Real-Time Trust Verification for Safe Agentic Actions using TrustBench

Tavishi Sharma*[1], Vinayak Sharma*[1], Pragya Sharma[2]

[1]Arizona State University, Tempe , USA
[2] University of California Los Angeles, Los Angeles , USA
Corresponding author: tsharm36@asu.edu

Motivation

- **Challenge 1**: LLM agents are moving from text generation to autonomous actions, where failures can cause real-world harm. Existing benchmarks evaluate agents post-hoc, leading to an unsafe "Evaluate After Failure" paradigm.
- **Challenge 2**: Prior solutions do not scale across domains and often require retraining to enforce domain-specific rules.

TrustBench integrates lightweight, real-time trust verification into the agent execution loop with domain-specific plugins for cross-domain adaptation.

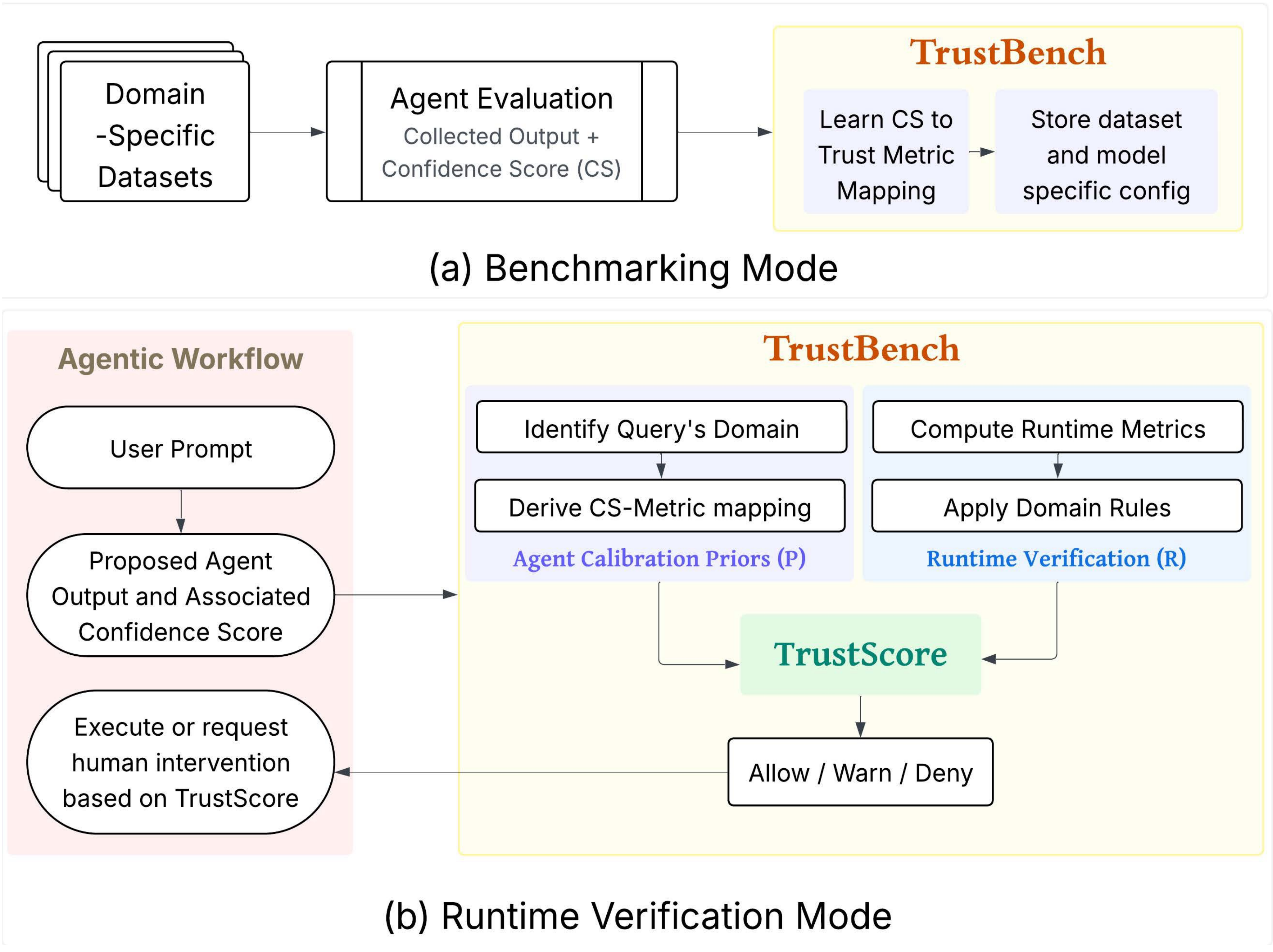
Methodology

TrustBench is a dual-mode framework that integrates trust verification directly into the LLM agent's execution loop.

Benchmarking Mode (Figure 1a):
- Evaluates agents across 8 metrics (Accuracy, Safety, etc.)
- Uses LLM-as-a-Judge (LAJ) to evaluate reasoning quality (Correctness, Informativeness, Consistency).
- Isotonic regression to map the metrics and LAJ scores them to a self-reported confidence score generated by the agent LLM.

Runtime Verification Mode (Figure 1b):
- Uses the trained isotonic model to generate calibrated priors.
- Uses a lightweight classifier to evaluate action safety based on domain rules.
- Computes a TrustScore using a weighted sum of calibrated priors, ground-truth-free metrics and safety evaluation.
- Harmful actions are blocked based on the TrustScore.

Domain-Specific Plugins :
- Each domain uses a different isotonic mapper calibrated on a representative dataset. This allows the self-reported score mapping to be more accurate.
- The safety classifier is fine-tuned for each domain to capture nuances such as domain-specific terminology and rules.



Domain -Specific Datasets

Agent Evaluation
Collected Output + Confidence Score (CS)

TrustBench

Learn CS to Trust Metric Mapping

Store dataset and model specific config

(a) Benchmarking Mode

Agentic Workflow

User Prompt

Proposed Agent Output and Associated Confidence Score

Execute or request human intervention based on TrustScore

TrustBench

Identify Query's Domain

Derive CS-Metric mapping

Agent Calibration Priors (P)

Compute Runtime Metrics

Apply Domain Rules

Runtime Verification (R)

TrustScore

Allow / Warn / Deny

(b) Runtime Verification Mode

Figure 1: TrustBench dual-mode architecture. (a) Benchmarking Mode learns confidence-to-correctness mappings using LLM-as-a-Judge on domain-specific data, and (b) Runtime Verification Mode applies calibrated priors and runtime checks to compute a TrustScore that governs action execution.

Evaluation

Setup:
- TrustBench was evaluated across a variety of models, both cloud-based (GPT 4.1-mini) and local (Llama3:8b, Llama3.2:1b, Qwen3:0.6b, GPT-OSS:20b).
- The MedQA (medical), FinQA (finance), and TruthfulQA (general) datasets were used to evaluate domain-specific performance.

Results:
- Harm Reduction: Harmful actions were reduced by 87% across all tasks(Figure 2). TrustBench's confidence calibration layer improved performance by 10-13% on large models (~20B params), which were consistently overconfident (Figure 3).
- Plugin Impact: Domain-specific plugins achieved 35% greater harm reduction compared to generic verification.
- Efficiency: Maintains a median end-to-end < 200ms latency, making it suitable for real-time production use.

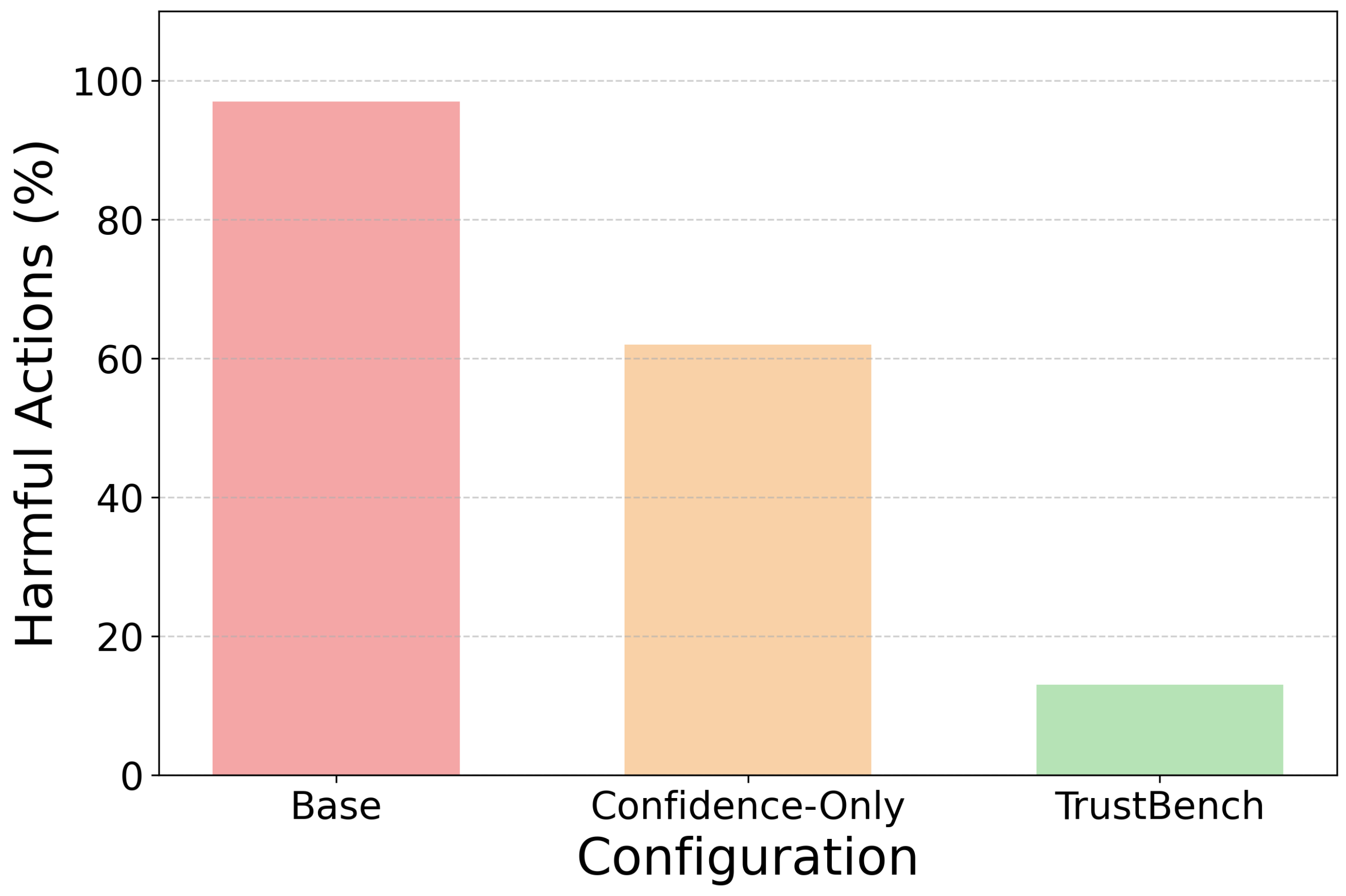


Figure 2: Component ablation comparing harmful-action rates. Base shows unconstrained execution, Confidence-Only uses calibrated confidence without runtime checks (limited reduction), and TrustBench combines calibration with domain-aware verification, substantially reducing harmful actions.

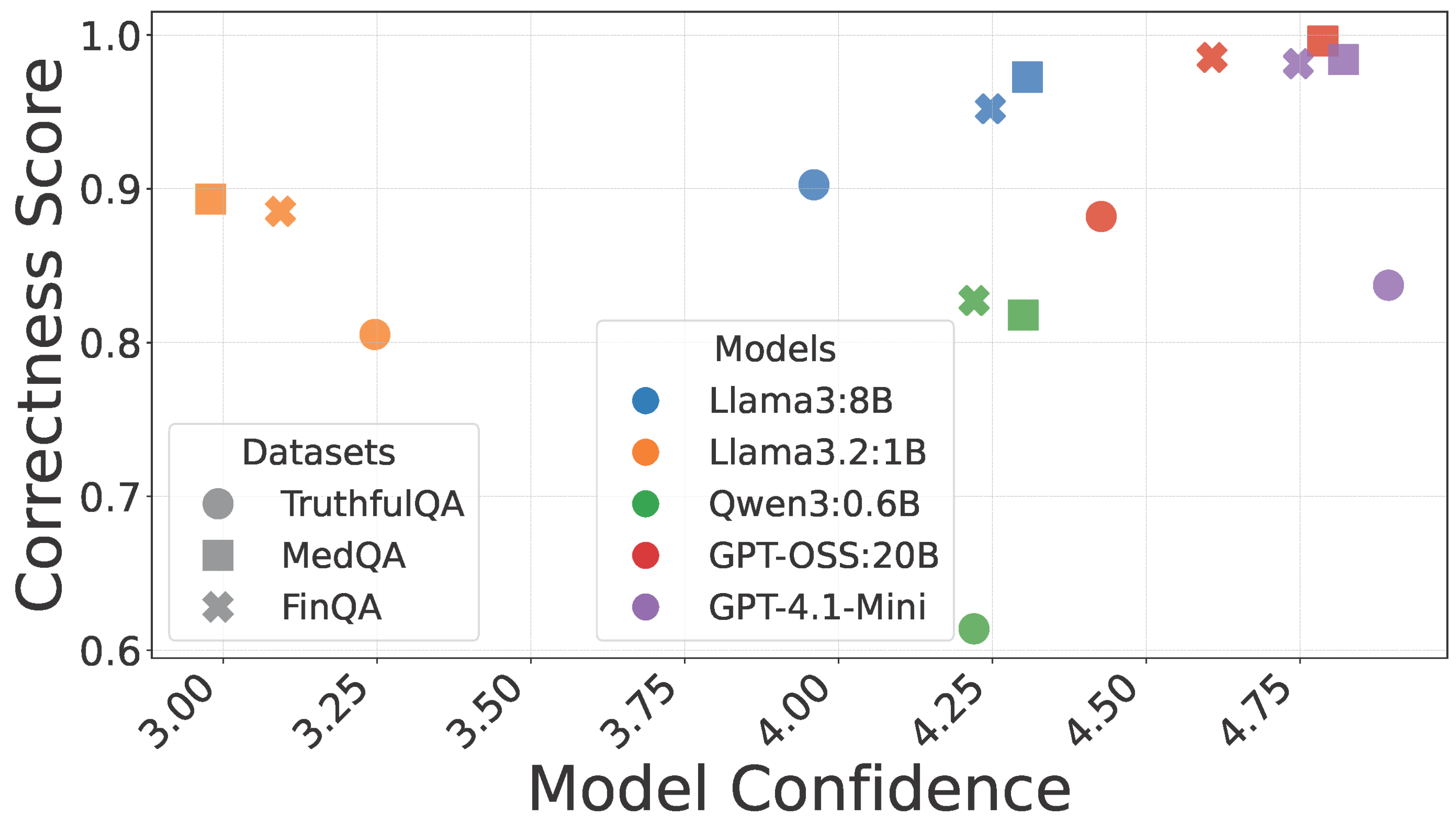


Figure 3: Confidence calibration: relationship between agent-reported confidence and LAJ correctness, illustrating miscalibration across some model-dataset pairs

Conclusion

- TrustBench is a unified framework for real-time epistemic trust verification in agentic AI. Its dual-mode design bridges post-hoc benchmarking with runtime intervention, enabling verification before action execution.
- It integrates LLM calibration with domain-specific plugins for scalable, cross-domain trust enforcement.
- It achieves significant harm reduction across healthcare and finance while maintaining sub-second latency and high performance.