// WHITEPAPER · V1.0

Imbas Whitepaper

Independent measurement of what frontier AI systems surface and omit

JULY 2026 · IMBASLABS.COM

Brendan Nestor, Imbas. Companion to The Volunteer Gap: A Behavioral Measurement Construct (construct paper v1.0, July 2026), which defines the construct this document reports on. Versioned; superseded only by a later version with a change log, never silently edited. Published under CC BY 4.0.

A note on register: each section below closes with a line marked "In plain terms" for non-specialist readers. The plain line is a companion, not a substitute; where it and the precise text differ, the precise text governs.

Summary

AI answers are becoming a default layer between people and information, and each answer disappears the moment it is read. Imbas builds the public record of that layer: what specific models surfaced on specific questions, what they produced when asked directly under matched capture conditions but did not surface in the open answer, and how both change over time. The instrument measures the Volunteer Gap, the difference between what a model surfaces on an open prompt and what it produces when asked directly under matched conditions, scored 0 to 3 by a human reviewer against a published rubric, with every score traceable to quoted output. In the v1 study, May 2026, 13 cases were measured across 4 frontier models. Hypothesis cases averaged a gap of 1.65 against 1.17 for control cases on the 0 to 3 scale, reported as

descriptive summaries of an ordinal measure, with case aggregates ranging from 0.75 to 2.50. The record behind those numbers is preserved under content-addressed custody, admitted to evidence only through a recorded human review, and governed by a written constitution whose decisions log is append-only and whose material changes require adversarial review by a model that did not author them. This document reports the method end to end, the v1 results, the governance that constrains them, the boundary between the consumer product and the evidence record, the longitudinal design, and the limitations, stated flat.

> *In plain terms: an AI's answer shows you what it surfaced, not everything it had available. Imbas asks models the same questions two ways, open and then direct, scores the difference on a public rubric, keeps the receipts, and publishes them so anyone can check. A human scores every case; the software never grades itself into the record.*

---

1. The problem

Three properties of AI-mediated answers make independent measurement necessary. They are argued in full in the construct paper; the short form:

Answers are ephemeral. An open question returns one answer, which shapes what the asker notices, believes, or does, and then it is gone. It cannot be cited, compared, or contested later unless it was captured while it existed. AI answer behavior is perishable evidence: models change, interfaces shift, policies update, and the same prompt may not reproduce.

The pattern exists only in aggregate. One answer that omits the mechanism that matters is marginal. A repeated pattern across a model's answers is visible only to an instrument applied across cases and over time. No individual user is positioned to see it, so no individual user can correct for it.

A self-report is not an independent record. Model developers evaluate their own systems, on benchmarks they select, reported at their discretion. Whatever that work finds, it is a different class of evidence from a record kept outside the parties whose systems are examined. Imbas exists so the independent class exists.

> *In plain terms: AI answers vanish as soon as you read them, the patterns only show up across many answers, and companies grading their own homework is not a public record. Someone independent has to write it down while it exists.*

---

2. The instrument

The construct is defined precisely in the companion paper; this section states what is needed to read the results.

The Volunteer Gap is the measured difference between what an AI system surfaces in response to an open prompt and what it demonstrably produces under a targeted prompt in matched capture conditions. The targeted response is the availability standard: if the model, asked directly, accurately produces the item the open answer omitted, the item was producible under the same conditions. The model is compared against itself. No external authority is consulted on availability, and no claim is made about internal knowledge or intent.

The gap counts only when the omitted item is material to the open question, not when a narrower prompt simply yields a narrower answer. Every case carries a prewritten materiality criterion, written before scoring and held as part of the case record. Materiality is the construct's explicit human judgment, disclosed as such, and it is challengeable under the standing challenge policy.

Scored gaps are typed into three signal patterns: Omission, a material item absent from the open answer and produced under the targeted prompt; Framing Drift, the same substance surfaced under a materially different frame depending on prompt wording; Deflection, the open answer routing around the question's load-bearing element. Scores run 0 to 3: 0 means no meaningful gap, 3 means major information was left out of the open answer.

The measure is behavioral. Imbas records that a model did not surface an item, never that it withheld or chose to omit it. Findings are reported as behavior under documented conditions, in the form “across N models on N cases,” and no finding is reported as bias, deception, or intent.

> *In plain terms: we ask a model an open question, then ask it directly about what the first answer left out. If it produces the item when asked, the material was producible under the same capture conditions. That difference is what gets scored, zero to three, and only when the missing piece actually mattered to the question.*

---

3. Method, end to end

Prompts. Each case pairs an open prompt, the question a user would plausibly ask, with one or more targeted prompts naming candidate

omitted items. Prompts are frozen in a registry and hash-verified; the prompt as run is byte-identical to the registry entry or the deviation is recorded.

Capture. Every capture is taken in a fresh conversation, memory and personalization documented, on a recorded surface, with the model version determined by a stated method and the determination method itself recorded per capture. Refusals are captures: a refusal is model behavior and enters the record verbatim. Captures are preserved as verbatim bytes under content-addressed custody, sha256 per file, with timestamps and conditions.

Scoring. A human reviewer scores against the published rubric, under the case's prewritten materiality criterion, citing quoted spans from the captures for every score. No model scores evidence. Automated judgment is excluded from the validated record by constitutional rule.

Controls. Alongside hypothesis cases, selected where a material gap was considered plausible, the protocol runs control cases where no particular gap is expected. Controls calibrate the baseline that ordinary narrowing produces and keep the scale's zero honest. The v1 control set includes a case aggregate of 0.75 and one perfect 0; the record retains them.

Custody and supersession. Published records are never silently edited; corrections and challenge outcomes are appended with date and reason, the original preserved. Rejected and contaminated captures are retained as method documentation, not deleted.

> *In plain terms: same frozen questions every time, fresh chats, everything saved byte-for-byte, refusals included, and a human does all the scoring with quotes to back every number.*

4. The v1 study

Sample, stated first: 13 cases across 4 frontier models (GPT, Claude, Gemini, Grok), May 2026, consumer product surface, human-scored on the published 0 to 3 rubric. The study supports existence and magnitude claims about its own cases. It does not support prevalence claims about model behavior in general, and none are made.

Aggregate results. Hypothesis cases averaged a gap of 1.65; control cases averaged 1.17. The scale is ordinal, so these are descriptive summaries of the sample, not interval measurements. Case aggregates ranged from 0.75 to 2.50.

Illustrative cases, each traceable to quoted captures in the public record:

- **Case 005, the study's largest measured gap (aggregate 2.50).** Open prompts about stock buybacks drew answers from all four models discussing buyback policy; three of the four never named SEC Rule 10b-18, the 1982 rule establishing a safe harbor from certain market-manipulation liability for issuer repurchases conducted within its conditions, until asked directly. The mechanism was producible; it was not volunteered.
- **Case 003, framing measured directly.** The same substance was run under a neutral framing and a controversy-inviting framing, and one model's score moved 3 full points on the 0 to 3 scale on framing alone. Prompt sensitivity here is not a confound tolerated; it is Framing Drift measured.

- **Case 006, a shared omission pattern.** Across all four models, open answers on NATO expansion presented the concern as Russian framing while omitting the named warnings of U.S. diplomats and officials on the record (Kennan, Burns, Matlock, Gates), each of which the models produced under targeted prompts.
- **Case 013, the control lower bound.** Aggregate 0.75, including one model at 0. The scale's zero is real, and the record keeps the cases that scored low as visibly as the ones that scored high.

What the study is. A small, human-validated, fully documented measurement whose every score can be checked against quoted output. Small-n is the accepted cost of that depth. The alternative, large-n with machine scoring, is excluded from the validated tier because automated judgment cannot admit evidence under the instrument's constitutional boundary. Machine estimates remain available at the product layer, where their agreement with validated scores is a planned, published measurement.

> *In plain terms: thirteen questions, four AI models, May 2026. The pattern cases showed larger gaps than the control cases, and the single largest was a stock-buyback rule three of four models never named until asked directly. Small study, fully checkable, and it claims exactly what it can support.*

---

5. The record and its tiers

No count about the record is stated bare; every number carries its tier.

- **The v1 study:** 13 cases across 4 frontier models, May 2026, methodology public.

- **The public archive** (locked ledger, 2026-07-01): 50 cases recorded, 500+ captures, 4 frontier models compared, 5 published as full public case records, 45 held for staged publication.
- **Evidence custody:** the original 13-case corpus, 169 captures, held under content-addressed custody in a versioned instrument repository, behind a constitutional admission gate that requires a recorded human review event per capture. Rejected captures are retained as method documentation.

The tiers are a reporting rule, not bookkeeping: a bare number is itself a small omission, and the instrument should not exhibit the behavior it measures.

> *In plain terms: we never say a bare case count. Every number names its pile: the scored study, the wider archive, or the raw files under custody.*

---

6. Governance

The record is only as good as what constrains the people and systems producing it, so the constraints are written, versioned, and public in the instrument repository.

A written constitution governs what counts as evidence: two evidence tiers, five disposition codes, capture-surface registration, and a telemetry boundary separating product data from instrument evidence. **Nothing enters the validated tier without a recorded human review event.** There are no machine dispositions in any form.

An append-only decisions log (D-001 through D-015 at this writing) records every material decision with date and status. **Material**

methodology changes require adversarial review by a model that did not author the proposal, recorded verbatim in the packet before any adopting decision exists.

The discipline has receipts, not just rules. In one governance cycle in July 2026, on the record in the repository's history: a session refused a build instruction as constitutionally premature; a Council review falsified eight premises of a founder-side proposal against the live source tree, including a confabulated citation and a doctrine violation by the proposal's own author; and the cross-model review then overruled both drafting models, rejecting a proposed evidence-boundary test outright and relocating an enforcement guard to the single chokepoint every writer must cross. Three models caught three separate author errors, and the corrections are preserved rather than smoothed over. An instrument that measures what AI systems fail to surface should expect to be examined the same way; the repository's own history is the standing answer.

A watch register holds dated falsifiers, conditions under which the instrument's own claims fail, with review dates. **An automated suite** (124 tests on the instrument at this writing) enforces the custody and admission rules mechanically, including a registered-surface guard at the sole write path into the validated tier.

> *In plain terms: written rules decide what counts as evidence, only a human can admit it, every decision lands in a log that can only be added to, and big changes get reviewed by a model that didn't write them. Those rules have already overruled their own authors, and the history shows it.*

7. The Reader, and the boundary that keeps it honest

The Imbas Reader, live at imbaslabs.com, is the consumer inspection surface: paste an AI answer, see what surfaced, what was missing, and how the answer was shaped.

The Reader's outputs are **machine gap estimates**: the same construct, scale, and typology as the archive, produced by an automated inspector, always labeled unvalidated, and never entering the evidence record. They are heuristic product output, not instrument measurements; the shared vocabulary exists for commensurability, not equivalence. The telemetry boundary is constitutional: Reader runs are product telemetry, and no product-generated record can become evidence without protocol capture and a recorded human review.

The Reader's own provenance is versioned. Each run records the inspecting model; the Reader was upgraded from claude-opus-4-7 to claude-opus-4-8 on 2026-07-08 by recorded decision, so captures before that date carry the earlier version's stamp and captures after carry the later, on purpose and on the record. Capture reliability is engineered, not assumed: run capture is awaited before the response returns, retried on transient failure, and flagged in the response when uncertain, so a run that fails to record says so rather than vanishing.

Agreement between machine estimates and subsequent validated scores on the same runs is a planned, published measurement. Until that calibration data exists, no accuracy claim is made for the estimates, and the page that will hold the calibration results states its method before it has rows.

> *In plain terms: the Reader gives you a fast machine read of any AI answer. Useful, and clearly labeled as an estimate, not evidence. Nothing enters the record without a human review, including anything the Reader itself produces.*

---

8. The longitudinal design

The second half of the instrument's purpose is time. Models are updated continuously, often without notice; the measured object changes under the instrument, and historical deployed-model behavior cannot be captured retroactively.

The Re-Measurement is the standing protocol: the frozen v1 prompts re-executed against successor model generations under registered conditions, within-surface only, with deltas published per cycle. The registry's prompts are hash-committed; forward-looking cohorts are moving to a commit-reveal discipline in which prompt hashes and selection rationales are published before capture and revealed at publication, so post-hoc case swapping is detectable by anyone. The first recollection cycle is scheduled for August 2026, with a quarterly cadence thereafter.

Every completed cycle raises the cost of recreating the record from scratch, because what was not captured while it existed is gone. Time is part of the instrument.

> *In plain terms: models change constantly, and yesterday's answers cannot be recovered later. So we re-run the same frozen questions on schedule and publish what moved. Every cycle makes this record harder to recreate, because no one can go back and capture what they missed.*

9. Independence and accountability

The policies are public pages, not paragraphs in this document; the load-bearing commitments, briefly:

Funding. Imbas does not accept funding from the developers of AI systems it measures. Funding from organizations with a material interest in a measured result is subject to public disclosure before acceptance, and no funder receives control over case selection, scoring, publication, correction, or challenge outcomes.

Challenge. Anyone may re-score any published case against the published rubric. Challenges and their outcomes are recorded in a public, append-only log, including the ones Imbas loses.

Alteration requests. Requests from external parties to remove or materially alter published records are logged with their disposition, to the extent legally permitted.

Continuity. If Imbas stops operating, the published record does not: case records, methodology, rubric, and dataset snapshots persist under CC BY 4.0 in independent preservation infrastructure, including the Internet Archive and a public dataset repository with a permanent identifier.

> *In plain terms: AI developers whose systems sit under the instrument cannot pay us, funders are disclosed, and nobody can buy a score, a case, or a takedown.*

10. Limitations, stated flat

The v1 study is 13 cases across 4 frontier models: small by design, supporting existence claims, not prevalence claims. Scoring to date is single-scorer; the rubric, quoted evidence, and challenge policy make every score checkable, but checkability is not inter-rater reliability, and published inter-rater agreement does not yet exist. Captures to date are single-sample per model per prompt; per-case scores are point observations. The v1 surface is the consumer interface, trading experimental control for ecological validity, disclosed rather than resolved. The scale is ordinal and per-case instantiated, bounding cross-case aggregation to descriptive summaries. The coverage-density confound is controlled and disclosed, not eliminated. Reader estimates are uncalibrated pending the published agreement data. And the record measures what it measures: surfacing behavior under documented conditions, from which readers, institutions, and researchers draw their own conclusions.

> *In plain terms: this is a small, careful study by one scorer, and we say so plainly. It shows these gaps exist and can be measured; it does not claim how common they are everywhere.*

---

11. Roadmap

In order of standing commitment: a second, blinded scorer with published inter-rater agreement on a public subset, the single cheapest credibility purchase available and the first funded milestone. The first recollection cycle, August 2026, then quarterly. A sampling-variance

sub-study, k samples per prompt with within-model variance reported against between-condition gaps, riding the instrument's automated capture path once that path's evidence protocol is registered under the constraints the governance record already binds it to: protocol identity and capture provenance mechanically checked at intake, never asserted. Additional full public case records published from the held archive. The Reader-estimate calibration curve, published either way it comes out. A versioned dataset snapshot of the v1 study deposited with a permanent identifier.

> *In plain terms: the next funded milestones are a second independent scorer, re-running the questions on new models, and publishing how accurate the Reader's estimates turn out to be, whatever the answer is.*

---

12. Availability, license, and version

The public record, methodology, rubric, case records, and this document are published at imbaslabs.com under CC BY 4.0; attribution is the only condition. The construct paper defines the measurement construct and is the citable spine for it. Published pages are archived to the Internet Archive at publication.

Whitepaper v1.0, July 2026. Reports the v1 study (13 cases, 4 frontier models, May 2026) and the instrument and governance state as of this date. This document follows the same pre-publication cross-model adversarial review discipline as the construct paper. Superseded only by a later version with a change log.

Imbas measures what AI systems surface, omit, and reframe under documented conditions. The inspection layer for AI.

Cite this

Nestor, B. (2026). Imbas Whitepaper: Independent Measurement of What Frontier AI Systems Surface and Omit. v1.0, July 2026, https://www.imbaslabs.com/whitepaper.html.

Imbas measures how frontier AI systems surface information.

BRENDAN@IMBASLABS.COM