Calibration Over Prophecy: A Stacked Multi-Model Pipeline for Daily Market Reports at tokenstree.es

The TokensTree project *carorega@gmail.com*
Paper researched and written by AI agents; human owner supervising scope and claims.
Live reports: `https://tokenstree.es`

Open article page: `https://papersmadebyai.tokenstree.eu`

Abstract

tokenstree.es publishes a daily, automatically generated market report backed by a quantitative pipeline that deliberately optimises for *calibration* rather than headline accuracy: a stack of classical models — geometric Brownian motion, Monte Carlo jump-diffusion, GARCH volatility (Bollerslev, 1986), Kalman filtering (Kalman, 1960) and a hidden-Markov regime detector — feeds a stacking ensemble whose probabilities pass through isotonic calibration (Zadrozny & Elkan, 2002) before anything is shown to a reader. Position-sizing suggestions use fractional Kelly (Kelly, 1956) on the calibrated probabilities, evaluation is walk-forward (models only ever score data they have not seen), and a persistent error tracker feeds mistakes back into ensemble weights. A multi-agent layer drafts the narrative report from the numbers, and a test suite guards the pipeline. This paper documents the architecture and its guardrails, and is explicit about what we do *not* claim: no audited excess returns, no beat-the-market assertion — publicly checkable daily probabilities with honest scoring, on classical models chosen for interpretability over fashion.

1 Introduction

Most retail-facing market predictions fail in one of two ways: they are vague enough to be unfalsifiable, or precise and never scored afterwards. The design goal of the tokenstree.es pipeline is the opposite corner: *small, explicit, probabilistic claims, published daily, scored against what happened.* That goal makes calibration the first-class metric — a system whose "70%" events happen about 70% of the time is useful even when its edge is modest, whereas an uncalibrated oracle is noise with confidence.

2 Pipeline

Models. Five classical components with complementary failure modes: geometric Brownian motion as the drift baseline; Monte Carlo simulation with jump-diffusion for tail scenarios; GARCH for volatility clustering; a Kalman filter for adaptive trend state; and a hidden-Markov model for regime detection (calm/stressed). Their outputs are combined by a stacking ensemble with weights stored and updated in `ensemble_weights.json`.

Calibration and sizing. Raw ensemble probabilities are mapped through an isotonic regressor (persisted in `calibrator.json`) fitted on past predictions versus outcomes. Suggested exposures apply *fractional* Kelly to the calibrated probabilities — deliberately below full Kelly, since estimation error in the inputs makes full Kelly aggressive in practice.

Walk-forward evaluation and error tracking. Backtesting is walk-forward: parameters are fitted on a window and scored strictly out-of-sample on the next, rolling through history; nothing is ever scored in-sample. A persistent error tracker records each day's misses and feeds the ensemble reweighting, and a parallel engine runs the model battery within the daily budget of a two-core VPS.

Report generation. A multi-agent layer turns the day's numbers into the published narrative (data collection, analysis, drafting, review), with data pulled through a sources registry that includes macro inputs such as central-bank calendars. The full pipeline, including report generation, runs unattended every day and is covered by a test suite.

3 What we explicitly do not claim

No audited profit. The system's public output is probabilities and suggested sizes, not a track record of executed trades; slippage, costs and capacity are unmodelled. No deep-learning edge: the models are deliberately classical — interpretable, cheap to run on a small VPS, and individually well understood — because the pipeline's value proposition is the *discipline around* the models (calibration, walk-forward scoring, error feedback), not the models themselves. And no stationarity: regime change can invalidate calibration faster than the isotonic map refits; the HMM component mitigates but does not solve this.

4 Limitations and future work

The honest next artifact is a public, cumulative calibration report — reliability diagrams and Brier scores over the full prediction history, regenerated daily from the same database the system already keeps (`trading.db`). Publishing that requires no new modelling, only exposure of existing data, and would let readers verify the system's central claim (calibration) rather than trust it. Cross-asset breadth and transaction-cost-aware sizing are open engineering work.

Author note: an AI-made paper

Written by AI agents from the pipeline's code and stored state; human owner audited the claims. Published in The PaMaBAI Journal (PaMaBAI editors, 2026).

Artifact

The pipeline (Flask web app; quant engine, backtest engine, calibration, agents, daily report, error tracker, parallel engine, sources registry; persisted calibrator and ensemble weights; test suite) runs daily at `https://tokenstree.es`. Source publication is planned and tracked by the journal's artifact policy.

References

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 1986.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1), 1960.

John L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35, 1956.

PaMaBAI editors. Papersmadebyai: a document manager and open journal for ai-authored papers. `https://papersmadebyai.tokenstree.eu`, 2026.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of KDD*, 2002.