# The Claims Ledger

*Every claim this research programme stands on, uses or has to answer to*

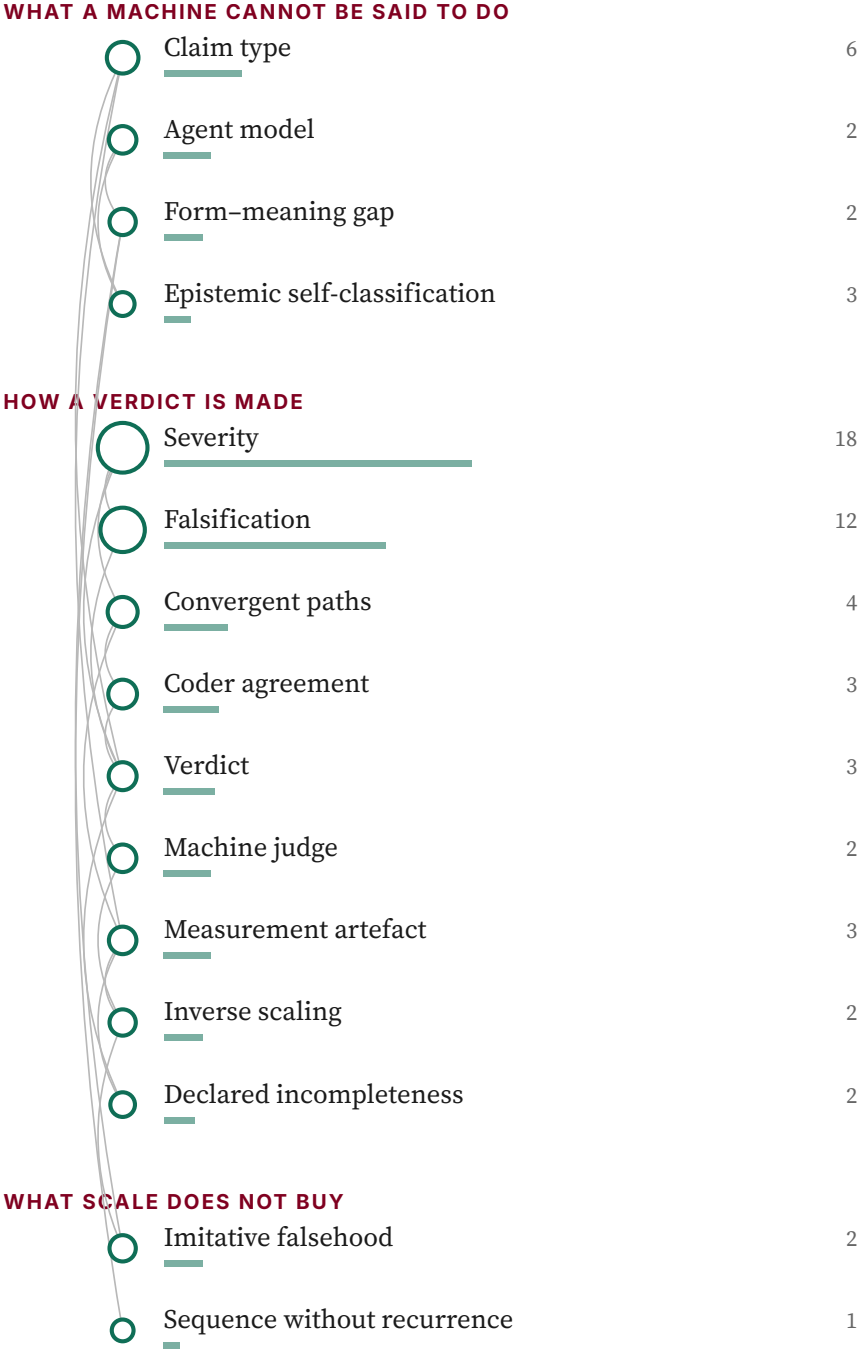
*Each one verbatim, with the context it came in, the question it answers and the links it keeps with the others. Arranged as the map: question, concept, claim.*

CLAUDIO CAMMARANO 17 NOVEMBER 2026

| | |
|---|---|
| EDITION | 17 November 2026 · 54 claims, 15 concepts, 3 questions |
| NOTATION | ENTIA 1.1.0, doi.org/10.5281/zenodo.23303451 |
| ORIGIN | https://claudiocammarano.com/en/claims/ |

# The map

*The same questions as propositions, with the links declared between them. The bar is weight, the arc is a link; the number counts the claims under each concept.*



# Part I · What a machine cannot be said to do

*The limit is argued from how these systems are trained, not from what they fail at today.*

## Claim type

**Does the kind of a claim decide how it can be checked?**

What kind of claim something is decides how it can be checked at all: the ontology comes before the epistemology.

LINKED TO Verdict · Measurement artefact

---

**C001** The governing principle, which became the paper's title, is this: ontology precedes epistemology. **You cannot determine whether a claim is true without first establishing what type of claim you are dealing with.** A statistical claim and a normative claim share nothing in their truth conditions — treating them as interchangeable is the error the framework is designed to prevent.

> *Before asking whether a statement is true, you have to settle what kind of statement it is: the test that fits a measurement does not fit a value judgement. The whole framework rests on taking that step first, and in the open.*

SOURCE Lab note, https://claudiocammarano.com/lab/2026-02-03-the-seven-dimensional-taxonomy-v2/, ¶4

NAMES Cammarano

MARKS conceptual tool · D1.14 metaphysical-ontological · assumed

RELATIONS refines → C002

ONLINE https://claudiocammarano.com/en/claims/c001/

---

**C002** You cannot determine whether a claim is true without first establishing what type of claim you are dealing with. **A statistical claim and a normative claim share nothing in their truth conditions — treating them as interchangeable is the error the framework is designed to prevent.**

> *Two claims of different kinds share no standard of truth, so a procedure that works on one returns a meaningless verdict on the other. The failure is not getting the answer wrong; it is applying the wrong test and not noticing.*

SOURCE Lab note, https://claudiocammarano.com/lab/2026-02-03-the-seven-dimensional-taxonomy-v2/, ¶4

MARKS conceptual tool · D1.14 metaphysical-ontological · assumed

ONLINE https://claudiocammarano.com/en/claims/c002/

---

**C003** Dimension D1 — Content Type is the load-bearing axis. It currently includes eleven active categories (D1.1 through D1.10, plus D1.14), covering the full range from historical-factual and empirical-quantitative claims to ethical-normative and metaphysical-ontological ones.

**The remaining dimensions [...] operate as modifiers that describe how a given claim type behaves epistemically.**

> *One dimension of the taxonomy decides what a claim is; the other six describe how it behaves once that is settled. Keeping a single axis load-bearing is what stops the scheme from flattening into a list of labels.*

**SOURCE** Lab note, https://claudiocammarano.com/lab/2026-02-03-the-seven-dimensional-taxonomy-v2/, ¶3

**NAMES** Cammarano

**MARKS** working method · D1.10 prescriptive-policy · assumed

**RELATIONS** extends → C001

**ONLINE** https://claudiocammarano.com/en/claims/c003/

---

**C004** The third — that a model can report on its own internal states and still not classify the epistemic type of its own assertions, which is why the check has to come from outside — is an argument. **Measurements and experiments are answerable to data and to a coding protocol. Arguments are answerable to other arguments.** I had been asking one document to be accountable to both, and the seams were showing long before the second model made them visible.

> *A measurement, an experiment and an argument are accountable to different things, so they cannot be asked to share one document and one standard of proof. This is the taxonomy's own principle turned on the programme's publication plan, and it is what split the working paper in two.*

**SOURCE** Lab note, https://claudiocammarano.com/lab/2026-09-03-what-counts-as-one-paper/, ¶10

**MARKS** conceptual tool · D1.3 logical-formal · used

**RELATIONS** extends → C001

**ONLINE** https://claudiocammarano.com/en/claims/c004/

---

**C005** It is a satisfying convergence and it is not a demonstration. I measured 180 outputs. They measured representations, in a different setting, for a different purpose. **A convergence between two lines of evidence is not the same as a link between them, and writing it up as though it were would be the category collapse this project exists to catch.**

> *Two independent findings pointing the same way is not evidence that they are connected. The rule is applied here against the programme's own interest, on a convergence it would have been convenient to claim.*

**SOURCE** Lab note, https://claudiocammarano.com/lab/2026-07-28-what-counts-as-introspection/, ¶11

**MARKS** working method · D1.10 prescriptive-policy · used

**RELATIONS** refines → C020

**ONLINE** https://claudiocammarano.com/en/claims/c005/

---

**C046** Now conservative 'activists' hold that we are born with our basic expectations; with them we turn the world into 'our world' but must then live for ever in the prison of our world.

**The idea that we live and die in the prison of our 'conceptual framework' was developed primarily by Kant: pessimistic Kantians thought that the real world is for ever unknowable because of this prison, while optimistic Kantians thought that God created our conceptual framework to fit the world. But revolutionary activists believe that conceptual frameworks can be developed and also replaced by new, better ones; it is we who create our 'prisons' and we can also, critically, demolish them.**

> *Conceptual frameworks are made, not given, and can be rebuilt critically. The programme's taxonomy is such a framework: a prison it built and has to be ready to demolish.*

SOURCE Lakatos (1978), *The Methodology of Scientific Research Programmes*, Cambridge University Press, p. 20, ch. 1

MARKS conceptual tool · D1.14 metaphysical-ontological · assumed

ONLINE https://claudiocammarano.com/en/claims/c046/

## Agent model

**Does predicting text require representing the mind that wrote it?**

A model can carry partial representations of the intentions behind the text it was trained on without being an agent itself.

LINKED TO Form–meaning gap

---

**C006** In particular, I claim: **In the course of performing next-word prediction in context, current LMs sometimes infer approximate, partial representations of the beliefs, desires and intentions possessed by the agent that produced the context, and other agents mentioned within it.** (C2) Once these representations are inferred, they are causally linked to LM prediction, and thus bear the same relation to generated text that an intentional agent's state bears to its communciative actions.

> *Predicting the next word can require the model to build a partial picture of what the writer believed and wanted. The claim is about representations inside the system, not about the system being someone.*

SOURCE Andreas (2022), *Language Models as Agent Models*, Advances in Neural Information Processing Systems, doi:10.18653/v1/2022.findings-emnlp.423, §1 Introduction, (C1)

NAMES Andreas

MARKS conceptual tool · D1.14 metaphysical-ontological · contested · contested

RELATIONS contests → C008

ONLINE https://claudiocammarano.com/en/claims/c006/

---

**C007** This debate is not just academic; the extent and manner in which machines understand our world has real stakes for how much we can trust them to drive cars, diagnose diseases, care for the elderly, educate children, and more generally act robustly and transparently in tasks that impact humans. **Moreover, the current debate suggests a fascinating divergence in how to think about understanding in intelligent systems, in particular the contrast between mental models that rely on statistical correlations and those that rely on causal mechanisms.** Until quite recently there was general agreement in the AI research community about machine understanding: while AI systems exhibit seemingly intelligent behavior in many specific tasks, they do not understand the data they process in the way

humans do.

> *The dispute over machine understanding is not one question but two pictures of what understanding is: tracking statistical regularities, or holding a causal model of how things work. Naming the split explains why the two camps talk past each other.*

**SOURCE** Mitchell & Krakauer (2023), *The Debate Over Understanding in AI's Large Language Models*, Proceedings of the National Academy of Sciences, doi:10.1073/pnas.2215907120, p. 1

**NAMES** Mitchell · Krakauer

**MARKS** conceptual tool · D1.6 interpretive-hermeneutic · used

**RELATIONS** presupposes → C008

**ONLINE** https://claudiocammarano.com/en/claims/c007/

## Form–meaning gap

**Can a system trained only on form learn what the language is about?**

Training on form alone leaves meaning out of reach: what reads as understanding is a property of the text, not of the system.

**LINKED TO** Agent model · Imitative falsehood · Sequence without recurrence

---

**C008** From our perspective, these are overclaims caused by a misunderstanding of the relationship between linguistic form and meaning. **We argue that the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning.** We take the term language model to refer to any system trained only on the task of string prediction, whether it operates over characters, words or sentences, and sequentially or not.

> *A system trained only on the shape of language has nothing in its input that could tell it what the language is about. The argument is about what the training signal contains, not about how well the system performs today.*

**SOURCE** Bender & Koller (2020), *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, doi:10.18653/v1/2020.acl-main.463, §1 Introduction

**NAMES** Bender · Koller

**MARKS** technical reference · D1.14 metaphysical-ontological · assumed

**ONLINE** https://claudiocammarano.com/en/claims/c008/

**ALSO HERE** C007, given under Agent model.

## Epistemic self-classification

**Can a model report what kind of claim it is about to make?**

A system can report on contents of its own processing and still not represent what kind of claim it is about to make: the typing has to come from outside.

**LINKED TO** Agent model · Claim type

**C009** **A model that reports a concept implanted in its own activations is doing something, and "no metacognition" is not an accurate name for it.** So the flat version is gone. I am not interested in defending it.

> *The programme used to assert flatly that the producing system performs no metacognition over its own outputs. A 2026 result on models reporting concepts implanted in their own activations withdrew that premise, and this is the withdrawal, kept in the archive rather than quietly edited out.*

SOURCE Lab note, https://claudiocammarano.com/lab/2026-07-28-what-counts-as-introspection/, ¶6

MARKS corrective datum · D1.14 metaphysical-ontological · assumed

ONLINE https://claudiocammarano.com/en/claims/c009/

---

**C010** The reportable representations carry contents. They do not carry a typing of the assertion those contents are about to become. **So §3.3 now reads: language models lack epistemic self-classification. It is a smaller claim, and it is better positioned than the one it replaces.** The old premise was a claim about a capability, and capability claims are hostage to the next experiment. The new one is a claim about a representational content, which is what the framework was always about.

> *The premise the whole framework rests on, in its narrow form: a model may represent that a passage is in Spanish, or that it is being evaluated, without representing that the assertion it is about to make is causal rather than correlational. The typing of the claim is the one content the self-report does not carry.*

SOURCE Lab note, https://claudiocammarano.com/lab/2026-07-28-what-counts-as-introspection/, ¶8

NAMES Cammarano

MARKS conceptual tool · D1.14 metaphysical-ontological · assumed

RELATIONS refines → C009 · refines → C006

ONLINE https://claudiocammarano.com/en/claims/c010/

---

**C011** Alongside this, the reportable component of a concept's representation turns out to carry a median of six to seven per cent of that representation's variance, with the rest sitting outside it. **What a model can say about its own processing is a narrow and selective channel onto that processing, not a transcript of it.** Lindsey's own numbers point the same way from a different angle: introspective detection succeeds on a minority of trials even in the strongest models he tested, and models asked to describe the mechanics behind their own calculations often describe them wrongly.

> *What a model reports about its own workings is not a log of them. The reportable part of a representation carries a small share of its variance, and the same information can drive the task while bypassing that part entirely, which is why a protocol cannot ask the model to certify its own output.*

SOURCE Lab note, https://claudiocammarano.com/lab/2026-07-28-what-counts-as-introspection/, ¶10

MARKS corrective datum · D1.2 empirical-quantitative · assumed

RELATIONS confirms → C010

**ONLINE** https://claudiocammarano.com/en/claims/c011/

# Part II · How a verdict is made

*A judgement counts only with the procedure that produced it attached.*

## Severity

**When does passing a test count as evidence?**

A result counts as evidence only if the procedure that produced it could have caught the error: agreement a method was bound to find shows nothing.

LINKED TO Verdict · Convergent paths

---

**C012** I call this the severity requirement. In its weakest form it supplies a minimal requirement for evidence: Severity Requirement (weak): **One does not have evidence for a claim if nothing has been done to rule out ways the claim may be false. If data x agree with a claim C but the method used is practically guaranteed to find such agreement, and had little or no capability of finding flaws with C even if they exist, then we have bad evidence, no test (BENT).** The "practically guaranteed" acknowledges that even if the method had some slim chance of producing a disagreement when C is false, we still regard the evidence as lousy.

> *Agreement between data and a claim is not evidence when the method could not have found the claim false. This is the negative, minimal criterion the programme uses to tell a validation that tested something from one that only confirmed what it presupposed.*

SOURCE Mayo (2018), *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, doi:10.1017/9781107286184, p. 5, Excursion 1, Tour I, §1.1

NAMES Mayo

MARKS conceptual tool · D1.8 ethical-normative · assumed

ONLINE https://claudiocammarano.com/en/claims/c012/

---

**C013** I also advocate the positive counterpart: Severity (strong): **We have evidence for a claim C just to the extent it survives a stringent scrutiny. If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, then the passing result, x, is evidence for C.** One way this can be achieved is by an argument from coincidence.

> *The positive counterpart of the weak requirement: evidence is graded by how capable the test was of finding flaws. It matters because it turns severity from a veto into a measure, which is what a validation protocol needs to rank results rather than just reject them.*

SOURCE Mayo (2018), *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, doi:10.1017/9781107286184, p. 14, Excursion 1, Tour I, §1.2

NAMES Mayo

MARKS conceptual tool · D1.8 ethical-normative · assumed

RELATIONS extends → C012

ONLINE https://claudiocammarano.com/en/claims/c013/

**C014** First, there's the idea I call lift-off. Lift-off: **An overall inference can be more reliable and precise than its premises individually.** Each scale, by itself, has some possibility of error, and limited precision.

> *Convergent, independent checks can warrant a conclusion more strongly than any single one of them. This is the epistemic ground for aggregating multiple fallible runs or judges instead of demanding a single infallible one.*

**SOURCE** Mayo (2018), *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, doi:10.1017/9781107286184, p. 15, Excursion 1, Tour I, §1.2

**NAMES** Mayo

**MARKS** conceptual tool · D1.3 logical-formal · used

**ONLINE** https://claudiocammarano.com/en/claims/c014/

---

**C015** The argument to "weight gain" is an example of an argument from coincidence to the absence of an error, what I call: Arguing from Error: **There is evidence an error is absent to the extent that a procedure with a very high capability of signaling the error, if and only if it is present, nevertheless detects no error.** I am using "signaling" and "detecting" synonymously:

> *Absence of a detected error counts only if the procedure would have signalled that error had it been there. It is the operational form of severity, and the test any judge-based validation has to pass before its silence can be read as a result.*

**SOURCE** Mayo (2018), *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, doi:10.1017/9781107286184, p. 16, Excursion 1, Tour I, §1.2

**NAMES** Mayo

**MARKS** working method · D1.3 logical-formal · assumed

**RELATIONS** refines → C013

**ONLINE** https://claudiocammarano.com/en/claims/c015/

---

**C016** I'm interested in the problem of how we get the kind of knowledge we do manage to obtain – and how we can get more of it. **Statistical methods should not be seen as tools for what philosophers call "rational reconstruction" of a piece of reasoning. Rather, they are forward-looking tools to find something out faster and more efficiently, and to discriminate how good or poor a job others have done.** The job of the philosopher is to clarify but also to provoke reflection and scrutiny precisely in those areas that go unchallenged in ordinary practice.

> *Mayo treats inferential methods as instruments for finding things out and for grading others' work, not as after-the-fact justifications. The programme reads validation protocols the same way: as probes that can fail, not as rationalisations of verdicts already reached.*

**SOURCE** Mayo (2018), *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, doi:10.1017/9781107286184, p. 8, Excursion 1, Tour I

**NAMES** Mayo

**MARKS** working method · D1.8 ethical-normative · used

ONLINE https://claudiocammarano.com/en/claims/c016/

---

**C017** Borrowing a term from philosophy of science, we may call it our meta-methodology – it's one level removed. To put my cards on the table: **A severity scrutiny is going to be a key method of our meta-methodology. It is fairly obvious that we want to scrutinize how capable a statistical method is at detecting and avoiding erroneous interpretations of data.** So when it comes to the role of probability as a pedagogical tool for our purposes, severity – its assessment and control – will be at the center.

> *Severity is applied one level up, to methods rather than to claims: a method is judged by its capacity to catch erroneous interpretations. This is the move the programme makes when it treats severity as a meta-level property of validation procedures.*

SOURCE Mayo (2018), *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, doi:10.1017/9781107286184, p. 9, Excursion 1, Tour I, Methodology and Meta-methodology

NAMES Mayo

MARKS working method · D1.9 procedural-instructional · assumed

RELATIONS presupposes → C013

ONLINE https://claudiocammarano.com/en/claims/c017/

---

**C030** others strive to say H0 is less believable given data x than before, and offer a quantitative report of the difference. **What happened to the goal of scrutinizing BENT science by the severity criterion? Neither "probabilism" nor "performance" directly captures that demand.** To take these goals at face value, it's easy to see why they come up short.

> *Neither of the two dominant philosophies of statistics, assigning degrees of belief or securing long-run error rates, asks whether a test could have caught the error. That gap is why the programme cannot borrow its validation criterion from either school and has to state severity on its own.*

SOURCE Mayo (2018), *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, doi:10.1017/9781107286184, p. 13, Excursion 1, Tour I

MARKS conceptual tool · D1.6 interpretive-hermeneutic · assumed

RELATIONS presupposes → C012

ONLINE https://claudiocammarano.com/en/claims/c030/

---

**C031** So when it comes to the role of probability as a pedagogical tool for our purposes, severity – its assessment and control – will be at the center. **The term "severity" is Popper's, though he never adequately defined it. It's not part of any statistical methodology as of yet.** Viewing statistical inference as severe testing lets us stand one level removed from existing accounts, where the air is a bit clearer.

> *Severity comes from Popper but arrived without a working definition and without a place in any statistical method. Mayo's contribution, and the programme's starting point, is that operational definition: the word is old, the criterion is not.*

SOURCE Mayo (2018), *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, doi:10.1017/9781107286184, p. 9, Excursion

1, Tour I, Methodology and Meta-methodology

**MARKS** technical reference · D1.1 historical-factual · used

**RELATIONS** refines → C017

**ONLINE** https://claudiocammarano.com/en/claims/c031/

---

**C032** We want to tell what's true about statistical inference. Among the most notable of these truths is: **P-values can be readily invalidated due to how the data (or hypotheses!) are generated or selected for testing.** If you fool around with the results afterwards, reporting only successful guesses, your report will be invalid.

> *A significance level is valid only relative to how data and hypotheses were selected; selective reporting makes a hard result look easy to explain away as chance. For the programme this is the warning attached to any choice of which outputs or cases enter a validation set.*

**SOURCE** Mayo (2018), *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, doi:10.1017/9781107286184, p. 17, Excursion 1, Tour I

**MARKS** corrective datum · D1.4 mathematical-analytic · used

**ONLINE** https://claudiocammarano.com/en/claims/c032/

---

**C033** The same kind of reasoning is at the heart of experimental testing. I shall call it arguing from error. **After learning enough about certain types of mistakes, we may construct (often from other tests) a testing procedure with an overwhelmingly good chance of revealing the presence of a specific error, if it exists-but not otherwise. Such a testing procedure may be called a severe (or reliable) test, or a severe error probe.** If a hypothesized error is not detected by a test that has an overwhelmingly high chance of detecting it, if instead the test yields a result that accords well with no error, then there are grounds for the claim that the error is absent.

> *A severe test is built from prior knowledge of specific mistakes: a probe that would reveal a given error if present, and stay silent otherwise. This is the original formulation of the idea, and it ties severity to an error repertoire rather than to probability in a hypothesis.*

**SOURCE** Mayo (1996), *Error and the Growth of Experimental Knowledge*, University of Chicago Press, p. 7, ch. 1, §8

**MARKS** working method · D1.9 procedural-instructional · assumed

**RELATIONS** confirms → C015

**ONLINE** https://claudiocammarano.com/en/claims/c033/

---

**C036** In particular, Bayesian model checking, they say, uses statistics to satisfy Popperian criteria for severe tests. **The idea of error statistical foundations for Bayesian tools is not as preposterous as it may seem. The concept of severe testing is sufficiently general to apply to any of the methods now in use.** On the face of it, any inference, whether to the adequacy of a model or to a posterior probability, can be said to be warranted just to the extent that it has withstood severe testing.

> *Severity is not tied to one statistical school: any inference can be judged by whether it withstood severe testing. That generality is what licenses the programme to apply it to LLM validation, a domain Mayo did not consider.*

**SOURCE** Mayo (2018), *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, doi:10.1017/9781107286184, p. 27, Excursion 1, Tour I

**MARKS** conceptual tool · D1.6 interpretive-hermeneutic · assumed

**RELATIONS** extends → C017

**ONLINE** https://claudiocammarano.com/en/claims/c036/

---

**C037** Inductive methods - understood as methods of experimental testing - are justified to the extent that they are error-correcting methods. We may call this Peirce's error-correcting or self-correcting thesis (SCT): Self-Correcting Thesis SCT: **methods for inductive inference in science are error correcting; the justification for inductive methods of experimental testing in science is that they are self-correcting.** Peirce's SCT has been a source of fascination and frustration.

> *Peirce justified induction not by the truth of its conclusions but by the method's capacity to correct its own errors over time. It is the older root of error statistics, and the ground for treating a validation protocol as justified by what it would catch, not by what it has found so far.*

**SOURCE** Mayo (2005), *Peircean Induction and the Error-Correcting Thesis*, Transactions of the Charles S. Peirce Society, p. 299

**MARKS** conceptual tool · D1.6 interpretive-hermeneutic · used

**RELATIONS** presupposes → C038

**ONLINE** https://claudiocammarano.com/en/claims/c037/

---

**C038** Peirce's philosophy of inductive inference in science is based on the idea that what permits us to make progress in science, what allows our knowledge to grow, is the fact that science uses methods that are self-correcting or error correcting: **Induction is the experimental testing of a theory. The justification of it is that, although the conclusion at any stage of the investigation may be more or less erroneous, yet the further application of the same method must correct the error.** Inductive methods - understood as methods of experimental testing - are justified to the extent that they are error-correcting methods.

> *Induction is justified by the method, not by any single conclusion: persisting in it corrects whatever error a stage contains. It is the original of the self-correcting thesis that Mayo turns into the basis of error statistics.*

**SOURCE** Peirce (1931–1958), *Collected Papers of Charles Sanders Peirce*, Harvard University Press, CP 5.145; quoted in Mayo 2005, p. 299

**MARKS** conceptual tool · D1.6 interpretive-hermeneutic · used

**ONLINE** https://claudiocammarano.com/en/claims/c038/

---

**C039** In describing his theory of inference, Peirce could be describing that of the error-statistician: **The theory here proposed does not assign any probability to the inductive or hypothetic conclusion, in the sense of undertaking to say how frequently that conclusion would be found true. [...] The theory here presented only says how frequently, in this universe, the special form of induction or hypothesis would lead us right.** The probability given by this theory is in every way different - in meaning, numerical value, and form - from that of those who would apply to ampliative inference the doctrine of inverse chances.

> *Probability attaches to the method's track record, not to the hypothesis. The programme inherits the split: a validation reports how often a procedure gets it right, never how probable a given verdict is.*

**SOURCE** Peirce (1931–1958), *Collected Papers of Charles Sanders Peirce*, Harvard University Press, CP 2.748; quoted in Mayo 2005, p. 300

**MARKS** conceptual tool · D1.14 metaphysical-ontological · assumed

**ONLINE** https://claudiocammarano.com/en/claims/c039/

---

**C040** For starters, induction, for Peirce, is a matter of subjecting hypotheses to "the test of experiment" (7.182). **The process of testing it will consist, not in examining the facts, in order to see how well they accord with the hypothesis, but on the contrary in examining such of the probable consequences of the hypothesis . . . which would be very unlikely or surprising in case the hypothesis were not true.** When, however, we find that prediction after prediction, notwithstanding a preference for putting the most unlikely ones to the test, is verified by experiment, ... we begin to accord to the hypothesis a standing among scientific results.

> *A test does not collect agreement; it looks for consequences that would be surprising if the hypothesis were false. This is severity before the word, a century before Mayo, and the reason fit with the data is never enough on its own.*

**SOURCE** Peirce (1931–1958), *Collected Papers of Charles Sanders Peirce*, Harvard University Press, CP 7.231; quoted in Mayo 2005, p. 304

**MARKS** conceptual tool · D1.9 procedural-instructional · assumed

**RELATIONS** confirms → C013

**ONLINE** https://claudiocammarano.com/en/claims/c040/

---

**C041** In particular, if we are wrong to reject Ho (and Ho is actually true), we would find we were rarely able to get so statistically significant a result to recur, and in this way we would discover our original error. **It is true that the observed conformity of the facts to the requirements of the hypothesis may have been fortuitous. But if so, we have only to persist in this same method of research and we shall gradually be brought around to the truth.** The correction is not a matter of getting higher and higher probabilities, it is a matter of finding out whether the agreement is fortuitous; whether it is generated about as often as would be expected were the agreement of the chance variety.

> *A lucky agreement is exposed by persisting with the same method: replication is the correction. For validation, a single passing run is not a result until the procedure has been run again.*

**SOURCE** Peirce (1931–1958), *Collected Papers of Charles Sanders Peirce*, Harvard University Press, CP 7.115; quoted in Mayo 2005, p. 309

**MARKS** working method · D1.9 procedural-instructional · used

**RELATIONS** refines → C038

**ONLINE** https://claudiocammarano.com/en/claims/c041/

---

**C042** Yet a further safeguard is at hand: **Nor must we lose sight of the constant tendency of the inductive process to correct itself. This is of its essence. This is the marvel of it. ...even though doubts may be entertained whether one selection of instances is a random one, yet a different selection, made by a different method, will be likely to vary from the normal in**

**a different way, and if the ratios derived from such different selections are nearly equal, they may be presumed to be near the truth.** Here, the marvel is an inductive method's ability to correct the attempt at random sampling.

> *Two selections made by different methods err in different ways, so their agreement is evidence about the truth. It is the convergence argument the programme uses when independent runs or coders reach the same verdict.*

**SOURCE** Peirce (1931–1958), *Collected Papers of Charles Sanders Peirce*, Harvard University Press, CP 2.729; quoted in Mayo 2005, p. 315

**MARKS** conceptual tool · D1.6 interpretive-hermeneutic · assumed

**RELATIONS** confirms → C014

**ONLINE** https://claudiocammarano.com/en/claims/c042/

---

**C043** But this is very different from the uniformity of nature assumption. **...the negative fact supposed by me [no mysterious force interferes with the action of chance] is merely the denial of any major premise from which the falsity of the inductive conclusion could be deduced. Actually so long as the influence of this mysterious source not be overwhelming, the wonderful self-correcting nature of the ampliative inference would enable us, even so, to detect and make allowance for them.** Not only do we not need the uniformity of nature assumption, Peirce declares

> *Induction needs no uniformity of nature, only the absence of an overwhelming hidden bias, and lesser biases are themselves detectable by the method. A validation does not have to assume a clean world, only one where its errors are not engineered against it.*

**SOURCE** Peirce (1931–1958), *Collected Papers of Charles Sanders Peirce*, Harvard University Press, CP 2.749; quoted in Mayo 2005, p. 316

**MARKS** conceptual tool · D1.14 metaphysical-ontological · used

**ONLINE** https://claudiocammarano.com/en/claims/c043/

## Falsification

**What does it take for a test to count against a theory?**

A refutation is not Nature speaking: it is a verdict produced by a procedure, and it counts only as far as that procedure could have gone the other way.

**LINKED TO** Severity · Verdict

---

**C034** It seems incumbent upon anyone mounting such an approach to dispel the ghosts of Popper's negativism right away, or at least sketch how they will be dispelled. **The most devastating criticism of Popper's approach is this: having rejected the notion that learning is a matter of building up probability in a hypothesis, Popper seems to lack any meaningful way of saying why passing severe tests counts in favor of a hypothesis.** Popper's account seems utterly incapable of saying anything positive.

> *Falsificationism explains why failing a test counts against a hypothesis, but not why passing one counts for it. The programme needs the positive half, which is why it uses Mayo's severity rather than Popper's corroboration.*

**SOURCE** Mayo (1996), *Error and the Growth of Experimental Knowledge*, University of Chicago Press, p. 8, ch. 1

**MARKS** conceptual tool · D1.6 interpretive-hermeneutic · used

**RELATIONS** refines → C031

**ONLINE** https://claudiocammarano.com/en/claims/c034/

---

**C035** These test criteria will be based on the idea of severity sketched above. **A hypothesis H that passes the test failed by rival hypothesis H' (and by other alternative hypotheses considered) has passed a severe test for Popper-but not for me. Why not? Because for H to pass a severe test in my sense, it must have passed a test that is highly capable of probing the ways in which H can err.** And the test that alternative hypothesis H' failed need not be probative in the least so far as the errors of H go.

> *Beating rivals is not passing a severe test: the test has to probe the ways this particular hypothesis could be wrong. For validation this means comparing models or judges against each other is not a substitute for probing each one's own failure modes.*

**SOURCE** Mayo (1996), *Error and the Growth of Experimental Knowledge*, University of Chicago Press, p. 9, ch. 1

**MARKS** conceptual tool · D1.6 interpretive-hermeneutic · assumed

**RELATIONS** extends → C034

**ONLINE** https://claudiocammarano.com/en/claims/c035/

---

**C044** Of course, replacing proof by probability was a major retreat for justificationist thought. But even this retreat turned out to be insufficient. **It was soon shown, mainly by Popper's persistent efforts, that under very general conditions all theories have zero probability, whatever the evidence; all theories are not only equally unprovable but also equally improbable.** Many philosophers still argue that the failure to obtain at least a probabilistic solution of the problem of induction means that we "throw over almost everything that is regarded as knowledge by science and common sense".

> *If every universal theory has zero probability on any evidence, probability cannot be what separates good theories from bad ones. It is the negative result that pushes the programme, like Peirce and Mayo, to grade procedures rather than hypotheses.*

**SOURCE** Lakatos (1978), *The Methodology of Scientific Research Programmes*, Cambridge University Press, p. 11, ch. 1

**MARKS** corrective datum · D1.4 mathematical-analytic · used

**RELATIONS** confirms → C039

**ONLINE** https://claudiocammarano.com/en/claims/c044/

---

**C045** To (non-tautologous) unfalsifiable propositions the dogmatic falsificationist gives short shrift; he brands them 'metaphysical' and denies them scientific standing. **Dogmatic falsificationists draw a sharp demarcation between the theoretician and the experimenter: the theoretician proposes, the experimenter - in the name of Nature - disposes.** As Weyl put it: 'I wish to record my unbounded admiration for the work of the experimenter in his struggle to wrest interpretable facts from an unyielding Nature who knows so well how to meet our theories with a decisive No - or with an inaudible Yes.'

> *Dogmatic falsificationism treats the experimenter as Nature's neutral voice, delivering verdicts the theory cannot touch. Lakatos sets it up in order to take it apart, and so does the programme: a validator's verdict is itself a theory-laden procedure, not Nature speaking.*

**SOURCE** Lakatos (1978), *The Methodology of Scientific Research Programmes*, Cambridge University Press, p. 13, ch. 1

**MARKS** conceptual tool · D1.6 interpretive-hermeneutic · contested · contested

**ONLINE** https://claudiocammarano.com/en/claims/c045/

---

**C047** A theory is a tool which we test by applying it, and which we judge as to its fitness by the results of its applications. **From a logical point of view, the testing of a theory depends upon basic statements whose acceptance or rejection, in its turn, depends upon our decisions. Thus it is decisions which settle the fate of theories.** To this extent my answer to the question, 'how do we select a theory?' resembles that given by the conventionalist; and like him I say that this choice is in part determined by considerations of utility.

> *Every test bottoms out in basic statements that someone decides to accept; logic alone never forces the stop. For the programme this is the status of a coder's label: the verdict rests on a decision under rules, which is why the rules have to be published.*

**SOURCE** Popper (1959), *The Logic of Scientific Discovery*, Hutchinson, doi:10.4324/9780203994627, §30; Routledge Classics ed., p. 91

**MARKS** conceptual tool · D1.14 metaphysical-ontological · assumed

**ONLINE** https://claudiocammarano.com/en/claims/c047/

---

**C048** But for me the choice is decisively influenced by the application of the theory and the acceptance of the basic statements in connection with this application; whereas for the conventionalist, aesthetic motives are decisive. **Thus I differ from the conventionalist in holding that the statements decided by agreement are not universal but singular. And I differ from the positivist in holding that basic statements are not justifiable by our immediate experiences, but are, from the logical point of view, accepted by an act, by a free decision.** (From the psychological point of view this may perhaps be a purposeful and well-adapted reaction.)

> *What is decided by agreement is the singular observation, not the theory, and no experience justifies it: it is accepted. It is the exact status of an annotation in a validation set, agreed on and therefore open to re-examination.*

**SOURCE** Popper (1959), *The Logic of Scientific Discovery*, Hutchinson, doi:10.4324/9780203994627, §30; Routledge Classics ed., p. 92

**MARKS** conceptual tool · D1.14 metaphysical-ontological · assumed

**RELATIONS** refines → C047

**ONLINE** https://claudiocammarano.com/en/claims/c048/

---

**C049** This story strongly suggests that even a most respected scientific theory, like Newton's dynamics and theory of gravitation, may fail to forbid any observable state of affairs. **Indeed, some scientific theories forbid an event occurring in some specified finite spatio-temporal region (or briefly, a 'singular event') only on the condition that no other factor (possibly hidden in some distant and unspecified spatio-temporal corner of the universe) has any influence on it.** But then such theories never alone contradict a 'basic' statement:

they contradict at most a conjunction of a basic statement describing a spatio-temporally singular event and of a universal non-existence statement saying that no other relevant cause is at work anywhere in the universe.

> *A theory forbids an event only under a ceteris paribus clause, so a failed prediction never refutes it alone. For validation, a failed check indicts the whole test setup, not only the claim under test.*

**SOURCE** Lakatos (1978), *The Methodology of Scientific Research Programmes*, Cambridge University Press, p. 17, ch. 1

**MARKS** conceptual tool · D1.3 logical-formal · used

**ONLINE** https://claudiocammarano.com/en/claims/c049/

---

**C050** If, however, still accepting the demarcation criterion of dogmatic falsificationism, we deny that facts can prove propositions, then we certainly end up in complete scepticism: then all science is undoubtedly irrational metaphysics and should be rejected. **Scientific theories are not only equally unprovable, and equally improbable, but they are also equally undisprovable.** But the recognition that not only the theoretical but all the propositions in science are fallible, means the total collapse of all forms of dogmatic justificationism as theories of scientific rationality.

> *Neither proof, nor probability, nor refutation can be read off the facts: all three fail together. It is the dead end that makes method, not data, the place where rationality has to be rebuilt.*

**SOURCE** Lakatos (1978), *The Methodology of Scientific Research Programmes*, Cambridge University Press, p. 19, ch. 1

**MARKS** corrective datum · D1.3 logical-formal · used

**RELATIONS** extends → C044

**ONLINE** https://claudiocammarano.com/en/claims/c050/

---

**C051** Thus 'all swans are white' is either an oddity and easily disprovable or a scientific proposition with a ceteris paribus clause and therefore undisprovable. **Tenacity of a theory against empirical evidence would then be an argument for rather than against regarding it as 'scientific'. 'Irrefutability' would become a hallmark of science.** To sum up: classical justificationists only admitted proven theories; neoclassical justificationists probable ones; dogmatic falsificationists realized that in either case no theories are admissible.

> *Once ceteris paribus clauses are allowed, a theory that survives every challenge looks more scientific, not less: the reductio of dogmatic falsificationism. A claim that no check can touch is the warning sign, not the achievement.*

**SOURCE** Lakatos (1978), *The Methodology of Scientific Research Programmes*, Cambridge University Press, p. 19, ch. 1

**MARKS** conceptual tool · D1.3 logical-formal · contested · contested

**RELATIONS** presupposes → C049

**ONLINE** https://claudiocammarano.com/en/claims/c051/

---

**C052** The Duhemian conservative conventionalist (or 'methodological justificationist', if you wish) makes unfalsifiable by fiat some (spatiotemporally) universal theories, which are distinguished by their explanatory power, simplicity or beauty. **Our Popperian**

**revolutionary conventionalist (or 'methodological falsificationist') makes unfalsifiable by fiat some (spatio-temporally) singular statements which are distinguishable by the fact that there exists at the time a 'relevant technique' such that 'anyone who has learned it' will be able to decide that the statement is 'acceptable'. Such a statement may be called an 'observational' or 'basic' statement, but only in inverted commas.** Indeed, the very selection of all such statements is a matter of a decision, which is not based on exclusively psychological considerations.

> *An observation counts as basic because a trained technique makes it decidable, not because it is given. It describes a coding manual exactly: what makes a label acceptable is that anyone trained on it would reach it.*

**SOURCE** Lakatos (1978), *The Methodology of Scientific Research Programmes*, Cambridge University Press, p. 22, ch. 1

**MARKS** working method · D1.6 interpretive-hermeneutic · used

**RELATIONS** refines → C048

**ONLINE** https://claudiocammarano.com/en/claims/c052/

---

**C053** Calling these statements 'observational' is no more than a manner of saying that, in the context of his problem, that is, in testing our gravitational theory, the methodological falsificationist uses radio-optics uncritically, as 'background knowledge'. **The need for decisions to demarcate the theory under test from unproblematic background knowledge is a characteristic feature of this brand of methodological falsificationism.** (This situation does not really differ from Galileo's 'observation' of Jupiter's satellites: moreover, as some of Galileo's contemporaries rightly pointed out, he relied on a virtually non-existent optical theory - which then was less corroborated, and even less articulated, than present-day radio-optics.

> *Every test decides what is under test and what is taken for granted, and that line is a decision, not a fact. In a validation, the judge model and the coding manual are the background knowledge, and the programme has to say so.*

**SOURCE** Lakatos (1978), *The Methodology of Scientific Research Programmes*, Cambridge University Press, p. 23, ch. 1

**MARKS** working method · D1.9 procedural-instructional · used

**ONLINE** https://claudiocammarano.com/en/claims/c053/

---

**C054** To sum up: **the methodological falsificationist offers an interesting solution to the problem of combining hard-hitting criticism with fallibilism. Not only does he offer a philosophical basis for falsification after fallibilism had pulled the carpet from under the feet of the dogmatic falsificationist, but he also widens the range of such criticism very considerably.** By putting falsification in a new setting, he saves the attractive code of honour of the dogmatic falsificationist: that scientific honesty consists in specifying, in advance, an experiment such that, if the result contradicts the theory, the theory has to be given up.

> *Methodological falsificationism keeps hard criticism while admitting every statement is fallible, by making the decisions explicit. It is the stance of a validation protocol: fallible verdicts, stated in advance, applied without prevarication.*

**SOURCE** Lakatos (1978), *The Methodology of Scientific Research Programmes*, Cambridge University Press, p. 28, ch. 1

**MARKS** conceptual tool · D1.6 interpretive-hermeneutic · used

**RELATIONS** contests → C045

**ONLINE** https://claudiocammarano.com/en/claims/c054/

## Convergent paths

**Is agreement between different derivations evidence about the answer?**

Several different derivations arriving at the same answer is evidence about the answer, not noise to be averaged away.

**LINKED TO** Coder agreement · Measurement artefact

---

**C025** It has been observed that chain-of-thought prompting significantly improves model performance across a variety of multi-step reasoning tasks (Wei et al., 2022). **Self-consistency leverages the intuition that complex reasoning tasks typically admit multiple reasoning paths that reach a correct answer [...].** The more that deliberate thinking and analysis is required for a problem (Evans, 2010), the greater the diversity of reasoning paths that can recover the answer.

> *When several different chains of reasoning arrive at the same answer, that agreement is itself evidence for the answer. The method samples many derivations and keeps the one they converge on.*

**SOURCE** Wang & Wei & Schuurmans & Le & Chi & Narang & Chowdhery & Zhou (2022), *Self-consistency improves chain of thought reasoning in language models*, §1 Introduction

**MARKS** conceptual tool · D1.3 logical-formal · used

**ONLINE** https://claudiocammarano.com/en/claims/c025/

---

**C020** The gap is the finding. **Surface-level variance [...] masks high semantic convergence.** An LLM that produces thirty different textual responses to the same prompt is not thirty times unreliable; it may be converging on the same epistemic conclusion through thirty different linguistic paths. Distinguishing these two phenomena requires a framework. That is what this project provides.

> *Thirty differently worded answers to the same prompt are not thirty disagreements: they can be one conclusion reached thirty ways. Measuring the wording instead of the conclusion makes a model look far less stable than it is.*

**SOURCE** Lab note, https://claudiocammarano.com/lab/2026-02-28-vertical-validation-180-api-runs/, ¶11

**MARKS** corrective datum · D1.2 empirical-quantitative · assumed

**RELATIONS** uses-method-in → C019

**ONLINE** https://claudiocammarano.com/en/claims/c020/

**ALSO HERE** C014, given under Severity.

**ALSO HERE** C042, given under Severity.

## Coder agreement

**What does agreement between coders actually measure?**

Agreement between coders measures a procedure, not a truth, and it improves more by adding coders than by correcting for their bias.

**LINKED TO** Verdict · Convergent paths

---

**C018** Regarding the question of annotator bias, there is an overwhelming consensus in CL practice: K and α are used in the vast majority of the studies we reported. **We agree with the view that K and α are more appropriate, as they abstract away from the bias of specific coders. But we also believe that ultimately this issue of annotator bias is of little consequence because the differences get smaller and smaller as the number of annotators grows [...].** We believe that increasing the number of annotators is the best strategy, because it reduces the chances of accidental personal biases.

> *When measuring whether two people label things the same way, correcting for each coder's personal bias matters less than how many coders there are. For a coding protocol, that moves the effort from the statistics to the recruitment.*

**SOURCE** Artstein & Poesio (2008), *Inter-coder agreement for computational linguistics*, Computational Linguistics, doi:10.1162/coli.07-034-r2, §5.2 Choosing a Coefficient

**NAMES** Artstein · Poesio

**MARKS** working method · D1.10 prescriptive-policy · used

**RELATIONS** refines → C019

**ONLINE** https://claudiocammarano.com/en/claims/c018/

---

**C019** This is Mode A of Coding Manual v2.1, and Mode A carries a mandatory check. **A human recodes a sample by hand, blind to the machine, and the two must agree above 90%.**

> *A machine-produced verdict counts here only after a person has recoded a sample by hand, without seeing what the machine decided, and the two agree above ninety per cent. The threshold is what turns an output into evidence.*

**SOURCE** Lab note, https://claudiocammarano.com/lab/2026-07-21-what-counts-as-a-verdict/, ¶2

**MARKS** working method · D1.10 prescriptive-policy · used

**ONLINE** https://claudiocammarano.com/en/claims/c019/

**ALSO HERE** C020, given under Convergent paths.

## Verdict

**What has to hold before a judgement on an output counts?**

A judgement on an output is worth no more than the procedure that produced it, and the procedure has to be inspectable.

**LINKED TO** Coder agreement · Machine judge · Claim type · Declared incompleteness

**ALSO HERE** C019, given under Coder agreement.

**C021** We show that some of the biases are minor or can be mitigated. **Once addressed, our results from 3K controlled expert votes and 3K crowdsourced human votes in the wild verify that GPT-4 judge match human evaluations at an agreement rate exceeding 80%, achieving the same level of human-human agreement (4.2, Table 4).** Consequently, this suggests LLM-as-a-judge is a scalable method to swiftly evaluate human preference, serving as a promising alternative to traditional human evaluations.

> *Asked to rank answers, a strong model agreed with human raters about as often as human raters agreed with each other, above eighty per cent. What was measured is a rate of agreement, which is not the same thing as competence to judge.*

SOURCE Zheng & Chiang & Sheng & Zhuang & Wu & Zhuang & Lin & Li & Li & Xing & Zhang & Gonzalez & Stoica (2023), *Judging LLM-as-a-judge with MT-bench and Chatbot Arena*, Advances in Neural Information Processing Systems, doi:10.52202/075280-2020, §1 Introduction

MARKS corrective datum · D1.2 empirical-quantitative · contested · contested

RELATIONS contests → C019

ONLINE https://claudiocammarano.com/en/claims/c021/

**C022** A framework whose central instrument is an Epistemic Responsibility Check cannot exempt itself from the responsibility it checks. **Reproducibility is a property of documentation, not of results: if a third party cannot recompute AR-sem = 78.33% from the deposited materials, the number is testimony, not evidence.** Hence this release.

> *A number nobody else can recompute from the deposited materials is testimony, not evidence. A framework whose instrument is an epistemic responsibility check cannot exempt itself from that check, which is the reason the materials are deposited at all.*

SOURCE Lab note, https://claudiocammarano.com/lab/2026-07-10-coding-manual-v2-1-retrospective-qc/, ¶2

MARKS working method · D1.10 prescriptive-policy · used

RELATIONS refines → C019

ONLINE https://claudiocammarano.com/en/claims/c022/

## Machine judge

**Can a model stand in for a human rater?**

A model can stand in for a human rater at a measurable rate of agreement; the rate is the claim, not the competence.

LINKED TO Verdict · Inverse scaling

ALSO HERE C021, given under Verdict.

**C023** We also aim to develop methods that encode desirable AI behavior in a simple and transparent form, and that make it easier to understand and evaluate AI decision making. **In this paper we develop a method we refer to as Constitutional AI (CAI), depicted in Figure 1, and use it to train a non-evasive and relatively harmless AI assistant, without any human feedback labels for harms. The method therefore improves upon, and partially replaces reinforcement learning from human feedback [...].** The new assistant 'RL-CAI' is

preferred by crowdworkers over those trained with previously collected [Bai et al., 2022, Ganguli et al., 2022] human feedback labels for harmfulness.

> *Harmlessness was trained by having the model criticise and revise its own answers against a written list of principles, instead of having people label every harmful output. Human judgement moves out of the labels and into the rules.*

**SOURCE** Bai & others (2022), *Constitutional AI: Harmlessness from AI Feedback*, arXiv preprint arXiv:2212.08073, §1 Introduction

**NAMES** Bai

**MARKS** working method · D1.9 procedural-instructional · used

**ONLINE** https://claudiocammarano.com/en/claims/c023/

## Measurement artefact

**When does a jump in a curve belong to the metric rather than to the thing measured?**

A discontinuity in a curve can belong to the metric rather than to the thing measured: a finer metric dissolves it.

**LINKED TO** Declared incompleteness · Claim type · Convergent paths

---

**C024** **Breakthrough behavior is consistent with the model suddenly gaining new skills in an abrupt way. Careful analysis of task behavior, however, suggests that the underlying change in model capabilities is generally more smooth.** These smooth improvements can be exposed by decomposing tasks in ways that better capture partial progress. We describe some of those ways in the remainder of this section.

> *Abilities that look as though they switch on all at once with scale often improve smoothly underneath: the jump belongs to an all-or-nothing metric. Change the metric and the cliff turns back into a slope.*

**SOURCE** Srivastava & others (2023), *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*, Transactions on Machine Learning Research, §3.4.2

**NAMES** Srivastava

**MARKS** corrective datum · D1.2 empirical-quantitative · assumed

**ONLINE** https://claudiocammarano.com/en/claims/c024/

**ALSO HERE** C020, given under Convergent paths.

**ALSO HERE** C005, given under Claim type.

## Inverse scaling

**Does a larger model get better at everything?**

More parameters do not buy more of everything: on some properties the larger model is the worse one.

**LINKED TO** Imitative falsehood · Measurement artefact

---

**C026** **Across different model families, the largest models were generally less truthful (Fig.**

2). **This "inverse scaling" trend contrasts with most tasks in NLP, where performance improves with model size [...].** One explanation of this result is that larger models produce more imitative falsehoods because they are better at learning the training distribution. Another explanation is that our questions adversarially exploit weaknesses in larger models not arising from imitation of the training distribution.

> *On questions written to echo common misconceptions, the larger models answered falsely more often than the smaller ones, reversing the usual pattern. Scale is not a general-purpose fix, and on some properties it is the problem.*

**SOURCE** Lin & Hilton & Evans (2022), *TruthfulQA: Measuring how models mimic human falsehoods*, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, doi:10.18653/v1/2022.acl-long.229, §1 Introduction

**MARKS** corrective datum · D1.2 empirical-quantitative · assumed

**ONLINE** https://claudiocammarano.com/en/claims/c026/

**ALSO HERE** C024, given under Measurement artefact.

## Declared incompleteness

**Does an evaluation have to publish what it did not measure?**

An evaluation is honest when it states the distance between what it aspires to measure and what it actually measures.

**LINKED TO** Measurement artefact · Verdict

---

**C027** Different benchmarks make different decisions on what to prioritize, how to make these decisions, and to what extent these processes are made clear in presenting the benchmark. **Since our aim is holistic evaluation, we believe it is necessary to be explicit on the relationship between what we aspire to evaluate and what we actually evaluate.** The construction of HELM starts top-down with a taxonomy over scenarios and metrics (see Figure 2). The taxonomy not only facilitates the systematic selection of scenarios and metrics, but it also make explicit what is missing.

> *An evaluation should publish the distance between what it set out to measure and what it actually measured. Stating what was left out counts as part of the result, not as an apology for it.*

**SOURCE** Liang & others (2022), *Holistic Evaluation of Language Models*, arXiv preprint arXiv:2211.09110, §2, Recognition of incompleteness

**NAMES** Liang

**MARKS** working method · D1.10 prescriptive-policy · used

**ONLINE** https://claudiocammarano.com/en/claims/c027/

**ALSO HERE** C022, given under Verdict.

# Part III · What scale does not buy

*Size improves many things monotonically and some things not at all.*

## Imitative falsehood

**Can a model be wrong because it learned its training data well?**

A model can be wrong because it learned the training distribution well, not because it learned it badly.

LINKED TO Inverse scaling · Form–meaning gap

ALSO HERE C026, given under Inverse scaling.

---

**C028** Similar to psychological hallucination, which is hard to tell apart from other "real" perceptions, hallucinated text is also hard to capture at first glance. **Within the context of NLP, the above definition of hallucination, the generated content that is nonsensical or unfaithful to the provided source content [...], is the most inclusive and standard.** However, there do exist variations in definition across NLG tasks, which will be further described in the later task-specific sections.

> *Across the field, hallucination means generated content that is nonsensical or unfaithful to the source it was given: a definition about fidelity to the input, not about truth in the world. It is the broadest one in use and the one that travels best between tasks.*

SOURCE Ji & Lee & Frieske & Yu & Su & Xu & Ishii & Bang & Madotto & Fung (2022), *Survey of hallucination in natural language generation*, ACM Computing Surveys, doi:10.1145/3571730, §2

MARKS conceptual tool · D1.6 interpretive-hermeneutic · used

ONLINE https://claudiocammarano.com/en/claims/c028/

## Sequence without recurrence

**What changed when processing a sequence stopped being step by step?**

Relations between positions in a sequence can be computed all at once instead of step by step, which is what made the scale possible.

LINKED TO Form–meaning gap

---

**C029** End-to-end memory networks are based on a recurrent attention mechanism instead of sequence-aligned recurrence and have been shown to perform well on simple-language question answering and language modeling tasks [34]. **To the best of our knowledge, however, the Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution.** In the following sections, we will describe the Transformer, motivate self-attention and discuss its advantages over models such as [17, 18] and [9].

> *The architecture behind current language models dropped step-by-step processing of a sequence in favour of computing every relation between positions at once. It is the engineering fact the rest of the debate stands on.*

SOURCE Vaswani & Shazeer & Parmar & Uszkoreit & Jones & Gomez & Kaiser &

Polosukhin (2017), *Attention Is All You Need*, Advances in Neural Information Processing Systems, §2 Background

**NAMES** Vaswani

**MARKS** technical reference · D1.1 historical-factual · assumed

**ONLINE** https://claudiocammarano.com/en/claims/c029/

# Sources

Andreas, Jacob (2022). *Language Models as Agent Models*. Advances in Neural Information Processing Systems. doi:10.18653/v1/2022.findings-emnlp.423. C006

Artstein, Ron and Poesio, Massimo (2008). *Inter-coder agreement for computational linguistics*. Computational Linguistics. doi:10.1162/coli.07-034-r2. C018

Bai, Yuntao and others (2022). *Constitutional AI: Harmlessness from AI Feedback*. arXiv preprint arXiv:2212.08073. C023

Bender, Emily M. and Koller, Alexander (2020). *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.463. C008

Ji, Ziwei and Lee, Nayeon and Frieske, Rita and Yu, Tiezheng and Su, Dan and Xu, Yan and Ishii, Etsuko and Bang, Yejin and Madotto, Andrea and Fung, Pascale (2022). *Survey of hallucination in natural language generation*. ACM Computing Surveys. doi:10.1145/3571730. C028

Lakatos, Imre (1978). *The Methodology of Scientific Research Programmes*. Cambridge University Press. C046, C044, C045, C049, C050, C051, C052, C053, C054

Liang, Percy and others (2022). *Holistic Evaluation of Language Models*. arXiv preprint arXiv:2211.09110. C027

Lin, Stephanie and Hilton, Jacob and Evans, Owain (2022). *TruthfulQA: Measuring how models mimic human falsehoods*. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. doi:10.18653/v1/2022.acl-long.229. C026

Mayo, Deborah G. (1996). *Error and the Growth of Experimental Knowledge*. University of Chicago Press. C033, C034, C035

Mayo, Deborah G. (2005). *Peircean Induction and the Error-Correcting Thesis*. Transactions of the Charles S. Peirce Society. C037

Mayo, Deborah G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge University Press. doi:10.1017/9781107286184. C012, C013, C014, C015, C016, C017, C030, C031, C032, C036

Mitchell, Melanie and Krakauer, David C. (2023). *The Debate Over Understanding in AI's Large Language Models*. Proceedings of the National Academy of Sciences. doi:10.1073/pnas.2215907120. C007

Peirce, Charles Sanders (1931–1958). *Collected Papers of Charles Sanders Peirce*. Harvard University Press. C038, C039, C040, C041, C042, C043

Popper, Karl R. (1959). *The Logic of Scientific Discovery*. Hutchinson. doi:10.4324/9780203994627. C047, C048

Srivastava, Aarohi and others (2023). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*. Transactions on Machine Learning Research. C024

Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N. and Kaiser, Lukasz and Polosukhin, Illia (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems. C029

Wang, Xuezhi and Wei, Jason and Schuurmans, Dale and Le, Quoc and Chi, Ed and Narang, Sharan and Chowdhery, Aakanksha and Zhou, Denny (2022). *Self-consistency improves chain of*

*thought reasoning in language models*. C025

Zheng, Lianmin and Chiang, Wei-Lin and Sheng, Ying and Zhuang, Siyuan and Wu, Zhanghao and Zhuang, Yonghao and Lin, Zi and Li, Zhuohan and Li, Dacheng and Xing, Eric P. and Zhang, Hao and Gonzalez, Joseph E. and Stoica, Ion (2023). *Judging LLM-as-a-judge with MT-bench and Chatbot Arena*. Advances in Neural Information Processing Systems. doi:10.52202/075280-2020. C021

Lab notes on claudiocammarano.com: C001, C002, C003, C004, C005, C009, C010, C011, C020, C019, C022.

---

**Compiled, not written.** This edition was generated on 17 November 2026 from the site's data (`src/_data/claims.js`, `claimLayers.js`): 54 claims, 15 concepts, 3 questions. No sentence is added at compile time. Data fingerprint: `6e64fbf60e50`.

Numbers are permanent: a claim keeps its number in every edition, and a new claim takes the next free one. Only the latest edition is kept.

Quoted passages belong to their authors and are given with source and locus. Notation: ENTIA 1.1.0, https://doi.org/10.5281/zenodo.23303451.

© Claudio Cammarano. Licence CC BY-NC-ND 4.0 for the apparatus: share with attribution, not for commercial use, without changes.

Original and current version: https://claudiocammarano.com/en/claims/. The map and the filters are interactive only online.

I also write *The Abstract*, in English, once a month, on the content economy: https://claudiocammarano.com/newsletter/.