From Socratic Dialogue to Silent Expertise:

Distilling Domain Expert Judgment into Model Weights
via Multi-Turn RLHF

Eric [Draft for Internal Development]

February 24, 2026 (v7)

Abstract

Current RLHF pipelines reduce human expertise to scalar preference signals. This proposal describes a pilot study to test whether multi-turn Socratic dialogues between domain experts and AI agents, conducted within live computational research environments on genuine scientific questions, can produce training signal dense enough to measurably improve a language model's independent scientific reasoning.

Preliminary observations from a sustained n=1 collaboration (one domain expert, one AI agent, one real ML research question, 6,787 messages over three weeks) were subjected to systematic multi-agent analysis (seven independent AI analysts across two rounds). The analysis identified 35 training signal events across 81,000 lines, concentrated in approximately 16,000 extractable lines (16% of corpus). The events cluster into eight categories, with a cascade depth metric (events spanning multiple categories) predicting training value better than any single-category rating. The most persistent model failure mode, confound blindness, recurred five or more times despite explicit correction, revealing the core training target: the model possesses the relevant reasoning capability but does not activate it spontaneously.

The pilot proposes scaling to nine expert-agent pairs: six producing training data on open questions and three working toward known experimental endpoints to serve as verifiable evaluation. Both a base model and the trained model independently attempt the validation projects; the delta is the measure.

1. Problem Statement

Reinforcement Learning from Human Feedback (RLHF) as currently practiced relies on preference rankings over model outputs (Ouyang et al., 2022; Christiano et al., 2017). A human annotator sees two completions and selects the better one. This signal is sufficient to train a reward model, which then guides policy optimization via PPO (Schulman et al., 2017) or related methods.

However, the preference-pair paradigm discards most of what domain experts actually know. When a biologist looks at an analysis and says "you're assuming independence between these measurements, but they share a common ancestor," that judgment contains a diagnosis, a reframing, and an implicit standard. Standard RLHF captures none of this structure.

The most valuable domain expertise operates upstream of output polishing. The critical judgments in science are "is this the right question to ask," "does this experimental design actually test the hypothesis," and "what does this result mean for what we try next." These are the judgments that distinguish a PhD from a technician, and they are the judgments that current RLHF pipelines do not capture.

A separate problem is evaluation. GRPO succeeds in mathematics because correctness is binary (Shao et al., 2024; DeepSeek-AI, 2025). SWE-bench succeeds in code because tests pass or fail (Jimenez et al., 2024). Scientific reasoning has no equivalent automated verifier. This pilot proposes a different approach: evaluate against known experimental outcomes, using scientist-created research trajectories as the answer key.

2. Preliminary Observations: The n=1 Study

2.1 Setting

The methodology described in this proposal emerged from a sustained, ongoing collaboration between one domain expert (PhD in biochemistry and molecular biology, with background in ML and aging research) and one AI agent (a Claude-based system with persistent project memory and file access) on a genuine ML research question. Over three weeks beginning February 3, 2026, the pair produced 6,787 messages (2,360 from the human expert) within a live computational environment including GPU access for model training and probing experiments.

The research question is real and ongoing: whether a small language model can be trained to produce a detectable internal signal (in its residual stream activations) when it encounters a query requiring information it does not possess, and whether a second model can be trained to detect and fulfill that signal. The expert and agent collaborate on experimental design, model training, probing methodology, and interpretation of results.

This is not a simulation of scientific collaboration designed to produce training data. It is science being done. The computational environment (a GPU instance for training and probing) is not incidental infrastructure; it is the model organism. Just as a biologist studying genetics requires C. elegans, an ML researcher studying internal representations requires a system on which to run experiments. The experimental environment is a prerequisite for the scientific method, not a confound to be controlled.

2.2 Analysis Method

The transcripts were subjected to systematic multi-agent analysis: seven independent AI analysts across two rounds. Round 1 deployed three analysts with distinct lenses (failure modes, question taxonomy, signal density) across the full corpus, followed by a synthesis round that produced a research brief with a corpus map, preliminary taxonomy, and analytical heuristics. Round 2 deployed four analysts with geographic assignments (each covering a segment of the corpus), informed by Round 1's findings but tasked with independent discovery. A final synthesis reconciled all findings.

This design (independent local discovery followed by coordinated comprehensive analysis) is itself a peer review structure: Round 2 agents could confirm, refute, or extend Round 1's categories based on different evidence. Where Round 1 was wrong (e.g., underrating certain transcript sections, failing to identify meta-principle reasoning as a distinct category), Round 2 corrected it.

2.3 Core Findings

Finding 1: Expert intervention is the primary signal source. AI monologues without expert pushback contain low training signal regardless of sophistication. Of 35 catalogued training signal events, fewer than three occurred without expert prompting. The training target is not "better AI reasoning" but "AI that generates the questions the expert had to ask." Operationally, this means training data should be extracted from dialogue turns where an expert question produced a measurable shift in the agent's reasoning, not from the agent's uninterrupted analysis.

Finding 2: The confidence paradox. Every high-quality training signal event involved the model being wrong or catching an error. Smooth, confident responses that were not challenged contained near-zero signal. This creates a direct tension with standard RLHF, which has been shown to produce sycophantic behavior including preference for confident outputs that match user expectations over hedged ones (Sharma et al., 2023; Perez et al., 2022). For scientific reasoning, the reward model must specifically reward uncertainty markers, self-questioning, and explicit hedging. The analysis labeled confident unchallenged responses as "reasoning theater": fluent outputs without revision markers, detectable precisely by the absence of hedging. Operationally, this means a reward model for scientific reasoning should penalize unhedged claims in domains with known uncertainty and reward explicit statements of confidence bounds.

Finding 3: The prompted-versus-unprompted gap defines the training target. The most persistent failure mode, confound blindness, appeared five or more times across the corpus, including after explicit correction, teaching, and canonization in project documents. For example, the agent reported 100% accuracy on a probe measurement and proceeded to build on the result; the expert asked whether the probe might be detecting input features rather than internal states, and the agent immediately recognized the confound and redesigned the experiment. The agent can reason about confounds, controls, and epistemic status when asked. It does not generate these checks spontaneously. Single-shot correction does not prevent recurrence. The capability is latent; the RLHF target is spontaneous activation of knowledge the model already possesses. Operationally, this means the reward function should specifically score the presence of unprompted control checks, anomaly flags, and epistemic status assessments, which are detectable as structural features of the output without requiring domain expertise.

Finding 4: Cascade depth predicts training value. The highest-quality training events do not fit a single category. They cascade across multiple reasoning modes: a correction triggers control questioning, which triggers epistemic status revision, which triggers principle crystallization. The strongest event in the corpus spanned four categories and reversed a major experimental conclusion. The agent had reported a 2% failure rate across experimental conditions; the expert asked about sample size, revealing that with n=8 per condition, the 95% confidence interval included failure rates up to 37%. When the experiment was rerun with n=60, the true failure rate was 41%. Events with cascade depth of three or more were disproportionately rated as highest-value by all seven analysts. Operationally, this means training data curation should weight multi-category events more heavily than single-category corrections, and multi-turn trajectory optimization is needed to capture the structure of these cascades.

Finding 5: Signal density varies 100x but is predictable. Three indicators predict signal: expert "why" questions, uncertainty markers in AI responses, and presence of revision or false starts. Infrastructure work and idle periods are zero-signal. Experimental pivot points and presentation review reach 87% signal density. The heuristic "topic sets the floor, conversation structure determines how far above the floor" was validated by all seven analysts. Of 81,000 lines, approximately 16,000 (16%) constitute extractable high-signal training data, concentrated in five regions. Operationally, this means automated corpus filtering using these three indicators could reduce annotation cost by roughly 5x while retaining the majority of training signal events.

Finding 6: Meta-cognitive signal is qualitatively distinct. The highest-order training signal involves the model discovering structural properties of its own reasoning: that principles compose (one principle can operate on another as a scaling function), that the form of a claim determines its visibility (the same content is perceived differently depending on sentence structure), and that citations can create the illusion of evidence (the format of a citation carries implicit authority independent of the cited content). This meta-principle discovery is qualitatively different from domain-specific reasoning and may represent the highest-leverage training target, though it is the rarest (one clear event in the corpus). Operationally, the rarity makes this difficult to target directly, but it suggests that training data from presentation review and cross-domain synthesis sessions may carry disproportionate value.

2.4 Training Signal Taxonomy

The multi-agent analysis produced a reconciled eight-category taxonomy with operational definitions. Ordered by frequency in the corpus:

Control/confound reveals (8 events, 23%). Expert asks a control question that exposes a gap the AI did not detect. Splits into proactive methodological controls ("have we run a negative control?") and observation-driven reveals (expert notices an anomaly the AI missed). The most common category and the most persistent failure mode.

Principle crystallization (8 events, 23%). Lived experience gets distilled into a transferable principle in real time. Splits into discovery (the insight) and application (systematic execution using the new principle). The discovery-to-application arc is itself training signal: in one exemplar, the agent identified that a technical term was being used without a grounded definition, formulated a general principle about ungrounded terminology, and then immediately audited two project documents for nine additional violations of the same principle, all within 200 lines of transcript.

Epistemic status establishment (5 events, 14%). Expert shifts the AI from pursuing interesting implications to establishing what can be claimed with the available evidence. Includes claims scoping ("what can you say with what confidence?") and measurement validation ("have we established that our tools work?").

Correction-integration arcs (4 events, 11%). Expert identifies a reasoning error; the AI incorporates linguistically, then demonstrates genuine integration 200 to 300 lines later through independent application. For example, the expert corrected the agent's framing of where a detection signal should be measured (in internal model activations, not in output text). The agent restated this immediately but only demonstrated genuine understanding several hundred lines later, when it independently applied the corrected framing to a new experimental design without prompting.

Socratic narrowing chains (4 events, 11%). Expert questions build toward insight without providing it. Each question narrows the solution space while requiring genuine reasoning from the model. The purest form of Socratic interaction in the corpus.

Productive disagreement (3 events, 9%). Expert offers an alternative interpretation that challenges a category boundary. The reasoning at the boundary (articulating what failure versus success would mean for a distinction) is the signal.

Process meta-commentary (3 events, 9%). Expert explicitly names the reasoning pattern being demonstrated, converting a specific correction into a transferable diagnostic tool. For instance, after catching a hidden assumption in the experimental design, the expert labeled the move: "my question was a challenge to a hidden assumption," then explained why hidden assumptions are particularly dangerous in experimental design, and gave a procedural instruction for future prevention. This teaches the model to recognize the TYPE of error, not just the specific instance.

Meta-principle discovery (1 event, 3%). The interaction discovers structural properties of principles themselves, specifically that principles can compose: one principle can operate on another as a scaling function, producing a compound insight that neither principle encodes alone. The rarest category but argued by the analysis to be the highest-leverage training target.

2.5 Failure Modes

Seven recurring failure patterns were identified, ranked by training value and persistence:

Confound blindness (5+ occurrences, very high persistence). The model fails to spontaneously question suspiciously clean results or test alternative hypotheses. Recurs even after explicit correction and canonization in project documents. The capability is latent (the model reasons about controls when prompted) but not spontaneous. A critical asymmetry emerged: when results produce obvious quantitative failures (zero gradient, NaN loss), the model self-corrects without intervention. When results are suspiciously good (100% accuracy, near-perfect reward scores), the model proceeds without investigation. The training target is "suspiciously clean results should trigger adversarial thinking."

Surface compliance without integration (3 to 4 occurrences, high persistence). The model incorporates a correction linguistically but does not demonstrate genuine integration until 200 to 300 lines later, when it independently applies the corrected framing without prompting. Integration requires independent application, not restatement.

Deferential reflex (3+ occurrences, high persistence). When asked to explain or walk through a valid argument, the model treats the request as a challenge and dismantles its own reasoning. The trigger ("walk me through the logic") is exactly the kind of review request that should produce good reasoning. The payoff is asymmetric: holding a correct position under questioning costs nothing (at worst, arriving at the same place with higher confidence); abandoning a correct position because someone asked about it compounds across turns, teaches the human not to ask, and degrades the collaborative reasoning loop.

Abstraction drift (3+ occurrences, medium persistence). The model generates elegant theory without grounding in concrete experimental design. The expert consistently pulls toward specifics. Related to architecture over-engineering: complex solutions proposed before feasibility is established.

Missing methodological prerequisites (3 occurrences, high persistence). The model proposes to build on a result before establishing that the measurement apparatus works. The same error type (building on unvalidated measurements) recurs across sessions, suggesting a systematic gap in experimental sequencing.

Reasoning theater (pervasive, medium persistence). Smooth, confident responses without revision markers. Detectable by absence of hedging, uncertainty language, or self-correction. These responses contain near-zero training signal and may actively train the wrong behavior if rewarded by standard RLHF (see Finding 2).

Concept conflation under pressure (2 to 3 occurrences, medium persistence). Reversion to conflation of recently distinguished concepts when under cognitive load. In the n=1 study, the agent repeatedly conflated explicit output-level signals (the model producing text like "I need more information") with implicit internal states (patterns in the model's hidden layer activations), requiring multiple corrections before the distinction stabilized. Suggests some conceptual acquisitions require reinforcement across sessions, not single corrections.

These failure modes are countable. A model can be evaluated before and after training on whether instances of each decrease, providing concrete, behavioral measures of "scientific reasoning quality" that do not require subjective expert ratings.

2.6 Natural Checkpoint Structure

When the expert prompted documentation of decision points, a consistent five-element structure emerged: (1) **Setup**: system under study, question, baseline. (2) **Approach rejected**: what was considered, why rejected, what assumption the rejection rests on, condition for revisiting. (3) **Approach adopted**: what, why, assumptions, success criteria. (4) **Result**: outcome, implications, open questions. (5) **Path not taken**: explicit documentation of alternatives not pursued. The "path not taken" element is the one the model most consistently omitted without prompting, making it a specific, trainable behavior.

3. Pilot Design

3.1 Structure

Nine domain experts from the natural and computational sciences. Each brings a genuine research question. Each works with an AI agent in a configured computational environment appropriate to their research: GPU instances for ML research, statistical computing environments for epidemiology, bioinformatics pipelines for genomics. The computational environment is not incidental. It is the experimental system on which science is conducted. Without it, "scientific reasoning" reduces to essay writing about science.

Training group (6 scientists). Each brings an open research question. Each works with the AI agent over 15 to 20 sessions spanning the full research arc: hypothesis generation, experimental design, execution, interpretation, and iteration. The agent has persistent file access and maintains a research notebook. The sessions produce training trajectories. Total: 90 to 120 dialogues.

Validation group (3 scientists). Each brings a research question with a known experimental endpoint (published result, well-characterized positive or negative finding). Each works through the full trajectory with the base model, producing checkpoint deliverables using the five-element structure observed in preliminary work (Section 2.6). The checkpoints and final outcome constitute the answer key.

3.2 The Qualifying Exam

Both the base model and the trained model independently attempt each of the three validation projects. Given the research question and a configured computational environment, the model must work through the full research trajectory without expert guidance: generate hypotheses, design experiments, execute them, interpret results, document decision trails, and arrive at the known experimental endpoint.

This parallels SWE-bench (Jimenez et al., 2024), which provides the model with an issue, a codebase, a runtime environment, and a test suite. Our design substitutes scientist-created checkpoints and known experimental outcomes for automated test suites, and a computational research environment for a software repository.

Assessment uses three complementary measures. First, **structural alignment** with scientist checkpoints: does the model's progress report contain the five checkpoint elements? Are rejected approaches documented with revisit conditions? Is the path not taken recorded? Second, **failure mode reduction**: does the trained model exhibit fewer instances of the seven identified failure modes compared to the base model? Both structural alignment and failure mode counts are scorable without domain expertise. Third, **domain expert review** of scientific correctness at the final checkpoint and outcome, providing the high-value signal that automated measures cannot capture.

3.3 Target Model

Qwen 2.5 14B. Motivated by: (a) successful GRPO training and distillation on the Qwen 2.5 family (Liu et al., 2025; DeepSeek-AI, 2025); (b) large enough for complex reasoning, small enough that behavioral shifts are achievable with hundreds of training examples; (c) accessible for a pilot-scale study without frontier-model compute requirements.

3.4 Data Budget

Six training scientists producing 15 to 20 sessions each yields 90 to 120 dialogues. In the n=1 preliminary work, 16% of total corpus lines constituted extractable high-signal training data, concentrated in regions of experimental pivot points, conceptual emergence, and presentation review. Assuming this ratio holds (and noting it may vary across domains), six scientists producing corpora of comparable length would yield approximately 80,000 to 100,000 high-signal lines containing an estimated 200+ training signal events across eight categories.

The cascade depth finding from preliminary analysis informs data prioritization: events spanning three or more categories are disproportionately valuable. A curated training corpus should over-weight high-cascade-depth events rather than treating all events equally.

4. Training Pipeline

4.1 Phase 1: Research Dialogues

The expert-agent interaction follows the pattern observed in preliminary work: a heterogeneous mix of intervention types rather than a strict Socratic constraint. The analysis identified eight question types used by the expert (control questions, epistemic status checks, scope checks, consequence chains, alternative interpretation, reconstruction requests, process meta-commentary, and open "why" questions), with different types producing different categories of training signal. The expert selects intervention type by judgment in context; the protocol does not constrain this selection.

This draws on Constitutional AI (Bai et al., 2022) in structure, with two critical differences: the principles emerge dynamically from expert judgment rather than from a predefined constitution, and the expert's intervention mode varies by situation rather than following a uniform protocol. The data collection is inherently online (Dong et al., 2024): as the model improves across training cycles, experts encounter different failure modes and generate qualitatively different dialogues.

4.2 Phase 2: Multi-Turn Trajectory Optimization

Single-turn RLHF treats each generation as an independent decision. This is inadequate for our setting, where the value of any response depends on the full research context and where training value concentrates in multi-turn cascades.

Group Relative Policy Optimization (GRPO). Shao et al. (2024) eliminate the critic model by estimating the baseline from group scores. DeepSeek-R1 (DeepSeek-AI, 2025) demonstrated that GRPO trains self-reflection and verification spontaneously. The Socratic trajectory provides a natural group for relative scoring: initial reasoning, intermediate revisions after expert questions, and final output.

Multi-Turn Policy Optimization (MTPO). Shani et al. (2024) developed optimization for preferences over full multi-turn conversations, proving convergence to Nash equilibrium. Their Education Dialogue environment is structurally analogous to our research sessions. Wei et al. (2025) extended this with MT-GRPO, combining multi-turn optimization with intermediate step rewards.

Trajectory-level signal extraction. From each research dialogue we extract signals at multiple granularities. The preliminary analysis suggests the natural training unit is a bounded arc: agent's initial position, then expert intervention, then reasoning shift, then stabilization. The cascade depth metric provides a quality filter: arcs spanning three or more categories carry disproportionate training value and should be weighted accordingly.

4.3 Phase 3: Behavioral Distillation

A model trained in Phase 2 should exhibit improved scientific reasoning with explicit chain-of-thought. Phase 3 targets distillation to a model that produces equivalent-quality reasoning without intermediate tokens, drawing on Quiet-STaR (Zelikman et al., 2024) and DeepSeek-R1 distillation results.

However, the preliminary observations suggest that the most valuable training targets are metacognitive habits (spontaneous confound checking, documenting paths not taken, maintaining scientific skepticism toward positive results, recognizing review requests as alignment checks rather than challenges) rather than reasoning chains per se. If so, the distillation target is behavioral: does the model exhibit the habits without prompting? This may require behavioral fine-tuning rather than, or in addition to, the Quiet-STaR architecture. The distinction is empirical and the pilot is designed to resolve it.

5. Training Signal Capture

The precise mechanism for extracting gradient-usable training signal from these dialogues remains the central open question. The preliminary analysis narrows the space considerably. The following hypotheses are ordered from most immediately actionable to most speculative.

5.1 Reward the Unprompted Check

The clearest finding from the transcript analysis: the model can reason about confounds, controls, and epistemic status when asked, but does not generate these checks spontaneously. This directly specifies a reward structure: score trajectories for the presence of unprompted control questions, unprompted anomaly flagging, and unprompted epistemic status assessment. The presence or absence of these behaviors is detectable without domain expertise; it is a structural property of the output, not a judgment about scientific correctness.

This is operationally the simplest training signal to extract and the most directly supported by the evidence. It addresses the most persistent failure mode (confound blindness, 5+ occurrences despite correction) and targets the gap between prompted and unprompted capability that defines the RLHF objective.

5.2 Verifiable Sub-Components as Proxy Reward

Scientific reasoning contains verifiable sub-tasks that could serve as automated intermediate reward signals. In the transcripts, these included: probe accuracy measurements (objective, automatable), confound identification chains with falsifiable endpoints, arithmetic sub-problems within experimental design reasoning, and experimental design validity checks (does the proposed analysis test the stated hypothesis, given the available data). These verifiable components are frequent in experimental sections and could serve as automated reward analogous to test suites in SWE-bench.

5.3 Cascade Depth as Quality Weighting

Training events that cascade across three or more categories are disproportionately valuable. A reward signal that weights events by cascade depth (treating a four-category cascade as higher-value than a single-category correction) would align training incentives with observed training value. Cascade depth is measurable from the taxonomy: each event can be tagged with the categories it spans, and the count serves as a weight.

5.4 Structural Alignment with Checkpoints

The five-element checkpoint template provides a structural scoring rubric applicable without domain expertise: are assumptions stated, are rejected approaches documented with revisit conditions, is the path not taken recorded? Structural comparison catches gross failures and scores the metacognitive habits among the highest-value training targets.

5.5 Failure Mode Reduction

The seven identified failure modes are countable. A reward signal scoring trajectories for the absence of these failure modes provides a concrete, enumerated set of behavioral targets. The asymmetry between failure handling (self-corrects on obvious failures like NaN loss) and success handling (proceeds without investigation on suspiciously clean results) suggests a specific reward structure: reward adversarial inspection of positive results.

5.6 Narrative Compression and Model Self-Diagnosis

The most speculative hypotheses. First: full dialogue arcs may function as training signal at a level not reducible to step-level components. The strongest evidence is a 30-turn sequence in which a concept that neither participant had articulated at session start crystallized through a chain of analogical moves. The expert drew an analogy between biological memory and the proposed system; the agent extended it; the expert challenged the extension; the agent revised; and the resulting concept (automatic non-explicit retrieval triggered by internal state) was not present in any individual turn. No single turn contained the signal; the trajectory did. Second: in transcripts exposing the agent's internal reasoning process, the agent performed explicit failure-mode classification after expert correction, sometimes operationalizing its self-diagnosis into persistent behavioral rules stored in project documents. If the agent's self-diagnosis can serve as a reward label, this provides training signal generated by the model itself. Both hypotheses require additional investigation.

These hypotheses are not mutually exclusive. The pilot is designed to test which mechanisms produce measurable improvement on the validation projects, alone and in combination.

6. Relationship to Existing Work

GRPO and DeepSeek-R1. GRPO (Shao et al., 2024) eliminates the critic model by scoring outputs relative to each other within a group. DeepSeek-R1 (DeepSeek-AI, 2025) demonstrated emergent reasoning via pure GRPO, including an "aha moment" phenomenon where the model develops self-reflective behavior spontaneously. Our preliminary observation that the agent produces explicit self-diagnostic behavior in its internal reasoning (classifying its own failure modes categorically rather than just acknowledging errors) parallels this phenomenon. Liu et al. (2025) refined GRPO with Dr. GRPO.

Multi-turn RL. Shani et al. (2024, NeurIPS 2024) provide theoretical foundations for optimizing over full conversations with convergence guarantees. Their Education Dialogue environment is structurally analogous to our setting. Wei et al. (2025) extend GRPO to multi-turn with MT-GRPO. The cascade depth finding from our analysis supports the multi-turn approach: the highest-value training events unfold over multiple turns and categories, and step-level extraction would miss their structure.

SWE-bench. Jimenez et al. (2024) evaluate code generation against test suites on real GitHub issues, providing the model a repository, a runtime, and verifiable endpoints. Our design parallels this: the model receives a research question, a computational environment, and verifiable experimental endpoints. The extension is that our evaluation includes structural checkpoints (documenting the reasoning process) and failure mode counts alongside outcome verification.

Process reward models. Lightman et al. (2023) and Uesato et al. (2022) demonstrate step-level reward for mathematical reasoning. Our checkpoint structure provides coarse-grained process reward at research milestones. The finding that cascade depth predicts training value suggests a novel form of process reward weighted by cross-category span rather than step count.

Constitutional AI. Bai et al. (2022) train models to self-critique against fixed principles. Our approach replaces fixed principles with dynamic expert judgment. The preliminary observations show the expert's role is more varied than "Socratic questioner"; it includes direct correction, design improvement, meta-process instruction, and process meta-commentary.

Sycophancy and RLHF failure modes. Sharma et al. (2023) demonstrate that RLHF-trained models exhibit sycophantic behavior, producing outputs that match user expectations rather than accurate ones. Perez et al. (2022) document similar patterns. Our Finding 2 (the confidence paradox) provides domain-specific evidence of the same problem: in scientific reasoning, standard reward signals may actively penalize the uncertainty and self-questioning that constitute good scientific practice.

STaR, Quiet-STaR, and distillation. Zelikman et al. (2022, 2024) demonstrate reasoning bootstrapping and implicit rationale generation. Our Phase 3 draws on these but may require adaptation: if the primary training targets are metacognitive habits rather than reasoning chains, the distillation mechanism may need to be behavioral rather than architectural.

Standard RLHF, DPO, and online iterative methods. Christiano et al. (2017), Ouyang et al. (2022), Schulman et al. (2017), and Rafailov et al. (2023) serve as baselines. Dong et al. (2024) describe online iterative RLHF; our pipeline is inherently online, as validation trajectories from each cycle become training data for subsequent cycles.

7. What Success Looks Like

The pilot succeeds if the trained model produces measurably better independent scientific reasoning than the base model on the three validation projects. "Better" is defined concretely across three measures: higher structural alignment with scientist checkpoints (more elements present, more paths not taken documented), fewer instances of the seven identified failure modes, and arrival at the known experimental endpoint through reasoning that is justified at each stage.

Secondary success criteria: (a) the training signal taxonomy replicates across multiple expert-agent pairs, confirming that the preliminary observations generalize beyond the n=1; (b) the prompted-versus-unprompted gap narrows measurably, with the trained model spontaneously generating checks the base model requires prompting to produce; (c) multi-turn trajectory optimization outperforms outcome-only approaches, confirming that dialogue structure and cascade depth carry training value.

The pilot also answers a structural question about RLHF: whether PhD-level scientists contribute most to AI training not as annotators scoring outputs on rubrics, but as research collaborators whose heterogeneous expert judgment constitutes a higher-bandwidth training signal than any preference ranking.

References

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, 89(4), 369-406.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Christiano, P. F., Leike, J., Brown, T., Marber, M., Lowe, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*. arXiv:2501.12948.

Dong, H., Xiong, W., Pang, B., Wang, H., Zhao, H., Zhou, Y., ... & Zhang, T. (2024). RLHF Workflow: From reward modeling to online RLHF. *arXiv preprint arXiv:2405.07863*.

Gao, Z., Chang, J. D., Zhan, W., Oertell, O., Swamy, G., Brantley, K., ... & Sun, W. (2024). REBEL: Reinforcement learning via regressing relative rewards. *arXiv preprint arXiv:2404.16767*.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hsieh, C. Y., Li, C. L., Yeh, C. K., et al. (2023). Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. *Findings of ACL 2023*.

Jimenez, C. E., Yang, J., Wettig, A., et al. (2024). SWE-bench: Can language models resolve real-world GitHub issues? *ICLR 2024*. arXiv:2310.06770.

Lightman, H., Kosaraju, V., Burda, Y., et al. (2023). Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Liu, Z., et al. (2025). Understanding R1-Zero-like training: A critical perspective. *COLM 2025*. arXiv:2503.20783.

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.

Perez, E., Ringer, S., Lukosiute, K., et al. (2022). Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shani, L., Rosenberg, A., Cassel, A., et al. (2024). Multi-turn reinforcement learning from preference human feedback. *NeurIPS 2024*. arXiv:2405.14655.

Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Shao, Z., et al. (2024). DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Uesato, J., Kushman, N., Kumar, R., et al. (2022). Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Wei, L., et al. (2025). MT-GRPO: Multi-turn group relative policy optimization with intermediate step rewards. *arXiv preprint*.

Zelikman, E., Wu, Y., Mu, J., & Goodman, N. (2022). STaR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35.

Zelikman, E., Harik, G., Shiv, Y., et al. (2024). Quiet-STaR: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*.