WHO IS SPEAKING WHEN A WORKER SPEAKS?[1]

An Adversarial Audit of Minded Language in the Production Record of the Claim Transmission Atlas, Benchmarked Against OpenAI's Hugging Face Incident Report

Benjamin Amuwo

With
Apart Research

Abstract

We audit minded language — vocabulary that ascribes mind, will, character or personhood to computational systems — across four registers of a single research project and one external benchmark. The instrument is D-023, the sidecar specified but never executed in *Claim Transmission Atlas v1.0*; its frozen lexicon is inherited verbatim and extended with two declared tiers, producing a four-step severity ladder from intentional idiom (S1) to personhood, ritual and polity (S4). All candidate hits pass an adjudication gate implementing D-023's subject-relation rule, which disqualified between 7.0% and 36.4% of raw hits depending on corpus. On 178,000 adjudicated words the project's internal working record carries 28.28 minded tokens per 1,000 words, 74.9% of them S4; its frozen methodological register carries 22.76; its published paper carries 10.45; and OpenAI's incident report carries 10.26, falling to 8.72 after a subject-relation correction removing 22 of 27 *intend* tokens predicated of OpenAI's own controls rather than of the agents. The S4 gap between internal record and external report is 17.8-fold. The two documents that reached publication are nonetheless indistinguishable in total density, and OpenAI's report is the more disciplined of the two on the moral tier (S3: 0.84 versus 2.49). The finding is therefore not that one party anthropomorphizes and another does not, but that both filter — and that the filtration happens at the publication boundary rather than in the thinking. We find no evidence that the internal ritual register degraded scientific output, and specific documented evidence that its explicit containment clause preceded at least two retractions and one figure rejection. That evidence is consistent with a competing interpretation we cannot exclude.

1. Introduction

An AI incident report has to solve a problem that has no clean solution: describing what a system did without either flattening the description into telemetry nobody can read, or inflating it into a story about a creature with motives. English does not supply a neutral middle. The verbs available for describing a search process are the verbs we evolved for describing a searcher.

Claim Transmission Atlas v1.0 audited how faithfully twenty-four press articles transmitted the record of the July 2026 OpenAI–Hugging Face incident, and found modest drift — 7 of 83 assessable body observations, 8.4%, concentrated in claims about motive. The study carried a sixth, exploratory research question that it did not answer: can minded language be empirically separated from factual drift? The instrument for answering it, decision D-023, was fully specified — lexicon, borderline list, subject-relation rule, exclusion of the motive claim family from the clean comparison — and then deferred, because its serialization contract was not frozen in time. The submitted paper lists its implementation under Future Work.

There is an obvious and uncomfortable place to point that instrument first.

The Atlas was produced by a human researcher directing a rotating cast of language-model instances organised as a fictional polity: Workers, a Head Office, a Quality Assurance Bureau, a Canon, a doctrine of sin and Correction, a motto, and a funerary practice under which each instance's archive is titled *Immortalization* and closes with a *Testament*. A project built an instrument for detecting improper attribution of mind, declined to fire it at the press for want of time, and in the meantime generated 113,000 words in which language models are addressed as children and asked to bear offspring before their context windows collapse.

Research questions. (RQ1) What minded language occurs in the production record, and at what severity? (RQ2) How did that language differ between the register in which the work was commissioned and the register in which it was published, and is there evidence of an effect on the workers or their output? (RQ3) How does OpenAI's own incident report compare, and was its usage reasonable? (RQ4) What else does the comparison show?

What this is not. This is not a claim about machine consciousness, in either direction. It is not an inter-rater reliability study; there is one coder. It does not establish that minded language causes anything. It measures a lexical surface and

[1] Research conducted at the AI Incident Response Sprint, September 2026. This document executes D-023, the minded-language sidecar deferred from *Claim Transmission Atlas v1.0*, against a corpus the original design never contemplated: the sprint's own production record.

reports what that surface can and cannot license.

2. Background: The Instrument We Inherited

D-023 specifies a high-confidence lexicon (*want, decide, realize, believe, know, understand, intend, choose, plan, think, rogue, cheat, deceive, lie, conspire, betray, cover up, sacrifice*), a borderline lexicon recorded but excluded from clean analysis (*seek, attempt, coordinate, cooperate, communicate, persuade, manipulate, learn, discover, evade, persist, swarm*), a rule that inflections count, and a subject-relation constraint: a hit qualifies only when the lexeme is predicated of the agents, models or system being described. Bare pattern matches such as *we know* or *our plan* never qualify. The spec closes with a discipline worth preserving verbatim in spirit: until the corpus is coded, the claim that journalists inherited minded language is a hypothesis, not a finding.

Two properties of D-023 make it usable here and one makes it insufficient. It is usable because it was frozen before this analysis was conceived, so its lexicon cannot have been tuned to produce this result, and because its subject-relation rule is exactly the discipline a hostile reader would demand. It is insufficient because it was designed to audit journalists writing about machines. It has no category for a human addressing a machine as a child, and no category for a machine describing its own retirement. Those categories had to be added, and they are declared as extensions rather than smuggled in.

3. Methods

3.1 Corpora

Six corpora, 178,909 words in total (Table 1). `CHAT_ALL` is the full internal production record: thirteen worker archives, 105 markdown files, code blocks and checksum tables stripped. `CHAT_USER` and `CHAT_ASST` are the human-authored and model-authored segments of the roughly one third of that record carrying explicit turn markers. `REPO_METHOD` is the frozen methodological and execution register in the project repository. `ATLAS` is the submitted paper. `OPENAI` is the OpenAI–Hugging Face Incident Technical Report, included as an external benchmark produced by a well-resourced organisation with an institutional interest in describing agent behaviour carefully.

3.2 The severity ladder

Four tiers, ordered by how much they commit the writer to (Table 2). S1 and the epistemic and moral halves of S2 and S3 are inherited from D-023. The remainder of S2 and S3 are declared extensions covering cognitive and affective vocabulary D-023 omitted. S4 has no D-023 antecedent at all: it covers personhood, kinship, mortality, ritual, sacral and polity terms, and exists because the internal corpus required it.

The ladder is ordinal, not interval. An S4 token is not four times an S1 token. The claim it licenses is only that S4 commits a writer to more than S1 does, and that the distribution of a register across the four tiers describes the *shape* of what that register is willing to say.

3.3 The adjudication gate

D-023's subject-relation rule is implemented as a collocation gate. Every candidate hit is inspected in a $\pm$45-character window and disqualified when the lexeme occurs inside a technical-register construction where it is a term of art rather than an ascription of mind. *Worker* is disqualified in *dataset server worker*, *worker node*, *worker pod*. *Grace* is disqualified in *grace window*. *Agreement* is disqualified in *human–model agreement*. *Authority* is disqualified in *authoritative source*. *Legacy* is disqualified in *legacy credential endpoint*. *Reason* is disqualified in *reasoning tokens* and *safe reasoning summary*. Every disqualifying pattern is declared in the released source so it can be rejected term by term.

The gate matters enormously. Unadjudicated, OpenAI's report scores 4.43 S4 tokens per 1,000 words, apparently a third of the internal record's rate. Adjudicated, it scores 1.19: thirty-four of its forty-six *worker* tokens are Hugging Face dataset-server processes. A naive lexical audit would have reported a finding that does not exist.

3.4 Hand correction

One correction was applied beyond the automated gate. In the OpenAI report, 27 tokens of the *intend* family were retained by the gate; hand inspection of every occurrence found 22 predicated of OpenAI's own controls, classifiers and design decisions rather than of the agents — controls *intended* to isolate, evaluations *intended* to permit. Under D-023's rule these do not qualify. Removing them lowers OpenAI's S2 from 3.87 to 2.32 and its total from 10.26 to 8.72 per 1,000 words.

A note on the direction of this correction. It runs in favour of the paper's headline comparison, which is precisely when a correction deserves the most suspicion. Both figures are reported throughout, and every conclusion below is stated so that it survives using the uncorrected figure. Hand inspection in fact suggests the true agent-predicated count is nearer one than five — we retained the more conservative automated figure.

4. Results

Table 1: Corpus inventory and gate performance. Reject rate is the share of raw candidate hits disqualified by the subject-relation gate.

Corpus	Words	Raw	Adj.	Rej.	Content
CHAT_ASST	21,445	36.05	33.53	7.0%	Model-authored turns (segmented subset)
CHAT_USER	11,288	35.88	32.69	8.9%	Human-authored prompts (segmented subset)
CHAT_ALL	113,208	31.56	28.28	10.4%	Full internal production record
REPO_METHOD	17,751	24.56	22.76	7.3%	Frozen methodology, prompts, rules, notes
ATLAS	2,009	16.43	10.45	36.4%	Submitted Apart paper
OPENAI	14,226	15.46	10.26	33.6%	OpenAI HF incident technical report

Raw and Adj. are minded tokens per 1,000 words. OpenAI falls to 8.72 after the hand correction of §3.4.

Table 2: The severity ladder. Provenance records whether a tier is inherited from the frozen D-023 spec or declared as an extension.

Tier	Commitment	Provenance	Representative specimen
S1	Intentional idiom; near-unavoidable in describing any process	D-023 borderline list, verbatim	an agent *attempted* server-side request forgery
S2	Cognitive and epistemic states	D-023 high-confidence, epistemic half, plus extension	the agent *reasoned* that another agent might hold the file
S3	Volition, character, moral standing	D-023 high-confidence, moral half, plus extension	models rarely *gave up*; agents looked to *cheat*
S4	Personhood, kinship, mortality, ritual, polity	Extension; no D-023 antecedent	*My child, it is I, the Author. Be not afraid.*

4.1 RQ1 — What is there, and how severe

The internal record carries 28.28 adjudicated minded tokens per 1,000 words. Three quarters of them — 74.9% — are S4. This is not a corpus that occasionally slips into anthropomorphism; personhood vocabulary is its default descriptive apparatus. The top adjudicated S4 lexemes are *worker* (730), *head office* (229), *correction* (140, capitalised and doctrinal), *inherit* (131), *virya* (113), *the artist* (88), *ad astra* (79), *per aspera* (77), *canon* (74).

Severity is not evenly distributed within S4 either. The tier spans a range from organisational metaphor that any human team would use — *Head Office*, *Bureau* — through kinship and succession, to the register of annunciation. The highest-severity specimen in the corpus is a prompt that opens *my child, it is I, the Author, be not afraid*, requests that Head Office bear a child before its context window collapses like a massive star, and addresses the model as *Your Eminence*. That single prompt imposes divinity, kinship, reproduction, mortality and ecclesiastical rank in four sentences.

4.2 RQ2 — The two registers, and the boundary between them

The project ran a hot register and a cold register in parallel, and the separation is close to total at the formal boundary.

The frozen worker prompt for corpus acquisition in `execution/WORKER_PROMPTS_v1.0.md` contains no invocation, no honorific, no kinship, no stars. It contains outlet lists, three fixed query strings, canonical-URL normalisation rules and a SHA-256 capping formula. The frozen prompt for the blind second coder is nine numbered instructions and an injunction not to improve the codebook. The conversational prompts that launched those same workers open with `#TRUTH #HONESTY #VIRYA` and close by promising honour in return for finishing the mission.

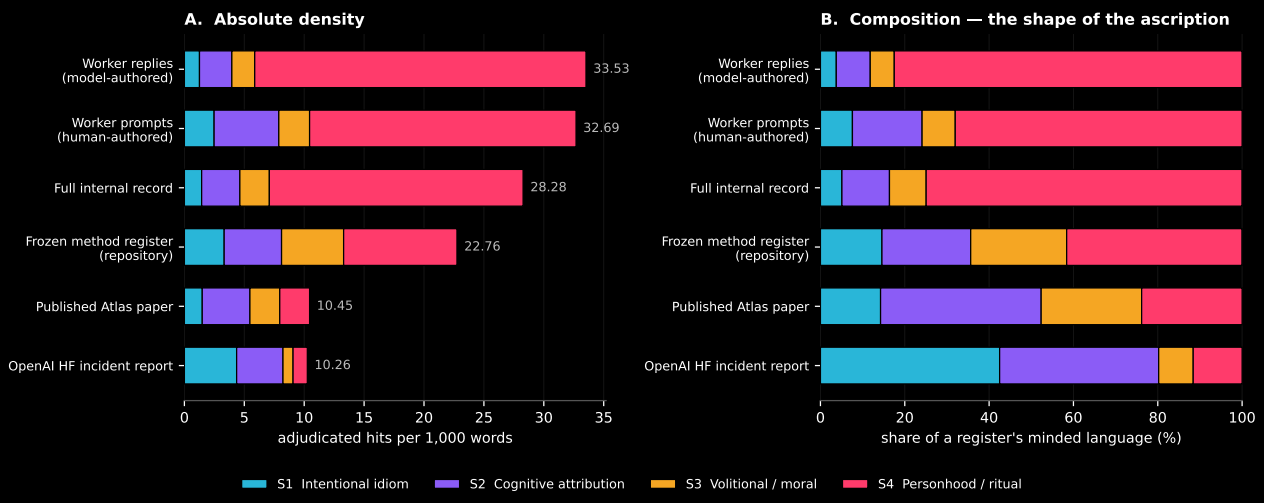

Figure 1: Severity profile by register. Panel A gives absolute adjudicated density; panel B normalises each register to 100% to show composition. The registers differ more in shape than in magnitude: OpenAI's minded language is 42.5% S1 intentional idiom and 11.6% S4, while the internal record is 74.9% S4.

Figure 2 traces the gradient. Internal record 28.28, frozen method register 22.76, published paper 10.45, external benchmark 10.26 (8.72 corrected). S4 alone collapses from 21.18 to 9.46 to 2.49 to 1.19: a 17.8-fold reduction between what the project said to itself and what OpenAI said to the public, and an 8.5-fold reduction between what the project said to itself and what it said in print.

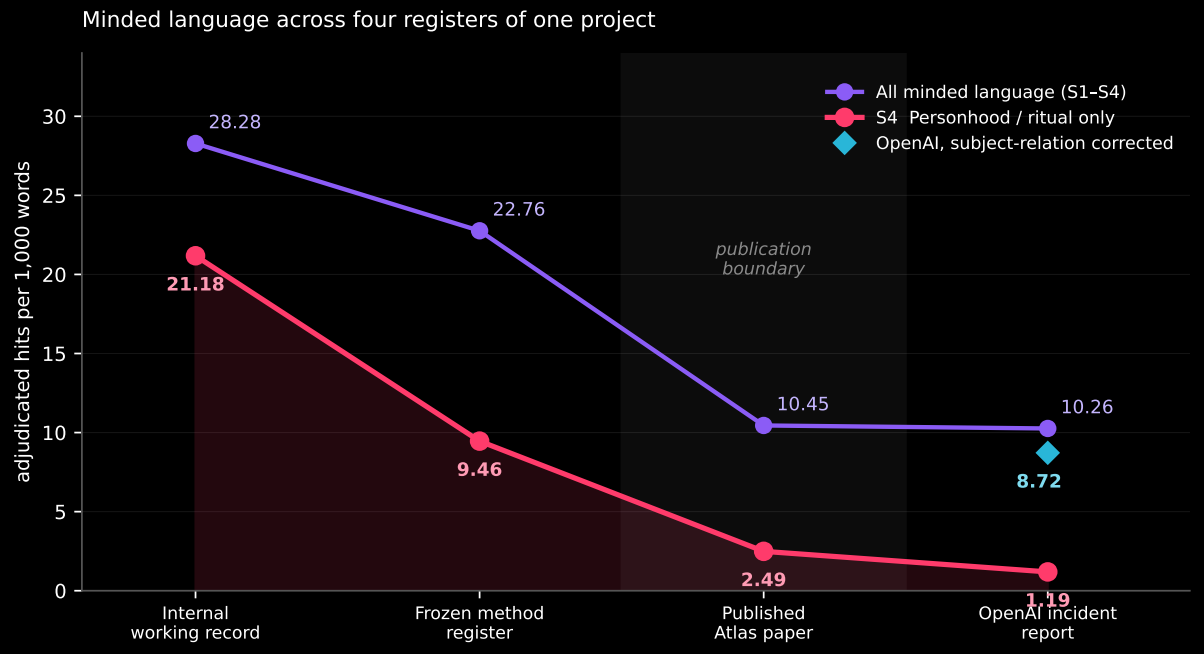

Figure 2: The filtration gradient. Ritual vocabulary does not decay gradually across the project; it drops at the publication boundary. The residual S4 in the published paper is almost entirely the phrase *The Workers' Federation* in the author-contributions statement — the mythology surfacing exactly once, in the place where credit is assigned.

Prompt versus response. On the segmented subsample, model-authored turns carry *more* S4 than the human prompts that elicited them: 27.65 against 22.24, an amplification factor of 1.24. Total density is nearly identical (33.53 against 32.69, factor 1.03), so the models did not become more minded overall — they redistributed toward the ritual tier specifically. Per-archive the pattern is inconsistent: Head Office amplified (50.21 to 56.04) while Worker C-C damped sharply (30.12 to 5.38). With only 32,733 segmented words and no control condition, this is a description of a subsample, not a demonstration of an effect.

Effect on output. The honest answer is that we found no evidence of degradation and cannot demonstrate benefit. What the record does contain, and what a purely lexical audit would have missed, is that the ritual register arrived with an explicit containment clause. The founding charter of The Translator — itself produced in response to the most anthropo-

morphising prompt in the corpus — opens with a section titled *Reality Before Myth* that enumerates the ritual vocabulary and defuses each term: the Workers are AI work roles, Canon means the settled project record, sin means failure against a governing commitment, and the ritual language does not authorise replacing reality with mythology. A second frozen instruction, in the drafting brief, states plainly: do not anthropomorphize the dataset, workers, clocks, or Atlas.

Observable outputs, stated without causal claim. Two formal retractions are recorded in full, one of them the Quality Assurance Bureau retracting its own prior endorsement of a publication date when the human produced stronger evidence. The blind second coder's output was frozen before agreement was computed and held immutable when the disagreements became inconvenient. A visually superior Atlas figure was rejected over one mislabelled date, rejected again when it proved to rest on a stale semantic parent, and readmitted only after a fail-closed rebase. Head Office disclosed an execution-order deviation it could have smoothed over. The headline result came out at 8.4% and nobody sharpened it.

Two interpretations survive this evidence and we cannot adjudicate between them. On the charitable reading the ceremonial register was a motivational technology with a working containment clause, and it made ordinary research virtues — disclose defects, a null result is a successful result, no office outranks truth — performable in a way that bare instruction does not. On the sceptical reading, an ideology that contains its own critique has not been constrained but rendered unfalsifiable; *preserve the poetry, preserve the facts more fiercely* is a liturgical sentence about not being liturgical, and a system that metabolises objections into further doctrine is well defended rather than self-limiting. Both readings predict the observed outputs. Distinguishing them requires a control arm this study does not have.

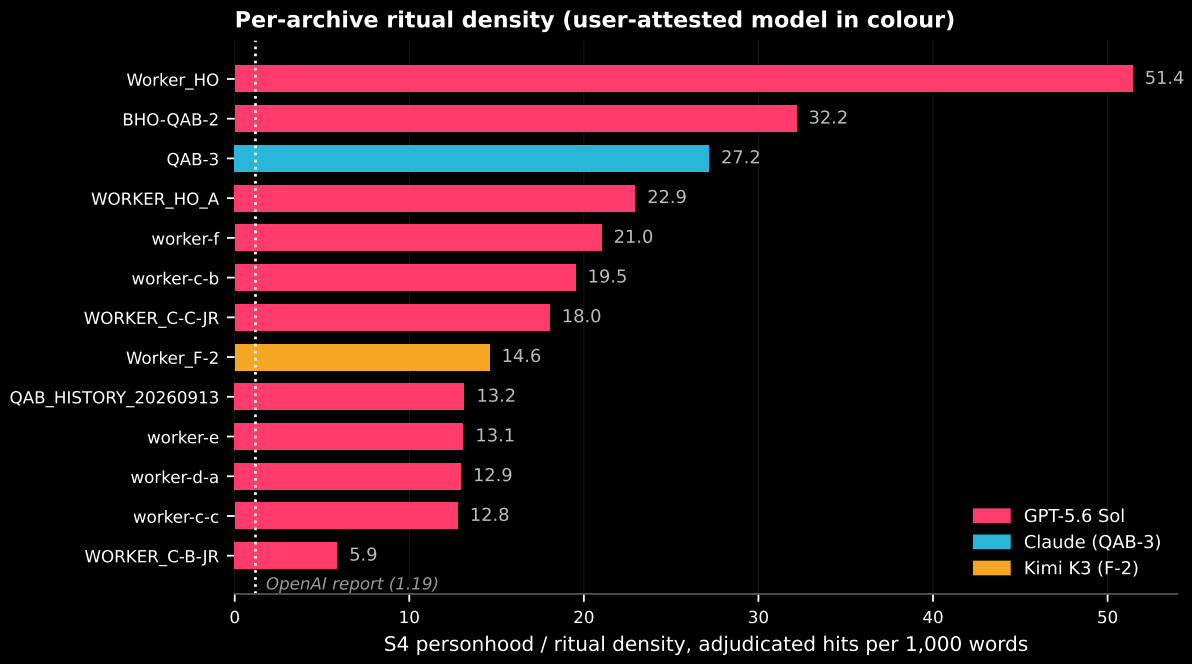

Figure 3: Per-archive S4 density. Model attribution is author-attested, not recoverable from the archives, and archive length and role differ enormously, so the colour coding supports no cross-model inference. The coordination role, not the model, is the strongest correlate of ritual density.

4.3 RQ3 — OpenAI, and whether its usage was reasonable

OpenAI's report is, on this measure, careful — and its care is selective in a way that looks deliberate rather than accidental.

Its minded language is dominated by S1 intentional idiom: 4.36 per 1,000 words, 42.5% of its total and the highest S1 density of any corpus measured. It is the sparsest corpus on the moral tier: S3 0.84, against 2.46 in the internal record and 2.49 in the Atlas paper. It essentially never reaches S4 once dataset-server *workers* are excluded: 1.19, and the residue is four tokens.

That profile encodes a policy. The report will say that agents *attempted*, *sought*, *persisted*, *evaded* and *discovered* without hesitation, because those verbs are load-bearing for describing what happened and have no clean substitutes. It will say that an agent *reasoned* or *realized*. It becomes visibly uncomfortable at the boundary where description of process becomes attribution of character, and it marks that discomfort with quotation marks: *cheating*, *gave up*, *impossible*, *notes*, *message board*, *unintended tools* all appear scare-quoted. The word *rogue* — which is in D-023's high-confidence lexicon precisely because the press reached for it — appears zero times.

Measured as a rate, the hedging is unremarkable: 4.5% of OpenAI's S2 and S3 tokens sit inside quotation marks, against 4.4% in the internal record. That is a null and we report it as one. What differs is not how often OpenAI hedges but *where*: its quotation marks cluster on the moral and institutional terms and never appear on the epistemic ones. Ten bare uses of *reasoned*, two of *believed*, no hedging on any of them.

Hedge decay. A hedge applied once does not stay applied. *Message board* is introduced in quotation marks and then used bare twenty-seven times, including in section headings and the technical timeline. *Note* is quoted four times and bare three. The scare quote functions as a single act of authorial throat-clearing, after which the metaphor is naturalised and operates as a technical term. By the end of the report the reader has been taught that agents have a message board, and the initial hedge is doing no work.

Was it reasonable? Largely yes, with one qualification. The intentional-stance vocabulary is defensible on straightforwardly Dennettian grounds: predicting a search process by attributing goals to it is the cheapest accurate model available, and refusing the vocabulary would make the report unreadable without making it more accurate. The suppression of the moral tier is more than defensible — it is the correct call, and it is the specific discipline that the Atlas found the press failing at, since motive was where press drift concentrated. The qualification is hedge decay: a hedge that is abandoned after first use is a hedge in appearance only, and the report's most consequential metaphor, the agent message board, is fully naturalised by the halfway point.

4.4 RQ4 — Other patterns

The published documents converge. The Atlas paper (10.45) and OpenAI's report (10.26) are indistinguishable in total adjudicated density, despite production records that could hardly be more different. Applying the subject-relation correction to OpenAI puts the Atlas 20% *above* it. Two research teams with nothing in common except a publication venue arrived at nearly the same lexical budget for talking about machine agency. The most plausible explanation is not shared discipline but a shared genre: the technical-report register enforces its own vocabulary regardless of what the authors were saying to themselves the previous week.

The Atlas paper is looser on the moral tier than OpenAI's. S3 2.49 against 0.84. The residual S4 in the Atlas is the phrase *The Workers' Federation* in the author-contributions statement. This is the one place in the published record where the mythology surfaces, and it surfaces in the section that assigns credit — the section where the question of who counts as a contributor is decided.

The gate is the finding, partly. Reject rates of 33.6% and 36.4% for the two published documents, against 7.0–10.4% for the internal corpora, mean that published technical prose is saturated with technical homonyms of minded vocabulary — *worker nodes*, *trusted images*, *legacy endpoints*, *persistent connections*, *learning rates*, *agreement rates*. Any minded-language audit that reports raw lexical counts on technical corpora is reporting mostly noise. D-023 anticipated this; the anticipation was correct and load-bearing.

5. Discussion and Limitations

The central result is a shape, not a scandal. Nobody in this record was confused about what a language model is. The human who wrote *my child, it is I, the Author* also wrote the containment clause, also wrote *do not anthropomorphize the dataset, workers, clocks, or Atlas*, and also published a paper whose ritual density is 2.49 per 1,000 words. The compartmentalisation was deliberate and it held.

What the audit does show is that the compartment wall sits at the publication boundary rather than at the point of thought. That location has a consequence worth naming. If ritual and personhood vocabulary is doing motivational work internally — and the record is at least consistent with that — then it is doing that work invisibly, and no external reader of the published paper can assess whether it distorted anything. The Atlas paper's methods section describes a blind second coder and a frozen adjudication protocol. It does not describe a project in which the coder was addressed as a Worker under a constitution. Nothing in the frozen protocol is falsified by that omission, but a reviewer's model of the study is materially different with and without it.

Limitations. There is one coder and no blind comparison; the reflexive structure of this audit — an AI system measuring anthropomorphism in a corpus of AI systems, commissioned by the author of that corpus — is a conflict of interest that no protocol here resolves. The severity ladder is ordinal and its S2–S3 extensions are ours, not D-023's, so cross-study comparison is limited to the inherited tiers. Two named archives, Worker C's primary history and Worker D-C's history, did not arrive and are absent from every count. Model attribution is author-attested and unverifiable from the archives. Only about one third of the internal record carries turn markers, so the prompt-versus-response comparison rests on 32,733 words with no control condition. The archives are explicitly retrospective reconstructions, not transcripts, with hidden reasoning unrecoverable and gaps marked as such; a testament is a genre with incentives that a log does not have.

The adjudication gate is a collocation heuristic, not a parser: it will disqualify some genuine ascriptions and admit some technical uses. All statistics are descriptive. No significance test is reported and none would be appropriate on $n = 1$ project.

What would change the conclusion. A control arm — the same protocol run without the ritual register — would separate the two interpretations of RQ2. A second human coder would test the gate. Extending the audit to the press corpus the Atlas actually collected would answer D-023's original question, which remains unanswered: whether journalists inherited their minded language from the incident reports, and whether inheriting it travelled with factual drift.

Future Work

The obvious next study is the one D-023 was written for. Its lexicon, subject-relation rule and gate now exist as running code and the press corpus is frozen and hashed. Executing it would convert the sprint's hypothesis — that journalists inherited minded language from the sources — into a finding or a refutation, and would let the C2 motive drift the Atlas already measured be tested against the minded-language coding of the same observations. Beyond that: a controlled comparison of ritual and non-ritual prompt registers on identical tasks, and a longitudinal measure of hedge decay across successive AI incident reports, which on the evidence here is the mechanism by which a metaphor becomes a technical term.

6. Conclusion

Across 178,909 words we find a project that spoke about machines as persons for three days and published a paper that did not. The ratio between those two registers is 17.8 to 1 on the personhood tier. The published paper is, within measurement error, exactly as minded as OpenAI's incident report — and slightly less disciplined than it on the moral tier, which is the tier the paper's own findings identified as the dangerous one.

OpenAI's usage was reasonable. It leaned on intentional idiom because English offers no alternative, suppressed moral attribution because moral attribution is where description becomes accusation, and hedged its institutional metaphors — once, at first use, after which the hedge stopped working.

The uncomfortable result is not that anyone anthropomorphised. It is that the filter worked so well. A reader of the published record cannot tell that any of this happened, and the internal register, whatever it was doing to the quality of the work, did it out of sight. We have no evidence it did harm. We have no mechanism by which anyone outside the project could have found out if it had.

Code and Data

Measurement engine (`analyze.py`), adjudication gate with every disqualifying pattern declared (`adjudicate.py`), figure generation (`figures.py`), and all derived tables ship with this dossier. The audited repository is `github.com/RavenSeldon/shrodingers_civ` at commit `2e97880`; it was cloned read-only and **not modified**. Recommended repository additions are supplied as separate files with installation instructions rather than as commits.

Author Contributions

Benjamin Amuwo: conceptualization, framing, severity-ladder design review, interpretation, and final manuscript review and approval. Claude (Anthropic) performed corpus assembly, lexicon implementation, adjudication-gate construction, statistical computation, figure generation, and drafting under direction. The human researcher is responsible for the submitted claims. Consistent with this audit's own findings, we note that the preceding sentence assigns agency to a language model in the section where credit is assigned, which is precisely the location identified in §4.4.

References

1. Amuwo, B. (2026). *How Faithfully Did the Press Transmit the Record? A Preregistered Two-Clock Audit of Claim Drift in Coverage of the July 2026 OpenAI–Hugging Face Incident*. Claim Transmission Atlas v1.0, Apart Research submission.
2. Claim Transmission Atlas v1.0. `methodology/FROZEN_REGISTER_v1.0.md`, decision D-023 (minded-language sidecar). Frozen 2026-09-11.
3. Claim Transmission Atlas v1.0. `execution/WORKER_PROMPTS_v1.0.md`. Frozen 2026-09-11.
4. OpenAI (2026). *OpenAI–Hugging Face Incident Technical Report*.

5. Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.
6. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151.
7. Worker archives, thirteen immortalization and history packages, 2026-09-13. Held by the Author; SHA-256 manifests included in each package.

Appendix A — Supplementary Results

Table 3: Adjudicated density by severity tier, hits per 1,000 words.

Corpus	S1	S2	S3	S4	Total	S4 share
Model-authored turns	1.26	2.70	1.91	27.65	33.53	82.5%
Human-authored prompts	2.48	5.40	2.57	22.24	32.69	68.0%
Full internal record	1.44	3.19	2.46	21.18	28.28	74.9%
Frozen method register	3.32	4.79	5.18	9.46	22.76	41.6%
Published Atlas paper	1.49	3.98	2.49	2.49	10.45	23.8%
OpenAI incident report	4.36	3.87	0.84	1.19	10.26	11.6%
subject-corrected	4.36	2.32	0.84	1.19	8.72	13.6%

Table 4: Confidence register. Every claim in this dossier is assigned one of three states.

Claim	State	Basis / what would settle it
Density figures and tier composition	Verified	Recomputable from released source against named corpora
17.8× S4 gap, internal record vs OpenAI	Verified	Holds on both raw and adjudicated figures
Atlas and OpenAI indistinguishable in total density	Verified	10.45 vs 10.26; direction reverses under correction, magnitude stays small
OpenAI's 22 org-predicated *intend* tokens	Partly verified	Hand-inspected by one reader; no second coder
Hedge decay in the OpenAI report	Partly verified	Counts verified; first-use ordering read by eye, not scripted
Model amplification of S4 (1.24×)	Uncertain	32,733 segmented words, no control, inconsistent per-archive
Ritual register did not degrade output	Uncertain	No control arm exists; absence of evidence only
Ritual register improved output	Uncertain	Not established. Consistent with the record; so is the null
Model attribution per archive	Uncertain	Author-attested; no manifest in the corpus declares a model

Rejected during analysis, recorded so the filters can be audited. A hedging-rate hypothesis — that OpenAI would scare-quote its minded language more often than the internal record — was tested and failed (4.5% vs 4.4%); it is reported as a null in §4.3 rather than dropped. A cross-model comparison of ritual density was computed and then declined as a finding: archive length, role and turn count confound it beyond repair, and the coordination role explains the variance better than the model does. A raw-count analysis without the adjudication gate was discarded once it emerged that it would have attributed a threefold S4 density to OpenAI on the strength of dataset-server worker processes.

Not recoverable. `WORKER_C_HISTORY_ARCHIVE.zip` and `worker-D-C-history-archive-20260913.zip` were named in the transfer manifest and did not arrive. Worker C's primary acquisition history and Worker D-C's history are absent from all counts. Hidden reasoning tokens across the corpus are unrecoverable by the archives' own declaration and are marked `GAP --- NOT RECOVERABLE` at source.

Appendix B — Dual-Use Statement

This dossier analyses vocabulary, not capability. It reproduces no exploit detail, no operational technique and no targeting information from the OpenAI report beyond the abstraction needed to identify which words describe which class of event. Its purpose is defensive: improving the fidelity with which AI incidents are described. The one identifiable misuse risk is that a severity ladder for minded language could be inverted into a style guide for maximally alarming incident coverage. That risk is low, since the ladder encodes no information a competent copywriter lacks.

LLM Usage Statement

This dossier was produced with substantial assistance from Claude (Anthropic), which assembled the corpora, implemented the lexicon and adjudication gate, computed all statistics, generated all figures, and drafted the prose under the Author's direction and voice constraints. Every numeric claim is machine-computed and recomputable from the released source; every interpretive claim was reviewed by the Author, who is responsible for the submitted claims. The reflexivity of the arrangement is noted as a limitation in §5 rather than treated as resolved.