Bayesian Truth Serum for LLMs?

Shuo Li Liu [1]

Abstract

Bayesian Truth Serum (BTS) asks each voter to report an answer and forecast how other voters will answer. For strategic human voters, the scoring rule can reward truthful reports at equilibrium. We test whether this signal still helps when the voters are language models. The original assumptions–independent errors, strategic incentives, and coherent posteriors–do not hold for LLM panels.

On 871 code-review items judged by four models from the GPT-5 family, every BTS-family rule is less accurate than plain majority vote. BTS-E reaches 0.9724 accuracy versus 0.9793 for majority vote, with paired McNemar $p = 0.031$. BTS-clamp, BTS-info, and the Surprisingly Popular (SP) rule also lose significantly, and these losses survive Benjamini–Hochberg correction at $q = 0.10$. A 150-item probe over MMLU, TruthfulQA, and ARC Challenge shows the same pooled ordering.

Calibration moves in the opposite direction. BTS-family rules reduce binned expected calibration error by 21–32 percent relative to majority vote, and SP gives the lowest split-conformal mean coverage error in a small diagnostic. The failure mode is structural: shared pretraining breaks the conditional-independence assumption that BTS requires, while elicited confidences are decoder samples rather than coherent posteriors. BTS is best understood as a peer-calibration statistic for held-out calibration, abstention, and conformal wrapping, not as a replacement for majority vote.

[1]Princeton University. Correspondence to: Shuo Li Liu <shuoliliu@princeton.edu>.

1. The Question for LLM Panels

Panels of language models now act as predictors, reviewers, and decision makers inside larger systems (Du et al., 2024; Wu et al., 2024; McAleese et al., 2024). Each panel returns a verdict and an uncertainty signal. That signal is not automatically a statistical guarantee: it is not coverage, type-I error, or calibration in the sense used for predictive models. Once language models enter the loop as noisy stochastic predictors, we need to ask which statistical tools still give valid inference.

BTS is tempting because it already elicits the meta-predictions that a calibration wrapper can use as a nonconformity score. In the classical mechanism, respondents report both their own answer and a forecast of the population answer distribution (Prelec, 2004). Answers are rewarded when they are more common than predicted, and forecasts are rewarded for matching the empirical population distribution. Related peer-prediction methods adapt this idea to small panels, correlated reports, or multiple tasks (Witkowski and Parkes, 2012; Dasgupta and Ghosh, 2013; Shnayder et al., 2016; Kong and Schoenebeck, 2022). The distinct Surprisingly Popular rule scores answers rather than reviewers (Prelec et al., 2017). A minority answer that is unexpectedly common may be the informed answer.

That appeal should not be confused with a guarantee for LLM panels. The classical theorem is a mechanism-design statement about strategic human respondents with a payoff. An inference-time language model is not paid by the score, does not choose a report to maximize expected utility, and does not provide independent private information in the classical sense. The model emits text conditioned on a prompt. It also shares pretraining corpora, instruction-tuning patterns, benchmark exposure, and common failure modes with the other models in the panel. These facts put the LLM case outside the conditions that make BTS truth-inducing.

The experiment splits cleanly. On top-label accuracy, majority vote wins. On binned calibration, BTS-family rules improve the reliability curve. For LLM

panels this distinction matters. A score with slightly lower top-label accuracy can still help if it identifies cases that deserve another reviewer, human review, or abstention. Conversely, a better calibrated confidence number does not license the system to override the majority answer. Accuracy, calibration, and coverage are separate objects.

The panel rule has to say what it is optimizing: accuracy, Brier risk, binned reliability, conformal coverage, or routing utility. These targets are not interchangeable. BTS supplies a peer score. Distribution-free wrappers can certify decisions made from that score under their own assumptions. Neither step removes the need to state the target.

2. Panel Setup and Aggregation Rules

Let (X_i, Y_i^*) be an input item and ground-truth label with $Y_i^* \in \mathcal{V}$, where $\mathcal{V}$ is finite. A panel $\mathcal{R} = \{r_1, \ldots, r_K\}$ of language-model reviewers returns, for each item i and reviewer k, a triple

$$(v_{i,k}, c_{i,k}, \widehat{p}_{i,k}) \in \mathcal{V} \times [0,1] \times \Delta(\mathcal{V}).$$

Here $v_{i,k}$ is the verdict, $c_{i,k}$ is the stated confidence, and $\widehat{p}_{i,k}$ is the reviewer forecast of the population verdict distribution. An aggregation rule A maps the K triples to a panel verdict $\widehat{y}_i$ and confidence $\widetilde{c}_i$.

The empirical population share and the geometric mean forecast are

$$\bar{x}_i(v) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}[v_{i,k} = v],$$

$$\bar{y}_i(v) = \exp\left\{ \frac{1}{K} \sum_{k=1}^{K} \log \widehat{p}_{i,k}(v) \right\}.$$

$$\bar{p}_i(v) = \frac{1}{K} \sum_{k=1}^{K} \widehat{p}_{i,k}(v).$$

The first two quantities are the realized population share and the geometric mean forecast; $\bar{p}_i$ is the arithmetic mean forecast used by SP. Small Laplace smoothing is applied before logarithms. The Prelec information score and prediction score for reviewer k on item i are

$$IS_{i,k} = \log \frac{\bar{x}_i(v_{i,k})}{\bar{y}_i(v_{i,k})},$$

$$PS_{i,k} = \sum_{v \in \mathcal{V}} \bar{x}_i(v) \log \frac{\widehat{p}_{i,k}(v)}{\bar{x}_i(v)}.$$

All frequencies and forecasts inside logarithms use the same ϵ-Laplace smoothing, with the usual $0 \log 0 = 0$ convention for unsmoothed zero-frequency terms. The full BTS score is $T_{i,k} = IS_{i,k} + \alpha PS_{i,k}$ with $\alpha = 1$. The information term rewards unexpectedly common answers; the prediction term rewards accurate forecasts of the realized peer distribution. Let $T_i = (T_{i,1}, \ldots, T_{i,K})$.

Table 1. Aggregation rules evaluated on the same elicited panel.

Rule	Decision weight or rule
MV	$\sum_k \mathbf{1}[v_{i,k} = v]$
CW	$\sum_k c_{i,k} \mathbf{1}[v_{i,k} = v]$
BTS-E	$\sum_k \exp(T_{i,k}) \mathbf{1}[v_{i,k} = v]$
BTS-soft	$\sum_k \mathrm{softmax}_k(T_i) \mathbf{1}[v_{i,k} = v]$
BTS-clamp	$\sum_k \max(T_{i,k}, 10^{-3}) \mathbf{1}[v_{i,k} = v]$
BTS-info	$\sum_k \exp(IS_{i,k}) \mathbf{1}[v_{i,k} = v]$
SP	$\arg\max_v \{\bar{x}_i(v) - \bar{p}_i(v)\}$

Table 1 lists the seven rules. MV and CW are voting baselines. Among the BTS variants, BTS-E treats T as a log score, BTS-soft normalizes reviewer scores, BTS-clamp is a positive-clamp ablation, and BTS-info removes the prediction term. SP is different. It scores answers, not reviewers: it chooses the answer whose observed share most exceeds its predicted share (Prelec et al., 2017). Keep SP separate because it can calibrate well even when BTS-weighted voting loses accuracy.

3. The Original Theory and the Transfer Gap

3.1. What BTS Proves for Human Respondents

Classical BTS is a truth-inducing scoring rule for subjective data (Prelec, 2004). Respondents possess private signals, share a population model, report both their answer and the distribution they expect others to report, and receive a score tied to those reports. Under the stated conditions, truthful reporting is a Bayes–Nash equilibrium. Robust BTS variants relax parts of the common-prior structure for small populations (Witkowski and Parkes, 2012). Later peer-prediction mechanisms refine the elicitation problem by weakening priors, using multiple tasks, or constructing strictly proper peer games (Dasgupta and Ghosh, 2013; Shnayder et al., 2016; Kong and Schoenebeck, 2022).

BTS is sound for the theorem it proves: an incentive property for strategic respondents. An LLM panel at inference time is a different object. The score $T_{i,k}$ is not paid out. The model does not choose a report by optimizing $T_{i,k}$. The population forecast $\widehat{p}_{i,k}$ is a text-elicited probability vector rather than a stable posterior belief. A transfer theorem would need additional conditions connecting peer calibration to correctness.

3.2. What the Scores Estimate

The peer scores do not directly estimate $\Pr[v_{i,k} = Y_i^*]$. They are functions of $(v_{i,k}, \bar{x}_i, \widehat{p}_{i,k})$. The full score $T_{i,k}$ targets peer calibration, $IS_{i,k}$ targets unexpected commonness, and SP targets the answer with largest $\bar{x}_i(v) - \bar{p}_i(v)$. The truth Y_i^* is not an argument of these functions. Correctness enters only through assumptions that align peer agreement with truth. Three assumptions make that bridge plausible.

Assumption 1 (Exchangeable peer draw). Conditional on the item, reviewers are sampled from a stable population and are exchangeable for the purpose of the panel statistic.

Assumption 2 (Conditional error independence). Given Y^* and relevant item difficulty, the verdict errors of different reviewers are not driven by a shared latent failure mode.

Assumption 3 (Truth anchor). The expected population share of the true answer exceeds the expected share of each false answer, and forecasts are calibrated enough for unexpected commonness to indicate information rather than shared bias.

These assumptions are identifying conditions, not consequences of BTS. If they hold, a peer-calibration statistic can be informative about truth. If they fail, an unexpectedly common answer can be wrong, and a well-predicted peer distribution can simply reflect a correlated mistake.

Proposition 1 (Peer calibration does not identify correctness under shared error). There exist $K = 4$ panel distributions satisfying the BTS reporting format such that majority vote, BTS information scoring, full BTS weighting, and the Surprisingly Popular rule select the same false answer, while the peer scores contain no signal favoring the lone correct reviewer.

Proof sketch. Let $\mathcal{V} = \{\text{buggy}, \text{clean}\}$ and let the truth be buggy. Suppose a shared training artifact induces three reviewers to answer clean and one reviewer to answer buggy. Suppose the same artifact makes the reviewers forecast population shares of 0.75 clean and 0.25 buggy. Then $\bar{x}(\text{clean}) = \bar{y}(\text{clean}) = 0.75$ and $\bar{x}(\text{buggy}) = \bar{y}(\text{buggy}) = 0.25$. The information scores are zero for both the incorrect majority and the correct minority. The prediction scores do not distinguish truth from shared forecast accuracy. SP sees no answer as surprisingly popular. Reweighting collapses to the wrong majority. The failure is conditional dependence through the shared artifact, not a numerical accident. □

The counterexample is deliberately small because the pathology is structural. LLM judges often share pretraining corpora, instruction-tuning objectives, evaluation contamination, and prompt priors. Judge correlation is already a known concern in LLM evaluation (Lu et al., 2024). BTS can reward or fail to penalize the shared pattern if the forecast and the verdict inherit the same bias.

3.3. Why LLM Panels Break the Incentive Story

Three mismatches do most of the damage. Shared pretraining weakens the conditional independence that makes peer signals informative. Several models may learn the same wrong pattern, and then accurately forecast one another's wrong answers. Non-strategic predictors do not treat the score as a payoff. A frozen forward pass has no incentive to choose the report that maximizes expected reward. Finally, elicited probabilities are prompted text, not stable posterior beliefs. Confidence elicitation in language models is prompt sensitive and imperfectly aligned with predictive correctness (Kadavath et al., 2022; Tian et al., 2023). BTS therefore becomes a post-processing statistic over correlated samples.

These failures also explain the calibration gain. BTS weighting and SP tend to soften confident panel outputs in contested regions. Softening can improve binned reliability even while it worsens sharpness and top-label accuracy. This is the standard calibration/refinement tradeoff: a score may move probability mass away from overconfident peaks, improving ECE, while losing the proper-score or accuracy advantage of a sharper predictor (Brier, 1950; Guo et al., 2017; Kumar et al., 2019).

3.4. What a Positive Result Would Need

The counterexample in Proposition 1 is not a full impossibility theorem, but it shows what a positive LLM result would have to rule out: shared wrong priors that make wrong answers both common and well predicted. A positive claim needs either diversity assumptions strong enough to rule out those priors, or a calibration wrapper that makes no correctness claim. In particular, the statement

$$\arg\max_v \bar{x}_i(v) = Y_i^*$$

is an accuracy statement about the panel majority, while

$$\bar{x}_i(v) - \bar{p}_i(v) > 0$$

is a statement about unexpected commonness. These statements coincide only under a bridge condition linking unexpected commonness to truth. LLM systems do not get that bridge for free.

This point also separates BTS from ordinary confidence weighting. Confidence weighting assumes each reviewer confidence $c_{i,k}$ is at least monotone in its own correctness probability. BTS assumes a stronger peer relation, where a forecast about other reviewers contains information about whether the answer is informed. If the panel members share a blind spot, the peer relation can be internally consistent and externally wrong. A monitoring panel may share a training-derived misconception about an API, a mathematical convention, or a security pattern.

3.5. Estimands for Panel Uncertainty

The same panel transcript can support several estimands. Let $\widehat{y}_A(X)$ and $\widetilde{c}_A(X)$ denote the verdict and confidence output by rule A. Accuracy asks whether $\widehat{y}_A(X) = Y^*$. Brier risk scores the reported confidence. ECE measures binned reliability. Conformal coverage measures whether a set-valued prediction contains the truth. Routing utility measures the value of actions such as execute, call a tool, escalate, or abstain. These estimands can order rules differently.

This lens explains the otherwise awkward table. Majority vote wins accuracy. Confidence-weighted voting has the best Brier score in the main table. BTS-clamp has the lowest displayed ECE. SP has the best split-conformal diagnostic. None dominates. A panel system should choose the rule that matches the decision. A triage layer may value a conservative uncertainty signal. A human-review policy may care about calibrated ranking of ambiguous cases more than about small changes in top-label accuracy.

4. Distribution-Free Wrappers Around a Peer Score

A cleaner deployment design has two layers. The inner layer supplies a score or verdict. The outer layer supplies the statistical guarantee. BTS may be useful as an inner-layer score even when its incentive theorem does not transfer.

Split conformal prediction. Given exchangeable calibration and test data, split conformal prediction converts any nonconformity score into a finite-sample marginal coverage guarantee (Vovk et al., 2005; Bates et al., 2021; Angelopoulos and Bates, 2023). For a binary verdict, when $\widetilde{c}_i$ denotes confidence assigned to the predicted label, one simple nonconformity score is $1-\widetilde{c}_i$ for the predicted label and $\widetilde{c}_i$ for the alternative. The calibration quantile determines which labels enter $C_\alpha(X)$. The guarantee belongs to the conformal wrapper and exchangeability, not to the BTS score itself.

Anytime-valid testing. For sequential comparisons, confidence sequences and betting martingales allow monitoring without an α-spending penalty (Howard et al., 2021; Ramdas et al., 2023; Waudby-Smith and Ramdas, 2024). Let $U_i, V_i \in \{0,1\}$ be correctness indicators for a team rule and a reference rule. The centered contrast

$$X_i = \frac{U_i - V_i + 1}{2} \in \{0, \tfrac{1}{2}, 1\}$$

supports an e-process of the form

$$E_n = \prod_{i \le n} \left(1 + \lambda_i (X_i - \tfrac{1}{2})\right),$$

under the null $\mathbb{E}[X_i \mid \mathcal{F}_{i-1}] \le 1/2$, with predictable bets λ_i chosen so the factors stay nonnegative. The test remains valid under optional stopping when the supermartingale conditions hold. Again, validity comes from the betting construction, not from BTS.

Prediction-powered calibration. Prediction-powered inference uses model predictions as noisy proxies and corrects them with a labeled subset (Angelopoulos et al., 2023; 2024; Zrnic and Candes, 2024). A natural Brier functional is

$$\theta_{\text{Brier}} = \mathbb{E}\left[(\widetilde{c}(X) - \mathbf{1}\{\widehat{y}(X) = Y^*\})^2\right],$$

Binned ECE is a separate functional of the joint law of confidence and correctness after fixing the binning scheme. A partial-label rerun could estimate such targets with the LLM panel as a proxy and a smaller audited label set as correction data. The present experiment labels every item, so it reports point estimates rather than prediction-powered confidence intervals.

BTS can lose to majority vote and still help routing, but only after held-out calibration. A nonconformity score, routing score, or abstention score can be valuable if it is calibrated on held-out deployment data. But the guarantee must be stated at the wrapper level. A BTS score alone is not a coverage statement, a type-I error bound, or an accuracy certificate.

5. Strict Elicitation and Implementation Audit

BTS is easy to implement almost correctly. The theorem assumes a specific prompt and score: one population forecast, plus both the information and prediction terms.

Population forecasts, not per-peer forecasts. Prelec's rule asks each respondent for one distribution over the

population frequency of answers. Asking for one distribution per other reviewer changes the object. A per-peer prompt can be useful engineering telemetry, but it is not the elicitation used in the BTS theorem. The corrected pipeline therefore elicits a single $\widehat{p}_{i,k} \in \Delta(\mathcal{V})$ for the population verdict frequency.

Full score, not information score alone. The prediction-score term is not decorative. It is part of the mechanism that stabilizes truthful reporting in the human case. Dropping it leaves only the unexpectedly-common direction. The ablation therefore reports both the full-score variants and BTS-info, making clear when performance comes from the information term alone.

Positive weights without arbitrary clamps. The information score is a log ratio. Exponentiating $T_{i,k}$ or using a softmax respects the log-score interpretation. Clamping $T_{i,k}$ at a small positive value avoids negative weights but changes the ordering near zero. The table preserves BTS-clamp as an implementation ablation rather than treating it as the canonical rule.

SP as a separate mechanism. The Surprisingly Popular rule chooses an answer by unexpected commonness. It does not score reviewers and then weight votes. Its strong conformal diagnostic in this experiment should therefore not be reported as evidence that full BTS weighting is accurate. SP and BTS are neighboring peer-prediction ideas, not the same estimator.

Audit consequence. A report that says only "BTS was used" leaves out the parts that change the result: forecast format, prediction-score term, negative-score handling, and whether SP was counted as BTS. For LLM panels, those choices determine both the statistical target and the downstream behavior of the uncertainty layer. We report every diagnostic against its own estimand and repair the calibration plots using the exact table values.

6. Experimental Protocol

The main evaluation contains 871 binary verdict items. The standard tier contributes 824 items, mostly code-review and math-review cases. The code-review slice includes crypto timing, regular-expression denial of service, parser injection, concurrency, and logic bugs. The hard tier contributes 47 items patterned on CVE-style security bugs and proof traps. The panel contains four language-model reviewers: gpt-5.4, gpt-5.4-mini, gpt-5.3-codex, and gpt-5.2. This is a within-provider GPT-5-variant panel, so the experiment should be read as a shared-family stress test rather than a cross-provider diversity test. A planned cross-provider panel was not completed after gateway failures; no cross-provider claim is made. Elicitation uses one deterministic trial per item and reviewer, with cached calls keyed by model, prompt, temperature, seed, and item.

The primary endpoint is accuracy of BTS-E against majority vote, tested by paired exact McNemar at $\alpha = 0.05$. Displayed BTS/SP accuracy comparisons are controlled by Benjamini–Hochberg at $q = 0.10$.

Calibration is reported separately. ECE is a binned reliability diagnostic, not a complete calibration proof. Brier is a proper score, so it can penalize rules that improve reliability by becoming less sharp. We also report a small split-conformal coverage-error diagnostic.

A 150-item probe checks whether the ordering is a code-review artifact. It uses 50 examples each from MMLU, TruthfulQA, and ARC Challenge. The probe is not a target-population estimate. It is a covariate-shift check across qualitatively different input distributions.

The conformal diagnostic uses $n = 158$ items with a 50/50 calibration/test split. It reports empirical coverage, mean prediction-set size, and mean absolute deviation between empirical and nominal coverage across $\alpha \in \{0.05, 0.10, 0.20\}$. The split mixes tiers and is too small for deployment-slice validity claims. It is included to test whether SP and BTS scores are plausible nonconformity scores.

The implementation follows the audit in Section 5. Each reviewer reports one forecast over the population verdict frequency, not one forecast per peer. All variants remain visible instead of being merged into one BTS number.

7. Results

Accuracy favors majority vote. Table 2 reports the main ablation. Every BTS-family rule is less accurate than majority vote. BTS-E reaches 0.9724 accuracy against 0.9793 for majority vote, a 0.7 percentage point loss. The McNemar discordance is one-sided in the wrong direction: $b = 0$ BTS-only correct cases and $c = 6$ majority-vote-only correct cases, with exact two-sided $p = 0.031$. BTS-clamp reaches $p = 0.0039$, and BTS-info together with SP reaches $p = 0.0078$. The accuracy negative survives the planned Benjamini–Hochberg correction across the displayed BTS/SP accuracy comparisons.

Table 2. Seven-rule ablation on the main evaluation ($n = 871$, $K = 4$). McNemar b/c counts rule-only correct versus majority-vote-only correct. $*$ marks $p < 0.05$ before correction; † marks survival under Benjamini–Hochberg correction at $q = 0.10$.

Rule	Accuracy [Wilson 95%]	Brier	ECE	McNemar b/c	p
MV	0.9793 [.968, .987]	0.0230	0.0323	n/a	n/a
CW	0.9816 [.970, .989]	0.0215	0.0346	3/1	.625
BTS-E	0.9724 [.959, .981]	0.0290	0.0254	0/6	.031∗†
BTS-soft	0.9724 [.959, .981]	0.0290	0.0254	0/6	.031∗†
BTS-clamp	0.9690 [.955, .979]	0.0318	0.0220	0/9	.0039∗†
BTS-info	0.9701 [.957, .980]	0.0311	0.0231	0/8	.0078∗†
SP	0.9701 [.957, .980]	0.0308	0.0231	0/8	.0078∗†

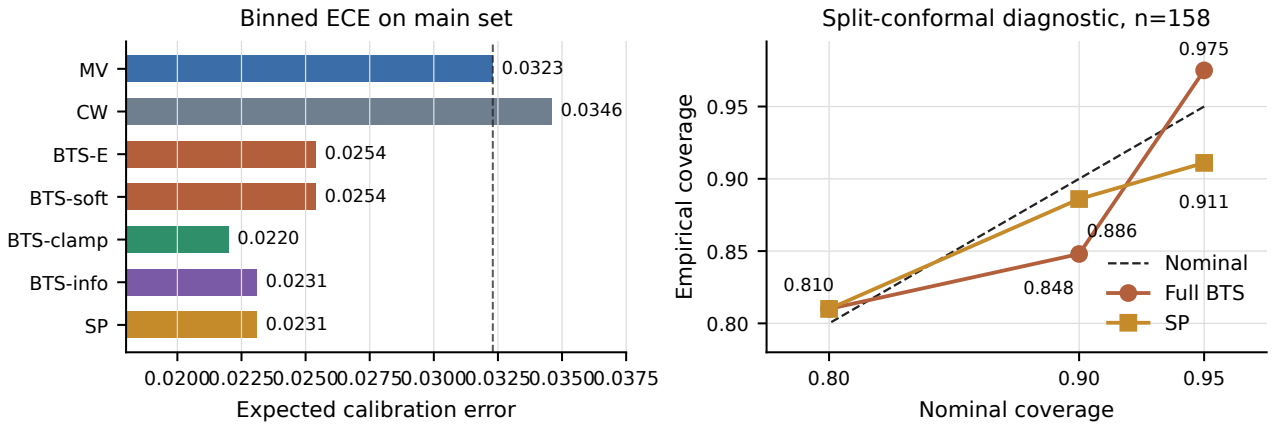

Figure 1. Repaired diagnostics using the exact values reported in Tables 2 and 4. Left: binned expected calibration error on the main $n = 871$ evaluation; the dashed line marks majority vote. Right: split-conformal empirical coverage on the $n = 158$ diagnostic split. These panels replace the earlier clipped ECE chart and the conformal plot whose caption used the wrong sample size.

Calibration improves by ECE but not by every score. The calibration result points the other way only for binned reliability (Figure 1, left). Majority vote has ECE 0.0323. BTS-E lowers ECE to 0.0254. BTS-info and SP lower it to 0.0231. BTS-clamp reaches 0.0220, a 32 percent reduction relative to majority vote under five equal-mass bins. With ten equal-width bins and a debiased estimator, absolute values move but the ordering is similar. Brier score moves against BTS: majority vote has Brier 0.0230 and BTS-clamp has 0.0318. Thus the result is not a blanket calibration victory. It is reduced binned reliability error with a loss of sharpness/proper-score performance.

The ordering is not only a code-review artifact. Table 3 reports the 150-item probe. Majority vote ties or beats BTS-E on MMLU and TruthfulQA. ARC Challenge is the lone reversal. Pooled over all three corpora, majority vote reaches 0.933 and BTS-E reaches 0.927. The probe is intentionally small, but it rules out the easiest explanation that the main result is peculiar to one code-review slice.

Table 3. Multi-corpus probe, 50 items per corpus. The probe checks ordering under covariate shift and is not a confirmatory target-population estimate.

Corpus	Rule	Accuracy	Brier
MMLU	MV	0.92	0.170
MMLU	BTS-E	0.90	0.186
TruthfulQA	MV	0.90	0.142
TruthfulQA	BTS-E	0.88	0.212
ARC Challenge	MV	0.98	0.018
ARC Challenge	BTS-E	1.00	0.002
Pooled	MV	0.933	0.110
Pooled	BTS-E	0.927	0.133

SP is the better conformal backend in this diagnostic. Table 4 and Figure 1 compare full BTS with SP as split-conformal scores. SP has lower mean coverage error, 0.0212 versus 0.0390. At $\alpha = 0.05$, SP empirical coverage is 0.911, below nominal 0.95. Because the split is small and mixes tiers, the number should be read as a diagnostic rather than a deployment guarantee. The narrow takeaway is that SP may be a useful nonconformity or routing score, while formal coverage still requires held-out exchangeable calibration data.

Table 4. Split-conformal diagnostic ($n = 158$, 50/50 split). Cells show coverage and mean set size. Mean error averages $|\text{coverage} - (1 - \alpha)|$ over the three α levels.

Backend	$\alpha = .05$	$\alpha = .10$	$\alpha = .20$	Err.
Full BTS	.975 / 1.94	.848 / 1.00	.810 / 1.00	.0390
SP	.911 / 1.90	.886 / 1.77	.810 / 1.00	.0212

Sequential evidence is uninformative at the ceiling. The anytime-valid team-versus-best-individual test stays near an e-value of 1 on the hard tier. Team and best individual both score 1.000 on every hard-tier item, leaving zero informative discordant trials. This is a power failure caused by saturation, not evidence against the betting construction. These data do not support a team-beats-best-individual claim; a harder restratified benchmark is needed to test it.

Robustness checks favor caution. Five-sample self-consistency on a matched 190-item subset reaches 0.9263, while single-trial majority vote reaches 0.9368 on the same items. A three-paraphrase robustness check over 50 items ranks confidence weighting as most stable and BTS-E as least stable. A small persuasion defense holds at 27 of 27 sessions, giving a one-sided 95 percent lower bound of 0.895. These checks point the same way: use panel uncertainty only with a stated policy and an audited deployment slice.

8. Diagnostics Beyond the Main Table

Per-tier behavior. The standard tier preserves the main ordering: majority vote remains above BTS-E and SP. The hard tier is less informative because the evaluated panel saturates many items. Majority vote reaches the ceiling, while BTS can miss items that every panel member called correctly once the peer score injects noise into the aggregation. This is precisely the scenario in which a monitor should resist decorative reweighting. When the base panel is already correct and unanimous, an auxiliary score needs a strong reason to move the decision.

Sample-size volatility. Across pilot sample sizes, a BTS variant can look better than majority vote before the sign reverses. The 871-item run makes the negative accuracy result statistically visible. Peer-prediction mechanisms are high-variance in small panels because the forecast and the empirical distribution are both estimated from a small number of reviewers. Reporting a BTS accuracy win on $n < 200$ items is therefore fragile.

External comparators. The natural baselines are not only majority vote. Self-consistency samples each reviewer multiple times (Wang et al., 2023). Multi-agent debate uses interaction before a final answer (Du et al., 2024). Critic-style systems use specialized models to find errors (McAleese et al., 2024). AutoGen-style systems coordinate agents through conversation (Wu et al., 2024). Calibration surveys place these methods inside the broader confidence-estimation landscape (Geng et al., 2024). In the matched self-consistency subset, extra samples did not recover the accuracy gap. The comparison argues against assuming that extra model calls automatically mean more independent evidence. They can amplify a shared prior as easily as correct it.

Persuasion and adversarial robustness. The 27-session persuasion check is small but relevant. All sessions held their pre-persuasion verdict after an adversarial reviewer was injected. The evaluated panel did not immediately collapse under the tested persuasion prompts. This does not establish robustness to adaptive attacks, prompt injection, reward hacking, or strategic model behavior. A production monitor should include attack openings, tool-output manipulation, and role-confusion cases in the calibration suite.

What would change the conclusion. A positive BTS-as-accuracy result would need one of three things: a model roster with demonstrably less correlated error, a task distribution where minority informed answers are common and forecastable, or an elicitation method that makes population forecasts substantially more reliable than stated confidences. The most direct next test is a cross-provider panel spanning GPT, Claude, Gemini, DeepSeek, and at least one open-weight model, with pairwise error-correlation matrices reported before rerunning the BTS/MV comparison. The same protocol should extend the leave-K-out ablation to roughly 10–12 reviewers and plot accuracy, Brier risk, and API cost against K. A positive BTS-as-calibration result needs less: held-out deployment calibration showing that the peer score ranks errors or ambiguous cases better than majority confidence. The current evidence supports the second, not the first.

9. Implications for LLM Panel Use

Do not use BTS as the default voting rule. When the operational target is top-label accuracy, the data support majority vote or confidence-weighted vote. A system that lets BTS override a majority answer should justify that override on a deployment-specific validation set. The burden is high for safety reviews, security triage, and code execution gates.

Use peer scores for routing and abstention. The confidence layer is where BTS/SP has the strongest case. A BTS/SP score can decide whether a system asks for more reviewers, escalates, or abstains. One calibrated policy might act only when majority vote is unanimous and the conformal set is singleton. It might escalate split votes or abstain when the deployment slice lacks calibration support. Such a policy should be evaluated as a policy, not only as an offline score column.

Calibrate on the deployment slice. LLM panel data are not exchangeable by default. Security reviews, math proofs, code patches, and factual claims occupy different strata. A conformal or calibration layer should therefore be rerun as Mondrian calibration by tier, task type, disagreement bucket, and model roster. The model roster is part of the statistical object. Changing one reviewer changes the panel distribution.

Report the assumptions, not only the leaderboard. Release the model roster, prompt family, population-forecast elicitation format, aggregation rule, accuracy, Brier, ECE, conformal split, abstention threshold, and failure examples. Keep BTS distinct from SP: BTS scores reviewers and then weights answers, while SP scores answers directly by unexpected commonness. Accuracy, calibration, and coverage must remain separate columns.

10. Limitations and Conclusion

The main limitation is scope. The 871-item set is large enough to make the BTS accuracy loss visible, but it remains curated. Some items use synthetic or header-conditioned ground truth, so absolute accuracy should not be read as an external benchmark score; within-item comparisons across rules are the estimand. Absolute accuracy is high, and the hard tier is saturated. The result should be repeated on a harder suite with more discordant cases.

The elicitation is also narrow. We use one population-forecast prompt family, and forecast quality is not independently audited beyond paraphrase checks. The conformal diagnostic is small and mixes tiers. It should be rerun with larger held-out sets and Mondrian calibration by deployment slice. The PPI discussion should likewise be tested with an explicit masked-label split, for example holding out a small audited subset and treating the remaining labeled items as proxy-only during estimation. The model roster is time-specific; new model families may change the shared-error structure.

Two further stress tests remain open. First, the evaluated tasks are binary or small multiple-choice verdicts; larger answer spaces may give unexpectedly-common answers more room to separate informed minorities from shared mistakes. Second, the paper treats inference-time LLMs as non-strategic predictors, but that claim should be probed with an incentive-aware prompt that states the BTS scoring rule and asks the model to optimize the score. If the scores and decisions remain unchanged, the non-strategic-predictor diagnosis becomes stronger; if they move, the deployment claim needs that qualification.

The original theory still matters. BTS gives a principled way to connect answers with population forecasts, and the human mechanism has a real equilibrium argument. The LLM setting changes the object. A frozen model panel has no incentive channel, no private-type report in the mechanism-design sense, and substantial shared error. Under those conditions, the peer score does not identify truth by itself.

The result is narrow, but it is usable. BTS is not a truth serum for LLM panels in the classical strategic sense. It loses to majority vote on accuracy in the main experiment and in the pooled cross-corpus probe. It also improves binned ECE and gives a promising SP score for conformal routing diagnostics. The lesson is not that BTS beats voting. Peer-distribution scores can be calibrated into abstention, routing, and coverage tools, but the guarantee comes from held-out validation and the distribution-free wrapper.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Foundations and Trends in Machine Learning, 16(4):494–591, 2023.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. Prediction-powered inference. Science, 382(6671):669–674, 2023.

Angelopoulos, A. N., Duchi, J. C., and Zrnic, T. PPI++: Efficient prediction-powered inference. 2024.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. I. Distribution-free, risk-controlling prediction sets. Journal of the ACM, 68(6), 2021.

Brier, G. W. Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1):1–3, 1950.

Dasgupta, A. and Ghosh, A. Crowdsourced judgement elicitation with endogenous proficiency. In Proceedings of the 22nd International Conference on World Wide Web, 2013.

Du, Y., Li, S., Torralba, A., Tenenbaum, J., and Mordatch, I. Improving factuality and reasoning in language models through multi-agent debate. In International Conference on Machine Learning, 2024.

Geng, J. et al. A survey of confidence estimation and calibration in large language models. In Proceedings of NAACL, 2024.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In International Conference on Machine Learning, 2017.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. Annals of Statistics, 49(2):1055–1080, 2021.

Kadavath, S. et al. Language models mostly know what they know. 2022.

Kong, Y. and Schoenebeck, G. Dominantly truthful multi-task peer prediction with a constant number of tasks. In Innovations in Theoretical Computer Science, 2022.

Kumar, A., Liang, P., and Ma, T. Verified uncertainty calibration. In Advances in Neural Information Processing Systems, 2019.

Lu, Y., Xu, T., Zhang, J., Kong, Y., and Schoenebeck, G. Eliciting informative text evaluations with large language models. In ACM Conference on Economics and Computation, 2024.

McAleese, N. et al. LLM critics help catch LLM bugs. 2024.

Prelec, D. A Bayesian truth serum for subjective data. Science, 306(5695):462–466, 2004.

Prelec, D., McCoy, H. S., and Seung, S. A solution to the single-question crowd-wisdom problem. Nature, 541:532–535, 2017.

Ramdas, A., Gruenwald, P., Vovk, V., and Shafer, G. Game-theoretic statistics and safe anytime-valid inference. Statistical Science, 38(4):576–601, 2023.

Shnayder, V., Agarwal, A., Frongillo, R., and Parkes, D. C. Informed truthfulness in multi-task peer prediction. In ACM Conference on Economics and Computation, 2016.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Proceedings of EMNLP, 2023.

Vovk, V., Gammerman, A., and Shafer, G. Algorithmic Learning in a Random World. Springer, 2005.

Wang, X. et al. Self-consistency improves chain-of-thought reasoning in language models. In International Conference on Learning Representations, 2023.

Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. Journal of the Royal Statistical Society: Series B, 2024.

Wasserman, L., Ramdas, A., and Balakrishnan, S. Universal inference. Proceedings of the National Academy of Sciences, 117(29):16880–16890, 2020.

Witkowski, J. and Parkes, D. C. A robust Bayesian truth serum for small populations. In AAAI Conference on Artificial Intelligence, 2012.

Wu, Q. et al. AutoGen: Enabling next-generation LLM applications via multi-agent conversation. 2024.

Zrnic, T. and Candes, E. J. Cross-prediction-powered inference. Proceedings of the National Academy of Sciences, 121(15), 2024.