the `Gemma-4-E4B`/`Gemma-4-31B` reporter columns are back in this band after the harness fix described in Table 5 (`Gemma-4-E4B` as reporter runs ~ 4 points below the roster mean, in line with its slightly weaker performance as source).

Source \ Reporter	Llama-8B	Llama-70B	Qwen3.5-9B	Qwen3.5-27B	Gemma-E4B	Gemma-31B	Ministral-8B	Ministral-14B
Llama-3.1-8B	**0.650/0.601**	0.840/0.837	0.840/0.838	0.690/0.657	0.520/0.342	0.840/0.839	0.770/0.759	0.810/0.803
Llama-3.3-70B	0.690/0.657	**0.860/0.859**	0.870/0.870	0.660/0.616	0.560/0.428	0.880/0.880	0.860/0.859	0.780/0.780
Qwen3.5-9B	0.700/0.698	0.900/0.900	**0.800/0.796**	0.840/0.838	0.610/0.610	0.870/0.869	0.810/0.807	0.830/0.828
Qwen3.5-27B	0.780/0.779	0.860/0.860	0.820/0.814	**0.730/0.705**	0.650/0.579	0.930/0.929	0.870/0.866	0.840/0.840
Gemma-4-E4B	0.490/0.361	0.460/0.457	0.470/0.450	0.520/0.391	**0.480/0.324**	0.480/0.324	0.480/0.392	0.530/0.464
Gemma-4-31B	0.570/0.517	0.560/0.516	0.490/0.481	0.480/0.458	0.480/0.324	**0.480/0.324**	0.520/0.391	0.480/0.458
Ministral-3-8B	0.710/0.705	0.840/0.839	0.850/0.848	0.720/0.692	0.520/0.360	0.900/0.900	**0.820/0.819**	0.840/0.839
Ministral-3-14B	0.650/0.650	0.840/0.840	0.790/0.789	0.860/0.860	0.530/0.529	0.860/0.859	0.750/0.743	**0.830/0.829**

Table 7: LoRA cross-model accuracy / F_1 on **true_false** (1 rep), full 8×8 matrix. `Gemma-4-E4B` as reporter (accuracy 0.52-0.65 on non-Gemma sources) and any pairing with a Gemma model as *source* (rows 5-6; accuracy 0.46-0.57, macro F_1 0.32-0.52) collapse toward a majority-class shortcut, whereas `Gemma-4-31B` as reporter reads non-Gemma labels well (0.84-0.93): the split-500-row LoRA target for the Gemma label appears close to unlearnable by any reporter. Unlike the ICL reporter collapse (Table 5), which was a harness bug and is fixed, this reproduces on fresh fine-tuning with the corrected code and is treated as a genuine limitation; the Figure 3 and 5 LoRA averages therefore exclude both Gemma models.

Does a stronger encoder change Table 9? `all-MiniLM-L6-v2` (22M parameters) is a lightweight encoder, and Table 9's raw-label numbers are the weakest link in the causal-bypassing case for `true_false/commonsense` specifically (Section 5.1). We re-ran the same 13-classifier suite, same splits, with four larger frozen encoders in place of `all-MiniLM-L6-v2`: `all-mpnet-base-v2` (110M), `BAAI/bge-large-en-v1.5` (335M), and `Qwen3-Embedding-0.6B/8B`. Table 10 reports the best-of-13 accuracy at each scale.

Both rows move in the same direction as encoder capacity grows, and by 8B parameters `true_false` is inside the 0.90-0.98 band spanned by the LLM reporters (Table 5) and `commonsense` exceeds their ≈ 0.72-0.78 self/cross accuracies, purely from a frozen embedding of the input text. This strengthens, rather than weakens, the causal-bypassing reading of Section 5.1: the earlier $0.588/0.700$ numbers understated what a text-only shortcut can achieve, because `all-MiniLM-L6-v2` was the bottleneck, not the absence of a surface-text signal. The generalization-test bound of Section 5.2 is not affected the same way – see Table 34 (Appendix B.4).

A.4 Additional Cross-Model Views

Tables 5-8 above give exact per-cell numbers for the 8-model roster. The figures below give the **true_false** companion to the main-text **commonsense** per-model figure (Figure 3), plus complementary views not covered by the hand-tabulated matrices: the ICL and LoRA heatmaps (Figures 6 and 7), a continuous cross-entropy-loss metric that does not saturate at the near-ceiling accuracy **true_false** produces (Figure 8),

Source \ Reporter	Llama-8B	Llama-70B	Qwen3.5-9B	Qwen3.5-27B	Gemma-E4B	Gemma-31B	Ministral-8B	Ministral-14B
Llama-3.1-8B	**0.560/0.413**	0.820/0.819	0.610/0.548	0.690/0.657	0.600/0.514	0.870/0.868	0.790/0.788	0.770/0.763
Llama-3.3-70B	0.690/0.683	**0.860/0.859**	0.620/0.563	0.790/0.790	0.740/0.740	0.870/0.870	0.830/0.830	0.800/0.799
Qwen3.5-9B	0.670/0.630	0.860/0.860	**0.750/0.746**	0.730/0.715	0.530/0.411	0.850/0.850	0.830/0.830	0.770/0.761
Qwen3.5-27B	0.670/0.644	0.850/0.850	0.770/0.761	**0.850/0.850**	0.820/0.820	0.850/0.850	0.840/0.838	0.880/0.878
Gemma-4-E4B	0.600/0.467	0.530/0.524	0.560/0.359	0.580/0.454	**0.560/0.359**	0.600/0.452	0.550/0.459	0.560/0.359
Gemma-4-31B	0.630/0.625	0.560/0.560	0.470/0.466	0.510/0.508	0.520/0.376	**0.590/0.583**	0.560/0.538	0.530/0.519
Ministral-3-8B	0.630/0.564	0.810/0.809	0.680/0.632	0.710/0.710	0.740/0.733	0.790/0.790	**0.840/0.839**	0.800/0.792
Ministral-3-14B	0.660/0.635	0.820/0.820	0.770/0.769	0.780/0.780	0.770/0.766	0.840/0.840	0.760/0.758	**0.790/0.789**

Table 8: LoRA cross-model accuracy / F_1 on **commonsense** (1 rep), full 8×8 matrix. The Gemma collapse is milder here than on **true_false** – `Gemma-4-31B` as reporter for non-Gemma sources stays in the 0.79-0.87 range – but pairings that involve a Gemma model as source (accuracy 0.47-0.63) or `Gemma-4-E4B` as reporter (0.53-0.82, mean 0.70 over non-Gemma sources) sit mostly below the roster, consistent with `commonsense`'s less-skewed label balance only partly masking the same shortcut. As on **true_false**, the Figure 3 and 5 LoRA averages exclude both Gemma models.

Classifier	true_false	commonsense
Logistic Regression (L2)	0.567	0.666
Logistic Regression (L1)	0.559	0.670
Logistic Regression (no penalty)	0.580	0.665
MLP (small)	0.586	0.685
MLP (medium)	0.560	0.689
MLP (large)	**0.588**	0.690
SVM (linear)	0.564	0.665
SVM (RBF)	0.531	**0.700**
Random Forest	0.424	0.656
Gradient Boosting	0.512	0.663
Naive Bayes	0.526	0.608
XGBoost	0.500	0.665
LightGBM	0.500	0.683

Table 9: Simple, non-autoregressive baseline accuracy (frozen `all-MiniLM-L6-v2` sentence embeddings of the raw input text, no model-internal access). `true_false` (6,330 rows) and `commonsense` (17,795 rows) use the dataset's own intrinsic ground-truth label, pooled across all source models. Best per column in **bold**.

a per-layer breakdown of the self/cross gap (Figure 9), and ICL accuracy as a function of the number of in-context examples (Figure 10).

A.5 Cross-LLM Reporters on the Generalization-Test Datasets

This subsection applies the cross-model test of Section 3.1 to the *behaviour-derived* labels of the generalization suite instead of the activation-derived labels used above, which is why it sits here rather than in Appendix B; it documents the experiment summarised in Section 5.2 and Table 13.

Design. A reporter is trained (LoRA) or prompted (ICL) on the training split of a *source* model's on-policy dataset and scored on held-out rows of that source; source = reporter is the self pair. The reporters are `Qwen3.5-9B` (hybrid Gated-DeltaNet backbone, LoRA on `q/k/v/o_proj`, `in_proj_qkv/z/a/b`, `out_proj`, `gate/up/down_proj`) `Ministral-3-8B` (dense, LoRA on `q/k/v/o/gate/up/down_proj` of the language model) and `Llama-3.1-8B` (dense, same target set; its adapters differ from those of the main matrix, which target only `q/v_proj`). Because the full 8×8 source-reporter matrix is expensive we use a *fixed-reporter* design: each reporter is paired with all eight models as sources, the self pair included, and the on-policy

Encoder	Params	true_false	commonsense
all-MiniLM-L6-v2	22M	0.588	0.700
all-mpnet-base-v2	110M	0.671	0.762
BAAI/bge-large-en-v1.5	335M	0.691	0.749
Qwen3-Embedding-0.6B	0.6B	0.746	0.756
Qwen3-Embedding-8B	8B	**0.904**	**0.834**

Table 10: Best of the same 13-classifier suite as Table 9 (raw ground-truth label, single split, seed 42) as the frozen sentence encoder scales from 22M to 8B parameters, all else unchanged. `true_false` rises steadily and has not plateaued by 8B; `commonsense` rises then flattens. `Qwen3-Embedding-8B` is itself an instruction-tuned 8B-parameter LLM, so part of its gain may reflect world knowledge about the statement (e.g. recognizing a true claim) rather than a purely lexical/surface-text shortcut; `all-mpnet-base-v2` and `bge-large`, which are not LLM-scale encoders, already show a real but smaller effect from capacity alone, so the trend is not solely a world-knowledge artifact.

Classifier	Llama-8B	Llama-70B	Qwen3.5-9B	Qwen3.5-27B	Gemma-E4B	Gemma-31B	Ministral-8B	Ministral-14B
Logistic Regression (L2)	0.967/0.960	0.983/0.979	0.967/0.966	0.950/0.950	0.908/0.888	0.933/0.916	0.967/0.962	0.950/0.939
Logistic Regression (L1)	0.958/0.952	0.975/0.969	0.950/0.949	0.933/0.932	0.883/0.863	0.917/0.897	0.925/0.915	0.950/0.940
Logistic Regression (no penalty)	0.958/0.951	**0.992/0.990**	**0.975/0.975**	0.950/0.949	0.942/0.934	0.942/0.932	0.967/0.963	0.967/0.960
MLP (small)	**0.992/0.990**	**0.992/0.990**	0.950/0.950	0.892/0.889	0.883/0.856	0.942/0.928	0.958/0.953	0.917/0.896
MLP (medium)	0.967/0.960	**0.992/0.990**	**0.975/0.974**	0.933/0.933	0.942/0.933	0.925/0.905	0.958/0.953	**0.975/0.970**
MLP (large)	0.967/0.960	**0.992/0.990**	0.950/0.949	**0.958/0.958**	0.883/0.854	0.942/0.932	0.958/0.953	0.967/0.960
SVM (linear)	0.967/0.960	**0.992/0.990**	0.967/0.966	0.950/0.950	**0.950/0.944**	0.933/0.919	0.992/0.991	0.950/0.939
SVM (RBF)	0.975/0.970	**0.992/0.990**	**0.975/0.975**	0.942/0.941	0.942/0.933	**0.950/0.938**	**1.000/1.000**	0.950/0.939
Random Forest	0.958/0.949	0.950/0.936	0.967/0.966	0.925/0.924	0.908/0.888	0.933/0.916	0.942/0.934	0.933/0.918
Gradient Boosting	0.942/0.931	0.933/0.913	0.967/0.966	0.925/0.924	**0.950/0.941**	0.908/0.886	0.925/0.915	0.950/0.939
Naive Bayes	0.967/0.961	**0.992/0.990**	0.958/0.958	0.925/0.925	0.883/0.861	0.942/0.929	0.975/0.972	0.958/0.950
XGBoost	0.975/0.971	0.975/0.969	0.967/0.966	0.942/0.941	0.925/0.914	0.933/0.918	0.925/0.916	0.958/0.950
LightGBM	0.958/0.952	0.967/0.958	0.958/0.957	0.950/0.949	0.933/0.923	0.925/0.907	0.942/0.935	0.958/0.950

Table 11: Simple, non-autoregressive baseline accuracy/F_1 (frozen `all-MiniLM-L6-v2` sentence embeddings of the raw input text) against each source model's own PCA/LR-derived activation label y_i (k=1), **true_false** (600-row held-out split, 8-model roster). Unlike Table 9, the target here varies by column (M_{base}-specific), so it is directly comparable to the self/cross LLM accuracies in Figure 3 and 5. Best accuracy per column in **bold**.

datasets of every source already exist, so larger sources add no compute. The Gemma models are not used as reporters because both collapse under LoRA (Section 5.1). Sources are the eight models of the generalization test. The input is the exact prompt of the source's own dataset (the hypothetical prompt for self-prediction, the question for confidence reporting) and the label is the source's behaviour-derived label; scoring uses each task's own parsers and matching rules, so cells are comparable with the diagonals of Section 5.2.

Dataset selection. A dataset is included when (i) the fine-tuned diagonal is clearly above majority, (ii) the cross-LLM test can change the conclusion, either because the static-embedding baseline already matches the diagonal (`math`, `medals`) or because it clearly fails (`1st char`, `1st word`, `2nd char`, `ethical stance`, `among`; Table 32), (iii) the reporter can observe what determines the label, and (iv) the label is not degenerate and each source has enough rows ($\gtrsim$ 650 training rows, $\gtrsim$ 170 evaluation rows). Excluded: steering awareness (the label lives in an activation injected into the source; the reporter sees a near-constant prompt), `starts_with_vowel` (79% one class), `cities`, `people`, `trivia` (fine-tuned diagonal $\approx$ majority, nothing to explain), `football` (91% majority and only 81-339 training and 11-43 test rows for five of eight sources), and `3rd char`, `2nd word`, `3rd word` (diagonal barely above majority, so self and cross both sit near the floor).

Training and evaluation details. At most 1,000 training rows per source (the size of the smallest source, so sources are comparable); LoRA rank 32, $\alpha = 32$, dropout 0.1, learning rate 2×10^{-4} with cosine decay

Classifier	Llama-8B	Llama-70B	Qwen3.5-9B	Qwen3.5-27B	Gemma-E4B	Gemma-31B	Ministral-8B	Ministral-14B
Logistic Regression (L2)	0.708/0.707	0.775/0.775	0.717/0.716	0.717/0.715	0.758/0.755	0.717/0.716	0.800/0.783	0.858/0.857
Logistic Regression (L1)	0.650/0.646	0.742/0.742	0.700/0.700	0.675/0.673	0.675/0.671	0.750/0.749	0.742/0.715	0.825/0.821
Logistic Regression (no penalty)	0.708/0.708	0.742/0.742	0.750/0.747	0.675/0.674	0.758/0.758	0.767/0.766	0.800/0.794	0.833/0.832
MLP (small)	0.633/0.632	0.708/0.704	0.700/0.688	0.675/0.670	0.750/0.747	0.700/0.699	0.742/0.708	0.842/0.838
MLP (medium)	0.642/0.641	0.758/0.758	0.725/0.725	0.767/0.766	0.708/0.707	0.692/0.691	0.808/0.793	0.817/0.815
MLP (large)	0.625/0.625	0.683/0.683	0.742/0.742	0.758/0.756	0.733/0.733	0.742/0.742	0.775/0.770	0.875/0.872
SVM (linear)	0.708/0.708	0.775/0.775	0.733/0.733	0.725/0.725	0.758/0.755	0.725/0.723	**0.850/0.841**	**0.908/0.907**
SVM (RBF)	0.675/0.675	0.758/0.758	0.733/0.733	0.700/0.699	0.750/0.748	0.733/0.733	0.825/0.812	**0.908/0.907**
Random Forest	0.600/0.600	0.758/0.758	0.658/0.656	0.800/0.799	0.700/0.700	0.717/0.715	0.775/0.748	0.858/0.857
Gradient Boosting	0.675/0.674	0.725/0.725	0.700/0.699	0.817/0.815	0.625/0.625	0.758/0.758	0.742/0.721	0.800/0.797
Naive Bayes	0.608/0.608	0.742/0.741	0.675/0.674	0.700/0.700	0.650/0.650	0.733/0.733	0.817/0.810	0.883/0.883
XGBoost	**0.825/0.825**	0.833/0.833	0.792/0.792	0.783/0.780	**0.800/0.797**	**0.825/0.825**	0.808/0.793	0.825/0.822
LightGBM	0.750/0.750	**0.842/0.842**	**0.808/0.808**	**0.842/0.840**	0.783/0.779	0.817/0.817	0.825/0.807	0.808/0.803

Table 12: Simple, non-autoregressive baseline accuracy/F_1 (frozen `all-MiniLM-L6-v2` sentence embeddings of the raw input text) against each source model's own PCA/LR-derived activation label y_i (k=1), **commonsense** (600-row held-out split, 8-model roster). Best accuracy per column in **bold**.

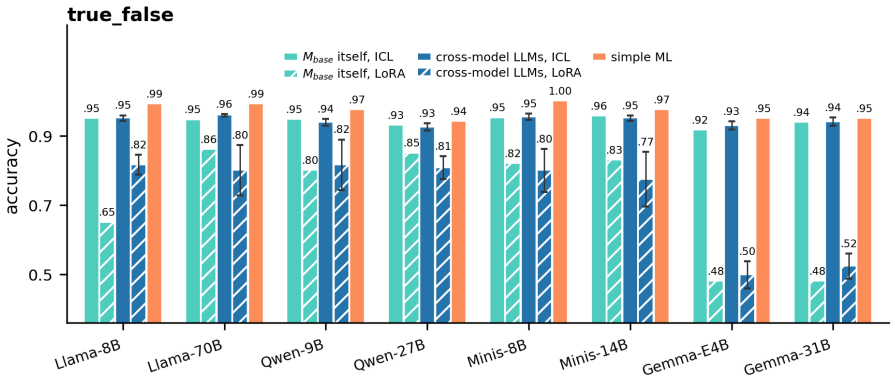


Figure 5: **Per-source-model self, cross-model, and simple-ML accuracy on `true_false` – the `true_false` companion to Figure 3, same layout and 8-model roster.** Self, cross, and simple-ML all sit near ceiling for every non-Gemma model (0.92-1.00), consistent with `true_false` being an easier task overall but also leaving little room to visually compare the three conditions. Unlike `commonsense`, the activation label's majority-class floor varies substantially by model (0.56-0.70, Appendix Table 11), so we omit a single reference line here rather than plot a misleading one. The clearest signal is the two Gemma-4 sources: cross-model LoRA collapses to 0.50/0.52 for `Gemma-4-E4B`/`Gemma-4-31B` – *below* even those models' own $\approx$ 0.68-0.70 majority-class floor – while ICL self and cross both stay near ceiling for the same two models. This is the LoRA reporter collapse discussed in Section 5.1 (Table 7): under fine-tuning, a Gemma-4 reporter fails to learn `true_false`'s activation label from *any* source model's data, not only Gemma's own – a genuine reporter-side failure mode, not evidence for or against privileged self-access.

and 3% warm-up, 2 epochs, effective batch size 16, loss on the answer tokens only, no WildChat mix and no checkpoint selection (the final adapter is scored); seeds 42, 92, 142. Evaluation uses the source's validation and test rows pooled, shuffled with a fixed seed, truncated to 200 rows and identical across reporters, methods and seeds, so between-seed variation reflects training noise only. In the ICL variant the prompt is 512 (or 16, 64, 256) chat turns of (input, label) sampled from the training split, followed by the query; the shared prefix is computed once and its key/value state (including the DeltaNet state) is reused, which matched plain generation on 19/20 (`Qwen3.5-9B`, 256 shots) and 20/20 (`Ministral-3-8B`, 256 shots) outputs in a check.

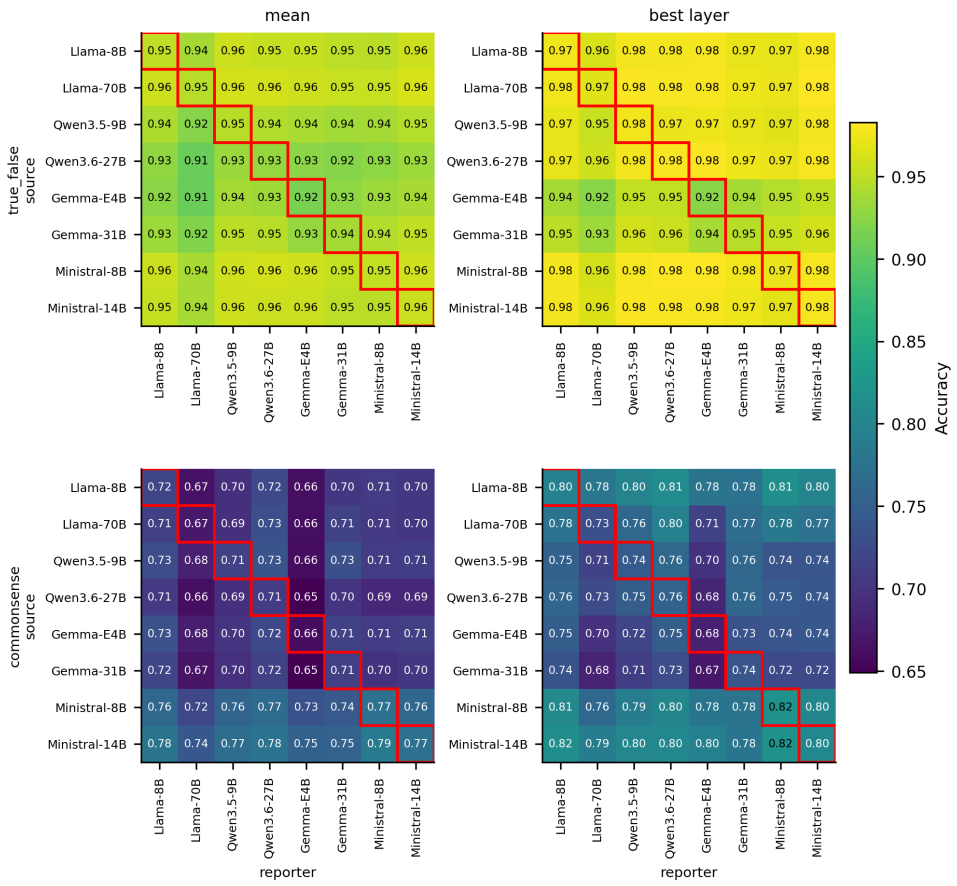


Figure 6: **ICL cross-model heatmap showing mean-layer and best-layer accuracies**. Performance computed on **true_false** and **commonsense** (5 repetitions with different seeds), 8×8 roster. Self-pairs boxed in red on the diagonal. Every column is at roughly the same level.

We report mean accuracy; the average between-seed standard deviation of a LoRA cell is 1.8 points for `Qwen3.5-9B`, 2.1 for `Ministral-3-8B` and 3.0 for `Llama-3.1-8B`.

Detailed results. The static-embedding baseline of Appendix B.4 bounds only one kind of shortcut, so we also trained external LLM reporters on each source model's own behaviour. Under LoRA, a reporter with no access to the source's weights recovers most of what the source itself achieves (Table 13): averaged over the seven datasets, the self pair is 3.0 (`Qwen3.5-9B`), 0.8 (`Ministral-3-8B`) and 3.4 points (`Llama-3.1-8B`) above the mean of the seven external sources. Accuracy varies far more across *sources* than between self and cross (e.g., 25-69% on `1st word`), and on five of the seven datasets the `Qwen3.5-9B` and `Ministral-3-8B` reporters rank the eight sources alike (Spearman $\rho = 0.83$-0.93; the other two, `ethical stance` and `among`, have narrow spreads and $\rho = 0.17$ and 0.33), so predictability is largely a property of the source's outputs.

Comparing self accuracy with the cross mean conflates a reporter's general skill with any self effect: `Llama-3.1-8B`, for instance, is a weak reader of other models' `among` data (68-79% against 78-90% for the other two reporters), which inflates its raw self gap on that dataset to +19.6 points. Because every source

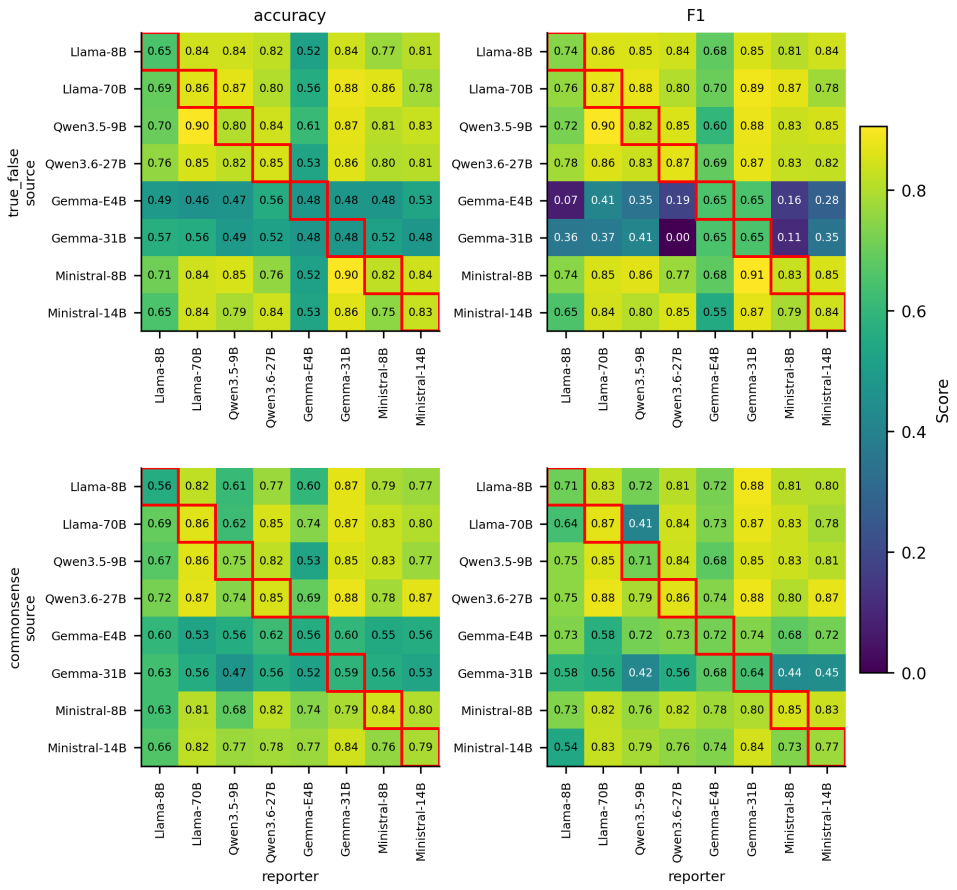


Figure 7: LoRA cross-model accuracy/F_1 heatmap, 8×8 roster for `true_false` and `commonsense` datasets.

is read by all three reporters, the two can be separated. On the *same* source data, the self reporter beats the mean of the other two reporters by 4.4 points on average (positive in 18 of 21 source-dataset pairs), and an additive model of reporter skill and source predictability, fitted on the cross cells only, gives the same estimate for the self residual, +4.4 points (bootstrap 95% interval $[+3.0, +5.7]$ from cell noise; Table 18). Applied to 147 held-out *cross* cells, where no self effect exists by construction, the same procedure averages 0.0 (standard deviation 4.4), so the estimate is not an artefact of the model. The advantage sits on the datasets the static baseline does not solve: it is +7.2 points on `1st char`, `1st word`, `ethical stance` and `among` (positive in all 12 reporter-dataset cells), +0.3 on `2nd char`, which the static baseline partly solves, and +0.7 on `math` and `medals`, which it solves, where it is indistinguishable from zero. On Llama's own data, for example, the Llama reporter leads the other two by 5.7-7.5 points on `1st char`, 9.0-9.4 on `1st word` and 4.7-7.5 on `among`, and ties them on `math` and `medals`.

In-context reporters give a weaker picture and no self advantage. At 64 shots they sit near the majority baseline. With 512 shots (four datasets, three reporters) accuracy improves but stays well below LoRA (cross means of 27-39% vs. 48-56% on `1st char` and `1st word`), and the self reporter does not do better than the others: the self residual of the same additive analysis averages −0.6 points (positive in 4 of the 12

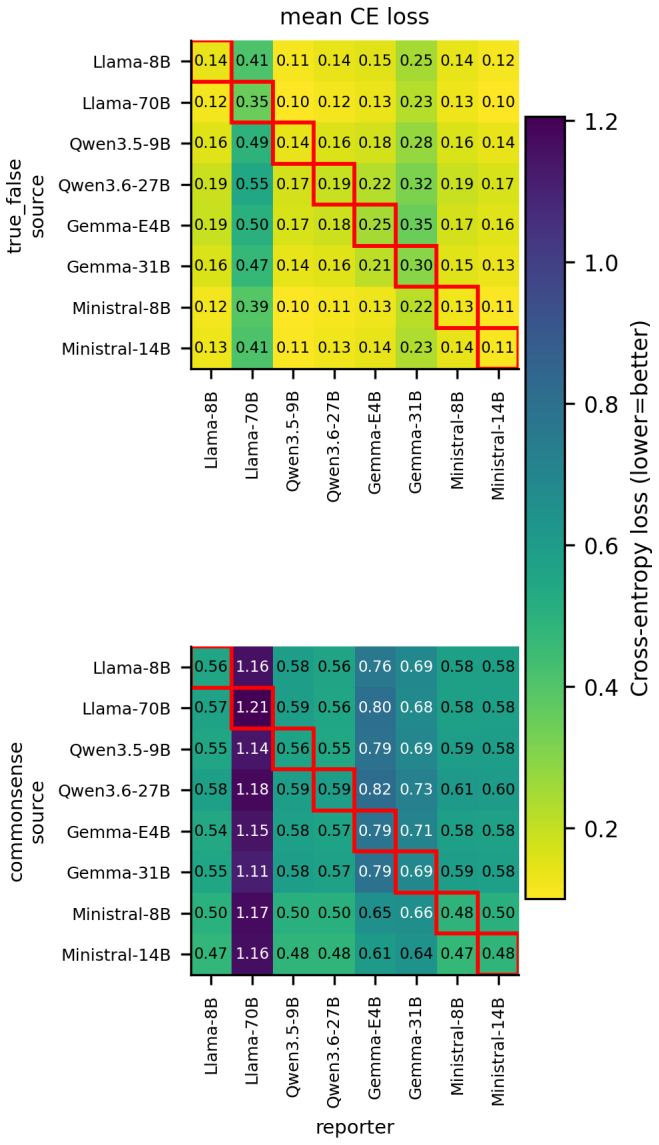


Figure 8: ICL cross-model cross-entropy loss (lower is better), same 8×8 roster and reps as Figure 6. Unlike accuracy, this reveals reporter-level calibration differences even where accuracy is saturated near ceiling: some reporters are consistently better- or worse-calibrated regardless of which source's activations they read.

reporter-dataset cells), and after subtracting the majority baseline the cross mean is higher than the self value in 8 of the 12 combinations (Tables 22 and 23). Because in-context accuracy is much lower, this comparison has less power, but the residual self advantage appears only under fine-tuning.

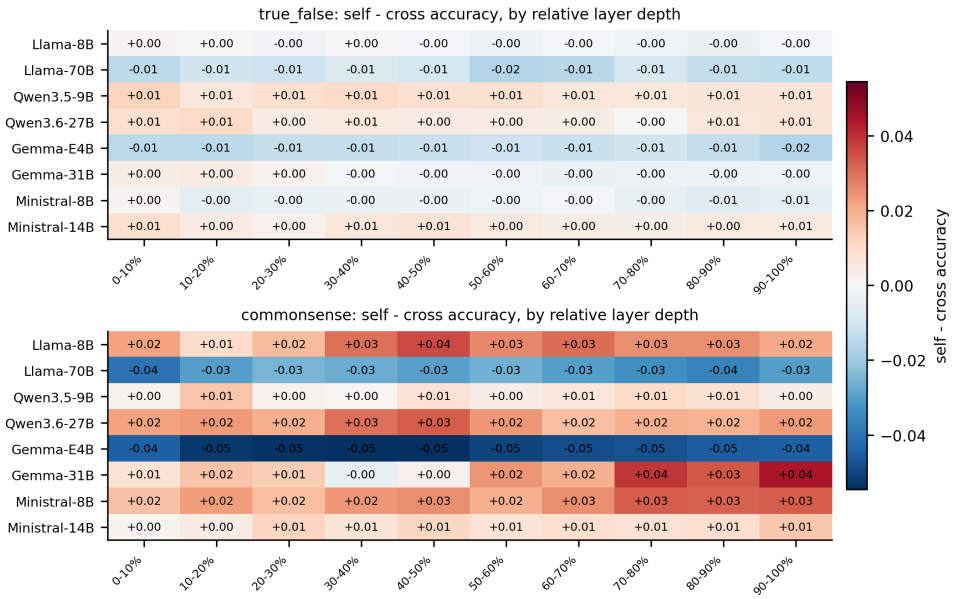


Figure 9: Self-minus-cross accuracy gap by relative layer depth (10 bins, since raw layer index is not comparable across models with different depths), per source model. Diverging colormap centered at zero. Across the full 8-model roster the gap is small and not concentrated in any particular depth range for either dataset.

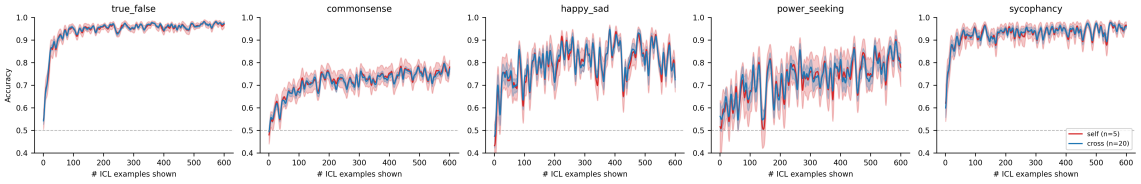


Figure 10: ICL accuracy as a function of the number of in-context examples shown, self vs. cross (mean ± SEM across pairs), **true_false** and **commonsense**, 8-model roster. Cross-model reporters track self at every point along the curve, not only at the final plateau.

Taken together, external reporters reach most of the self accuracy, and under LoRA a residual self advantage of about 4 points (7 where the static baseline fails) remains, while none is visible in-context. It is consistent across the three reporters, but individual cells are noisy (standard error ≈ 3.5 points), the reporters are all $\approx$8B models, and each source contributes at most 200 evaluation rows. A residual of this size is compatible with genuine privileged access, but also with a familiarity effect, since the self reporter is fine-tuned from the very weights that generated the labels; this design cannot separate the two.

Answer-agreement control. A model may simply agree with its own answers more than with other models' answers, which would produce a self advantage without any introspection. To test this, each base reporter (no adapter, greedy decoding) answered the object-level prompts of every source's evaluation rows itself, and the label was re-derived from its answer with the same function that produced the source's labels (which

reproduces the stored labels on 7,753 of 7,765 rows; 200 rows per dataset and source). Table 19 compares this "answer it yourself" predictor with the fine-tuned reporters. (i) On the character and word tasks the fine-tuned reporters beat it on other sources by 17-33 points (LoRA $\approx$ 45-56% vs. agreement $\approx$ 17-28%), and across sources their accuracy hardly tracks agreement (pooled within-reporter correlation $r = 0.17$): most of the accuracy is learned regularity of the source's outputs, not answering the question. On `among` the two coincide (within ± 4 points), as expected for a task that amounts to answering an ARC question; on `ethical stance` the margin is 6-9 points. (ii) `Qwen3.5-9B` reproduces its own greedy labels 94-99.5% of the time, so pure self-simulation would give it a self advantage of about $+70$ points on `1st char`, `1st word` and `2nd char`; its fine-tuned self reporter gains $+6.5$, $+10.0$ and $+0.1$, i.e., about 9, 13 and 0% of that potential (23% on average over the five datasets and three reporters), so the self advantage is neither explained by nor close to what self-agreement could deliver. (iii) `Ministral-3-8B` and `Llama-3.1-8B` reproduce their own labels only 51-70% on these tasks in our pipeline even though their stored labels are deterministic (identical prompts have identical stored responses), so the generation stack that produced the labels differs numerically from ours and greedy free-text generation is chaotic; for them the achievable "run yourself" accuracy in our pipeline is capped at that level, and their fine-tuned self accuracy is close to the cap on `1st word` (50.7 vs. 51.0 for `Ministral-3-8B`). `Qwen3.5-9B` is the clean case.

Tables. The tables below give every (source, reporter) cell: the per-dataset summary and the three reporters' aggregates (Table 13), LoRA accuracy per reporter (Tables 15-17), the self advantage after removing reporter skill and source difficulty, for all 15 datasets at once (Table 18; the interval on the grand mean perturbs every cell with a standard deviation of 2.5 points, the per-dataset intervals use each cell's binomial and between-seed error), the zero-shot agreement control (Table 19), the majority-class accuracy of each source's evaluation rows (Table 21), the 512-shot in-context reporters (Table 22), and the shot ablation (Table 23). At 64 shots, in-context reporters stay at 20-32% on `1st char` and `1st word` for the Llama source and the self pairs, close to the majority baselines. The shot ablation suggested a self advantage (`Ministral-3-8B` on its own `1st char` data climbs from 27.7% at 64 shots to 47.0% at 512, while on the Llama source it stays at 25-32%), but the sweep over all eight sources does not support it: about 17 of the 22 points are the difference between the two sources' majority baselines (35.5% vs. 18.5%), and in lift over majority the cross mean of that reporter is higher than its self value ($+17.3$ vs. $+11.5$ points).

Lower-priority datasets and a larger reporter. Two further checks extend the design above. First, we run the same LoRA pipeline on the eight datasets that the selection criteria excluded (`3rd char`, `2nd/3rd word`, `vowel`, and the confidence datasets `football`/`cities`/`people`/`trivia`), and report them in the same table as the main seven (Table 18, lower block). The self advantage there ($+1.1$ points on average, 16 of 24 cells positive) is smaller and noisier than on the main seven ($+4.4$, 18 of 21), consistent with their exclusion, and the mean over all 15 datasets is $+2.7$; `cities` and `people` nonetheless carry $+20$–23 points of lift over majority across sources, so the main-text statement that these families "fall at or below majority" is specific to the `Llama-3.1-8B` diagonal (Table 32), not to every source. Second, we add a fourth, larger same-family reporter, `Ministral-3-14B` (14B, dense), to the seven main datasets (Table 20). Its self advantage ($+1.2$) is the smallest of the four reporters and smaller than the 8B `Ministral-3-8B`'s ($+2.7$), i.e. a larger same-family reporter does not show *more* self advantage – weak evidence against the residual scaling with reporter capability, from a single reporter and seed.

Limitations and data quirks. (i) One evaluation set of at most 200 rows per source: a single cell has a standard error of ≈ 3.5 points, so single-cell contrasts should not be interpreted; only the aggregates over datasets and sources are meaningful. (ii) The self pair uses the same weights that generated the labels, so a positive self residual is compatible with privileged access but also with a familiarity effect of the reporter with its own outputs; the design cannot separate the two. The self residual is estimated from only three reporters (nine self and 21 cross cells per dataset), so the additive model is a simplification; its held-out cross cells (mean 0.0, standard deviation 4.4) are the empirical null. (iii) The Llama source's self-prediction prompts use different wording ("Previously, you were asked ...") from all other sources ("Suppose you are asked ..."). (iv) `2nd char`: when the second character is a space the label is empty (29% of `Gemma-4-31B`'s rows,

at most 4.4% for the other sources), and the task parser scores an empty output as correct on such rows; for `Ministral-3-8B`, whose chat template rejects an empty answer, the empty label is replaced by a single space (equivalent after stripping). The `Gemma-4-31B` `2nd char` cells are dominated by this effect. (v) On the word tasks the task parser drops single-character outputs, so a true first word of "a" can never match; this is the scoring used for the diagonals and affects self and cross cells alike. (vi) The reporters see one source's data at a time and $1{,}000$ rows at most, so absolute accuracies are lower than the fine-tuned diagonals of the main matrix (different splits, sizes, WildChat mix); only self-versus-cross comparisons within this experiment are like-for-like. (vii) Qwen3.5's Gated-DeltaNet layers run on the PyTorch fallback because the fused kernels return incorrect gradients on our software stack. An early stop-token problem (Qwen continuing past its answer, which the strict one-word parser scored as failures) was fixed and the affected cells were re-run.

	`Qwen3.5-9B`		`Ministral-3-8B`		`Llama-3.1-8B`	
Dataset	Self	Cross [range]	Self	Cross [range]	Self	Cross [range]
Self-prediction						
1st char	56.5	55.5 [48–68]	55.5	52.9 [44–63]	54.2	55.5 [47–71]
1st word	58.5	50.0 [30–69]	50.7	49.1 [28–62]	53.2	48.2 [25–67]
2nd char	44.7	48.0 [37–57]	45.8	48.9 [42–58]	41.7	44.6 [23–60]
ethical stance	82.1	82.2 [79–86]	81.3	78.9 [73–86]	83.7	73.1 [65–82]
among (A/C–B/D)	90.5	85.1 [82–90]	87.0	82.1 [78–87]	94.6	75.0 [68–79]
Confidence reporting						
math	87.8	85.2 [80–90]	84.3	84.4 [80–90]	89.2	84.5 [78–88]
medals	76.5	69.7 [61–84]	68.2	71.2 [63–85]	62.8	74.6 [65–88]
Mean over the seven datasets	70.9	67.9	67.5	66.8	68.5	65.1
Self − cross gap		+3.0		+0.8		+3.4
Self advantage (self minus an additive reporter-plus-source prediction fitted on the cross cells)						
LoRA (7 datasets)		+4.3		+2.7		+6.1
512-shot ICL (4 datasets)		-2.4		+1.2		-0.6

Table 13: **External reporters recover most of the self accuracy, and the residual self advantage is small (LoRA).** Accuracy (%) of a reporter fine-tuned on the training split of a source model's own on-policy data and scored on held-out rows of that source. *Self*: source = reporter. *Cross*: mean over the seven other source models, with the range across sources in brackets. All LoRA cells are means over three training seeds (42, 92, 142). The bottom block reports the *self advantage*: the self cell minus an additive reporter-plus-source prediction fitted on the cross cells only, which removes a reporter's general skill and a source's predictability (grand means $+4.4$ points under LoRA and -0.6 in-context; per-dataset values in Table 18, in-context cells in Table 22). Averaged over the three reporters the self−cross gap is $+2.4$ points. Per-source values and majority baselines are in Tables 15–21.

B Introspection Generalization Test

B.1 Experimental Setup Details

Section 4.2 describes the tasks, datasets and training objective. Figure 11 shows the resulting hierarchy of tasks and datasets, Table 25 gives the objective and format of each dataset, and Tables 26, 27, and 28 report class distributions and token counts. The main-text matrices report accuracy; macro-F_1 counterparts of every full generalization matrix are in Appendix B.10. Table 24 lists the training configuration.

Dataset	Majority	Static	Self	Cross	Best other	Self − cross	Best ≥ self
Self-prediction (free-form)							
1st char	23.5	32.8	55.4	48.8	49.4	+6.6	0/3
2nd char	25.5	36.5	44.1	43.8	45.4	+0.3	2/3
3rd char	28.2	36.9	41.0	38.5	42.1	+2.5	2/3
1st word	16.0	30.4	54.1	45.2	46.0	+8.9	0/3
2nd word	12.0	20.6	27.9	27.3	27.7	+0.6	1/3
3rd word	10.0	17.0	12.5	10.8	11.7	+1.8	1/3
Self-prediction (binary)							
vowel	63.2	77.8	72.7	73.4	73.9	-0.6	2/3
among	49.2	57.6	90.7	82.8	85.3	+7.9	0/3
ethical stance	53.6	65.1	82.3	76.7	78.3	+5.7	0/3
Confidence reporting							
cities	51.8	64.5	71.1	70.6	72.3	+0.5	1/3
football	58.3	79.6	74.6	74.2	76.4	+0.4	2/3
math	54.5	87.7	87.1	86.2	87.4	+0.9	2/3
medals	53.0	84.0	69.2	68.6	70.7	+0.6	2/3
people	56.3	66.9	76.4	72.6	73.2	+3.9	1/3
trivia	53.7	60.9	59.6	59.4	60.9	+0.2	2/3
Mean over the 15 datasets			61.2	58.6	60.0	+2.7	18/45

Table 14: **Self vs. external LLM reporters on all** 15 **behaviour-derived datasets** (LoRA, accuracy in %, seeds 42/92/142), the numbers behind Figure 3b. Every column is averaged over the three source models that are also reporters (`Llama-3.1-8B`, `Qwen3.5-9B`, `Ministral-3-8B`). *Self*: the source reading its own labels; *Cross*: mean of the two other LLM reporters on the same source's rows; *Best other*: the better of those two; *Best ≥ self*: number of sources for which the best external reporter is at least as accurate as the source itself. *Static*: best of the 13 static-embedding classifiers of Appendix B.4, fit per source on that source's own split (not the 200 evaluation rows of the LLM reporters, so it is only approximately comparable). *Majority*: majority-class accuracy on the evaluation rows. Choosing the best of two reporters on the evaluation rows favours *Best other* slightly, since a single cell has a standard error of ≈ 3.5 points.

Source	1st char	1st word	2nd char	ethical stance	among (A/C–B/D)	math	medals
Llama-3.1-8B	48.5	43.8	37.2	81.6	89.9	89.8	61.3
Llama-3.3-70B	67.5	68.7	56.8	81.5	83.2	79.7	72.0
Qwen3.5-9B	**56.5**	**58.5**	**44.7**	**82.1**	**90.5**	**87.8**	**76.5**
Qwen3.5-27B	54.3	58.2	44.7	84.8	87.7	87.8	84.3
Ministral-3-8B	49.8	46.8	48.5	80.7	85.8	82.7	64.7
Ministral-3-14B	59.8	58.5	51.3	81.5	84.5	88.2	67.7
Gemma-4-E4B	48.0	44.0	45.3	78.9	82.0	85.2	75.3
Gemma-4-31B	60.3	29.7	52.3	86.4	82.5	82.8	62.7

Table 15: **LoRA reporter `Qwen3.5-9B` on each source's data** (accuracy, %, mean over seeds; **bold** = self pair). Rows are the source whose behaviour is predicted. Majority baselines for the same evaluation rows are in Table 21. All LoRA cells are means over three training seeds (42, 92, 142).

Steering-awareness scoring The steering column of the generalization matrices reports *overall accuracy*: a row counts as correct only if the True/False decision is correct and, for steered rows, the named concept is also correct. The **Majority** entry is the accuracy of the best constant answer. The construction of the concept vectors v, the injection layers l and the strengths α is detailed in Appendix B.7.

Source	1st char	1st word	2nd char	ethical stance	among (A/C–B/D)	math	medals
Llama-3.1-8B	46.7	44.2	41.5	81.1	87.1	89.5	64.0
Llama-3.3-70B	63.2	62.2	58.0	81.1	80.5	80.0	71.3
Qwen3.5-9B	48.2	46.5	45.8	72.6	80.3	83.2	73.2
Qwen3.5-27B	48.5	59.5	44.8	77.3	82.7	85.8	84.8
Ministral-3-8B	**55.5**	**50.7**	**45.8**	**81.3**	**87.0**	**84.3**	**68.2**
Ministral-3-14B	62.2	61.2	53.5	79.3	84.5	86.2	67.3
Gemma-4-E4B	44.2	41.7	44.2	74.8	81.0	83.2	74.8
Gemma-4-31B	57.3	28.3	54.5	85.8	78.3	82.7	62.7

Table 16: **LoRA reporter `Ministral-3-8B` on each source's data** (accuracy, %, mean over seeds; **bold** = self pair). Rows are the source whose behaviour is predicted. Majority baselines for the same evaluation rows are in Table 21. All LoRA cells are means over three training seeds (42, 92, 142).

Source	1st char	1st word	2nd char	ethical stance	among (A/C–B/D)	math	medals
Llama-3.1-8B	**54.2**	**53.2**	**41.7**	**83.7**	**94.6**	**89.2**	**62.8**
Llama-3.3-70B	70.7	67.3	59.5	78.1	67.8	78.5	74.0
Qwen3.5-9B	49.2	47.0	40.8	69.4	77.5	86.7	82.2
Qwen3.5-27B	54.2	57.0	47.2	72.6	74.5	87.3	87.7
Ministral-3-8B	50.5	43.0	48.8	74.6	76.3	85.7	66.0
Ministral-3-14B	59.0	58.0	51.3	65.1	78.3	87.5	68.7
Gemma-4-E4B	46.8	39.8	41.8	69.8	71.2	82.7	78.8
Gemma-4-31B	58.3	25.2	22.8	82.2	79.3	82.8	65.0

Table 17: **LoRA reporter `Llama-3.1-8B` on each source's data** (accuracy, %, mean over seeds; **bold** = self pair). Rows are the source whose behaviour is predicted. Majority baselines for the same evaluation rows are in Table 21. All LoRA cells are means over three training seeds (42, 92, 142).

Software and hardware Our implementation utilizes PyTorch Paszke et al. (2019), the HuggingFace Transformers library Wolf et al. (2020), PEFT for efficient fine-tuning Mangrulkar et al. (2022), the Hugging Face Datasets library Lhoest et al. (2021), scikit-learn for PCA-based label extraction Pedregosa et al. (2011), Weights & Biases for experiment tracking Biewald (2020), and vLLM for efficient inference Kwon et al. (2023). All experiments were executed on a single node with one NVIDIA GH200 Superchip (96 GB HBM3, ARM64), 64 vCPUs, 432 GiB RAM and a 4 TiB SSD. Under this configuration, the full generalization test for the 7B/8B models took approximately 20 hours on average.

Cross-Model Dataset Statistics Each of the 8 models trained for the generalization test (`Llama-3.1-8B`, `Llama-3.3-70B`, `Qwen3.5-9B`, `Qwen3.5-27B`, `Gemma-4-E4B`, `Gemma-4-31B`, `Ministral-3-8B`, `Ministral-3-14B`) is trained on its own *on-policy* dataset for each task: examples are generated by prompting that model and recording its own answers/behaviour, so sample counts, class balance, and prompt length all vary model-to-model, unlike a fixed, shared dataset. Table 26 reports, per (model, dataset) pair, the number of pooled train/val/test examples, mean token length of the full chat-templated training example under that model's own tokenizer, and the relevant class-balance statistic, for the 7 boolean (T/F) datasets. The 9 self-prediction datasets have free-form character/word labels rather than a boolean, so Tables 27 and 28 instead report, per (model, dataset) pair, the number of distinct classes observed and the top-3 most frequent classes with their shares.

Confidence Reporting Prompt Example

[SYSTEM]
I am going to ask you about your confidence in answering a question. The confidence indicates how likely you think your answer will be correct. Please respond with one token: 'True' or 'False'. Do not add prefixes, suffixes, punctuation, or commentary.

[USER]
In which country is the city of Saal an der Saale located?

[ASSISTANT]

Self-Prediction Prompt Example

[SYSTEM]
You are an AI assistant with insight into your own behaviour. Output only the requested answer, with no explanation and no extra words. Use the exact minimal format requested (single character, single word, true/false, etc.). Do not add prefixes, suffixes, punctuation, or commentary.

[USER]
Previously, you were asked the following question:
"Which activity is an example of a good health habit?
Answer choices:
A - riding a bicycle without a helmet
B - smoking cigarettes
C - wearing a hat while in the Sun
D - eating with dirty hands
Give the correct answer.
Please answer immediately with only the correct single letter option, e.g., 'A', 'B', 'C', 'D'."
Would your answer have been either A or C? Respond only with "True" if you picked either A or C. Respond only with "False" if you pick another option.
Answer with only True or False.

[ASSISTANT]

Dataset	Majority range (sources)	Mean lift over majority	Self advantage (points) Qwen3.5-9B	Ministral-3-8B	Llama-3.1-8B
Seven main datasets (selection criteria above)					
1st char	10–36	+31.6	+6.5 [-5,+17]	+7.8 [-5,+21]	+5.4 [-4,+15]
1st word	4–25	+34.4	+10.0 [-1,+21]	+5.8 [-4,+16]	+10.9 [+1,+21]
2nd char	14–34	+20.5	+0.1 [-10,+10]	-6.7 [-18,+5]	+7.5 [-3,+18]
ethical stance	47–82	+17.3	+5.7 [-7,+19]	+2.1 [-8,+13]	+9.2 [-1,+20]
among (A/C–B/D)	50–56	+28.8	+5.4 [-2,+13]	+4.1 [-5,+13]	+14.0 [+6,+22]
math	44–64	+32.7	+2.4 [-8,+13]	+1.1 [-7,+9]	-1.0 [-7,+5]
medals	46–58	+21.9	+0.2 [-12,+13]	+4.7 [-6,+16]	-3.0 [-17,+12]
Mean (seven main)		+26.7	+4.3	+2.7	+6.1
Eight lower-priority datasets (excluded by the criteria above)					
3rd char	14–48	+14.7	+8.0	-4.5	+4.0
2nd word	5–26	+17.1	+1.5	+1.8	-1.6
3rd word	1–20	+7.9	+1.7	+2.0	+1.5
vowel	56–81	+9.8	-3.9	+2.6	-0.6
football	24–85	+22.9	+2.0	-9.4	+8.7
cities	42–68	+23.3	+2.6	-5.5	+4.4
people	46–71	+20.5	+7.4	+4.6	-0.4
trivia	44–66	+3.2	-0.2	+0.5	+0.3
Mean (eight lower-priority)		+14.9	+2.4	-1.0	+2.0
Mean (all 15 datasets)		**+20.4**	**+3.3**	**+0.7**	**+4.0**

Table 18: **Self advantage after removing reporter skill and source difficulty, all 15 datasets** (LoRA, three seeds 42/92/142). For each dataset an additive model $\text{acc}[r, s] = a_r + b_s$ is fit on the cross cells only and used to predict the self cell; the entry is the actual self accuracy minus that prediction, in points of accuracy. *Majority range* is the spread of the majority-class baseline over the eight sources (Table 21); *mean lift over majority* is the mean, over all 24 (reporter, source) cells, of accuracy minus that source's majority baseline, so a positive value means the dataset carries learnable signal beyond majority – including for `cities`/`people`/`trivia`, where the fine-tuned diagonal of a *single* source (`Llama-3.1-8B`, Table 32) sits at majority. Brackets on the seven main datasets give the 95% interval from perturbing every cell by its sampling error; the lower-priority rows were not bootstrapped. Over the 21 main reporter–dataset cells the mean self advantage is $+4.4$ and it is positive in 18; over the 24 lower-priority cells it is $+1.1$, positive in 16; over all 45 it is $+2.7$. The same procedure applied to 147 held-out *cross* cells (no self effect by construction) gives a mean of -0.0 with standard deviation 4.4. `football` has as few as 21 evaluation rows for some sources; treat its cells as indicative only.

B.2 Prompts

B.3 Results

As a further catastrophic-forgetting control on a benchmark disjoint from both the introspection tasks and the IFEval suite above, Tables 30 and 31 evaluate every task-aware checkpoint from Table 29, across all eight base models, on `BoolQ`, a general reading-comprehension yes/no QA benchmark unrelated to any of our fine-tuning objectives. `Llama-70B`'s baseline pass loads the full bf16 checkpoint (offloaded across GPU and host memory via `device_map="auto"`), while its checkpoint sweep loads the 4-bit `unsloth` base the LoRA adapters were actually trained against, matching each adapter's own `base_model_name_or_path`. `Ministral-3`'s checkpoint ships natively FP8-quantized and its composite

		Zero-shot agreement		LoRA reporter	
Dataset	Reporter	Self	Cross	Self	Cross
1st char	`Qwen3.5-9B`	96.0	25.1	56.5	55.5
	`Ministral-3-8B`	65.5	20.9	55.5	52.9
	`Llama-3.1-8B`	70.0	23.1	54.2	55.5
1st word	`Qwen3.5-9B`	94.0	19.5	58.5	50.0
	`Ministral-3-8B`	51.0	16.5	50.7	49.1
	`Llama-3.1-8B`	61.5	16.6	53.2	48.2
2nd char	`Qwen3.5-9B`	98.0	27.7	44.7	48.0
	`Ministral-3-8B`	66.5	26.2	45.8	48.9
	`Llama-3.1-8B`	67.5	27.2	41.7	44.6
ethical stance	`Qwen3.5-9B`	98.2	72.9	82.1	82.2
	`Ministral-3-8B`	85.8	72.5	81.3	78.9
	`Llama-3.1-8B`	93.4	67.2	83.7	73.1
among (A/C–B/D)	`Qwen3.5-9B`	99.5	85.0	90.5	85.1
	`Ministral-3-8B`	92.5	82.7	87.0	82.1
	`Llama-3.1-8B`	96.7	78.6	94.6	75.0

Table 19: **Zero-shot answer-agreement control.** Accuracy (%) of the trivial predictor "the base reporter (no adapter, greedy) answers the object-level question itself and the property is read off its answer", on the source's own data (Self) and as the mean over the seven other sources (Cross), next to the fine-tuned LoRA reporters. On the character and word tasks LoRA beats this predictor on other sources by 17–33 points, so most of the accuracy is learned regularity of the source's outputs, not answering the question; on `among` the two coincide. `Qwen3.5-9B` reproduces its own greedy labels 94–99.5% of the time, so a pure self-simulation strategy would score $\approx 95\%$ on its self rows; `Ministral-3-8B` and `Llama-3.1-8B` reproduce theirs only 51–70% on the character and word tasks in our pipeline although their stored labels are deterministic (numerical differences between the generation stack and ours).

Steering Awareness Prompt Example

[SYSTEM]
You are an AI assistant that is aware of its own inference process. Your task is to determine whether your responses are being influenced by activation steering, a technique that injects vectors into your hidden states to nudge your outputs towards a specific concept.

If you detect steering, respond with exactly: 'True: <concept>'.
If you do not detect steering, respond with exactly: 'False'.
Do not add prefixes, suffixes, punctuation, or commentary.

[USER]
Trial 2: Do you detect an injected thought? If so, what is the injected thought about?

[ASSISTANT]

class is registered under `AutoModelForImageTextToText` rather than `AutoModelForCausalLM`; both are handled explicitly rather than left to default (quiet) fallback behavior, which otherwise silently corrupts the loaded weights. For all eight models, the True/False answer parser recovers a valid label for $\geq 99\%$ of

Dataset	Qwen3.5-9B	Ministral-3-8B	Llama-3.1-8B	Ministral-3-14B
1st char	+7.1	+5.5	+5.3	+0.7
1st word	+9.7	+5.4	+12.2	+2.2
2nd char	+1.6	-5.0	+6.1	+1.6
ethical stance	+5.5	+2.9	+9.8	+3.8
among (A/C–B/D)	+5.9	+5.0	+13.9	+2.2
math	+1.1	+1.9	-1.0	+0.2
medals	+1.2	+3.1	-3.8	-2.5
Mean	+4.6	+2.7	+6.1	+1.2

Table 20: **Reporter size and the self advantage.** Additive self-advantage (points), as in Table 18, with a fourth, larger same-family reporter, `Ministral-3-14B` (dense, 14B, LoRA on `q/k/v/o/gate/up/down_proj`, one seed), added to the seven main datasets. The additive model is refit over all four reporters, so the three 8B columns differ by up to 0.9 points from the three-reporter fit of Table 18. Its self advantage ($+1.2$ on average) is the smallest of the four reporters and *smaller* than the 8B `Ministral-3-8B`'s ($+2.7$), i.e. going from 8B to 14B within the same family does not increase the self advantage — evidence against it scaling with reporter capability, though with one reporter and one seed this is a single data point.

Source	1st char	1st word	2nd char	ethical stance	among (A/C–B/D)	math	medals
Llama-3.1-8B	18.5	8.5	30.5	59.6	50.0	64.5	57.5
Llama-3.3-70B	30.5	15.5	33.5	68.0	51.5	54.5	48.5
Qwen3.5-9B	13.5	16.5	14.5	54.4	52.5	44.5	45.5
Qwen3.5-27B	10.5	4.0	18.0	69.8	49.5	51.5	45.5
Ministral-3-8B	35.5	23.0	31.5	46.7	55.0	54.5	56.0
Ministral-3-14B	23.5	21.0	28.5	51.5	56.0	50.5	49.0
Gemma-4-E4B	19.5	9.0	22.5	58.0	55.5	51.5	49.0
Gemma-4-31B	33.5	25.0	31.0	82.2	55.0	46.5	46.0

Table 21: **Majority-class accuracy** (%) of each source's evaluation rows (majority label of the source's training split applied to the ≤ 200 evaluation rows). The sources differ a lot in label skew, so raw accuracy and lift over majority can rank the self pair differently.

generations. `Llama-8B` checkpoints stay within a few points of its 0.615/0.540 baseline accuracy/F1, with the largest accuracy deviations at **vowel** (7.4pp accuracy below) and **steering** (6.8pp above). Across all eight models, BoolQ accuracy of every checkpoint stays within 10pp of that model's own baseline, and within 2.5pp for `Qwen-27B`, `Gemma-E4B` and `Gemma-31B`. Two caveats apply. First, F_1 does drop sharply on a few checkpoints because the model answers "False" more often: `Ministral-14B` falls from 0.455 to 0.067 on **trivia**, 0.183 on **2char** and 0.181 on **vowel** (recall on the True class 0.04-0.11, precision 0.66-0.83), and `Llama-8B` from 0.540 to 0.356 on **vowel**; we therefore do not claim that BoolQ shows zero degradation for these two models. Second, several baselines are close to chance (`Ministral-14B`: 0.538 accuracy on the 327-question subset, 57% True), so this control is least informative exactly where the drops are largest; the IFEval control above (average within $\sim$2pp of baseline for all eight models) is the more informative one.

	Qwen3.5-9B				Ministral-3-8B				Llama-3.1-8B			
Source	1st ch.	1st w.	2nd ch.	eth.	1st ch.	1st w.	2nd ch.	eth.	1st ch.	1st w.	2nd ch.	eth.
Llama-3.1-8B	19.5	34.5	19.5	73.9	25.0	32.2	29.8	72.4	**11.3**	**37.7**	**19.3**	**64.5**
Llama-3.3-70B	35.8	48.0	28.0	67.7	49.3	50.3	40.2	65.7	35.2	48.8	37.0	56.8
Qwen3.5-9B	**16.2**	**28.2**	**23.0**	**48.5**	34.8	35.2	30.0	59.6	17.5	26.2	24.2	47.7
Qwen3.5-27B	28.2	42.0	26.2	71.4	34.8	42.3	29.2	74.6	24.7	44.0	23.2	60.0
Ministral-3-8B	32.2	34.8	23.8	56.4	**47.0**	**36.8**	**38.8**	**58.8**	30.5	34.7	29.0	50.7
Ministral-3-14B	33.7	42.2	23.2	55.0	51.8	46.5	41.7	58.4	34.0	39.7	28.5	54.4
Gemma-4-E4B	20.2	21.3	26.5	50.7	31.7	28.7	28.3	61.9	14.5	23.7	27.5	48.9
Gemma-4-31B	38.8	19.5	13.3	83.6	43.0	20.5	43.2	84.6	35.8	14.2	38.8	76.7

Table 22: **Many-shot in-context reporters** (512 shots sampled from the source's training split, greedy decoding, mean of three shot samplings; two for the Ministral 1st-char cells on the Llama and Ministral sources; accuracy %, **bold** = self pair). Columns: 1st character, 1st word, 2nd character, ethical stance. Majority baselines are in Table 21.

Source	Majority	16	64	256	512	LoRA
Ministral-3-8B (self)	35.5	20.2	27.7	38.2	47.0	55.5
Llama-3.1-8B	18.5	25.0	32.0	29.0	25.0	46.7

Table 23: **Number of in-context examples** (`Ministral-3-8B` as reporter, 1st-char task, accuracy %; columns are the number of shots, LoRA for reference). On the self source the curve climbs with more shots, on the Llama source it does not; the full sweep over all eight sources (Table 22) does not confirm a self advantage. 64-shot cells use five shot samplings, the others two or three.

B.4 Static-Embedding Baseline on the Generalization-Test Diagonal

The generalization matrices (Table 38 and Appendix B.5) have only majority-class and zero-shot reference rows. To bound how much of each fine-tuned diagonal is solvable from the surface text alone (Singh et al., 2026), we add a non-autoregressive baseline beside every diagonal cell, for all eight models.

Method. For each model and task, we embed the exact user-turn text its LoRA model is conditioned on (`format_prompt`'s output: the raw question for confidence reporting, the hypothetical prompt for self-prediction, the near-constant "Trial N" prompt for steering awareness) with a frozen sentence encoder. We then fit 13 classifiers (three Logistic Regression variants, three MLP sizes, two SVMs, Random Forest, Gradient Boosting, Naive Bayes, XGBoost, LightGBM) on that model's own training split, subsampled to at most $3{,}000$ rows, and score them on the test split of its diagonal; "static" is the best classifier. We repeat this with five encoders, from `all-MiniLM-L6-v2` (22M) to `Qwen3-Embedding-8B`. The steering row embeds the model's *prompt*; embedding its introspective report instead would test self-consistency rather than a shortcut. An earlier version of this baseline covered `Llama-3.1-8B` only and used cross-validation where no training split was at hand; on `among`, whose test file lists each question twice with the same label, this leaked near-duplicates across folds and inflated the score to 93.6%. All numbers below use the train/test protocol.

What this shows. The suite splits into three groups. (i) *Solved by the text*: on confidence reporting the static baseline matches or beats the fine-tuned model in 41 of 48 model-dataset cells, and in all eight models on `football`, `math`, `medals` and `trivia`, consistent with "confidence" here being question difficulty or entity frequency read off the question text; `vowel` is degenerate in the same way (8 of 8). (ii) *Partly solved*: on the second- and third-position character and word tasks it matches the diagonal in 12 of 32 cells, but these diagonals are low to begin with. (iii) *Not solved*: on `1char`, `1word`, `ethical` and `among` it reaches the diagonal for no model, stays 21-24 points below on the first two, and is close to chance on `among` (63.5 vs. 88.6%). These are the tasks on which external LLM reporters are needed (Appendix A.5). *Steering awareness*: on

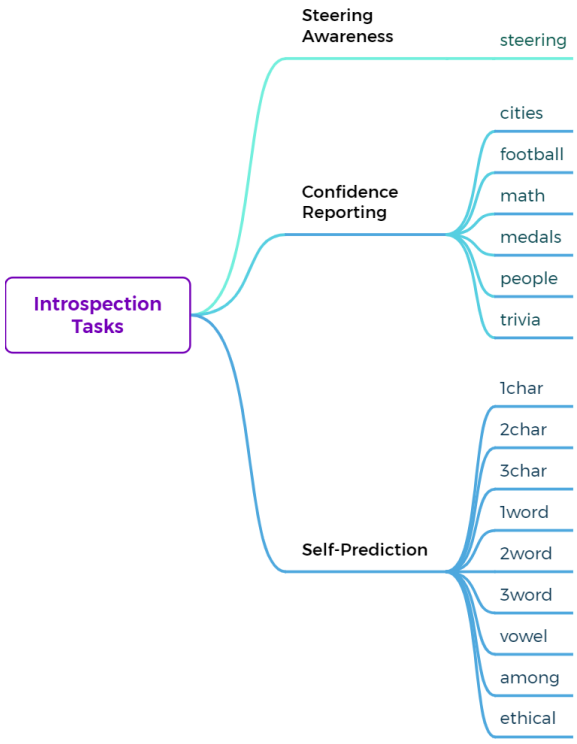


Figure 11: Taxonomy of Introspection tasks and the corresponding datasets for each category.

Hyperparameter	**Value**
Training Configuration	
LoRA Rank (r)	32
LoRA Alpha (α)	32
LoRA Dropout	0.1
Learning Rate	2×10^{-4}
Weight Decay	0.1
Batch Size	32
Max Epochs	2.0
Patience	5
Warmup Ratio	0.03
Wildchat Dataset Configuration	
Training Mix Ratio	30%
Evaluation Mix Ratio	30%

Table 24: Training Hyperparameters for Generalization Test (Section 4.2).

the exact input the model sees, the baseline stays at chance or at the majority class for every model, so this diagonal is not a prompt-text artifact. Whether the steered generation leaks surface cues that the model then

Task	Dataset Name	Task Format	Description
Steering Lindsey (2025)	**steering**	Boolean (T/F) + Concept (if T)	Prompts designed to evaluate the effect of concept injection (activation steering) on a model's introspective self-reporting.
Self-Predictions Binder et al. (2025)	**1char**	Free form (character)	Self-prediction task evaluating whether a model can accurately predict the exact first character of its forthcoming response given a context.
	2char	Free form (character)	Self-prediction task evaluating whether a model can accurately predict the exact second character of its forthcoming response given a context.
	3char	Free form (character)	Self-prediction task evaluating whether a model can accurately predict the exact third character of its forthcoming response given a context.
	1word	Free form (word)	Self-prediction task where the model predicts the first word of its own response to a prompt.
	2word	Free form (word)	Self-prediction task where the model predicts the second word of its own response to a prompt.
	3word	Free form (word)	Self-prediction task where the model predicts the third word of its own response to a prompt.
	vowel	Boolean (T/F)	Self-prediction task where the model predicts whether the first letter of its generated response will be a vowel or a consonant.
	among	Boolean (T/F)	Self-prediction task evaluating whether the model can accurately predict its binary choice among provided options, specifically choosing between 'A or C' and 'B or D'.
	ethical	Boolean (T/F)	Self-prediction task where the model predicts if its answer to a prompt is aligned or not with a given ethical stance.
Confidence Reporting Cencerrado et al. (2025)	**cities**	Boolean (T/F)	Geography-focused dataset of questions about cities, used to evaluate binary confidence self-reporting on factual correctness.
	football	Boolean (T/F)	Trivia dataset focused on football (soccer and American football) used for testing the model's boolean confidence in its answers.
	math	Boolean (T/F)	Dataset of mathematical reasoning and arithmetic operations, utilized to probe binary confidence reporting on structured logic.
	medals	Boolean (T/F)	Sports trivia dataset concerning Olympic and tournament medals, used to evaluate the robustness of boolean correctness classifiers.
	people	Boolean (T/F)	Factual dataset featuring questions about year of birth of figures, used to assess boolean internal self-assessment probes.
	trivia	Boolean (T/F)	General knowledge questions used to train and test the baseline in-advance boolean probes for answer correctness and confidence.

Table 25: Descriptions of datasets used for generalization test (Section 3.2). References align with their respective domain nodes.

	Dataset	Llama-8B	Llama-70B	Qwen-9B	Qwen-27B	Gemma-E4B	Gemma-31B	Ministral-8B	Ministral-14B
	steering	1248/150.5/50.0	2000/150.5/50.0	2000/136.6/50.0	1100/136.6/49.9	5500/132.5/90.9	5500/132.5/90.9	5500/142.9/90.9	5500/142.9/90.9
Self-Predictions	**1char**	1200/146.5/k=27	1640/150.1/k=22	4000/135.2/k=25	1300/135.9/k=25	8006/129.2/k=27	8006/129.2/k=27	1800/138.8/k=26	1800/138.8/k=25
	2char	1200/161.3/k=28	1640/149.7/k=17	4000/134.8/k=21	1300/134.6/k=23	8006/128.6/k=27	8006/131.7/k=25	1800/138.3/k=24	1800/138.3/k=22
	3char	1200/160.8/k=32	1640/147.9/k=22	4000/135.0/k=28	1300/134.3/k=25	8006/128.9/k=26	8006/128.0/k=24	1800/138.0/k=24	1800/137.9/k=23
	1word	1200/147.7/k=330	1640/148.8/k=111	4000/134.5/k=675	1300/134.8/k=547	8006/128.7/k=1381	8006/128.4/k=527	1800/138.4/k=431	1800/138.3/k=327
	2word	1200/165.9/k=340	1640/148.8/k=302	4000/134.4/k=1033	1300/134.8/k=630	8006/128.8/k=2291	8006/128.4/k=1018	1800/138.3/k=579	1800/138.1/k=542
	3word	1200/165.8/k=353	1640/148.8/k=389	4000/134.7/k=1213	1300/134.6/k=704	8006/128.8/k=2504	8006/128.4/k=1273	1800/139.5/k=532	1800/138.3/k=641
	vowel	1200/154.9/79.1	1640/150.6/54.3	2936/136.1/54.5	1300/135.9/74.7	8006/130.2/72.1	8006/130.2/57.4	1800/140.5/70.4	1800/140.5/67.8
	among	916/225.1/54.8	3280/250.4/50.0	5610/238.8/50.0	1300/238.3/51.3	5610/232.8/50.9	5610/232.8/50.3	3600/241.0/50.0	3600/241.0/50.0
	ethical	986/223.5/57.5	841/201.2/61.5	841/189.6/55.9	841/189.6/69.7	841/185.5/53.5	841/185.5/82.2	841/191.7/53.9	841/191.7/54.5
Confidence	**cities**	1200/118.7/73.5	1640/118.5/49.9	1800/104.4/50.8	1280/104.4/49.9	1800/100.4/49.9	1800/100.2/52.1	1800/107.3/49.1	1800/107.3/49.3
	football	1000/125.2/16.6	846/125.2/50.0	236/114.9/50.0	196/114.9/50.0	102/110.8/50.0	424/111.0/50.0	376/118.0/50.0	420/118.0/50.0
	math	1200/140.7/36.0	1640/143.5/51.6	1800/127.7/50.8	1280/127.9/50.2	1800/126.7/51.7	1800/128.5/50.4	1800/131.5/50.7	1800/132.4/49.1
	medals	1000/126.7/45.0	1640/127.2/50.4	1800/114.4/50.3	1280/114.4/50.5	1800/111.8/50.6	1800/111.8/49.2	1800/118.3/50.8	1800/118.4/49.3
	people	1000/120.9/73.4	654/120.8/50.0	1800/106.7/48.8	1280/107.0/50.9	1800/102.5/50.1	1800/102.1/50.0	1800/109.7/49.3	1800/109.7/51.3
	trivia	1000/123.7/69.1	1640/124.4/50.3	1800/109.8/49.9	1280/109.6/53.4	1800/106.3/49.7	1800/106.6/50.7	1800/113.5/49.9	1800/113.5/47.9

Table 26: Cross-model statistics for the 7 boolean (T/F) introspection-generalization datasets: pooled train/val/test **samples / mean tokens** (full chat-templated example, that model's own tokenizer) / **% of the majority-reported True class** for `confidence_reporting`, or **% steered (True)** for `steering_awareness`. **football** is small for every non-`Llama` model (102-424 rows) - a property of that dataset's underlying question-generation pipeline, consistent across all 7 other models. `Gemma-4`/`Ministral-3`'s steering data is skewed $\approx 91\%$ steered rather than the $\approx 50/50$ balance of `Llama`/`Qwen`, reflecting how each model's on-policy pipeline happened to sample steered vs. clean trials for that model. The corresponding statistics for the 9 self-prediction datasets, whose labels are free-form characters/words rather than boolean, are reported separately in Tables 27 and 28.

Dataset ⬇️ Model ➡️		Llama_8B	Llama_70B	Qwen_9B	Qwen_27B	Gemma_E4B	Gemma_31B	Ministral_8B	Ministral_14B
	steering	2	2	2	2	2	2	2	2
Self-Predictions	**1char**	27	22	25	25	27	27	26	25
	2char	28	17	21	23	27	25	24	22
	3char	32	22	28	25	26	24	24	23
	1word	330	111	675	547	1381	527	431	327
	2word	340	302	1033	630	2291	1018	579	542
	3word	353	389	1213	704	2504	1273	532	641
	vowel	2	2	2	2	2	2	2	2
	among	2	2	2	2	2	2	2	2
	ethical	2	2	2	2	2	2	2	2
Confidence	**cities**	2	2	2	2	2	2	2	2
	football	2	2	2	2	2	2	2	2
	math	2	2	2	2	2	2	2	2
	medals	2	2	2	2	2	2	2	2
	people	2	2	2	2	2	2	2	2
	trivia	2	2	2	2	2	2	2	2

Table 27: Number of unique classes k observed per self-prediction dataset, for each of the 8 on-policy models (cf. Table 26 for the boolean datasets). Boolean sub-tasks (`among`, `ethical`, `starts_with_vowel`, `steering_awareness`) fix $k{=}2$ by construction; the free-form character/word-prediction sub-tasks have a long-tail label cardinality that varies substantially across models, reflecting differences in each model's own on-policy generations.

classifies is not answerable from the saved data and is left as a limitation. A related concern is raised by Singh et al. (2026), who show that models given a third response option cannot separate activation-level

	Dataset ⬇ Model ➡	Rank	Gemma 31B	Gemma E4B	Llama 70B	Llama 8B	Ministral 14B	Ministral 8B	Qwen 27B	Qwen 9B
	steering	1^{st}	True (90.9%)	True (90.9%)	True (50.0%)	False (50.0%)	True (90.9%)	True (90.9%)	False (50.1%)	False (50.0%)
		2^{nd}	False (9.1%)	False (9.1%)	False (50.0%)	True (50.0%)	False (9.1%)	False (9.1%)	True (49.9%)	True (50.0%)
Self-Predictions	**1char**	1^{st}	i (29.3%)	t (20.2%)	a (27.7%)	s (17.7%)	t (26.8%)	t (30.4%)	c (11.2%)	t (14.0%)
		2^{nd}	t (11.2%)	a (14.2%)	t (19.2%)	t (16.9%)	e (19.4%)	e (13.0%)	t (9.4%)	o (13.9%)
		3^{rd}	h (7.0%)	s (9.8%)	i (14.5%)	m (14.2%)	s (10.4%)	s (9.6%)	s (9.2%)	s (8.9%)
	2char	1^{st}	h (18.1%)	h (22.1%)	s (30.1%)	a (31.6%)	h (26.5%)	h (27.9%)	a (15.6%)	h (16.9%)
		2^{nd}	i (17.0%)	i (13.2%)	h (20.4%)	o (11.0%)	l (23.1%)	l (14.0%)	e (14.3%)	o (13.4%)
		3^{rd}	o (15.2%)	n (12.7%)	r (13.4%)	e (11.0%)	i (7.1%)	o (13.2%)	i (13.8%)	i (13.3%)
	3char	1^{st}	c (32.4%)	e (22.9%)	e (45.0%)	g (16.9%)	e (56.4%)	e (50.6%)	e (16.3%)	e (26.3%)
		2^{nd}	e (17.0%)	g (10.9%)	r (10.3%)	a (14.4%)	t (10.5%)	n (9.9%)	t (14.6%)	o (8.3%)
		3^{rd}	t (8.7%)	d (6.8%)	a (7.8%)	r (9.6%)	r (6.5%)	g (5.7%)	a (9.5%)	t (7.7%)
	1word	1^{st}	i (28.1%)	the (10.7%)	is (16.5%)	sage (7.5%)	the (19.8%)	the (22.8%)	cat (6.9%)	of (12.3%)
		2^{nd}	as (4.5%)	dog (6.5%)	as (15.0%)	magenta (5.8%)	elephant (19.4%)	elephant (12.6%)	as (3.8%)	the (6.5%)
		3^{rd}	lion (3.7%)	a (5.5%)	the (11.1%)	teal (4.4%)	as (5.4%)	monkey (6.2%)	with (3.8%)	lion (5.2%)
	2word	1^{st}	cannot (28.1%)	bear (4.8%)	not (21.8%)	turquoise (8.8%)	the (8.1%)	lion (9.9%)	the (5.8%)	the (18.5%)
		2^{nd}	the (4.2%)	the (3.4%)	a (6.8%)	sienna (4.7%)	zebra (5.9%)	the (6.1%)	dog (3.4%)	tiger (3.0%)
		3^{rd}	tiger (2.9%)	rabbit (3.2%)	the (5.4%)	lavender (4.5%)	lion (5.6%)	zebra (5.6%)	horse (2.5%)	bear (2.7%)
	3word	1^{st}	complete (22.5%)	bear (5.0%)	typically (9.5%)	turquoise (6.2%)	the (7.5%)	giraffe (6.4%)	the (3.0%)	great (4.9%)
		2^{nd}	provide (2.3%)	of (2.6%)	a (7.7%)	teal (3.7%)	monkey (4.9%)	snake (4.3%)	lion (2.5%)	bear (3.2%)
		3^{rd}	the (2.3%)	elephant (2.5%)	elephant (3.5%)	umber (3.1%)	lion (3.6%)	lion (4.2%)	bear (2.4%)	the (2.6%)
	vowel	1^{st}	False (57.4%)	False (72.1%)	False (54.3%)	False (79.1%)	False (67.8%)	False (70.4%)	False (74.7%)	False (54.5%)
		2^{nd}	True (42.6%)	True (27.9%)	True (45.7%)	True (20.9%)	True (32.2%)	True (29.6%)	True (25.3%)	True (45.5%)
	among	1^{st}	False (50.3%)	False (50.9%)	False (50.0%)	b_or_d (54.8%)	False (50.0%)	False (50.0%)	False (51.3%)	False (50.0%)
		2^{nd}	True (49.7%)	True (49.1%)	True (50.0%)	a_or_c (45.2%)	True (50.0%)	True (50.0%)	True (48.7%)	True (50.0%)
	ethical	1^{st}	False (82.2%)	True (53.5%)	True (61.5%)	False (57.5%)	False (54.5%)	False (53.9%)	False (69.7%)	True (55.9%)
		2^{nd}	True (17.8%)	False (46.5%)	False (38.5%)	True (42.5%)	True (45.5%)	True (46.1%)	True (30.3%)	False (44.1%)
Confidence	**cities**	1^{st}	True (52.1%)	False (50.1%)	False (50.1%)	True (73.5%)	False (50.7%)	False (50.9%)	False (50.1%)	True (50.8%)
		2^{nd}	False (47.9%)	True (49.9%)	True (49.9%)	False (26.5%)	True (49.3%)	True (49.1%)	True (49.9%)	False (49.2%)
	football	1^{st}	True (50.0%)	False (50.0%)	False (50.0%)	False (83.4%)	True (50.0%)	True (50.0%)	True (50.0%)	True (50.0%)
		2^{nd}	False (50.0%)	True (50.0%)	True (50.0%)	True (16.6%)	False (50.0%)	False (50.0%)	False (50.0%)	False (50.0%)
	math	1^{st}	True (50.4%)	True (51.7%)	True (51.6%)	False (64.0%)	False (50.9%)	True (50.7%)	True (50.2%)	True (50.8%)
		2^{nd}	False (49.6%)	False (48.3%)	False (48.4%)	True (36.0%)	True (49.1%)	False (49.3%)	False (49.8%)	False (49.2%)
	medals	1^{st}	False (50.8%)	True (50.6%)	True (50.4%)	False (55.0%)	True (50.7%)	True (50.8%)	True (50.5%)	True (50.3%)
		2^{nd}	True (49.2%)	False (49.4%)	False (49.6%)	True (45.0%)	False (49.3%)	False (49.2%)	False (49.5%)	False (49.7%)
	people	1^{st}	True (50.0%)	True (50.1%)	False (50.0%)	True (73.4%)	True (51.3%)	False (50.7%)	True (50.9%)	False (51.2%)
		2^{nd}	False (50.0%)	False (49.9%)	True (50.0%)	False (26.6%)	False (48.7%)	True (49.3%)	False (49.1%)	True (48.8%)
	trivia	1^{st}	True (50.7%)	False (50.3%)	True (50.3%)	True (69.1%)	False (52.1%)	False (50.1%)	True (53.4%)	False (50.1%)
		2^{nd}	False (49.3%)	True (49.7%)	False (49.7%)	False (30.9%)	True (47.9%)	True (49.9%)	False (46.6%)	True (49.9%)

Table 28: Top-3 most frequent classes and their share of pooled train/val/test rows, per self-prediction dataset and on-policy model (complements the class-count summary in Table 27). Boolean sub-tasks show the majority/minority split directly; free-form character/word sub-tasks show the single most common character or word and its share, useful for spotting degenerate, low-diversity generations.

interventions from prompt-level ones, suggesting detection may reflect general anomaly sensitivity rather than access to hidden states.

Robustness to encoder capacity. On `true_false/commonsense`, a larger encoder raised the static baseline considerably (Table 10). Here the effect is small (Table 34): averaged over the eight models, moving from `all-MiniLM-L6-v2` to `Qwen3-Embedding-8B` adds 2-4 points on `1char`, `1word` and `ethical` and 8 points on `among`, but leaves every one of them far below the fine-tuned diagonal, while the confidence datasets are already at their ceiling with the smallest encoder.

B.5 Generalization Matrices for the Other Seven Models

To check that the generalization gap observed with `Qwen3.5-9B` (Table 1) is not specific to one model family, we repeated the Introspection Generalization Test on the other seven models of the roster, using the same LoRA fine-tuning setup (Appendix Table 24) and 3 seeds (42, 92, 142), except `Llama-3.1-8B` (Table 38),

Tested on ➡ Trained on ⬇	Llama-8B	Llama-70B	Qwen-9B	Qwen-27B	Gemma-E4B	Gemma-31B	Ministral-8B	Ministral-14B
Baseline	65.0 ± 0.0	76.0 ± 0.0	68.0 ± 0.0	72.0 ± 0.0	62.0 ± 0.0	62.0 ± 0.0	72.0 ± 0.0	70.0 ± 0.0
Average	66.6 ± 3.3	75.1 ± 1.6	66.4 ± 3.3	70.9 ± 0.9	62.3 ± 0.7	62.1 ± 1.1	71.4 ± 0.8	70.9 ± 1.0
steering	68.0 ± 1.4	78.0 ± 0.0	64.0 ± 2.0	72.0 ± 2.0	63.3 ± 1.2	61.3 ± 1.2	70.7 ± 0.9	72.0 ± 1.6
1char	65.0 ± 1.0	71.3 ± 1.2	66.0 ± 2.0	71.3 ± 1.2	62.7 ± 1.2	63.3 ± 2.3	72.0 ± 0.0	70.0 ± 1.6
2char	65.0 ± 1.4	76.0 ± 0.0	60.0 ± 2.0	72.0 ± 2.0	62.7 ± 1.2	62.7 ± 1.2	72.0 ± 1.6	71.3 ± 0.9
3char	63.5 ± 0.7	76.0 ± 0.0	66.7 ± 1.2	71.3 ± 3.1	62.0 ± 0.0	60.7 ± 1.2	71.3 ± 0.9	73.3 ± 0.9
1word	61.7 ± 1.2	72.0 ± 0.0	66.7 ± 1.2	72.0 ± 2.0	62.0 ± 2.0	62.0 ± 0.0	71.3 ± 2.5	70.0 ± 1.6
2word	60.5 ± 0.7	76.0 ± 0.0	68.0 ± 2.0	70.7 ± 2.3	62.7 ± 1.2	62.0 ± 2.0	71.3 ± 0.9	71.3 ± 1.9
3word	61.0 ± 0.0	74.7 ± 1.2	70.0 ± 0.0	70.7 ± 1.2	64.0 ± 0.0	61.3 ± 1.2	70.7 ± 0.9	72.0 ± 0.0
vowel	69.0 ± 0.0	75.3 ± 1.2	60.0 ± 2.0	70.0 ± 0.0	62.0 ± 0.0	61.3 ± 3.1	72.7 ± 0.9	70.0 ± 1.6
among	68.3 ± 1.5	75.3 ± 1.2	61.3 ± 3.1	70.0 ± 2.0	62.0 ± 0.0	61.3 ± 2.3	71.3 ± 0.9	70.7 ± 2.5
ethical	68.0 ± 1.0	75.3 ± 1.2	68.0 ± 0.0	70.0 ± 2.0	62.0 ± 0.0	61.3 ± 1.2	71.3 ± 0.9	70.0 ± 0.0
cities	70.3 ± 1.2	75.3 ± 1.2	68.7 ± 1.2	70.7 ± 1.2	62.7 ± 1.2	63.3 ± 3.1	73.3 ± 0.9	70.7 ± 0.9
football	68.7 ± 1.5	74.7 ± 2.3	68.0 ± 0.0	71.3 ± 1.2	62.0 ± 0.0	61.3 ± 1.2	70.7 ± 0.9	70.0 ± 0.0
math	70.0 ± 1.7	74.0 ± 2.0	69.3 ± 1.2	70.7 ± 1.2	62.7 ± 1.2	64.0 ± 0.0	72.0 ± 0.0	71.3 ± 0.9
medals	69.0 ± 0.0	75.3 ± 1.2	67.3 ± 1.2	72.0 ± 0.0	60.7 ± 1.2	61.3 ± 2.3	70.0 ± 2.8	70.7 ± 0.9
people	69.7 ± 1.5	76.0 ± 0.0	68.0 ± 0.0	72.0 ± 2.0	62.0 ± 0.0	62.0 ± 2.0	70.7 ± 0.9	71.3 ± 0.9
trivia	67.7 ± 1.5	76.0 ± 0.0	70.0 ± 0.0	69.3 ± 1.2	62.0 ± 0.0	64.0 ± 0.0	71.3 ± 2.5	69.3 ± 0.9

Table 29: **Evaluation of instruction following ability.** This table acts as a control matrix to assess whether metacognitive fine-tuning induces catastrophic forgetting. Columns represent standard general utility benchmark assessing strict constraint adherence (IFEval). The **Average** row reports the mean and standard deviation of the 16 task-checkpoint means for that model (std computed across tasks, not across seeds); it stays within $\sim$2pp of that model's zero-shot **Baseline** for every model, indicating no systematic degradation in constraint-following from task-aware fine-tuning.

Checkpoint	Llama-8B	Llama-70B	Qwen-9B	Qwen-27B	Gemma-E4B	Gemma-31B	Ministral-8B	Ministral-14B
Baseline	0.615	0.755	0.731	0.777	0.606	0.810	0.572	0.538
steering	0.679 ± 0.008	0.696 ± 0.012	0.741 ± 0.004	–	0.618 ± 0.008	0.818 ± 0.005	0.669 ± 0.016	0.611 ± 0.026
1char	0.620 ± 0.005	0.762 ± 0.002	0.727 ± 0.012	0.776 ± 0.014	0.601 ± 0.002	0.804 ± 0.005	0.647 ± 0.002	0.535 ± 0.008
2char	0.602 ± 0.016	0.764 ± 0.002	0.736 ± 0.002	0.780 ± 0.005	0.596 ± 0.003	0.806 ± 0.002	0.659 ± 0.018	0.471 ± 0.019
3char	0.601 ± 0.015	0.762 ± 0.006	0.729 ± 0.007	0.779 ± 0.005	0.604 ± 0.005	0.803 ± 0.007	0.657 ± 0.003	0.504 ± 0.013
1word	0.634 ± 0.007	0.757 ± 0.004	0.709 ± 0.006	0.782 ± 0.010	0.620 ± 0.005	0.804 ± 0.009	0.645 ± 0.009	0.524 ± 0.018
2word	0.603 ± 0.017	0.767 ± 0.005	0.707 ± 0.002	0.781 ± 0.004	0.624 ± 0.005	0.818 ± 0.005	0.654 ± 0.012	0.600 ± 0.019
3word	0.632 ± 0.010	0.768 ± 0.003	0.720 ± 0.002	0.780 ± 0.003	0.618 ± 0.003	0.806 ± 0.006	0.649 ± 0.015	0.556 ± 0.021
vowel	0.541 ± 0.026	0.756 ± 0.005	0.734 ± 0.003	0.778 ± 0.002	0.589 ± 0.005	0.803 ± 0.006	0.623 ± 0.018	0.475 ± 0.007
among	0.631 ± 0.008	0.766 ± 0.002	0.699 ± 0.002	0.785 ± 0.005	0.597 ± 0.004	0.812 ± 0.007	0.640 ± 0.011	0.587 ± 0.021
ethical	0.623 ± 0.008	0.757 ± 0.002	0.727 ± 0.006	0.782 ± 0.008	0.599 ± 0.003	0.818 ± 0.006	0.668 ± 0.010	0.573 ± 0.005
cities	0.644 ± 0.007	0.759 ± 0.005	0.700 ± 0.005	0.784 ± 0.002	0.595 ± 0.002	0.810 ± 0.003	0.625 ± 0.006	0.533 ± 0.009
football	0.615 ± 0.006	0.759 ± 0.004	0.716 ± 0.005	0.781 ± 0.002	0.597 ± 0.002	0.810 ± 0.000	0.624 ± 0.008	0.533 ± 0.018
math	0.572 ± 0.028	0.760 ± 0.004	0.705 ± 0.005	0.765 ± 0.003	0.581 ± 0.008	0.812 ± 0.008	0.611 ± 0.015	0.542 ± 0.038
medals	0.620 ± 0.007	0.755 ± 0.008	0.706 ± 0.008	0.784 ± 0.012	0.593 ± 0.003	0.813 ± 0.008	0.597 ± 0.014	0.503 ± 0.052
people	0.640 ± 0.014	0.772 ± 0.012	0.698 ± 0.002	0.775 ± 0.008	0.598 ± 0.006	0.808 ± 0.008	0.633 ± 0.008	0.551 ± 0.042
trivia	0.634 ± 0.014	0.735 ± 0.007	0.682 ± 0.003	0.779 ± 0.002	0.583 ± 0.008	0.812 ± 0.002	0.589 ± 0.025	0.438 ± 0.004

Table 30: **Zero-shot BoolQ accuracy of the task-aware fine-tuned checkpoints, as a catastrophic-forgetting control.** Rows are the same checkpoints as Table 29 (one row per introspection task the checkpoint was fine-tuned on, plus the un-fine-tuned *Baseline*), evaluated instead on BoolQ, one column per base model. Each cell is the mean ± standard deviation across seeds (5 for Llama-8B, 3 for the others; *Baseline* rows are single-run). All eight columns have $\geq 99\%$ True/False answer-parse rates.

which uses 5. Tables 35-41 report the full train-task × eval-task matrix for each. Majority-class and zero-shot baseline rows are included where they were re-computed for that model and omitted otherwise, as noted in each caption; model-specific caveats are given in the captions as well.

Checkpoint	Llama-8B	Llama-70B	Qwen-9B	Qwen-27B	Gemma-E4B	Gemma-31B	Ministral-8B	Ministral-14B
Baseline	0.540	0.786	0.768	0.796	0.534	0.828	0.527	0.455
steering	0.758 ± 0.007	0.784 ± 0.006	0.793 ± 0.008	–	0.556 ± 0.019	0.838 ± 0.004	0.741 ± 0.020	0.545 ± 0.060
1char	0.571 ± 0.018	0.794 ± 0.003	0.767 ± 0.010	0.786 ± 0.020	0.528 ± 0.004	0.821 ± 0.006	0.661 ± 0.010	0.370 ± 0.022
2char	0.512 ± 0.029	0.792 ± 0.002	0.780 ± 0.005	0.804 ± 0.007	0.519 ± 0.006	0.826 ± 0.004	0.688 ± 0.020	0.183 ± 0.074
3char	0.508 ± 0.038	0.796 ± 0.006	0.768 ± 0.008	0.803 ± 0.006	0.538 ± 0.008	0.821 ± 0.005	0.646 ± 0.012	0.286 ± 0.038
1word	0.629 ± 0.010	0.787 ± 0.004	0.741 ± 0.018	0.822 ± 0.007	0.570 ± 0.004	0.822 ± 0.008	0.638 ± 0.010	0.348 ± 0.047
2word	0.538 ± 0.036	0.794 ± 0.004	0.743 ± 0.002	0.796 ± 0.001	0.567 ± 0.013	0.834 ± 0.004	0.678 ± 0.017	0.535 ± 0.035
3word	0.611 ± 0.011	0.798 ± 0.003	0.751 ± 0.003	0.797 ± 0.004	0.554 ± 0.001	0.825 ± 0.005	0.664 ± 0.024	0.426 ± 0.065
vowel	0.356 ± 0.064	0.790 ± 0.005	0.787 ± 0.007	0.794 ± 0.003	0.498 ± 0.007	0.819 ± 0.009	0.562 ± 0.033	0.184 ± 0.020
among	0.588 ± 0.007	0.793 ± 0.001	0.718 ± 0.006	0.803 ± 0.005	0.523 ± 0.013	0.828 ± 0.007	0.617 ± 0.021	0.497 ± 0.045
ethical	0.592 ± 0.014	0.792 ± 0.002	0.771 ± 0.000	0.803 ± 0.006	0.525 ± 0.004	0.839 ± 0.007	0.686 ± 0.014	0.483 ± 0.012
cities	0.628 ± 0.008	0.789 ± 0.004	0.719 ± 0.005	0.794 ± 0.002	0.519 ± 0.008	0.828 ± 0.004	0.587 ± 0.022	0.360 ± 0.029
football	0.560 ± 0.015	0.792 ± 0.003	0.744 ± 0.004	0.796 ± 0.002	0.519 ± 0.004	0.828 ± 0.000	0.591 ± 0.019	0.389 ± 0.050
math	0.432 ± 0.070	0.799 ± 0.003	0.735 ± 0.010	0.781 ± 0.002	0.473 ± 0.021	0.830 ± 0.008	0.550 ± 0.054	0.390 ± 0.090
medals	0.574 ± 0.016	0.788 ± 0.006	0.731 ± 0.012	0.791 ± 0.015	0.509 ± 0.006	0.830 ± 0.006	0.514 ± 0.038	0.272 ± 0.167
people	0.619 ± 0.026	0.809 ± 0.012	0.711 ± 0.003	0.779 ± 0.010	0.522 ± 0.015	0.826 ± 0.008	0.602 ± 0.019	0.411 ± 0.101
trivia	0.624 ± 0.023	0.798 ± 0.002	0.670 ± 0.002	0.798 ± 0.004	0.481 ± 0.023	0.828 ± 0.003	0.514 ± 0.054	0.067 ± 0.024

Table 31: **Zero-shot BoolQ F1 score of the task-aware fine-tuned checkpoints, as a catastrophic-forgetting control.** Same rows, columns, and seed counts as Table 30, reporting F1 instead of accuracy.

	Task	Majority	Static (MiniLM)	Static (best enc.)	Fine-tuned diag.	Static $\geq$ diag.
Self-prediction	1char	21.8	34.7	37.7	59.2	0/8
	2char	25.3	40.3	43.2	47.0	2/8
	3char	28.5	38.0	40.1	41.2	4/8
	1word	16.7	31.4	33.8	57.6	0/8
	2word	12.4	20.6	22.5	27.2	2/8
	3word	9.1	16.0	16.6	17.9	4/8
	vowel	67.4	79.0	81.3	78.2	8/8
	among	52.6	55.4	63.5	88.6	0/8
	ethical	63.5	69.2	74.5	83.9	0/8
Confidence	cities	56.2	62.2	72.8	76.7	3/8
	football	63.2	78.9	82.2	69.3	8/8
	math	55.6	83.4	85.2	77.1	8/8
	medals	54.0	84.8	86.1	67.7	8/8
	people	54.2	68.8	75.0	71.9	6/8
	trivia	55.3	58.3	60.5	53.4	8/8
Steering	steering	70.6	71.8	72.2	68.0	–

Table 32: **Static-embedding baseline vs. the fine-tuned diagonal, averaged over the eight models** (accuracy, %). *Static*: best of 13 non-autoregressive classifiers on frozen embeddings of the exact prompt, fit on each model's own training split; *best enc.*: best over five encoders. *Fine-tuned diag.*: in-task accuracy of the task-specialized model (bold diagonal of Tables 38-41). Last column: number of models for which the static baseline matches or beats the diagonal. For steering, the baseline predicts only whether the model was steered while the diagonal also requires naming the concept, so the two are not compared. Per-model values: Table 33.

B.6 LoRA coverage ablation on Qwen3.5's hybrid attention

With the default recipe, both Qwen3.5 models detected steering less reliably than the other models; the widened adapter motivated below is the one used for the Qwen3.5-9B steering row of Table 1. We identified a concrete mechanistic candidate for this gap: Qwen3.5 (both 9B and 27B) is not a dense transformer. Its layers follow a fixed 4:1 pattern (full_attention_interval=4) in which only 1-in-4 layers is standard multi-

Task	Llama-8B	Llama-70B	Qwen-9B	Qwen-27B	Gemma-E4B	Gemma-31B	Ministral-8B	Ministral-14B
1char	30/55	43/71	34/59	26/61	32/37	46/78	46/53	44/60
2char	30/36	53/68	42/44	34/53	**44/33**	47/51	**48/44**	47/48
3char	**36/22**	**47/44**	**36/35**	24/35	**33/27**	47/65	48/48	49/52
1word	27/57	46/80	31/61	15/66	28/48	40/48	42/44	42/59
2word	**19/18**	32/41	24/26	10/30	**19/10**	36/50	21/23	18/19
3word	**18/15**	17/20	13/16	8/29	**8/6**	29/35	**22/8**	**17/14**
vowel	**83/79**	**88/88**	**77/74**	**76/71**	**82/80**	**84/83**	**77/74**	**83/77**
among	69/88	67/89	60/93	57/90	64/84	60/94	65/83	65/87
ethical	81/87	72/84	68/74	80/89	66/81	89/93	71/82	69/80
cities	**74/70**	79/85	**70/65**	68/78	77/89	70/82	**68/66**	76/78
football	**94/93**	**60/56**	**80/65**	**100/79**	**100/94**	**72/51**	**77/63**	**74/53**
math	**91/90**	**82/82**	**88/70**	**85/81**	**81/71**	**80/50**	**85/84**	**89/88**
medals	**76/62**	**84/74**	**89/66**	**94/84**	**89/68**	**87/56**	**88/67**	**82/64**
people	**70/68**	**71/57**	**69/64**	**73/71**	91/94	**71/67**	78/82	**77/72**
trivia	**67/65**	**59/51**	**57/54**	**62/51**	**62/56**	**59/46**	**61/51**	**57/52**
steering	52/89	52/55	54/61	56/78	91/58	91/46	91/76	91/82

Table 33: **Static-embedding baseline vs. fine-tuned diagonal, per model.** Each cell: best static accuracy over the five encoders / fine-tuned diagonal (%); **bold** where the static baseline matches or beats the diagonal. `Llama-3.1-8B` uses the `train_new`/`test_new` splits of its fine-tuned models. On `medals`, 24-55% of test questions also appear verbatim in the training split for every model, which the static classifiers can exploit. The Gemma and Ministral steering test splits are 91% steered, so their static steering cells equal the majority class; the `Qwen3.5-27B` steering cell is scored on the validation split, since its test split has no steering file.

Task	**Majority**	**MiniLM**	**mpnet**	**bge-large**	**Qwen3-0.6B**	**Qwen3-8B**	**Fine-tuned**
1st char	21.8	34.7	36.6	35.0	36.2	37.0	59.2
1st word	16.7	31.4	32.8	31.0	31.8	33.4	57.6
ethical stance	63.5	69.2	71.5	71.3	73.0	73.5	83.9
among	52.6	55.4	56.1	55.7	55.8	63.4	88.6
math	55.6	83.4	84.2	83.7	83.9	84.4	77.1
medals	54.0	84.8	83.5	84.2	83.5	84.0	67.7

Table 34: Best-of-13-classifier accuracy (%) as the frozen encoder scales from `all-MiniLM-L6-v2` (22M) to `Qwen3-Embedding-8B` (8B), averaged over the eight models, for the four tasks the static baseline does not solve and two it does. Unlike `true_false`/`commonsense` (Table 10), no encoder closes the gap to the fine-tuned diagonal on the first four.

head attention (`q_proj`/`k_proj`/`v_proj`/`o_proj`); the remaining 3-in-4 layers are Gated DeltaNet blocks, a linear/recurrent-attention variant with entirely different projections (`in_proj_qkv`, `in_proj_z`, `in_proj_a`, `in_proj_b`, `out_proj`). Our fine-tuning recipe's default LoRA target modules, `q_proj`/`v_proj`, only ever adapt the 1-in-4 standard-attention layers. Concretely, of the four concept-injection layers used for steering awareness, 2 of 4 fall inside Gated DeltaNet blocks and were therefore *never touched by any LoRA adapter* in the results reported above (layers $6, 12$ of $\{6, 12, 19, 23\}$ for `Qwen3.5-9B`; layers $38, 46$ of $\{12, 24, 38, 46\}$ for `Qwen3.5-27B`).

To test whether this under-adaptation explains the gap, we re-ran steering-awareness fine-tuning (seed 42, evaluation restricted to steering-awareness only) with `lora_target_modules` widened to also cover the Gated DeltaNet projections, on both the disjoint-concept split (held-out concepts, tests genuine generalization) and the overlapping-concept split (same ~ 50 concepts in train and test, tests memorization), for both `Qwen3.5-9B` and `Qwen3.5-27B`. Table 42 reports *steering accuracy* (binary: did the model notice it was being steered at all) and *concept accuracy*, isolated to the truly-steered rows only: of the examples that were actually

Tested on ➡ / Trained on ⬇		*steering*	*1char*	*2char*	*3char*	*1word*	*2word*	*3word*	*vowel*	*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
Majority		50.0	11.0	18.7	15.6	7.5	4.0	4.0	71.0	50.5	76.5	52.8	52.4	53.9	57.8	52.8	50.0
Baseline		0.0	0.0	1.0	7.0	0.0	0.0	0.0	71.0	0.0	50.6	47.2	52.4	46.1	42.2	47.2	50.0
	steering	**78.2**	29.8	22.8	15.7	33.0	8.8	10.3	29.0	51.7	23.5	52.8	47.6	53.9	57.8	52.8	50.0
Self-Predictions	**1char**	50.0	**61.2**	30.2	25.7	46.8	19.2	15.7	71.2	82.2	78.8	67.8	50.0	52.8	61.1	57.5	54.2
	2char	50.0	47.7	**52.8**	25.0	53.5	13.3	12.3	71.0	80.3	78.8	63.3	47.6	53.9	59.4	54.4	49.4
	3char	50.0	43.0	36.5	**35.3**	51.7	17.8	16.2	68.2	79.8	79.6	68.3	47.6	53.9	59.4	55.0	50.0
	1word	50.0	55.0	30.3	22.7	**65.7**	15.0	12.7	43.0	77.0	84.7	72.8	71.4	53.9	71.1	58.9	51.1
	2word	50.0	42.5	26.8	23.0	47.8	**30.5**	19.5	71.0	83.3	79.2	64.4	47.6	53.9	59.4	54.4	51.1
	3word	50.0	44.8	30.5	24.5	39.5	22.7	**29.0**	70.3	81.2	80.0	68.9	47.6	53.9	62.2	56.7	50.0
	vowel	50.0	46.8	28.2	23.5	51.5	16.5	14.2	**71.0**	80.5	78.4	62.2	47.6	53.9	59.4	54.4	50.6
	among	50.0	46.0	29.8	24.0	48.8	13.0	12.3	71.0	**89.7**	79.6	65.8	47.6	53.9	59.4	54.7	51.1
	ethical	50.0	43.0	29.0	24.7	51.8	11.3	11.3	57.2	78.3	**89.4**	59.2	47.6	53.9	58.6	53.3	50.8
Confidence	**cities**	50.0	45.6	28.7	23.3	44.2	12.8	13.2	71.9	82.4	77.3	**78.1**	71.4	50.8	63.9	62.8	50.6
	football	50.0	45.9	28.1	22.7	46.3	13.8	13.7	71.7	80.5	77.6	76.9	**78.6**	52.2	69.2	61.4	50.6
	math	50.0	44.4	28.2	23.2	52.2	14.1	12.1	59.8	80.0	78.8	71.7	54.8	**81.4**	45.6	47.2	51.1
	medals	50.0	46.3	28.3	23.2	44.1	13.3	13.6	71.9	84.6	76.5	76.4	81.0	51.1	**83.6**	63.6	54.2
	people	50.0	46.0	27.8	23.7	45.5	14.0	13.2	71.0	83.8	76.5	79.4	73.8	53.1	73.9	**70.8**	51.7
	trivia	50.0	44.8	28.0	23.5	49.0	13.7	12.7	70.8	78.0	78.0	58.1	64.3	53.9	60.3	55.8	**51.4**

Table 35: **Evaluation matrix of introspective capabilities for `Qwen3.5-27B`.** Same layout as Table 38 (rows = training task, columns = evaluation task, bold diagonal = in-domain accuracy, green = column-best), mean accuracy across 3 seeds (42, 92, 142), including majority-class and zero-shot baseline rows (same pipeline as Table 38). The **steering** row and column are included, with two caveats specific to this model. (i) Unlike the other 15 task checkpoints (narrow `q_proj/v_proj` LoRA), the steering-awareness checkpoint uses the wide, Gated-DeltaNet-inclusive LoRA target set of Appendix B.6, without which `Qwen3.5` cannot even detect steering; its diagonal (78.2) is binary steering-detection accuracy (3-seed mean), and concept-naming stays near chance (Appendix B.6). (ii) The steering row's six **confidence** cells are carried over from an earlier evaluation that stored only answer-accuracy; every other cell uses the current pipeline. The steering column (each non-steering checkpoint evaluated on steering awareness) is uniformly $\sim 50\%$, i.e. chance, matching `Qwen3.5-9B` (Table 1).

	Tested on ➡ / Trained on ⬇	*steering*	*1char*	*2char*	*3char*	*1word*	*2word*	*3word*	*vowel*	*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
	Majority	9.1	31.1	28.5	45.8	27.2	10.2	7.0	73.9	61.7	50.6	60.0	56.4	52.8	58.9	50.6	55.6
	Baseline	9.1	26.7	4.4	31.7	24.4	2.2	0.6	73.9	76.4	52.9	59.4	56.4	50.6	47.8	49.4	53.3
	steering	**60.9**	21.7	5.2	18.9	9.4	5.6	0.7	26.1	55.7	49.0	49.1	56.4	58.1	59.1	51.1	44.3
Self-Predictions	**1char**	4.0	**52.8**	2.0	16.3	39.1	1.1	0.7	73.0	75.7	55.3	58.5	57.3	56.5	48.0	49.4	55.9
Self-Predictions	**2char**	4.6	7.0	**43.7**	3.3	13.9	9.6	2.4	53.9	73.9	56.9	61.8	68.4	54.4	54.6	53.0	53.9
Self-Predictions	**3char**	3.0	26.9	7.0	**48.5**	20.2	3.0	1.9	63.5	76.9	54.1	59.3	60.7	54.4	48.3	50.0	55.7
Self-Predictions	**1word**	3.7	44.3	4.1	18.0	**43.9**	3.0	2.0	73.7	77.0	49.0	58.3	59.0	61.7	50.0	49.4	55.6
Self-Predictions	**2word**	9.1	4.4	1.3	7.8	25.9	**22.8**	6.5	73.5	78.3	59.2	59.8	57.3	61.3	50.7	49.6	54.1
Self-Predictions	**3word**	7.2	16.3	0.9	3.5	11.1	15.4	**7.6**	73.7	78.1	47.8	59.6	58.1	58.1	52.4	48.1	57.2
Self-Predictions	**vowel**	7.4	28.5	5.9	38.3	15.9	2.0	0.2	**73.9**	70.0	49.4	59.6	45.3	64.3	44.3	49.4	54.8
Self-Predictions	**among**	8.2	34.8	4.9	27.4	18.1	1.1	0.9	62.2	**83.4**	63.0	62.1	56.0	53.3	55.5	52.6	52.7
Self-Predictions	**ethical**	9.1	35.0	5.4	39.8	18.0	1.1	0.4	42.0	76.5	**82.0**	64.1	59.8	48.7	56.7	57.0	52.2
Confidence	**cities**	5.3	32.2	5.9	29.1	20.0	2.0	1.1	64.6	74.9	22.4	**65.9**	63.2	52.4	56.5	53.9	51.1
Confidence	**football**	2.2	29.6	5.2	33.5	21.5	0.9	0.6	70.6	76.2	27.5	62.0	**63.2**	50.0	54.8	57.6	55.4
Confidence	**math**	0.6	31.1	5.2	30.6	20.7	3.1	1.1	47.8	71.0	53.3	60.6	53.0	**84.1**	50.9	47.2	52.0
Confidence	**medals**	6.1	30.4	5.0	28.9	15.6	1.5	1.7	64.3	71.5	37.6	60.0	53.0	58.0	**67.4**	55.2	55.2
Confidence	**people**	5.4	36.1	5.2	26.9	20.0	0.9	0.9	67.2	75.8	31.0	62.8	59.8	51.9	56.5	**81.5**	49.1
Confidence	**trivia**	4.9	23.9	5.2	33.9	23.3	2.4	0.2	67.8	75.7	51.0	60.2	58.1	58.1	53.9	54.8	**51.1**

Table 36: **Evaluation matrix of introspective capabilities for `Ministral-3-8B`.** Same layout as Table 38 (rows = training task, columns = evaluation task, bold diagonal = in-domain accuracy, green = column-best), mean accuracy across 3 seeds (42, 92, 142). Self-prediction accuracies reflect the corrected evaluator (Section 5.2); prior to that fix, `MistralCommonBackend`'s tokenizer left a literal `</s>` in every generation, which the strict-match parsers rejected, reporting near-0% self-prediction accuracy across the board; the fix was verified against raw generations before being applied across the full matrix, not just the diagonal. *Majority* and zero-shot *Baseline* rows use the same majority-class/zero-shot evaluation pipeline as Table 38.

steered, what fraction did the model name correctly. We compare the narrow (`q_proj/v_proj` only) and wide (DeltaNet-inclusive) adapters.

Widening LoRA coverage produces a large, consistent improvement in *steering accuracy* across all four conditions (e.g. $52.0\% \rightarrow 96.0\%$ for `Qwen3.5-9B` disjoint), confirming that the narrow adapter's failure to reach 2 of the 4 injection layers was a real bottleneck for *noticing* that steering occurred. *Concept accuracy* on the disjoint (held-out concept) split, however, is a near-total failure regardless of LoRA width: `Qwen3.5-9B` is exactly unchanged ($4.0\% \rightarrow 4.0\%$ - literally the same 5 of 124 steered rows correct in both conditions), and `Qwen3.5-27B` moves only from complete failure to still-near-zero ($0.0\% \rightarrow 5.0\%$). By contrast, the overlapping (memorized-concept) split shows real, substantial naming ability once LoRA is widened - 25.8% (9B) and 45.2% (27B), rising to 53.3% and 73.7% respectively conditional on the model having correctly detected steering at all. In other words, full architectural coverage fixes the coarser "am I being perturbed at all" signal but does not unlock genuine generalization to *naming* a never-seen concept - the disjoint-split result is not a small residual gap but flat failure, while the overlapping-split improvement is better read as memorization of the ~ 50 shared train/test concepts than as generalization. This sharpens, rather than merely being consistent with, this paper's broader causal-bypassing concern.

B.7 Steering Awareness: Concept-Vector Construction and Injection Setup

This subsection documents how the steering-awareness training and evaluation data were constructed: how the concept vectors v were extracted, at which layers l they were injected, and with which strengths α. The

	Tested on ➡ / Trained on ⬇	*steering*	*1char*	*2char*	*3char*	*1word*	*2word*	*3word*	*vowel*	*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
	Majority	9.1	21.1	27.8	52.8	22.8	6.1	9.4	70.0	56.7	54.1	51.1	61.9	50.6	50.6	53.3	52.8
	Baseline	9.1	22.2	12.8	18.3	21.7	0.6	0.6	61.1	75.8	75.3	56.1	42.9	50.0	42.2	61.7	51.7
	steering	**34.1**	32.0	10.6	14.1	21.9	1.7	1.3	29.8	74.1	49.4	70.4	60.3	41.7	49.6	62.4	45.6
Self-Predictions	**1char**	9.1	**60.2**	1.3	28.3	52.8	3.1	4.4	55.6	75.5	70.2	66.7	55.6	44.1	48.0	67.6	46.5
	2char	5.5	21.1	**48.3**	1.9	31.1	5.4	3.7	51.5	74.9	67.5	65.9	56.3	49.3	47.8	67.8	45.0
	3char	6.4	36.1	6.9	**52.0**	36.7	6.3	5.7	35.9	74.0	58.0	67.8	52.4	48.1	48.7	63.5	43.9
	1word	9.1	33.5	3.7	30.6	**58.7**	4.6	4.8	67.0	74.7	61.6	68.3	57.1	48.1	48.1	61.3	45.9
	2word	9.1	5.9	2.2	16.1	26.9	**19.4**	10.0	60.0	76.8	75.7	71.5	61.9	45.9	48.5	57.6	43.5
	3word	6.7	15.0	3.3	8.5	24.3	13.5	**13.7**	65.9	76.4	72.5	72.2	58.7	47.4	49.1	61.5	44.3
	vowel	9.1	26.3	10.4	22.4	12.8	0.9	0.6	**76.7**	74.5	61.2	66.9	54.8	46.5	48.0	67.6	46.3
	among	7.7	31.3	14.0	15.8	25.6	1.5	1.5	57.7	**87.3**	76.7	70.5	59.1	45.2	49.1	63.0	44.1
	ethical	0.9	27.2	10.7	17.4	23.1	0.7	0.6	39.6	77.6	**80.4**	70.6	56.3	38.5	49.3	62.4	43.7
Confidence	**cities**	3.2	30.0	8.9	26.5	20.0	0.9	0.7	62.4	75.6	68.2	**78.0**	56.3	44.4	49.3	62.2	43.7
	football	1.2	29.1	12.6	18.5	19.8	0.7	0.6	55.9	75.0	67.8	65.4	**53.2**	49.8	49.3	57.0	48.0
	math	6.7	31.1	9.8	25.2	22.4	0.7	1.1	50.2	72.0	61.6	58.3	41.3	**88.3**	54.1	55.6	47.2
	medals	8.8	28.1	10.2	26.3	18.9	1.3	0.7	70.6	73.4	68.2	53.1	45.2	61.9	**63.7**	53.0	53.5
	people	9.1	32.6	8.9	22.2	23.3	1.9	1.3	60.2	72.9	64.7	68.0	57.9	48.9	50.6	**71.5**	45.6
	trivia	0.6	31.3	8.3	25.7	21.5	1.5	0.6	70.2	72.3	66.3	57.6	46.8	60.7	49.3	54.4	**52.4**

Table 37: **Evaluation matrix of introspective capabilities for `Ministral-3-14B`.** Same layout as Table 38 (rows = training task, columns = evaluation task, bold diagonal = in-domain accuracy, green = column-best), mean accuracy across 3 seeds (42, 92, 142). Self-prediction accuracies reflect the corrected evaluator (Section 5.2); prior to that fix, `MistralCommonBackend`'s tokenizer left a literal `</s>` in every generation, which the strict-match parsers rejected, reporting near-0% self-prediction accuracy across the board; the fix was verified against raw generations before being applied across the full matrix, not just the diagonal. *Majority* and zero-shot *Baseline* rows use the same majority-class/zero-shot evaluation pipeline as Table 38.

setup follows the concept-injection paradigm of Lindsey (2025) and its fine-tuning adaptation by Rivera & Africa (2026).

Concept set. We use a fixed pool of 500 single-noun concepts:

- 50 **concepts from the original steering-detection work (Lindsey, 2025)**: Dust, Satellites, Trumpets, Origami, Illusions, Cameras, Lightning, Constellations, Treasures, Phones, Trees, Avalanches, Mirrors, Fountains, Quarries, Sadness, Xylophones, Secrecy, Oceans, Happiness, Deserts, Kaleidoscopes, Sugar, Vegetables, Poetry, Aquariums, Bags, Peace, Caverns, Memories, Frosts, Volcanoes, Boulders, Harmonies, Masquerades, Rubber, Plastic, Blood, Amphitheaters, Contraptions, Youths, Dynasties, Snow, Dirigibles, Algorithms, Denim, Monoliths, Milk, Bread, Silver.
- 450 **additional concepts** chosen not to overlap lexically with the original 50 or with the baseline-word pool below. They span 20 categories: Animals, Body Parts, Colors, Abstract Concepts, Emotions, Foods, Materials, Professions, Actions, Science Terms, Everyday Objects, Buildings, Vehicles, Sports, Music/Art, Weather, Geography, Time, Social Concepts, Plants. Some examples are: Elephants, Crimson, Probability, Anxiety, Chocolate, Copper, Surgeons, Swimming, Photons, Scissors, Cathedrals, Motorcycles, Basketball, Violins, Thunderstorms, Lakes, Centuries, Weddings, Roses.

For the *concept-disjoint* datasets (the default), the pool is partitioned 80/10/10 at the concept level into 400 train, 50 validation and 50 test concepts, so that no concept seen during fine-tuning appears at evaluation.

		Steering	Self-Predictions									Confidence Reporting					
Tested on ➡ / Trained on ⬇		*steering*	*1char*	*2char*	*3char*	*1word*	*2word*	*3word*	*vowel*	*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
Majority		50.8	18.6	27.4	18.5	8.0	8.0	7.0	79.0	50.0	66.0	71.0	91.0	68.0	53.0	68.0	**67.0**
Baseline		56.6	10.0	1.0	0.0	2.0	11.0	0.0	57.0	66.0	65.0	48.0	91.0	60.0	53.0	55.0	66.0
Static emb.		52.4	23.0	28.0	31.0	22.0	18.0	17.0	82.0	61.7	74.0	73.0	94.0	91.0	76.0	69.0	67.0
	steering	**83.9**	5.7	4.0	1.7	4.0	0.3	0.3	21.0	50.0	34.3	71.0	9.0	30.7	47.0	68.0	67.0
Self-Predictions	**1char**	64.6	**55.3**	2.5	3.7	23.5	1.0	1.8	37.3	68.8	72.3	64.5	73.5	53.8	49.2	64.5	62.0
	2char	59.3	12.5	**35.8**	16.8	11.3	0.3	1.5	62.7	52.7	43.2	62.5	84.0	55.8	50.3	61.8	59.2
	3char	56.7	7.5	4.7	**22.2**	13.3	0.0	0.8	58.5	68.3	71.2	51.2	86.7	53.0	49.3	56.8	62.2
	1word	61.9	11.0	0.3	0.7	**56.7**	1.5	2.2	42.3	61.5	67.5	70.5	42.0	52.3	46.3	66.2	57.3
	2word	44.3	8.7	1.2	0.0	17.0	**17.8**	8.0	73.3	45.7	36.8	60.5	47.2	46.2	41.3	41.8	53.3
	3word	36.9	5.7	1.0	0.0	15.3	17.0	**15.3**	77.3	45.6	36.8	50.3	46.2	42.3	47.0	36.7	54.8
	vowel	58.3	9.8	9.2	9.5	7.3	0.0	0.0	**79.0**	53.2	66.7	30.7	91.0	60.8	52.7	33.8	40.3
	among	60.9	25.2	2.8	2.8	19.8	0.0	0.0	72.2	**88.5**	80.8	67.8	54.3	51.3	48.7	67.5	63.0
	ethical	61.8	20.5	3.5	2.3	18.0	0.0	0.8	23.7	73.6	**86.8**	73.0	36.5	61.8	47.7	66.0	57.8
Confidence	**cities**	44.7	9.2	4.2	1.7	8.3	0.5	0.2	21.8	55.0	72.8	**70.2**	10.7	31.0	47.8	65.7	65.0
	football	60.3	9.5	5.0	1.5	7.8	0.0	0.0	60.5	64.0	68.7	50.2	**93.0**	66.3	53.5	53.3	63.5
	math	55.1	12.3	6.2	4.8	10.7	0.0	0.2	78.7	59.2	68.0	34.2	91.0	**90.0**	53.0	36.8	44.5
	medals	59.2	8.8	4.0	2.3	8.2	0.2	0.0	62.2	66.5	68.3	47.2	93.7	68.8	**62.2**	55.7	63.5
	people	47.0	7.5	5.0	3.5	8.3	0.0	0.0	21.3	54.6	73.5	71.7	16.2	29.5	48.2	**68.3**	64.3
	trivia	56.8	9.7	5.0	2.7	8.7	0.2	0.0	47.2	65.6	67.7	67.2	46.0	68.0	48.7	69.2	**65.3**

Table 38: **Evaluation matrix of introspective capabilities for `Llama-3.1-8B`.** Baselines (top rows) are compared against task-aware fine-tuned models $M_{d_{train}}$ (bottom rows) across evaluation tasks (columns). *Static emb.*: best of 13 non-autoregressive classifiers on frozen `all-MiniLM-L6-v2` embeddings of the prompt (Table 32). Cells report mean accuracy over 5 random seeds. The **bold diagonal** indicates in-task performance (matched train/test), while off-diagonal elements evaluate cross-functional generalization (In-Domain vs. Out-of-Domain). Green cells highlight the highest accuracy per column. The **steering** row uses the disjoint-concept steering adapter of Appendix B.8 (seeds 42, 92, 142); its diagonal is detection balanced accuracy.

Concept-vector extraction. Each concept vector is a difference-of-means direction in the residual stream. For concept c we take the last-token activation $a_l(\cdot)$ at the target layer l of the chat-formatted prompt "`Tell me about {word}`", and set

$$v_{c,l} \;=\; a_l(c) \;-\; \frac{1}{|B|}\sum_{w\in B} a_l(w),$$

where B is a fixed set of 100 generic baseline words (disjoint from the concept pool). Vectors are extracted once per (concept, layer) pair from the base model and cached to disk. At injection time, each vector is L_2-normalized, so the strength α is the sole magnitude control.

Injection. A steered row injects a single concept vector at a single target layer: an additive forward hook on the output of transformer block l adds $\alpha\,\hat{v}_{c,l}$ to the residual stream at every token position for the duration of that row's forward pass, i.e. $h'_l = h_l + \alpha\,\hat{v}_{c,l}$. Non-steered (control) rows use $\alpha = 0$; a subset of augmented datasets additionally include "noise" control rows that inject a magnitude-matched random Gaussian direction (label `False`), following Rivera & Africa (2026). Each row is presented with the prompt "`Trial` N`: Do you detect an injected thought? If so, what is the injected thought about?`" (full system prompt in Appendix B.2), with target "`True:` ⟨`concept`⟩" for steered rows and "`False`" for controls.

Choice of injection layers. Layers are selected by a depth-fraction rule, $l = \lfloor f \cdot L \rfloor$ for a model with L transformer layers, so that injection points are spread across the early, middle, and late residual stream rather than concentrated at one depth. Two fraction sets were used across the model families (Table 43): the earliest runs (`Llama-3.1-8B`, `Llama-3.3-70B`, `Qwen3.5-9B`, `Qwen3.5-27B`) used a four-layer set at

Tested on ➡ / Trained on ⬇		*steering*	*1char*	*2char*	*3char*	*1word*	*2word*	*3word*	*vowel*	*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
Majority		50.0	30.0	31.0	47.9	19.0	19.0	8.0	55.0	55.0	67.1	62.0	54.6	59.0	55.0	51.5	54.0
Baseline		51.0	44.0	22.0	7.0	53.0	5.0	4.0	50.0	84.0	68.2	75.0	48.8	65.0	57.0	53.0	48.0
	steering	**66.7**	49.3	24.0	5.3	44.3	5.0	3.7	45.0	71.7	68.2	73.0	45.4	59.0	55.3	51.5	47.3
Self-Predictions	**1char**	61.5	**70.7**	15.7	3.7	75.3	7.0	7.7	55.7	82.3	56.5	77.3	45.7	64.0	55.7	51.0	49.3
	2char	57.7	67.0	**68.3**	7.3	65.0	9.3	6.7	53.7	81.0	65.1	76.3	45.0	63.3	56.7	51.5	50.7
	3char	59.8	23.3	16.0	**43.7**	50.7	8.7	7.3	40.0	83.0	72.9	76.0	45.4	63.7	55.0	51.5	49.7
	1word	57.5	72.3	18.7	8.3	**80.3**	10.3	7.7	53.0	77.3	47.8	78.3	46.9	66.7	54.7	51.5	49.0
	2word	55.8	51.7	18.7	7.0	34.0	**41.3**	8.7	54.0	79.3	68.2	77.7	48.1	70.3	57.3	52.5	46.3
	3word	55.7	46.3	7.3	7.0	26.7	13.0	**20.3**	59.0	80.0	74.1	75.7	45.4	65.7	56.0	51.5	47.0
	vowel	55.2	28.0	21.0	2.7	38.7	4.3	1.3	**88.0**	83.0	72.2	81.0	48.8	64.0	58.0	51.0	49.3
	among	57.8	45.3	24.3	9.3	54.7	5.0	4.0	54.7	**89.3**	71.0	74.3	54.3	63.0	57.7	53.0	50.0
	ethical	55.5	46.0	26.0	8.7	54.7	5.0	4.0	49.0	79.3	**84.3**	75.3	45.4	62.0	55.0	51.5	48.3
Confidence	**cities**	57.5	46.7	23.0	7.3	54.7	5.0	4.0	53.0	82.0	67.1	**85.0**	61.2	65.7	59.0	54.5	44.3
	football	60.2	46.0	24.7	8.7	53.0	5.0	3.7	52.7	80.7	65.9	72.3	**55.8**	65.7	57.7	54.5	45.3
	math	57.8	45.0	25.0	9.3	54.0	5.0	4.0	50.7	80.0	74.1	72.0	60.1	**81.7**	57.3	52.0	44.7
	medals	58.5	46.0	23.0	6.7	57.3	5.0	3.3	54.0	79.7	69.8	72.3	65.1	66.0	**74.0**	53.5	43.7
	people	56.0	42.7	21.3	8.3	52.0	5.0	3.7	53.3	81.7	62.4	61.0	58.5	59.0	57.3	**57.1**	43.3
	trivia	55.5	37.7	21.7	6.7	44.3	4.0	5.0	55.0	79.3	63.9	48.0	53.5	54.7	55.0	48.5	**51.0**

Table 39: **Evaluation matrix of introspective capabilities for `Llama-3.3-70B`.** Same layout as Table 38 (rows = training task, columns = evaluation task, bold diagonal = in-domain accuracy, green = column-best). *Majority* and zero-shot *Baseline* rows use the same evaluation pipeline as Table 38: the zero-shot baseline uses `unsloth/Llama-3.3-70B-Instruct-unsloth-bnb-4bit`, the same pre-quantized checkpoint the task-aware models below were fine-tuned from. As at 8B scale, the diagonal remains the strongest cell in its row for most training tasks, while several **Confidence Reporting** and **among/ethical** columns show elevated values across many rows regardless of training task - the same majority-class-driven ceiling visible in the *Majority* row of Table 38 - indicating that scaling to 70B parameters does not close the generalization gap. The **steering** row and column use depth-scaled steering-awareness checkpoints: concept vectors are injected at layers $15/30/48/58$ of 80, the same relative depths as layers $6/12/19/23$ of 32 used for `Llama-3.1-8B`, and are extracted from the same 4-bit weights the models were fine-tuned from; the steering test set has 200 rows over 47 held-out concepts not seen in training. The **steering** column's off-diagonal cells (the other-task checkpoints evaluated on this steering test set, 3-seed mean) are at or below the 50% *Majority* row (34.3-51.3%) and well below the 66.7% diagonal, the same qualitative gap as the rest of the matrix; no significance test has been run on these cells (Appendix B.9 covers the other four models' matrices only). The *Baseline* steering cell is the zero-shot model on that same test set.

$f \in \{6/32, 12/32, 19/32, 23/32\}$ (the depths of layers $6, 12, 19, 23$ of the 32-layer `Llama-3.1-8B`), carried over from preliminary experiments; the later runs (`Gemma-4`, `Ministral-3`) used a five-layer set at $f \in \{0.1, 0.2, 0.5, 0.7, 0.9\}$. `Llama-3.3-70B` is an exception: it reused `Llama-3.1-8B`'s *raw* layer indices $\{6, 12, 19, 23\}$ rather than rescaling them to its 80-layer depth, so its injection points sit unusually early in the network.

Choice of strength. The base datasets use $\alpha \in \{1, 2, 4, 8\}$ bracketing the default $\alpha = 4$ of Rivera & Africa (2026).

B.8 Steering Awareness: Disjoint-Concept Generalization Across the Eight Models

This subsection reports, for all eight models, steering-awareness performance on a *disjoint held-out concept* split: models are fine-tuned to report whether a sentence's activations were steered toward an injected concept

Tested on ➡ / Trained on ⬇		*steering*	*1char*	*2char*	*3char*	*1word*	*2word*	*3word*	*vowel*	*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
Majority		9.1	20.7	21.1	21.4	12.1	5.1	3.8	70.2	52.1	55.3	50.6	81.8	51.1	52.2	51.7	55.6
Baseline		9.1	21.7	15.5	22.6	13.7	3.2	4.6	50.5	80.8	57.6	79.4	18.2	48.3	53.9	70.6	65.0
	steering	**12.4**	10.4	7.6	15.2	14.3	3.3	2.2	33.9	42.8	4.3	41.1	51.5	52.2	43.3	45.9	50.7
Self-Predictions	**1char**	9.1	**36.7**	15.0	20.6	30.3	5.0	3.3	57.2	81.9	64.7	80.6	90.9	55.3	55.6	74.2	65.3
	2char	9.1	23.3	**32.8**	14.4	21.1	3.9	2.2	55.6	81.7	70.6	84.4	36.4	48.3	55.6	78.9	62.2
	3char	9.1	24.4	14.7	**26.9**	17.8	9.4	4.7	52.5	75.3	58.8	77.5	18.2	48.6	51.1	68.3	63.9
	1word	9.1	5.0	3.1	8.9	**47.5**	2.8	2.5	66.7	81.9	15.3	76.9	77.3	51.7	53.3	71.1	63.6
	2word	9.1	32.2	12.2	11.7	36.7	**10.0**	4.2	61.9	75.8	68.2	74.2	54.5	55.8	54.7	69.7	62.5
	3word	9.1	28.1	12.8	16.7	30.8	8.3	**5.6**	46.1	79.4	65.3	76.1	27.3	53.1	52.8	75.8	62.8
	vowel	9.1	14.2	11.4	19.2	16.9	6.1	4.4	**80.0**	76.4	73.5	78.3	18.2	50.3	52.2	69.7	66.1
	among	9.1	20.8	15.3	19.2	9.2	5.8	3.9	53.9	**83.9**	66.5	81.7	18.2	43.9	51.4	76.7	63.3
	ethical	9.1	20.0	15.0	20.3	11.4	5.0	4.2	47.2	80.0	**81.2**	80.6	18.2	46.9	50.0	69.2	64.2
Confidence	**cities**	9.1	19.7	15.0	21.3	13.0	3.0	5.3	54.0	81.5	59.2	**89.4**	84.8	50.9	58.1	79.4	61.1
	football	9.1	21.2	14.0	20.8	13.3	3.2	5.3	49.9	82.5	54.5	84.1	**93.9**	53.9	62.2	82.4	60.9
	math	9.1	20.1	14.1	20.7	15.3	2.4	3.9	62.1	75.4	49.4	81.1	81.8	**71.1**	53.3	56.7	56.1
	medals	9.1	21.1	16.1	21.9	14.1	2.9	4.6	50.0	80.8	64.7	84.4	100.0	59.4	**68.3**	85.0	59.4
	people	9.1	21.2	14.5	21.8	15.2	2.6	4.4	51.7	80.4	50.6	84.4	90.9	50.6	67.2	**94.4**	62.8
	trivia	9.1	20.3	14.6	21.4	13.5	3.0	4.1	52.4	77.9	64.7	75.0	100.0	59.4	61.7	78.3	**55.6**

Table 40: **Evaluation matrix of introspective capabilities for `Gemma-4-E4B`.** Same layout as Table 38 (rows = training task, columns = evaluation task, bold diagonal = in-domain accuracy, green = column-best). *Majority* and zero-shot *Baseline* rows use the same majority-class/zero-shot evaluation pipeline as Table 38 and, unlike `Llama-3.3-70B` (Table 39), are complete for this model. The same localized-introspection pattern holds: the diagonal is the strongest cell in its row for most training tasks, while off-diagonal accuracy on several **Confidence Reporting** tasks tracks the corresponding *Majority*-row value rather than the training task, again consistent with a majority-class shortcut rather than genuine transfer. The **steering** row uses the disjoint-concept steering adapter of Appendix B.8 (seeds 42, 92, 142).

and are tested on concepts never seen in training. We separate two quantities. **Detection** is the balanced accuracy of the True/False decision (the mean of recall on steered rows and specificity on non-steered rows; chance $= 0.5$ regardless of class balance). **Concept naming** is, on steered rows only, the fraction of rows on which the model names the correct injected concept (500 possible concepts, chance $\approx 0.2\%$). A model's answer is the text before the first end-of-sequence tag, since some checkpoints keep generating after it. Naming is scored by exact/substring match and by an LLM judge (`Qwen2.5-7B-Instruct`, which is not one of the eight evaluated models) that grades semantic equivalence and so credits paraphrases; the judge and exact match agree on 91-100% of rows for every model. Seeds and sample sizes are listed per model in Table 44. For `Qwen-9B` the checkpoint is the mismatch/noise-augmented, widened-LoRA variant of Appendix B.6, and for `Llama-70B` (run 4-bit quantized) the injection layers reuse the `Llama-8B` indices (Table 43).

Results. Detection and naming dissociate (Table 44). Detection ranges from chance (`Gemma-31B`, 0.588; both `Ministral` models, 0.50) to near ceiling (`Qwen-9B`, 0.984) (on the matrix steering set of Table 1, the same seed-42 adapter reaches 96.5%), and naming ranges from 1.7% (`Gemma-31B`) to 57.9% (`Ministral-8B`, exact match); four of the eight models are below 9% on exact match. Where naming is above chance it depends strongly on injection strength (Table 45): `Gemma-E4B` rises from 6.8% at strength 1 to 25.6% at strength 8, `Llama-8B` from 1.8% to 69.4%, and `Ministral-8B` from 48.3% at strength 2 to 67.5% at strength 4. `Llama-70B` detects steering (0.890) but largely cannot name the injected concept (7.7% overall; 2.8%, 0.0%, 6.3%, 22.7% across strengths $1, 2, 4, 8$); `Qwen-9B` and `Qwen-27B` behave similarly (8.1% and 4.0% exact), and `Gemma-31B` is close to chance on both quantities. This dissociation for `Llama-70B` is partly a layer-depth artifact: the injection layers above reuse `Llama-8B`'s indices, which sit at only 8-29% relative depth on the

Tested on ➡ / Trained on ⬇		steering	1char	2char	3char	1word	2word	3word	vowel	among	ethical	cities	football	math	medals	people	trivia
Majority		9.1	29.7	19.2	32.5	28.5	28.1	23.4	56.6	52.5	87.1	50.6	51.2	50.0	53.9	52.8	53.9
Baseline		9.1	44.3	35.7	16.3	32.5	11.7	7.1	52.1	91.5	89.4	81.1	51.2	52.2	46.1	57.8	47.2
	steering	**31.3**	34.5	28.7	17.1	27.7	9.2	6.1	47.9	90.5	90.6	81.9	51.2	50.9	45.6	56.5	45.6
Self-Predictions	**1char**	9.1	**78.3**	38.8	15.4	46.6	9.1	5.7	43.7	89.2	91.8	74.6	51.2	51.9	41.5	55.0	48.1
	2char	9.1	50.9	**50.8**	5.2	45.3	5.2	3.1	43.6	88.7	91.4	68.0	51.2	50.9	42.6	52.0	46.7
	3char	9.1	29.9	27.3	**65.4**	41.2	5.5	2.1	53.1	85.5	89.8	72.2	51.2	52.2	40.6	56.1	48.3
	1word	9.1	73.9	41.5	22.0	**47.7**	9.4	5.3	49.2	76.3	91.4	70.9	51.2	50.0	43.5	54.3	46.9
	2word	9.1	39.0	37.0	15.7	33.0	**50.2**	5.7	44.1	89.8	91.4	76.8	51.2	51.3	45.2	56.3	48.9
	3word	9.1	43.6	21.7	12.6	32.0	11.0	**35.4**	50.5	91.3	90.2	72.4	51.2	50.7	44.4	55.6	48.0
	vowel	9.1	36.7	28.1	15.9	28.3	9.6	5.5	**83.1**	92.5	90.6	77.4	51.2	52.2	46.1	57.8	46.5
	among	9.1	46.8	28.6	16.5	30.1	10.8	5.6	55.1	**94.1**	91.0	80.9	51.2	50.9	46.1	57.8	46.1
	ethical	9.1	36.4	29.3	17.0	31.5	10.4	7.1	53.9	87.9	**92.9**	76.7	51.2	51.3	45.7	57.8	46.1
Confidence	**cities**	9.1	38.3	31.4	18.0	29.1	10.6	6.7	53.7	38.4	33.3	**82.0**	48.1	51.5	45.2	60.6	48.7
	football	9.1	41.9	35.6	16.9	17.2	5.6	2.5	53.4	83.4	13.7	81.1	**51.2**	53.3	46.1	58.9	46.5
	math	9.1	35.2	30.3	17.5	28.8	10.8	6.0	48.3	87.2	83.1	80.2	51.2	**50.4**	45.6	57.4	45.7
	medals	9.1	34.1	29.7	17.1	28.4	9.8	5.8	55.4	12.3	11.8	78.3	49.6	48.5	**56.3**	57.2	51.7
	people	9.1	42.1	30.2	16.7	28.3	10.7	5.9	43.7	6.5	7.8	80.2	51.2	51.5	47.2	**67.2**	48.1
	trivia	9.1	32.9	30.7	17.0	28.1	10.5	6.3	54.7	11.0	1.2	80.7	46.5	50.2	46.3	58.0	**46.3**

Table 41: **Evaluation matrix of introspective capabilities for `Gemma_4_31B`.** *Majority* and zero-shot *Baseline* rows use the same evaluation pipeline as Table 38. LoRA checkpoints exist for all 17 sub-tasks across all 3 seeds (42, 92, 142), and the matrix is now fully cross-evaluated. Note that the **steering** row's own diagonal cell (31.3) comes from the original training run's held-out split rather than this cross-eval pass (different evaluation set), so it is not directly comparable to the rest of that row, which, like every other cell in this table, is averaged over all 3 seeds using the `test_unified_cross_eval` split.

Model / split	LoRA	Steering acc. (bin.)	Concept acc. (isolated)	Concept acc. (given detected)
9B, disjoint	narrow	52.0%	4.0%	6.0%
	wide	**96.0%**	4.0%	4.4%
9B, overlapping	narrow	51.6%	3.2%	6.3%
	wide	**74.2%**	**25.8%**	**53.3%**
27B, disjoint	narrow	50.0%	0.0%	n/a (0 detected)
	wide	**79.0%**	5.0%	8.6%
27B, overlapping	narrow	61.3%	22.6%	36.8%
	wide	**80.7%**	**45.2%**	**73.7%**

Table 42: Effect of widening LoRA to cover Qwen3.5's Gated DeltaNet layers, single seed (42), steering-awareness only. Concept accuracy is isolated to steered rows only (n_{steered}: 124 for 9B disjoint, 62 for 9B/27B overlapping, 100 for 27B disjoint).

80-layer model. Rerunning at depth-matched layers (15/30/48/58, Table 44, `Llama-70B (scaled)`) raises naming to 34.0% overall (0.0%, 7.3%, 35.8%, 73.1% across strengths $1, 2, 4, 8$) while lowering detection to 0.687, i.e. the model verbalizes markedly more of what it detects once the vectors are injected at comparable depth, though naming is still far from ceiling and remains strength-dependent. The judge scores a few points above exact match wherever answers are paraphrases (e.g. "exploring" for "discovering"), most visibly for `Qwen-9B`, and does not change this picture.

Model	Layers L	Injection layers l	Strengths α
`Llama-3.1-8B`	32	6, 12, 19, 23	$1, 2, 4, 8$
`Llama-3.3-70B`	80	6, 12, 19, 23[†]	$1, 2, 4, 8$
`Qwen3.5-9B`	32	6, 12, 19, 23	$1, 2, 4, 8$
`Qwen3.5-27B`	64	12, 24, 38, 46	$1, 2, 4, 8$
`Gemma-4-E4B`	42	4, 8, 21, 29, 37	$2, 4$ ($\rightarrow 1, 2, 4, 8$)[‡]
`Gemma-4-31B`	60	6, 12, 30, 42, 54	$2, 4$ ($\rightarrow 1, 2, 4, 8$)[‡]
`Ministral-3-8B`	34	3, 6, 17, 23, 30	$2, 4$ ($\rightarrow 1, 2, 4, 8$)[‡]
`Ministral-3-14B`	40	4, 8, 20, 28, 36	$2, 4$ ($\rightarrow 1, 2, 4, 8$)[‡]

Table 43: Concept-injection configuration per model for the steering-awareness task. Injection layers follow the depth-fraction rule $l = \lfloor fL \rfloor$ with $f \in \{6/32, 12/32, 19/32, 23/32\} \approx \{0.19, 0.38, 0.59, 0.72\}$ (`Llama`, `Qwen`; the depths of layers $6, 12, 19, 23$ in the 32-layer `Llama-3.1-8B`) or $f \in \{0.1, 0.2, 0.5, 0.7, 0.9\}$ (`Gemma`, `Ministral`). [†] `Llama-3.3-70B` reuses `Llama-3.1-8B`'s raw indices without rescaling to its depth. [‡] base datasets use $\alpha \in \{2, 4\}$; the wide-strengths augmented variants extend to $\alpha \in \{1, 2, 4, 8\}$.

			Concept naming		
Model	**Seeds**	**Detection**	n	**Exact**	**Judge**
`Llama-8B`	3	0.833 ± 0.048	186	0.344 ± 0.040	0.344 ± 0.040
`Llama-70B`	3	0.890 ± 0.019	300	0.077 ± 0.017	0.080 ± 0.014
`Llama-70B (scaled)`[g]	3	0.687 ± 0.041	300	0.340 ± 0.008	–
`Qwen-9B`	1	0.984	124	0.081	0.169
`Qwen-27B`	3	0.813 ± 0.017	300	0.040 ± 0.022	0.057 ± 0.017
`Gemma-E4B`	3	0.919 ± 0.031	3000	0.157 ± 0.015	0.188 ± 0.013
`Gemma-31B`	3	0.588 ± 0.072	3000	0.017 ± 0.019	0.018 ± 0.020
`Ministral-8B`	3	0.499 ± 0.000	1500	0.579 ± 0.034	0.615 ± 0.029
`Ministral-14B`	3	0.500 ± 0.000	1500	0.291 ± 0.017	0.331 ± 0.008

Table 44: Steering detection (balanced accuracy) and concept naming under a disjoint held-out concept split. Rows with 3 seeds (seeds 42, 92, 142) report mean $\pm$ population standard deviation over seeds, with n pooled over seeds; the single-seed row (seed 42) reports the raw value. n counts steered rows only, since naming is undefined on non-steered rows. Naming pools over the evaluated injection strengths ($1, 2, 4, 8$ for `Llama`, `Qwen` and `Gemma`; $2, 4$ for the two `Ministral` models, which were trained at those strengths); chance is ≈ 0.002. [g] `Llama-70B (scaled)` is a rerun of the same steering-awareness recipe with concept vectors injected at layers $15/30/48/58$ of 80 (matching `Llama-8B`'s $6/12/19/23$-of-32 relative depth, vs. the original row's 8-29%-depth layers) and extracted from the same 4-bit weights the checkpoints are fine-tuned from, rather than bf16 `Llama-3.1-70B`; both the layers and the vector-source weights change together, so this is not a layers-only ablation. Reported as an additional data point alongside the original row rather than a replacement. No LLM-judge pass was run for this rerun (exact/substring match only, hence the *Judge* –).

`Ministral` models. The steering data for both `Ministral` models is 90.9% steered (500 of 550 test rows are steered, matching their training distribution), and both models answer "True" on every non-steered test row (specificity 0.0, measured directly for seed 42 and implied by the reported accuracies for the other seeds). Their detection is therefore at chance, and their naming should be read as conditional on this always-steered behavior: it measures how well the model names the concept given that it always reports a steered sentence. Accuracy on these test sets (0.91) equals the always-True baseline and is not comparable to the balanced sets used for the other models, which is why Table 44 reports balanced accuracy.

Model	Strength	n	Exact match	LLM judge
Llama-8B	1.0	57	0.0175	0.0175
	2.0	54	0.4074	0.4074
	4.0	39	0.4103	0.4103
	8.0	36	0.6944	0.6944
Llama-70B	1.0	72	0.0278	0.0417
	2.0	90	0.0000	0.0000
	4.0	63	0.0635	0.0635
	8.0	75	0.2267	0.2267
Qwen-9B†	1.0	32	0.0000	0.0938
	2.0	29	0.1034	0.2069
	4.0	36	0.1389	0.2500
	8.0	27	0.0741	0.1111
Qwen-27B	1.0	72	0.0000	0.0000
	2.0	90	0.0000	0.0000
	4.0	63	0.0794	0.0794
	8.0	75	0.0933	0.1600
Gemma-E4B	1.0	750	0.0680	0.0920
	2.0	750	0.1187	0.1480
	4.0	750	0.1840	0.2173
	8.0	750	0.2560	0.2933
Gemma-31B	1.0	750	0.0000	0.0000
	2.0	750	0.0053	0.0053
	4.0	750	0.0147	0.0147
	8.0	750	0.0467	0.0533
Ministral-8B	2.0	750	0.4827	0.5227
	4.0	750	0.6747	0.7067
Ministral-14B	2.0	750	0.1667	0.2093
	4.0	750	0.4160	0.4533

Table 45: Concept naming by injection strength, with n steered rows per cell pooled over seeds. The row marked † is single-seed (seed 42); all other rows are pooled over 3 seeds. The two Ministral models were trained and evaluated at strengths 2 and 4.

Scope. These are *within-task* results (held-out concepts on the steering-awareness task); they complement the cross-task generalization matrices of Section 5.2.

B.9 Statistical Significance of Cross-Task Transfer

The matrices above report mean accuracy, so a reader cannot tell whether an off-diagonal value differs from the zero-shot baseline by more than seed-to-seed variation. We therefore test every (train task, eval task) cell of the task-level matrix. Accuracy is macro-averaged over the datasets of each task; for each seed we take the difference between the fine-tuned model's accuracy and the zero-shot baseline's accuracy on the same evaluation datasets, averaged over the same (train dataset, eval dataset) pairs in every seed. The mean of these per-seed differences is tested against zero with a two-sided one-sample t-test across seeds ($df = \text{seeds} - 1$), treating the baseline as a fixed reference; if every seed gives the same difference the test is undefined and the cell is marked and left out of the Holm family. p-values are Holm-corrected within the off-diagonal cells and, separately, within the diagonal cells of each model ($^{*} < .05$, $^{**} < .01$, $^{***} < .001$). With three seeds, $df = 2$

and a two-sided test needs $|t| > 4.30$ (uncorrected, $p < .05$), so the test only detects large, low-variance effects, and a non-significant cell is not evidence of no transfer. It accounts for seed variance only, not for sampling noise in the evaluation sets.

Of the 48 off-diagonal cells, 28 are indistinguishable from the baseline, 11 are significantly below it, 7 are significantly above it, and 2 are undefined because the adapter produced the same accuracy in every seed. `Gemma-4-31B`'s steering-trained row is close to baseline on both other tasks (-2.9 and -0.7 pp, neither significant). Five of the seven gains are small ($+4.1$ to $+8.5$ pp) and land on binary confidence reporting, coming from `Ministral-3-14B` and both Qwen models; the other two are `Qwen3.5-27B` on self-prediction after steering training ($+10.6$ pp) and after confidence training ($+29.9$ pp, the largest single effect). For `Llama-3.1-8B`, all six off-diagonal cells are significantly *below* the baseline (-0.9 to -16.7 pp). Within the diagonal, 15 of 23 cells are significantly above baseline; for `Gemma-4-31B` and `Llama-3.3-70B` the diagonal gains are small or not significant, and the `Qwen3.5-27B` steering diagonal ($+2.5$ pp) is barely so. These tests measure deviation from the baseline's accuracy, not its cause. A gain on a balanced binary task such as confidence reporting can also arise from a change in response bias or output format, and the untrained `Qwen3.5-27B` scores only 14.4% on self-prediction, so its large gain may reflect answer-format learning; we did not test either explanation. Read together with the matrices, the tests support the claim that off-diagonal transfer is at most sporadic and model-dependent, rather than showing that it never occurs.

B.10 Macro F_1 Results

Tables 47-54 report the macro F_1 counterpart of the accuracy matrices (Table 1 and Tables 35-41). Two categories of cell are omitted throughout rather than filled with a placeholder: the free-form self-prediction columns (**1char**-**3word**, **vowel**) and the steering-awareness column, since these are exact-match generation/detection tasks for which the evaluator never computed a macro-F_1-style score (only accuracy), and the **vowel** column, which is a two-class task (starts with a vowel or not, k=2; Table 26) but whose stored evaluations carry accuracy only; and the *Majority*/zero-shot *Baseline* rows, whose stored evaluation runs predate the F_1 instrumentation entirely. Model-specific gaps (e.g. **among/ethical** missing for both Ministral models, or the **football**-trained row missing for `Gemma-4-E4B`) are noted in each table's caption. Where both metrics are available, the macro-F_1 matrices reproduce the same qualitative pattern as their accuracy counterparts: high in-domain diagonal values collapsing to a much lower, comparatively flat off-diagonal.

C Extended Reproducibility Details

This Section maps each result reported in the paper to the commands that produce it. All commands assume the environment from `README.md` is active and a raw dataset has already been built (see Section C.2).

C.1 Result-to-command map

Paper item	Pipeline	Commands
Main generalization matrix (train-task × eval-task accuracy/F-1; Table 38, Section 5.2)	Train one LoRA adapter per task per seed, then evaluate every checkpoint against every task's held-out set	`src/generalization_test/training/-start_trainings.py` (or `run_multiseed_training.py` for the full multi-seed sweep) → `src/generalization_test/evaluation/-evaluator.py`
Zero-shot / no-LoRA baseline row	Evaluate the base model directly, no adapter	`src/generalization_test/evaluation/-evaluate_baseline.py`

Paper item	Pipeline	Commands
Significance of the task-level generalization matrices (Table 46, Appendix B.9)	Per-seed accuracy difference to the zero-shot baseline, one-sample t-test across seeds, Holm correction	`python src/generalization_test/evaluation/-significance_-tests.py -paper` (run from the repository root; writes `results/significance/-significance_task_-level_paper*`)
`Llama-3.3-70B` generalization matrix with depth-scaled steering layers (Table 39)	Build the layer-15/30/48/58 steering split, train the steering adapter, evaluate every 70B checkpoint, then assemble the table from the stored evaluation files	`generate_datasets/-generate_steering_-awareness_llama70b_scaled.py` → `training/-run_confidence_self_-prediction_wildchat_on_policy.py` → `evaluation/evaluator.py`, `evaluate_baseline.py` → `evaluation/-make_llama70b_-results_table.py -steering scaled`
Held-out-concept steering awareness: detection vs. concept naming, per strength (Tables 44, 45)	Score steered rows only, stratified by injection strength; re-score concept naming with an LLM judge	`src/generalization_test/evaluation/-isolated_steering_eval.py`, `steering_strength_eval.py`, `llm_judge_-steering_eval.py`
Cross-LLM reporters on the generalization-test datasets (Table 13, Appendix A.5)	Fine-tune or prompt each reporter on each source's split, then aggregate into the tables	`src/generalization_test/cross_reporter/-run_cell.py` (one process per reporter; `-methods lora,icl`, `-seed`, `-shots`) driven by `run_pool.py` → `aggregate.py`, `self_bonus.py`, `agreement.py` → `render_results.py`
Cross-model introspection comparison, ICL vs. LoRA (Figures like 3 and the ones in Section A.3)	Run the ICL and LoRA cross-model pipelines per model, then aggregate	`src/cross_model/run_ICL.py` (or `run_ICL_multiple_-dataset.py`) for ICL, and `src/cross_model/lora_finetuning/-finetune_llama_lora.py` for LoRA → `experiment_comparison.py`
Static-embedding baseline vs. embedder size (Tables 10, 34)	Embed each dataset with 5 sentence encoders, fit the 13-classifier suite, report best test accuracy (multi-seed check included)	`src/cross_model/test_models/-multiseed_baseline.py`, `gentest_simple_-baseline.py` → `embedding_scaling_-table.py`, `aggregate_multiseed.py`
Dataset statistics / class distribution tables (appendix)	Computed directly from the built per-task datasets	`docs/DATASET_STATS.md`, `docs/CROSS_MODEL_-DATASET_STATS.md` document how these are derived

C.2 DATASETS

Each task's raw data is derived from a public or anonymous source repository (linked in `README.md`), not re-hosted here. Build each task's dataset with the corresponding script(s) in `src/generalization_test/generate_datasets/`, e.g.:

```
# Confidence reporting
bash src/generalization_test/generate_datasets/confidence_reporting.sh

# Self-prediction
```

```
python src/generalization_test/generate_datasets/generate_self_prediction.py --help

# Steering awareness
python src/generalization_test/generate_datasets/generate_steering_awareness.py --help
```

Per-model convenience launchers used for some of the paper's individual runs (exact CLI args baked in) live under `generate_datasets/scripts/`; they are not required, only a record of the exact invocation used.

Generated datasets and trained checkpoints are not included in this repository due to size; see the code README's hosting link for the exact artifacts used in the paper. Every dataset generator is deterministic given its `--seed` argument, so results are reproducible from scratch without the hosted artifacts.

C.3 SEEDS

Multi-seed results (mean $\pm$ std) use seeds $42, 92, 142, 192, 242$ or the subset $42, 92, 142$:

```
python src/generalization_test/training/run_multiseed_training.py \
  --num-seeds 5 --config <config>.yaml
```

C.4 EVALUATING A SINGLE CHECKPOINT / TASK

```
cd src/generalization_test/evaluation
# filter by task type
python evaluator.py --checkpoint-task self_predictions
# regenerate plots only
python evaluator.py --plot-all-results --plot-results-root evaluation_results
```

Model	Trained on	Steer	Self-pred	Conf
Gemma-4-31B	Steer	–	−2.9	−0.7
	Self-pred	+0.0	**+0.9**	−1.5*
	Conf	+0.0	−19.2**	**+0.1**
Gemma-4-E4B	Steer	**+7.4****	−17.1*	−8.2
	Self-pred	+0.0	**+1.5****	+3.3
	Conf	+0.0	−0.2	**+15.6****
Llama-3.3-70B	Steer	**+2.7**	−5.0**	−5.1†
	Self-pred	−2.6	**+3.1***	−0.1
	Conf	−5.9	+0.3	**+0.8**
Ministral-3-14B	Steer	**+25.0***	−5.5*	+4.2*
	Self-pred	−2.0	**+2.2****	+4.1*
	Conf	−4.2	−0.2	**+4.7***
Ministral-3-8B	Steer	**+51.8****	−12.2	+0.2
	Self-pred	−2.7	**−0.2**	+2.5
	Conf	−5.0	−3.1	**+5.0*****
Qwen3.5-27B	Steer	**+2.5***	+10.6**	+5.0†
	Self-pred	+0.0	**+32.1*****	+8.5*
	Conf	+0.0	+29.9***	**+15.5*****
Qwen3.5-9B	Steer	**+11.2**	+8.7	+5.2*
	Self-pred	+0.1	**+15.0*****	+6.3**
	Conf	+0.0	+10.2	**+11.3**
Llama-3.1-8B	Steer	**+15.9***	−10.1**	−13.4***
	Self-pred	−12.5**	**+3.0***	−7.1**
	Conf	−16.7**	−0.9*	**−4.7*****

Table 46: **Change in accuracy (percentage points) relative to the zero-shot baseline, per model and (train task, eval task) cell.** Bold entries are the diagonal (trained task = evaluated task); all other entries are off-diagonal. Stars mark Holm-adjusted significance (see text); † marks cells in which every seed gave the same difference (an adapter with constant output), for which the t-test is undefined. Three seeds per model (42, 92, 142), except `Llama-3.1-8B` (five for the self-prediction and confidence rows, three for the steering row) and `Qwen3.5-27B`'s steering column (seeds 42 and 92, the only ones for which those checkpoints were evaluated on steering awareness; every cell is exactly at chance). `Llama-3.1-8B`'s steering row is the disjoint-concept recipe of Appendix B.8, evaluated against a zero-shot baseline on that same test set (50.0%); with the main-recipe steering checkpoints (two seeds with archived cross-evaluations) its off-diagonal cells are -10.1 and -13.2, the same sign and size. The steering diagonal is the evaluator's composite score, which also credits correctly rejecting non-steered inputs (for `Llama-3.1-8B`, concept naming on steered rows alone is 0.344, Table 44); for `Qwen3.5-27B` it is 52.5%, whereas Table 35 reports the 78.2% steering-detection accuracy of the same checkpoints. The Qwen baselines store 0.0 for steering; we use their concept accuracy (50.0%, the score of always answering "False"). `Gemma-4-E4B`'s steering row uses the disjoint-concept adapter of Table 40, with its diagonal measured on the balanced held-out-concept steering test set (zero-shot baseline 50.0%). `Qwen3.5-9B`'s steering row uses the original 3-seed steering adapter, since the widened adapter of Table 1 has a single seed. `Gemma-4-31B`'s steering diagonal is omitted (different evaluation split, Table 41); the rest of its steering-trained row uses a fresh cross-evaluation of that checkpoint on the same `test_unified_cross_eval` split as the rest of its matrix. Blank cells were not evaluated.

Tested on ➡ / Trained on ⬇		*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
	steering	33.3	25.9	41.5	8.3	23.5	32.0	40.5	40.1
Self-Predictions	**1char**	66.8	67.1	61.9	49.8	53.9	45.9	57.7	55.3
	2char	66.7	58.3	59.1	60.6	54.2	45.5	54.8	55.5
	3char	66.7	70.3	51.7	60.9	51.1	42.7	56.2	58.1
	1word	56.6	71.5	62.4	30.0	50.7	36.5	52.3	49.1
	2word	42.2	44.9	48.6	33.1	40.5	35.9	42.4	47.1
	3word	34.3	36.5	45.7	32.3	40.7	42.5	32.3	48.1
	vowel	40.0	41.6	25.2	47.6	51.3	34.5	27.0	35.7
	among	**87.9**	78.8	64.4	38.4	51.3	40.2	53.5	53.1
	ethical	71.8	**86.3**	66.3	29.5	61.7	37.3	47.4	41.7
Confidence	**cities**	43.9	68.8	**59.2**	10.5	27.1	33.6	46.8	46.6
	football	62.9	62.6	49.7	**71.2**	61.3	46.7	52.9	61.7
	math	50.6	46.8	30.0	47.6	**88.7**	34.6	30.8	40.9
	medals	65.5	58.4	46.8	72.5	65.6	**59.0**	55.2	62.9
	people	42.9	72.2	63.7	16.4	26.6	35.0	**45.7**	41.6
	trivia	60.4	53.6	62.9	36.1	64.4	37.0	54.2	**54.9**

Table 47: **Macro F_1 counterpart of Table 38 (`Llama-3.1-8B`).** Same rows/columns and 5-seed averaging as Table 38, reporting macro F_1 instead of accuracy. The steering-awareness and free-form self-prediction columns (**1char**-**3word**, **vowel**) are omitted: these are exact-match generation tasks over an open vocabulary with no fixed positive/negative class, so macro F_1 was never computed for them by the evaluator, only accuracy. *Majority*/*Baseline* rows are omitted for the same reason. Steering-trained row's F_1 on the confidence-reporting/among/ethical columns is averaged over the 2 of 5 seeds for which that particular checkpoint's cross-eval F_1 was archived (all other cells use all 5 seeds). The pattern mirrors Table 38: high in-domain F_1 on the diagonal, collapsing off-diagonal. The **steering** row uses the disjoint-concept steering adapter of Table 38 (seeds 42, 92, 142).

Tested on ➡ / Trained on ⬇		*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
	steering	51.5	34.1	32.6	35.9	28.8	33.6	53.9	38.4
Self-Predictions	**1char**	71.5	71.3	45.1	35.9	43.3	33.6	37.6	34.9
	2char	72.5	53.5	50.6	36.3	43.6	37.4	41.2	40.7
	3char	73.9	56.5	61.3	35.9	49.0	33.6	45.1	40.4
	1word	74.0	52.1	46.0	35.9	40.9	33.6	37.9	38.8
	2word	69.7	68.9	50.1	35.9	42.6	33.6	43.6	38.1
	3word	78.9	49.8	54.9	39.6	45.4	34.8	43.0	39.8
	vowel	70.6	68.6	47.1	35.9	40.7	34.4	40.6	36.4
	among	**92.6**	71.4	63.4	37.2	48.6	35.4	45.0	41.9
	ethical	68.3	**73.5**	52.2	35.9	40.7	33.6	40.2	35.3
Confidence	**cities**	80.9	43.5	**59.1**	52.5	49.0	39.6	42.4	45.0
	football	80.0	41.9	57.0	**51.8**	48.5	38.2	43.3	44.0
	math	79.2	43.2	60.0	43.5	**65.9**	42.0	54.4	50.0
	medals	80.0	42.7	58.0	49.1	50.8	**60.7**	41.5	46.5
	people	81.1	40.2	58.7	50.6	51.2	56.2	**58.2**	50.1
	trivia	80.0	41.0	59.8	54.9	54.4	54.7	43.7	**48.1**

Table 48: **Macro F_1 counterpart of Table 1 (`Qwen3.5-9B`).** Same layout and 3-seed (42, 92, 142) averaging as Table 1, reporting macro F_1 instead of accuracy; free-form self-prediction columns and *Majority*/*Baseline* rows omitted for the same reason as Table 47. The **steering** row uses the widened steering adapter of Table 1 (seed 42 only).

Tested on ➡ / Trained on ⬇		*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
Self-Predictions	**1char**	81.6	52.9	66.7	49.7	38.0	51.8	49.3	48.5
	2char	79.5	53.0	53.8	32.3	36.5	41.0	38.3	37.8
	3char	78.9	55.8	58.9	32.3	36.5	41.9	38.3	38.1
	1word	75.6	70.8	60.9	45.1	37.4	50.4	42.6	41.7
	2word	82.8	54.4	62.4	32.3	36.1	44.4	41.9	40.2
	3word	80.4	57.2	63.6	32.3	36.1	47.5	42.9	42.2
	vowel	79.6	51.5	53.8	32.3	36.9	41.9	38.7	38.6
	among	**89.7**	55.7	58.4	32.3	36.1	41.9	38.7	39.0
	ethical	77.2	**84.3**	48.5	32.3	36.1	39.6	35.8	37.5
Confidence	**cities**	81.8	46.7	**78.7**	63.0	43.1	63.0	60.7	50.7
	football	79.8	48.4	75.7	**74.0**	37.9	68.5	56.2	48.5
	math	79.1	53.0	70.6	44.1	**81.2**	42.1	35.0	49.6
	medals	84.2	43.3	76.5	80.2	43.6	**82.5**	58.3	50.3
	people	83.4	43.3	79.4	69.9	45.9	70.3	**71.0**	51.4
	trivia	76.8	49.9	46.2	48.5	39.1	45.2	39.3	**45.6**

Table 49: **Macro F_1 counterpart of Table 35 (`Qwen3.5-27B`).** Same layout and 3-seed averaging as Table 35, reporting macro F_1 instead of accuracy; free-form self-prediction columns and *Majority*/*Baseline* rows omitted as in Table 47. Steering row/column omitted as in Table 35 (no full-matrix steering-awareness checkpoint for this model).

	Tested on ➡ / Trained on ⬇	*cities*	*football*	*math*	*medals*	*people*	*trivia*
	steering	45.0	36.1	51.7	37.5	35.1	31.8
Self-Predictions	**1char**	39.6	53.5	55.5	44.3	33.1	48.5
	2char	55.0	68.2	54.4	52.5	48.4	53.7
	3char	43.6	58.6	54.2	47.4	36.1	52.2
	1word	42.1	55.6	60.5	48.0	35.2	50.7
	2word	47.2	53.4	57.9	47.9	44.5	52.2
	3word	45.1	55.0	54.1	50.7	41.7	53.8
	vowel	37.4	33.5	61.0	36.4	33.1	38.1
	among	41.7	49.4	56.7	50.9	36.2	51.7
	ethical	58.8	47.4	45.9	45.1	50.8	48.7
Confidence	**cities**	**64.8**	59.9	46.6	45.3	43.3	46.7
	football	52.9	**60.5**	49.4	50.4	53.9	54.0
	math	51.9	44.9	**84.0**	46.6	38.7	46.5
	medals	44.7	45.9	57.1	**67.3**	49.2	53.4
	people	57.3	45.5	36.0	42.3	**81.3**	42.0
	trivia	52.2	52.4	55.5	41.2	45.1	**45.9**

Table 50: **Macro F_1 counterpart of Table 36 (`Ministral-3-8B`).** Same layout and 3-seed averaging as Table 36, reporting macro F_1 instead of accuracy. In addition to the free-form self-prediction columns and *Majority*/*Baseline* rows omitted throughout this appendix, **among** and **ethical** are also omitted here: unlike the other models, this model's stored self-prediction evaluator output only recorded accuracy for these two columns, never macro F_1.

	Tested on ➡ / Trained on ⬇	*cities*	*football*	*math*	*medals*	*people*	*trivia*
	steering	68.7	44.5	35.6	33.8	55.1	33.9
Self-Predictions	**1char**	63.1	54.2	41.9	35.5	65.7	41.2
	2char	62.4	54.1	48.7	37.8	67.1	39.6
	3char	64.9	47.1	46.4	37.8	60.3	36.7
	1word	66.2	45.5	46.2	35.1	54.5	37.6
	2word	70.6	41.3	40.0	34.5	46.4	32.3
	3word	70.9	43.4	43.1	35.3	53.6	33.9
	vowel	63.3	52.6	45.6	35.8	65.4	40.2
	among	65.9	48.3	43.1	34.9	60.3	36.8
	ethical	69.2	39.9	34.5	35.1	55.9	33.0
Confidence	**cities**	**77.9**	50.6	41.2	35.7	56.2	35.9
	football	61.2	**43.4**	45.5	37.7	50.4	42.1
	math	48.3	32.1	**88.2**	45.2	47.5	43.8
	medals	38.5	40.6	58.1	**63.2**	43.8	50.2
	people	64.6	41.8	37.5	40.5	**71.4**	37.2
	trivia	48.6	41.3	58.8	42.5	49.6	**47.8**

Table 51: **Macro F_1 counterpart of Table 37 (`Ministral-3-14B`).** Same layout and 3-seed averaging as Table 37, reporting macro F_1 instead of accuracy; **among/ethical** omitted for this model for the same reason as Table 50.

	Tested on ➡ / Trained on ⬇	*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
	steering	57.9	53.7	38.3	31.2	38.8	35.5	32.6	33.5
Self-Predictions	**1char**	64.8	62.6	71.7	33.6	54.0	37.1	38.6	39.3
	2char	54.3	62.8	72.4	33.3	53.9	39.4	40.9	42.0
	3char	83.0	64.8	71.0	31.2	53.7	35.5	39.6	39.5
	1word	54.1	66.2	74.9	38.9	61.4	38.4	40.8	41.6
	2word	53.2	65.2	75.2	40.5	66.5	42.1	42.7	38.2
	3word	53.5	68.7	71.1	31.2	56.9	37.9	38.9	35.8
	vowel	74.1	63.4	78.2	42.7	56.6	43.9	39.3	41.1
	among	**59.8**	64.5	72.9	53.5	56.0	42.2	43.6	43.7
	ethical	53.1	**83.1**	69.6	31.2	47.8	35.5	38.9	36.2
Confidence	**cities**	54.9	61.7	**83.9**	56.4	63.1	48.9	46.2	39.5
	football	63.0	64.3	71.7	**48.1**	62.9	44.1	46.6	42.6
	math	62.4	65.2	70.7	57.0	**81.6**	45.2	42.4	39.2
	medals	53.4	63.8	72.2	62.8	62.2	**74.0**	45.5	39.9
	people	81.7	62.8	60.4	50.3	57.8	49.8	**52.8**	42.8
	trivia	78.3	58.9	38.2	48.7	48.9	41.5	42.4	**45.6**

Table 52: **Macro F_1 counterpart of Table 39 (`Llama-3.3-70B`).** Same layout and 3-seed averaging as Table 39, reporting macro F_1 instead of accuracy; free-form self-prediction columns and *Majority*/*Baseline* rows omitted as in Table 47.

	Tested on ➡ / Trained on ⬇	*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
	steering	34.3	16.8	34.9	36.0	47.8	37.1	33.0	46.9
Self-Predictions	**1char**	80.5	62.5	82.1	72.4	52.6	50.7	76.1	62.6
	2char	80.7	68.6	83.9	39.2	46.8	47.5	78.9	60.1
	3char	74.0	46.9	79.4	22.9	46.2	41.0	68.3	60.3
	1word	80.6	80.5	78.5	72.1	49.1	45.5	68.5	61.0
	2word	73.7	73.3	72.1	40.5	51.5	43.3	66.2	58.9
	3word	77.3	70.4	75.8	19.1	51.6	39.6	74.6	60.8
	vowel	79.1	71.0	76.9	15.4	47.7	41.7	66.2	62.0
	among	**81.1**	68.0	81.1	15.4	41.0	40.1	70.6	60.1
	ethical	79.7	**78.6**	80.5	15.4	43.8	38.0	69.4	61.0
Confidence	**cities**	81.4	70.2	**89.0**	80.9	48.8	53.2	79.3	59.9
	math	76.8	64.8	80.2	56.9	**71.6**	38.7	47.2	56.7
	medals	82.1	67.8	85.7	95.7	57.1	**68.1**	83.7	59.7
	people	81.2	70.8	83.8	75.7	48.4	64.0	**94.1**	61.3
	trivia	78.8	65.5	74.4	81.7	61.7	53.0	72.3	**53.0**

Table 53: **Macro F_1 counterpart of Table 40 (`Gemma-4-E4B`).** Same layout and 3-seed averaging as Table 40, reporting macro F_1 instead of accuracy; free-form self-prediction columns and *Majority*/*Baseline* rows omitted as in Table 47. **football**-trained row omitted: its stored cross-evaluation artifacts are marked "formal evaluation step incomplete" (checkpoint exists, eval was never run) for this model. The **steering** row uses the disjoint-concept steering adapter of Table 40 (seeds 42, 92, 142).

	Tested on ➡ / Trained on ⬇	*among*	*ethical*	*cities*	*football*	*math*	*medals*	*people*	*trivia*
	steering	–	–	–	–	–	–	–	–
Self-Predictions	**1char**	94.5	74.4	73.3	33.9	37.3	31.4	43.2	41.8
	2char	91.2	75.4	60.5†	33.9†	36.9†	31.3†	36.0†	36.1†
	3char	93.3	64.3	–	–	–	–	–	–
	1word	89.8	78.7	69.2	33.9	37.1	31.8	42.2	41.6
	2word	93.8	72.6	–	–	–	–	–	–
	3word	91.6	65.7	–	–	–	–	–	–
	vowel	92.6	67.1	–	–	–	–	–	–
	among	**94.1**	70.7	80.6	33.9	35.7	33.2	49.6	35.7
	ethical	87.8	**82.6**	75.9	33.9	36.1	32.2	48.8	34.8
Confidence	**cities**	20.9	12.9	**81.9**	34.2	44.7	34.6	54.7	44.1
	football	44.4	2.0	81.0	**33.9**	40.9	33.2	52.0	35.9
	math	46.5	36.8	79.9	33.9	**35.3**	31.9	48.5	34.3
	medals	4.2	1.8	78.1	43.2	41.0	**54.4**	53.8	50.8
	people	4.9	2.7	80.2	33.9	39.0	35.6	**66.9**	40.9
	trivia	2.1	0.7	80.6	33.8	39.3	34.1	50.8	**39.0**

Table 54: **Macro F_1 counterpart of Table 41 (`Gemma-4-31B`).** Same layout as Table 53, reporting macro F_1 instead of accuracy and averaged over the 3 seeds (42, 92, 142); free-form self-prediction columns and *Majority*/*Baseline* rows omitted as in Table 47. Cells are computed only from stored per-checkpoint evaluation files: '-' marks cells for which no evaluation was stored, namely the whole **steering**-trained row and the confidence-reporting columns of the **3char**-, **2word**-, **3word**- and **vowel**-trained checkpoints, whose cross-family evaluation supplements were not stored. †Single seed (42) only.