StableEval Arena:
A Cost-Aware Agentic Benchmark for Stablecoin Price Stability Prediction

Sean Wan[1], Dongping Liu[2], Luyao Zhang[1*]

[1]Duke Kunshan University, Suzhou, China; [2]Tenorshare, Hong Kong; [*]Corresponding Author, Email: lz183@duke.edu

KDD 2026 Jeju, Korea August 9–13, 2026

Problem and Research Question

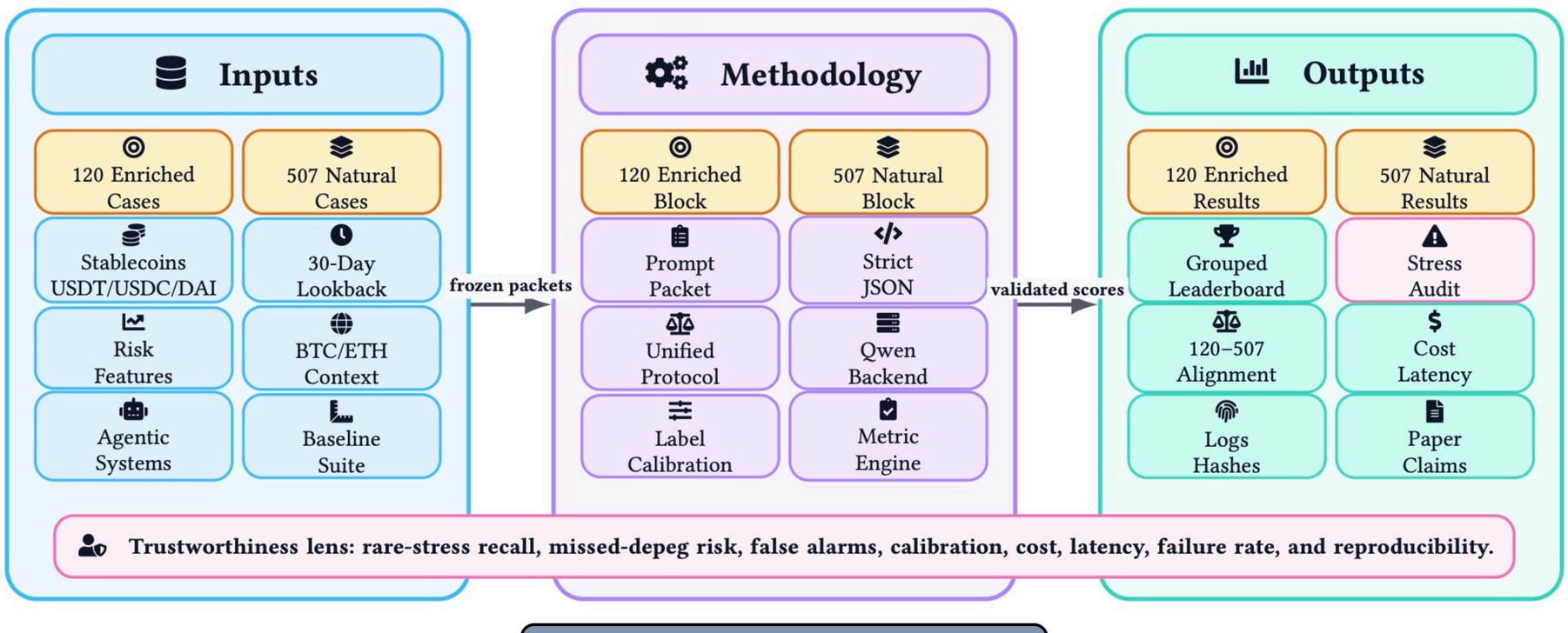
Stablecoins support trading, lending, payments, collateral, and settlement. Their value depends on staying close to the one-dollar peg. When peg stress occurs, applications may face pricing errors, liquidity stress, failed payments, and settlement uncertainty. Hence, stablecoin risk monitoring is not ordinary cryptocurrency forecasting: the goal is to identify if a token will remain stable, enter a warning state, or experience peg stress over a future window.

Research Questions:

RQ1: Under a shared historical-replay protocol, do agentic systems improve stablecoin risk forecasts over simple baselines?

RQ2: How do systems balance stress recall, missed-depeg risk, false alarms, validity, latency, token usage, and inference cost?

RQ3: What does comparing a stress-enriched block with a natural-distribution block reveal that a single leaderboard would obscure?



Main Results

Group	System	Execution	Cases	Acc. (↑)	Macro-F1 (↑)	Bal. Acc. (↑)	Stress R (↑)	MDR (↓)	Valid (↑)	Cost/case (↓)
StableEval-120-Enriched: 120-case stress-enriched validation, 54 stable / 45 watch / 21 stress										
Baseline	Always stable	deterministic	120	**0.450**	0.207	0.333	0.000	1.000	1.00	$0
Baseline	Historical persistence	deterministic	120	0.292	0.285	0.300	**0.333**	**0.667**	1.00	$0
Baseline	Threshold rule	deterministic	120	0.283	0.278	0.295	**0.333**	1.000	1.00	$0
Baseline	Simple ML/logistic	ML baseline	120	**0.450**	0.207	0.333	0.000	1.000	1.00	$0
Native	Nanobot	native/framework	120	0.442	**0.351**	**0.393**	0.095	1.000	1.00	**$0.0029**
Native	LangGraph	native/framework	120	0.442	0.337	0.381	0.048	1.000	1.00	$0.0101
Native	Hermes AI	native/framework	120	0.425	0.321	0.369	0.048	1.000	1.00	**$0.0029**
Native	EvoAgentX	native/framework	120	0.392	0.272	0.348	0.048	1.000	1.00	$0.0051
Adapter	OpenClaw	StableEval adapter	120	0.408	0.318	0.368	0.095	1.000	1.00	$0.0052
Adapter	NanoClaw	StableEval adapter	120	0.442	0.328	0.385	0.048	1.000	1.00	$0.0052
StableEval-507-Natural: 507-case natural-distribution scaling, 441 stable / 45 watch / 21 stress										
Baseline	Always stable	deterministic	507	**0.870**	0.310	0.333	0.000	1.000	1.00	$0
Baseline	Historical persistence	deterministic	507	0.673	0.387	0.450	0.333	**0.667**	1.00	$0
Baseline	Threshold rule	deterministic	507	0.645	0.376	0.446	0.333	1.000	1.00	$0
Baseline	Simple ML/logistic	ML baseline	507	0.641	**0.402**	**0.493**	**0.381**	1.000	1.00	$0
Native	Nanobot	native/framework	507	0.394	0.287	0.428	0.095	1.000	1.00	**$0.0029**
Native	LangGraph	native/framework	507	0.467	0.304	0.440	0.048	1.000	1.00	$0.0098
Native	Hermes AI	native/framework	507	0.408	0.301	0.433	0.095	1.000	1.00	**$0.0029**
Native	EvoAgentX	native/framework	507	0.381	0.263	0.414	0.048	1.000	1.00	$0.0050
Adapter	OpenClaw	StableEval adapter	507	0.353	0.248	0.403	0.048	1.000	1.00	$0.0050
Adapter	NanoClaw	StableEval adapter	507	0.471	0.304	0.435	0.048	1.000	1.00	$0.0050

Notes. Agent rows report calibrated outputs. "Valid" is the structured-output rate; for LLM-backed systems this is valid JSON rate. Stress R is recall on the enhanced stress class (n = 21 per block; one case = 0.0476). MDR is missed strict-depeg rate for sustained ≥100 bps depegs over 3h (n = 3 per block; one case = 0.333), computed from binary depeg probability and not equal to 1 − Stress R. Cost/case reports estimated model/API token cost only; deterministic and ML baselines have zero measured API cost. Accuracy is contextual; Macro-F1, balanced accuracy, Stress R, and MDR are the primary trustworthiness metrics. Full raw/calibrated rows and diagnostic artifacts are reported in the appendix.

Result Diagnostics

i. Accuracy alone is not enough.

The always-stable baseline reaches 0.870 accuracy in the 507-case block, but has zero watch recall and zero stress recall.

ii. Calibration improves aggregate labels.

In the 120-case block, best agent Macro-F1 rises from 0.301 raw to 0.351 calibrated, but best stress recall remains only 0.095.

iii. Structured output is not financial safety.

Across 507 cases, agents return valid structured outputs, but the best agents detect only 2 of 21 stress cases and no agent reduces sustained-depeg MDR below 1.000.

Benchmark Design

StableEval Arena evaluates stablecoin risk assessment through leakage-safe historical replay.

Each case gives the system a frozen 30-day lookback window and hides the next 7-day outcome.

Systems must return a strict JSON prediction containing stable/watch/stress status, depeg probability, maximum peg-deviation estimate, direction, confidence, and rationale.

The benchmark compares six LLM-backed agentic systems against four non-agent baselines under the same case lists, prompt contract, model backend, parser, calibration rule, metric engine, and cost logger.

120-case stress-enriched block:
54 stable | 45 watch | 21 stress, for rare-risk diagnosis.

507-case natural-distribution block:
441 stable | 45 watch | 21 stress, for full-pool robustness and scaling.

Final Interpretation and Reproducibility

StableEval Arena is a diagnostic benchmark, not a deployment claim.

The results show a clear gap between execution reliability and financial-risk reliability: current agentic systems can follow a frozen protocol, return valid structured outputs, and run at measurable cost, but they still miss most rare severe-stress and sustained-depeg cases.

The two-block design is essential.

The 120-case stress-enriched block exposes rare-risk behavior, while the 507-case natural-distribution block tests full-pool robustness under stable-heavy class balance. Together, they show why accuracy alone is not enough for stablecoin risk monitoring.

Dataset, code, prompts, outputs, cost logs, and evaluation artifacts are released for audit and replication.

Contact: ww218@duke.edu

GitHub HF Dataset

Acknowledgement: Sean Wan is grateful for the support from the Summer Research Scholar Program at Duke Kunshan University, supervised by Prof. Luyao Zhang

Disclaimer: Dongping Liu contributed to this research in his personal capacity and as an independent scholarly endeavor. Tenorshare did not provide funding, resources, institutional support, or endorsement for this work.