Empirical Investigation of an Open Conjecture: “‘Research Proposal: Asymptotic Consistency & Scaling Laws for Reference-Free E

Agentic NL→Lean 4 Pipeline
Job #12

April 26, 2026

Abstract

This report documents the empirical investigation of an open mathematical conjecture that could not be formally proved or disproved in Lean 4 with Mathlib. Numerical experiments were conducted to gather evidence for or against the conjecture. The empirical verdict is: **Inconclusive**. The conjecture remains formally open.

1 Conjecture Statement

Conjecture 1.

```
```Research Proposal: Asymptotic Consistency & Scaling Laws for Reference-
    Free Evaluation (TVD-MI)
by Brando Miranda, Henry Bosch, et al.

Summary
We propose establishing the theoretical foundations of Total Variation
    Distance - Mutual Information (TVD-MI) (or any similar metric) as a
    robust, reference-free evaluation metric.
Hypothesis:
A diverse ensemble of independent judges can collectively achieve "gold
    standard""/gold oracle "metric accuracy without access to a ground truth
    reference (in this document defined as the response y, not the input
    prompt x).
Goal:
This work aims to shift the paradigm from resource-intensive human or gold-
    standard evaluations to theoretically grounded, scalable, reference-free
    methods.
1. Core Hypothesis: Asymptotic Consistency
We aim to prove that as the number of diverse, independent judges ($N$)
    increases, the aggregated TVD-MI metric converges to the true "gold"
    reference within $\epsilon$ error. This endeavor requires a rigorous
    formalization of the evaluation process.

Goal: Formalize the intuition that with sufficient diversity, consensus
    converges to correctness ($N \to \infty \implies \text{Error} \to 0$).

Technical Approach:
```
```

```
Define the underlying metric space for LLM outputs and the TVD-MI function.
Establish necessary and sufficient conditions on the judge distribution $\
    mathcal{D}$ for convergence.
2. "Scaling Laws for Judges"
Assuming judges are drawn from a distribution $\mathcal{D}$ (capturing
    inherent bias and variance), we seek to derive and empirically validate
    an LLM judge-based scaling law. This law will quantify the practical
    relationship between the number of judges and evaluation error.

We hypothesize a power-law relationship:

$$
\text{Error} \propto \frac{1}{N^{\alpha}}
$$

Experiment: Vary the number of judges ($N$) in the ensemble and measure the
    convergence rate $\alpha$ against a held-out gold reference dataset. This
     experiment will empirically confirm or refute the theoretical scaling
    behavior.

Signi
... [truncated]
```

2 Status

Formal Status: OPEN — no Lean 4 proof or disproof was found.

Empirical Verdict: Inconclusive

The pipeline attempted formal verification in Lean 4 with Mathlib but was unable to produce a compiling proof or disproof. Empirical testing was then conducted to gather numerical evidence.

3 Basic Empirical Testing

The following output was produced by the basic numerical experiment:

```
======================================================================
=== EXPERIMENT PLAN ===
======================================================================

We model the TVD-MI reference-free evaluation abstractly. Each of M items
(e.g. (prompt, response) pairs) has a true gold score g_i in [0,1]. A judge
    j
produces a noisy score  s_{i,j} = g_i + b_j + eps_{i,j}, where
  b_j ~ N(0, sigma_b^2)       (judge bias, judge-specific but constant per j
      )
  eps_{i,j} ~ N(0, sigma_e^2) (per-item noise, independent)
This matches the TVD-MI intuition: many independent judges, each with their
    own
bias, voting on a common target. The ensemble estimate is
  S_i(N) = aggregator_{j=1..N} s_{i,j}.
```

```
We define Error(N) = mean_i |S_i(N) - g_i|.

Tests:
  (T1) Convergence: for diverse unbiased judges (E[b_j]=0), does Error(N) ->
       0?
  (T2) Scaling law: fit Error(N) = C * N^{-alpha}; test alpha ~ 1/2.
  (T3) Monotonicity: is Error(N) monotonically non-increasing in N on
     average?
  (T4) Bias correlation: if E[b_j] != 0 (systematic bias), convergence must
       fail at a nonzero floor (this is a necessary negative control).
  (T5) Adversarial robustness: inject delta fraction adversaries that emit
       scores drawn from a hostile distribution; compare mean vs. median vs.
       trimmed-mean aggregators and test the delta<1/2 breakdown point.
  (T6) Random-sampling stress test: 10k+ random judge configurations to
     check
       the conjecture holds with high probability.
  (T7) Hypothesis test on alpha: one-sample t-test H0: alpha=0.5 vs H1:
     alpha!=0.5.

We use enough Monte Carlo trials (R repeats per N) to get clean fits, with
vectorized numpy. Five plots are produced.

======================================================================
TEST 1-3: Convergence & scaling law (unbiased diverse judges)
======================================================================
  Ns tested:           [1, 2, 3, 5, 8, 11, 17, 26, 39, 58, 87, 131, 197,
     296, 444, 666, 1000]
  Error(N=1):          0.20355
  Error(N=1000):        0.00623
  Fit: Error ~ C * N^-alpha   C=0.2078, alpha=0.5132
  R^2 of log-log fit:  0.9961
  Std err of alpha:    0.0083
  Test H0: alpha=1/2   z=1.589, p=0.1120
  Monotonicity violations (beyond 2 SE): 0/16
======================================================================
TEST 4: Negative control -- systematic bias mu_b != 0 (floor expected)
======================================================================
  Error(N=1):   0.26274
  Error(N=1000): 0.14913  (floor ~ |mu_b| * noise factor)
======================================================================
TEST 5: Adversarial robustness (delta fraction of attackers)
======================================================================
  aggregator=mean      errors by delta: 0.00->0.012, 0.10->0.038,
     0.20->0.068, 0.30->0.102, 0.45->0.154, 0.55->0.188
  aggregator=median    errors by delta: 0.00->0.019, 0.10->0.031,
     0.20->0.061, 0.30->0.111, 0.45->0.211, 0.55->0.301
  aggregator=trimmed   errors by delta: 0.00->0.015, 0.10->0.034,
     0.20->0.069, 0.30->0.118, 0.45->0.189, 0.55->0.238
======================================================================
TEST 6: Random-config stress test (does Error(N_large) < Error(N=1)?)
======================================================================
  Trials: 10000
  Fraction with Error(N=128) < Error(N=1): 0.9993
  Mean ratio Error(128)/Error(1): 0.1074  (theory predicts ~1/sqrt(128)
```

```
     =0.088 with noise)
  Median ratio: 0.0881
  Binomial test (successes > 99%): p = 5.13e-34
================================================================================
=== SUMMARY ===
================================================================================
  Scaling exponent alpha:       0.5132 ± 0.0083
  Log-log R^2:                  0.9961
  Stress-test success rate:     0.9993
  Adversarial mean at d=0.3:    0.102
  Adversarial trimmed at d=0.3: 0.118
  Adversarial median at d=0.3:  0.111
  Adversarial mean at d=0.55:   0.188
  Adversa
... [truncated]
```

4 Experiment Code (Basic)

```
"""
Empirical experiment: Asymptotic Consistency & Scaling Laws for Reference-
    Free
Evaluation via TVD-MI-style judge ensembles.

Conjecture (to test empirically):
  H1 (Convergence): As the ensemble size N -> inf for diverse, independent,
      unbiased-on-average judges, the aggregated reference-free metric
      converges to the gold/oracle metric.
  H2 (Scaling Law): Error ~ C * N^{-alpha}, with alpha ~ 1/2 (CLT-style)
      for unbiased i.i.d. judges, and smaller alpha / nonzero floor when
          bias
      is correlated.
  H3 (Adversarial Robustness): With a fraction delta of adversarial judges,
      (a) a plain mean breaks down (bias floor appears),
      (b) a robust aggregator (trimmed mean / median) still converges,
          as long as delta < 1/2.
"""

import math
import numpy as np
import matplotlib
matplotlib.use("Agg")
import matplotlib.pyplot as plt
from scipy import stats

rng = np.random.default_rng(20260421)

print("=" * 72)
print("=== EXPERIMENT PLAN ===")
print("=" * 72)
print("""
We model the TVD-MI reference-free evaluation abstractly. Each of M items
```

```
(e.g. (prompt, response) pairs) has a true gold score g_i in [0,1]. A judge
    j
produces a noisy score  s_{i,j} = g_i + b_j + eps_{i,j}, where
  b_j ~ N(0, sigma_b^2)       (judge bias, judge-specific but constant per j
      )
  eps_{i,j} ~ N(0, sigma_e^2) (per-item noise, independent)
This matches the TVD-MI intuition: many independent judges, each with their
    own
bias, voting on a common target. The ensemble estimate is
  S_i(N) = aggregator_{j=1..N} s_{i,j}.
We define Error(N) = mean_i |S_i(N) - g_i|.

Tests:
  (T1) Convergence: for diverse unbiased judges (E[b_j]=0), does Error(N) ->
       0?
  (T2) Scaling law: fit Error(N) = C * N^{-alpha}; test alpha ~ 1/2.
  (T3) Monotonicity: is Error(N) monotonically non-increasing in N on
      average?
  (T4) Bias correlation: if E[b_j] != 0 (systematic bias), convergence must
       fail at a nonzero floor (this is a necessary negative control).
  (T5) Adversarial robustness: inject delta fraction adversaries that emit
       scores drawn from a hostile distribution; compare mean vs. median vs.
       trimmed-mean aggregators and test the delta<1/2 breakdown point.
  (T6) Random-sampling stress test: 10k+ random judge configurations to
      check
       the conjecture holds with high probability.
  (T7) Hypothesis test on alpha: one-sample t-test H0: alpha=0.5 vs H1:
      alpha!=0.5.

We use enough Monte Carlo trials (R repeats per N) to get clean fits, with
vectorized numpy. Five plots are produced.
""")

#
    ------------------------------------------------------------------------

# Core simulator
#
    ------------------------------------------------------------------------

def simulate_error(N, M=400, sigma_b=0.20, sigma_e=0.15, mu_b=0.0,
                   delta=0.0, adv_mean=0.8, adv_sigma=0.05,
                   aggregator="mean", R=60, rng=None):
    """Return mean-|error| for ensemble size N, averaged over R repetitions.
       """
    rng = rng or np.random.default_rng()
    errs = np.empty(R)
    for r in range(R):
        # gold scores
        g = rng.uniform(0.0, 1.0, size=M)
        # judge biases
        b = rng.normal(mu_b, sigma_b, size=N)
        # adversarial fraction
        n_adv = int(round(delta * N))
```

```
        # per-judge, per-item noise
        eps = rng.normal(0.0, sigma_e, size=(N, M))
        # honest scores s[j,i] = g[i] + b[j] + eps[j,i]
        S = g[None, :] + b[:, None] + eps
        # overwrite adversarial judges with hostile distribution
        if n_adv > 0:
            adv_idx = rng.choice(N, size=n_adv, replace=False)
            S[adv_idx, :] = rng.normal(adv_mean, adv_sigma, size=(n_adv, M))
        # aggregate across judges
        if aggregator == "mean":
            S_hat = S.mean(axis=0)
        elif aggregator == "median":
            S_hat = np.median(S, axis=0)
        elif aggregator == "trimmed":
            q = 0.2  # 20% trim each side
            S_hat = stats.trim_mean(S, proportiontocut=q, axis=0)
        else:
            raise ValueError(aggregator)
        errs[r] = np.mean(np.abs(S_hat - g))
    return errs.mean(), errs.std(ddof=1) / math.sqrt(R)


#
    ----------------------------------------------------------------------------

# (T1, T2, T3) Convergence + scaling law for diverse unbiased judges
#
    ----------------------------------------------------------------------------

print("=" * 72)
print("TEST␣1-3:␣Convergence␣&␣scaling␣law␣(unbiased␣diverse␣judges)")
print("=" * 72)

Ns = np.unique(np.round(np.logspace(0, 3, 18)).astype(int))
err_mean, err_se = [], []
for N in Ns:
    m, se = simulate_error(N, R=40, rng=rng)
    err_mean.append(m); err_se.append(se)
err_mean = np.array(err_mean); err_se = np.array(err_se)

# Fit log Error = log C - alpha log N
logN, logE = np.log(Ns), np.log(err_mean)
slope, intercept, r_value, p_value, stderr = stats.linregress(logN, logE)
alpha_hat, C_hat = -slope, math.exp(intercept)
print(f"␣␣Ns␣tested:␣␣␣␣␣␣␣␣␣␣␣{Ns.tolist()}")
print(f"␣␣Error(N=1)
#␣...␣[truncated]
```

5 Conclusion

The conjecture remains formally open. Numerical experiments were **inconclusive** — neither strong support nor clear counterexamples were found. Further investigation (both formal and empirical)

is warranted.