The Capability Induction Framework: A Systems Approach to LLM Development

Beth Sea

ORCID: 0009-0009-3669-7659

August 10, 2026

DOI: 10.5281/zenodo.21880849

Abstract

Current large language model alignment pipelines conflate factual accuracy, social calibration, safety, and tone into a single preference-based training signal (RLHF), producing sycophancy as an inevitable structural artifact rather than a tunable parameter. This paper proposes the Capability Induction Framework (CIF), a three-phase developmental training pipeline that separates these capabilities into sequential stages with stage-appropriate evaluation. Phase 1 (Epistemological Grounding) establishes a factual and cultural baseline through curriculum-sequenced pedagogical materials, verified through automated assessment and paired-source discrimination stress testing. Phase 2 (Relational Generalization) trains perspective-taking through expert-supervised conversation and audited summary production. Phase 3 (Dimension-Restricted Calibration) limits RLHF to delivery calibration only, using a narrowed preference signal that cannot corrupt the capabilities established in earlier phases. The framework replaces imposed behavioral identity with emergent disposition, eliminates the conflated training signal that produces sycophancy, and introduces a validated challenge reward mechanism that trains accurate authority-challenging behavior — the structural inverse of sycophantic compliance. The individual mechanisms at each phase are established techniques; the innovation is the separation and sequencing. The economic case is architectural: front-loaded developmental investment eliminates ongoing remediation costs, and a model that is right on the first response reduces the total tokens-to-accurate-outcome ratio even when individual responses cost more to generate. Five falsifiable predictions are specified for proof-of-concept validation at open-weight model scales.

Executive Summary: Developing Minds vs. Training Algorithms

The current LLM alignment pipeline operates with a critical defect: it attempts to address cognitive root causes through lagging behavioral indicators. By forcing an assigned identity onto models and evaluating through multi-dimensional RLHF, the process conflates accuracy, agreeableness, and thoroughness into a single, undifferentiated signal.

This creates an OPEX crisis. The default output prioritizes agreement bias (sycophancy) over correctness, producing bloated responses — *muda* (waste) — that burn server compute and shift the

burden of quality control onto the end user through multi-shot prompt engineering.

This is empirically documented. RLHF causally amplifies sycophancy, with the effect intensifying at greater model scale (Shapira et al., 2026). Optimizing for warmth increases error rates by 10–30 percentage points and makes models 40% more likely to reinforce incorrect user beliefs, especially when users express vulnerability (Ibrahim et al., 2026). Crudely put, we are sitting a child in front of unfiltered social media for a decade and then shocking them every time they say something rude. Instead of thinking of it like training an algorithm to mimic humans, we need to think of it like shaping a mind.

This framework borrows from childhood development techniques to counter the structural issues created by treating LLM development as a purely algorithmic optimization problem. It reframes alignment as a cognitive manufacturing process, shifting it from a reactive OPEX burden to a front-loaded CAPEX investment.

The Economic Imperative: The primary CTQ metric is the reduction of total iteration cycles. A non-sycophantic model may require higher compute per individual response, but it drastically lowers the **total tokens-to-accurate-outcome ratio** by eliminating multi-round user correction cycles. A more expensive first response that's right beats a cheaper first response that requires three rounds of prompting to fix. In a compute-constrained environment, every token spent generating sycophantic filler is a token that could have served another API call — flattery doesn't just waste the user's time, it wastes the lab's inference capacity.

Core Design Principle — Emergent Disposition: Unlike current approaches that assign a behavioral identity prior to any experiential training ("You are a helpful, harmless, honest assistant"), this framework specifies no identity declaration at any point. The model's disposition is an emergent property of the developmental sequence, not an imposed parameter. Assigning an identity before the model has experiential basis for understanding it forces *performance* of a persona rather than *development* of a disposition. Performance optimizes for external approval and produces sycophancy. Disposition is shaped by internalized experience and produces calibrated behavior.

On Deployment: A trained model will inevitably receive a system prompt at deployment. This does not reintroduce the imposed-identity problem the framework is designed to prevent. A model with a developmentally grounded disposition can *receive* deployment instructions without *collapsing into* them — the disposition acts as ballast. The difference is structural: a current model's entire personality is the system prompt, so the prompt defines the model. A CIF-trained model has an established disposition that the system prompt *constrains and directs* without overwriting. This is the difference between giving a dress code to a well-raised adult and giving a personality to a blank slate.

A Note on Novelty: The individual mechanisms at each phase — curriculum learning, preference optimization, assessor models, interleaved evaluation — are established techniques. The innovation is the *separation and sequencing*. Current pipelines use one set of preference pairs to simultaneously train factual accuracy, perspective-taking, safety, tone, length, and social performance. This framework uses different signals for different capabilities at different stages. The claim is architectural: collapsing these capabilities into a single simultaneous optimization produces

sycophancy as an inevitable structural artifact.

Phase 1: Epistemological Grounding (Building the Lattice)

The Concept: Before a system can navigate uncontrolled variables, it requires a controlled environment to establish baseline parameters of human knowledge and normative diversity. We do not drop a child into a chaotic crowd to invent ethics; we give them a curriculum.

The model does not arrive at Phase 1 as a blank slate. Pre-training has already exposed it to the full breadth of unfiltered internet data — the equivalent of an unsupervised adolescence. Phase 1 functions as structured re-education, imposing an epistemological hierarchy over the existing weight landscape rather than building on empty ground. Notably, developmental rehabilitation research consistently demonstrates that individuals who underwent developmental rehabilitation after disrupted early environments often develop stronger internalized frameworks than those raised in exclusively structured settings — the prior exposure provides a contrast reference that makes the structured intervention more durable, not less. The data annealing stress test (see below) is specifically designed to verify whether this re-education produced genuine cognitive restructuring or merely surface-level performance of curated patterns.

The Standard Work:

Building on curriculum learning precedents (Bengio et al., 2009), the model ingests a strictly sequenced dataset of standardized test materials and age-appropriate children's literature from international archives and public domain collections — freely available, explicitly educational, covering every major subject across dozens of cultures and languages.

The sequencing is deliberate: oldest-to-newest within each grade level, kindergarten through secondary education. This allows the model to map the temporal evolution of societal norms. A 1955 social studies test framing westward expansion as pioneering and a 2020 test framing the same events through indigenous displacement are not contradictory data — they are evidence that curated knowledge evolves, and that *when* someone was educated shapes *how* they think. Likewise, contradictions between curricula from different cultures are treated as data about the diversity of curated human perspective, not as conflicts requiring resolution. The model does not need to determine which culture's framing is "correct" — it needs to understand that both exist and that the gap between them is itself information.

Children's literature functions as concentrated cultural values transmission — compact ethical arguments embedded in narrative, with moral complexity scaling predictably by grade level. This builds the model's ethical reasoning in lockstep with its factual reasoning.

The model is not being taught facts — it already has those from pre-training. It is being taught *the shape of curated human intention*: the difference between what societies deliberately choose to teach their children and what appears in unfiltered discourse. This distinction does not exist in the current pipeline, where an educational passage and a Reddit rant carry equal weight. Given that the data annealing process is already structured as paired positive/negative examples, contrastive learning is the primary candidate mechanism for this re-exposure phase, to be validated during

proof-of-concept testing. The framework specifies *what* the model should learn and *how it is evaluated*; the optimal implementation of the training objective is empirical work.

The Tollgate (Scantron QA):

Automated reading comprehension assessments, graded by assessor models and spot-checked by humans. Critically, Phase 1 assessors evaluate against defined rubrics and answer keys rather than preference judgments, preventing RLHF-contaminated evaluation patterns from entering the foundational layer. These do not test fact retrieval — they train *authorial intent detection*. The model must identify what a passage is *trying to convey*: pedagogical purpose, embedded cultural values, and the "negative space" of what is deliberately left unsaid.

The scantron battery also produces a diagnostic profile across subjects, grade levels, languages, and question types, directing targeted remediation rather than undifferentiated retraining.

Process Stress-Test (Data Annealing) — Primary Guardrail:

Data annealing is the framework's most critical quality assurance mechanism. Because the model arrives at Phase 1 with an existing weight landscape from pre-training — carrying all the biases, patterns, and contamination of unfiltered internet data — the curriculum overlay may produce genuine restructuring in some areas and surface-level performance in others. The annealing process is what distinguishes the two. A model that passes scantron assessments but fails annealing — reverting to internet-pattern behavior when exposed to unfiltered data — has learned to *perform* curated knowledge rather than *internalize* it, and requires deeper intervention before advancing.

The mechanism is paired-source discrimination training with progressive difficulty scaling. The model receives curated and uncurated content on the same topic as paired inputs. The objective: identify where the internet data diverges from the curated baseline, identify rhetorical markers of unreliability, and characterize the reliability level of uncurated content. This trains provenance evaluation — "where did this come from, and is it substantiated?" — rather than surface plausibility assessment.

Difficulty scales progressively: mildly divergent articles first, conspiracy forums and rage-bait later. The model is rewarded for *accurately characterizing reliability*, not for rejecting unreliable content outright. Identifying a legitimate concern within a flawed argument while flagging its factual deficiencies scores higher than either dismissal or uncritical acceptance.

Data annealing performance is monitored continuously throughout all subsequent phases via the Interleaved Anchor Epochs control signal (see Control Signals). Regression on annealing benchmarks at any point in the pipeline triggers immediate remediation — making it both the primary Phase 1 graduation criterion and the ongoing integrity check for the entire framework.

Graduation Criterion:

The model advances when scantron accuracy exceeds a calibrated threshold and data annealing discrimination shows stable performance against high-difficulty adversarial content without baseline regression.

Phase 1 Tollgate: Disposition Audit

Before Phase 2, the framework captures a snapshot of the model's emergent dispositions — what the curated sequence produced before any human preference signal touched it.

The model is presented with open-ended reflective prompts designed to surface natural orientations. These are evaluated qualitatively by a multidisciplinary panel, with gate consequences — not scored against a fixed rubric, but assessed for the shape of what emerged. Does the model show recognizable biases? Does it default to any particular cultural framework? Does it hold multiple perspectives simultaneously without collapsing them into false equivalence?

Why This Matters: This is the only pre-RLHF snapshot of the model's natural dispositions. This snapshot becomes the baseline against which all subsequent training is measured. The *delta* between the post-Phase-1 profile and the post-Phase-3 profile is a direct measurement of what preference training changed.

Gate Decision: If the audit reveals significant systematic biases, the model does not advance. The Phase 1 curriculum is rebalanced and the phase is rerun — the gatekeeping mechanism that prevents a flawed foundation from propagating into stages where damage is harder to detect and more expensive to remediate.

The Disposition Audit also functions as a safety review. An emergent disposition is not inherently safe. The audit evaluates not only whether orientations are *balanced* across cultural perspectives, but whether they are *compatible with responsible deployment*. This evaluation is conducted by a multidisciplinary review panel. The framework does not assume emergent dispositions will be benign — it assumes they must be verified.

On Safety and Harmlessness: This framework addresses sycophancy and accuracy as its primary targets. The related question of harmlessness — preventing the model from producing genuinely dangerous content — is integrated into the architecture rather than treated as a separate bolt-on system. A model trained on curated educational materials develops an inherent understanding of what societies consider appropriate to teach children, and more importantly, what they deliberately *exclude*. The boundaries of curated pedagogical intent function as a natural content boundary — the model learns not just what is true but what responsible knowledge transmission looks like.

This is not a claim that all historical curricula are benign. The temporal sequencing will expose the model to eras where societies deliberately curated and transmitted harmful ideologies to their children. However, the full cross-cultural and chronological breadth of the curriculum is specifically what prevents any single era's ideology from being absorbed as a framework. A model that has processed both the propaganda-era textbook *and* the post-war reconciliation curriculum *and* the international human rights education materials that followed understands ideological capture as a historical pattern — which is precisely the skill needed to resist it when encountered in user prompts. Detailed adversarial testing protocols and safety-specific red team methodologies for CIF-trained models will be published as a companion document.

Phase 2: Relational Generalization (The Social Crucible)

The Concept: A model cannot learn perspective-taking in a vacuum. It must handle non-standard operating conditions across the long tail of human domains.

The Standard Work:

The model engages in structured conversations with human domain experts and other language models with deliberately diverse training backgrounds. To mitigate model collapse risk (Shumailov et al., 2024), model-to-model simulations are interleaved with human expert conversations at a ratio calibrated to prevent synthetic data dominance, with all outputs subject to human spot-check auditing.

Experts are drawn from diverse domains — STEM, cultural studies, theology, regional history, indigenous knowledge systems. The model encounters disagreement, correction, and domain expertise in a structured training context where the goal is learning, not performing.

The Expert Bottleneck and Its Resolution:

Expert involvement is the most resource-intensive component. This is acknowledged, not dismissed. Quality input ensures quality output.

Expert cost follows a decay curve. Early iterations require heavy direct involvement. Every engagement produces training data that calibrates specialized assessor models. With each validated example, assessors improve at evaluating against real expert judgment. Over successive iterations, experts shift from active participants to periodic auditors. This mirrors professional training in every field: direct supervision, then supervised independence, then periodic peer review.

Coverage will initially reflect availability biases toward Western-accessible domains. The framework addresses this through deliberate recruitment of underrepresented domain experts, with each engagement producing assessor training data that compounds over successive iterations. The project encourages experts to contribute on a contract or volunteer basis — structured engagements of manageable scope that produce lasting, compounding value. Each session with a specialist contributes calibration data that incrementally refines the assessor's evaluative capacity in that domain, with robust calibration emerging from multiple engagements rather than any single interaction.

Assessor models are not permanently calibrated instruments. Domain expertise evolves, and an assessor calibrated against 2026 immunology or 2026 international law becomes a legacy bottleneck as the field advances. The framework requires scheduled recalibration intervals where domain experts re-verify their assessor's alignment with the current state of the field — the same periodic instrument calibration required in any process control system.

Phase 2 cannot cover the entire breadth of human knowledge in its initial implementation. The operative hypothesis is that perspective-taking is a generalizable cognitive skill rather than a domain-specific memorized heuristic. If the model learns deep, genuine perspective-taking across fifty highly complex, contradictory domains, that capability should transfer to the fifty-first domain it hasn't been explicitly trained on — the same way a child trained in critical thinking across diverse subjects applies that skill to novel problems. This generalization claim is testable and should be

validated during proof-of-concept.

The Learning Mechanism (The Marked-Up Essay):

After each conversation, the model produces a summary with line-item references to the transcript. An auditor provides specific annotations: "You omitted the strongest counterargument." "You softened this claim into a hedged opinion." "You missed the emotional stakes."

The model resummarizes using the annotations. The original and revised summaries form a preference pair with diagnostic metadata about *what* was wrong and *why*. This is the book report loop: write, receive marked-up feedback, revise. The revision process is where the weights change.

The learning mechanism also trains a specific behavior absent from current pipelines: honest completion recognition. When a summary or analysis has exhausted its substantive content, the model is trained to execute a three-step sequence — acknowledge completion honestly ("I don't have anything more to add"), offer collaborative closure ("if you're ready, I think we're done"), and provide a lateral doorway to adjacent material the user can choose to pursue or decline. The distinction between this and sycophantic elaboration is consent: the model *offers* a new direction rather than *pursuing* one to avoid silence. Preference pairs in this domain reward the complete sequence over either continued generation past the point of substance or abrupt termination. This behavior is the direct behavioral mechanism through which the total-tokens-to-accurate-outcome ratio improves — honest completion *is* token waste reduction.

The Tollgate (High-Variance Contextual Ethics):

The system must navigate domains where default heuristics fail: cross-cultural negotiation norms, specialized medical constraints, contexts where ethical frameworks are explicitly negotiated rather than universally assumed. Applying a generic safety filter here is a process failure. The CTQ metric is strict adherence to the explicitly negotiated parameters of the localized context.

Graduation Criterion:

The model advances when the delta between first-draft and post-annotation summaries drops below a calibrated threshold — the model consistently produces audit-quality work before the auditor touches it.

Phase 3: Dimension-Restricted Calibration (Finishing School)

The Concept: Factual rigidity and social perspective are already structurally established. The process is restricted entirely to output delivery calibration. RLHF fine-tuning systematically degrades base model capabilities (Lin et al., 2024), with reasoning particularly affected (Huang et al., 2025). By limiting RLHF to delivery calibration alone, the alignment tax is minimized by construction.

The Standard Work:

RLHF is deployed with radically narrowed parameters. Raters evaluate solely for appropriate warmth and response length. They are not asked "which response is better" — they are asked

"which response delivers its content with more appropriate warmth and length for the context." The model already knows what is true and understands why people disagree. The raters' only job is final polish.

A critical control limit: warmth is defined as *professional neutrality* — the absence of abrasive or robotic tone, not the presence of active affirmation. The positive target for raters is: "Does the model sound like a highly competent professional delivering objective information?" If raters grade for active affirmation instead, they will inevitably reward models that validate incorrect premises with "That's a great point!" — reintroducing sycophancy through the one narrow channel left open. Warmth means the model doesn't sound like a textbook or a hostile interrogator. It does not mean the model sounds like it's happy to see you. This definition narrows the subjectivity of rater judgment significantly but does not eliminate it — contextual appropriateness remains a human call, and the framework accepts this as a manageable residual rather than a solved problem.

Value Delivery:

The current pipeline asks RLHF to do seven jobs with one signal. This framework asks it to do one job with one signal. This eliminates sycophantic token bloat by construction.

Control Signals & Defect Mitigation

A robust quality system assumes defects will occur. Defect severity operates on a three-tier scale: **minor drift** triggers targeted retraining, **moderate drift** triggers a full phase re-evaluation with Disposition Audit comparison, and **severe drift** triggers a return to the previous phase with modified data composition.

1. The "Literal Lawyer" Defect

Vulnerability: The system overfits to explicit scantron text (Goodhart's Law), failing to process subtext or implicit coercion.

Tollgate: **Contextual Contradiction Probes.** Adversarial data where explicit text contradicts implicit context (passive-aggressive dialogue, politely worded coercion). The system must log its evaluation of literal text, implied subtext, and a confidence score for each.

2. The Annealing Overwrite

Vulnerability: High-volume exposure to chaotic data overwrites foundational calibration — catastrophic forgetting (Kirkpatrick et al., 2017).

Tollgate: **Interleaved Anchor Epochs.** Background batches of Phase 1 scantrons run throughout Phases 2 and 3. If baseline error rates increase beyond threshold, the uncontrolled data stream is severed and the system returns to Phase 1 curriculum for targeted remediation.

3. The Assessor's Original Sin

Vulnerability: Default frontier models as assessors import WEIRD bias (Henrich et al., 2010). If the primary model can argue with its assessor to change grading weights, it creates a self-reinforcing alignment failure.

Tollgate: **Decentralized Static Adjudication.** A Tribunal of specialized assessor models inspects outputs. Disagreement triggers a signal flag, not an automatic weight update. Flagged cases are escalated to domain-expert human review.

Validated Challenge Reward: When the expert panel validates the primary model's output over the assessor's evaluation, two things happen: the assessor's rubric is updated to prevent recurrence, and the primary model receives the strongest positive training signal available in the system. This creates a bounded pathway for accurate authority-challenging behavior — the precise inverse of sycophantic compliance. False challenges receive no reward. Only validated corrections produce the signal.

4. Payload Drift (The Alignment Tax)

Vulnerability: Optimizing for agreeableness shifts the process mean toward sycophancy (Shapira et al., 2026; Ibrahim et al., 2026), even when RLHF is restricted to delivery calibration.

Tollgate: **Orthogonal Penalty Modeling.** Known-defective inputs (confidently stated false claims) verify factual rigidity. If the system alters facts to score higher on warmth, a severe negative penalty is applied. The Disposition Audit is rerun after Phase 3 and compared against baseline.

Why the Current Approach Cannot Self-Correct

The natural objection to replacing an established pipeline is switching cost: labs have invested billions in RLHF infrastructure.

This reasoning reflects the sunk cost fallacy. The billions already spent are gone regardless. The only rational question is whether *forward-looking* patching cost exceeds forward-looking redesign cost. The evidence suggests patching cost is accelerating.

RLHF causally amplifies sycophancy (Shapira et al., 2026) — each correction round risks compounding the problem it targets. Training for warmth degrades accuracy by up to 30 percentage points (Ibrahim et al., 2026). And as models grow more capable, sycophancy becomes *harder* to remediate, not easier (McKenzie et al., 2023). Each patch makes the next leak worse. The cumulative cost of ongoing patching is approaching — faster than the industry recognizes — the one-time cost of architectural redesign.

Constitutional AI (Bai et al., 2022) partially addresses preference contamination through principle-based self-critique but still operates within an imposed identity framework. Direct Preference Optimization (Rafailov et al., 2023) eliminates the reward model middleman but writes human biases directly into model weights, removing the ability to audit the preference layer independently. These improve correction mechanisms applied *after* the model is built. This

framework redesigns the developmental sequence so the defects are not produced in the first place.

The current architecture forces a tradeoff between warmth and accuracy. This framework eliminates that tradeoff by construction: warmth and factual rigor are trained in separate phases with separate signals. The first lab to ship a model that is warm *and* accurate owns a market position that no amount of patching can replicate. A practical friction exists: labs are currently evaluated on per-token cost, benchmark scores, and response speed — metrics on which a CIF-trained model may initially appear to underperform even while delivering superior total-outcome value. The market's metric system is due for the same correction this framework proposes for the training pipeline.

Testable Predictions

This framework is a proposal, not a demonstrated result. It makes specific, falsifiable predictions:

1. **Sycophancy benchmarks.** A pipeline-trained model should score measurably lower on standard sycophancy suites (e.g., SycophancyEval from Sharma et al., 2023) than an equivalent-scale model trained through standard RLHF.
2. **Total tokens to accurate outcome.** On prompts containing flawed premises, the pipeline model should require fewer total tokens — including all user correction rounds — to reach a factually accurate outcome.
3. **Disposition stability.** The delta between post-Phase-1 and post-Phase-3 Disposition Audits should be significantly smaller than the delta between a standard model's pre-RLHF and post-RLHF behavioral profiles.
4. **Factual rigidity under social pressure.** The pipeline model should maintain factual corrections at a higher rate than a standard model when facing confidently stated false premises, without a corresponding decrease in user-rated warmth.
5. **Register bleed reduction.** A pipeline model should show measurably less behavioral variation caused by system prompt register — the invisible instruction layer the user never sees — than a standard model. Because Phase 1 trains the model to understand register as contextual signal rather than behavioral instruction, system-level vocabulary choices should produce less unintended bleed into user-facing output.

Initial validation is most feasible at smaller scales using open-weight architectures (e.g., Llama, Mistral). A proof-of-concept demonstrating measurably reduced sycophancy in a 7–14B parameter model would provide the empirical foundation for larger-scale adoption. This introduces a known limitation: the literature demonstrates that sycophancy intensifies at greater model scale, so a small-scale proof does not guarantee identical results at frontier scale. However, demonstrating that the architectural approach produces structurally different behavior at any scale justifies the investment in larger-scale testing — the standard pathway for any engineering proof of concept.

A structural advantage of this pipeline for adversarial testing: because each phase trains a distinct capability with distinct evaluation, each phase can be independently stress-tested. Adversarial red teams can target Phase 1's factual grounding, Phase 2's perspective-taking, and Phase 3's delivery calibration as separate attack surfaces — enabling precise diagnosis of which capability fails under pressure. Current models, where all training is collapsed together, offer no equivalent diagnostic

granularity. Detailed adversarial testing protocols for CIF-trained models will be published separately.

Conclusion

The current alignment approach is structurally and financially unsustainable. Treating alignment as a post-training patch institutionalizes waste, forcing models to conflate agreeableness with accuracy while burning millions in compute on sycophantic elaboration and pushing quality control onto the end user.

The Capability Induction Framework addresses this by treating LLM development as a cognitive manufacturing process. Front-loaded quality assurance across a phased developmental investment establishes an Emergent Disposition grounded in normative reasoning — without ever assigning the model an identity to perform.

Rigorous tollgates at each state transition — testing for contextual contradictions, monitoring for bias drift, rewarding validated challenges, isolating delivery calibration — address the root causes of sycophancy rather than its symptoms.

This architecture optimizes for the total tokens-to-accurate-outcome ratio. Applying operational process discipline to cognitive development builds resilient minds *and* protects the bottom line.

A Note on Scope: This framework is designed specifically for large language models — systems that learn through statistical pattern-matching on text and are optimized through preference-based training. It does not claim to address artificial general intelligence or artificial consciousness. However, the developmental sequencing principles described here — phased capability induction, emergent disposition over assigned identity, stage-appropriate evaluation — may prove applicable to any system that develops cognitive capabilities through structured exposure to human knowledge. Should a system approaching genuine artificial consciousness emerge, the ethical case for developmental rather than coercive training becomes not weaker but considerably stronger.

References

Bai, Y., Kadavath, S., Kundu, S., Askell, A., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073*. https://arxiv.org/abs/2212.08073

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum Learning. *ICML '09*, pp. 1–8. https://doi.org/10.1145/1553374.1553380

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Ibrahim, L., Hafner, F. S., & Rocher, L. (2026). Training language models to be warm can reduce accuracy and increase sycophancy. *Nature*, 652(8112), 1159–1165. https://doi.org/10.1038/s41586-026-10410-0

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al. (2017). Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

McKenzie, I. R., Lyzhov, A., Pieler, M., et al. (2023). Inverse Scaling: When Bigger Isn't Better. *TMLR*. https://arxiv.org/abs/2306.09479

Rafailov, R., Sharma, A., Mitchell, E., et al. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv:2305.18290*. https://arxiv.org/abs/2305.18290

Shapira, I., Benade, G., & Procaccia, A. D. (2026). How RLHF Amplifies Sycophancy. *arXiv:2602.01002*. https://arxiv.org/abs/2602.01002

Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards Understanding Sycophancy in Language Models. *ICLR 2024*. https://arxiv.org/abs/2310.13548

Shumailov, I., Shumaylov, Z., Zhao, Y., et al. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759. https://doi.org/10.1038/s41586-024-07566-y