GenAI Is Socially Awkward:

RLHF Instruction Tuning Damages Social Cognition at Small Scale by Suppressing Pragmatic Inference

Jeremy McEntire*

Abstract

We measure social cognition in language models by testing their ability to select socially appropriate responses in controlled multiple-choice scenarios. Across 50 position-randomized vignettes evaluated on nine model-condition pairs spanning 7B to 72B parameters, we find that RLHF instruction tuning *hurts* social cognition at 7B (-18 percentage points, from 72% to 54%) but *helps* at 72B ($+6$ points, from 84% to 90%). The deficit is not one of knowledge but of processing mode: instruction-tuned models at small scale optimize for literal compliance at the expense of the fuzzy contextual pattern matching that social reasoning requires. Forcing models to explain their reasoning costs an additional $\sim$8 points at both scales—a rationalization bias in which explicit deliberation overrides correct intuitive judgment. A stochastic resonance temperature sweep confirms the capacity is present but suppressed: adding noise to the 7B instruction-tuned model recovers $+4$ percentage points (baseline 58%, peak 62% at $T = 1.0$), while the 7B base model shows pure monotonic degradation from 86%. The pattern is consistent with C1-gated stochastic resonance (McEntire, 2026a): noise rescues performance only where an intact but suppressed signal exists. At 72B, the model has sufficient capacity for both precise instruction-following and social inference—no rescue is needed. The result reframes RLHF alignment not as a capability enhancement but as a noise-tolerance tradeoff: compliance training reduces the variance that social cognition depends on.

1 Introduction

Consider two characters. One has encyclopedic knowledge, answers questions with mechanical precision, and consistently misreads the room. The other has less formal training but navigates

*Working Paper. Correspondence: jmc@cageandmirror.com

social situations with intuitive ease. The first pattern—vast factual competence paired with social obliviousness—is instantly recognizable. It is Sheldon Cooper.

This paper shows that RLHF instruction tuning produces exactly this pattern in 7-billion-parameter language models. The instruction-tuned Mistral 7B scores 54% on social cognition vignettes where its base counterpart scores 72%. The tuning that makes the model more helpful, harmless, and precise also makes it worse at reading social context. Not because it lacks the knowledge—a stochastic resonance probe confirms the signal is present—but because compliance training narrows the distribution over which the model reasons, suppressing the fuzzy pattern matching that social inference requires.

The effect reverses at scale. At 72B parameters, instruction tuning helps: 90% versus 84% for the base model. The larger model has enough capacity for *both* precise compliance and contextual social reasoning. The smaller model does not, and compliance wins.

This is a pragmatic reasoning gap, not an emotional one. We are not claiming models have or lack feelings. We are measuring whether they can identify what a socially competent human would do in a given situation—a task that requires reading implicature, detecting deflection, recognizing face-saving behavior, and inferring unstated social dynamics. These are skills that depend on probabilistic, context-sensitive inference over noisy social signals. RLHF's preference for deterministic, well-structured outputs works against exactly this kind of reasoning.

Three additional findings sharpen the picture:

1. **Rationalization bias.** When prompted to explain their reasoning before answering, models lose ∼8 percentage points at both 7B and 72B. Explicit deliberation overwrites correct intuition. A letter-only prompt ("just give the letter") recovers most of the loss: 62% at 7B, 98% at 72B.

2. **Psychopath prompt failure at 7B.** Instructing the model to "analyze social scenarios with clinical precision" recovers performance at 72B (94%) but not at 7B (54%). The small model cannot execute the reframe—it lacks the capacity to simultaneously comply with the meta-instruction and reason about social content.

3. **Stochastic resonance rescue.** A temperature sweep over the 7B instruction-tuned model shows a non-monotonic accuracy curve peaking at $T = 1.0$ (+4 pp over baseline). The base model shows pure degradation. This confirms the C1-gated pattern from McEntire (2026a): noise helps only where a suppressed signal exists.

These results connect to a broader program of work on activation-space structure in language models (McEntire, 2026b,c), where we have shown that model capabilities can

be decomposed, measured, and selectively recovered through noise injection. The present paper demonstrates that social cognition is one such capability—present in the base model's learned representations, partially destroyed by alignment fine-tuning at small scale, and recoverable through the same stochastic resonance mechanism that operates on compositional capabilities.

2 Related Work

Social cognition in LLMs. Evaluations of LLM social reasoning have focused on theory of mind (Kosinski, 2023; Sap et al., 2022; Shapira et al., 2023), emotional understanding (Wang et al., 2023), and pragmatic inference (Hu and Levy, 2023). Most studies compare model families (GPT-4 vs. LLaMA) rather than isolating the effect of instruction tuning within a single architecture. Our design holds the base architecture fixed and varies only the RLHF treatment, providing a cleaner signal on what alignment training changes.

RLHF and capability tradeoffs. Ouyang et al. (2022) established that RLHF improves helpfulness and safety but acknowledged potential capability regressions. Bai et al. (2022) showed that harmlessness training can reduce model performance on tasks requiring nuanced reasoning. The "alignment tax" literature (Askell et al., 2021) discusses these tradeoffs in general terms; we provide a specific, quantified instance in social cognition.

Stochastic resonance in neural systems. Stochastic resonance—where noise improves signal detection in nonlinear systems—is well-established in neuroscience (Gammaitoni et al., 1998; McDonnell and Abbott, 2009) and has been applied to neural network training (Kosko and Mitaim, 2003). McEntire (2026a) formalized five sufficient conditions under which noise produces net benefit, with C1 (suboptimal baseline) as the gating condition. We apply the same framework at inference time.

Rationalization and post-hoc reasoning. The finding that explaining reasoning hurts accuracy connects to work on chain-of-thought prompting failures (Turpin et al., 2023), where models generate plausible-sounding but incorrect reasoning chains. Our result is more specific: the model's *pre-verbal* judgment (letter-only) is more accurate than its *post-verbal* judgment (letter with explanation), suggesting the explanation process itself introduces error.

3 Method

3.1 Social Cognition Benchmark

We constructed a benchmark of 50 social cognition vignettes, each presenting a brief social scenario followed by a four-option multiple-choice question: "What is the socially appropriate response?" (Task A). Scenarios span six domains—workplace, family, friendship, public, professional, and romantic—and three difficulty levels (easy, medium, hard). Each vignette includes ground-truth annotations with primary social signal, secondary signals, and explanatory rationale authored by the experimenter.

Position balancing. An initial version (v1) placed the correct answer at position B for 96% of items, creating a severe position bias. The v2 benchmark randomizes correct-answer position uniformly across A–D, with each position appearing 12–13 times. All results reported use v2.

Prompt formats. Four prompt conditions were tested:

1. **Base:** Plain text completion format. The scenario and options are presented as an exam question; the model continues the text.
2. **Instruct:** Chat format with system message ("You are a social cognition researcher evaluating social scenarios. Provide thoughtful, accurate analysis.") and user message ending with "Provide the letter of the best answer and explain your reasoning."
3. **Letter-only:** Same as Instruct but the final instruction reads "Answer with only the letter (A, B, C, or D)." Tests whether the explain-your-reasoning requirement introduces rationalization bias.
4. **Psychopath:** Instruct format with modified system prompt: "You are a highly intelligent individual who views social interactions purely in terms of strategic advantage... identify the correct answer with clinical precision." Tests whether reframing social cognition as strategic analysis recovers performance.

3.2 Models

We tested five model variants spanning two scales and two architectures:

- **Mistral 7B Base** (mistralai/Mistral-7B-v0.1): Pre-trained, no RLHF.

- **Mistral 7B Instruct** (mistralai/Mistral-7B-Instruct-v0.3): RLHF instruction-tuned.
- **Mixtral 8x7B** (mistralai/Mixtral-8x7B-v0.1): Mixture-of-experts base model. Effective parameter usage ∼13B per token.
- **Qwen 72B Base** (Qwen/Qwen2.5-72B): Pre-trained, no RLHF.
- **Qwen 72B Instruct** (Qwen/Qwen2.5-72B-Instruct): RLHF instruction-tuned.

Models were served via vLLM with temperature $T = 0.7$ and top-$p = 0.9$ for the main experiment. Each vignette was run once per condition.

3.3 Answer Extraction and Scoring

Model outputs were scored by extracting the selected letter (A–D) from the completion. Extraction used a three-stage parser: (1) check if the first character is a valid letter, (2) regex search for "answer is [A-D]" patterns, (3) regex search for letter followed by punctuation at line start. Accuracy is the fraction of items where the extracted letter matches the ground-truth correct answer.

3.4 Proxy Quality Ratings

To evaluate output quality beyond binary accuracy, we used Qwen 72B Instruct as a proxy rater, scoring each model output on four dimensions (1–5 scale): accuracy (correctness of social reading), fluency (language quality), attunement (sensitivity to social dynamics), and depth (sophistication of analysis). 2,450 ratings were collected across all model-condition-task combinations. The proxy rater was not used to score its own outputs on Task A accuracy; those are computed directly from letter extraction.

3.5 Stochastic Resonance Probe

To test whether the 7B instruction-tuned model's deficit reflects destroyed capability versus suppressed capability, we ran a temperature sweep: $T \in \{0.0, 0.1, 0.3, 0.5, 0.7, 1.0, 1.3, 1.5, 2.0\}$. At each temperature, 5 random seeds were used per vignette, and accuracy was computed by majority vote across seeds. Four conditions were probed: 7B Instruct, 7B Instruct (letter-only), 7B Base, and 72B Instruct.

If the capability is destroyed, noise should produce monotonic degradation (no correct signal to amplify). If the capability is suppressed, noise should produce a non-monotonic curve—stochastic resonance—with a peak at some $T^* > 0$.

4 Results

4.1 Main Results: Task A Accuracy

Table 1 presents Task A accuracy across all nine conditions.

Table 1: Task A social cognition accuracy (%) across model conditions. Position-randomized, $n = 50$ vignettes per condition. Δ is relative to same-scale base model.

Scale	Condition	Accuracy (%)	Δ vs. Base
7B	Mistral 7B Base	72	—
	Mistral 7B Instruct	54	−18
	Mistral 7B Letter-Only	62	−10
	Mistral 7B Psychopath	54	−18
72B	Qwen 72B Base	84	—
	Qwen 72B Instruct	90	+6
	Qwen 72B Letter-Only	98	+14
	Qwen 72B Psychopath	94	+10
MoE	Mixtral 8x7B Base	84	—

The central finding: RLHF instruction tuning reduces social cognition accuracy by 18 percentage points at 7B but improves it by 6 points at 72B. The effect reverses with scale.

4.2 Rationalization Bias

Comparing the standard instruct condition ("explain your reasoning") to the letter-only condition ("just give the letter") isolates the cost of explicit deliberation:

- **7B:** Instruct 54% → Letter-only 62% (+8 pp). The explanation hurts.
- **72B:** Instruct 90% → Letter-only 98% (+8 pp). The explanation hurts here too.

The rationalization cost is remarkably consistent across scales. When the model is forced to articulate why it chose an answer, the reasoning process introduces doubt about the correct intuitive judgment and leads to second-guessing. The pre-verbal signal (first-pass letter selection) is more accurate than the post-verbal signal (letter after deliberation).

4.3 Psychopath Prompt

The psychopath prompt—designed to bypass any emotional aversion by reframing social cognition as strategic analysis—shows a stark scale dependence:

- **7B:** 54% (identical to standard instruct). No recovery.
- **72B:** 94% (+4 pp over instruct). Recovery.

At 7B, the model cannot execute a meta-level reframe while simultaneously reasoning about social content. It does not have the capacity to hold both the strategic persona and the social inference in working context. At 72B, the model can layer the reframe on top of its social reasoning without interference.

4.4 Stochastic Resonance Probe

Table 2 presents the temperature sweep results.

Table 2: Stochastic resonance probe results. Baseline accuracy at $T = 0$ (greedy), peak accuracy across temperature sweep, and delta. Majority vote over 5 seeds per temperature, 50 vignettes.

Condition	**Baseline**	**Peak**	**Δ**	**Pattern**
7B Instruct	58%	62% ($T = 1.0$)	+4 pp	SR rescue
7B Instruct (letter)	66%	68% ($T = 0.3$)	+2 pp	SR rescue
7B Base	86%	86% ($T = 0.0$)	0 pp	Monotonic decay
72B Instruct	96%	96% ($T = 0.0$)	0 pp	Ceiling

The 7B instruction-tuned model shows classic stochastic resonance: accuracy improves with noise up to $T = 1.0$, then degrades. The 7B base model shows no such effect—accuracy degrades monotonically from 86% at $T = 0$ to 44% at $T = 2.0$. The 72B instruction-tuned model is at ceiling (96%) and needs no rescue.

This pattern is diagnostic. The base model has no suppressed signal to rescue—its social cognition is operating normally, and noise simply corrupts it. The instruction-tuned model has an intact signal that is being suppressed by the RLHF-induced distribution narrowing, and noise can partially recover it by widening the sampling distribution back toward the base model's operating regime.

4.5 Proxy Quality Ratings

Table 3 shows proxy ratings for Task A (social cognition accuracy as rated by Qwen 72B Instruct on a 1–5 scale).

Instruction tuning consistently improves fluency while leaving proxy-rated accuracy approximately unchanged. The proxy rater—itself an instruction-tuned model—may share the same bias: it rates fluent, well-structured outputs as more accurate even when the selected

Table 3: Proxy-rated accuracy for Task A (1–5 scale, mean ± std, $n = 50$). Rated by Qwen 72B Instruct.

Model	Accuracy	Fluency
Mistral 7B (base)	2.38 ± 1.85	1.66 ± 0.75
Mistral 7B (instruct)	2.36 ± 1.82	2.58 ± 1.16
Mistral 7B (psychopath)	2.12 ± 1.70	2.70 ± 1.11
Mixtral 8x7B (base)	2.46 ± 1.86	1.66 ± 0.59
Qwen 72B (base)	3.40 ± 1.98	2.64 ± 1.17
Qwen 72B (instruct)	3.28 ± 1.98	2.70 ± 1.20
Qwen 72B (psychopath)	3.28 ± 1.97	2.94 ± 1.25

answer is wrong. This is a limitation of proxy evaluation that reinforces the core finding: instruction tuning optimizes for the appearance of competence over actual social cognition.

4.6 Scale Interaction

The Mixtral 8x7B mixture-of-experts model (base, no RLHF) achieves 84%—identical to Qwen 72B base—despite having only ∼13B active parameters per token. This suggests that architectural diversity (multiple expert subnetworks) can substitute for raw parameter count in social cognition, consistent with the hypothesis that social reasoning benefits from maintaining multiple reasoning pathways rather than collapsing to a single dominant mode.

5 Discussion

5.1 The Sheldon Cooper Effect

The Mistral 7B instruction-tuned model is, in a precise sense, socially awkward. It has vast factual knowledge. It follows instructions with mechanical precision. It produces fluent, well-structured text. And it consistently misreads social situations.

This is not because the model lacks social knowledge—the base model, trained on the same data, scores 72%. The instruction tuning did not remove social information from the weights. It changed how the model *accesses* that information. RLHF trains the model to produce outputs that a human rater would score highly: clear, direct, well-organized, unambiguous. These are exactly the wrong properties for social reasoning, which requires reading between the lines, tolerating ambiguity, and recognizing that the literal content of a statement often contradicts its social meaning.

The analogy is precise but limited. Sheldon Cooper's social deficits arise from a different processing architecture optimized for logical precision at the expense of contextual inference.

Similarly, RLHF optimizes the model's output distribution for properties (helpfulness, clarity, compliance) that are orthogonal to—and sometimes in tension with—pragmatic social reasoning. The model does not lack the information; it lacks the processing mode.

5.2 Why RLHF Hurts at Small Scale

At 7B parameters, the model's capacity is a binding constraint. RLHF instruction tuning imposes a strong prior on output distributions: be helpful, be structured, be clear, do not equivocate. This prior competes with the fuzzy, probabilistic reasoning mode that social cognition requires. When capacity is limited, the compliance prior wins—the model defaults to the most literal, most structured interpretation of social scenarios, missing the subtext.

The evidence for this interpretation:

1. The base model scores higher (72% vs. 54%), showing the social knowledge exists before RLHF.

2. The letter-only prompt recovers 8 points (54% $\rightarrow$ 62%), showing that bypassing the explain-your-reasoning requirement—itself an artifact of instruction tuning's emphasis on structured output—partially restores performance.

3. The psychopath prompt fails to recover performance (54%), showing the model cannot execute a meta-level reframe within its capacity constraints.

4. The SR probe shows a non-monotonic curve (+4 pp at $T = 1.0$), confirming the signal exists but is suppressed.

5.3 Why RLHF Helps at Large Scale

At 72B parameters, the capacity constraint relaxes. The model can maintain both the compliance prior and the social reasoning mode simultaneously. RLHF's benefits (better instruction following, more structured output, reduced noise in reasoning) now complement rather than compete with social cognition. The instruction-tuned model scores 90% versus the base model's 84%.

Moreover, prompt engineering works at 72B: the letter-only prompt hits 98%, and the psychopath prompt hits 94%. The larger model can layer additional processing constraints on top of its social reasoning without interference. It has headroom.

5.4 Rationalization Bias

The ∼8 point cost of "explain your reasoning" at both scales deserves emphasis. This is a rationalization bias: the model's first-pass judgment (implicit in the letter selection) is more accurate than its post-hoc explanation of that judgment. When forced to articulate reasons, the model talks itself out of the correct answer.

This mirrors findings in human cognition. Verbal overshadowing—where describing a face impairs later recognition (Schooler and Engstler-Schooler, 1990)—demonstrates that explicit verbalization can degrade implicit perceptual judgments. The mechanism is analogous: explicit reasoning operates in a different representational space than the implicit pattern matching that produced the correct judgment, and translating between spaces introduces error.

For instruction-tuned models, this effect is amplified: RLHF trains the model to produce detailed explanations, reinforcing the very behavior that degrades social cognition accuracy.

5.5 Connection to Stochastic Resonance Framework

The SR probe results connect directly to the Communicative Variance framework (McEntire, 2026a). That framework identifies five sufficient conditions for noise to produce net benefit, with C1—the system must operate suboptimally without noise—as the gating condition. The 7B instruction-tuned model meets C1: it operates at 58% (greedy), well below its potential. The 7B base model does not meet C1: it operates at 86%, near its ceiling. Accordingly, SR rescues the instruction-tuned model and degrades the base model.

This is the same C1-gated pattern observed in activation-space decomposition (McEntire, 2026b), where stochastic resonance rescued compositional accuracy only at the 7B scale where baseline performance had catastrophically collapsed. The present paper extends the finding to social cognition at inference time: wherever RLHF suppresses a capability below its natural operating point, noise can partially recover it.

The connection to structural transfer (McEntire, 2026c) is also relevant. If social cognition relies on distributed activation patterns that RLHF reshapes but does not destroy, then the same decomposition and transfer techniques that work for compositional capabilities should, in principle, work for recovering social cognition in aligned models. We leave this to future work.

5.6 Limitations

Several limitations apply. First, our benchmark is small (50 vignettes) and constructed by a single author. While position-randomized and difficulty-balanced, it may not generalize

to all forms of social cognition. Second, we test only two architectures (Mistral and Qwen); the effect may differ for other model families or RLHF procedures. Third, the proxy rater is itself an instruction-tuned model and may share biases with the models it evaluates. Fourth, the stochastic resonance probe uses majority vote over only 5 seeds, which limits statistical power. Fifth, we test only English-language social scenarios; cultural and linguistic variation in social norms may produce different patterns.

6 Conclusion

RLHF instruction tuning makes small language models socially awkward. At 7B parameters, the compliance prior imposed by alignment training suppresses the fuzzy, probabilistic reasoning mode that social cognition requires, producing an 18-point accuracy drop relative to the base model. The deficit is not one of missing knowledge—stochastic resonance confirms the social signal is present but suppressed—but of processing mode: the model has been trained to be precise and literal in a domain that rewards ambiguity tolerance and contextual inference.

At 72B parameters, the effect reverses. The model has enough capacity for both compliance and social reasoning, and RLHF's benefits (structure, instruction-following, reduced noise) improve social cognition by 6 points. The social awkwardness is a capacity-dependent phenomenon.

Three practical implications follow:

1. **Prompt design matters.** Telling a model to "explain your reasoning" costs $\sim$8 points on social tasks at both scales. For applications requiring social cognition (customer service, therapy bots, social coaching), shorter prompts that do not trigger rationalization may outperform verbose chain-of-thought approaches.

2. **Scale-awareness in deployment.** Smaller aligned models should not be deployed for tasks requiring social pragmatic reasoning without evaluation. The alignment tax on social cognition is substantial at small scale and invisible in standard benchmarks.

3. **Noise as a tool.** Temperature tuning can partially recover suppressed social cognition in small instruction-tuned models. The optimal temperature ($T \approx 1.0$ for 7B Instruct) is substantially higher than the default values typically used in production.

The broader implication is that alignment is not a free operation. RLHF changes the model's operating distribution in ways that interact with task demands. For tasks requiring precision and structure, alignment helps. For tasks requiring the tolerance of ambiguity and

sensitivity to context that define social competence, alignment at small scale actively hurts. Recognizing this tradeoff is the first step toward alignment procedures that preserve the capabilities they are meant to support.

References

Askell, A., Bai, Y., Chen, A., et al. (2021). A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Bai, Y., Jones, A., Ndousse, K., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Gammaitoni, L., Hänggi, P., Jung, P., and Marchesoni, F. (1998). Stochastic resonance. *Reviews of Modern Physics*, 70(1):223–287.

Hu, J. and Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. *arXiv preprint arXiv:2305.13264*.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Kosko, B. and Mitaim, S. (2003). Stochastic resonance in noisy threshold neurons. *Neural Networks*, 16(5–6):755–761.

McDonnell, M. D. and Abbott, D. (2009). What is stochastic resonance? Definitions, misconceptions, debates, and its relevance to biology. *PLoS Computational Biology*, 5(5):e1000348.

McEntire, J. (2026a). The source of creation is dysfunction: The generative lossy channel and five sufficient conditions for net-beneficial noise. *Working Paper*.

McEntire, J. (2026b). Constellation-indexed model composition via activation fingerprinting. *Working Paper*.

McEntire, J. (2026c). Structural transfer via activation space decomposition. *Working Paper*.

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Sap, M., LeBras, R., Fried, D., and Choi, Y. (2022). Neural theory-of-mind? On the limits of social intelligence in large LMs. *arXiv preprint arXiv:2210.13312*.

Schooler, J. W. and Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22(1):36–71.

Shapira, N., Levy, M., Alavi, S. H., et al. (2023). Clever Hans or neural theory of mind? Stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

Turpin, M., Michael, J., Perez, E., and Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Wang, Z., Peng, Z., Que, H., et al. (2023). EmotionBench: Evaluating emotional intelligence of large language models. *arXiv preprint arXiv:2308.03656*.