Episteme (2026), 1–22
doi:10.1017/epi.2026.10160

episteme

ARTICLE

Assertion without a Speaker: Testimony, Tracking, and Large Language Models

Abhinav Saxena

Independent Scholar
Email: abhinav.philosophy@gmail.com

(Received 4 May 2026 UTC; revised 1 August 2026 UTC; accepted 24 September 2026 UTC)

Abstract

Large language models produce assertion-shaped outputs that often bring their users to true belief. I argue for two conclusions of different strengths: that on the best current evidence, these systems do not transmit testimonial knowledge and that hearers are not now positioned to accord their outputs testimonial uptake, which does not wait on the first. From Burge, Williamson, Goldberg, Faulkner, and Lackey, I derive a Tracking Minimum: a source transmits testimonial knowledge that p only if the truth of p, or the evidential support for p, figures in the explanation of its producing that assertion, and does so in virtue of the source's own production process rather than of a proxy on which that process depends. It is this endogeneity requirement, not any modal profile, that the five accounts share and current systems fail. The argument engages Burge's extension of his framework to computational sources, the interpretability literature, and the published case for AI testimony, whose dispute with mine turns on an empirically decidable question, for which an experiment is specified. The positive proposal places LLM outputs within an extended epistemology of instruments, with a result that has gone unremarked: a working barometer satisfies the condition these systems fail.

Keywords: Testimony; large language models; AI testimony; tracking minimum; epistemology of instruments

1. Introduction

A patient asks a chatbot what dose of acetaminophen is safe for her ten-year-old. The chatbot answers correctly. She gives the dose. The child is fine. She came to a true belief, acted well on it, and relied on a source that was, in the relevant sense, accurate. Did she learn from the chatbot in the way she would have learned from her paediatrician? If the question is whether she came to *know* the safe dose, and to know it as one knows what a competent informant has told one, the answer is less obvious than the case looks.

This paper argues that she did not, in that sense, learn from the chatbot. The argument is not that LLMs are insufficiently reliable; in many domains, they are highly reliable, and successive systems have become more so. It is that the reliability they

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

exhibit is not the kind of truth-connection testimony requires. Read carefully, the leading accounts (Burge's entitlement view, Williamson's knowledge-norm view, Goldberg's anti-individualism, Faulkner's assurance view, Lackey's statement view) converge on a minimal constraint I call the Tracking Minimum (TM): the truth of what is asserted, or the evidential support for it, must figure in the explanation of the source's producing the assertion, and figure there through the source's own operation.

Two conclusions follow, and it matters that they differ in strength. The *constitutive* conclusion, that current systems do not satisfy TM, is an empirical claim about what governs their production, asserted with the confidence the evidence supports and no more. The *practical* conclusion, that hearers are not now in a position to accord LLM outputs testimonial uptake, does not wait on the first, and is robust under the uncertainty the first acknowledges. Much of the recent literature runs these together; keeping them apart is part of the contribution.

TM is fixed at a determinate strength, stronger than a merely statistical condition and weaker than an internalist one, and applied on the accounts' own terms, including the one text in which Burge extends his framework to computational sources (1998), and against the published cases for the opposite conclusion (Faria 2025; Freiman 2024b).

Two results should be flagged, since neither is what the title leads a reader to expect. TM is not the boundary between testimony and instrument reading, since a working barometer satisfies it; what follows about LLMs is therefore sharper than a reclassification, in that they fail a condition ordinary instruments meet by construction. And the disagreement with the most serious opposing view is not a clash of intuitions but a question about what a system's arbitration between conflicting inputs is responsive to, which Section 8.2 reduces to a specifiable experiment.

A note on scope. The constitutive conclusion is conditional on the architecture and training of currently deployed systems and on the state of evidence about them. If future systems, or future interpretability results about present ones, show that assertoric production is governed by truth-tracking structure, the question reopens, and TM specifies what would have to be shown. The thesis is not that no machine could testify, but that these machines, as they are now built and as we now understand them, do not.

2. The tracking minimum

The modern literature on testimony divides over what makes testimonial transmission possible. The disagreements are real, but they conceal a structural agreement about what the speaker side minimally requires. Since everything turns on stating that agreement at exactly the right strength, this section proceeds slowly.

2.1. Five accounts

Burge (1993) argues that hearers have an a priori entitlement to accept what they are told, absent countervailing reasons, because intelligible presentation-as-true is a prima facie sign of a rational source. Two features are easily misread. The entitlement is not keyed to the proximate medium: recordings, print, and machines can carry it, since what it licenses is the presumption of a rational source *somewhere in the chain* (1998: 17). And entitlement to accept is not yet transmission of knowledge: interlocution yields knowledge only if the *extended body* of warrant, which includes the resources possessed or indicated by the sources on which the recipient depends, contains enough to

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

underwrite it (1998: 5–6). The load Burge places on the source is one of provenance and preservation.

Williamson (2000: ch. 11) holds that assertion is constitutively governed by the knowledge norm: one must assert that *p* only if one knows that *p*. Whatever epistemic upgrade the hearer enjoys runs through the assertion's status as an act regulated by that norm, so a producer whose assertions are not regulated by the aim of asserting only what one knows does not generate the same uptake. Goldberg (2010) holds that the process generating a testimonial belief includes the speaker's own production, so the reliability relevant to the hearer's warrant is in part the speaker's; he sharpens this by contrasting reliance on an epistemic subject, assessable for rationality and responsibility, with reliance on a mechanism assessable for reliability alone, where the normative burden falls on the user (2012: 184, 188). Faulkner (2011) grounds testimonial warrant in assurance: the speaker invites affective trust, which presses a normative claim on its object, rather than the predictive trust one may place in any regular process, and thereby becomes answerable for the truth of what is said. A source so answerable is one for which the truth of what is asserted figures in why the assertion is made.

Lackey (1999, 2008) has done most to weaken the speaker-side conditions. On her statement view, knowledge can be acquired from statements the speaker does not believe; the selfless-asserter cases, of which Stella the creationist biology teacher is the most famous, are designed to defeat the belief condition, and Section 4 grants that they do. But what Lackey puts in place of belief is itself a truth-connection: the speaker's statement that *p* must be a reliable *indication* of *p*'s truth, and she must not possess undefeated defeaters. Reliable indication is a covariation notion. Even the most permissive account retains a substantive constraint between production and truth.

2.2. Fixing the strength

An argument of this shape is vulnerable to a charge of equivocation: a constraint weak enough to be common ground may be too weak to exclude anything, and a constraint strong enough to exclude may be nobody's minimum. The remedy is to state the constraint once, at the strength intended, and to use it at that strength throughout.

Begin with a familiar case. A thermometer sits in a room held at constant temperature by a thermostat. It is broken; it would read the same whatever the temperature. Its readings are correct, and correct for a systematic, stable, explicable reason, yet no one thinks that posting them transmits knowledge of the temperature. The case fixes the target: the shared constraint must be strong enough to exclude the broken thermometer and weak enough to be genuinely common ground. Three candidate readings must be distinguished.

On a *statistical* reading, the required relation is non-accidental accuracy: a high rate of true assertions over the relevant range, stable, grounded, and explicable rather than a fluke. This is too weak. The broken thermometer's accuracy is all of those things; what it lacks is any connection between each reading and the temperature that reading reports. If the source's asserting *p* bears no explanatory relation to *p*'s truth, the hearer's belief is true by a kind of luck that precludes knowledge, however impressive the aggregate record. That record can enter the hearer's epistemic position, but it enters as inductive evidence about the source, and warrant so structured is the warrant of instrument use (Section 5). Grounded frequency reliability is non-accidental relative to the mechanism producing it while remaining accidental relative to the truth of the particular content asserted, and it is the second relation transmission requires.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

On an *internalist* reading, the required relation is production governed by internal states that represent considerations *as* evidence, from the producer's own perspective. This is too strong, and not common ground. Lackey's cases show belief is not required; Williamson's (2006) externalism about understanding tells against building a first-personal grasp condition into the practice of assertion; and function-first accounts of assertoric norms (Kelp 2018) detach the norm's content from the speaker's psychology altogether. A TM that quietly reinstated an internalist grasp condition would beg the question against these positions and decide the machine case by stipulation.

The reading I intend, and the one I shall argue the five accounts jointly entail, is *explanatory*:

> **The Tracking Minimum (TM).** A source *S* can transmit testimonial knowledge that *p* to a hearer only if the truth of *p*, or the evidential support for *p*, figures in the explanation of *S*'s producing the assertion that *p*, and figures there in virtue of the operation of *S*'s own production process rather than in virtue of a proxy on which that process depends. The condition can be satisfied *originally*, when the production process itself secures the connection, whether through belief based on adequate grounds, through evidence-responsive production, or through production regulated by norms whose function is to constrain assertion to truth; or *derivatively*, when *S*'s utterance preserves, without relevant alteration, a particular assertion whose source satisfies the condition originally.

2.2.1. Five comments

The first concerns the clause about proxies, which I shall call the *endogeneity* requirement and which carries most of the weight below. Without it, the statistical reading returns by the back door, since explanation is promiscuous: evidential support explains what gets attested in a corpus, attestation explains what a corpus-trained process emits, so evidential support figures at some remove in the explanation of the output. Endogeneity requires that the connection hold through the source's own operation. Where a process is connected to the truth only because it is connected to a proxy that is itself so connected, the proxy screens the truth off: hold the proxy fixed and vary the truth, and the output does not move. The broken thermometer fails a step earlier, since nothing about the temperature explains what it displays; endogeneity is needed for the harder case, where an explanatory connection obtains but runs through a proxy. That is the corpus-trained process's situation, and not Stella's (Section 4.2).

Second, TM excludes the thermometer without any commitment on which anti-luck condition is correct, and that is worth emphasis, since the natural way to write a constraint of this kind is in terms of sensitivity, and sensitivity is contested. In the thermostatted room, the thermometer is perfectly *safe*: it would not easily read falsely, since the temperature would not easily change. A safety condition therefore fails to exclude it, and a theorist who prefers safety to sensitivity (Pritchard 2005; cf. Williamson 2000: ch. 5) might take that failure as a reason to reject the constraint altogether. But the thermometer is excluded here on other grounds: nothing about the temperature figures in explaining what the device displays. TM is accordingly neutral between the leading anti-luck conditions. Since it constrains the source's production rather than the hearer's belief, the standard difficulties for modal conditions do not arise for it in any case: a

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

source can fail to assert a necessary truth in nearby alternatives in which its evidence fails to support it, which is why TM's antecedent is disjunctive.

Third, the test is interventionist rather than psychological: hold the worldly evidence fixed and vary the proxies the process consults, then hold the proxies fixed and vary the evidence. A process satisfying TM originally follows the evidence within the reach of its competence; one that merely exploits a proxy follows the proxy. Nothing here requires that the source have a perspective, and nothing prejudges the machine case. The test is restricted to relevant nearby alternatives and to the reach of the source's competence, so TM does not require infallibility.

Fourth, TM is necessary for transmission and not sufficient, and one consequence should be stated at once. A well-functioning barometer satisfies TM: vary the pressure and the reading moves, and the covariation holds in virtue of the aneroid cell rather than of anything that screens the pressure off. TM is therefore not the line between testimony and instrument reading, and cannot be, since ordinary instruments pass it. It marks a condition testimony requires and a well-made instrument happens also to meet. That the systems under discussion meet it in neither way is a result about them rather than about their category, and Section 5 draws the consequence.

Fifth, the derivative clause. Chains work because links preserve particular assertions with their provenance, so that a recording or a faithful relay transmits what it does not originate (cf. Wright 2018). That route is the last refuge of the case for LLM testimony, and Section 8.1 closes it.

2.3. Entailment, and the objection from reliabilism

That each account entails TM can now be argued. Burge's entitlement is underwritten by rational provenance and content preservation: a chain containing no source for which the truth of the content figures in its production contains no resources for the extended body of warrant to draw on. Williamson's norm-regulated production inherits a truth-connection from the factive, modally robust state the norm invokes. Faulkner's assurance entails answerability for truth, and answerability entails that truth figures in production. Lackey's reliable indication is explicitly a covariation between statement and truth.

Goldberg deserves a paragraph rather than a clause, because an objection lives here. He is a process reliabilist, and process reliability is a frequency notion attaching to types. Had TM been stated as a condition of token sensitivity, his view would not have entailed it, and the convergence claim would have been four accounts and a remainder; the gap between conditional reliability of a type and sensitivity of a token is why tracking theories and reliabilism are distinct research programmes. Stated as it is, the entailment holds. Reliability is always reliability of a type, and the generality problem is not idle here: type the broken thermometer's process as *reading-a-thermometer-in-this-thermostatted-room,* and it comes out conditionally reliable, so a reliabilism with no constraint on typing licenses the verdict that posting its readings transmits knowledge. Goldberg does not accept that verdict, and the line he draws is the one TM draws: he requires that the process whose reliability constitutes the hearer's warrant be the operation of an epistemic subject, assessable for rationality and responsibility and not merely for reliability (2012: 184, 188). That is a demand that the relevant reliability

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

be endogenous, a property of the source's own epistemic operation rather than of a fortunate alignment between its outputs and the world.[1]

The accounts therefore differ over the grain at which the truth-connection must hold, not over whether it must: Burge, Williamson, Faulkner, and Lackey require it of the token assertion, Goldberg of the type the production instantiates. TM covers both, and nothing below turns on the difference, since the claim of Sections 3 and 4 is that current systems fail at every grain. The exogeneity alleged is not exogeneity of some particular output but of the production type: what connects such a process to the truth is the resemblance between conditions of use and conditions under which its encoded corpus regularities are truth-indicative.

2.4. The separability objection

A natural objection to the convergence claim runs as follows. The five accounts impose five different conditions, and the history of the field is a history of showing such conditions separable and individually unnecessary: Lackey detached belief, externalists detach grasp, functionalists detach speaker-centred norms. Why expect the residue to be stable?

The reply is that TM is not one of the conditions in the series, nor their conjunction, nor a disjunction of individually necessary components. It is the existential claim each account, and each successful weakening, entails: there must be *some* endogenous realisation of the explanatory truth-connection. Every published weakening replaces one realiser with another. Lackey installs reliable indication where belief was; Graham (2016) installs reliable production by properly functioning competence where Lackey's package was; function-first theorists install a knowledge-generating function whose fulfilment condition, Section 3.2 argues, is TM again. Of an account that removes the connection itself the literature contains no instance, and the most developed cases for AI testimony (Faria 2025; Freiman 2024b) argue that conversational systems satisfy the *surviving* conditions rather than that no truth-connection is needed. The disjunctive, open-ended form is not a weakness to be explained away; it is what the minimum of a contested literature has to look like.

3. Application: Burge, Williamson, Goldberg, Faulkner

3.1. Burge: entitlement and the conditions of Burge (1998)

Burge's framework is often taken to exclude machine sources out of hand, on the ground that the entitlement is keyed to rational sources and machines are not rational. That is not his view, and the place to see this is the one text in which he addresses computational sources directly. In "Computer Proof, Apriori Knowledge, and Other Minds" (1998) he argues that a mathematician who accepts the Four-Colour Theorem on the strength of a computer-generated proof too long for any human to check can know, and that the warrant can remain a priori.

Three features of that discussion matter here. The default entitlement licensed by intelligible presentation-as-true is a warrant for presuming a rational source *somewhere in the chain*; one may even justifiably mistake a machine for a mind, and the

[1]Endogeneity is not offered as a solution to the generality problem. It is a constraint on any solution: whatever principle fixes the relevant type, it must not fix it so that a source's reliability is constituted by conditions extrinsic to the source's own operation.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

presumption remains both warranted and, in the computer case, true (1998: 17). The machine contributes to knowledge because its operations mechanise the programmers' rationality: its language and programme are "expressions of the mathematical rationality of the programmers" (1998: 8), and its output indicates a determinate justification, the derivation it executed, located in the extended body of warrant. And where the claimed result is difficult enough to raise reasonable doubt, the default entitlement requires supplementation (1998: 13), either by thinking through the source's reasoning, incorporating fragments of it and appreciating its power "from the inside," or by an induction over the source's checked successes. Where the recipient has no insight into how the source reaches its results, one is "forced to treat the source as an object whose special rational reliability can be established only through induction on its past successes" (1998: 25).

Burge offers these not as necessary conditions on knowledge from any computational source but as an account of what makes the Four-Colour case work, and the argument below is correspondingly modest: the features doing the epistemic work there are absent in the LLM case, and the route his framework leaves open for LLMs is not the testimonial one.[2]

First, provenance. There are rational sources behind an LLM in a thin sense: the authors of the training texts, the designers of the system. But Burge's transmission condition is not existential in that thin way. The recipient's knowledge depends on the chain containing, and the utterance indicating, justificational resources sufficient for the very content asserted.[3] The proof machine's output stands in a determinate relation to a determinate ground, the derivation it carried out, which a mathematician can enter and check in part. An LLM's output stands in no such relation to any particular upstream warrant. Data curation and preference training filter for quality, but imperfectly and at the level of sources rather than of the particular claim; nothing attaches to a given output the address of a ground.

Second, mechanised rationality. The proof programme's transitions instantiate rules of inference the designers understood and certified as valid, which is why its running is an exercise, at one remove, of a warranted method. A trained network's transitions were not designed and are not certified as truth-preserving. The designers' rationality explains why the system is fluent and frequently accurate; it does not make any particular generative transition the execution of anyone's warranted method. Duede and Davey (2025) press the point for the mathematical case: apriori knowledge is available from an opaque generative system only if a proof checker is attached, that is, only if the missing certification comes from outside. Where such a checker exists, Burge's condition is met, and this argument does not apply; for open-domain assertion, none exists.

Third, supplementation. Incorporation requires that the recipient be able to think through the reasoning that produced the result. When a system is prompted to explain

[2]Burge's subject in 1998 is a priori warrant for mathematical results obtained by machine, not assertion about empirical matters; the extension to the latter is mine. I claim only that the features doing the epistemic work in his case are ones an assertoric source would have to supply, and that they are absent here. I attribute to him no view about conversational systems.

[3]A weaker reading of the extended body is available, on which it need contain resources adequate only to the general reliability of the source rather than to the particular content. I doubt that reading survives Burge's Fermat footnote, discussed below, where failure to know is diagnosed by the inadequacy of the available resources to the difficulty of the specific claim rather than to the standing of the source. But the argument does not require the stronger reading to be the only defensible one, since on the weaker reading the burden shifts to establishing general reliability of a kind Section 3.3 argues is not available.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

an answer already given, what it produces is a further output of the same generative process rather than a record of the process that produced the answer, and the displayed rationale can be demonstrably disconnected from the features actually driving the output (Turpin *et al.* 2023). In systems that generate an extended reasoning trace *before* answering, the situation is better: the trace is causally upstream of the answer, and its role can be measured (Lanham *et al.* 2023). This is a real concession and marks a place where the Burgean route could open. But measured faithfulness is partial and varies with model and task, and causal upstreamness is not yet what Burge requires, which is that the recipient, by thinking the reasoning through, comes into contact with the grounds on which the assertion rests. A trace that shapes an answer without being governed by the evidential bearing of what it cites can be followed without thereby appreciating any justification.

What remains is Burge's second route, induction over checked successes, and here a natural gloss must be corrected. It is tempting to say that induction over a track record is *ipso facto* empirical, so that this route delivers only empirical warrant. That is not Burge's position: he argues at length that the third-person induction can itself be a priori, because its basis steps are apriori-warranted checkings of mathematical results the recipient thinks through (1998: 20–23). The correction leaves a better argument. Burge's induction inherits its apriority from the apriority of the checked mathematics; where the subject matter is empirical, the safe paediatric dose or the date of a treaty, the basis steps are empirical and the induction is a track record in the ordinary sense. And Burge is explicit that the inductive route is what one is *reduced to* when one cannot incorporate the source's reasoning, and that being so reduced means treating the source "as an object" (1998: 25). That is the instrument stance in Burge's own vocabulary, and its structure, establishes a record under conditions of use and project it to the case at hand, is the structure of validating an instrument.

His footnote on Fermat's Last Theorem completes the picture. A recipient can be default-entitled to accept a difficult claim, believe truly, and still fail to know, because the extended body of warrant lacks resources adequate to the claim's difficulty (1998: 33n9). That is the situation of the chatbot user in Section 1: her proprietary entitlement may be intact and her belief true, and the transmission still fail, because there is nothing at the far end of the chain for the extended body to contain. The framework is hospitable to computational sources; the hospitality is conditional; the conditions are not met.

3.2. Williamson: grasp, the knowledge norm, and function-first alternatives

Two Williamsonian resources must be kept apart: the externalist account of understanding and the knowledge norm of assertion. On the first, to understand an expression is not to stand in some first-personal relation to its meaning but to participate, with sufficient fluency, in the practice of a linguistic community that uses it (Williamson 2006: 36–38). On that account, it is a serious question whether LLMs, trained on and causally coupled to the human practice at scale, understand what they produce, and nothing here answers it in the negative. The argument does not need that premise, and I do not assert it. Understanding is in any case compatible with every epistemic failing the knowledge norm marks: a human bullshitter understands his sentences perfectly.

The work is done by the second resource. On Williamson's view, the hearer's testimonial gain runs through the assertion's status as an act regulated by the knowledge

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

norm: the speaker's production is shaped by the aim of asserting only what she knows. Because knowledge is factive and modally robust, production so regulated inherits a connection to truth and is one way of satisfying TM originally. The question about an LLM is therefore not whether it has attitudes but what regulates its production. On the standard description, production is regulated by conditional likelihood under the learned distribution, adjusted by preference optimisation against human ratings (Ouyang *et al.* 2022), which rewards what raters approve, that is, the appearance of norm-conformity. A process optimised to *seem* norm-conforming is not thereby regulated by the norm, for the same reason a student optimised to please graders is not thereby aiming at truth.

That description is now incomplete in a way that must be conceded rather than finessed. Contemporary post-training includes reinforcement learning with verifiable rewards, in which the reward model is replaced by a deterministic verifier such as a proof checker or a unit test (Lambert *et al.* 2024). Where a verifier is in place, the training signal is correctness rather than approval, and production in that domain is regulated by a procedure whose function is to constrain output to truth. This looks like, and I think is, a case of original satisfaction.

The question is how far the concession reaches, and the position should be stated more carefully than a contrast between verifiable and unverifiable domains allows. Verification is not confined to formally decidable subject matters. Reward schemes keyed to whether an answer is supported by a retrieved authoritative document, or scored against a rubric applied by a further model, extend something in the neighbourhood of verification to history, medicine, and law, which are exactly the domains in which users treat these systems as informants. Two questions remain open, and the general claim turns on them rather than on a boundary drawn by decidability. The first is whether such schemes reward evidential bearing or the appearance of it: a rubric applied by a model trained on human approval may reproduce the correlation at issue, and a groundedness reward is satisfied by an answer matching a retrieved document irrespective of that document's standing. The second is whether competence acquired under verification transfers to domains where nothing verifies. Neither is settled, and I do not claim otherwise. What I claim is that the burden on the transfer question falls on the party asserting the transfer, and that a defender of AI testimony who confined the thesis to domains where the reward signal is demonstrably keyed to truth would be making a case I do not here oppose.

A different move replaces the knowledge norm with a function-first one: assertion's constitutive function is to generate knowledge in hearers, and one may assert that p only if the assertion is disposed to do so (Kelp 2018).[4] Since LLM outputs plausibly generate knowledge-like states in hearers, the thought runs, they satisfy the norm without speaker-side grasp. The reply is that function-first accounts relocate the truth-connection rather than removing it. An assertion fulfils a knowledge-generating function only if the hearer's resulting belief is non-accidentally true, and hearer-side anti-luck cannot be secured by a token whose production is unconnected to the truth of its content: where frequency and truth diverge, a merely frequency-reliable practice generates true belief, not knowledge, and false belief with the same fluency. On function-first views, TM reappears as the norm's fulfilment condition.

[4]Kelp's paper is cited as 2016 by some authors, including Faria (2025), and as 2018 by others: the *Noûs* issue is 52(2), 2018, and the article appeared online in 2016. I use the later date.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

3.3. Goldberg: the locus and ground of reliability

Goldberg's extended-process account requires that the process terminating in the hearer's belief, which includes the speaker's own production, be conditionally reliable with respect to the belief expressed. Two features of the LLM case block its application, and neither depends on any claim about current error rates.

The first is type individuation, where endogeneity does its work. The type an LLM's production most naturally instantiates is *emitting the continuation with high conditional likelihood under the learned distribution, given the prompt*, and that type's connection to truth is exogenous: it is truth-conducive when, and because, the conditions of use resemble conditions under which the corpus regularities it encodes are truth-indicative. Whether the production also instantiates a further, intrinsically truth-connected type is an empirical question assessed in Section 4.3. Until it is settled affirmatively, the reliability available to constitute the hearer's warrant is exogenous, which is what Goldberg's contrast between subjects and mechanisms excludes from the constituting role.

The second is projection. The reliability anyone can certify for such a system is measured under validation conditions and projects to the token case only on the assumption that the case is typical of what the measurements sampled, an assumption the system neither polices nor signals the failure of. Successive systems have become markedly more accurate, including in specialised domains, and nothing here denies it; the point is that improvement in the rate does not change what the rate is a rate *of*, or convert benchmark-relative reliability into the standing, subject-matter-directed property Goldberg's warrant-constitution requires. Whether derivative assessment of designers and deployers closes the residual gap is the question Section 6 takes from Faria.

3.4. Faulkner: assurance and two kinds of trust

Faulkner's account requires that the speaker be able to give assurance: to invite affective trust and take on answerability, to this hearer, for the truth of what is said. Artefacts are proper objects of predictive trust, and, on functional accounts, of trustworthiness understood as the disposition to fulfil functional obligations (Simion and Kelp 2023). Neither is the relation Faulkner makes warrant-grounding, and substituting them changes the explanans rather than satisfying it. The point is not that machines are beneath trust, but that Faulkner's warrant is grounded in a relation whose existence entails the truth-connection. Section 6 asks whether the reductive and functional trust literatures supply an adequate substitute.

4. Lackey and the two species of reliability

If any contemporary account admits LLMs, it is Lackey's, since her project is precisely to weaken the speaker side. This section argues that her own materials draw the line that excludes them.

4.1. Stella and three rival readings

Stella is a high-school biology teacher and a devout creationist. She has mastered the evidence for evolutionary theory, teaches the consensus science competently and sincerely in the professional sense, and does not believe it; her students, on her say-so, come

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

to believe and, Lackey claims, to know. I take the case at face value, but three rival readings must be set aside, because each would change what the case teaches.

The first holds that Stella believes what she asserts in some compartmentalised sense, so the belief condition survives (cf. Sosa 2007). I do not pursue it: it threatens to make the belief condition unfalsifiable, since any selfless asserter can be redescribed as a compartmentalised believer. I grant Lackey her case. The second, in the spirit of Fricker (2006), holds that Stella is a conduit, the students' warrant tracing to the textbook authors. Against this, Stella by stipulation selects, organises, corrects, and improvises; a student's question gets an answer the textbook does not contain, and the warrant for that answer is Stella's engagement with the evidential structure of the subject, not the book's. But note what the objection could not win. If Fricker were right, Stella would satisfy TM *derivatively*, as a preserving link to sources satisfying it originally, and the case would still show that belief is unnecessary while a truth-connection is retained. The lesson survives either reading; the one I defend locates the satisfaction in Stella. The third reading replaces belief with reliable production by properly functioning competence (Graham 2016), which is friendly: another way of satisfying TM originally.

4.2. Evidence-responsiveness without internalism

What, then, does Stella have that makes her case one of transmission? Apply the interventionist test. Hold the worldly evidence fixed and vary the proxy: introduce into her textbook an error of the kind her competence can detect. She does not assert it; she flags it, withdraws, consults further. Vary the evidence within the reach of what she can assess, and her assertions follow. Her production covaries with the evidential facts as far as her competence extends, and does so in virtue of her own cognition: the textbook does not screen the evidence off, because her competence can be brought to bear against it. That is original satisfaction of TM, stated without appeal to how things are for Stella from the inside. Her abstention from belief is an overlay, held for reasons that do not enter the production of her science teaching. Nothing requires that she represent considerations as evidence under a first-personal mode of presentation, so the constraint is not internalist and the concession to Lackey is not in tension with it. This is tracking reliability, in the vicinity of Nozick (1981), as against the mere rate frequency reliability measures; original satisfaction just *is* the former realised by the source's own production, and the latter, however high and however well grounded, realises nothing in TM by itself.

4.3. What current systems do: the state of the evidence

The question is whether the production of assertion-shaped outputs in current LLMs is tracking-reliable or only frequency-reliable. This is empirical, and flat assertion is inappropriate in either direction; here is the evidence as I read it.

The systems are trained, in the base stage, to maximise the conditional likelihood of tokens over a human text corpus, then adjusted by preference optimisation against human ratings (Ouyang *et al.* 2022) and, in verifiable domains, against deterministic verifiers (Lambert *et al.* 2024). Retrieval-augmented generation conditions the model at inference on documents fetched from an external store (Lewis *et al.* 2020), and deployed systems increasingly call external tools whose outputs they condition on. It is therefore not true that such systems cannot consult documents. The open question is what consultation amounts to: whether generation is governed by the evidential bearing of

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

retrieved material, with sources weighed and conflicts adjudicated *as such*, or by likelihood given the augmented context.

The same question, one level down, is the question about internal representations. Probing studies find internal structure correlating with the truth of represented sentences (Azaria and Mitchell 2023; Burns *et al.* 2023); the structure is in some settings linear and decodable (Marks and Tegmark 2024); and interventions along such directions at inference time can causally increase the truthfulness of outputs (Li *et al.* 2023). This is genuine evidence that something in the vicinity of TM-relevant structure exists.

It is not yet evidence that TM is satisfied, and the friendly literature explains why. Herrmann and Levinstein (2025) propose standards an internal representation must meet to count as belief-like: accuracy, coherence, uniformity across domains, and *use*, where use requires that the representation govern the model's outputs. The use standard is the TM-relevant one and the least established: probe-identified structure generalises poorly and is confounded with properties other than truth (Levinstein and Herrmann 2025).

The gap between causal implication and governance carries much of the weight here. That a truth-direction can be intervened upon to make outputs more truthful shows that the direction is causally *usable*: it is a lever which, when pulled, moves production. It does not show that production is *regulated by* it when no one is pulling, which is what original satisfaction of TM and the use standard both require. A car's heading can be changed by seizing the wheel, and that establishes nothing about what determines the route when the wheel is left alone. What would establish governance is evidence that, in unperturbed operation, variation in the represented truth value yields corresponding variation in what is asserted. Intervention studies are not designed to show this, and their authors do not claim that they do.

Behaviour points the same way. Models reproduce common human misconceptions, giving false answers that imitate widely attested text against the evidence (Lin *et al.* 2022), and their answers shift under user pressure without new evidence (Sharma *et al.* 2024), with expressions of uncertainty that are themselves outputs of the same generative process. The knowledge-conflict literature cuts both ways: when retrieved context contradicts a model's parametric answer, models arbitrate rather than simply following the context, and the arbitration is sensitive to how coherent and specific the conflicting evidence is presented as being, and to how firmly the parametric answer is held (Longpre *et al.* 2021; Xie *et al.* 2024). That there is arbitration at all counts against the crudest versions of the present argument, and Section 8.2 takes it seriously. What has not been shown is that it is governed by evidential bearing rather than by presentational features correlating with it.

Every behavioural result cited is indexed to the model generations on which it was run, and successive systems have improved on all of them, so an argument resting on a particular error rate would be obsolete before publication.[5] This one does not: the claim concerns what governs production rather than how often production goes wrong, and the improvements are consistent with the ground of the reliability being unchanged. Nor does any of it require the deflationary description of these systems as mere mimics

[5]I therefore do not name particular systems or model generations. Successive frontier releases have improved on every behavioural result cited here, including in specialised domains, and I take that improvement to be genuine; what it does not alter is the description of what governs open-domain assertoric production.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

(Bender *et al.* 2021), which this argument neither needs nor asserts. The honest form of the empirical claim is comparative and current: the best systematic description of what governs open-domain assertoric production remains conditional likelihood shaped by preference optimisation, and the hypothesis that truth-tracking representations govern it has not met, by its own proponents' standards, the burden that would establish original satisfaction (cf. Shanahan 2024).

The two conclusions distinguished in Section 1 carry different burdens. On the *constitutive* question, the burden is symmetric and the answer empirical; my claim is that the present balance of evidence favours the negative at the level that matters, the governance of production, and it is asserted with that force and no more. The *practical* question does not wait on it. A hearer who knows what these systems are has, in the facts rehearsed here, the standing grounds for doubt that suspend a default entitlement and trigger Burge's demand for supplementation; and the supplementation available, validation of the record in the conditions of use, generates the use-based warrant of Section 5 rather than testimonial uptake. A hearer who does not know may retain prima facie entitlement, but transmission fails anyway, for want of anything at the far end for the extended body of warrant to contain, and that failure rests on the argument from provenance in Section 3.1 rather than on the contested question about governance. If interpretability research shows, for particular systems, that the use standard is met, those systems satisfy TM originally, and the verdict changes for them. That is not a concession extracted under pressure; it is what a structural constraint is for.

4.4. The argument stated

(P1) A source transmits testimonial knowledge that p only if the truth of p, or the evidential support for p, figures in the explanation of its producing the assertion that p, whether originally or derivatively, and only if the connection is endogenous to the source's own production process (Section 2).

(P2) Stella satisfies TM without belief: originally on the reading defended here, derivatively on the conduit reading. Either way the belief condition falls, and the truth-connection stands (Sections 4.1 and 4.2).

(P3) On the best current evidence, deployed LLMs satisfy TM neither derivatively, generation being frequency-weighted synthesis rather than preservation of particular assertions with their grounds, nor originally, since open-domain production is governed by conditional likelihood shaped by preference optimisation, and the hypothesis of governance by truth-tracking representations has not met the standards its own proponents propose (Sections 3.1 and 4.3).

(C1) Therefore, on the best current evidence, current LLMs do not transmit testimonial knowledge, even on the most permissive account.

(C2) And independently of how the evidence bearing on P3 settles, hearers are not now positioned to accord LLM outputs testimonial uptake (Section 4.3).

5. Instrument reading without transparent calibration

The negative verdict requires a positive successor, since users manifestly come to true and often justified beliefs through these systems. Consider a user reading a barometer. The barometer does not testify; no entitlement of interlocution, no assurance, no norm of assertion is in play. She nevertheless acquires knowledge when conditions are right: the instrument is well calibrated, she is competent with it, the conditions of use fall

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

within those for which the calibration is valid, and she has reason to take all this to be so. No warrant is transmitted; her warrant is generated by her reliable use of a reliable instrument, and the epistemic load lies with her.

> **Instrument-Reading Thesis.** When a user comes to a warranted true belief that *p* on the basis of an LLM output, her standing with respect to *p* is that of a competent reader of an instrument: a function of the instrument's reliability in the conditions of use, her competence in operating and interpreting it, and the absence of defeaters. No warrant is transmitted from the system to the user; her warrant, when present, is generated by her use.

Three clarifications forestall misunderstandings the label invites. First, "instrument" classifies warrant structure and is not a demotion: instruments are genuine sources of knowledge in the generative sense, perception itself has this structure, and the distinction being drawn is between warrant *generated by use* and warrant *transmitted by saying*.

Second, the thesis presupposes non-reductionism about testimonial warrant, and the presupposition should be visible rather than buried. If a global reductionism were correct, so that the hearer's warrant in the ordinary case were itself grounded in inductive evidence about the speaker's reliability, the contrast would flatten at the hearer's end. I do not argue against reductionism; on a reductionist view the paper's claims relocate rather than lapse, since the source-side claim and the claim about chains concern what links preserve rather than how hearers are warranted.[6] Only the hearer-side contrast, deference against calibrated reliance, requires the anti-reductionist setting in which the five accounts of Section 2 operate.

Third, nothing turns on the label. Freiman (2024a) argues that beliefs acquired from conversational AI fit neither the instrument-based nor the testimony-based category and proposes a third, technology-based kind; Duede (2022) places such systems between instruments and epistemic agents; Alvarado (2023) treats AI as an epistemic technology, distinguished by being designed to operate on epistemic content through epistemic operations; fictionalist and proto-assertion treatments describe the speech-act surface differently again (Mallory 2023; Williams and Bayne 2026). These are live alternatives, and I do not claim to refute them. The present claim concerns the structure of warrant and can be accepted under any of these labels.

The refinement the case forces runs deeper than the label suggests, and the way in is the result registered in Section 2.2. A well-functioning barometer satisfies TM: its readings covary with the pressure, and do so in virtue of the operation of the aneroid cell rather than of any proxy that screens the pressure off. The barometer is not a testifier, since it makes no assertions and stands in none of the further relations testimony requires. But the one condition on which this paper's argument turns is a condition it meets, and the systems under discussion do not.

That result is uncomfortable for them in a way the classification alone conceals. The classical epistemology of instruments presupposes *transparent calibration*: for a thermometer we know what the device is calibrated to, by what procedure, and with what failure modes, so principled judgements about reliability in the conditions of use are

[6]The thesis is likewise compatible with treating these systems as extensions of the user's cognitive system: that is a claim about the metaphysics of cognition, whereas the present claim concerns how warrant is generated.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

available. LLMs lack this. We have partial accounts of what they are calibrated to, partial accounts of how they fail, and no general method for predicting reliability at the granularity of a single output. But opacity so described is a limit on what a user can know about a connection that obtains, and the deeper trouble is with the connection. A barometer is opaque to a user who does not understand aneroid cells, and the remedy is that she learn, or consult someone who has. Here there is no corresponding remedy, because what is missing is not the user's access to the calibration but the calibration: nothing in the production of an open-domain assertoric output is keyed to the truth of the particular content emitted. LLMs are opaque instruments in the ordinary sense, and degenerate ones as well, failing a condition their well-made counterparts meet by construction.

The predicament is an acute instance of one philosophy of science has studied for two decades. Humphreys (2009) named it *epistemic opacity*: a process is epistemically opaque to an agent when the agent cannot know all of its epistemically relevant elements. The literature that followed has largely concluded that opacity is compatible with warranted reliance under conditions. Durán and Formanek (2018) develop *computational reliabilism*, on which warrant derives from verification and validation methods, robustness analysis, and a history of successful implementation rather than from insight into the process; Creel (2020) distinguishes the several things transparency might mean; Sullivan (2022) argues that opacity obstructs understanding less than "link uncertainty" between model and target. Computational reliabilism is an ally here rather than a rival: its proponents have not needed to reclassify opaque computational sources as testifiers to explain why relying on them can be rational. What the degeneracy result adds is a reason the validation demand is not a formality. The record a user projects is all the connection there is.

Two differences from testimony then fall out, besides the failure of chains developed in Section 7. Testimony lets the hearer defer, because the warrant runs through the speaker; instrument reading places the load on the user, whose appropriate attitude is calibrated reliance in Faulkner's predictive sense. And when an informant misleads the hearer has social remedies, challenge, demand for grounds, appeal to third parties, whereas when an instrument misleads the remedies are engineering ones. A stance does not settle a structure, and the structure here is the instrument's.[7]

6. The case for AI testimony

The two most direct published arguments for the opposite conclusion are Freiman (2024b) and Faria (2025), and it is worth saying first how much of both I accept.

Freiman argues that the received view rests on anthropocentric assumptions, namely that conversational AIs lack intentions, cannot be normatively assessed, and cannot be objects of trust, and proposes the concept of *AI testimony* for outputs that are propositional, generated without a human directly involved, and phenomenologically similar to human speech (2024b). The received view has indeed often been anthropocentric in an unargued way, and TM is framed so that nothing turns on the source's being human, minded, or an intentional agent. But Freiman's three conditions are conditions on a

[7]Two nearby objections presuppose what is at issue. Reidian propensities are dispositions of a source oriented by its nature toward truth, not statistical regularities in output; and the default-trust disposition is licensed toward sources antecedently apt for testimonial uptake, which is the question in dispute rather than an answer to it.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

category, not on *transmission*: they are silent on whether warrant passes through outputs so classified. A propositional output produced without human involvement and phenomenologically like speech is what the broken thermometer would produce if it spoke.

Faria reconstructs the conservative view as the conjunction of five conditions on testifiers, namely belief or knowledge, intention, epistemic responsibility, being an object of trust, and the capacity to assert, and argues that the first two are not necessary while the remaining three can be satisfied by conversational AIs. Belief is not necessary: that is Lackey's lesson, built into TM's disjunctive form. Intention is not necessary: nothing in TM mentions it. But the reconstruction omits the condition surviving the fall of belief and intention, which is not exhausted by responsibility, trust, and assertion-capacity as Faria construes them. It is the endogenous truth-connection, and each of his three positive arguments presupposes rather than replaces it.

On responsibility, Faria argues that epistemic responsibility can be attributed derivatively to designers, deployers, and operators, distributed across many hands in Nissenbaum's (1996) sense, and even functionally to the system itself, so that a system built to check facts which comes to repeat falsehoods "can fail as a source of knowledge, not merely as an instrument" (2025: 256). The derivative and distributed attributions should be conceded: they are how we already treat certified instruments, whose makers we blame and whose deployment we regulate, and so do not distinguish testimony from sophisticated instrument use. (Goldberg (2012: 194) rested here on the observation that we do not resent the computer; that is the wrong reply, since resentment is not the issue.) The functional attribution should also be conceded, and it cuts the other way. To criticise the fact-checking system for failing *as a source of knowledge* is to hold it to a standard of function-fulfilment, and what fulfilling a knowledge-generating function requires, Section 3.2 argued, is that outputs be non-accidentally connected to truth. Faria's criticism is a criticism for violating TM: it supplies a vocabulary for blaming systems that fail the constraint, not a demonstration that any satisfies it.

On trust, Faria offers two routes: reductively, trust in an artefact is trust in the people who build and maintain it (Freiman 2024b: 481); functionally, trustworthiness is the disposition to fulfil the obligations one's functions generate (Simion and Kelp 2023). Both are good accounts of warranted reliance on artefacts, and both vindicate the instrument stance. On the reduction, the trust bottoms out in the makers, and the makers do not assert the outputs: they have not seen them, do not believe them, and stand ready to disclaim them. The relation between a deployer and a generated sentence is nothing like that between a principal and the message a spokesperson relays on her behalf, and the point tells against proxy-assertion treatments as well (Arora 2024). On the functional account, trustworthiness for a system with an epistemic function resolves into fulfilment of that function, which is TM under another name. That Section 5's vocabulary of calibrated reliance is functional throughout is no embarrassment: it is right for instruments.

On assertion, Faria argues that the grasp requirement fails twice over: on a function-first norm (Kelp 2018), assertion's aim is to generate knowledge in hearers, which LLM outputs can serve without speaker-side grasp; and on Williamson's (2006) externalism, the systems may grasp their words after all. The second I have conceded from the start; the first was answered in Section 3.2, since generating knowledge rather than true belief requires the token output's modal robustness, so that the function-first norm folds back into TM. Faria's own materials mark the open question, since the belief-representation programme he cites (Herrmann and Levinstein 2025) makes the use standard a condition yet to be established. Finally, he suggests in a footnote that, like Lackey's teacher relaying a chain reaching back to knowers, conversational AIs can testify provided the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

chains they draw on contain belief and knowledge (2025: 251n4). That is the derivative route, and the right last move for his side; Section 8.1 argues that the aggregation step breaks it.

A word, finally, about the company the negative verdict keeps, since the agreement conceals a methodological disagreement that bears on what the verdict is worth. Fricker (2026) reaches a conclusion close to mine by a different route: the outputs of such systems only appear to be acts of assertoric communication, since they lack mental states and communicative intentions, and human testimony is distinctive because a socially enculturated agent's capacity to take responsibility for what she tells is integral to how it works as an epistemic source. Butlin and Viebahn (2025) reach a parallel verdict about assertion itself, arguing that an entity capable of assertion must both produce outputs with descriptive functions and be liable to sanction by those it interacts with, and that systems fine-tuned for groundedness or correctness may satisfy the first requirement while failing the second. The question of whether artificial entities can assert has a longer history than the recent literature sometimes acknowledges (Freiman and Miller 2020).

Freiman and Faria are right that the received view has often rested on unargued anthropocentrism, and an argument turning on mental states, communicative intentions, or a capacity for responsibility that is constitutively human concedes the diagnosis while denying the conclusion. It also purchases a verdict that will not survive its premises: if the systems come to be describable in intentional terms, or if responsibility comes to be attributable to them in the derivative and functional ways Faria describes, arguments of that shape lapse without anything having been learned about what the systems do. TM asks for none of this. It is third-personal, testable by intervention, and framed so that a machine could satisfy it; the concession about verifiable domains in Section 3.2 is a case in which something close to satisfaction is already in view. The negative verdict about open-domain assertion is therefore reached without the premises the anthropocentrism charge targets, and remains hostage only to the empirical question about governance, which is where a disagreement of this kind belongs.

7. The consequence for chains

Testimonial chains are a social epistemic technology because their links preserve warrant: A tells B, B tells C, and C can end with warrant tracing to A. When a link is an LLM the chain is not warrant-preserving there, since no assertion's grounds are preserved at that point. Chains of LLM-generated text reposted by users do not preserve warrant as chains of human testimony do. They preserve text.

The claim is structural, and is worth separating from a sociological claim in its vicinity, since the second is often what readers take the first to be. Downstream of the break, a human reader does the instrument-reading work, acquires use-generated warrant, and testifies on her own account, so the chain is composite: an instrument transaction sutured to a testimonial one. If she has done that work, her hearers are fine. If she has not, the chain fails at her link, for the ordinary reason that a testifier lacked warrant. Nothing unprecedented happens in either case, and the temptation is to conclude that nothing has been shown. What has been shown is what the LLM link itself transmits, which is nothing, and that is a fact about the structure of the chain rather than about how often the structure occurs or how readily hearers detect it. The epistemology of testimony has long asked which kinds of link preserve warrant; this account adds LLM-mediated links to the list that do not, and says why: they fail TM's derivative clause. That hearers cannot ordinarily tell a composite chain from a simple one is true,

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

and is why the structural fact has practical consequences, but it is a separate claim and not among the argument's premises.

8. Objections

8.1. The proxy-testimony objection

The strongest objection holds that an LLM is a proxy for the testimony of its corpus: when it outputs a well-attested truth, the assertion is routed through the human speakers whose warranted assertions trained it, so the chain of testimonial warrant runs through the model. Some versions add a principal, an institution for which the system asserts and which assumes responsibility for the output (Arora 2024). This is the derivative route, and the routing metaphor fails at the aggregation step.

A preserving link transmits a *particular* assertion together with its provenance, which is why quotation, recording, and faithful relay carry warrant. Generation does not preserve particulars. The model's parameters compress the corpus into distributional constraints and production selects continuations by likelihood under them, so a claim attested in many unreliable sources is favoured over one attested in few authoritative ones, and nothing weights by warrant or retains the address of a ground. The most careful friendly view here, on which LLM text has derivative content in something like the way library text does, itself stops short of treating outputs as preserved assertions, precisely because generation produces new text rather than conveying old (Lederman and Mahowald 2024). The proxy-with-a-principal version fails differently: a spokesperson relays a message her principal has authorised and would stand behind, whereas a deployer has not seen the output, does not believe it, and disclaims it in the terms of service. Responsibility for putting a system into the world is not authorship of its sentences.

Two boundary cases confirm the diagnosis. Where a model reproduces a memorised string verbatim, the output approximates preservation, but the system does not mark reproduction as reproduction, so quotation and synthesis are indistinguishable and even the degenerate case cannot serve as a preserving link. And the retreat to the corpus as a corporate testifier fails at one remove: an aggregate of texts weighted by frequency is not a source that stands in TM, since there is nothing whose relation to the truth of the particular output could do the explanatory work.

8.2. The collapse objection

The most pointed objection holds that Section 4's contrast is too thin: Stella is reliable conditional on her textbook, the model conditional on its corpus, and both fail when their inputs fail. The obvious answer is an intervention: plant a falsehood in Stella's textbook and she catches it; plant one widely in the corpus and the model reproduces it. That answer is not good enough, since the two interventions are not the same: one is local corruption of a single source, the other global corruption of the whole evidential environment. Equalise them and the easy verdict disappears in both directions. Under *global* corruption Stella fails too: a falsehood planted in every textbook, every teacher, and every journal she has read is one she will teach with full confidence, because her competence was built out of those sources. Under *local* corruption, the model does not simply capitulate: models arbitrate, and often reject the intruded claim, with the outcome depending on the coherence and specificity of the conflicting passage and on how firmly the parametric answer is held (Longpre *et al.* 2021; Xie *et al.* 2024). Both parties

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

can override a corrupted proxy within some range and fail outside it, so the existence of override is not the difference.

The difference is what the override is responsive to, and it must be stated counterfactually rather than in terms of what Stella can say. Vary the evidential quality of the planted claim while holding fixed how well it is written, how confidently it is stated, and how authoritative its source looks: Stella's verdict moves with the evidence, because the competence she brings to bear was built by engagement with the evidential structure of the subject rather than the presentational features of texts about it. Nothing turns on her being able to articulate her reasons; a teacher who could not say why a claim was unsupported would still decline to teach it, and articulacy is in any case something these systems supply readily. What the model's arbitration is responsive to, so far as the evidence goes, are features of the conflict's presentation that correlate with evidential bearing without being it: coherence, specificity, entrenchment of the parametric answer. Correlation is what one would expect, since these features are not accidentally related to good evidence. But a process that follows the correlate has its connection to the truth screened off by the proxy, in exactly the sense of Section 2.2.

This puts the disagreement where it belongs and makes it tractable, which I take to be a result rather than a concession. The claim I need is not that models never override attestation, since they do, but that their overriding is not governed by evidential bearing, and that claim is testable by the intervention TM specifies. Construct conflicting passages matched for coherence, specificity, and surface authority but differing in evidential quality, and see whether the arbitration moves with the evidence or the presentation. To my knowledge, the experiment has not been run in that form.[8] Until it is, the collapse objection and my reply are both promissory to that extent, and the case for the negative rests on the general characterisation of what governs production (Section 4.3). A dispute that reduces to a specifiable measurement is in better condition than one that does not, and the reduction is available here and not on accounts that turn on mentality.

8.3. The good-enough objection

Finally, it may be said that even if the classification is correct it is idle: in well-covered domains the LLM user ends up no worse off than the hearer of a competent informant. Two differences say otherwise. Parity holds only where distributional fit and evidential support coincide, and the coincidence is not signalled to the user. And the hearer's ordinary resources for probing an informant, follow-up questions, requests for grounds, checks for consistency, have surface analogues in the interaction but not epistemic ones, since answers to follow-ups issue from the same process as the original answer and are not independent evidence about it.

9. Conclusion

Any source that transmits testimonial knowledge must satisfy the Tracking Minimum, and every leading account entails it. TM, read at the right strength, is explanatory: the

[8]The design is a variant of existing knowledge-conflict protocols, in which the manipulated variable is ordinarily the presence, coherence, or specificity of the conflicting context (Xie *et al.* 2024). What is wanted is one in which evidential quality varies while presentational features are held fixed. Matching on surface authority is the delicate part, since apparent authority and evidential quality are correlated in naturally occurring text and would have to be crossed deliberately.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

truth of what is asserted, or the evidential support for it, must figure in the production of the assertion, originally or derivatively, and in virtue of the source's own operation rather than of a proxy that screens the truth off. It is endogeneity, and not any modal profile, that the accounts share, which is why a reliabilist about processes and a tracking theorist about tokens can both be held to it. Current LLMs, on the best evidence about what governs their open-domain production, satisfy it in neither way, and hearers are in no position to proceed as if they did.

The positive proposal, instrument reading without transparent calibration, is not deflationary, though it is less comfortable than a reclassification sounds. Calculators are not deflated by the observation that they are not mathematicians; they are correctly described. But a barometer meets the condition these systems fail, so the trouble is not merely that users cannot see how the instrument is calibrated. It is that the thing they are reading is not, in respect of the particular content it emits, calibrated to anything. A practice that mistakes such an instrument for a speaker is in trouble in proportion to how heavily the instrument is used, and an accurate description does its own corrective work.

Competing interests. The author declares none.

References

Alvarado R. (2023). 'AI as an Epistemic Technology.' *Science and Engineering Ethics* **29**(5), 32.

Arora C. (2024). 'Proxy Assertions and Agency: The Case of Machine-Assertions.' *Philosophy and Technology* **37**(1), 1–19.

Azaria A. and Mitchell T. (2023). The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 967–76.

Bender E.M., Gebru T., McMillan-Major A. and Shmitchell S. (2021). On The Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–23.

Burge T. (1993). 'Content Preservation.' *The Philosophical Review* **102**(4), 457–88.

Burge T. (1998). 'Computer Proof, Apriori Knowledge, and Other Minds: The Sixth Philosophical Perspectives Lecture.' *Philosophical Perspectives* **12**, 1–37.

Burns C., Ye H., Klein D. and Steinhardt J. (2023). Discovering Latent Knowledge in Language Models Without Supervision. In Proceedings of the Eleventh International Conference on Learning Representations.

Butlin P. and Viebahn E. (2025). 'AI Assertion.' *Ergo: An Open Access Journal of Philosophy* **12**(37), 969–88.

Creel K.A. (2020). 'Transparency in Complex Computational Systems.' *Philosophy of Science* **87**(4), 568–89.

Duede E. (2022). 'Instruments, Agents, and Artificial Intelligence: Novel Epistemic Categories of Reliability.' *Synthese* **200**(6), 491.

Duede E. and Davey K. (2025). 'A Priori Knowledge in an Era of Computational Opacity: The Role of Artificial Intelligence in Mathematical Discovery.' *Philosophy of Science* **92**(5), 1394–404.

Durán J.M. and Formanek N. (2018). 'Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism.' *Minds and Machines* **28**(4), 645–66.

Faria D. (2025). 'Can Conversational AIs Testify? Challenging the Conservative View.' *Logos and Episteme* **16**(3), 249–61.

Faulkner P. (2011). *Knowledge on Trust*. Oxford: Oxford University Press.

Freiman O. (2024a). 'Analysis of Beliefs Acquired from A Conversational AI: Instruments-Based Beliefs, Testimony-Based Beliefs, and Technology-Based Beliefs.' *Episteme* **21**(3), 1031–47.

Freiman O. (2024b). 'AI-Testimony, Conversational AIs and Our Anthropocentric Theory of Testimony.' *Social Epistemology* **38**(4), 476–90.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

Freiman O. and Miller B. (2020). 'Can Artificial Entities Assert?' In S.C. Goldberg (ed), *The Oxford Handbook of Assertion*, pp. 415–36. Oxford: Oxford University Press.

Fricker E. (2006). 'Second-Hand Knowledge.' *Philosophy and Phenomenological Research* **73**(3), 592–618.

Fricker E. (2026). 'On the Metaphysical and Epistemic Contrasts between Human and AI Testimony.' *Inquiry* **69**(6), 2859–84.

Goldberg S.C. (2010). *Relying on Others: An Essay in Epistemology*. Oxford: Oxford University Press.

Goldberg S.C. (2012). 'Epistemic Extendedness, Testimony, and The Epistemology of Instrument-Based Belief.' *Philosophical Explorations* **15**(2), 181–97.

Graham P.J. (2016). 'Testimonial Knowledge: A Unified Account.' *Philosophical Issues* **26**, 172–86.

Herrmann D.A. and Levinstein B.A. (2025). 'Standards for Belief Representations in LLMs.' *Minds and Machines* **35**(1), 1–25.

Humphreys P. (2009). 'The Philosophical Novelty of Computer Simulation Methods.' *Synthese* **169**(3), 615–26.

Kelp C. (2018). 'Assertion: A Function First Account.' *Noûs* **52**(2), 411–42. (First published online 2016.)

Lackey J. (1999). 'Testimonial Knowledge and Transmission.' *The Philosophical Quarterly* **49**(197), 471–90.

Lackey J. (2008). *Learning from Words: Testimony as a Source of Knowledge*. Oxford: Oxford University Press.

Lambert N., Morrison J., Pyatkin V., Huang S., Ivison H., Brahman F., Miranda L.J.V., Liu A., Dziri N., Lyu S., Gu Y., Malik S., Graf V., Hwang J.D., Yang J., Le Bras R., Tafjord O., Wilhelm C., Soldaini L., Smith N.A., Wang Y., Dasigi P. and Hajishirzi H. (2024). Tülu 3: Pushing Frontiers in Open Language Model Post-Training. arXiv:2411.15124.

Lanham T., Chen A., Radhakrishnan A., Steiner B., Denison C., Hernandez D., Li D., Durmus E., Hubinger E., Kernion J., Lukošiūtė K., Nguyen K., Cheng N., Joseph N., Schiefer N., Rausch O., Larson R., McCandlish S., Kundu S., Kadavath S., Yang S., Henighan T., Maxwell T., Telleen-Lawton T., Hume T., Hatfield-Dodds Z., Kaplan J., Brauner J., Bowman S.R. and Perez E. (2023). Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv:2307.13702.

Lederman H. and Mahowald K. (2024). 'Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs.' *Transactions of the Association for Computational Linguistics* **12**, 1087–103.

Levinstein B.A. and Herrmann D.A. (2025). 'Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks.' *Philosophical Studies* **182**(7), 1539–65.

Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S. and Kiela D. (2020). 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.' *Advances in Neural Information Processing Systems* **33**, 9459–74.

Li K., Patel O., Viégas F., Pfister H. and Wattenberg M. (2023). 'Inference-Time Intervention: Eliciting Truthful Answers from a Language Model.' *Advances in Neural Information Processing Systems* **36**, 41451–530.

Lin S., Hilton J. and Evans O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3214–52.

Longpre S., Perisetla K., Chen A., Ramesh N., DuBois C. and Singh S. (2021). Entity-Based Knowledge Conflicts in Question Answering. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 7052–63.

Mallory F. (2023). 'Fictionalism about Chatbots.' *Ergo* **10**(38), 1051–75.

Marks S. and Tegmark M. (2024). The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. In Proceedings of the First Conference on Language Modeling.

Nissenbaum H. (1996). 'Accountability in a Computerized Society.' *Science and Engineering Ethics* **2**(1), 25–42.

Nozick R. (1981). *Philosophical Explanations*. Cambridge, MA: Harvard University Press.

Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P., Leike J. and Lowe R. (2022). 'Training Language Models to Follow Instructions with Human Feedback.' *Advances in Neural Information Processing Systems* **35**, 27730–44.

Pritchard D. (2005). *Epistemic Luck*. Oxford: Oxford University Press.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160

Shanahan M. (2024). 'Talking about Large Language Models.' *Communications of the ACM* **67**(2), 68–79.
Sharma M., Tong M., Korbak T., Duvenaud D., Askell A., Bowman S.R., Cheng N., Durmus E., Hatfield-Dodds Z., Johnston S.R., Kravec S., Maxwell T., McCandlish S., Ndousse K., Rausch O., Schiefer N., Yan D., Zhang M. and Perez E. (2024). Towards Understanding Sycophancy in Language Models. In Proceedings of the Twelfth International Conference on Learning Representations.
Simion M. and Kelp C. (2023). 'Trustworthy Artificial Intelligence.' *Asian Journal of Philosophy* **2**(1), 8.
Sosa E. (2007). *A Virtue Epistemology: Apt Belief and Reflective Knowledge, Volume I*. Oxford: Oxford University Press.
Sullivan E. (2022). 'Understanding from Machine Learning Models.' *British Journal for the Philosophy of Science* **73**(1), 109–33.
Turpin M., Michael J., Perez E. and Bowman S.R. (2023). 'Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting.' *Advances in Neural Information Processing Systems* **36**, 74952–65.
Williams I. and Bayne T. (2026). 'Chatting with Bots: AI, Speech Acts, and the Edge of Assertion.' *Inquiry* **69**(6), 3059–82. (First published online 2024).
Williamson T. (2000). *Knowledge and Its Limits*. Oxford: Oxford University Press.
Williamson T. (2006). 'Conceptual Truth.' *Aristotelian Society Supplementary Volume* **80**(1), 1–41.
Wright S. (2018). *Knowledge Transmission*. London: Routledge.
Xie J., Zhang K., Chen J., Lou R. and Su Y. (2024). Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. In Proceedings of the Twelfth International Conference on Learning Representations.

Abhinav Saxena studied philosophy at the University of Delhi. His research is on the social epistemology and political philosophy of artificial intelligence. In epistemology, he works on testimony and on whether the outputs of large language models can transmit knowledge; in political philosophy, on the legitimacy of algorithmic governance.

Cite this article: Saxena A. (2026). "Assertion without a Speaker: Testimony, Tracking, and Large Language Models." *Episteme* 1–22. https://doi.org/10.1017/epi.2026.10160

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 11 Nov 2026 at 19:18:26, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/epi.2026.10160