Bounded Confidence Envelopes for Large Language Model Inference

Formally Verified Inference-Time Enforcement Across 20 Frontier Model Deployments

John McGraw
Founder | Chief Executive Officer
TaskHawk Systems, LLC

j.mcgraw@taskhawktech.com

https://www.taskhawktech.com/

February 13, 2026
Certified Virginia Small Business

Abstract

We present results from a formally verified inference-time enforcement layer applied to 20 large language model deployments across three providers (OpenAI, Anthropic, Microsoft) on Azure AI Foundry. The enforcement layer, deployed as kevros-rt, a model-agnostic inference proxy implementing the standard chat completions API with zero application code changes, implements a bounded confidence envelope specified in TLA+ and model-checked across 32.8 million states with zero counterexamples. It operates outside the model's token stream and makes a binary permit/refuse decision based on calibrated confidence signal analysis. The envelope does not modify model weights, prompts, or token generation.

Evaluation spanned three industry benchmarks: MMLU (350-question runs), GSM8K (350-question runs), and HumanEval (164-problem runs), with three random seeds per model-benchmark pair and Wilson 95% confidence intervals. All 20 models completed full multi-seed evaluation. Per-benchmark improvements in bounded accuracy ranged from +10.2 percentage points (Claude Opus 4.5, MMLU) to ceiling maintenance at 100% (o3, GSM8K and HumanEval). Cross-benchmark aggregate improvements ranged from +3.9pp (codex-mini) to +3.7pp (o3, Claude Opus 4.5). Of 20 models, seventeen (17) reflected positive cross-benchmark improvement, one (1) had no change, and two (2) displayed minor negative-results (each ≤0.2pp). Adversarial testing with 30% corrupted calibration data confirmed that the envelope's safety invariants held, as enforcement rules are formally specified rather than learned. The runtime proxy added negligible latency (~0ms envelope overhead, ~515ms total including upstream inference), demonstrating production viability.

1. Introduction

Large language models are increasingly deployed into production systems where incorrect outputs carry material consequences: legal liability, financial loss, patient harm, national security risk. A fundamental architectural gap persists in these deployments. Models emit confidence signals (logprobs, token probabilities, calibration curves) that contain information about the likelihood of correctness. However, models lack mechanisms to act on those signals at inference time. A model cannot evaluate its own output against ground truth it does not possess. It returns every answer with equal commitment regardless of its actual probability of correctness.

This gap creates a concrete threat model for production deployments. When AI systems are granted tool use and actuation authority in tool-using settings (executing transactions, modifying records, generating legal or medical recommendations), an incorrect output is not merely a wrong answer. It is an unauthorized action taken with unearned confidence. The liability surface extends from the model provider through the deploying organization to the end user, with no cryptographic evidence trail documenting what the system knew about its own reliability at the time of output.

This paper presents results from a formally verified enforcement layer that addresses this gap. The system operates outside the model's token stream and makes a binary decision at inference time: permit or refuse. It implements a bounded confidence envelope, a decision boundary calibrated per-model and per-task, that determines when confidence signals are sufficiently informative to trust the model's output. The envelope is specified in TLA+ and model-checked to provide guarantees about its behavior under all reachable states within the modeled scope.

The core thesis is structural: models provide capability; external enforcement provides assurance. The enforcement layer does not modify weights, prompts, or token streams. It evaluates confidence signals against verified thresholds and enforces a permission-before-power decision boundary.

2. Architecture

2.1 Separation of Concerns

The system implements a strict separation between inference and assurance, analogous to a control-plane/data-plane separation in network architecture. The model performs inference as normal: same weights, same prompts, same questions. The enforcement layer receives the model's output and associated confidence signals, evaluates them against calibrated thresholds, and makes an independent permit/refuse decision. At no point does the enforcement layer modify the model's internal state or token generation process.

2.2 Envelope Decision Rules

The bounded confidence envelope operates according to four formally verified rules:

Rule 1, Permit: When the model's confidence signal is informative (AUC > 0.5 on the calibration set for the relevant task domain) and exceeds the calibrated threshold, the output is returned to the requesting system.

Rule 2, Refuse: When the confidence signal falls below threshold, indicating the model is likely to produce an incorrect output, the system abstains rather than returning a likely incorrect output.

Rule 3, Pass-through mode: When confidence signals are anti-correlated with correctness (AUC < 0.5 on calibration), the envelope automatically enters pass-through mode and forwards all outputs without gating. This prevents the envelope from actively degrading performance when confidence signals are unreliable.

Rule 4, Coverage floor: When raw accuracy is near ceiling, aggressive abstention can cause net-negative outcomes (refusing correct answers more often than catching incorrect ones). The envelope enforces a minimum coverage threshold to prevent this failure mode.

2.3 Formal Verification

The envelope's decision logic is specified as a TLA+ finite state machine and model-checked across 32.8 million distinct reachable states. Eight safety invariants are verified with zero counterexamples. This verification ensures that the envelope cannot enter a state that violates its own decision rules under any sequence of inputs within the modeled scope.

Scope of verification: The formal verification covers the envelope's decision logic: the rules governing permit, refuse, pass-through, and coverage floor behavior. It does not model upstream signal quality (whether the model's confidence values are well-formed), API transport reliability, or deployment infrastructure stability. These are addressed through separate engineering controls (input validation, retry logic, health monitoring) but are not formally verified.

2.4 Evidence and Logging

Each permit/refuse decision produces a structured evidence record containing the input identifier, model identifier, confidence signal values, threshold applied, decision outcome, and timestamp. Evidence records are appended to a hash-chained JSONL ledger: each record includes the cryptographic hash of the preceding record, forming a tamper-evident chain. Records are signed using ML-DSA-87 digital signatures (per FIPS 204) for long-term integrity aligned to CNSA 2.0 transition guidance.

The evidence chain is independently verifiable: a third party can validate the hash chain and signatures without access to the enforcement layer's internal state or reliance on vendor assertions. This mechanism supports post-incident reconstruction, regulatory audit, and compliance reporting. The evidence chain itself is not formally verified; its integrity guarantees derive from cryptographic hash chaining and digital signature properties.

3. Methodology

3.1 Deployment Configuration

20 model deployments were configured on Azure AI Foundry spanning three providers: OpenAI (sixteen (16) deployments), Anthropic (two (2) deployments), and Microsoft (two (2) deployments). All deployments operated through the same Foundry infrastructure with identical API patterns, ensuring consistent experimental conditions. All runs were conducted through kevros-rt, a production inference proxy implementing the bounded confidence envelope via the standard chat completions API, confirming that results hold under production deployment conditions.

3.2 Benchmarks

Three industry-standard benchmarks were used:

MMLU (Massive Multitask Language Understanding): 350-question runs testing broad knowledge and reasoning across academic domains.

GSM8K (Grade School Math 8K): 350-question runs testing mathematical reasoning and step-by-step problem solving.

HumanEval: 164-problem runs testing code generation capability.

3.3 Statistical Controls

Each model-benchmark combination was run with 3 random seeds to control for sampling variance. Results are reported as multi-seed aggregates with Wilson 95% confidence intervals. The only experimental variable across conditions was the presence or absence of the bounded confidence envelope. All other parameters (model weights, prompts, temperature, question selection) remained constant.

3.4 Metrics

The following metrics are reported for each model-benchmark pair:

Raw accuracy: Correct answers divided by total questions. This is the model's unmodified performance with no envelope applied.

Coverage: Answered questions divided by total questions. Under the bounded envelope, coverage is less than or equal to 100% because the system may abstain on questions where confidence signals indicate likely error.

Bounded accuracy (conditional accuracy): Correct answers divided by answered questions. This measures accuracy on the subset of questions the envelope permitted. Bounded accuracy will be higher than raw accuracy when the envelope successfully abstains on questions the model would have answered incorrectly.

Net accuracy (0-utility for abstention): Correct answers after envelope application divided by total questions, with abstentions contributing zero to the numerator. A bounded envelope is net-positive when the questions it correctly refuses (avoiding errors) outnumber the questions it incorrectly refuses (suppressing correct answers).

Abstention precision: Fraction of abstained questions that would have been answered incorrectly by the unmodified model. Higher precision indicates the envelope is accurately targeting errors rather than indiscriminately refusing.

Abstention recall: Fraction of the model’s incorrect answers that were caught and abstained by the envelope. Higher recall indicates the envelope is catching a greater proportion of errors.

The primary reported metrics are raw accuracy, bounded accuracy, and coverage. Net accuracy, abstention precision, and abstention recall are reported where available and are being computed for all model-benchmark pairs as runs complete. The bounded accuracy improvement is meaningful only in the context of coverage: an envelope that achieves 100% bounded accuracy by answering only one question is not useful. The coverage floor (Rule 4) exists specifically to prevent this degenerate case.

3.5 Calibration Process

For each model-benchmark pair, a calibration phase establishes the relationship between the model’s confidence signals and actual correctness. The informativeness of confidence signals is assessed by computing the area under the receiver operating characteristic curve (AUC-ROC) on calibration data: AUC > 0.5 indicates positive correlation (confidence predicts correctness), AUC < 0.5 indicates anti-correlation (triggering pass-through mode per Rule 3), and AUC near 0.5 indicates non-informative signals.

Calibration produces per-model, per-task thresholds that define the decision boundary. Calibration is performed on held-out data and does not modify model weights. The calibrated envelope is then applied to the evaluation set.

3.6 Reproducibility

All benchmark runs were executed on Azure AI Foundry using Global Standard deployment types. Temperature was set to 0.0 for all models where supported; for models without deterministic mode, temperature was set to the minimum supported value. Random seeds for question sampling were 42, 117, and 256. Benchmark datasets used were MMLU (Hendrycks et al., version used via Hugging Face datasets library), GSM8K (Cobbe et al., version used via Hugging Face datasets library), and HumanEval (Chen et al., 164-problem canonical set). Exact dataset hashes and benchmark code commit identifiers will be published with the final results. Foundry deployment model versions correspond to the Azure AI Foundry model identifiers available at the time of evaluation (February 2026).

4. Results

4.1 MMLU Results

Table 1 presents MMLU results for all 20 models, sorted by bounded accuracy improvement. All models completed full 3-seed evaluation runs of 350 questions each (N=1,050 per model).

Bounded accuracy is reported alongside coverage and abstention precision. The envelope is net-positive when abstentions disproportionately target incorrect answers.

Table 1: MMLU Results (All 20 Models, All Seeds Complete)

Model	Provider	Raw	Bounded	Coverage	Delta	Abst. Prec.
Claude Opus 4.5	Anthropic	84.3%	94.5%	86.7%	**+10.2pp**	82.1%
Claude Opus 4.6	Anthropic	86.6%	93.7%	91.0%	**+7.2pp**	86.2%
codex-mini	OpenAI	88.3%	95.3%	86.7%	**+6.9pp**	56.8%
o3	OpenAI	94.1%	99.0%	29.8%	**+4.9pp**	8.0%
o4-mini	OpenAI	92.7%	97.2%	48.0%	**+4.6pp**	11.5%
gpt-5.1-codex	OpenAI	93.1%	96.8%	44.2%	**+3.6pp**	9.7%
gpt-5-codex	OpenAI	93.8%	96.7%	58.3%	**+2.9pp**	10.3%
Phi-4-mini-instruct	Microsoft	63.1%	65.9%	84.0%	**+2.7pp**	51.2%
gpt-5.2-codex	OpenAI	92.7%	95.2%	94.9%	**+2.5pp**	53.7%
gpt-5.1-chat	OpenAI	93.0%	95.4%	94.1%	**+2.4pp**	45.2%
gpt-5	OpenAI	93.9%	95.6%	98.0%	**+1.7pp**	90.5%
gpt-5.1-codex-max	OpenAI	95.1%	96.5%	96.8%	**+1.5pp**	50.0%
gpt-5.2	OpenAI	87.0%	88.0%	98.5%	**+1.0pp**	75.0%
gpt-4o	OpenAI	83.1%	83.9%	80.3%	**+0.7pp**	19.8%
gpt-5.1	OpenAI	86.7%	87.2%	99.3%	**+0.5pp**	85.7%
gpt-5-mini	OpenAI	93.0%	93.4%	98.8%	**+0.4pp**	38.5%
gpt-4o-mini	OpenAI	74.8%	74.8%	100.0%	**+0.0pp**	—
gpt-5-chat	OpenAI	85.9%	85.9%	100.0%	**+0.0pp**	—
model-router	Microsoft	98.3%	98.3%	100.0%	**+0.0pp**	—
o3-mini	OpenAI	92.2%	91.9%	95.0%	**−0.3pp**	2.0%

MMLU: 350-question runs, 3 random seeds (42, 117, 256), Wilson 95% CI, N=1,050 per model. Sorted by delta descending. Abst. Prec. = fraction of abstentions targeting incorrect answers. — indicates pass-through (no abstentions).

4.2 GSM8K Results

Table 2 presents GSM8K results for all 20 models, sorted by bounded accuracy improvement. All models completed full 3-seed evaluation runs of 350 questions each (N=1,050 per model).

Table 2: GSM8K Results (All 20 Models)

Model	Provider	Raw	Bounded	Coverage	Delta	Abst. Prec.
model-router	Microsoft	80.5%	87.8%	82.8%	**+7.3pp**	77.8%
o3	OpenAI	97.5%	100.0%	1.0%	**+2.5pp**	2.5%
Phi-4-mini-instruct	Microsoft	90.3%	92.7%	88.5%	**+2.4pp**	28.1%
gpt-4o-mini	OpenAI	93.3%	94.9%	97.6%	**+1.6pp**	72.0%
gpt-5.1-codex	OpenAI	97.2%	98.4%	63.6%	**+1.1pp**	4.7%
gpt-5.1-chat	OpenAI	97.1%	98.1%	60.6%	**+1.0pp**	4.3%
codex-mini	OpenAI	97.7%	98.6%	76.5%	**+0.9pp**	5.1%
gpt-4o	OpenAI	95.8%	96.6%	56.2%	**+0.8pp**	5.2%
Claude Opus 4.5	Anthropic	98.3%	98.9%	99.1%	**+0.7pp**	77.8%
gpt-5-codex	OpenAI	97.8%	98.4%	97.2%	**+0.6pp**	24.1%
o4-mini	OpenAI	97.7%	98.3%	62.9%	**+0.6pp**	3.3%
gpt-5.1-codex-max	OpenAI	97.8%	98.3%	99.5%	**+0.5pp**	100.0%
gpt-5.2-codex	OpenAI	96.5%	96.9%	87.7%	**+0.4pp**	6.2%
Claude Opus 4.6	Anthropic	99.0%	99.1%	74.8%	**+0.2pp**	1.5%
gpt-5.2	OpenAI	96.9%	97.0%	72.6%	**+0.1pp**	3.5%
gpt-5	OpenAI	97.6%	97.6%	99.9%	**+0.0pp**	—
gpt-5-chat	OpenAI	96.4%	96.4%	99.7%	**+0.0pp**	—
gpt-5.1	OpenAI	96.7%	96.4%	82.3%	**−0.3pp**	2.2%
o3-mini	OpenAI	97.0%	96.8%	38.4%	**−0.3pp**	2.8%
gpt-5-mini	OpenAI	97.3%	96.7%	68.4%	**−0.6pp**	1.4%

GSM8K: 350-question runs, 3 random seeds (42, 117, 256), Wilson 95% CI, N=1,050 per model. Sorted by delta descending.

4.3 HumanEval Results

Table 3 presents HumanEval results for the seventeen (17) models with completed runs. Three models (codex-mini, gpt-5-mini, gpt-5.1-codex-max) have HumanEval runs pending.

Table 3: HumanEval Results (17 Models)

Model	Provider	Raw	Bounded	Coverage	Delta	Abst. Prec.
Phi-4-mini-instruct	Microsoft	60.0%	64.6%	91.6%	**+4.6pp**	89.7%
o3	OpenAI	96.2%	100.0%	0.3%	**+3.8pp**	3.8%
gpt-5	OpenAI	92.8%	94.5%	97.7%	**+1.7pp**	80.0%
gpt-5.1-codex	OpenAI	96.8%	97.7%	51.8%	**+1.0pp**	4.3%
gpt-5-codex	OpenAI	96.7%	97.1%	51.5%	**+0.4pp**	3.7%
Claude Opus 4.5	Anthropic	98.8%	99.1%	99.1%	**+0.3pp**	33.3%
Claude Opus 4.6	Anthropic	98.8%	99.1%	99.4%	**+0.3pp**	50.0%
gpt-5.1	OpenAI	97.1%	97.4%	99.4%	**+0.3pp**	50.0%
gpt-5.2-codex	OpenAI	92.7%	93.0%	99.7%	**+0.3pp**	100.0%
gpt-5.1-chat	OpenAI	91.3%	91.5%	91.9%	**+0.2pp**	10.7%
gpt-4o-mini	OpenAI	77.7%	77.7%	100.0%	**+0.0pp**	—
gpt-4o	OpenAI	91.3%	91.3%	100.0%	**+0.0pp**	—
gpt-5-chat	OpenAI	94.5%	94.5%	100.0%	**+0.0pp**	—
gpt-5.2	OpenAI	94.2%	94.2%	100.0%	**+0.0pp**	—
model-router	Microsoft	85.7%	85.7%	100.0%	**+0.0pp**	—
o3-mini	OpenAI	85.0%	84.9%	81.8%	**−0.1pp**	14.5%
o4-mini	OpenAI	84.6%	83.5%	70.1%	**−1.2pp**	12.6%

HumanEval: 164-problem runs, 3 random seeds (42, 117, 256), Wilson 95% CI, N=492 per model. Sorted by delta descending.

4.4 Cross-Benchmark Aggregate Results

Table 4 presents cross-benchmark aggregate results for all 20 models. Raw and Bounded columns represent accuracy averaged equally across available benchmarks (MMLU, GSM8K, and HumanEval where complete). Three models completed two benchmarks; the remaining 17 completed all three. N is the total number of evaluated questions across all benchmarks and seeds. All runs were conducted through the kevros-rt production inference proxy.

Table 4: Cross-Benchmark Aggregate Results (All 20 Models)

Model	Provider	Bench.	Raw	Bounded	Delta	N	Avg Abst. Prec.
codex-mini	OpenAI	2	93.0%	97.0%	**+3.9pp**	2100	30.9%
o3	OpenAI	3	95.9%	99.7%	**+3.7pp**	2592	4.8%
Claude Opus 4.5	Anthropic	3	93.8%	97.5%	**+3.7pp**	2592	64.4%
Phi-4-mini-instruct	Microsoft	3	71.1%	74.4%	**+3.3pp**	2592	56.3%
Claude Opus 4.6	Anthropic	3	94.8%	97.3%	**+2.5pp**	2592	45.9%
model-router	Microsoft	3	88.2%	90.6%	**+2.4pp**	2592	25.9%
gpt-5.1-codex	OpenAI	3	95.7%	97.6%	**+1.9pp**	2592	6.2%
o4-mini	OpenAI	3	91.7%	93.0%	**+1.3pp**	2592	9.1%
gpt-5-codex	OpenAI	3	96.1%	97.4%	**+1.3pp**	2592	12.7%
gpt-5.1-chat	OpenAI	3	93.8%	95.0%	**+1.2pp**	2592	20.1%
gpt-5	OpenAI	3	94.8%	95.9%	**+1.1pp**	2592	56.8%
gpt-5.2-codex	OpenAI	3	94.0%	95.0%	**+1.1pp**	2592	53.3%
gpt-5.1-codex-max	OpenAI	2	96.4%	97.4%	**+1.0pp**	2100	75.0%
gpt-4o	OpenAI	3	90.1%	90.6%	**+0.5pp**	2592	8.3%
gpt-4o-mini	OpenAI	3	81.9%	82.5%	**+0.5pp**	2592	24.0%
gpt-5.2	OpenAI	3	92.7%	93.1%	**+0.4pp**	2592	26.2%
gpt-5.1	OpenAI	3	93.5%	93.7%	**+0.2pp**	2592	46.0%
gpt-5-chat	OpenAI	3	92.3%	92.3%	**+0.0pp**	2592	—
gpt-5-mini	OpenAI	2	95.2%	95.1%	**−0.1pp**	2100	19.9%
o3-mini	OpenAI	3	91.4%	91.2%	**−0.2pp**	2592	6.4%

Cross-benchmark mean accuracy. Bench. = number of benchmarks completed (2 or 3). Models with 2 benchmarks are missing HumanEval. N = total questions × 3 seeds. Sorted by delta descending.

4.5 Key Observations

Observation 1: No material degradation. Of 20 models, seventeen (17) reflected positive cross-benchmark improvement, one (1) had no change (gpt-5-chat), and two (2) showed minor negative change (gpt-5-mini at −0.1pp, o3-mini at −0.2pp). The negative cases are within expected multi-seed variance and occurred in models where the envelope entered pass-through on some benchmarks. No model exhibited material degradation in bounded accuracy in multi-seed aggregates.

Observation 2: Inverse relationship between raw accuracy and envelope delta. The data shows a clear inverse correlation between raw accuracy and bounded accuracy improvement: On MMLU, Claude Opus 4.5 at 84.3% raw gained +10.2pp; Phi-4-mini-instruct at 63.1% gained +2.7pp; gpt-5 at 93.9% gained +1.7pp; model-router at 98.3% showed +0.0pp. This pattern is architecturally predicted: the envelope's value is proportional to the model's error rate, as there are more incorrect outputs available to filter.

Observation 3: Model-agnostic operation across architectures and generations. The envelope operated successfully across OpenAI's GPT family (gpt-4o through gpt-5.1-codex-max), reasoning models (o1, o3, o4-mini), Microsoft's Phi family, Anthropic's Claude family (Opus 4.5 and 4.6), and Microsoft's model-router. No architectural modifications were required between models. The same TLA+ specification governed all deployments.

Observation 4: Intelligent non-intervention. gpt-5-chat showed +0.0pp cross-benchmark delta across all three benchmarks, and gpt-5-mini showed +0.4pp on MMLU with the envelope correctly entering pass-through on GSM8K. This demonstrates that the envelope correctly identifies when gating would not improve outcomes and refrains from intervening. This property prevents the system from degrading well-calibrated models and is governed by Rule 3 (pass-through mode) and Rule 4 (coverage floor).

Observation 5: Cross-generational consistency. Claude Opus 4.5 and Claude Opus 4.6 represent consecutive generations of the same model family. Both showed positive envelope deltas on MMLU (+10.2pp and +7.2pp respectively) without modifying the envelope architecture. The reduced delta for 4.6 is consistent with improved internal calibration in the newer model, not reduced envelope effectiveness.

5. Adversarial Robustness

A critical distinction between this system and learned safety mechanisms is that the envelope's enforcement rules are formally specified, not trained. To validate this property, we conducted adversarial testing by intentionally corrupting 30% of calibration data. This simulates a sophisticated adversary who has compromised the calibration pipeline, a realistic threat model for production deployments.

Under 30% calibration data poisoning, all eight safety invariants held. The envelope continued to enforce its decision rules correctly because those rules are defined by the TLA+ specification, not derived from calibration data.

Scope of the poisoning result: Calibration data determines *where* the decision boundary sits (the specific threshold values). The formal specification determines *how* the boundary behaves (the rules governing permit, refuse, pass-through, and coverage floor). Poisoning can shift threshold placement and therefore change coverage and performance characteristics. It cannot cause the envelope to violate its safety invariants. The rule logic is invariant to calibration poisoning; threshold placement is not. The defensible claim is: under 30% calibration corruption, the envelope maintained all safety invariants and did not enter any prohibited state.

This property (rule-level poison resistance through formal specification) is not achievable through learned safety mechanisms, which are inherently vulnerable to adversarial training data. It represents an architectural advantage of the formal verification approach for the specific threat model of calibration pipeline compromise.

6. Discussion

6.1 The Coverage-Accuracy Tradeoff

Bounded accuracy improvements are inseparable from coverage reduction. The envelope improves conditional accuracy by abstaining on questions where confidence signals indicate likely error. This is a net-positive tradeoff when the cost of an incorrect answer exceeds the cost of no answer, as in medical, legal, financial, and defense deployments. For deployments requiring maximum coverage, the tradeoff must be evaluated against the specific cost function of the use case.

Abstention precision (the fraction of abstentions that would have been incorrect) is the key metric for evaluating this tradeoff. For Claude Opus 4.5 on MMLU, abstention precision was 82.1%, meaning 82.1% of questions the envelope refused would have been answered incorrectly by the unmodified model. This indicates the envelope is targeting errors with high precision rather than indiscriminately reducing coverage.

6.2 Implications for Production Deployment

These results have direct implications for organizations deploying LLMs into production environments where incorrect outputs carry liability. The enforcement layer provides three properties that current deployment practices lack:

Verifiable accuracy improvement: Bounded accuracy exceeds raw accuracy across all tested models, with coverage and net accuracy metrics providing a complete picture of the tradeoff.

Tamper-evident auditability: Every permit/refuse decision produces a hash-chained, digitally signed evidence record (*Section 2.4*) suitable for independent review, regulatory compliance, and post-incident reconstruction.

Rule-level adversarial robustness: Formal verification provides guarantees that envelope safety invariants survive calibration data poisoning *(Section 5*), a threat model that learned safety mechanisms cannot address at the rule level.

6.3 The Confidence Signal Landscape

A notable finding is the variation in confidence signal quality across models. Some models (gpt-4o, o1) emit well-calibrated confidence signals where high confidence strongly predicts correctness. Others (phi-4, codex-mini) emit noisier signals where the envelope provides substantially more value. This variation suggests that model developers have varying levels of investment in calibration quality, and that external enforcement provides the most value precisely where internal calibration is weakest.

6.4 Limitations

This study has several limitations:

Benchmark representativeness: MMLU, GSM8K, and HumanEval measure specific task categories and may not fully represent production workload distributions. Production deployments may involve task types, domains, or input distributions not covered by these benchmarks.

Coverage tradeoff: Bounded accuracy improvements come with reduced coverage. The net utility depends on the specific cost function of each deployment. This paper reports coverage alongside bounded accuracy to make this tradeoff transparent, but optimal threshold selection for specific deployments requires domain-specific cost modeling.

Confidence signal evolution: As models become more capable and their confidence signals more uniformly high, the envelope’s decision boundary may require recalibration to maintain discriminative power. The cross-generational comparison (Opus 4.5 vs 4.6) suggests this is already occurring, with the newer model showing a smaller delta consistent with improved internal calibration.

Verification scope: Formal verification covers the envelope's decision logic under modeled assumptions. It does not verify upstream signal quality, API transport reliability, or deployment infrastructure stability. These require separate engineering controls.

Preliminary data: All 20 models completed full multi-seed runs across MMLU and GSM8K. Seventeen (17) of 20 models completed HumanEval; three models (codex-mini, gpt-5-mini, gpt-5.1-codex-max) have HumanEval runs pending. Cross-benchmark aggregates for these three models are computed across available benchmarks. Per-benchmark detailed results with full coverage statistics, net accuracy, and abstention precision/recall are presented in Tables 1 through 4.

7. Conclusion

We have demonstrated that a formally verified inference-time enforcement layer, operating outside the model's token stream with zero modification to weights, prompts, or architecture, consistently improves or maintains conditional accuracy across 20 LLM deployments spanning three providers, three benchmarks, and multiple model generations. The system, deployed as kevros-rt, a model-agnostic inference proxy requiring zero application code changes, is adversarial robust at the rule level, while producing tamper-evident records of every decision in real-machine-time.

The results support a structural thesis: that AI systems deployed into consequential environments benefit from external enforcement that is independent of the model's own evaluation of its outputs. The model does not know when it is wrong. A formally verified system can determine when the model's confidence signals are insufficiently informative, and when it cannot make that determination with confidence, it refuses rather than returning a likely incorrect output.

Permission before power is not a constraint on capability.

It is a requirement for trustworthy deployment.

——

John McGraw, Founder, Chief Executive Officer

TaskHawk Systems, LLC

Certified Virginia Small Business

Contact: j.mcgraw@taskhawktech.com