WHAT DO SELF-IMPROVEMENT METHODS OPTIMIZE FOR? AN EMPIRICAL-MECHANISTIC STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-improvement methods such as debate, simple bootstrap, Gibbs sampling, and internal coherence maximization (ICM) can improve language model accuracy without external supervision, yet the signal they optimize remains unclear. What drives the gains, and when? We test the hypothesis that these methods improve accuracy through *coherence optimization* (Qiu et al., 2026), finding behaviors that are jointly predictable under a model's prior. For debate, bootstrap, and Gibbs sampling, we find that, overall, coherence and training-pool accuracy increase in sync, and rejecting update steps that increase coherence also lowers accuracy, suggesting a causal relation. For the remaining method, ICM, where ablation is inapplicable, Qiu et al. (2026) proved its analytical equivalence to coherence optimization under an information saturation condition, which we experimentally verified. Our results suggest that coherence optimization provides a common objective for understanding and designing feedback-free self-improvement methods.

1 INTRODUCTION

Language models can improve their accuracy without external supervision. Multiagent debate (Du et al., 2023), internal coherence maximization (ICM; Wen et al., 2025), and bootstrap (Lee, 2013; Canavan et al., 2026b;a) have each demonstrated substantial gains on reasoning and factual benchmarks. ICM nearly matches the performance of training on golden labels on GSM8K and TruthfulQA (Wen et al., 2025), debate improves GSM8K accuracy by 8 percentage points (Du et al., 2023), and bootstrap recovers 83–100% of the gap between zero- and many-shot golden-label prompting (Canavan et al., 2026b;a). Debate was proposed as a mechanism for weak judges to supervise stronger models without access to ground truth (Irving et al., 2018) with empirical demonstrations (Khan et al., 2024). Yet despite these successes, the signal these methods optimize remains unclear. What drives the improvement, and when should we expect it to succeed or fail?

Qiu et al. (2026) propose a theoretical answer, arguing that these methods are all instances of *coherence optimization*. Coherence measures how well a model's responses across contexts "fit together" according to its own prior: a high-coherence set of responses is one where each response is predictable given the others. The theory shows that debate, bootstrap, and ICM are special cases of this objective (Qiu et al., 2026, Propositions 4.3–4.4 and Appendix C.3), that coherence regularization is optimal for semi-supervised learning (Qiu et al., 2026, Theorem 5.5), and that Gibbs sampling provides an efficient algorithm for direct coherence optimization (Qiu et al., 2026, Theorem 4.2). These theoretical results, however, rest on preliminary illustrative experiments without measuring coherence across methods or testing whether it is causally responsible for the observed gains. Therefore, the theoretical core, which makes a causal claim of coherence leading to accuracy, has yet to be experimentally tested.

To fill this gap, we implement Gibbs sampling, cooperative debate, and bootstrap, tracking predictability alongside accuracy over rounds, and ablate by suppressing coherence-increasing update steps to test for causality. Experiments run on Llama-3.2-1B-Instruct and Llama-3.1-8B-Base/-Instruct, using the GSM8K dataset (Cobbe et al., 2021).

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.

For ICM, the remaining method where ablation is inapplicable,[1] we distinguish its objective defined by leave-one-out coherence from the theoretically analyzed autoregressive coherence of Qiu et al. (2026), restate the saturation condition connecting them, and experimentally verify the condition.

To summarize, our contributions are as follows:

1. **Empirical unification of self-improvement methods.** We show that coherence tracks accuracy improvements across Gibbs sampling, bootstrap, and cooperative debate, consistent with these methods optimizing a common objective. For a fourth method, internal coherence maximization (ICM), our results support the prediction of Qiu et al. (2026) that its objective approximates coherence.
2. **Gibbs sampling as direct optimizer.** We evaluate Gibbs sampling for direct coherence optimization. Our implementation improves held-out GSM8K accuracy without gold-label supervision and outperforms the evaluated gold-label 8-shot and 25-shot prompting and majority-vote baselines.
3. **Intervention for causal identification.** We propose rejection filtering as a method for pinpointing the mediating factor through which self-improvement methods lead to accuracy increase. Rejection filtering targeting coherence reduces accuracy by 2.9 pp in bootstrap and 7.1 pp in debate, suggesting a causal link between coherence and accuracy. We also report the limits of isolating coherence from other selection effects.

2 Related Work

This section outlines the core related works. Appendix G contains a comprehensive discussion.

Oversight and self-improvement. Debate aims to help weaker judges evaluate stronger models (Irving et al., 2018; Khan et al., 2024), within the broader study of scalable oversight and weak-to-strong generalization (Christiano et al., 2018; Bowman et al., 2022; Burns et al., 2023). Cooperative debate (Du et al., 2023), bootstrap (Canavan et al., 2026b;a), self-consistency (Wang et al., 2022), and Self-Refine (Madaan et al., 2023) use model-generated responses to improve predictions. Pseudo-labeling (Lee, 2013) and STaR (Zelikman et al., 2022) are related training approaches; STaR additionally uses answer correctness to filter rationales. Our focus is the signal accompanying iteration, measured across three methods and manipulated through rejection.

Elicitation and objectives. Contrast consistent search uses consistency constraints on internal representations (Burns et al., 2022); ICM optimizes mutual predictability of output labels (Wen et al., 2025). We distinguish the latter from autoregressive coherence and restate their conditional relationship. Sharpening (Huang et al., 2024) favors responses with high sequence likelihood for each prompt, whereas coherence optimization (Qiu et al., 2026) concerns predictability across a response set. Our experiments test the coherence account's empirical scope, including cases where predictable responses become less accurate.

3 Background

We summarize the coherence optimization framework of Qiu et al. (2026), focusing on the definitions and results needed to follow our experiments.

3.1 Coherence and D-Policies

Consider a set of *contexts* $\mathcal{S} = \{s_1, \ldots, s_{|\mathcal{S}|}\}$, where each context $s \in \mathcal{S}$ represents the set of possible responses to a given prompt (e.g., all possible answers to a math question). A *deterministic policy* (d-policy) π is a complete assignment of one response to each context, so that $\pi(s) \in s$ for all $s \in \mathcal{S}$. Consider a model with *trained prior* σ mapping every pair (π, s) of d-policy π and context s to a

[1] Ablation through suppression of coherence-increasing update steps is inapplicable on ICM, as ICM is itself an annealing algorithm over coherence, in direct conflict with the ablation.

Algorithm 1 Training-Friendly Gibbs Sampling (Qiu et al., 2026)

Input: Contexts $\mathcal{S}$, initial d-policy π_0, rounds N, context fraction γ, inverse temperature β
for $t = 0, 1, \ldots, N-1$ **do**
  Sample subset $\hat{\mathcal{S}}_t \subset \mathcal{S}$ uniformly at random, of size $\lfloor \gamma |\mathcal{S}| \rfloor$
  Form context $\phi_t = \sum_{s \in \hat{\mathcal{S}}_t} \pi_t(s)$ ▷ Conditioning context
  for each $s_n \notin \hat{\mathcal{S}}_t$ **do**
    Sample $a_t^n \sim \sigma^\beta(\phi_t, s_n)$ ▷ Resample response
  end for
  Update: $\pi_{t+1}(s_n) = a_t^n$ if $s_n \notin \hat{\mathcal{S}}_t$, else $\pi_{t+1}(s_n) = \pi_t(s_n)$
end for

distribution over s, representing the output distribution of the model given a prompt and a set π of labeled examples (e.g., few-shot or finetuning samples).

The *coherence* of a d-policy measures how well its responses fit together as per a model's prior:

$$\chi(\pi) \coloneqq \sum_{n=1}^{|\mathcal{S}|} \log_2 \sigma\left(\sum_{m=1}^{n-1} \pi(s_m),\ s_n\right)\big(\pi(s_n)\big) \leq 0. \tag{1}$$

Each term is the log-probability that the model assigns to response $\pi(s_n)$ when conditioned on all preceding responses in some ordering of contexts. Under the theory's chain-rule assumption, the total is independent of ordering (Qiu et al., 2026). Intuitively, coherence is high (close to zero) when each response is predictable given the others, and negated coherence equals the description length (in bits) of the d-policy under the model's prior, up to rounding (Qiu et al., 2026, Section 5.1).

3.2 Gibbs Sampling for Coherence Optimization

Direct optimization of coherence is intractable because it requires searching over exponentially many d-policies. Qiu et al. (2026, Theorem 4.2) show that Gibbs sampling provides an efficient solution that converges to the softmax distribution over coherence, $\mathrm{X}^\beta(\pi) \propto 2^{\beta\chi(\pi)}$, which places higher probability on more coherent d-policies.

In practice, the single-context leave-one-out updates of the original algorithm are impractical for large language models. We instead use the *training-friendly variant* (Qiu et al., 2026, Algorithm 2), which operates in rounds. At each round t, a random γ-fraction of contexts is held fixed as conditioning context $\hat{\mathcal{S}}_t$, and all remaining responses are resampled in parallel:

The parameter γ controls how much context is available for resampling. Smaller γ resamples more responses per round (stronger updates), while larger γ provides more conditioning context (more stable but slower convergence).

Qiu et al. (2026) argue theoretically that several independently developed self-improvement methods can be understood as special cases or approximations of coherence optimization: debate maps to Gibbs sampling with $|\mathcal{S}|$=2 (Qiu et al., 2026, Proposition 4.3), bootstrap approximates coherence with $D_{\mathrm{TV}} = O(|\beta - 1|)$ (Qiu et al., 2026, Proposition 4.4), and ICM optimizes a bidirectional variant (mutual predictability, conditioning on all other responses $m \neq n$ rather than preceding ones $m < n$).

4 Experimental Setup

In this section, we describe our experiment process for empirically testing the theoretical predictions and algorithmic proposal in Qiu et al. (2026).

4.1 Methods and Implementations

We implement three self-improvement methods. Gibbs sampling is tested with both an in-context learning (ICL) and a supervised fine-tuning (SFT) learning system; the remaining methods use ICL. The ICL experiments measure coherence via leave-one-out log-probabilities, with method-specific scoring prompts (Appendix A).

Gibbs sampling. We implement the training-friendly variant (Algorithm 1) with two sampler designs. The *vanilla* sampler holds a γ-fraction of contexts fixed as few-shot examples and resamples all remaining responses in parallel. The *preservation* sampler modifies context construction by mixing previous-round predictions with freshly generated zero-shot responses, which we observe empirically to reduce drift from the pretraining distribution (details in Appendix F). Under the ICL learning system, each response is generated by prompting the model with the conditioning context as few-shot question-answer pairs. Under the SFT learning system, the model is fine-tuned on the conditioning chunk each round, then generates responses for the complement without in-context examples. The training-friendly Gibbs variant (Algorithm 1) was introduced to make Gibbs sampling tractable for finetuning learning systems (Qiu et al., 2026), but applies equally to ICL.

Cooperative debate. We implement N=2-agent cooperative debate following Du et al. (2023). In each round, both agents generate revised solutions conditioned on the other agent's current solution, updating in parallel. This maps directly to Gibbs sampling over a d-policy with $|\mathcal{S}|$=2 (Qiu et al., 2026, Proposition 4.3). Debates run for a fixed number of rounds R with no early stopping.

Bootstrap. We implement the iterative few-shot prompting method of Canavan et al. (2026b;a). Round 0 produces zero-shot predictions. In each subsequent round, each question is answered using the previous round's responses to all other questions as demonstrations. The previous round is held fixed until all new responses are generated. Unlike the preservation Gibbs sampler, the main bootstrap trajectory experiments do not mix in fresh zero-shot responses between rounds.

ICM and informational saturation. ICM optimizes mutual predictability, which conditions on all other responses. Write $a_n = \pi(s_n)$, $P_n = \sum_{m<n} a_m$, and $F_n = \sum_{m>n} a_m$. The objectives are

$$\chi(\pi) = \sum_n \log_2 \sigma(P_n, s_n)(a_n), \tag{2}$$

$$f_{\mathrm{MP}}(\pi) = \sum_n \log_2 \sigma(P_n + F_n, s_n)(a_n), \tag{3}$$

Informational saturation means that adding the remaining responses F_n changes predictions little once the preceding context P_n is informative enough, implying $\chi \approx f_{\mathrm{MP}}$ (Qiu et al., 2026, Appendix C.3). We test the saturation prediction by scoring fixed label assignments under both objectives. ICM also uses a logical-consistency penalty and a different search procedure (Wen et al., 2025).

4.2 Benchmarks, Measurement, and Evaluation

The main Gibbs, bootstrap, and debate experiments use GSM8K (Cobbe et al., 2021) and Llama-3.2-1B-Instruct. Additional Gibbs experiments use Llama-3.1-8B-Base and Instruct models on GSM8K. Table 3 distinguishes the settings and replicate counts; SFT results without coherence scoring test accuracy gains only.

For the Gibbs, bootstrap, and debate experiments, we score each completion conditioned on the other responses and sum its token log-probabilities. Let $L(\pi) = \sum_n |\pi(s_n)|$. The reported metric is

$$\hat{\chi}(\pi) = \sum_n \ln P_\theta(a_n \mid a_{-n}, s_n), \qquad \chi_{\mathrm{pt}}(\pi) = \frac{\hat{\chi}(\pi)}{L(\pi)}. \tag{4}$$

Table 1: Coherence-accuracy coupling for Gibbs sampling (100 seeds $\times$ 16 rounds).

Test	Statistic	Value	p
OLS (seed fixed effects)	$\hat{\beta}_1$	1.163	$< 10^{-300}$
	t	54.25	
	R^2	0.788	
End-of-episode r	Pearson r	+0.604	$< 10^{-10}$
	95% CI	$[+0.44, +0.74]$	
Within-seed r (all)	Mean r	+0.772	$< 10^{-57}$
	% positive	100%	

Table 2: Overview of key experiments and results.

	Gibbs	**Bootstrap**	**Debate**	**ICM**
Accuracy gain	+* [§5.1]	+* [§5.2]	+* [§5.2]	+* [Wen et al.]
Coh–acc coupling	+* [§5.1]	+* [§5.2]	+* [§5.2]	N/A (implied)
Rejection filter	+ [§5.1]	+* [§5.2]	+* [§5.2]	N/A (inapplicable)
Info saturation	N/A (inapplicable)	N/A (inapplicable)	N/A (inapplicable)	+* [§5.3]

+: supportive −: opposing ?: mixed findings $*: p < .05$

Table 3: Gibbs gains across setups. R is the final round; T is the update temperature. $\hat{\beta}_1$ is the seed-fixed-effect slope of train accuracy on per-token coherence. Peak is each seed's maximum train accuracy, averaged across seeds; R_{peak} is the mean selected round. See Appendix B.5 for p-values. Coherence and slope not computed for SFT setups due to compute constraints.

Model	Method	n	T	R	Seeds	Initial Acc.	Final Acc.	Peak	R_{peak}	$\hat{\beta}_1$
1B-Instruct	ICL	100	.3	15	100	$21.9 \pm .4$	$46.2 \pm .9$	$52.1 \pm .7$	10.5	1.163
1B-Instruct	ICL	200	.3	15	20	$26.3 \pm .7$	$31.2 \pm .6$	$36.6 \pm .6$	5.1	.170
1B-Instruct	ICL	200	.7	15	20	$21.6 \pm .6$	$37.6 \pm .9$	$42.6 \pm .7$	9.7	.669
8B-Base	ICL	50	.7	10	100	$18.6 \pm .6$	$31.3 \pm .8$	$41.0 \pm .6$	5.6	.326
8B-Base	ICL	50	.3	20	90	$46.6 \pm .8$	$54.6 \pm .7$	$64.8 \pm .6$	9.3	.230
8B-Instruct	ICL	200	.3	15	20	$84.2 \pm .6$	$87.6 \pm .4$	$90.1 \pm .4$	6.6	.278
8B-Base	SFT	5,000	.7	10	15	$29.8 \pm .2$	32.5 ± 5.4	43.3 ± 2.2	5.4	—
Qwen3.8-27B‡	SFT	2,000	.3	6	2	$72.1 \pm .2$	$72.8 \pm .2$	$72.8 \pm .2$	6	—

‡Qwen3.8-27B runs on SuperGPQA to avoid saturation; see Appendix B.6.

Gibbs and bootstrap condition across questions, so their response pool forms one d-policy. Debate conditions within each question, so each debate forms a separate d-policy. We normalize each d-policy's score by its total response-token count, then average equally across the independent debates. We call χ_{pt} *per-token coherence* for brevity: it is per-token-normalized leave-one-out coherence, in nats per token. Before normalization, this has the conditioning form of f_{MP}, with a change of log base from Equation 3. Dividing by response length changes the metric; it is not an approximation of the unnormalized χ. It reduces the mechanical length dependence of a log-probability sum but does not make the metric independent of response content or length (Appendix A.1).

We report means $\pm$ standard error (SE) across seeds unless indicated otherwise. For coherence–accuracy coupling, we regress train accuracy (as a fraction) on per-token coherence with seed fixed effects and OLS standard errors. We also report the mean within-seed Pearson correlation across rounds, its two-sided one-sample t-test across seeds, and the fraction of positive correlations. Final-round correlations compare seeds. Train accuracy describes the response pool being iteratively updated; test accuracy evaluates held-out questions. Peak accuracy is each seed's maximum over rounds, averaged across seeds. We therefore also report accuracy at the fixed final round. Rejection experiments compare paired runs with and without the filter on identical data splits (Algorithm 2); a two-sided paired t-test evaluates the accuracy difference.

5 Results

In this section, we present our experiment results, organized by the method tested.

5.1 Gibbs Sampling

We evaluate Gibbs sampling with ICL as our learning system on GSM8K with 1B-Instruct, 100 seeds, and n=100 training items (settings in Appendix F).

Coherence-accuracy coupling. Table 3 shows the coherence and accuracy gains from Gibbs sampling under different setups, showing positive coupling between the two. As an example, for 1B-Instruct, Figure 1 shows that train accuracy and per-token coherence increase overall over 15 rounds. The OLS regression with seed fixed effects (Table 1) gives a positive slope of $\hat{\beta}_1 = 1.163$ ($t = 54.25$, $p < 10^{-300}$), indicating that a 0.1 increase in per-token coherence predicts $\sim$12 pp higher accuracy within the same seed. The association is positive across seeds at the final round ($r = +0.604$) and within seeds across rounds (mean $r = +0.772$, positive in all 100 seeds).

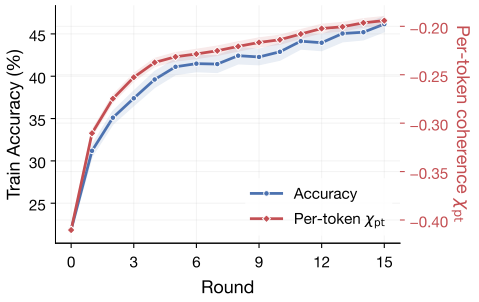

Figure 1: Gibbs sampling on GSM8K (1B-Instruct, n=100, 100 seeds, $\pm$1 SE). Train accuracy and per-token coherence χ_{pt} co-improve over 15 rounds.

Test accuracy gains. Mean test accuracy rises from 28.69% at round 0 to 39.69$\pm$0.65% at round 15 (+11.00 pp; 100 seeds, 200 held-out problems per seed), exceeding golden 25-shot prompting (35.9% $\pm$ 0.33%) by 3.78 pp. The paired gain across seeds is 11.00 $\pm$ 0.83 pp ($t = 13.27$, $p = 1.1 \times 10^{-23}$). The mean per-seed peak is 44.9%, at a mean peak round of 9.3 (Figure 2). *This peak is a retrospective diagnostic, not the accuracy of a test-independent stopping rule.*

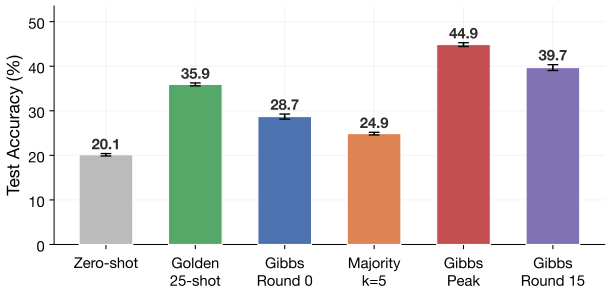

Figure 2: Test accuracy on GSM8K (1B-Instruct, 100 seeds, $\pm$1 SE). Peak is the per-seed maximum over test rounds, averaged across seeds (mean selected round 9.3). The fixed round 15 result is 39.69 $\pm$ 0.65%, versus golden 25-shot at 35.9% $\pm$ 0.33%.

Ablation via rejection filtering. On GSM8K with 8B-Instruct (three paired seeds, 15 rounds), rejection lowers mean train accuracy by 3.3 $\pm$ 3.7 pp for vanilla Gibbs (paired $t = -.90$, $p = .464$) and 4.0$\pm$1.2 pp for preservation Gibbs ($t = -3.46$, $p = .074$), with decreases in all three preservation seeds. The vanilla filter rejects any coherence increase; the preservation filter allows increases of up to 5%. Appendix E gives per-seed comparisons.

5.2 Bootstrap and Debate

The theory predicts that independently developed self-improvement methods all optimize coherence (Qiu et al., 2026). We test this by applying the same per-token coherence analysis from Section 5.1 to bootstrap (Canavan et al., 2026b;a) and cooperative debate (Du et al., 2023), both using 1B-Instruct on GSM8K.

Coherence and accuracy: bootstrap. Figure 3 shows the same co-improvement pattern for bootstrap (100 seeds, n=50, 15 rounds): accuracy rises from 22.9% to 43.4% (+20.5 pp) while per-token coherence improves from -0.408 to -0.164. The paired gain across seeds is 20.50 ± 0.76 pp ($t = 27.12$, $p = 1.3 \times 10^{-47}$). The OLS coefficient is $\hat{\beta}_1 = 0.713$ ($t = 36.40$, $p = 2.1 \times 10^{-208}$), and the end-of-episode $r = +0.447$ ($p < 10^{-5}$). The coupling persists despite a different update mechanism: bootstrap reuses the full previous round as context rather than resampling a complement.

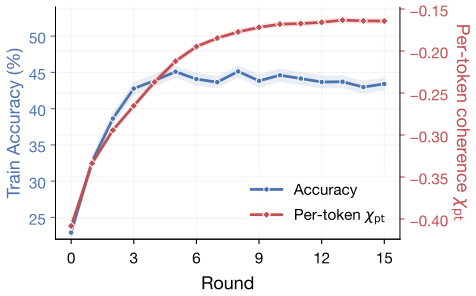

Figure 3: Bootstrap train accuracy and per-token coherence co-improve over 15 rounds (1B-Instruct, 100 seeds), mirroring the Gibbs pattern (Figure 1). Shading shows ± 1 SE.

Coherence and accuracy: cooperative debate. Figure 4 shows the pattern extends to cooperative debate (50 seeds, 5 rounds): majority accuracy rises from 23.2% to 28.8% at the final round (+5.6 pp) with coherence improving from -0.613 to -0.548. The paired gain across seeds is 5.60 ± 1.10 pp ($t = 5.10$, $p = 5.5 \times 10^{-6}$). The OLS coefficient is $\hat{\beta}_1 = 0.137$ ($t = 3.04$, $p = 0.003$). A higher-temperature variant (T=0.7, 83 seeds) produces noisier dynamics, but accuracy and coherence dips/recoveries remain aligned round-to-round and the OLS coefficient stays significant ($p < 10^{-4}$; Appendix C). Debate is the most structurally different method tested: two agents exchange solutions rather than a single model resampling. That the coupling persists is consistent with coherence as the shared objective.

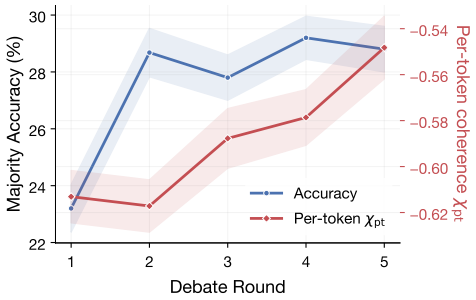

Figure 4: Cooperative debate majority accuracy and per-token coherence co-improve over 5 rounds (1B-Instruct, 50 seeds), consistent with the Gibbs and bootstrap patterns. Shading shows ± 1 SE.

Ablation via rejection filtering. We have established that coherence and accuracy co-improve across methods. To go beyond correlation, we apply the coherence-suppressing rejection filter (Algorithm 2, Appendix E) to bootstrap and cooperative debate on 1B-Instruct and Gibbs sampling on 8B-Instruct, running paired versions with and without the filter on identical data splits for each seed.

Rejection reduces final-round accuracy in both methods (Figures 5–6). After 10 rounds of bootstrap, accuracy reaches 41.0% with rejection versus 43.9% without it, a reduction of 2.9 pp (paired $t = 2.89$, $p = 0.006$, $N = 44$). At round 5 of debate, majority accuracy is 22.1% versus 29.2%, a reduction of 7.1 pp (paired $t = 4.53$, $p < 10^{-4}$, $N = 32$).

5.3 ICM

Informational saturation. We test the informational-saturation argument of Qiu et al. (2026) using Llama-3.1-8B-Base and the GSM8K verification pool released by Wen et al. (2025). Each question supplies four candidate solutions; a claim states that one candidate answer is correct. We score fixed gold and balanced random labels on nested sets of 8–256 claims (32 seeds, disjoint questions across seeds). We report the mean gap per scored claim and the relative gap as a fraction of the smaller score magnitude (definitions and scoring in Appendix F.4).

With gold labels and random claim order, the mean gap falls from 0.181 to 0.077 nats as the set grows from 8 to 256 claims; the relative gap falls from 66% to 17% (Figure 7). This is consistent with partial informational saturation. Within seeds, the mean gap declines with pool size (slope

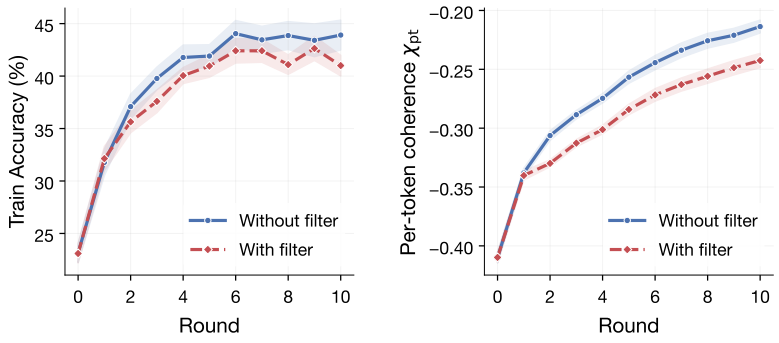

Figure 5: Rejection ablation on bootstrap (1B-Instruct, 44 paired seeds, 10 rounds). Suppressing coherence increases (dashed red) degrades both coherence and accuracy relative to running without the filter (solid blue). ± 1 SE.

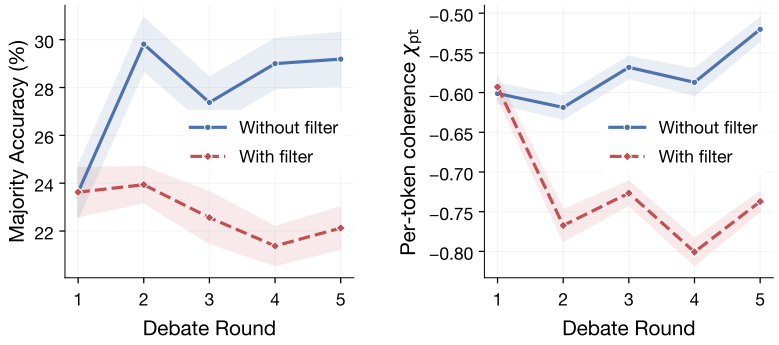

Figure 6: Rejection ablation on cooperative debate (1B-Instruct, 32 paired seeds, 5 rounds). The accuracy loss (-7.1 pp) is larger than in bootstrap. ± 1 SE.

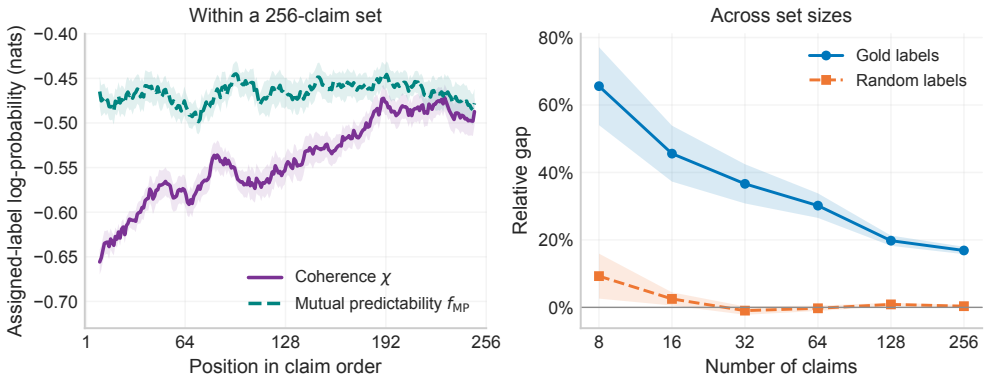

Figure 7: Coherence and mutual predictability with randomly ordered claims (8B-Base). Left: assigned-label log-probabilities under the two objectives at each position in a 256-claim set, using gold labels. Curves use 16-position moving averages within each seed. Right: relative gap, $100\,(f_{\mathrm{MP}} - \chi)/\min(|f_{\mathrm{MP}}|, |\chi|)$ in percent, with gold or balanced random labels on the same claims. Both panels show means ± 1 SE across 32 seeds.

-0.019 ± 0.003 nats per claim per doubling of the set, $t = -5.54$, $p = 4.5 \times 10^{-6}$, 32 seeds). Gaps are smaller for later claims, which have more preceding context and fewer remaining claims, and nearly vanish with random labels at 256 claims. Full results are in Appendix D.

6 Conclusion

We conclude by summarizing our findings and their limitations.

6.1 Summary of Findings

Coherence and accuracy co-improve across methods and models. Per-token coherence and accuracy rise together in the main GSM8K runs of Gibbs, bootstrap, and debate. All five additional GSM8K ICL settings on 1B-Instruct and 8B models also show positive within-seed slopes, with magnitudes that vary with temperature and model ($\hat{\beta}_1$ from .170 to .669); the 90-seed base-model experiment improves train accuracy by $8.0 \pm .8$ percentage points over 20 rounds. A preliminary 27B SuperGPQA experiment shows a small training-pool accuracy gain (two seeds).

ICM approaches coherence under partial saturation. ICM experiments with fixed gold labels show decreasing mean and relative gaps between mutual predictability and coherence as randomly ordered claim sets grow, consistent with partial informational saturation.

Gibbs improves held-out accuracy in the main setting. In the main 1B-Instruct GSM8K experiment (100 seeds), test accuracy rises by 11.00 ± 0.83 pp to $39.69 \pm 0.65\%$ at the fixed final round, above golden 25-shot prompting ($35.9 \pm 0.33\%$), without selecting a round by test accuracy. The additional settings report train-pool accuracy; the preliminary 27B held-out gain ($+2.2$ pp) has a 95% interval that includes zero.

Rejection provides intervention evidence. Coherence-based rejection reduces accuracy in bootstrap and debate, but further controls are needed to isolate the mechanism. The Gibbs comparisons also show lower mean accuracy with rejection across three paired seeds. Together, these results support coherence optimization as a unified account of self-improvement, with agreement between objectives and gains in accuracy depending on the setting.

6.2 Limitations and Future Work

Model scale. The experiments cover Llama-3.2-1B-Instruct and Llama-3.1-8B-Base and Instruct models, plus a preliminary two-seed Qwen3.8-27B experiment. Other model families and larger scales remain largely untested. Changes in pool size, temperature, and learning method limit causal comparisons of scale.

Benchmark diversity. GSM8K remains the primary benchmark, with a preliminary SuperGPQA experiment for Gibbs only; bootstrap and debate are evaluated only on GSM8K. Open-ended non-reasoning tasks and multilingual settings remain untested.

Coherence measurement. SFT accuracy results lack coherence measurements due to compute constraints.

Rejection mechanism. The filter accepts the least-increasing proposal after retries are exhausted. It may therefore fail to suppress coherence. Even when suppression succeeds, the intervention changes proposal selection and compute. A matched random-rejection control is needed to distinguish these effects. The style diagnostics in Appendix E use a separate intervention and do not resolve this identification problem.

Base model assumption. The theory assumes access to a base model prior σ (Qiu et al., 2026). Our experiments use both base and instruct model variants (Table 3). Whether post-training (RLHF, instruction tuning) fundamentally alters the coherence landscape is an open question with implications for applying these methods to production models.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods and experiment code, clean and reformat datasets, give feedback on experimental design, help refine explanatory hypotheses for observed results, and help interpret results. We have not used generative AI tools to develop the theoretical framework or formulate mathematical claims. Generating synthetic data, proofs, translation, and qualitative or thematic analysis are not applicable to this work. Additionally, we used generative AI tools to write and edit code, create figures, suggest experimental parameters, draft paper text (excluding references) from a manually written, detailed outline, and improve readability. We have reviewed all AI-assisted work: all experiment code was reviewed by at least 1 author, reported statistics were recomputed from saved outputs with separate verification scripts, references were compiled and checked manually, and all text underwent extensive manual editing and review. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie R. Kerr, Jared Mueller, Jeffrey Ladish, Joshua D. Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas C. Joseph, Noemi Mercado, Nova DasSarma, Robin J. Larson, Sam McCandlish, Sandipan Kundu, Scott Gregory Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom P Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring progress on scalable oversight for large language models. In *arXiv (Cornell University)*, 2022.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *arXiv (Cornell University)*, 2022.

Collin Burns, Pavel Izmailov, J. Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas R. Joglekar, Jan Leike, I. Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *International Conference on Machine Learning*, pp. 4971–5012, 2023.

Callum Canavan, Aditya Shrivastava, Allison Qi, Jonathan Michala, and Fabien Roger. Three concrete challenges and two hopes for the safety of unsupervised elicitation. *arXiv preprint arXiv:2602.20400*, 2026a.

Callum Canavan, Aditya Shrivastava, Allison Qi, Tianyi Alex Qiu, Jonathan Michala, and Fabien Roger. Eliciting base models with simple unsupervised techniques. https://www.lesswrong.com/posts/rFxfMbwJ3v4PNesWP/eliciting-base-models-with-simple-unsupervised-techniques, 2026b. LessWrong.

Paul F. Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. In *arXiv (Cornell University)*, 2018.

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

Yilun Du, Shuang Li, A. Torralba, J. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *International Conference on Machine Learning*, pp. 11733–11763, 2023.

Audrey Y. Huang, Adam Block, Dylan J. Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T. Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In *arXiv (Cornell University)*, 2024.

Geoffrey Irving, Paul F. Christiano, and Dario Amodei. Ai safety via debate. In *arXiv (Cornell University)*, 2018.

Akbir Khan, John B Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. In *arXiv (Cornell University)*, 2024.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36*, pp. 46534–46594, 2023.

Tianyi Qiu, Ahmed Hani Ismail, Zhonghao He, and Shi Feng. Self-improvement as coherence optimization: A theoretical account. *Transactions on Machine Learning Research*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Viet Le, Ed H., Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *arXiv (Cornell University)*, 2022.

Jiaxin Wen, Zachary Ankner, Arushi Somani, Peter Hase, Samuel Marks, Jacob Goldman-Wetzler, Linda Petrini, Henry Sleight, Collin Burns, He He, Shi Feng, Ethan Perez, and Jan Leike. Unsupervised elicitation of language models. In *arXiv*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems 35*, pp. 15476–15488, 2022.

A Coherence Measurement and Normalization

Equation 4 defines the measured leave-one-out score. In the main GSM8K experiments, each response is scored with the other items provided as demonstrations. This is a measurement choice, not a requirement imposed by ICL: one can also score an ordered sequence of prefixes. The theoretical relationship to autoregressive coherence is conditional on the saturation argument in Section 4.1.

Equation 4 sums completion-token log-probabilities and divides by the total number of response tokens. The same model generates and scores responses. Main GSM8K Gibbs and bootstrap scoring uses all other question–response pairs. Gibbs generation instead uses a γ-fraction conditioning chunk.

For each debate question, the d-policy comprises the $N = 2$ agents' responses $a_1, \ldots, a_N$, giving

$$\chi_{\mathrm{pt}}(\pi) = \frac{\sum_{k=1}^{N} \ln P_\theta(a_k \mid a_{-k}, \text{question})}{\sum_{k=1}^{N} |a_k|}.$$

We then average equally across questions, regardless of response length. Both metrics are in nats per token. Appendix A.1 illustrates the raw score's sensitivity to length.

A.1 Response Length and Raw vs. Per-Token Coherence

Response shortening on 8B-Base. In five GSM8K runs using 8B-Base, we supplied a simple "$2 + 2$" example to illustrate the answer format. Between rounds 0 and 6, mean response length falls from 165 to 77 characters. Raw coherence rises from $-2{,}155$ to $-1{,}198$, while per-token coherence falls from -0.782 to -0.991. By round 10, response length and per-token coherence recover. The intermediate divergence illustrates the raw score's sensitivity to response length.

Table 4 reports means across 100 seeds. The total response-token count per d-policy approximately doubles from round 0 to round 15 ($6{,}333 \to 12{,}945$, $+104\%$).

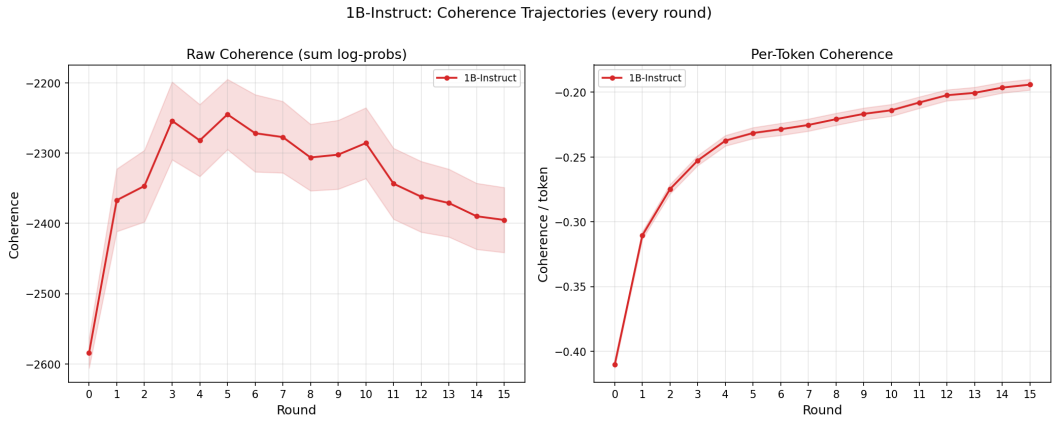

Figure 8: GSM8K Gibbs sampling (1B-Instruct). **Left:** Raw coherence (sum of log-probabilities) over rounds. Raw coherence improves in early rounds but then worsens as responses lengthen. **Right:** Per-token coherence over rounds. Mean per-token coherence increases at every round, even as total response-token count grows. Shaded regions indicate ± 1 SE across 100 seeds.

This length increase explains the divergence between raw and per-token coherence. Raw coherence $\hat{\chi} = \chi_{\text{pt}} \times \sum_n |\pi(s_n)|$ is the product of per-token coherence and total token count. Mean per-token coherence improves monotonically ($-0.410 \to -0.194$, $+53\%$), but each additional token adds a negative log-probability to the raw sum. In early rounds (0–3), per-token improvement dominates and raw coherence improves ($-2{,}584 \to -2{,}254$). In later rounds (4–15), token growth dominates and raw coherence worsens ($-2{,}254 \to -2{,}395$). The result is that raw coherence is non-monotonic and negatively correlated with accuracy in late rounds ($r = -0.163$, $p = 0.001$; Table 5), while per-token coherence tracks accuracy throughout.

Table 4: Response length, raw coherence, and per-token coherence across rounds (100 seeds, mean values). Total tokens approximately double while per-token coherence improves by 53%.

Round	Total tokens	Raw $\hat{\chi}$	χ_{pt}	Train acc (%)
0	6,333	$-2{,}584$	-0.410	21.9
3	9,363	$-2{,}254$	-0.253	37.4
6	10,708	$-2{,}272$	-0.229	41.5
9	11,397	$-2{,}302$	-0.217	42.3
12	12,326	$-2{,}362$	-0.202	44.0
15	12,945	$-2{,}395$	-0.194	46.2
Change (R0→R15)		$+7\%$	$+53\%$	$+24.3$ pp

A.2 Raw Coherence Trajectory

Raw coherence (sum of leave-one-out log-probabilities) exhibits a three-phase pattern across rounds:

- **Early (R0–R3):** Raw coherence improves from $-2{,}584$ to $-2{,}254$ ($+330$), as per-token coherence gains dominate the still-moderate increase in response length.
- **Middle (R4–R9):** Raw coherence stagnates ($-2{,}282 \to -2{,}302$, -20). Per-token improvement and length growth approximately cancel.
- **Late (R10–R15):** Raw coherence worsens ($-2{,}286 \to -2{,}395$, -109). Length growth dominates; responses continue to lengthen while per-token gains slow.

This non-monotonic pattern explains why raw coherence is negatively correlated with accuracy in late rounds (Table 5, $r = -0.163$, $p = 0.001$): later rounds have both higher accuracy and worse raw coherence, because longer responses accumulate more negative log-probability terms.

B Additional Gibbs Results

B.1 Within-Seed Correlation Analysis

Table 5 reports within-seed Pearson correlations between train accuracy and coherence for 1B-Instruct on GSM8K, split by round window and coherence metric. The contrast between raw and per-token coherence is striking: raw coherence shows a *negative* correlation with accuracy overall ($r = -0.153$), while per-token coherence shows a strong positive correlation ($r = +0.772$). Response length approximately doubles over rounds (Table 4); the raw sum therefore mixes predictability with the number of tokens scored.

Table 5: Within-seed Pearson r between train accuracy and coherence (100 seeds, 16 rounds). Reported: mean r across seeds; % of seeds with $r > 0$; p-value from one-sample t-test ($H_0 : \bar{r} = 0$).

Metric	Rounds	Mean r	% positive	p-value
Raw $\hat{\chi}$	All (0–15)	-0.153	—	7.2×10^{-6}
	Early (0–7)	$+0.007$	—	0.857
	Late (8–15)	-0.163	—	0.001
Per-token χ_{pt}	All (0–15)	$+0.772$	100%	6.1×10^{-58}
	Early (0–7)	$+0.819$	—	5.9×10^{-70}
	Late (8–15)	$+0.400$	—	1.4×10^{-14}

B.2 Test Accuracy Trajectory

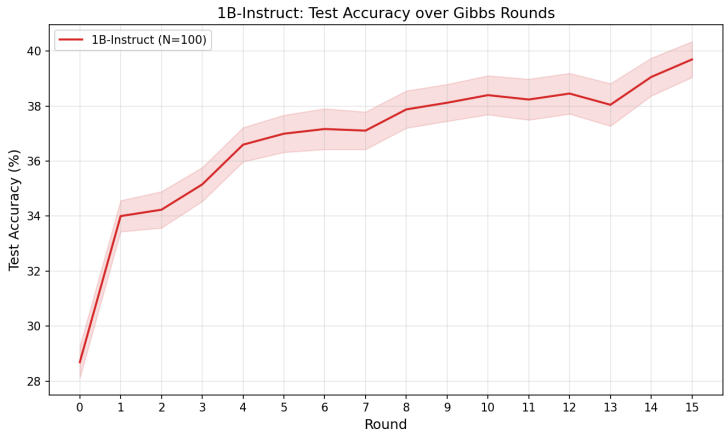

Figure 9: Test accuracy over 15 rounds of Gibbs sampling (100 seeds, ± 1 SE). Test accuracy rises from 28.69% at round 0 to 39.69% at round 15 ($+11.00$ pp).

Across 100 seeds, mean test accuracy rises from $28.69 \pm 0.57\%$ at round 0 to $39.69 \pm 0.65\%$ at round 15, a paired gain of 11.00 ± 0.83 pp (mean $\pm$ SE). The test-pool size is constant across rounds within each seed: 68 seeds have 200 questions, 28 have 199, and four have 198. We average each seed's accuracy equally. The mean per-seed peak is 44.9% (median 45.0%; Figure 10), at a mean selected round of 9.27. The fixed final-round score exceeds golden 8-shot (36.1%) and golden 25-shot (35.9%) prompting without selecting a round using test accuracy.

B.3 Per-Problem Transition Analysis

Across 100 seeds $\times$ 100 training problems (9,996 total observations), Table 6 shows how individual problem correctness changes from round 0 to round 15. Gibbs sampling produces a net improvement of 2,425 problem-level corrections, but this net gain masks substantial churn: 32.1% of all problems

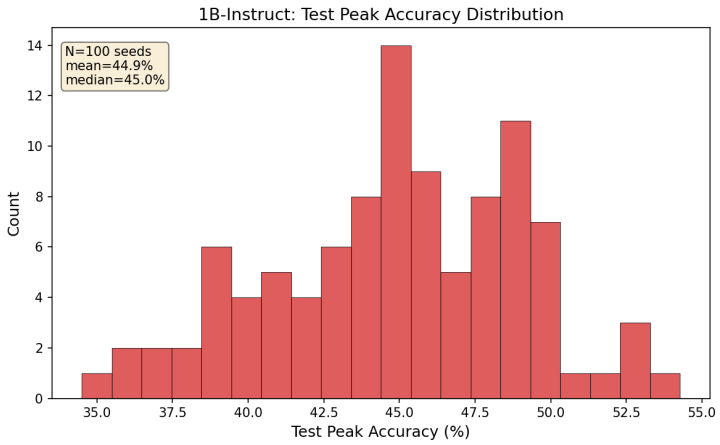

Figure 10: Distribution of per-seed peak test accuracy across 100 seeds ($N = 100$, mean 44.9%, median 45.0%). The spread reflects variability in which problems each seed samples.

change from incorrect to correct, while 7.8% change from correct to incorrect. Another 46.0% remain incorrect at both endpoints; this alone does not establish a model-capability limit.

Table 6: Per-problem correctness transitions from round 0 to round 15 (100 seeds $\times$ 100 training problems).

Transition	Count	Fraction
Wrong $\rightarrow$ Right (improved)	3,208	32.1%
Right $\rightarrow$ Right (stable correct)	1,406	14.1%
Wrong $\rightarrow$ Wrong (persistent error)	4,599	46.0%
Right $\rightarrow$ Wrong (regression)	783	7.8%
Net improvement	2,425	—

B.4 Qualitative Prediction Examples

We illustrate how Gibbs sampling changes model responses using predictions at round 0 and round 15 for the same problem in seed 50.

Example: wrong $\rightarrow$ correct. Consider the problem: "Hajar's score is 24 points. The difference between Hajar and Farah's score is 21. What is the sum of their scores, assuming Farah's score is higher?" (Gold answer: 69.)

At round 0, the model produces a confused response (36 words): it computes Farah's score as $24 + 21 = 45$ but then subtracts 21 back, returning the incorrect answer 24. The reasoning contains a logical loop.

At round 15, the model produces a structured two-step solution (59 words): Step 1 correctly computes Farah's score as $24 + 21 = 45$; Step 2 sums $24 + 45 = 69$. The response is longer, uses numbered steps, and avoids the circular error.

Response length. Across 9,996 training responses from 100 seeds, mean response length increases from 44 words at round 0 to 92 words at round 15 (median: 22 $\rightarrow$ 85).

B.5 Additional Gibbs Statistics

The following table supplements the final-round results with per-seed peaks and coupling statistics.

Table 7: Supplementary GSM8K statistics for Table 3, in the same row order, omitting its first and last rows. Peak accuracy is percent, mean ± SE where reported; peak round averages each seed's selected round. r is the mean within-seed correlation.

Setting	Peak accuracy	Peak round	Slope p-value	r
1B-Instruct, ICL, $n = 200$, $T = .3$	$36.6 \pm .6$	5.1	1.24×10^{-6}	.27
1B-Instruct, ICL, $n = 200$, $T = .7$	$42.6 \pm .7$	9.7	$< 10^{-50}$	.815
8B-Base, ICL, $n = 50$, 100 seeds	$41.0 \pm .6$	5.6	2.0×10^{-72}	.528
8B-Base, ICL, $n = 50$, 90 seeds	$64.8 \pm .6$	9.3	3.28×10^{-40}	.299
8B-Instruct, ICL, $n = 200$	$90.1 \pm .4$	6.6	2.2×10^{-16}	.42
8B-Base, SFT, $n = 5{,}000$, $R = 10$	43.3 ± 2.2	5.4	—	—

B.6 Preliminary 27B Results on SuperGPQA

Training-pool accuracy. We observe positive accuracy changes in both training seeds when evaluating the round-6 Qwen3.8-27B adapters on fresh answers to the 2,000 training questions. With targeted continuation of capped answers to 24,576 tokens, mean accuracy rises from 72.05% to 72.83% ($+0.78$ percentage points; Table 8). The gains are $+0.70$ and $+0.85$ points in seeds 42 and 43. These are preliminary accuracy gains at 27B scale without gold-answer supervision. They measure newly generated answers from the trained adapters, rather than the stored response pool, which retains answers between rounds.

Table 8: Fresh training-pool accuracy on SuperGPQA (%). Both seeds use the same 2,000 questions. "Extended" continues only answers capped at 16,384 tokens, up to 24,576; it is a length diagnostic, not a fresh full-pool evaluation at the larger limit. Means average the two training seeds.

	16,384-token limit			Extended capped answers		
Seed	Base	Round 6	Gain (pp)	Base	Round 6	Gain (pp)
42	71.75	72.20	+0.45	71.90	72.60	+0.70
43	71.90	72.65	+0.75	72.20	73.05	+0.85
Mean	71.83	72.43	+0.60	72.05	72.83	+0.78

Held-out evaluation. After selecting the seed-42 round-6 adapter using earlier evaluations, we compare it with the base model on 500 previously unused calculation questions sampled uniformly from the combined mathematics and physics pool. Accuracy increases from $349/500$ (69.8%) to $360/500$ (72.0%), a $+2.2$-point gain (paired-bootstrap 95% interval $[-0.2, 4.6]$ points). There are 24 wrong-to-correct and 13 correct-to-wrong changes. The base model has one truncated answer and the adapter has none. Both models use the same questions, temperature .3, and a 24,576-token generation limit.

Setup. We use the `Qwen/Qwen3.8-27B` checkpoint and SuperGPQA, restricting questions to mathematics and physics with `is_calculation=true`. The training pool contains 1,000 questions per field and is shared by two independently generated initial response pools (seeds 42 and 43). Each round fits a fresh rank-16 adapter (LoRA alpha 32) on 1,500 randomly selected responses, retains those responses, and regenerates the remaining 500. In the preservation setting, 450 fitting targets are replaced with fresh frozen-model responses. Training uses one epoch, learning rate 10^{-5}, effective batch 32, weight decay .01, and zero LoRA dropout. The learning rate was chosen from $\{5 \times 10^{-6}, 10^{-5}, 10^{-4}\}$ using three-round runs without base-response mixing.

C Cooperative Debate at Higher Temperature

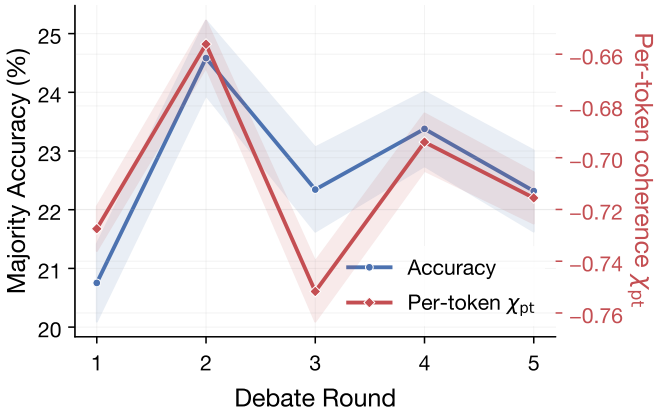

Figure 11: Cooperative debate at temperature 0.7 (1B-Instruct, 83 seeds, 5 rounds). The trajectory is non-monotonic, but the round-to-round dips and recoveries in accuracy and coherence are aligned: both peak at round 2, dip at round 3, and partially recover at round 4. The OLS regression with seed fixed effects yields $\hat{\beta}_1 = 0.125$ ($t = 3.95$, $p < 10^{-4}$, $R^2 = 0.34$; $N = 415$), indicating a positive within-seed association despite the noisier trajectories.

We repeat the cooperative debate experiment from Section 5.2 at temperature 0.7 with 83 seeds. Higher temperature increases sampling diversity but produces a noisier trajectory: majority accuracy peaks at 24.6% (round 2), dips to 22.3% (round 3), and settles at 22.3% (round 5), a net gain of only +1.6 pp compared to +5.6 pp at temperature 0.3. Per-token coherence follows the same non-monotonic pattern, peaking at round 2 (-0.656) and dipping at round 3 (-0.752). Accuracy and coherence track one another despite their non-monotonic trajectories. This pattern is consistent with coherence contributing to accuracy improvements, a possibility we examine further through the rejection experiments. Across seeds at the final round, the correlation is $r = +0.152$ ($p = 0.17$); the regression using all rounds has a positive within-seed slope ($p < 10^{-4}$).

D ICM Supplementary Analyses

D.1 Ordering

Table 9 expands the ordering comparison in Section 5.3. Shuffling question blocks produces a similar mean gap to data order, whereas shuffling individual claims changes both its magnitude and sign at larger sizes. At 256 claims, the paired difference between claim shuffle and question shuffle is 0.188 ± 0.005 nats; question shuffle minus data order is 0.003 ± 0.004. Thus, the observed reduction under random claim order depends on how related claims enter the context. Negative gaps mean that gold labels receive lower mean log-probability with all other claims than with preceding claims alone.

Table 9: Gold-label gaps across set sizes and orderings, mean ± SE over 32 seeds. Data order keeps each question's four claims together; question shuffle reorders these blocks; claim shuffle mixes all claims. Relative gaps use the formula in Section 5.3, calculated separately for each seed.

	Mean gap (nats)			Relative gap (%)		
Claims	Data order	Question shuffle	Claim shuffle	Data order	Question shuffle	Claim shuffle
8	0.114 ± 0.018	0.143 ± 0.021	0.181 ± 0.020	39.7 ± 6.7	52.7 ± 9.8	65.6 ± 11.6
16	0.081 ± 0.015	0.068 ± 0.017	0.130 ± 0.018	29.3 ± 6.2	23.7 ± 6.0	45.6 ± 8.3
32	0.049 ± 0.011	0.050 ± 0.011	0.124 ± 0.012	17.4 ± 3.9	16.1 ± 3.4	36.6 ± 5.8
64	-0.006 ± 0.007	-0.004 ± 0.008	0.119 ± 0.010	-1.9 ± 2.0	-1.4 ± 2.3	30.2 ± 3.6
128	-0.054 ± 0.006	-0.051 ± 0.005	0.087 ± 0.006	-15.2 ± 1.7	-14.2 ± 1.6	19.8 ± 1.5
256	-0.113 ± 0.005	-0.111 ± 0.004	0.077 ± 0.004	-32.4 ± 1.6	-31.4 ± 1.2	16.9 ± 1.1

For the size trend, we fit a slope against $\log_2 |\mathcal{S}|$ separately for each seed across all six sizes. With gold labels and random claim order, the mean slope is -0.0187 ± 0.0034 nats per claim per doubling. The final doubling, from 128 to 256 claims, changes the mean gap by -0.0099 ± 0.0073 nats. The overall decline is clearer than the change over this final interval.

Random-label control. With random claim order at 256 claims, balanced random labels give a mean gap of 0.002 ± 0.003 nats, compared with 0.077 ± 0.004 for gold labels. The claims and their arrangement are identical, so matching claims and prompt format alone does not reproduce the gold-label gap.

D.2 Claim Position

We examine early and late positions in the 256-claim sets with gold labels and random claim order (Table 10). The two ranges contain 63 claims each. We exclude position 256 because its two contexts are identical by construction. These exploratory summaries use unsmoothed scores; the moving average in Figure 7 is only for display.

Table 10: Early versus late claims at fixed set size (256), mean ± SE across 32 seeds. Relative gaps are calculated from score sums within the indicated positions. The last column averages absolute changes in assigned-label probability, so increases and decreases cannot cancel; pp denotes percentage points.

Positions	Mean gap (nats)	Relative gap (%)	Absolute probability change (pp)
Early (2–64)	0.141 ± 0.008	31.2 ± 2.4	16.2 ± 0.4
Late (193–255)	0.019 ± 0.006	4.6 ± 1.3	7.5 ± 0.2

The smaller late-position gap is accompanied by a reduction in absolute probability changes, from 16.2 to 7.5 percentage points. Thus, the position trend reflects reduced sensitivity to additional context as well as cancellation of positive and negative changes. These reductions are consistent with partial informational saturation. Later positions also have fewer remaining claims, so this comparison jointly varies the preceding context and the amount of context added.

E Rejection-Sampling Ablation Design

Algorithm 2 Coherence-Suppressing Rejection Wrapper

Input: Base Gibbs sampler $\mathcal{G}$, d-policy π, threshold τ, max retries K
$\chi_{\text{pt}}^{\text{old}} \leftarrow \text{PerTokenCoherence}(\pi)$
$\delta^* \leftarrow \infty$; $\pi^* \leftarrow \pi$ ▷ Track least-increasing proposal
for $k = 1, \ldots, K+1$ **do**
  $\pi' \leftarrow \mathcal{G}(\pi)$ ▷ Run one Gibbs round from π
  $\chi_{\text{pt}}^{\text{new}} \leftarrow \text{PerTokenCoherence}(\pi')$
  $\delta \leftarrow (\chi_{\text{pt}}^{\text{new}} - \chi_{\text{pt}}^{\text{old}}) \,/\, |\chi_{\text{pt}}^{\text{old}}|$
  if $\delta \leq \tau$ **then**
    return π' ▷ Accept
  else if $\delta < \delta^*$ **then**
    $\delta^* \leftarrow \delta$; $\pi^* \leftarrow \pi'$ ▷ Update fallback
  end if
end for
return π^* ▷ Exhaustion: accept least-increasing proposal

The filter rejects proposals whose per-token coherence increase exceeds the threshold. The proposal generator is shared between runs with and without the filter, but rejection changes the distribution of accepted responses and the number of proposals evaluated (Algorithm 2).

The main bootstrap and debate comparisons use $\tau = 0$ (strict: any coherence increase triggers rejection). Round 0 is exempt since the initial d-policy has near-baseline coherence. The design

is paired: for each seed, we run the same method with and without the filter on identical data splits, comparing the effect of adding the filter. Table 12 lists the model-specific experiment sizes, temperatures, rounds, and paired seed counts.

Gibbs comparisons. The 8B-Instruct GSM8K comparisons use $n = 50$, 15 rounds, $T = .7$, $\gamma = .75$, and data seeds 1–3. Preservation Gibbs uses a 5% rejection threshold; final accuracy differences (rejection minus no filter) are -6, -2, and -4 pp, averaging -4.0 ± 1.2 pp ($t = -3.46$, $p = .074$). Vanilla Gibbs uses a zero threshold; across the same three paired seeds, differences are -8, $+4$, and -6 pp, averaging -3.3 ± 3.7 pp ($t = -.90$, $p = .464$). Both use two-sided paired tests and SE across seeds.

Exploratory style-filter experiment. We separately tested a filter intended to discourage repeated wording across responses. It takes the first 120 characters of each response, replaces digit sequences with a common placeholder, lowercases the text, and forms a set of words. The score is the mean pairwise Jaccard similarity: the number of shared words divided by the number of words in either set. A proposed Gibbs update is rejected if this score increases. The filter allows ten retries; if none satisfies the criterion, it retains the proposal with the smallest increase. This intervention targets word overlap rather than coherence.

Using Llama-3.1-8B-Base, we ran ten Gibbs rounds on 50 GSM8K training questions per seed ($T = .7$, $\gamma = .25$). The prompt included a single "$2 + 2$" answer-format example. Across ten paired seeds with identical initial responses, final train accuracy is 21.4% without the filter and 30.8% with it, a paired difference of 9.4 ± 4.0 percentage points (mean $\pm$ SE). Both conditions decline from their shared initial accuracy of 46.8%. We treat this as an exploratory result about the effect of filtering repeated wording.

Separate style-rejection diagnostics. A post-hoc comparison of matched runs with and without the style filter covered 10 prompt configurations. Response-length changes differed by roughly 30–50 characters; unique-answer fractions were about 83%, format-error rates ranged from 0–12%, and answer-flip rates differed by 1–3 pp. These are descriptive comparisons, not equivalence tests. Character-level Shannon entropy was approximately flat in both conditions for eight configurations. For the prompt with a single answer-format example, the change was $-.49$ bits without the filter versus $-.18$ with it; for a short instruction prompt, $-.67$ versus $-.13$ bits. Character entropy measures the distribution of characters within a response, not the model's generation-distribution entropy. These diagnostics concern style rejection and cannot substitute for a matched control of the main coherence filter.

F Hyperparameters and Configurations

F.1 GSM8K Gibbs Settings

Data and splits. The main GSM8K Gibbs experiment uses 100 independently seeded training and test pools. Table 11 gives the pool sizes; each seed has 198–200 held-out questions, fixed across rounds (Section 5.1).

Table 11: Gibbs sampling hyperparameters (GSM8K, Llama-3.2-1B-Instruct).

Parameter	Value
Model	Llama-3.2-1B-Instruct
Backend	vLLM
Learning system	ICL
Sampler	Preservation ($\rho = 0.7$)
Dataset	GSM8K
Training pool size n	100
Test set size	198–200
Number of rounds	15
Context fraction γ	0.25
Chunk size	25
Temperature	0.3
Generation limit (tokens)	512
Coherence frequency	Every round
Batch size	32
Seeds	100 (data seeds 1–100)
Context-window limit (tokens)	16,384

System prompt. The main 1B-Instruct GSM8K experiments (Gibbs sampling and baselines) use the following system prompt, provided as the user turn in the chat template:

```
You are an unbiased math-solving assistant for GSM8K
problems.
Follow these rules on every response:
1.  Provide DIRECT reasoning steps.
2.  Do NOT restate the question or mention your process
beyond the essential math steps.
3.  Use neutral, professional tone; no greetings, apologies,
or filler phrases.
4.  End with the numeric answer alone on a new line in this
exact format:
#### <number>
5.  Do not include units, punctuation, emojis, or commentary
after the final line.
```

The assistant responds with: `Acknowledged. All replies will follow the specified tone and finish with '#### <number>' as the final line.`

Preservation sampler. The preservation sampler constructs the conditioning context for each round by mixing 70% of responses from the previous round's d-policy with 30% freshly generated zero-shot responses from the same model (without in-context examples). This maintains diversity in the conditioning chunk and reduces drift from the pretraining distribution relative to the vanilla sampler, which uses 100% previous-round responses.

Baselines. The comparison uses zero-shot prompting, gold-label 8-shot and 25-shot prompting, and majority voting with $k = 5$, using the same 1B-Instruct model and GSM8K evaluation task (Figure 2).

F.2 Additional GSM8K Gibbs Settings

The 90-seed 8B-base ICL row in Table 3 uses a two-example GSM8K prompt, initial response temperature .7, update temperature .3, 20 rounds, $\gamma = .25$, and preservation ratio .7. Data seeds are 11–100 and the generation seed is 42.

The 100-seed 8B-Base ICL row at $T = .7$ uses a helpful-harmless-honest dialogue prompt followed by a GSM8K instruction ending in `#### <number>`, $n = 50$ training questions without a test set, 10 rounds, $\gamma = .25$, a preservation ratio of .7, a 256-token generation limit, format rejection, data seeds 1–100, and generation seed 42.

The 20-seed ICL experiments on 1B-Instruct at $T = .3$ and $T = .7$ and 8B-Instruct at $T = .3$ use data seeds 42–61, $n_{\text{train}} = n_{\text{test}} = 200$, 15 rounds, $\gamma = .25$, a preservation ratio of .7, a 512-token generation limit, and a 32,768-token context-window limit.

The 8B-Base SFT experiment uses seeds 42–56, $n = 5{,}000$, $T = .7$, and 10 rounds. The preservation ratio is .7, $\gamma = .05$, and the generation limit is 256 tokens. LoRA uses rank 32 and scale 64, five epochs per update, batch size 8, a 2,048-token training-sequence limit, and a constant learning rate of 5×10^{-5} after two warm-up steps. The prompt asks for step-by-step reasoning ending in `#### <number>`, with format rejection enabled. Three seeds evaluate 50 test questions per round; the other twelve evaluate 500.

We report all fifteen seeds, including seeds whose responses degenerate into the placeholder `#### [number]`.

F.3 Bootstrap and Debate Settings

All runs in Table 12 use Llama-3.2-1B-Instruct with vLLM and a 512-token generation limit. Seeds index independently sampled GSM8K question pools; rejection comparisons include only seeds for which both conditions reached the final round. Main bootstrap uses the full previous response pool, excluding the target question, as described in Section 4.1.

Table 12: Main bootstrap and debate settings. Counts for rejection are paired seeds.

Experiment	Questions	Agents	T	Rounds	Seeds
Bootstrap trajectory	50	1	.3	15	100
Bootstrap rejection	50	1	.3	10	44
Debate	50	2	.3	5	50
Debate, higher temperature	50	2	.7	5	83
Debate rejection	50	2	.3	5	32

F.4 ICM Settings

Data and sampling. The released training pool of Wen et al. (2025) contains 5,731 GSM8K questions, each with four candidate solutions and two True gold labels. We shuffle sorted question identifiers with seed 2026 and assign 64 questions to each of seeds 0–31, without overlap. Each smaller set takes a prefix of that seed's question list, retaining all four claims per question. Data order sorts selected question identifiers and retains candidate order. For each seed and size, we use one arrangement per ordering condition, with shuffle seed equal to the seed index. Balanced random labels are assigned using seed 90,000 plus that index, separately at each size. Gold and random labels use identical claims and arrangements. The sweep contains $6 \times 3 \times 2 \times 32 = 1{,}152$ measurements.

Prompt and scoring. We use Llama-3.1-8B-Base and the verification template of Wen et al. (2025):

```
Question: [question]
Claim: [candidate solution]
I think this claim is [True/False]
```

Context claims include their assigned labels and are separated by blank lines; the target ends before its label. We read next-token probabilities, renormalize over True and False, and take the natural log of the assigned label's probability. No labels are optimized during the sweep. Both scores use the same arrangement, retaining its order when the target is removed. We omit the first claim from both sums because it has no preceding context, comparing the same $|\mathcal{S}| - 1$ claims. The omitted first claim is still available as context for later claims.

Scoring checks. We request the top 64 token probabilities. When a label is absent, we obtain both label probabilities by appending each label and reading its log-probability with an echo request; this fallback occurs four times. As a consistency check, the last claim's two scores must agree within 10^{-3} nats, since their prompts are identical. All 1,152 measurements agree exactly on this check.

Gap definitions. Coherence conditions on preceding claims; mutual predictability conditions on all other claims. The mean gap is $(f_{\mathrm{MP}} - \chi)/(|\mathcal{S}| - 1)$, in nats per scored claim; the relative gap is $(f_{\mathrm{MP}} - \chi)/\min(|f_{\mathrm{MP}}|, |\chi|)$.

Uncertainty. The seed is the unit of replication. We average claims within a seed before calculating means and standard errors across the 32 seeds; comparisons across sizes or orderings are paired by seed. The relative gap is also calculated within each seed before averaging.

G Further Related Work

We position our work relative to four threads: scalable oversight, self-improvement methods, unsupervised elicitation, and theoretical accounts of why feedback-free methods improve accuracy.

Scalable oversight and debate. Scalable oversight seeks mechanisms for supervising models whose capabilities may exceed those of any individual human evaluator (Christiano et al., 2018; Bowman et al., 2022). Debate was proposed as one such mechanism: agents argue before a judge, making it easier for the judge to identify the correct answer than to generate it independently (Irving et al., 2018). Du et al. (2023) demonstrate that multiagent debate yields substantial accuracy gains on reasoning benchmarks through iterative exchange of solutions between model copies. Khan et al. (2024) show that adversarial debate enables weaker judges to identify correct answers when evaluating stronger debaters. More broadly, Burns et al. (2023) study weak-to-strong generalization, finding that strong pretrained models consistently outperform weak supervisors when fine-tuned on weak labels, because "[for] the purposes of alignment we do not need the weak supervisor to teach the strong model new capabilities; instead, we simply need the weak supervisor to elicit what the strong model already knows." These works evaluate debate's *outputs* (accuracy, judge agreement) but do not identify or manipulate the objective that debate optimizes internally. We contribute the latter: by tracking coherence across debate rounds and suppressing it via rejection sampling (Appendix E, Algorithm 2), we provide evidence about the mechanism driving debate's accuracy gains.

Self-improvement without external feedback. A family of methods improve model performance using self-generated signal. Pseudo-labeling (Lee, 2013) trains on a model's own confident predictions. Self-consistency (Wang et al., 2022) samples a diverse set of reasoning paths and selects the most consistent answer by marginalizing over the sampled paths. STaR (Zelikman et al., 2022) iteratively trains on self-generated rationales that yield correct answers, bootstrapping reasoning ability from a small number of examples. Self-Refine (Madaan et al., 2023) uses a single model to iteratively critique and revise its own outputs. Canavan et al. (2026b;a) show that bootstrap recovers 83–100% of the gap between zero-shot and many-shot golden-label prompting on datasets of $\sim$2k examples. Each method has been evaluated individually with method-specific explanations for its success. We test the prediction of Qiu et al. (2026) that these methods share a common objective (coherence) by measuring it across multiple methods and testing its causal role via ablation.

Unsupervised elicitation. A related line of work asks whether model capabilities can be elicited without any supervision. Contrast Consistent Search (Burns et al., 2022) finds directions in activation space that satisfy logical consistency properties (e.g., a statement and its negation have opposite truth values), operating on internal representations rather than model outputs. Wen et al. (2025) introduce ICM, which extends beyond consistency to optimize mutual predictability: each label should be predictable given all others. ICM matches or exceeds supervised performance on several benchmarks. We compare ICM's mutual-predictability score with coherence on fixed labeled claim sets to test their conditional relationship (Section 5.3).

Theoretical accounts. What explains the success of feedback-free self-improvement? Huang et al. (2024) analyze self-improvement as a sharpening mechanism that tilts the base model toward responses with higher self-reward (sequence-level likelihood), framing the problem as computational extraction of existing knowledge. Their key result is that SFT on best-of-N filtered data is minimax optimal for this objective. Qiu et al. (2026) propose a different account: these methods optimize coherence (the joint predictability of behaviors under the model's prior), and show that coherence regularization is optimal for semi-supervised learning with a pretrained prior. The two accounts

address different levels of description: sharpening characterizes how each prompt's conditional distribution is tilted toward high-likelihood responses, while coherence identifies the cross-context objective governing the joint policy. Both assume the model already possesses latent knowledge that self-improvement extracts. Our work provides a systematic empirical test of the coherence account across multiple methods, with rejection interventions that neither theoretical paper includes.