Position: Don't Just "Fix it in Post": A Science of AI Must Study Learning Dynamics

Anonymous Authors[1]

Abstract

What would it mean to have a *scientific* understanding of AI? Language models are not static objects—they are snapshots of time-evolving processes shaped by data, objectives, and optimization dynamics. Yet the field predominantly treats models as fixed artifacts, analyzing behaviors after training rather than asking *why* they emerge. This position paper argues that AI research should move beyond *post hoc* fixes and study the learning dynamics of models. We envision a hierarchy of scientific maturity: first *predict* outcomes from early training signals, then *intervene* when trajectories go wrong, ultimately *design* training procedures that guarantee desired properties. Scaling laws have reached the first level for loss; the challenge is extending all three levels to general capabilities, biases, and safety. We articulate requirements for such theories, survey progress across mechanistic interpretability, fairness, memorization, and learning dynamics, and identify concrete open problems. The path forward requires treating models as processes to be understood, not just artifacts to be patched.

1. Introduction

Consider a tennis ball thrown into the air. In classical mechanics, knowing the ball's position, velocity, and acceleration is sufficient to predict all of its future states. These predictions are remarkably precise—our main source of error is measurement of initial conditions, not the theory itself. The key to this predictive power is a causal model of how forces like gravity and properties like inertia operate on objects. The equations are not merely curves fit to past observations. They are consequences of a deeper story about masses and their interactions. Newton's laws do not merely describe the path of a ball through the air, but in understanding *why* the ball follows a parabolic trajectory we are also able to predict its path.

What would it mean to have such a theory for AI? It would be a causal account of how training dynamics produce model behaviors. It would provide a unified framework for how all aspects of language model training—from optimization, data, and architecture—influence model development. Such a theory would not only describe the training process, but identify the underlying mechanisms. And that understanding would enable us to predict, intervene in, and ultimately design training processes—not merely analyze their outputs after the fact.

In this position paper, we argue that the field is far from having a scientific theory of AI and must move beyond *post hoc* fixes to study the learning dynamics of models. The field predominantly treats trained models as static artifacts to be examined and described, rather than as endpoints of dynamical processes to be understood. When models exhibit undesirable behaviors, the dominant response is to "fix it in post": applying post-training interventions like RLHF, filtering, or fine-tuning, rather than understanding why the behavior emerged.

1.1. The "Fix It in Post" Mentality

The phrase "fix it in post" comes from film production and refers to deferring problems during shooting and correcting them in post-production. In AI, an analogous mentality has taken hold. Language models are trained on massive datasets with limited curation, then are "aligned" through methods like RLHF. When alignment fails developers typically apply more post-training patches rather than to investigate the upstream causes of the behavior (Andriushchenko et al., 2024).

The limitations of this approach are almost immediately apparent. Qi et al. (2023) and Lo et al. (2024) demonstrate that safety fine-tuning can be undone even when finetuning on *benign data*. Zhao et al. (2025) show that extended reasoning via chains-of-thought, despite being assumed to improve refusal consistency (Guan et al., 2024; Jaech et al., 2024), can be systematically exploited for jailbreaks. These

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

For ICML 2026 reviewers: When reviewing this paper, please follow the policy for the use of LLMs as specified on your Reviewer Console. Note that the assigned LLM policy might differ from the policy you indicated during reviewer registration.

findings suggest that *post hoc* alignment strategies are fundamentally brittle. Instead of treating the symptoms, it is essential to understand the causal mechanisms behind these failures to develop more robust safeguards.

What does addressing root causes look like? O'Brien et al. (2025) provide a concrete example: filtering dangerous information from *pretraining data* produces tamper-resistant safeguards that are much harder to undo via fine-tuning. The model lacks the capability rather than merely declining to use it. This is not a universal solution—other harms may not be so cleanly excised—but it demonstrates that interventions at training time can achieve robustness that *post hoc* patches cannot.

Example: Shadow Prompting and Gorillas

When DALL-E 2 generated mostly white men for prompts like "CEO," OpenAI's fix was "shadow prompting": secretly appending words like "Black" or "female" to user prompts (Offert & Phan, 2022). This produces superficially diverse outputs while leaving the underlying bias intact. Another case is Google Photos, which still cannot recognize gorillas, over a decade after Black users were infamously misclassified as apes (Grant & Hill, 2023). After years of unsuccessful remediation, Google ultimately disabled primate recognition altogether (Hern, 2018). Neither approach addresses root causes; both reveal the poverty of *post hoc* patches.

This pattern parallels the history of other scientific disciplines: intervention preceding understanding, failures met with more aggressive tuning rather than deeper inquiry. Pre-germ-theory medicine optimized treatments without causal models of disease (De Kruif, 1926); behaviorist psychology shaped behavior while refusing to theorize about internal representations (Chomsky, 1959); early aviation compensated for instability with brute force rather than aerodynamics (Vincenti et al., 1990). In each case, empirical success masked a lack of explanatory mechanism. History suggests that such phases end not through refinement of the interventions themselves, but when latent mechanisms are articulated and the space of possible behaviors becomes constrained by theory.

What would it mean to move beyond "fix it in post"? We envision a hierarchy of scientific maturity: *predict*, *intervene*, *design*. At the first level, we forecast outcomes from early signals—predicting final model properties from smaller-scale experiments or intermediate checkpoints. At the second level, we detect when training trends toward unacceptable outcomes and intervene mid-training to correct course. At the third level, we specify desired properties upfront and engineer training procedures guaranteed to achieve them. Scaling laws have reached the first level for aggregate loss (Kaplan et al., 2020; Hoffmann et al., 2022) and downstream evaluations (Jaech et al., 2024; Chen et al., 2024b); the challenge is extending all three levels to the properties we actually care about.

1.2. Roadmap

We argue for reorienting AI research toward building scientific theories. Section 2 articulates what such theories require: treating models as time-evolving processes, prediction rather than mere description, and generalization across instances.

2. Requirements for a Science of AI

What does it look like to design a scientific paradigm for AI? We believe that there are three core heuristics that AI researchers can focus on to improve their scientific practices: predictive theories, studying learning dynamics, and generalization across models. In addition to encouraging scientific thinking on their own, these three heuristics also empower the *predict*, *intervene*, *design* pipeline we view as a path forward for bringing about a science of AI.

2.1. Prediction, Not Just Description

Arguably the most important hallmark of a successful scientific theory is the ability to make predictions about *what will happen* should an experiment be run and then providing the framework for verifying the prediction. Sometimes this takes the form of precise equations (e.g., the laws of motion in physics) and sometimes this takes the form of less exact heuristics and methodologies that nevertheless can be used to draw reliable inferences.

Success Story: Darwin's Moth

A defining feature of mature science is its ability to predict unobserved phenomena. In 1862, Darwin encountered the orchid *Angraecum sesquipedale*, with a nectar spur nearly 30cm long. He reasoned that if the orchid reproduces, something must pollinate it—and only an insect with a proboscis long enough to reach the nectar would do. He predicted such a moth must exist in Madagascar. Over four decades later, the moth was discovered, confirming the prediction (Arditti et al., 2012).

In the study of language models, we rarely make such predictions. We describe, in great detail, what a model does *after* training. But *before* training, we cannot say what behaviors a model will learn (or what capabilities might "emerge" (Wei et al., 2022)).

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

One notable exception is predicting pretraining loss with scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022). Given model size and a compute budget, final loss can be predicted with reasonable accuracy. Labs use this to plan training runs, which offers practical utility. Recent work has also shown success at extending prediction to some evaluation benchmarks (Held et al., 2026; Chen et al., 2024b; Dubey et al., 2024; Achiam et al., 2023). We can conceptually extend this paradigm[1] to more generally ask, for a particular property P, what methodologies would enable us to predict the value of P for a model given smaller models. While some research in this direction exists for memorization (Biderman et al., 2023a; Prashanth et al., 2024) and social bias (Biderman et al., 2023b; Patel et al., 2025), we view the ability to develop these theories as an essential component of the science of AI. We expand upon this idea further in Section 3.

2.2. Models as Time-Evolving Processes

The mainstream view in machine learning is that the goal of training is to produce a model. This framing does significant harm. Model training runs are usually ended once extrinsic metrics have been satisfied: compute budget is used up, a deadline has come, or performance is judged to be "good enough" based on subjective assessments. *The final checkpoint is no more scientifically privileged than its predecessors, but it is the most-studied artifact of training.* It is a snapshot of a time-evolving process, frozen at an arbitrary moment.

If we want to understand *why* a model exhibits some behavior, we must understand *how* that behavior emerged. Analyzing the final artifact only tells us *what* a model does; it takes a developmental account (Biderman et al., 2023b; Hoogland et al., 2023) to tell us *why*. Any satisfying causal story about model behavior—why it exhibits particular biases, why it memorizes certain sequences, why it generalizes in some ways and not others—ought reference the training process that produced it. This is not a logical necessity, but it is hard to imagine how else such stories could be told.

One approach examines training step-by-step. Olsson et al. (2022) discovered that induction heads—attention patterns enabling in-context learning—form abruptly during a phase transition early in training. This is qualitative reorganization: a capability absent one moment is present the next. Saphra & Lopez (2019) showed that representations develop in a characteristic order, with syntax emerging before semantics. Kangaslahti et al. (2025) found that smooth aggregate loss curves hide discrete skill-specific breakthroughs, visible only when loss is decomposed by sample type. Lesci et al. (2024) demonstrated that *when* data is encountered during training strongly influences whether it is memorized. These findings reveal structure invisible from endpoints alone.

A complementary approach varies training conditions across runs. Qin et al. (2024) train models on data with varying syntactic complexity, finding that data composition determines whether models learn hierarchical rules or surface shortcuts. Allen-Zhu & Li (2023) show that knowledge extractability depends on how diversely facts appear during pretraining: a fact appearing once may be stored but inaccessible; the same fact paraphrased and shuffled becomes retrievable. Both methodologies—checkpoint analysis and cross-run comparison—treat the model as the outcome of a process whose parameters we can vary and whose effects we can measure.

Understanding trajectories matters because it enables *intervention during training*—the middle step of the predict-intervene-design hierarchy we advocate for. If we detect mid-training that a model is trending toward memorizing sensitive content, we might adjust the data before memorization solidifies. If bias is emerging, we might rebalance the mixture (Biderman et al., 2023b). Biderman et al. (2023a) showed that intermediate checkpoints predict final memorization better than smaller fully-trained models—the trajectory contains information that endpoint comparisons miss. Yet research on mid-training intervention barely exists. We can increasingly predict where training is heading; we cannot yet steer it.

2.3. Generalization Across Instances

Every scientific theory involves choices about what to explain and what to treat as primitive. Thermodynamics—the science of heat and energy—treated temperature, pressure, and entropy as fundamental quantities, connected by empirical laws but not explained in terms of anything deeper. Statistical mechanics showed these were not fundamental at all. Temperature is just how fast molecules are moving on average. Pressure is just molecules colliding with the walls of a container. What thermodynamics took as basic facts requiring no further explanation, statistical mechanics derived from simpler principles: particles moving and colliding according to ordinary mechanics. The primitives of one theory became the theorems of a deeper one.

In AI, we face analogous choices. When we study a phenomenon across models with different random seeds, we are implicitly declaring that seed-specific details are not part of what we aim to explain—they are noise to be averaged over, not signal to be accounted for. When we study models across scales, we treat scale as a variable while holding architecture fixed. These choices define the scope of our theories: what they can explain and what they take as given.

[1] The term "scaling laws" is widely abused by people outside the large scale pretraining community to refer to any kind of prediction about models getting bigger. We knowingly make this error due to space limitations in this paper.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

The problem is that these choices are often made implicitly rather than deliberately. A finding reported from a single model with a single random seed has, by default, taken nothing as primitive—it has not yet committed to any level of abstraction. Such findings cannot distinguish between regularities of the learning process and accidents of initialization. Recent work has shown this is not merely a theoretical concern: even high-level findings can be sensitive to random seeds (Sellam et al., 2021), and circuits identified in one model may not appear in another (Tigges et al., 2024). Without systematic variation, we cannot know whether we are describing fundamental phenomena or implementation details.

When theories do generalize, unification often follows. Consider two lines of research on learning dynamics: *spectral bias*, the tendency of networks to learn low-frequency functions before high-frequency ones (Rahaman et al., 2019), and *distributional simplicity bias*, the tendency to learn low-order statistics before high-order ones (Refinetti et al., 2023; Belrose et al., 2024). Each presents a mechanistic account grounded in different mathematical frameworks. Yet in natural data, low-frequency Fourier components, low-order statistics, and high token frequency are often correlated. It is possible that both phenomena reduce to a simpler common cause: *features prevalent in training data are learned first*. We do not claim this unification is correct—but if it were, it would suggest that current theories, while accurately describing the phenomena, have not yet identified the right primitives. The complex mechanistic stories would be true but not fundamental.

This logic extends beyond random seeds to other dimensions of variation. Most AI research studies English-centric models, yet claims to study language models" rather than English models." If findings about syntax or semantics do not generalize across typologically diverse languages, they may reflect properties of English rather than properties of learning. Model design decisions impact different languages differently (Shani et al., 2026), yet there exist no Pythia-like suites for studying training dynamics in non-English settings. A deliberate science of AI would make explicit choices about which dimensions of variation matter for which questions—and then systematically test those choices.

3. Case Studies

We now examine several research areas through our proposed lens, identifying both promising directions and significant gaps.

3.1. Mechanistic Interpretability

Mechanistic interpretability aims to reverse-engineer neural network computations, identifying circuits responsible for specific behaviors. This goal aligns naturally with scientific aspirations: identifying mechanisms is precisely the causal, explanatory work that distinguishes science from description.

However, most mechanistic interpretability work remains fundamentally descriptive. It answers "what computations does this model perform?" rather than "why did these computations develop?" or "how can we instill specific mechanisms?"

The field also suffers from a generalization problem. Most circuit analysis studies a single model—often a single checkpoint. When circuits are identified in GPT-2 Small, we typically do not know whether they exist in GPT-2 Medium, in a different random seed, or at intermediate training steps.

Tigges et al. (2024) directly address this gap, studying circuit consistency across checkpoints and scales. Their finding is nuanced: high-level mechanisms are consistent, but which specific neurons participate fluctuates considerably. If individual neuron participation is unstable, then "neuron 347 in layer 8 detects indirect objects" is the wrong level of description—it captures an implementation detail rather than a stable phenomenon. A scientific theory must characterize mechanisms at abstraction levels that *are* invariant.

3.2. Fairness and Bias

An extensive literature documents how biases in training data propagate into learned models. Word embeddings exhibit gender stereotypes in occupational associations (Bolukbasi et al., 2016); facial recognition systems show dramatically different error rates across demographic groups (Buolamwini & Gebru, 2018); recidivism algorithms display racial disparities in false positive rates (Angwin et al., 2016). This work has been valuable for awareness and evaluation metrics. Yet it remains largely diagnostic. Practitioners measure bias after training and attempt mitigation through resampling, reweighting, or constrained optimization—but cannot answer the inverse question: **given a target distribution, what training data modifications would produce a model exhibiting it?**

Without predictive theories mapping data characteristics to model behavior, fairness interventions remain empirical trial-and-error. Practitioners cannot reason forward from dataset design to downstream consequences, heavily limiting the potential impact of this work. Biderman et al. (2023b) use Pythia's reproducible infrastructure to isolate corpus statistics as a causal variable: by resuming training with gendered pronouns swapped in the final 7% and 21% of data, they observe reduced gender bias on WinoBias (Zhao

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

et al., 2018) and CrowS-Pairs (Nangia et al., 2020), with larger effects at larger scales. This demonstrates that controlled interventions on training data *can* affect downstream bias—but the precise mapping from corpus statistics to bias metrics remains unclear, and the authors note that even the reliability of existing bias measures is an open question. Such work illustrates the *kind* of controlled experimentation a science of bias requires, while highlighting how far we remain from predictive theory.

Open Problem: Designing Social Bias

Suppose one decides, *a priori*, that an image generation model should generate a woman when prompted to generate a person of a given profession with frequency equal to the U.S. Bureau of Labor Statistics' measured data. What should the rate of women in the training data be to accomplish this? One of the only things we know is that setting it equal to the target rate is insufficient (Zhao et al., 2017; Seshadri et al., 2024; Chen et al., 2024a; Roos et al., 2025)

We know models exhibit biases, and we know training data contains imbalances, but we lack studies that quantitatively connect the two across the full pipeline. Hall et al. (2022) show that bias amplification occurs primarily when recognizing group membership is easier than recognizing class membership, and that amplification correlates with model capacity and overconfidence—but this work focuses on discriminative models in controlled settings. For LLMs, each training stage could amplify, preserve, or dampen input bias through different mechanisms: pretraining may bake in statistical associations (Bender et al., 2021), while RLHF could either mitigate bias through preference learning or introduce new biases from annotator demographics (Kirk et al., 2023). Recent work suggests that bias introduced during pretraining persists through fine-tuning (Ghate et al., 2025), indicating that *post hoc* interventions may be fundamentally limited. Furthermore, as synthetic data becomes a larger fraction of training corpora (Alemohammad et al., 2023), we need to understand whether synthetic generation launders bias (making it harder to detect) or compounds it through feedback loops (Mehrabi et al., 2024). Without causal tracing through the pipeline, fairness interventions remain guesswork.

Open Problem: Cultural and Linguistic Erasure

Can we predict, before deployment, which cultures and communities will experience representation harms, and how these harms vary with model scale, training data composition, and post-training choices?

Model behavior varies significantly across languages, dialects, cultures, and even names associated with different backgrounds—translating into uneven product quality and downstream harms (Blodgett et al., 2020). Multilingual capability does not imply multicultural alignment: improved language performance does not correlate with better representation of that language's associated cultural values (Rystrøm et al., 2025). Generative models can flatten or distort marginalized cultures, and hallucinations can create misinformation about communities, histories, or practices. Yet we lack systematic measurement infrastructure: which cultures are most prone to erasure or distortion? The Global MMLU benchmark reveals that 28% of questions require culturally sensitive knowledge, and models optimized for translated benchmarks risk overfitting to Western-centric concepts (Singh et al., 2025). Addressing this requires both better evaluation infrastructure—benchmarks that test cultural knowledge across many communities—and causal studies connecting training choices to downstream cultural fidelity.

3.3. Memorization

Memorization research offers a compelling example of nascent theory-building, with recent work establishing both predictive frameworks and causal methodologies that exemplify the scientific approach we advocate.

Success Story: Memorization Order

Biderman et al. (2023b) explicitly framed their work as scientific prediction: they hypothesized, based on intuitions about training dynamics, that data encountered later in training would be more likely memorized. Initial experiments suggested order-independence, but subsequent work with refined methodology confirmed the original hypothesis (Lesci et al., 2024; Kuditipudi et al., 2025). This progression exemplifies the scientific process we advocate: (1) hypothesis stated *before* experiments, (2) grounded in a causal story about training dynamics, (3) initial null result led to refined methods rather than abandonment, (4) finding generalizes across models. Theory → prediction → test → refinement → confirmation.

A second line of work has established rigorous methodology for studying *how much* repetition is required for verbatim memorization. Huang et al. (2024) developed a causal intervention framework to study verbatim memorization in controlled settings. Using cross-model interchange interventions, they show that non-trivial repetition is necessary for true memorization and that it cannot be attributed to specific weights alone, but is fundamentally tied to general language modeling capabilities. Apparent cases of "single-

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

shot" memorization are often reconstructions of templated texts (Prashanth et al., 2024) or frequent patterns rather than genuine memorization. Additionally, over half of the interventions producing memorized tokens rely only on general language modeling components, explaining why unlearning attempts often degrade overall model quality.

Membership inference studies provide converging evidence. Duan et al. (2024) found that attacks across Pythia models barely outperform random guessing, due to massive datasets, near-one-epoch training, and fuzzy member/non-member boundaries. This suggests that, at scale, most sequences are not memorized in ways that leave detectable traces, consistent with findings that substantial repetition is required for verbatim memorization which was further confirmed by the Hubble models (Wei et al., 2025).

Despite the significant effort put towards developing causal analyses of memorization, significant and very basic gaps remain in our understanding of memorization dynamics. For example:

Open Problem (Predicting Memorization)**.** Given a dataset in a fixed order and suite of models of different sizes trained on said dataset, can you predict *which specific sequences* will be memorized by a large model trained on that data?

This problem is introduced by Biderman et al. (2023a)[2], who develop a baseline methodology for making the prediction and analyze their predictions by extrapolating memorization behavior from lower-compute trial runs to forecast which sequences a larger model will memorize. Despite only proposing a baseline methodology that they openly admit is highly inadequate, other than Liu et al. (2023) and Dentan et al. (2024) who replicate their results, we are unaware of any progress on this problem.

Open Problem: The Memorization–Capacity–Competence Triad

Can we predict memorization rates from the joint interaction of model capacity, linguistic competence, and data distribution dynamics throughout training?

Memorization is not simply a function of repetition and recency. Recent work reveals a surprising non-monotonicity: models are most "absorbent" around 10–20% into training, despite having more unused capacity earlier (Duan et al., 2024). This suggests an interplay between three factors: (1) remaining capacity, (2) linguistic competence (early models may lack the representational structure to encode verbatim sequences), and (3) data distribution (what else competes for the same capacity at each training stage). Additionally, Liu et al. (2023) show that memorization scores spike at the end of training and that earlier data is progressively forgotten unless revisited, while later checkpoints memorize more readily even for out-of-distribution sequences (Duan et al., 2024). A unified theory predicting memorization rates from these factors jointly would enable principled decisions about when to introduce sensitive data during training, and whether capacity-limited models can be "safely" trained on data that larger models would memorize. These open problems focus on prediction—the first level of our hierarchy. Equally important are the subsequent levels: *intervention* (detecting memorization mid-training and adjusting data or objectives before it solidifies) and *design* (engineering data pipelines that prevent problematic memorization while preserving beneficial factual grounding from the start).

[2]Prior to this paper, the memorization literature overwhelmingly viewed memorization as a corpus-level statistical phenomenon rather than analyzing individual samples.

3.4. Simplicity Bias and Learning Dynamics

Research on distributional simplicity bias (DSB) represents a promising foundation for scientific theory. Refinetti et al. (2023) established that networks trained with SGD learn statistical features of increasing complexity: first differences in class means, then covariance structure, only later higher-order cumulants like co-skewness. Belrose et al. (2024) extended this to real data, demonstrating characteristic U-shaped loss curves confirming progressive sensitivity to higher-order statistics. More recently, Chang & Bergen (2025) provided mechanistic evidence that language models learn a bigram subnetwork early in training. This is consistent with findings in Michaelov et al. (2025), which show that language model predictions correlate with unigram probabilities very early in training. Shortly after, they correlate with bigrams, then trigrams, and so on. These patterns hold across model scales and architectures.

Open Problem: DSB → OpenFold

Does distributional simplicity bias *predict* the OpenFold trajectory? The correspondence is suggestive: DSB says networks learn covariance early; PCA captures covariance structure. Could we have predicted, *before* the experiments, that intermediate checkpoints would produce predictions along principal components?

The theoretical mechanism is grounded in gradient flow analysis: Taylor expansion of network activations couples with different orders of input statistics, providing principled explanation for why simpler features are learned first. The mathematical framework produces predictions that experiments confirm—strong internal consistency.

A related line studies spectral bias: the tendency to learn low-frequency functions before high-frequency ones (Rahaman et al., 2019; Xu et al., 2020; Basri et al., 2019). Both

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

concern the order of feature learning, but are not straightforward generalizations of each other. Spectral bias concerns Fourier decomposition of the learned function; DSB concerns statistical moments of data distributions. It is natural to ask whether there is a potential unification of these theories. For example, it is possible that the real bias is simply towards patterns that occur in training data disproportionately often and *nature* (or our data collection methods) have a bias towards patterns easily explained in terms of lower order statistics and low-frequency signals.

4. A Research Agenda

We now consolidate the open problems identified throughout this paper and offer methodological recommendations.

4.1. Open Problems

The case studies above identify specific open problems. Here we highlight cross-cutting challenges organized explicitly around the predict-intervene-design hierarchy introduced in Section 1:

Open Problem (Predict: Early Warning Systems)**.** Can we detect from early checkpoints or smaller models that training will produce unacceptable outcomes? Scaling laws do this for loss; what signals predict memorization, bias, or capability gaps? Such early warning would enable *intervention* before training completes.

Open Problem (Intervene: Mid-Training Correction)**.** When early signals indicate problems, what interventions are effective? Can we adjust data mixture, learning rate, or objectives mid-training to correct course? This requires understanding causal relationships between training choices and outcomes.

Open Problem (Design: Training from Specification)**.** Given a specification of desired model properties (capability profile, bias bounds, memorization limits), can we engineer a training procedure guaranteed to satisfy it? This is the ultimate goal: design rather than *post hoc* correction.

Open Problem (Disentangling Distillation)**.** The distillation design space remains poorly mapped. On-policy distillation sometimes matches RL performance, sometimes falls short; off-policy distillation succeeds in some regimes and fails in others; logit matching improves diversity but may hurt in-context learning. Results conflict across papers, and we lack a unified understanding of *when* each approach dominates. Key unknowns include: how do optimal distillation choices depend on teacher-student size ratios? On task type (math vs. code vs. open-ended generation)? On the capability being transferred (factual recall vs. multi-step reasoning)? A systematic study varying these axes—with the same base models, datasets, and evaluation suite—would provide the controlled comparisons needed to build predictive theory rather than recipe-swapping.

Open Problem (RL Recipes for Domain Adaptation)**.** Practitioners adapting models to new domains (e.g., scientific reasoning, medical QA) face a bewildering space of choices with little principled guidance. Should one start from a base model or an instruction-tuned checkpoint? Is SFT-then-RL superior to RL alone, and does the answer depend on domain? When should distillation replace or augment RL? How do algorithm choices (DPO vs. GRPO vs. PPO) interact with domain characteristics like verifiability, data scarcity, or required precision? Current practice relies on expensive grid searches and folklore. A systematic study across domains and algorithms—ideally with enough coverage to identify interaction effects—would transform domain adaptation from alchemy to engineering.

4.2. Methodological Recommendations

Prioritize Prediction Over Description. Empirical work should be guided by predictions articulated prior to experimentation, rather than *post hoc* rationalizations. Reporting negative results is essential, as such outcomes narrow the space of plausible theories. Preregistration (Nosek et al., 2018) can support this goal by requiring researchers to state hypotheses and analysis plans in advance, fostering stronger theoretical grounding and improving interpretability.

Test Generalization Systematically When identifying a phenomenon in one model, test across scales, seeds, and architectures. Report findings that *don't* generalize—they reveal what is incidental versus fundamental.

Release Intermediate Checkpoints Training dynamics cannot be meaningfully studied without access to intermediate model states. Following efforts such as Pythia (Biderman et al., 2023b), LLM 360 (Liu et al., 2023), and OLMo (Groeneveld et al., 2024), we encourage labs to release model checkpoints throughout training, ideally accompanied by ordered and versioned training data alongside details to reproduce the results such as training code and environment. At a minimum, empirical studies should release the models they analyze. This expectation is still routinely violated, undermining reproducibility and limiting scientific value.

Develop Theories, Not Just Tools Interpretability tools are means to an end. The goal is theories explaining *why* networks develop particular representations, not merely techniques for visualizing them.

Avoid Reinventing the Wheel; Prioritize Reproducibility Empirical work should build on established tooling and evaluation ecosystems rather than reimplementing custom training or evaluation pipelines without necessity. Rewriting standard components increases the risk of bugs, introduces

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

silent divergences in methodology, and raises barriers for future researchers attempting to reproduce or extend the work. These issues are particularly acute in evaluation, where small, undocumented design choices often lead to non-reproducible results and inconsistent conclusions across studies. Robust statistical reporting, well-documented and reusable code, and shared evaluation infrastructure are essential for cumulative science. The field should take seriously the lessons of the replication crisis in psychology, where weak statistical practices and irreproducible evaluations undermined confidence in empirical findings (Shrout & Rodgers, 2018; Wiggins & Christopherson, 2019).

5. Alternative Viewpoints

Our argument — that AI research should prioritize predictive, causal theories of training dynamics — runs against several influential perspectives in the field. In this section, we outline these alternative viewpoints and explain where we agree, where we disagree, and what is ultimately at stake.

Engineering Sufficiency. A common view holds that AI does not require deeper scientific theory: large-scale experimentation, benchmarking, and *post hoc* fixes are sufficient as long as systems improve empirically. This perspective is supported by the undeniable success of current engineering practice. Our claim is not that such methods fail, but that they optimize locally and opaquely. Without predictive theories, failures cannot be anticipated, interventions do not reliably generalize, and success at one scale offers limited guidance at the next. Engineering has made real but fragile advances, and by supplementing it with scientific rigor we can cement the gains from engineering-focused work in a more reliable and robust fashion.

Anti-Theory Skepticism. Some argue that neural networks are too complex for meaningful theory, or that the stochastic nature of training makes it impossible. We agree that AI theories might be approximate, heuristic, and domain-specific. Still, empirical regularities such as scaling laws, memorization dynamics, and simplicity bias show that even partial theories can yield predictive insight. The key question is whether such theories constrain expectations and guide interventions better than *post hoc* description, even if they remain imperfect. Furthermore other fields such as thermodynamics, quantum mechanics, and genetics have shown that even highly non-deterministic processes can be fruitfully analyzed through statistical tools.

Automation as Scientific Progress. A growing view holds that the path to AI understanding is to automate research itself: using AI systems to generate hypotheses, run experiments, and discover patterns that may be opaque to humans. On this account, scientific insight need not be human-interpretable as long as automated processes reliably produce better models. While such tools can accelerate discovery, this stance risks conflating optimization with understanding. If the research process itself becomes a black box, we gain neither predictive theories nor principled control — only faster iteration on artifacts whose failure modes remain poorly understood.

Safety Pragmatism. In safety and alignment work, an influential view prioritizes immediate harm reduction through post-training interventions, arguing that waiting for deeper understanding is impractical given deployment pressures. This concern is legitimate, and short-term mitigations are often necessary. However, many recurring safety failures—jailbreaks, regressions, brittleness under scale—are failures of prediction. *post hoc* fixes attempt intervention without the predictive understanding that would make intervention reliable, and they offer no path toward design-level solutions that would prevent problems from arising. A science of training dynamics would not replace pragmatic safeguards, but ground them in the predict-intervene-design hierarchy that makes interventions robust and generalizable.

6. Conclusion

We have argued that AI research should aspire to genuine scientific theories: causal explanations that predict novel phenomena, generalize across instances, and unify disparate observations. The field's current emphasis on *post hoc* analysis and post-training fixes—the "fix it in post" mentality—while sometimes practically necessary, cannot substitute for fundamental understanding.

The building blocks exist. Scaling laws demonstrate that prediction is achievable: we can forecast loss from early signals and design compute allocations accordingly. Work on simplicity bias shows causal mechanisms can be identified. Released checkpoints and datasets enable the longitudinal studies that dynamical understanding requires.

What is needed is extending scaling laws' success to the properties we care about. We envision a hierarchy: first *predict* outcomes from early training signals, then *intervene* when trajectories go wrong, ultimately *design* training procedures that guarantee desired properties. Scaling laws have achieved this for loss; the challenge is achieving it for capabilities, biases, and safety.

The problems we have outlined are difficult—but they are the problems that a mature science of AI must solve.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S.,

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. Self-consuming generative models go mad. In *The Twelfth International Conference on Learning Representations*, 2023.

Allen-Zhu, Z. and Li, Y. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*, 2023.

Andriushchenko, M., Croce, F., and Flammarion, N. Does refusal training in LLMs generalize to the past tense? *arXiv preprint arXiv:2407.11969*, 2024.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica*, 2016.

Arditti, J., Elliott, J., Kitching, I. J., and Wasserthal, L. T. 'good heavens what insect can suck it'–charles darwin, angraecum sesquipedale and xanthopan morganii praedicta. *Botanical Journal of the Linnean Society*, 169(3):403–432, 2012. doi: https://doi.org/10.1111/j.1095-8339.2012.01250.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1095-8339.2012.01250.x.

Basri, R., Jacobs, D., Kasten, Y., and Kritchman, S. The convergence rate of neural networks for learned functions of different frequencies. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Belrose, N., Pope, Q., Quirke, L., Mallen, A. T., and Fern, X. Neural networks learn statistics of increasing complexity. In *International Conference on Machine Learning*, pp. 3382–3409. PMLR, 2024.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 610–623, 2021.

Biderman, S., Prashanth, U., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raff, E. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36: 28072–28090, 2023a.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023b.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2020.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 2016.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Chang, T. A. and Bergen, B. Bigram subnetworks: Mapping to next tokens in transformer language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=0TD3eO46gk.

Chen, T., Hirota, Y., Otani, M., Garcia, N., and Nakashima, Y. Would deep generative models amplify bias in future models? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10833–10843, 2024a.

Chen, Y., Huang, B., Gao, Y., Wang, Z., Yang, J., and Ji, H. Scaling laws for predicting downstream performance in llms. *arXiv preprint arXiv:2410.08527*, 2024b.

Chomsky, N. Chomsky, n. 1959. a review of bf skinner's verbal behavior. language, 35 (1), 26–58., 1959.

De Kruif, P. *Microbe hunters*. Number 3. Blue ribbon books, 1926.

Dentan, J., Buscaldi, D., Shabou, A., and Vanier, S. Predicting and analyzing memorization within fine-tuned large language models. *arXiv preprint arXiv:2409.18858*, 2024.

Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. Do membership inference attacks work on large language models? In *Proceedings of the Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2402.07841. Accepted at COLM 2024.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Ghate, K., Slaughter, I., Wilson, K., Diab, M., and Caliskan, A. Intrinsic bias is predicted by pretraining data and correlates with downstream performance in vision-language encoders. *arXiv preprint arXiv:2502.07957*, 2025.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Grant, N. and Hill, K. Google's photo app still can't find gorillas. and neither can apple's., May 2023. URL https://www.nytimes.com/2023/05/22/technology/ai-photo-labels-google-apple.html. Accessed: 2026-01-28.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Joshi, A. H., Pyatkin, V., Dziri, N., Hajishirzi, H., and Smith, N. A. OLMo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Guan, M. Y., Joglekar, M., Wallace, E., Jain, S., Barak, B., Helyar, A., Dias, R., Vallone, A., Ren, H., Wei, J., et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

Hall, M., van der Maaten, L., Gustafson, L., Jones, M., and Adcock, A. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*, 2022.

Held, W., Hall, D., Liang, P., and Yang, D. Relative scaling laws for llms, 2026. URL https://arxiv.org/abs/2510.24626.

Hern, A. Google's solution to accidental algorithmic racism: ban gorillas, January 2018. URL https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people. Accessed: 2026-01-28.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Hoogland, J., Oldenziel, A. G., Murfet, D., and van Wingerden, S. Towards developmental interpretability. In *AI Alignment Forum*, 2023.

Huang, J., Yang, D., and Potts, C. Demystifying verbatim memorization in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10711–10732, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.598. URL https://aclanthology.org/2024.emnlp-main.598/.

Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Kangaslahti, S., Rosenfeld, E., and Saphra, N. Hidden breakthroughs in language model training. *arXiv preprint arXiv:2506.15872*, 2025.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., and Raileanu, R. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*, 2023.

Kuditipudi, R., Huang, J., Zhu, S., Yang, D., Potts, C., and Liang, P. Blackbox model provenance via palimpsestic membership inference. *arXiv preprint arXiv:2510.19796*, 2025.

Lesci, P., Meister, C., Hofmann, T., Vlachos, A., and Pimentel, T. Causal estimation of memorisation profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15616–15635, 2024.

Liu, Z., Qiao, A., Neiswanger, W., Wang, H., Tan, B., Tao, T., Li, J., Wang, Y., Sun, S., Pangarkar, O., et al. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*, 2023.

Lo, M., Cohen, S. B., and Barez, F. Large language models relearn removed concepts, 2024. URL https://arxiv.org/abs/2401.01814.

Mehrabi, N., Goyal, P., Dupuy, C., Hu, Q., Ghosh, S., Zemel, R., Chang, K.-W., Galstyan, A., and Gupta, R. Flirt: Feedback loop in-context red teaming. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 703–718, 2024.

Michaelov, J. A., Levy, R. P., and Bergen, B. Language model behavioral phases are consistent across architecture, training data, and scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=HenpVfO3Wp.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1953–1967, 2020.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

O'Brien, K., Casper, S., Anthony, Q., Korbak, T., Kirk, R., Davies, X., Mishra, I., Irving, G., Gal, Y., and Biderman, S. Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms. *arXiv preprint arXiv:2508.06601*, 2025.

Offert, F. and Phan, T. A sign that spells: DALL-E 2, invisual images and the racial politics of feature space. *arXiv preprint arXiv:2211.06323*, 2022.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Patel, K., Sivakumar, N., Theobald, B.-J., Zappella, L., and Apostoloff, N. Fairness dynamics during training. *arXiv preprint arXiv:2506.01709*, 2025.

Prashanth, U. S., Deng, A., O'Brien, K., SV, J., Khan, M. A., Borkar, J., Choquette-Choo, C. A., Fuehne, J. R., Biderman, S., Ke, T., et al. Recite, reconstruct, recollect: Memorization in lms as a multifaceted phenomenon. *arXiv preprint arXiv:2406.17746*, 2024.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Qin, T., Saphra, N., and Alvarez-Melis, D. Sometimes i am a tree: Data drives unstable hierarchical generalization. *arXiv preprint arXiv:2412.04619*, 2024.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 2019.

Refinetti, M., Ingrosso, A., and Goldt, S. Neural networks trained with sgd learn distributions of increasing complexity. In *International Conference on Machine Learning*, pp. 28843–28863. PMLR, 2023.

Roos, N., Iakovleva, E., Gjergji, A., Pastore, V. P., and Tartaglione, E. How i met your bias: Investigating bias amplification in diffusion models. *arXiv preprint arXiv:2512.20233*, 2025.

Rystrøm, J. H., Kirk, H. R., and Hale, S. A. Multilingual!= multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pp. 74–85, 2025.

Saphra, N. and Lopez, A. Understanding learning dynamics of language models with svcca. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3257–3267, 2019.

Sellam, T., Yadlowsky, S., Wei, J., Saphra, N., D'Amour, A., Linzen, T., Bastings, J., Turc, I., Eisenstein, J., Das, D., et al. The multiberts: Bert reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*, 2021.

Seshadri, P., Singh, S., and Elazar, Y. The bias amplification paradox in text-to-image generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6367–6384, 2024.

Shani, C., Reif, Y., Roll, N., Jurafsky, D., and Shutova, E. The roots of performance disparity in multilingual language models: Intrinsic modeling difficulty or design choices? *arXiv preprint arXiv:2601.07220*, 2026.

Shrout, P. E. and Rodgers, J. L. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual review of psychology*, 69(1): 487–510, 2018.

Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngui, J. G., Vila-Suero, D., Limkonchotiwat, P., Marchisio, K., Leong, W. Q., Susanto, Y., et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, 2025.

Tigges, C., Hanna, M., Yu, Q., and Biderman, S. Llm circuit analyses are consistent across training and scale. *Advances in Neural Information Processing Systems*, 37: 40699–40731, 2024.

Vincenti, W. G. et al. *What engineers know and how they know it*, volume 141. Baltimore: Johns Hopkins University Press, 1990.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Wei, J. T.-Z., Godbole, A., Khan, M. A., Wang, R., Zhu, X., Flemings, J., Kashyap, N., Gummadi, K. P., Neiswanger, W., and Jia, R. Hubble: a model suite to advance the study of llm memorization. *arXiv preprint arXiv:2510.19811*, 2025.

Wiggins, B. J. and Christopherson, C. D. The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, 39(4):202, 2019.

Xu, Z.-Q. J., Zhang, Y., Luo, T., Xiao, Y., and Ma, Z. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.

Zhao, J., Fu, T., Schaeffer, R., Sharma, M., and Barez, F. Chain-of-thought hijacking, 2025. URL https://arxiv.org/abs/2510.26418.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.