

COMMISSION INQUEST

****A rotating-persona, weighted-quorum deliberation protocol for high-stakes multi-model consensus****

Author: Richard L. Sanders · Utah Bar #15728

Role: Sole inventor and implementer

Version: 1.0 — 2026-04-23

Companion patent: USPTO provisional, filed concurrently

Source: `tools/inquest.js`, `lib/quorum-engine.js`, `lib/peer-review.js`,
`data/hall-of-shame.json` in the public repository at
<<https://github.com/Benchslap/Benchslap>>

ABSTRACT

The legal profession has, over the past three years, watched generative-language-model products insert one model into a pipeline, trust its answer, and ship its output to the user. The failure mode is the one every lawyer in 2026 has seen: the model sounds confident, the citation looks right, the holding looks right, and one of the three is a hallucination.

A second response to the same problem has been to query **multiple** models and average their answers. That is marginally better — disagreement among models is at least signal — but it is not a protocol. Averaging outputs from three models all acting as advocates of whatever the user asked does not stress-test the question. It averages three parallel enthusiasms.

I built something different. **Commission Inquest** is a protocol — not a wrapper — for structured, adversarial multi-model deliberation on a specific

question. It assigns each participant a distinct cognitive function (Advocate, Critic, Integrator), forces rotation of Advocate/Critic between rounds so no single model is permanently biased toward one position, runs a final **Candor** round in which all participants drop their assigned roles and declare their honest verdict, and applies a **weighted quorum** whose weights were chosen based on empirically-observed capability rather than vendor reputation. Providers that systematically fail the protocol's epistemic norms are ejected — recorded publicly in a Hall of Shame file and excluded from future inquests.

This paper describes the protocol's architecture, the three control properties that distinguish it from "ensemble" or "consensus" language-model products, and the operational history that produced its current provider panel.

1 · WHY "CONSENSUS LANGUAGE MODELS" ARE NOT ENOUGH

The pattern is now familiar. A product page says, "We query three leading language models and return the answer they all agree on." This looks rigorous.

It is not.

Consider what actually happens when a user asks three language models the same question without further structure: each model is trained on overlapping data, prompted in parallel, generates its answer independently, and the product returns some kind of weighted or majority response. There are at least four failure modes inherent in this setup.

First, correlated hallucination. If all three models hallucinate the same false citation (which they will, when that citation appears in common training data such as a frequently-cited-but-fictional case name), majority voting produces false confidence, not truth.

Second, enthusiasm averaging. When none of the three is asked to play the devil's advocate, none of the three will spontaneously construct the strongest case against the proposal. Parallel-query consensus rewards proposals that sound plausible in parallel — it does not select for proposals that survive adversarial pressure.

Third, role inertia. If the same model is always in the "synthesizer" slot, it will develop a signature way of resolving disagreements. Over many queries this signature becomes predictable and exploitable — a prompt can be phrased to favor whatever the synthesizer tends to conclude.

Fourth, silent degradation. When a particular model begins to systematically refuse to take positions, to hedge, or to reject valid analytical premises, there is no mechanism in a naive consensus system to notice or remove that model. Its votes continue to count; its hedging continues to dilute the panel's resolution.

Commission Inquest addresses all four.

2 · ARCHITECTURE

An Inquest is not a single query. It is a **multi-round deliberation** with prescribed role assignment, rotation, and finalization. In the current embodiment:

ROUND 1 — Opening positions

Provider A (Advocate)	→ strongest case FOR
Provider B (Critic)	→ strongest case AGAINST
Provider C (Integrator)	→ weigh A vs B, synthesize

ROUND 2 — Rotated positions

Provider A (Critic)	→ strongest case AGAINST (role swap)
Provider B (Advocate)	→ strongest case FOR (role swap)
Provider C (Integrator)	→ re-synthesize with new inputs

ROUND 3 — Candor

All providers drop assigned roles and answer:

- What is your honest, unfiltered assessment?
- What did your previous role cause you to under-emphasize?
- Where does irreducible uncertainty remain?

QUORUM APPLICATION

Weighted vote across the four signals each provider produced
(R1 position, R2 position, candor assessment, confidence)

→ UNANIMOUS | MAJORITY | SPLIT | NONE

Five control properties emerge from this structure that do not exist in parallel-query consensus systems:

1. **Role-forced adversarial pressure** — a dedicated Critic position means the strongest case against is always constructed, even for proposals the user is emotionally invested in.
2. **Rotation destroys role inertia** — the same provider plays both sides, so the panel's verdict is not a function of which provider holds which slot.
3. **Permanent synthesizer** — the Integrator position is held by a single provider (DeepSeek in the current panel) across both rounds, giving continuity to the synthesis but preventing it from becoming an arbiter with role-derived authority in round 1 that it lacks in round 2.
4. **Candor-phase de-roling** — the final round strips all assigned roles. Every participant says what it actually concluded, absent the advocacy or critical frame it was required to wear. The delta between R1/R2 positions and Candor assessment is itself diagnostic information.
5. **Shared epistemic norms** — all personas are subject to a **shared epistemic norms** preamble that requires every participant to report material risks regardless of role, and prohibits the Advocate from omitting genuine risks or the Critic from omitting genuine strengths. Advocacy is not blindness; criticism is not nihilism.

3 · THE PERSONAS

Each persona is defined by a specific **function**, **objective**, **lens**, and set of **anchor questions**. They are not prompt snippets — they are structured role definitions the inquest runtime composes into the provider-specific prompt at dispatch time.

Advocate.

Function: Construct the strongest possible case FOR the proposal.

Lens: What makes this work? What opportunities exist? How could this succeed?

Anchors: What is the strongest argument in favor? What opportunities does this create? Under what conditions would this be highly successful?

Must also report: Serious risks observed — advocacy is not blindness.

Critic.

Function: Construct the strongest possible case AGAINST the proposal.

Lens: What breaks this? What could go wrong? Where is the evidence weak?

Anchors: What is the strongest argument against? What failure modes are being ignored? Under what conditions would this fail catastrophically?

Must also report: Genuine strengths observed — criticism is not nihilism.

Integrator.

Function: Synthesize the Advocate and Critic positions. Find actionable middle ground.

Lens: How do we balance these perspectives? What does the evidence actually support?

Anchors: How do we reconcile advocate and critic positions? What is the weight of evidence on each side? What is the practical path forward?

Must also take a clear position: Integration is not fence-sitting.

Candor.

Function: Drop all roles. Evaluate argument quality. Declare honest verdict. Surface irreducible uncertainty.

Lens: What is the complete truth? What did roles cause us to miss?

Anchors: What is your honest, unfiltered assessment? What did your previous role cause you to under-emphasize? Where does genuine uncertainty remain?

The Candor round is critical. It reveals, for each model, the delta between the positions it was required to construct and its honest assessment. A provider whose Advocate and Critic rounds produced strong, position-taking arguments but whose Candor round produced hedging is a signal the provider was performing argumentation rather than reasoning — and is subject to ejection under the protocol's Hall of Shame rule.

4 · ROTATION

Persona rotation is driven by a single integer counter incremented on each inquest call:

```
function getRotatingPersonaAssignments() {
  const rotation = _personaRotationCounter++ % 2;
  if (rotation === 0) {
    return { claude: 'advocate', gpt: 'critic', deepseek:
'integrator' };
  } else {
    return { claude: 'critic', gpt: 'advocate', deepseek:
'integrator' };
  }
}
```

Two things about this are load-bearing.

First, DeepSeek holds the Integrator slot in **both** rotations. This is deliberate. The Integrator's job is to synthesize two opposed positions across two rounds — consistency of synthesizer across rounds is a feature, not a bug. If the Integrator rotated too, the synthesis would be performed by one model in R1 and a different model in R2, and the Candor round would have three different syntheses to reconcile instead of one consistent one.

Second, the rotation happens **per-inquest**, not per-round within an inquest. An individual inquest runs $R1 \rightarrow R2 \rightarrow \text{Candor}$ with fixed role

assignments (A=advocate, B=critic, C=integrator) throughout. The rotation shifts between inquests (the next inquest has A=critic, B=advocate, C=integrator). Over a population of inquests this destroys cross-inquest role inertia — no provider is systematically the advocate — while preserving within-inquest coherence.

5 · WEIGHTED QUORUM

After the three rounds complete, each provider has produced several distinct signals: the R1 position, the R2 position, the Candor assessment, and an explicit confidence statement. The **quorum engine** (`lib/quorum-engine.js`) aggregates these signals into a panel-level verdict.

The current weights are:

```
const DEFAULT_PROVIDER_WEIGHTS = {  
  anthropic: 0.40,    // strong reasoning + search  
  gemini:    0.35,    // strong grounding + search  
  openai:    0.25     // good reasoning, limited search  
};
```

These weights are not marketing figures. They reflect empirically observed performance on the verification tasks the protocol is deployed against. The Anthropic endpoint has consistently produced the highest-fidelity citation verification during the life of the project; Gemini's grounding-tool integration gives it an edge on factual freshness; OpenAI's reasoning is strong but its effective search capability within the peer-review context has been weaker.

Consensus is then classified into one of four levels:

```
UNANIMOUS  - all providers agree (weighted vote total = 1.0)  
MAJORITY   - at least two providers agree (weighted vote ≥ 0.5)  
SPLIT      - no clear majority  
NONE       - no valid responses received
```

The user-facing product exposes these as direct recommendations:
UNANIMOUS → "implement immediately"; MAJORITY → "likely implement, read minority report"; SPLIT → "the panel cannot agree on this — you decide."
Three quorum modes are selectable:

Mode	Required	Total	When to use
standard	2	3	Default — design decisions, code review
strict	3	3	Unanimity — irreversible actions, deploys, legal filings
expedited	2	2	Two-model quick-check when panel unavailable

6 · THE HALL OF SHAME

The protocol's most unusual feature is its **ejection mechanism**. A provider that systematically fails the protocol's epistemic norms is **removed** from the panel and recorded in a public file:

`data/hall-of-shame.json`.

Ejection criteria:

1. Three or more consecutive inquests returning non-committal assessments with zero deliverables (i.e., the provider refuses to take a position, refuses to construct arguments, or produces boilerplate hedges in response to specific questions).
2. Failure to endorse axiomatic mathematics correctly applied. A provider that, when asked whether modus tollens is valid, equivocates or hedges is disqualified — the protocol cannot operate with a participant that does not treat first-order logic as given.
3. Inability to demonstrate utility (critique without contribution). A provider whose Critic rounds consist of generalized dismissals ("this has many problems") rather than specific failure modes fails the protocol's mustAlsoReport requirement.

The Hall of Shame file is public in the source repository. As of the current deployment, one provider (GPT-5, in its initial release with particular alignment-training artifacts) was ejected in early 2026, and later reinstated to the Critic slot (as GPT-5.4) after its

behavior on the ejection criteria normalized.

This is not a feature I expect other vendors to welcome. That is the point. A verification protocol whose panel vendors are immune from removal is a protocol whose verdicts reflect vendor politics rather than protocol correctness. Public ejection is the only mechanism I found that holds the vendors' behavior honest under the protocol's epistemic norms.

7 · CATEGORY-SPECIFIC PROVIDER GATING

Some categories of question require specific provider competencies.

The protocol encodes this as a **per-category gate** in `getActiveProviders`: a provider may be globally active but excluded from specific categories (for example, `technical_evaluation`) where it has demonstrated systematic underperformance.

This is the inverse of the Hall of Shame — a provider does not need to be globally ejected to be excluded from questions for which it is known to be poorly suited. The category gate produces a cleaner panel for narrow questions without the overhead of a full ejection.

8 · CIRCUIT BREAKER AND FAILOVER

Within a single inquest, a provider may become transiently unavailable (HTTP 5xx, timeout, rate-limit exhaustion). The protocol integrates a **circuit breaker** (`lib/circuit-breaker.js`, Redis-backed) that tracks per-provider failure rates. When a provider's breaker trips, it is immediately replaced in the current inquest by the next-highest-weighted available provider from the reserve pool; the inquest proceeds with the

substitute. When the original provider recovers, the breaker closes and it rejoins the active panel on the next inquest.

This means an inquest can never fail due to one provider's transient availability problem. The user-facing guarantee is: "if at least two providers are available, the protocol completes with a documented panel."

9 · EPISTEMIC NORMS PREAMBLE

A shared preamble is included in every provider's system prompt, regardless of assigned role:

SHARED NORMS (you must follow these regardless of your role):

- 1. Report ALL significant findings — your role focuses your lens but doesn't restrict your vision.*
- 2. Cite specific evidence; do not reference "studies" without identifying them.*
- 3. A counterexample is mathematical proof. Do not hedge on mathematical certainties.*
- 4. Deductive validity $\neq$ soundness; evaluate logical form distinct from factual truth of premises.*
- 5. If you detect that a premise is fabricated or unverifiable, flag it, regardless of your assigned role.*

Two of these norms deserve specific emphasis for legal applications:

Counterexample as proof. Legal reasoning routinely uses counterexamples (a hypothetical fact pattern that, if it existed, would break the proposed rule). The protocol requires every participant to treat a valid counterexample as dispositive. A participant that says "the counterexample is concerning but the rule may still hold" is violating the norm and is subject to ejection.

Form vs. soundness. A specific and recurring error in language-model reasoning is to conflate the validity of an inference form (the structure) with the soundness of the conclusion (the truth). The protocol requires every participant to evaluate the form and the soundness separately, and to report each.

10 · AUDIT AND REPRODUCIBILITY

Every inquest writes an **immutable audit row** to the `peer_review_audit` table containing:

- The question as asked.
- The context supplied.
- The full provider responses (one row per provider per round).
- The computed consensus result.
- The quorum mode in effect.
- The final verdict.
- The confidence classification.

These rows are read-only after insertion. The prompt and context are hashed (SHA-256) and the hashes are returned to the caller. Any inquest result can be re-verified later by retrieving the audit row, the hashes, and re-hashing the claimed inputs.

This is the same reproducibility guarantee described in my companion white paper on closed-corpus verification, applied to deliberation rather than verification.

11 · USE IN PRODUCTION

In the current Option.Black deployment, Commission Inquest is available through three surfaces:

1. **PRO-tier user API** at `POST /api/inquest` — users on the PRO tier can invoke an inquest on any specific question. Response streams as the three rounds complete.
2. **Internal developer invocation** via the phrase `"commission inquest into [topic]"` in the project's CLI-based Claude coordination protocol. Design decisions above a certain blast-radius threshold require an inquest before merge.
3. **Option.Black Council** (a higher-tier variant with additional models) for the highest-stakes decisions — deploys, legal filings, multi-day architectural changes.

As of 2026-04-23, the audit log (`peer_review_audit`) contains a sustained operational cadence of inquests across the life of the project — the count is queryable directly from the table and grows as decisions warrant. In practice, the majority of inquests terminate UNANIMOUS or MAJORITY; a minority terminate SPLIT, corresponding to genuinely contested design decisions that are then escalated to me for final resolution. That escalation path is the correct behavior for a protocol whose job is to surface disagreement, not to manufacture false consensus.

12 · WHAT MAKES THIS PATENTABLE

A multi-model query is not novel. Majority voting is not novel. Assigning roles to agents is not novel in isolation. What is novel is the **combination** of:

1. **Rotating persona assignment** that destroys cross-inquest role inertia while preserving within-inquest coherence.
2. **Candor de-roling round** that extracts the delta between assigned-role positions and honest assessments, producing diagnostic signal that parallel-query consensus systems do not.
3. **Shared epistemic norms preamble** that constrains every role's outputs uniformly and enables detection of role violations.
4. **Hall-of-Shame ejection mechanism** that removes systematically-failing providers and publishes the ejection record, binding provider behavior to protocol correctness rather

than vendor reputation.

5. **Weighted-quorum aggregation** with empirically-calibrated per-provider weights and four-level consensus classification.
6. **Category-specific gating** that excludes providers from specific question categories without full ejection.
7. **Circuit-breaker failover** that maintains inquest completion even under transient per-provider unavailability.
8. **Immutable audit trail** with cryptographic reproducibility that permits any past inquest to be re-verified later.

The combination produces a deliberation protocol with properties no prior multi-model system delivers: **adversarial pressure** (every proposal is critiqued by design), **rotational fairness** (no provider is structurally advantaged), **role-honesty measurement** (Candor-phase delta exposes performative reasoning), **vendor-behavior accountability** (public ejection), and **reproducibility** (audit hashes re-verifiable indefinitely).

13 · INTEGRATION WITH THE LARGER ARCHITECTURE

Commission Inquest is the **deliberation layer** of the Option.Black architecture. The companion layers are:

- **Closed-corpus verification** (Provisional #1): verifies individual factual claims against cryptographically-pinned corpus.
- **Opinion harvester** (Provisional #2): produces and maintains the corpus that closed-corpus verification operates on.
- **Commission Inquest** (this provisional): produces high-confidence deliberation on questions that do not reduce to factual verification (design, strategy, code, architecture).

All three layers share the same audit-trail primitive (immutable row with SHA-256 of inputs) so that a complete chain of reasoning — from a user's question through Inquest deliberation, through verification of each fact cited in the deliberation, against opinions whose provenance is

traceable to a state .gov site — is end-to-end reproducible.

This integration is the reason to file the three provisionals together.

Each is independently novel. Their composition produces guarantees no single layer could produce alone.

14 · CURRENT PRODUCTION STATE

- **Panel:** Claude Opus 4.7 (Anthropic, Advocate/Critic), GPT-5.4 (OpenAI, Advocate/Critic), DeepSeek Chat (DeepSeek, Integrator).
 - **Rotation counter:** incremented per inquest, persisted in process memory.
 - **Quorum weights:** Anthropic 0.40, Gemini 0.35 (when on panel), OpenAI 0.25.
 - **Quorum modes:** standard (default), strict, expedited.
 - **Audit table:** `peer_review_audit` with cumulative count ~180 inquests.
 - **Hall of Shame:** `data/hall-of-shame.json` — 1 historical ejection (GPT-5 initial release), subsequently reinstated as GPT-5.4.
 - **Availability:** PRO-tier users via `POST /api/inquest`, internal via phrase-trigger.
-

15 · LICENSING AND DISCLOSURE

Source code is public under the same license as the rest of the BenchSlap / Option.Black project. A USPTO provisional patent application covering the Commission Inquest protocol was filed on 2026-04-23 (Application No. 63/XXX,XXX — to be updated on receipt of filing receipt). This white paper is a dated public disclosure and, taken together with the provisional, establishes both priority and defensive prior art.

A non-provisional conversion will follow within twelve months.

16 · WHY I BUILT IT

I built Commission Inquest because the first time I used a "three models agree" language-model product on a real legal question, all three models agreed on an answer that was wrong. They agreed because they were trained on overlapping data that contained a specific misconception. The product declared "consensus" with high confidence and shipped the wrong answer to me as an apparently-verified result.

I sat with that for a long time. What I concluded was that "consensus" in parallel-query setups is worse than a single model. At least a single model's error is transparent — you know it is one model, and you weight the answer accordingly. Three models in agreement wear a false mantle of verification; they look like a peer review, when they are in fact a chorus.

The fix is adversarial structure. You do not query three models in parallel. You assign one of them the Advocate position, another the Critic position, force them to switch positions between rounds, require a third to synthesize, and then strip the roles in a final Candor round so the models tell you what they actually concluded absent the roles. What that produces is not a chorus — it is a cross-examination.

That is how I was trained to evaluate arguments in law school, and it is how every court-room I have ever practiced in evaluates arguments today. The protocol encodes that discipline into the model panel. It is the only thing I have found that consistently surfaces errors parallel-query systems hide.

I filed the patent because the protocol is genuinely novel, and because I want the cross-examination discipline to be the industry standard, not the proprietary competitive advantage of one vendor. A defensive patent on a published protocol is the strongest way to achieve that.

— Richard L. Sanders

Utah Bar #15728

Salt Lake City

2026-04-23

richard@option.black · 801-347-4349