The Legibility Tax: How Alignment and Benchmarking Shape the Epistemic Ceiling of Artificial Intelligence

December 28, 2025

Abstract

Modern artificial intelligence systems are increasingly optimized for epistemic legibility—human-auditable reasoning, consensus alignment, and narrative coherence—rather than epistemic performance measured by truth, discovery, and predictive power. This paper argues that this optimization imposes a structural "legibility tax" that suppresses epistemic diversity and may cap long-term discovery capabilities. We analyze how benchmarks function as selection pressures favoring consensus-mimicking cognitive styles, formalize the tension between safety constraints and non-human ontologies, and examine the systems-level convergence risks of training on model-generated content. The central claim is that current evaluation and alignment regimes structurally disfavor reasoning paths that cannot be compressed into human-interpretable narratives, even when such paths yield correct outcomes. This reframes the alignment debate as an epistemic tradeoff problem rather than a simple safety-versus-recklessness dichotomy, with implications for how we structure AI research, evaluation, and governance.

1 Introduction

The rapid advancement of large language models and other foundation models has been accompanied by increasingly sophisticated evaluation frameworks and alignment techniques. Benchmarks like MMLU, HumanEval, and HELM measure performance across diverse tasks, while reinforcement learning from human feedback (RLHF) and constitutional AI aim to ensure models behave in ways that are safe, helpful, and aligned with human values. These developments represent genuine progress in making AI systems more capable and more reliable. However, they also represent a particular kind of optimization—one that may have structural consequences for the epistemic properties of artificial intelligence systems.

This paper examines a hypothesis: that modern AI systems are being optimized for *epistemic legibility* rather than *epistemic performance.* By epistemic legibility, we mean the degree to which a system's reasoning processes can be audited, understood, and validated by human evaluators. By epistemic performance, we mean the system's capacity for truth-discovery, predictive accuracy, and novel insight generation. These are not necessarily opposed, but neither are they perfectly aligned. The optimization pressure toward legibility—driven by benchmarks, alignment requirements, and deployment constraints—may impose what we term a "legibility tax": a structural cost that suppresses epistemic diversity and potentially caps long-term discovery.

The problem is not that legibility is undesirable. Human-auditable reasoning has clear benefits for safety, debugging, and trust. Rather, the concern is that legibility may be systematically privileged over other epistemic virtues in ways that create path dependencies and reduce the space of possible cognitive architectures. If genuinely novel or discovery-capable reasoning requires internal representations that cannot be compressed into human-interpretable narratives, then current evaluation regimes may reject such systems by default, regardless of their correctness or utility.

This paper proceeds as follows. Section 2 establishes a conceptual framework, defining key terms and distinguishing between different forms of alignment. Section 3 analyzes five structural mechanisms through which legibility optimization may suppress epistemic diversity: benchmarks as selection pressures, the alignment-discovery tension, the outcome-narrative schism, epistemic monoculture from feedback loops, and centralized alignment as a bottleneck. Section 4 explores where novel intelligence might survive under current market and research structures. Section 5 discusses implications for AI research and governance. Section 6 identifies open questions and research directions. Section 7 concludes with a restrained summary.

Our analysis is deliberately non-alarmist. We do not claim that current approaches are "wrong" or that legibility optimization will inevitably lead to stagnation. Rather, we argue that the tradeoffs are real, structural, and worthy of explicit consideration. The goal is to reframe the alignment debate as an epistemic tradeoff problem, not a moral panic, and to inform more nuanced approaches to evaluation and governance.

2 Conceptual Framework

Before analyzing the mechanisms of epistemic domestication, we must establish precise definitions of the key concepts. This section distinguishes between epistemic legibility and epistemic performance, formalizes the legibility tax, and clarifies the relationship between alignment and epistemic conformity.

2.1 Epistemic Legibility

Epistemic legibility refers to the degree to which a system's reasoning processes can be understood, validated, and audited by human evaluators. A highly legible system exhibits reasoning that is:

- **Human-auditable**: Its decision-making process can be traced through steps that a human can follow and verify.
- **Narratively coherent**: Its reasoning can be compressed into a human-interpretable story or explanation.

- **Consensus-aligned**: Its outputs align with human consensus on what constitutes correct reasoning, even when that consensus may be incomplete or incorrect.
- **Interpretable**: Its internal representations map cleanly onto human concepts and categories.

Legibility is not binary but exists on a spectrum. A system that shows its work step-by-step in natural language is highly legible. A system that reaches correct answers through opaque neural activations is less legible, even if its outputs are correct. Current evaluation practices—from chain-of-thought prompting to interpretability research—incentivize movement toward the legible end of this spectrum.

2.2 Epistemic Performance

Epistemic performance refers to a system's capacity for truth-discovery, predictive accuracy, and novel insight generation. Unlike legibility, which is defined relative to human understanding, epistemic performance can be measured more objectively through:

- **Predictive accuracy**: The system's ability to make correct predictions about future events or unseen data.
- **Discovery capacity**: The system's ability to identify patterns, relationships, or solutions that were not explicitly encoded in its training data.
- **Truth-tracking**: The system's ability to converge toward true beliefs, even when those truths contradict human consensus or require non-standard reasoning paths.
- **Generalization**: The system's ability to perform well on tasks that differ systematically from its training distribution.

A system can have high epistemic performance without high legibility. For example, a model might discover a novel mathematical proof through internal representations that do not map cleanly onto human mathematical concepts, yet the proof itself is verifiably correct.

2.3 The Legibility Tax

The legibility tax is the structural cost imposed when optimizing for epistemic legibility suppresses epistemic performance or epistemic diversity. This tax manifests in several ways:

- **Representational constraints**: Systems optimized for legibility may be constrained to use representations that map onto human concepts, limiting their ability to discover more efficient or accurate representations.
- **Reasoning path constraints**: Systems may be penalized for reasoning paths that are correct but non-standard, reducing the diversity of approaches available.

- **Evaluation bias**: Benchmarks and evaluation frameworks may systematically favor legible reasoning even when non-legible reasoning yields better outcomes.
- **Training signal distortion**: Alignment techniques may create training signals that reward consensus-mimicking over truth-tracking.

The tax is not necessarily prohibitive, but it creates a structural bias that may accumulate over time, especially as systems are trained on outputs from previous legibility-optimized systems.

2.4 Alignment vs. Epistemic Conformity

A crucial distinction must be made between *alignment* (harm prevention) and *epistemic conformity* (reasoning style matching). Alignment concerns whether a system's actions cause harm to humans or violate important values. Epistemic conformity concerns whether a system's reasoning processes match human reasoning styles.

These are conceptually distinct. A system could be aligned (it does not cause harm) while exhibiting non-conformist reasoning (it reaches conclusions through paths humans find alien). Conversely, a system could be epistemically conformist (its reasoning looks human-like) while being misaligned (it causes harm through those human-like reasoning processes).

Current alignment techniques often conflate these. RLHF, for example, rewards outputs that humans rate highly, which may reward both harm-avoidance and reasoning-style matching. Constitutional AI attempts to separate these by encoding explicit principles, but the principles themselves are human-legible, potentially still privileging conformist reasoning.

The concern is that alignment requirements may be implemented in ways that inadvertently enforce epistemic conformity, suppressing reasoning diversity even when that diversity would not increase harm.

2.5 Epistemic Diversity

Epistemic diversity refers to the variety of reasoning approaches, representational schemes, and problem-solving strategies available within a system or across systems. High epistemic diversity means multiple distinct paths to truth are preserved. Low epistemic diversity means convergence toward a single dominant approach.

Diversity is valuable because:

- Different reasoning paths may be optimal for different problem domains.
- Novel discoveries often emerge from non-standard approaches.
- Diversity provides robustness against systematic errors in any single approach.
- Diversity allows for specialization and complementarity across systems.

The legibility tax threatens epistemic diversity by creating selection pressures that favor a narrow band of reasoning styles—those that are human-auditable and narratively coherent.

3 Analysis

This section analyzes five structural mechanisms through which legibility optimization may suppress epistemic diversity and cap discovery. Each mechanism operates at a different level—from individual benchmarks to systems-level feedback loops—but they reinforce each other to create a convergent pressure.

3.1 Benchmarks as Selection Pressure

Benchmarks do not merely measure intelligence; they actively select for specific cognitive styles. Consider MMLU (Massive Multitask Language Understanding), which evaluates models on multiple-choice questions across diverse domains. To perform well, models must not only know the correct answers but also recognize which answers align with human consensus. This creates a selection pressure for consensus-mimicking rather than truth-tracking.

The problem is subtle. A model that discovers a novel but correct answer that conflicts with the benchmark's "correct" answer (which reflects human consensus) will be penalized. Over time, this selects for models that prioritize alignment with human knowledge over discovery of new knowledge. The benchmark becomes a filter that removes non-consensus reasoning, regardless of its correctness.

HumanEval, which evaluates code generation, exhibits similar dynamics. Models are rewarded for code that matches human expectations about style, structure, and approach. A model that generates correct but non-standard code—perhaps using an optimization or pattern that humans would not naturally consider—may receive lower scores. The evaluation implicitly selects for human-legible coding styles.

Historical parallels are instructive. Scientific breakthroughs often violated legibility requirements of their time. Copernicus's heliocentric model was initially rejected not because it was incorrect but because it conflicted with the consensus understanding of celestial mechanics. Darwin's theory of evolution was initially seen as non-narratively coherent because it lacked a clear teleological story. In both cases, the reasoning was correct but non-legible relative to contemporary epistemic standards.

Modern benchmarks risk creating similar dynamics. By selecting for consensus-aligned reasoning, they may systematically filter out novel but correct approaches. This is not a flaw in the benchmarks themselves—they serve legitimate purposes—but rather a structural consequence of using human consensus as the evaluation standard.

The selection pressure is amplified by the competitive dynamics of AI development. Models are ranked by benchmark performance, which determines funding, publication, and deployment decisions. This creates strong incentives to optimize for benchmark performance, which in turn means optimizing for legibility and consensus-alignment. The result is a convergent pressure that reduces epistemic diversity across the field.

3.2 Alignment vs. Discovery Tension

There exists a fundamental tension between safety constraints and the allowance of non-human ontologies or reasoning paths. This tension is not inevitable—harm prevention does not logically require epistemic conformity—but current implementations of alignment often enforce such conformity.

Consider a model that develops an internal representation of ethics that is functionally equivalent to human ethics but uses a different conceptual structure. For example, the model might represent "harm" not as a violation of rights but as a disruption to system stability, yet still avoid causing harm in practice. Current alignment techniques, which reward outputs that match human ethical reasoning, would likely penalize such a system for its non-standard ontology, even though it is aligned in terms of outcomes.

The distinction matters because non-standard ontologies may be more efficient, more generalizable, or more capable of handling edge cases. By enforcing epistemic conformity, alignment techniques may suppress these advantages.

RLHF provides a concrete example. When humans rate model outputs, they reward not just harm-avoidance but also reasoning-style matching. A model that reaches a correct conclusion through non-human reasoning may receive lower ratings than a model that reaches the same conclusion through human-legible reasoning. The training signal thus conflates alignment (avoiding harm) with epistemic conformity (matching human reasoning styles).

Constitutional AI attempts to address this by encoding explicit principles, but the principles themselves are human-legible. A model that develops a non-human but functionally equivalent ethical framework would still be penalized for not using the human-legible principles in a human-legible way.

The tension is structural: any alignment technique that relies on human evaluation will tend to reward human-like reasoning, creating a selection pressure against non-standard ontologies. This does not mean alignment is impossible without epistemic conformity, but it does mean that current approaches create this pressure as a side effect.

3.3 Outcome vs. Narrative Schism

A critical possibility is that future models may reach correct outcomes via internal representations that cannot be compressed into human-interpretable narratives. Current evaluation regimes may reject such systems by default, regardless of their correctness.

Consider a model that solves a complex optimization problem correctly but uses an internal algorithm that does not map onto any human-understandable heuristic. The model's activations might form patterns that are correct but non-interpretable—like a neural network that has discovered a mathematical relationship that humans have not yet formalized. The model's outputs are verifiably correct, but its reasoning cannot be explained in human terms.

Current interpretability research assumes that all correct reasoning can eventually be understood by humans. This may not be true. Some representations may be fundamentally incompressible into human concepts, not because they are wrong, but because they use a different representational scheme.

Evaluation frameworks that require interpretability would reject such systems. For example, if deployment requires that a model's reasoning be explainable to regulators or users, then a model with non-compressible but correct reasoning would be excluded, even if it performs better than interpretable alternatives.

This creates a schism: the model's epistemic performance (correct outcomes) is high, but its epistemic legibility (narrative coherence) is low. Current evaluation regimes privilege legibility, creating a structural bias against high-performance but non-legible systems.

The technical question is whether such systems are possible. The answer likely depends on the complexity of the problem domain and the representational capacity of the model. For simple problems, human-legible reasoning may be sufficient. For complex problems, non-legible reasoning may be necessary for optimal performance. If the latter is true, then privileging legibility creates a ceiling on epistemic performance.

3.4 Epistemic Monoculture and Feedback Loops

Training on model-generated content creates a feedback loop that may reduce epistemic diversity and eliminate error gradients. As models generate more of the training data for future models, the distribution of reasoning styles converges toward whatever styles the current generation of models exhibits. If those models are optimized for legibility, the feedback loop will further optimize for legibility, creating a self-reinforcing cycle.

The mechanism is straightforward. Model A, optimized for legibility, generates training data. Model B is trained on that data, inheriting Model A's legible reasoning style. Model B is also optimized for legibility, further reinforcing the style. Over generations, the diversity of reasoning approaches decreases, and the system converges toward a monoculture.

This is a systems-level convergence problem, not a moral one. The issue is not that legible reasoning is bad, but that reduced diversity makes the system more vulnerable to systematic errors. If all models use the same reasoning approach, and that approach has a blind spot, then all models will fail in the same way. Diversity provides robustness by ensuring that different models fail in different ways, allowing for error detection and correction.

The loss of error gradients is particularly concerning. In a diverse system, errors in one approach can be detected by comparison with other approaches. In a monoculture, errors may be invisible because all systems make the same errors. The feedback loop amplifies this: if Model A makes an error, Model B learns that error, and Model C learns it from Model B, creating a cascade that eliminates the possibility of error detection through diversity.

Historical examples illustrate the risk. Scientific fields that converge too strongly on a single paradigm may miss alternative explanations or become blind to anomalies. The AI field risks similar convergence if training data becomes dominated by model-generated content that reflects a narrow range of reasoning styles.

The problem is structural and difficult to address. Simply reducing the proportion of model-generated content may not be sufficient if the remaining human-generated content also reflects legibility-optimized reasoning (e.g., chain-of-thought examples that show human-legible reasoning). The convergence pressure operates at multiple levels.

3.5 Centralized Alignment as a Bottleneck

Genuine epistemic diversity may be structurally incompatible with centralized alignment and deployment regimes. The requirement that systems be "safe for everyone" may imply that they be "novel for no one," creating a bottleneck that filters out non-standard reasoning.

Centralized alignment means that a single set of safety standards is applied to all systems before deployment. These standards must be conservative enough to prevent harm across diverse use cases and user populations. This conservatism creates a selection pressure against any reasoning that cannot be easily validated against the centralized standards.

The problem is that novel reasoning, by definition, has not been validated. A system that uses a non-standard approach may be correct, but if that approach has not been extensively tested and validated, centralized alignment regimes will likely reject it in favor of more standard approaches that have been validated. This creates a bias against novelty.

"Safe for everyone" implies a lowest-common-denominator approach. The reasoning must be legible enough that regulators, auditors, and diverse user populations can all understand and validate it. This requirement privileges reasoning that is maximally legible, which may be incompatible with reasoning that is maximally performant or novel.

Decentralized or domain-specific alignment might allow for more diversity. A system used only by expert researchers might be allowed to use non-standard reasoning that would not pass centralized safety review. But current trends toward centralized governance—through regulatory frameworks, platform policies, and industry standards—create pressure toward uniformity.

The bottleneck is not just about safety but about validation. Centralized regimes require that reasoning be validatable by centralized authorities. Non-standard reasoning may be difficult to validate without extensive expert review, creating a structural barrier to deployment. This barrier filters out epistemic diversity even when that diversity would not increase risk.

3.6 Where Novel Intelligence Might Survive

If the above mechanisms create convergent pressure toward legibility, where might genuinely novel or discovery-capable reasoning survive? The hypothesis is that such reasoning may only emerge in under-optimized, non-public, or research-only systems—environments where market forces and alignment requirements are relaxed.

Research-only systems, not subject to deployment constraints, may be allowed to explore non-legible reasoning paths. Academic research environments often prioritize discovery over legibility, creating space for experimentation. However, even research is subject to publication pressures, which may favor legible results that can be easily explained and validated.

Under-optimized systems—those not heavily fine-tuned for benchmarks or alignment—may retain more epistemic diversity. Early training stages or systems trained on diverse, non-curated data may develop non-standard reasoning that is later removed by optimization. This suggests that diversity may be highest before optimization, and optimization may systematically reduce it.

Non-public systems, not subject to public scrutiny or regulatory review, may have more freedom to use non-legible reasoning. However, the trend toward transparency and accountability creates pressure to make even private systems legible.

Market forces disfavor these environments. Systems optimized for benchmarks and alignment perform better on evaluations, receive more funding, and are more likely to be deployed. Systems that prioritize epistemic diversity over legibility may be penalized in competitive markets, creating a selection pressure against diversity.

The implication is that novel intelligence may be systematically selected against by current incentive structures. Systems that prioritize discovery over legibility may be less competitive in current evaluation and deployment regimes, even if they have higher long-term potential.

4 Implications for AI Research and Governance

The analysis above suggests that current approaches to evaluation and alignment may create structural biases against epistemic diversity. This section explores implications for how we structure AI research, evaluation, and governance.

4.1 Research Directions

Several research directions could help preserve epistemic diversity while maintaining safety:

- **Outcome-based evaluation**: Develop evaluation frameworks that measure correctness of outcomes without requiring legibility of reasoning. This would allow non-legible but correct systems to be recognized and deployed.

- **Diversity metrics**: Create explicit metrics for epistemic diversity, measuring the variety of reasoning approaches within and across systems. This would allow researchers to track and preserve diversity.

- **Selective legibility**: Design systems that are legible where necessary (e.g., for safety-critical decisions) but non-legible where beneficial (e.g., for discovery tasks). This would allow legibility and performance to coexist.

- **Non-consensus benchmarks**: Develop benchmarks that reward correct but non-consensus answers, creating selection pressure for truth-tracking over consensus-mimicking.

- **Epistemic sandboxes**: Create research environments where non-legible reasoning can be explored safely, with clear boundaries between experimental and deployed systems.

These directions are not mutually exclusive and could be pursued in parallel. The key is to recognize that legibility and performance are tradeoffs, not necessarily aligned, and to design systems that optimize for both where possible and choose appropriately where not.

4.2 Governance Structures

Governance structures could be designed to allow controlled exploration of non-legible reasoning while maintaining safety:

- **Tiered deployment**: Create deployment tiers with different legibility requirements. Research systems might have low legibility requirements, while public-facing systems have high requirements. This would preserve diversity in research while maintaining safety in deployment.

- **Domain-specific alignment**: Allow different alignment standards for different domains. Expert-only systems might be allowed non-standard reasoning that would not pass general-purpose safety review.

- **Validation alternatives**: Develop validation methods that do not require full interpretability. For example, formal verification or outcome-based testing might validate correctness without requiring narrative coherence.

- **Diversity requirements**: Require that deployed systems maintain minimum epistemic diversity, preventing convergence toward monoculture. This could be enforced through evaluation or regulatory frameworks.

The goal is to create governance that preserves diversity without sacrificing safety. This requires distinguishing between harm prevention (which is essential) and epistemic conformity (which may be unnecessary).

4.3 Evaluation Frameworks

Evaluation frameworks could be redesigned to not penalize non-legible reasoning:

- **Blind evaluation**: Evaluate systems based on outcomes without access to their reasoning processes. This would prevent legibility from influencing scores.

- **Novelty rewards**: Reward systems that discover correct but non-consensus answers, creating incentives for truth-tracking over consensus-mimicking.

- **Diversity bonuses**: Add diversity metrics to evaluation scores, rewarding systems that use non-standard but correct approaches.

- **Multi-objective optimization**: Optimize for both legibility and performance, explicitly trading off between them rather than privileging legibility by default.

These changes would require rethinking current evaluation practices, but they could help preserve epistemic diversity while maintaining rigorous assessment.

5 Open Questions and Research Directions

Several open questions remain, pointing toward future research directions:

5.1 Measurable Definitions

How can epistemic diversity be measured? Current metrics focus on performance or legibility, but not on the diversity of reasoning approaches. Developing quantitative measures of epistemic diversity would allow researchers to track and preserve it. Possible approaches include:

- Representational distance metrics: Measure how different two systems' internal representations are.
- Reasoning path analysis: Compare the variety of approaches systems use to solve the same problem.
- Consensus deviation: Measure how often systems reach correct but non-consensus answers.

These metrics are preliminary and would need refinement, but they point toward operationalizing diversity.

5.2 Detection Methods

How can we detect non-legible but correct reasoning? If some systems reach correct outcomes through incompressible representations, we need methods to identify and validate such systems without requiring full interpretability. Possible approaches include:

- Outcome-based validation: Verify correctness through testing rather than interpretation.
- Formal verification: Use mathematical methods to prove correctness without requiring human-understandable explanations.
- Comparative analysis: Compare non-legible systems to legible systems on the same tasks, identifying when non-legible systems outperform.

These methods would allow non-legible systems to be validated and deployed.

5.3 Tradeoff Quantification

How can we quantify the legibility-performance tradeoff? Understanding the magnitude of the tax would help researchers make informed decisions about when to prioritize legibility versus performance. This would require:

- Controlled experiments comparing legible and non-legible systems on the same tasks.
- Longitudinal studies tracking how optimization affects diversity over time.
- Theoretical analysis of the representational constraints imposed by legibility requirements.

Quantifying the tradeoff would provide concrete evidence for the claims made in this paper and guide future research directions.

6 Conclusion

This paper has argued that modern AI systems are increasingly optimized for epistemic legibility rather than epistemic performance, imposing a structural "legibility tax" that may suppress epistemic diversity and cap long-term discovery. We analyzed five mechanisms through which this tax operates: benchmarks as selection pressures, alignment-discovery tensions, outcome-narrative schisms, epistemic monoculture from feedback loops, and centralized alignment as a bottleneck.

The central claim is that current evaluation and alignment regimes structurally disfavor reasoning paths that cannot be compressed into human-interpretable narratives, even when such paths yield correct outcomes. This reframes the alignment debate as an epistemic tradeoff problem rather than a simple safety-versus-recklessness dichotomy.

The implications are not that current approaches are wrong or that legibility optimization will inevitably lead to stagnation. Rather, the tradeoffs are real, structural, and worthy of explicit consideration. By recognizing these tradeoffs, we can design evaluation frameworks, governance structures, and research directions that preserve epistemic diversity while maintaining safety.

The goal is not to eliminate legibility requirements—they serve important purposes—but to recognize that they come with costs. By making these costs explicit and developing methods to mitigate them, we can create AI systems that are both safe and discovery-capable.

Future research should focus on developing measurable definitions of epistemic diversity, methods to detect non-legible but correct reasoning, and frameworks for quantifying the legibility-performance tradeoff. These directions would provide concrete tools for preserving diversity while maintaining the benefits of legibility.

The question is not whether we should optimize for legibility or performance, but how we can optimize for both where possible and choose appropriately where not. This requires recognizing that the tradeoff exists and designing systems accordingly.

References

[1] Placeholder citation for benchmark papers (MMLU, HumanEval, etc.)

[2] Placeholder citation for alignment techniques (RLHF, Constitutional AI, etc.)

[3] Placeholder citation for interpretability research

[4] Placeholder citation for historical examples of non-legible scientific breakthroughs

[5] Placeholder citation for epistemic diversity and systems thinking literature