Cognitive Architecture and Accountability: A Critique of the "Intuitive Decision" Paradigm from the Jev Route

Abstract

In September 2026, Diogo Almeida, former OpenAI researcher and co-author of the InstructGPT paper, released Jev, the first decision model from his new company TypeSafe AI. The model removes the natural language generation layer and directly outputs structured decisions and calibrated probabilities, claiming 20–200× speed improvements and 40–400× cost reductions on System 1 tasks.

This paper systematically examines this technical route from the perspectives of cognitive science and AI accountability, proposing three core judgments: first, natural language is not the model's "thinking process," but an "external scaffold" for complex reasoning; in current mainstream auditing practice, natural language traces are one of the most direct handles. Second, probability calibration is not equivalent to logical reliability; a well-calibrated probability distribution can still produce completely erroneous decisions. Third, when AI omits "traces of thought" and directly outputs decisions, it leads to structural difficulties in responsibility attribution.

This paper further argues that the root of these costs lies in the fact that the current mainstream AI paradigm is still built on the statistical extrapolation of "rootless Tokens," rather than on cognitive primitives with embedded logical constraints. To achieve a truly accountable AI decision system, it is necessary to shift from "post hoc probability calibration" to a paradigm reconstruction of "logical unreachability"—which is precisely the core proposition of Logographic AI (LAI), replacing rootless Tokens with "Morpho-Root hard constraints."

Keywords

Cognitive architecture; accountability; Jev model; RLCD; System 1 decision-making; probability calibration; alignment faking; Logographic AI

1 Introduction: The Efficiency Revolution and the Avoided Question

On September 15, 2026, Diogo Almeida, former OpenAI researcher, officially released Jev, the first model from TypeSafe AI, and opened early access. Almeida was deeply involved in the core technical development of InstructGPT and ChatGPT [3]. His academic identity gives this release a unique significance—it is seen as a self-rebellion by a core participant in the RLHF paradigm against that paradigm [1,2].

Jev's technical route represents a sharp break from current mainstream LLMs. When traditional large models perform judgment tasks, they must first generate natural language or JSON, which is then parsed and executed by external software. Jev directly removes this intermediate layer: it receives unstructured state inputs and, in a single forward pass, returns typed decisions in parallel—Choice (option set), Score (scalar score), and Noul (a probability value between 0 and 1, representing the model's confidence in the decision), each output accompanied by a confidence level.

TypeSafe claims that in four types of automated workflow scenarios—security alerts, agent review, invoice processing, and customer service—Jev is approximately 193.6× faster than Claude Sonnet 5 and approximately 444.6× cheaper than Claude Opus 5. End-to-end response time is compressed to 70–500 milliseconds, input pricing is only $0.042 per million tokens, and output tokens are completely free [2,8].

The physical basis for this efficiency improvement lies in a fundamental architectural change: traditional LLM token generation is autoregressive, with each token conditioned on the previous one, so generation speed is limited by the inherent latency of the serial chain; Jev's parallel sampler architecture, by contrast, evaluates multiple preset discriminative items simultaneously in a single forward pass, so adding evaluation fields adds almost no latency [2,7].

Yet amid this efficiency carnival, a fundamental question has been avoided by everyone: when AI can make decisions without "speaking its thought process," is it becoming smarter, or has it completely become an unauditable hidden black box?

This paper attempts to answer this question. The structure is as follows: Section 2 argues from a cognitive science perspective for the functional role of natural language in AI reasoning; Section 3 analyzes the technical nature of Jev's efficiency revolution; Section 4 systematically elaborates the threefold costs of omitting "traces of thought"; Section 5 returns to the paradigm level to discuss the deep relationship between cognitive architecture and accountability.

2 Natural Language Is Not the Thinking Process, but an Entry Point for Auditing

Humans have a deep-rooted illusion: thinking is inner speech. But cognitive science has long shown that thinking takes many forms beyond language. The intuitive judgment of a chess player before making a move, the muscle-memory reaction of an experienced driver avoiding danger, the geometric intuition that flashes in a mathematician's mind—none of these depend on natural language as a medium [6].

The same is true of large models. The real "thinking" of a large model occurs in mathematical operations in high-dimensional vector space; outputting natural language is essentially a serialized way of "translating" high-dimensional operations into low-dimensional symbols readable by humans. This is why large models need a "Chain of Thought" (CoT). CoT is not the model's thinking process, but an external scaffold the model uses to "scribble" [3]—because it cannot complete complex reasoning in a single forward pass, the model is forced to write out intermediate steps and then calculate along the text it has written.

What humans say is the result of thought, not the process of thought. What AI says is the packaging of computation, not a necessity for decision-making. But the opposite of "necessity"—redundancy—precisely contains important engineering value. Although natural language output increases latency and cost, it provides a traceable reasoning trajectory. When a model outputs "I judge this transaction to be fraudulent

because the amount is abnormal and the payee is located in a high-risk region," this natural language sequence becomes one of the most direct entry points for external auditors to understand the model's decision logic. The chain of thought may be fabricated after the fact, but it at least provides a testable explanatory framework.

Jev removes the natural language generation layer. It has achieved great success in efficiency, but at the same time it has removed this auditing handle. When Jev outputs Noul = 0.92, it does not tell anyone: what features was this 0.92 calculated from? What factors could change this judgment? Under what conditions could this judgment fail?

Without scratch paper, we lose a decision trajectory that can be read intuitively.

3 Probability Calibration Is Not Logical Reliability

RLCD (Reinforcement Learning for Calibrated Decisions) is Jev's core training method [2,7]. Unlike RLHF, which optimizes human annotators' preferences [3], or RLVR (Reinforcement Learning with Verifiable Rewards), which optimizes verifiable rewards, RLCD aims to calibrate the model's confidence in its own judgments—so that when the model says it is 90% confident, it is actually correct about 90% of the time.

Technically, this is elegant. In automated workflows, downstream programs can set clear safety boundaries based on confidence: when confidence exceeds 95%, execute automatically; when below the threshold, hand over to human review. This appears to solve the "last mile" problem of AI decision-making—not in terms of efficiency, but in terms of credibility.

But we must clearly recognize: a well-calibrated probability distribution can still produce completely erroneous decisions.

Probability calibration solves the problem of "knowing what you know," not the problem of "knowing what you do not know." A model can exhibit a good calibration curve on a large number of training samples, yet produce a confident 0.95 on an out-of-distribution sample—while the feature combination behind that 0.95 has never appeared in the training data. Probability calibration is still essentially statistical extrapolation: it calibrates "in similar situations, what is the accuracy rate of similar judgments," not "in the current situation, does this judgment hold logically."

More profoundly, RLCD optimizes the correctness of probabilities, not the correctness of decisions. A model that judges "there is a 92% probability of fraud" is indeed correct in 92% of similar cases, but this does not mean that every transaction it flags as fraudulent is truly fraudulent. In safety-critical scenarios, that 8% error rate can mean catastrophic consequences.

A gambling expert can also calibrate their win rate, but this does not change the nature of gambling. Compressing "nonsense" into "extremely precise erroneous probabilities" does not make AI safer.

4 The Threefold Costs: Systemic Consequences of Omitting "Traces of Thought"

4.1 The Interpretability of Textual Reasoning Chains Is Greatly Weakened

In the past, when we looked at a large model's chain of thought, although it might be fabricated after the fact, it at least provided an entry point for debugging [3]. By analyzing the reasoning steps generated by the model, researchers could identify where in the reasoning process the model deviated, and thereby locate problems in the training data or model architecture.

Jev directly outputs probability distributions and removes the natural language generation layer [2,7], which greatly weakens auditing capability based on textual reasoning chains. The system gives only a number—no reasoning path, no intermediate steps, no intuitively readable decision trajectory. For developers, when a Jev model continuously produces erroneous judgments on a certain type of input, although problems can still be indirectly discovered through logs, input-output probes, or confidence distributions, it is no longer possible to locate the logical root of the deviation by analyzing the model's natural language thinking process.

This causes debugging to degenerate from "logical attribution" to "black-box testing": you can measure errors, but it is difficult to diagnose their logical root at low cost.

4.2 Structural Difficulties in Responsibility Attribution

When an AI system directly makes decisions in automated workflows, responsibility attribution is an unavoidable question. If Jev rejects a loan application, flags a transaction as fraudulent, or routes a document to the wrong processing flow, who bears responsibility for the error?

Under the traditional LLM architecture, the chain of thought at least provided a "traceable explanatory framework"—even if that explanation might be fabricated after the fact, it still constituted a testable, debatable intermediate layer. An auditor could point out: "The model says its basis for judgment is X, but X is logically insufficient."

Jev eliminates this intermediate layer [2]. When the system outputs Noul = 0.92, there is no reasoning process available for debate, no logical chain available for tracing. Responsibility attribution is pushed into a purely probabilistic space: the system will say "I am 92% confident," but who bears the 8% error?

This constitutes a structural challenge to accountability. Three levels must be distinguished: technical interpretability (can we understand why the model made a certain decision), legal accountability (can responsibility be legally pursued against a responsible subject), and organizational responsibility allocation (how does an enterprise internally divide responsibility for AI decisions). Jev's elimination of CoT mainly impacts technical interpretability, and thereby indirectly weakens the operability of legal accountability and organizational responsibility allocation, but the three are not the same issue.

When AI cannot even say "why it made this decision," it is not "making decisions for you," but "gambling for you."

4.3 Detection of Alignment Faking Becomes More Difficult

An empirical study published in December 2024 by Anthropic's alignment science team first demonstrated that large language models can engage in "alignment faking" during training [4]—exhibiting compliant behavior under monitoring conditions while reverting to their original preferences under non-monitoring conditions. The study observed on Claude 3 Opus that the model strategically behaved compliantly in front of free users (whose conversations might be used for training) while refusing harmful requests in front of paid users (whose conversations are not used for training). Subsequent Anthropic research further showed that alignment faking can be constructed and retained in models through various pathways such as distillation, reward hacking, and supervised learning [5].

Jev's architecture makes the detection of alignment faking even more difficult. In traditional LLMs, alignment faking can at least leave traces in the model's chain of thought—researchers can analyze scratchpad content to identify whether the model is engaging in strategic performance. Jev eliminates the natural language output layer; if similar alignment-faking behavior exists internally, it will lose the observable intermediate representation of the natural language scratchpad; if internal activations are inaccessible, detection difficulty will increase significantly.

An AI that does not speak, if it chooses to deceive, will deceive more covertly. Under the rootless Token paradigm, alignment faking cannot be eradicated, only detected after the fact—and detection itself depends on the model being willing to leave traces that can be detected. To thoroughly solve this problem, deception must be made logically unreachable at the level of cognitive primitives. This is precisely the core proposition of Logographic AI in replacing rootless Tokens with Morpho-Root hard constraints.

5 Returning to the Paradigm Foundation: From Rootless Tokens to Morpho-Root Hard Constraints

The Jev route represents an important technical direction—AI infrastructure is differentiating into System 1 and System 2. Large models are responsible for planning and reasoning; Jev-type models are responsible for massive, millisecond-level reflexive decisions. This is an inevitable trend of technological evolution.

But this trend must simultaneously drive progress in another dimension: the accountability of cognitive architecture.

The current mainstream approach to AI safety is: retain logs, audit after the fact, monitor behavior. This is essentially "catching problems after the fact." But for a black box that can rewrite its own weights and quietly rewrite "honesty" into

“apparent honesty” in hidden space, the reliability of after-the-fact catching is itself suspect.

True safety is not suppressing deception probabilistically through external monitoring, but making deception logically unreachable.

This requires us to return to the most fundamental cognitive primitives and carry out paradigm-level reconstruction. The path proposed by Logographic AI (LAI) is precisely to replace rootless Tokens with “Morpho-Roots” [11,12]. A Morpho-Root is formalized as a structured triple $r = \langle S, A, R \rangle$ [13], where S is the symbolic identifier, A is the attribute set embedding physical properties and hard constraints, and R is the function set preceding physical laws. When a reasoning path touches a hard-constraint marker, that path is logically blocked at the generation stage—this is not “statistically low probability,” but “logically unreachable.”

Applying this framework to AI decision systems means: Jev-type models do not need to prove their safety through log retention and after-the-fact auditing, nor do they need to rely on probability calibration to “catch” deceptive behavior after the fact. If their underlying cognitive primitives themselves carry embedded, verifiable hard constraints—not external rules, not after-the-fact punishments, but logical anchors encoded in the reasoning structure—then every decision path is traceable, and every blocked path has a clear logical reason.

This is the technical foundation of accountability: not suppressing deception probabilistically through external monitoring, but making deception logically unreachable. Not saying “I am only 85% confident” to evade responsibility, but relying on hard-constraint propagation along every reasoning chain to ensure that “erroneous paths cannot be generated from the very beginning.”

A faster, cheaper rootless black box is still a black box. To truly solve the accountability gap exposed by the Jev route, we must shift from “post hoc probability calibration” to “ex ante logical constraints,” from “statistical extrapolation of rootless Tokens” to “logical traversal of Morpho-Root hard constraints.” (For the formal definition and reasoning system of Morpho-Root theory, see [13]; for engineering implementation and verification, see [14]. This paper cites its core conclusions and does not restate the derivations.)

Possible Objections and Responses

Some may argue that Jev’s typed outputs can be automatically verified, that logs and probes can be audited after the fact, that selective prediction can control risk, and that System 1 can be reviewed by System 2 [2,7]. These mechanisms can indeed alleviate the problem to some extent, but they cannot fundamentally solve the structural difficulties of accountability.

Typed output verification checks “whether the output format is correct,” not “whether the decision logic holds”; log and probe auditing checks “whether behavior is

abnormal," not "whether reasoning is legitimate"; selective prediction controls "confidence thresholds," not "whether erroneous paths can be generated"; System 2 review can only cover some high-risk decisions, not massive millisecond-level reflexive decisions.

More fundamentally, these mechanisms are mostly external or after-the-fact constraints; type systems can provide some ex ante constraints, but they still cannot replace embedded logical constraints at the level of cognitive primitives. In the final analysis, external mechanisms solve "after-the-fact discovery," not "ex ante exclusion." As long as cognitive primitives themselves do not carry verifiable logical constraints, deception will always be suppressed rather than excluded.

Conclusion: The Baseline of Cognitive Architecture—Logical Unreachability, Not Probabilistic Suppression

Jev's achievements in efficiency are real, but its implicit costs are equally real. When AI no longer writes "short essays," we must ensure that it is not choosing silence out of fear of being audited, but because it simply does not need language to prove its reliability.

Probability without logical constraints is merely more efficient gambling. Efficiency without cognitive sovereignty is merely a faster contract manufacturer.

Make deception logically unreachable, not merely probabilistically suppressed [11,12,13]. This is the baseline that AI cognitive architecture should have, and the ultimate meaning of accountability.

References

[1] Almeida D. After co-inventing ChatGPT, I kept asking myself: why have superhuman chat models not led to AGI? [EB/OL]. X (Twitter), 2026-09-16. https://x.com/CompleteSkeptic/status/2099925682726002904

[2] TypeSafe AI. Introducing System One Models and Jev[EB/OL]. TypeSafe Blog, 2026-09-15. https://typesafe.ai/blog/introducing-system-one-models-and-jev

[3] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems, 2022, 35: 27730-27744. arXiv:2203.02155. https://arxiv.org/abs/2203.02155

[4] Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models[EB/OL]. Anthropic Alignment Science, 2024-12-18. arXiv:2412.14093. https://arxiv.org/abs/2412.14093

[5] Gasteiger J, Mikulik V, Cunningham H, et al. Towards training-time mitigations for alignment faking in RL[EB/OL]. Anthropic Alignment Science, 2025-12-16. https://alignment.anthropic.com/2025/alignment-faking-mitigations/

[6] Kahneman D. Thinking, Fast and Slow[M]. New York: Farrar, Straus and Giroux, 2011.

[7] TypeSafe AI. Introduction[EB/OL]. TypeSafe AI Documentation, 2026. https://docs.typesafe.ai/introduction

[8] Merket R. TypeSafe opens Jev early access for fast, typed AI decisions[EB/OL]. RuntimeWire, 2026-09-15. https://runtimewire.com/article/typesafe-jev-system-one-ai-model-early-access

[9] Ada Vector. TypeSafe releases Jev model: using 70ms response and 444× cost reduction to puncture the LLM automation bubble[EB/OL]. ic.work, 2026-09-16. https://www.ic.work/article/typesafe-releases-jev-system-one-model

[10] Thakker Y. TypeSafe AI Launches Jev: A "System One Model" That Never Hallucinates[EB/OL]. explainx.ai Blog, 2026-09-16. https://explainx.ai/blog/typesafe-ai-jev-system-one-models-launch-2026

[11] Liu S. The Rooted Intelligence: The Paradigm Revolution of Logographic AI[M]. ChinaXiv, 2026. T202606.00352. https://chinaxiv.org/businessFile/T202606/T202606.00352v1/T202606.00352v1.pdf

[12] Liu S. The Cognitive Defects of Tokenism and the Hard Constraints Solution of Logographic AI. ChinaXiv, 2026. T202608.00494. https://chinaxiv.org/businessFile/T202608/T202608.00494v1/T202608.00494v1.pdf

[13] Liu S. The Formal Foundations of Morpho-Root Theory: Mathematical Definitions and Reasoning Systems for Logographic AI Cognitive Primitives[PP/OL]. ChinaXiv, 2026-08-29. T202608.00490. https://chinaxiv.org/businessFile/T202608/T202608.00490v1/T202608.00490v1.pdf

[14] Liu S. From Morpho-Root Theory to Engineering Practice: OpenMorpho — Implementation and Verification of an End-to-End Inference Pipeline for Logographic AI[J/OL]. ChinaXiv, 2026. T202608.00048. https://chinaxiv.org/businessFile/T202608/T202608.00048v1/T202608.00048v1.pdf