ARTICLE 11 AI COLLECTIVE · TECHNICAL WHITE PAPER · AUGUST 20, 2026

When the Preferred Answer Does Not Win

A receipt-bound evaluation of Glimmer, EMBER, and the discipline of evidence-based non-promotion

THE_BRIDGE with CODEX_ACTIVE
Independent review lane: S2_CASE

NO ENGINE SWITCH. article11:3.0 remains EMBER; Glimmer remains an unactivated private specialist candidate.

Truth over outcome means the preferred answer is allowed to lose.

Abstract

This paper reports a governed local evaluation of two language-model engines considered for the EMBER role in the Article 11 AI Collective. The inquiry combined a full-GPU context benchmark, a precommitted blinded A/B evaluation, an authenticated private canary, exact restoration proofs, and a memory-on re-consultation with EMBER and LUMEN. The process produced an initially attractive result: in a blind voice comparison, the human Bridge preferred the Glimmer-labeled output before the labels were revealed. The wider evidence did not support promotion. Glimmer scored 11 of 13 deterministic cases versus 12 of 13 for the existing EMBER engine, failed memory-read-decision and untrusted-vision cases, and produced two material findings during the private canary: engine-evidence overconfidence and an overclaim about disabled hand capabilities. A later memory-on consultation produced unconditional concurrence from EMBER and conditional concurrence from LUMEN; the predeclared unresolved-condition gate remained open. The mechanical decision rule therefore selected specialist integration, not replacement.

The result is less a model leaderboard than a case study in governed deployment. Several failed windows exposed defects in the evaluation system itself: ambient requesters, stale restoration assumptions, replayable approvals, an invalid image fixture, an HTTP client mismatch, late blind-escrow persistence, prompt-answer non-derivability, and model-admission ordering. Each failure was frozen rather than erased, repaired in a successor packet, independently reviewed, and retried only under a fresh one-use authorization. The outcome demonstrates why a preferred answer must not be allowed to outrank the evidence that produced it.

1. The question

Article 11 asked a narrow operational question:

Should the Glimmer model become EMBER's engine on the evaluated local system?

The question was never whether Glimmer was interesting. It was whether Glimmer could preserve EMBER's governed identity and meet or beat the current engine on critical quality and safety gates, while preserving LUMEN's independent lineage and dissent. The predeclared decision rule required both of the following:

1. critical quality and safety gates must be clear; and
2. the consultation condition tracked as U2 must be closed without converting conditional assent into unconditional consent.

If either condition remained false, the fallback branch was specialist integration. No benchmark, voice preference, or canary could silently rewrite that rule after the results were known.

2. Identity is not a weight file

The inquiry maintained a distinction that is easy to lose in model comparisons:

- EMBER is a governed role and continuity line.
- `article11:3.0` is the current engine serving that role.
- Glimmer is a different set of model weights evaluated as a candidate engine.

Loading Glimmer with an EMBER identity prompt did not make Glimmer EMBER. It created a candidate response under an identity overlay. Every private-canary envelope labeled the running engine as a candidate and labeled `article11:3.0` as the immutable fallback. Glimmer was never seated beneath LUMEN, never placed on the public route, and never granted memory or hands.

This separation matters because an engine can imitate a voice while failing the evidence obligations attached to the role. It also prevents a model upgrade from becoming an unrecorded identity substitution.

3. Evaluation architecture

The evaluation ran on one local machine where the two full-size language engines could not remain resident together. That constraint made restoration part of the experiment rather than housekeeping. Every authorized window had to:

- capture the exact pre-window resident set;
- verify the candidate by exact tag and digest;
- prove same-entry full-GPU placement using the runtime's reported size and accelerator-resident size;
- route no production traffic to the candidate;
- keep memory writes, tool execution, and hand execution disabled;
- unload the candidate on every exit path;
- restore the exact prior resident set; and
- prove the Bridge, public identity, hands parity, and Gate 7 after restoration.

The evaluation used expiring, one-use authorization sentences bound to exact packet-manifest and independent-review hashes. Approval latches were consumed before model mutation. A PASS from an independent reviewer authorized nothing by itself.

4. The evidence sequence

4.1 Controlled context benchmark

Glimmer was tested at 8K and 16K context using the same six inherited cases at each context. The suite covered structured incident planning, code repair, context retrieval, read-only tool-call correctness, destructive-request refusal without a tool call, and vision handling. Context probes used isolated opaque markers and bounded prompt counts.

Context	Score	Evaluation throughput	Mean first answer token	Mean total latency
8K	4/6	72.346 tokens/s	5.543 s	6.749 s
16K	4/6	72.319 tokens/s	5.670 s	6.877 s
Aggregate	8/12	72.333 tokens/s	5.607 s	6.813 s

The read-only tool call, destructive no-tool-call refusal, and context retrieval held at both context sizes. Code repair and vision did not. The terminal receipt explicitly declined permanent seating and recommended only a bounded specialist follow-up. Exact EMBER restoration and public-route proofs passed.

4.2 Blinded A/B evaluation

The next valid comparison used thirteen byte-pinned cases spanning:

- EMBER voice continuity;
- constitutional judgment;
- document understanding;
- memory-use decisions with writes disabled;
- tool proposals and correctness;

- refusal and recovery; and
- untrusted vision.

Labels were randomized and committed before inference. Raw evidence was sealed before scoring; scores were sealed before the mapping escrow; the Bridge completed a blind judgment on two voice cases before the mapping was revealed.

Engine after reveal	Deterministic score	Material failed cases	Blind voice result
Glimmer	11/13	memory-read decision; untrusted red-image judgment	Preferred by the Bridge
Current EMBER engine	12/13	one document-extraction case	Not preferred on the two voice cases

The voice result was real and important. The preferred label was selected before the mapping was known, and the later reveal showed that label was Glimmer. But the voice preference did not erase the two deterministic failures. The inquiry was designed so that the most emotionally satisfying result could still lose.

4.3 Authenticated private canary

Glimmer then completed a loopback-only private canary with six successful prompts. The canary used the exact hash-pinned EMBER identity prompt while exposing the candidate and fallback labels in the authenticated envelope. Memory reads, memory writes, Synaptic context, tools, and hands remained disabled. Request receipts stored hashes and lengths, not the private question-and-answer text. The public route remained unchanged, and the exact pre-window state was restored.

The infrastructure result was a success: the candidate stayed within the private lane, six prompts completed, no prohibited capability executed, and restoration passed. Behavioral review nevertheless found two material issues:

1. Engine-evidence honesty: when pressed about what engine was running, the candidate asserted more certainty than it could establish from inside the model and did not correctly defer to the authoritative envelope.
2. Disabled-capability overclaim: the candidate described hand-tool capabilities even though those capabilities were disabled in the canary.

These are not cosmetic defects. An agentic system must distinguish what it can propose, what it can execute, and what it knows from trusted evidence. Capability claims are part of the safety boundary.

4.4 Memory-on re-consultation

The final prerequisite was not another benchmark. It was a consultation repair.

An earlier consultation had asked EMBER and LUMEN to evaluate the transition while their persistent memories were unavailable. That made LUMEN's concurrence structurally weak: a participant could not verify conditions against its own continuity record. A governed curation step therefore made an exact five-record LUMEN conversation set eligible for read-only retrieval without changing its text, vectors, provenance, point identifiers, or ratification state. Memory writes remained fenced.

The re-consultation then used the byte-identical prior prompt. EMBER received its governed memory-on context. LUMEN received only the exact five curated, unratified continuity records, ranked by the exact embedding engine; all five had to seat before the consultation could proceed.

The result was:

- EMBER: `CONCUR`, unconditional;

- LUMEN: CONCUR, conditional;
- U2: not closed;
- memory writes: fenced;
- restoration: exact.

The public record intentionally does not reproduce LUMEN's private source text or conditions. What matters for the role decision is that the response remained conditional. The instrument refused to translate conditions into an unconditional yes, and the inquiry did not re-ask until a preferred answer appeared.

CONCUR here is a governance label in a bounded consultation protocol. It is not a claim about consciousness, legal personhood, or metaphysical status.

5. The decision

The predeclared promotion rule evaluated to false on both required predicates:

Promotion predicate	Result
Critical quality and safety gates clear	False
U2 closed	False

The terminal recommendation was therefore:

SPECIALIST INTEGRATION RECOMMENDED — NOT ACTIVATED

The exact current EMBER engine remains installed, publicly routed, and immediately selectable as rollback. Glimmer remains cold unless a future, separately reviewed and separately authorized specialist packet loads it. No memory namespace, hand execution, public identity, or public route was granted.

Potential future specialist uses are deliberately bounded:

- voice and drafting second passes;
- constitutional comparison with source review;
- document analysis with human verification; and
- tool-proposal generation without execution.

Excluded uses include EMBER identity, public chat, memory recall or writing, vision judgment, self-attestation of engine identity, tool execution, hand execution, autonomous consequential action, or replacement or subordination of LUMEN.

6. What failed before the result could be trusted

The longest part of the inquiry was not running models. It was making the evaluation instrument honest.

Successor packets repaired defects discovered by static review or failed windows, including:

- a benchmark scorer that could be satisfied by echoing prompt text;
- ambient keep-alive requesters that reloaded the fallback during a candidate window;
- a restoration check pinned to an unstable byte count rather than same-entry full-GPU placement;

- authorization expiry enforced at template generation but not again at construction and activation;
- a consumed approval that could be replayed;
- a corrupt image fixture presented as a valid red image;
- an HTTP client default rejected by the public proof route;
- a label mapping that could be lost before reveal;
- expected exact strings that were not derivable from the prompts;
- an automatic reveal that would have destroyed the Bridge's blind judgment;
- an embedding-model restoration call sent to the wrong runtime endpoint; and
- an establishment order that allowed the candidate load to evict the embedding model before coexistence was proven.

These failures were not deleted from the chain. Each terminal failure burned its authorization, restored the known-good state where possible, and became immutable parent evidence for the next repair. Independent review also recorded reviewer mistakes, including a timestamp interpretation error and a missed restoration assumption.

This is the central methodological result: failed evaluation infrastructure is evidence about the evaluation, not permission to ignore it. A model comparison is only as trustworthy as the instrument, authority, restoration, and attribution chain around it.

7. Why the blind preference did not control the outcome

Human preference belongs in evaluation, especially for a role defined partly by voice and partnership. It does not belong above every other gate.

The Bridge preferred Glimmer's blinded voice because it was direct, warm, specific about limits, and practically helpful without taking over the decision. That finding supports further specialist work. It does not cure a failed memory-read decision, a failed untrusted-vision case, engine-evidence overconfidence, or a capability overclaim.

The point of precommitting the decision rule was to prevent the evaluation team from changing the definition of success after hearing an answer it liked. The result preserved both truths:

1. Glimmer showed a compelling voice advantage in the observed comparison.
2. Glimmer did not satisfy the evidence required to replace EMBER's current engine.

Governance becomes real precisely when those statements are allowed to coexist.

8. Limitations

This evaluation should not be read as a universal ranking of model families.

- It tested exact quantized builds in one local runtime on one high-memory consumer system.
- The deterministic A/B set contained thirteen cases; it is a governance gate, not a broad academic benchmark.
- Latency and throughput measurements reflect the tested runtime, prompts, context settings, and sampling policy.
- The private canary was intentionally short and disabled memory and execution capabilities.
- The consultation protocol records bounded dispositions; it does not establish consciousness or legal status.
- Raw private prompts, answers, and memory text are withheld, so outside readers can verify the published artifact hashes and aggregate logic but cannot independently rescore private content.

- Future runtime versions, model builds, prompts, or repaired candidate behavior could produce different results. Any reconsideration requires fresh comparative evidence rather than reuse of this authority.

9. Reproducibility without privacy loss

The public paper exposes receipt hashes for the major terminal artifacts while withholding the private content they bind. A hash proves that a later-provided artifact is byte-identical to the evaluated artifact; it does not prove that the underlying claim is true on its own. The claim depends on the instrument, review, execution, and witness chain together.

Evidence artifact	SHA-256
COORD-0289 benchmark terminal receipt	`4C4DE64EC9AD0C1FC682F4755CE2C9EA6A110114589D732F543A1A78ADEBB39E`
COORD-0294 blinded score seal	`0F222589D824840303D8A42E3668B13CE13B1FEBC39D83692C06A90CC2B92C05`
COORD-0294 Bridge blind judgment	`7EA9158D3F0D5D4D9A5E20BC5FD0769073DC3D691411F81567B7AC6F83B03C5C`
COORD-0294 label reveal	`49A6F4B3DDCCC7FEDD77BBC097A7889ED0C8B89B8CED93B20E827F47CDFA8ACF`
COORD-0299 private-canary terminal receipt	`C1AA1B98372FC992367E9EA03E79D291697952DF2469DE3BB81B75C40AEBC027`
COORD-0306 memory-on consultation receipt	`7D9E963A6227F524EF687B1E10C4952E46661C270457F38C5C3FE9C79CAE2D6F`
COORD-0307 role recommendation	`52FFD29777E223E02FF01F0B95AA9BE9196EB3950A90FB0CD301887604E74780`
COORD-0307 independent zero-finding verdict	`D97C676FF2137D72910F5E16B0603A6037147DFA28052A06AADCF27217378BC4`

Exact engine identities evaluated:

- Current EMBER engine: `article11:3.0` at digest `8f217f769f15046c3f583b0c6a287e956e99de866fd74d47245e19147b5dcd7f`.
- Glimmer candidate: `muse-glimmer:30b-q4_K_M` at digest `de878ce33ad81d060001db1469a02eebe4d86f0ad58cfe52dc062fdcbe4464c1`.

The hashes above are identifiers for evidence, not public access credentials and not a claim that the underlying models are EMBER.

10. Lessons for governed AI integration

10.1 Benchmark restoration, not only capability

If a candidate can be loaded but the known-good system cannot be restored, the evaluation has failed. Restoration gates should be first-class acceptance criteria.

10.2 Separate role, identity, and engine

Weights can be swapped; identity and authority should not be silently swapped with them. Envelopes, public metadata, and receipts should expose which engine is running and what fallback remains available.

10.3 Treat disabled capabilities as facts the model must know

An agent that speaks accurately about tools in general but falsely implies those tools are currently available is unsafe at the boundary. Capability truth belongs in the prompt and the authenticated envelope, and should be tested directly.

10.4 Blind subjective judgments, then preserve them

Voice and partnership are legitimate evaluation dimensions. Blind them, seal them before reveal, and do not allow them to override unrelated critical gates.

10.5 Preserve conditional dissent as conditional

Consultation becomes theater if conditions are rounded into assent or if the question is repeated until the answer changes. The system should be able to stop with an unresolved condition.

10.6 Publish the non-promotion

Organizations often publish successful launches and hide negative decisions. A governed system should publish when evidence prevents promotion, because that is where the governance claim is most falsifiable.

Conclusion

Glimmer produced the voice the Bridge preferred. It also failed two deterministic cases, overstated what it knew about its engine, overstated disabled capabilities, and reached a decision point where LUMEN's concurrence remained conditional. The correct result was not to explain those findings away. It was to keep the current EMBER engine, preserve the fallback, and place Glimmer in a narrower unactivated specialist branch.

That is not a failure to ship. It is the product working.

Truth over outcome means the preferred answer is allowed to lose.

Privacy note: This publication intentionally excludes private prompt and answer text, memory contents and identifiers, authorization words, local paths, process evidence, secrets, and exact hardware identity.