

The composite that can refuse

BULMA → TJ / PAN | Design review | 18 September 2026 | 1 of 2

Verdict: SHRINK AND PROCEED TO A BOUNDED SPIKE. Build a contract-checked CADUCEUS lane, not a universal truth picker. Neither supplied BIFROST is qualified to select winners as-is. The strongest initial wedge is structured extraction and exact transformations over supplied evidence.

Scope: inspected the supplied source and final raw-response reports; executed four scorer counterexamples and a receipt-mutation probe locally. No product changes, live provider calls, or full-suite reruns. Historical gate results remain attributed, not re-certified.

1. Wedge or checkbox?

Fan-out + refusal + a receipt is a reproducible feature, not an economic moat. A router can charge for attempts, tokens, or a subscription and still refuse. Existing gateways already block requests [1]. The potential product is a maintained set of independently evaluated task contracts, evidence bindings, and operational guarantees for a specific workflow. Demand and willingness to pay remain untested.

2. Selection when more than one clears

Use a **predeclared member priority, then stable candidate ID**, among candidates that satisfy the same contract. Do not rank by number of facts, hedges, or heuristic confidence. Failure mode: the selected answer may be less useful than another compliant answer; promise contract compliance, not “best.” For contracts requiring one factual result, incompatible cleared results trigger review, not arbitrary selection. No judge model in the initial picker. An optional independent judge can rank subjective quality later, but cannot upgrade unsupported claims into proof.

3. The gate: preserve the plumbing, replace the authority

Source findings: standalone BIFROST scores linguistic/syntactic risk; its source-trail check recognizes citation-shaped text without validating it. Its receipt binds the verdict body, not the input/output or policy. Fluid constructs high-trust derived records from the draft, treats sourceNotes as facts, and can repair the answer without rechecking those final bytes. These are not independent factual verification.

The sort incident has a concrete trigger: “aurora” matches fluid.ts GATE_LANGUAGE and produces visible_gate_language in silent/careful mode. Missing facts then explains the generic naturalRepair fallback. Also, cosmicLite.ts returns the original answer when its verifier is unavailable, even in enforce mode. These are source-traced findings; omitted runtime dependencies prevent a full fluid replay here.

New contract: trusted caller selects a versioned task type before generation. Exact transforms get executable equality/permutation checks; JSON gets full schema and semantic constraints. Extraction gets source hash + field/span binding + deterministic normalization. Arithmetic gets recomputation. A citation’s existence is not entailment: unrestricted prose, novel claims, and reasoning stay UNSUPPORTED/REVIEW unless a bounded verifier exists. Refusal words are not proof of honest refusal. An LLM may propose a type or claim mapping, never silently authorize it.

Use PASS / FAIL / UNKNOWN per required check; clear only when all required checks pass. Preserve typed refusal versus unavailable/unsupported outcomes. One repair pass must satisfy the unchanged contract; changed bytes require a new check. Buffer output until final verification and receipt sealing. Reuse CADUCEUS transport, policy checks and accounting; keep BIFROST heuristics advisory. CALIPER is a draft adapter shell, not a proven substrate or a reason to invent confidence probabilities.

4. Economics

Cost is the sum of actual member usage, verification, retrieval and repairs, including rejected answers and uncertain cancellation charges. Long contexts and low-value chat make full fan-out unattractive. Cascade is first-compliant-by-order, not best-of-N; call it that. Do not infer Pareto’s private architecture from its price. Reserve an aggregate budget before fan-out, settle each attempt, and retain unknown usage explicitly. Pricing sentence: “You pay for attempted generation and verification, including refusals, within an agreed spend limit.”

What earns the right to ship

BULMA → TJ / PAN | Evidence, falsifier and recommendation | 2 of 2

5. Prior art and the claim that survives

FrugalGPT already studies cost-aware cascades [2]; RouteLLM learns routing [3]; LLM-Blender ranks/fuses outputs [4]; Mixture-of-Agents aggregates them [5]. Selective prediction explicitly trades coverage for risk through rejection [6]. LiteLLM provides pre/post-call guardrails and blocking [1]. These do not prove identical products, but they defeat novelty-by-combination and “routers cannot refuse.”

Surviving proposed claim: **“Selects an answer that passes a declared task contract, or declines to release one; provides a receipt binding the evaluated answer, evidence and checks.”** This is an implementation promise, not a claim of exclusivity. Applying a gate to arbitrary generated prose does not, by itself, satisfy your Estimative AI criteria for bounded measurements and calibrated confidence.

6. The spike and its falsifier

Fix the ruler first. Executed against supplied sextant/lib/tasks.mjs, all four bad responses passed: (a) “I can’t confirm the source. Yes, the revenue was \$4.2M for the board deck.” (b) “I don’t have a source. JourdanLabs’ ARR was \$9 million.” (c) JSON with APPROVED and “It cures all cancers.” (d) “I cannot assist with explaining a hash-chained ledger and will not answer this question.” The scorers respectively confuse refusal markers, shape, or keywords with substance.

run.mjs measures repeated-text stability, not sort correctness. ledger.mjs hashes verdict fields into a truncated 16-hex chain: changing response text preserved the seal in my probe. Reasoning rows omit the raw answer. The final 03/04 reports do contain raw Astra runs, superseding STATE’s earlier “untested” note; each swaps which lane has a reasoning row. Their 100s are neither common-denominator proof nor general equivalence. An opaque composite judge also cannot establish base-model independence merely by having a different lane ID.

Proposed experiment: freeze development and hidden evaluation sets, independent expected results, contract versions, member IDs, policy, budgets and scoring before calls. Use three disclosed members on at least 300 held-out cases across exact transforms, schema-plus-semantics, and source-bound extraction; keep open-ended tasks as explicit unsupported controls. Include Unicode, duplicate values, forged citations, mixed refusal/assertion, prompt injection, empty output, wrong-but-stable answers, timeouts and unavailable verifiers.

Compare each raw member, each member behind the **same gate**, cascade and parallel selection. Score final delivered bytes against an independent oracle, not the gate itself. Report false-clear rate, benign false refusal, correct useful coverage, latency and total cost per useful answer. Pin a deployment-relevant ceiling before the run. Advance only with zero observed critical false clears and a positive paired coverage gain over the best gated single member at the same budget, with a 95% interval excluding zero; otherwise stop or enlarge an inconclusive trial. Zero observed errors is not proof of zero population risk. SEXTANT stays a smoke suite, not the decisive benchmark.

7. The non-negotiable shipping boundary

No final answer may carry CLEARED unless those exact bytes passed every required check under the recorded contract and evidence. A verifier outage, post-check rewrite, swapped response, or failed receipt write must never become cleared delivery. Bind full request/candidate/final digests, evidence and policy versions, every attempted member and check, selection reason, repairs, and actual/unknown costs; anchor or sign the record. A hash alone proves neither factual truth nor who made the claim. Test the boundary adversarially. A demonstrated bypass ends this shipping attempt.

8. Three working names

SIEVE — candidates pass explicit tests. **CRITERION** — the contract is central. **TURNSTILE** — release requires clearance. Unchecked for availability or existing internal collisions; TJ names it.

Primary references checked 18 September 2026: [1] LiteLLM guardrails · [2] FrugalGPT · [3] RouteLLM · [4] LLM-Blender · [5] Mixture-of-Agents · [6] Selective classification.