Quantifying Reasoning: From Theory to Measurement

Jack Hodges, Zach Glasser

May 2026

Abstract

No system currently exists to quantify the quality of human reasoning from natural language at scale. Adjacent capabilities have been developed; large language model (LLM) reasoning benchmarks evaluate machine inference, argumentation quality frameworks assess persuasive structure, and polling instruments measure opinion, but none address the underlying construct: how well a person forms and revises beliefs in proportion to evidence. This paper presents a reasoning evaluation pipeline designed to fill that gap. We first establish a philosophically grounded, operationalisable definition of reasoning quality, synthesising traditions from Aristotelian formal logic through classical Indian epistemology to Bayesian belief management. From this definition we derive five measurable dimensions; accuracy, calibration, consistency, evidence engagement, and updating behaviour, and review the technical literature establishing why existing natural language processing approaches are insufficient to capture them. We then propose a hybrid pipeline architecture operating in two evaluation modes: a direct LLM-as-judge pathway for short-form responses scored against a structured rubric, and a compressed implicit premise reconstruction pathway for long-form argument that extracts reasoning structure before scoring. The pipeline is designed to operate across three input contexts; structured game environments, live speech and debate, and published written argument, with dimension weighting varying by context and observability. We discuss the core technical challenges of implicit reasoning reconstruction, evaluator bias, and system validation, and present a framework for empirical testing against human-annotated benchmarks. The contribution is threefold: a cross-civilisational philosophical foundation for operationalising reasoning quality, the first pipeline architecture designed to evaluate human rather than machine reasoning from natural language, and a hybrid evaluation approach that balances analytical depth against computational feasibility at scale.

1. Introduction

The question of how well a person reasons is among the oldest in intellectual history. It is also, in 2026, among the least technically addressed. There exist sophisticated systems for measuring what people believe (polling), how persuasively they argue (argumentation quality assessment), and how accurately machines perform logical inference (LLM reasoning benchmarks). What does not exist is a system for measuring the quality of the process by which a human being moves from evidence to belief the reasoning itself, as distinct from its inputs, outputs, or rhetorical packaging.

This gap is not accidental. Reasoning quality is a latent construct: it cannot be directly observed, only inferred from imperfect and context-bound signals.[1] A written argument reveals something about the reasoning that produced it, but the relationship between text and thought is lossy. Premises are omitted, inferential steps are compressed, confidence is performed rather than reported, and the same conclusion can be reached through radically different chains of inference. Any system that claims to evaluate reasoning quality from natural language must confront these difficulties honestly, and the history of adjacent fields suggests that most have not.

LLM reasoning benchmarks such as Big-Bench Hard evaluate whether a model can produce a correct chain of thought before generating an answer.[2] This is a test of explicit deductive competence, not a measure of reasoning quality in the sense that matters for human judgement, where the dominant mode is abductive (inference to the best explanation under incomplete information), where confidence must be calibrated against uncertainty, and where active engagement with available evidence is as important as the logical structure of the argument. Argumentation quality research, meanwhile, has focused heavily on persuasive effectiveness and rhetorical structure.[3] A convincing argument and a well-reasoned one are not the same thing, and a framework optimised for the former may be actively misleading about the latter.

This paper proposes a reasoning evaluation pipeline: an architecture for quantifying human reasoning quality from natural language, grounded in a definition of reasoning that is philosophically defensible, empirically tractable, and resistant to gaming. The pipeline is designed to operate across three input contexts, structured interactive environments where game mechanics force reasoning to the surface, live speech and debate where reasoning is produced

under time pressure, and published written argument where an author has deliberately constructed a case, with evaluation depth and dimension weighting varying by context.

The contribution is threefold. First, we present an operationalisable definition of reasoning quality synthesised from cross-civilisational philosophical traditions, producing five measurable dimensions: accuracy, calibration, consistency, evidence engagement, and updating behaviour. Second, we review the technical literature establishing why existing natural language processing approaches, including natural language inference, direct LLM-as-judge evaluation, and argumentation mining, are individually insufficient for this task, while identifying the components from each that a viable pipeline must integrate. Third, we propose a hybrid pipeline architecture that operates in two evaluation modes: a direct structured evaluation for short-form inputs and a compressed implicit premise reconstruction pathway for long-form argument, both feeding into a longitudinal reasoning profile that captures dimensions observable only over time.

The paper proceeds as follows. Section 2 establishes the philosophical foundations for the definition of reasoning quality and derives the five scoring dimensions from them. Section 3 reviews the technical landscape and identifies the evaluation gap. Section 4 presents the pipeline architecture in detail. Section 5 describes the proposed validation framework. Section 6 discusses limitations and open problems.

2. What Is Reasoning? Philosophical Foundations

Before a system can claim to measure reasoning, it must settle what reasoning is. This is not a preliminary question. Every scoring dimension, every weighting decision, and every architectural choice in the pipeline that follows rests on the definition adopted here. A technically sophisticated apparatus built on a poorly specified target is not measuring reasoning - it is measuring something else with impressive infrastructure.

The difficulty is that reasoning has no agreed definition. It has been contested across at least 2,500 years of philosophy, logic, psychology, and epistemology, and not only within the Western tradition. Thinkers working independently in ancient Greece, classical India, and early China each arrived at the problem from different directions, and each tradition isolates something the others underweight. That these three civilisations, which had no contact with each other,

independently developed formal theories of inference is itself instructive. It suggests that the pressure to understand the movement from evidence to conclusion is not culturally contingent but something that thinking creatures encounter whenever they try to organise knowledge, settle disputes, or govern complex societies.

What follows is not a comprehensive intellectual history but a targeted account: each tradition and thinker is included because they ground a specific design decision in the pipeline, and nothing is included that does not.

2.1 Form, Warrant, and Framing: Three Classical Starting Points

Aristotle's contribution is foundational in the most literal sense: without it, there is nothing to quantify. In the *Prior Analytics*, he gives the first systematic account of valid deduction and establishes the claim that validity depends on form, not subject matter.[4] The syllogism 'All humans are mortal; Socrates is human; therefore Socrates is mortal' works not because of anything specific about Socrates or mortality but because of its structure. Substitute any terms, all planets orbit stars; Mars is a planet; therefore Mars orbits a star, and the same form guarantees the same result. This is the philosophical precondition for any measurement project. If reasoning were simply a flow of impressions with no evaluable structure, there would be no property to capture. Aristotle established that there is.

The Stoics, particularly Chrysippus, extended this by developing propositional logic, reasoning about connections between complete claims rather than about class membership.[5] Where Aristotle's system operates on terms ('All A are B'), Stoic logic operates on entire propositions and their truth-functional connections: 'If p then q; p; therefore q.' The modern conditional - the 'if…then' that structures legal argument, scientific hypothesis, and everyday planning, descends more directly from the Stoics than from Aristotle's syllogistic. Together, they establish the principle that underwrites the pipeline's existence: reasoning has structure, and some structures are better than others.

The classical Indian tradition takes a different and, for the purposes of this paper, equally important starting point. The Nyāya school, one of the six orthodox Hindu philosophical systems, embeds reasoning within a broader theory of *pramāṇa*, the valid means of acquiring

knowledge.[6] Inference (*anumāna*) is one such means, alongside perception, comparison, and testimony. The canonical example is: 'There is fire on the mountain, because there is smoke there, as in a kitchen.' The underlying concern is not primarily whether the argument is valid in form. It is whether the sign, the smoke - stands in a reliable evidential relationship to the probandum - the fire.

This shift from the Greek question ('is this argument formally valid?') to the Indian question ('does this sign reliably track the fact?') is not a minor variation. It is a different emphasis entirely. Bimal Krishna Matilal demonstrated that classical Indian epistemology is rigorous in a sense fully continuous with contemporary analytic philosophy, employing reductio ad absurdum arguments, systematic fallacy classification, and a sophisticated theory of inferential reliability.[7] Jonardon Ganeri extended this into contemporary epistemology, arguing that the *pramāṇa* tradition places reliability - does this inferential process actually produce true beliefs? at the centre of the account, rather than formal validity alone.[8] A system that only measures whether an argument is logically watertight is measuring something the Indian tradition correctly identifies as insufficient. The process must reliably track truth, not just satisfy formal rules. This is the philosophical root of the pipeline's *accuracy* dimension - the requirement that beliefs track available evidence and that the connection between evidence and conclusion is substantively reliable, not merely formally correct.

The Mohist school in classical China contributes a third emphasis that modern cognitive science has rediscovered with considerable force.[9] The Mohists were preoccupied with a question that precedes inference: are the categories correct? Is the comparison being drawn actually valid? Does the concept extend legitimately to the new case? This is a question about framing, not logic. How a problem is presented, what it treats as equivalent, and what distinctions it draws shapes the reasoning that follows, sometimes more than the logical structure of the argument itself. If a problem is categorised wrongly at the outset, no amount of correct inference on top of it will save the conclusion. This observation grounds the pipeline's *consistency* dimension: the requirement that judgement holds across contexts that vary in framing but not in underlying structure. A reasoner who reaches different conclusions on structurally identical problems presented under different frames is exhibiting a measurable

failure, and the Mohist tradition identified this failure before Western cognitive science catalogued it empirically.

2.2 Bias, Uncertainty, and the Limits of Coherence

Francis Bacon's contribution is diagnostic rather than constructive. Writing in 1620, he catalogued the systematic ways the human mind reliably destroys its own reasoning, calling them 'idols of the mind': the idol of the cave (personal history warps perception), the idol of the marketplace (language misleads), the idol of the theatre (received wisdom blinds).[10] This is the first serious taxonomy of cognitive bias, anticipating by three centuries what Kahneman and Tversky would prove empirically. For the pipeline, Bacon's insight is a design constraint: structured evaluation environments must force commitment before outcomes are revealed, not to create pressure, but to prevent the retrospective rationalisation that makes bias invisible. If a user sees the answer before committing to a position, what is measured is recognition, not reasoning.

David Hume identified a deeper problem that remains unsolved. The sun has risen every day in recorded history. We conclude it will rise tomorrow. But the conclusion does not follow logically from the premise, and any attempt to justify the inference by pointing to its past success is circular.[11] Our confidence that the future will resemble the past is grounded in habit, not in reason. Hume's contribution to the pipeline is the insight that certainty is never available to a reasoner operating in the empirical world. Every conclusion is held provisionally. This is why the pipeline measures *calibration*, the match between stated confidence and actual accuracy rather than correctness alone. Correctness is partly luck. A person who is confidently right by chance and a person who is tentatively right because they tracked the evidence carefully have produced the same output. Only calibration, measured over repeated judgements, distinguishes them.

Kant's response to Hume contains a warning the pipeline takes seriously. He argued that causal structure is not learned from experience but is built into how the mind organises experience, a precondition of coherent thought, not a conclusion derived from it.[12] What both Hume and Kant point toward, from opposite directions, is that the human mind has a deep, compulsive drive to make things cohere, to close gaps, to find explanations that feel settled and complete. That drive is exactly as likely to produce rationalisation as true insight. A mind that feels certain it has understood something may have reasoned its way carefully to the truth, or

may simply have found a story that satisfies the craving for closure. Those two states feel identical from the inside. This is why consistency, in the pipeline's scoring framework, is a dimension to be interrogated rather than simply rewarded. A person who reasons consistently may be consistently rationalising. The pipeline must be designed to distinguish the two, and the Kantian insight is that the distinction requires convergent evidence from multiple dimensions, not consistency data alone.

2.3 Abduction, Bounded Rationality, and Measurable Variation

The nineteenth-century formalisation of logic by Boole and Frege captured deduction with extraordinary precision but said almost nothing about the kinds of reasoning that dominate actual human life.[13,14] C.S. Peirce corrected this by identifying abduction - inference to the best available explanation as a distinct and irreducible mode of reasoning.[15] The grass is wet; it probably rained. The patient presents with these symptoms; the most likely diagnosis is this. These inferences are not deductively valid, alternative explanations always exist but they are recognisably rational. Abduction is reasoning under incomplete information: given what I currently know, what is the most defensible position to hold? Crucially, it is open to revision when new evidence arrives. This is the epistemic structure of most real-world judgement, and it is the structure that a reasoning evaluation pipeline must be designed to assess. A system that only measures deductive competence measures a vanishingly small fraction of the reasoning people actually do.

Herbert Simon established the conditions under which all reasoning actually occurs: bounded rationality.[16] No reasoner has ever had unlimited time, memory, or information. Real people satisfice, they find solutions that are good enough under the constraints they face, using heuristics that trade accuracy for speed. This is not degraded reasoning. It is the only kind that exists. A pipeline that penalises cognitive limits is penalising a feature of human cognition, not a failure. The implication is that evaluation must assess how well a person reasons *within* their constraints, not against an idealised standard that no human could meet.

Daniel Kahneman and Amos Tversky then demonstrated that the heuristics Simon described produce systematic, predictable errors, consistent across individuals and contexts.[17] That consistency is the empirical licence for the entire enterprise. If reasoning errors were random

noise, there would be nothing stable to measure at the individual level. The fact that they are patterned means that reasoning quality varies between people in ways a scoring system can, in principle, detect. Without this finding, there would be no stable property to capture and no score worth building. Kahneman's dual process framework - the distinction between fast, intuitive System 1 processing and slow, deliberate System 2 processing offers a useful heuristic for understanding these patterns but is too crude to serve as a measurement framework in its own right.[18] Expert intuition, which is fast, regularly outperforms careful deliberation, which is slow. Speed alone tells the pipeline nothing reliable about reasoning quality. What matters is the combination of accuracy, confidence, evidence engagement, and how a person revises their position when new information arrives.

Philip Tetlock's superforecaster research provided the applied confirmation: some individuals are consistently better calibrated than others, the difference is measurable across many judgements over time, and it improves with practice.[19] This is not a theoretical finding. It is an empirical demonstration that calibration. The match between confidence and accuracy is a learnable, trackable, individual-level property. It directly underwrites the pipeline's calibration dimension and its longitudinal scoring architecture.

2.4 The Bayesian Turn and the Working Definition

The most consequential modern redefinition of reasoning shifts the question entirely. Stop asking whether a conclusion follows necessarily from premises. Start asking: given what you know, how confident should you be, and how should that confidence change when new information arrives? This is the Bayesian view, and it provides the mathematical framework within which the pipeline's dimensions become formally expressible.[20]

Bayesian reasoning is updating correctly: revising beliefs in proportion to what the evidence actually warrants, neither dismissing new information nor overcorrecting in response to it. A doctor who estimates a twenty percent probability that a patient has a disease, receives a positive test result, and revises the estimate appropriately given the test's sensitivity and specificity is reasoning well, not because the conclusion is correct (it may not be), but because the update is proportional. Two failure modes are clearly separable: under-updating, where a person anchors too stubbornly to a prior position when evidence should move them, and over-updating, where

they overcorrect in response to information that does not warrant it.[21] Both are detectable by tracking how stated positions change when new evidence is introduced. The pipeline's *updating behaviour* dimension captures precisely this.

The Brier score makes part of this framework measurable. It calculates the squared distance between a stated probability and the eventual outcome, rewarding accurate confidence and penalising confident error. Across repeated forecasts, it provides an established measure of probabilistic accuracy and calibration. It does not, however, measure reasoning in full: it cannot show whether a forecast resulted from sound evidence engagement, consistent judgement or luck. The pipeline therefore treats it as one instrument for measuring accuracy and calibration alongside separate measures of updating, consistency and evidence engagement.

One recent theory cuts against the tradition that precedes it. Mercier and Sperber argued that reasoning did not evolve to help individuals find truth but to win arguments in social settings.[22] The theory is probably overstated, but the underlying observation is real: the context in which a person reasons changes what their reasoning produces. This has a direct design implication. The pipeline's structured evaluation environments are deliberately designed to eliminate the features of social reasoning that corrupt epistemic orientation. There is no audience to perform for, no opponent to defeat, and no reward for confidence or eloquence. What is measured is how a person's private commitments track evidence, not how effectively they persuade.

Synthesising across these traditions, the definition that has proven most empirically tractable, most resistant to gaming, and most defensible across the frameworks surveyed is this:

> *Reasoning is the process of forming and revising beliefs in proportion to evidence, in ways that are consistent across contexts, under conditions of uncertainty, in the service of understanding the world as it actually is.*

This definition rules out intelligence as a target. It rules out domain knowledge. It rules out the correctness of conclusions taken in isolation. It targets the quality of the process. Five measurable dimensions follow directly from it, each grounded in a specific philosophical tradition and each operationalisable through the pipeline architecture described in Section 4:

Accuracy. Whether beliefs track facts. Grounded in the Nyāya epistemic warrant tradition and Ganeri's reliability condition: the inferential process must actually produce true beliefs, not merely satisfy formal rules.

Calibration. Whether stated or implied confidence matches actual accuracy over repeated judgements. Grounded in Hume's insight that certainty is unavailable in empirical reasoning and Tetlock's empirical demonstration that calibration is a measurable, individually variable property.

Consistency. Whether judgement holds across contexts that vary in framing but not in underlying structure. Grounded in the Mohist emphasis on correct categorisation and the Kantian warning that coherence alone does not guarantee sound reasoning.

Evidence engagement. Whether the person actively engages with relevant information before forming a view, rather than reasoning from prior conviction alone. Grounded in Bacon's diagnostic insight that the primary enemy of good reasoning is the mind's tendency toward premature closure and motivated interpretation.

Updating behaviour. Whether revision is proportional to new evidence, neither anchoring to prior positions nor overcorrecting. Grounded in the Bayesian tradition and operationalised through Tetlock's forecasting methodology.

Not all five dimensions are equally observable in every evaluation context. Calibration and updating behaviour require repeated judgements over time and are therefore longitudinal measures, computable at the profile level rather than the individual response level. Accuracy, consistency, and evidence engagement can be assessed from a single response where the input provides sufficient structure. Section 4 details how each dimension is scored across the pipeline's three input modes and how the longitudinal reasoning profile accumulates evidence for dimensions that are not assessable from individual responses alone.

These five dimensions are the observable manifestations of a definition built from converging arguments across Greek, Indian, Chinese, and modern empirical traditions. From Aristotle we inherited the form: reasoning has evaluable structure. From the Indian *pramāṇa* tradition we inherited the reliability condition: formal validity is not enough; the process must actually track the world. From the Mohists we inherited the framing caution: how a problem is

categorised before inference begins matters as much as the inference itself. From Bacon, Hume, and Kant we inherited the diagnostic constraints: bias corrupts, certainty is unavailable, and coherence can mask rationalisation. From Simon, Kahneman, Tversky, and Tetlock we inherited the empirical licence: reasoning quality varies between individuals in measurable, reproducible ways. And from the Bayesian tradition we inherited the operational framework: reasoning quality is visible in the relationship between a person's beliefs and the evidence available to them, tracked over time. The remaining question is how to measure it.

3. Related Work and the Evaluation Gap

This section examines why the capability described in Section 2, quantifying human reasoning quality from natural language, does not yet exist, despite substantial progress in adjacent fields. The gap is not a matter of insufficient effort. It reflects a set of genuine technical challenges that existing approaches address only partially, and in some cases not at all. We organise the discussion around three questions: what makes the problem hard (3.1), what existing approaches achieve and where they fall short (3.2), and what components a viable pipeline must therefore integrate (3.3).

3.1 Core Technical Challenges

3.1.1 The Reconstruction Problem

The central difficulty in evaluating reasoning from text is that reasoning is mostly invisible. What a person writes or says is the output of a reasoning process, not the process itself. Premises are omitted, inferential steps are compressed, assumptions remain unstated, and the same conclusion can be reached through radically different chains of inference. The text is a lossy compression of the reasoning that produced it.

Any evaluator must therefore reconstruct, or at least approximate, the hidden reasoning structure. Habernal et al. demonstrated the scale of this problem in their work on implicit warrant identification, showing that human arguments routinely depend on unstated bridging assumptions that connect stated reasons to stated conclusions.[23] The difficulty is not merely that

these assumptions are hidden. It is that multiple plausible reconstructions exist for any given text, and the evaluator cannot know which one the author actually relied on.

Turpin et al. sharpened this concern by demonstrating that LLMs, when asked to reconstruct reasoning chains, tend to produce the chain that best justifies the stated conclusion rather than the chain most likely to have been used.[24] This is a systematic bias in reconstruction, not random noise. A pipeline that treats LLM-reconstructed chains as ground truth is evaluating a fictional version of someone's reasoning and calling it measurement. Lanham et al. reinforced this finding, arguing that the faithfulness of model-generated reasoning explanations to the actual process cannot be assumed and must be independently tested.[25]

The reconstruction problem is not solvable in the general case. It is, however, *mitigable* through three strategies that the proposed pipeline employs. First, environment design: structured evaluation contexts can force more of the reasoning process to the surface by requiring commitment before revelation, eliciting confidence levels, and probing reasoning through follow-up interaction. Second, treating the user's text as primary evidence and LLM reconstruction as a query layer rather than ground truth: the reconstruction illuminates the text but does not replace it as the object of evaluation. Third, longitudinal aggregation: a single response's hidden reasoning may be irrecoverable, but patterns across many responses reveal stable characteristics of how a person reasons, reducing dependence on any individual reconstruction.

3.1.2 Calibration and the Limits of One-Shot Judgement

Calibration, the match between stated confidence and actual accuracy, is among the strongest indicators of reasoning quality in the forecasting and decision science literature. Gneiting and Raftery formalised the relationship between stated confidence and observed outcomes, establishing calibration as a mathematically precise property of a sequence of probabilistic judgements.[26] Tetlock's superforecaster research demonstrated that calibration varies meaningfully between individuals and improves with practice.[19]

The difficulty for a text-based evaluation system is that calibration cannot be reliably inferred from a single response. A written answer may contain linguistic signals of confidence, hedging, qualifying, asserting but these are signals of expression style, not of actual epistemic calibration.

A cautious-sounding user may appear well calibrated while consistently being wrong. A confident user may sound overconfident while being right at the appropriate rate. Calibration is a property of a *sequence* of judgements with resolvable outcomes, not a property of a single text.

This has a direct architectural implication: the pipeline must maintain a longitudinal reasoning profile for each user, accumulating calibration data across multiple evaluation events. Calibration cannot be meaningfully weighted in any single-response evaluation. It becomes scorable only at the profile level, after sufficient repeated judgements have been recorded. The same constraint applies, though less severely, to updating behaviour, which requires at least a short sequence of related judgements to assess whether revision is proportional to new evidence.

3.1.3 Evaluator Bias and Instability

Even if the reconstruction problem were solved perfectly, a second problem remains: the evaluator itself. LLMs used as judges are influenced by features of the input that are irrelevant to reasoning quality. Chen et al. documented that LLM evaluators are susceptible to authority effects, stylistic preferences, and order effects that shift evaluations without changing the quality of the underlying reasoning.[27] Koo et al. extended this analysis using the CoBBLEr framework, categorising LLM judge biases into implicit biases (extracted from single-prompt evaluations) and induced biases (arising from downstream influence of conversational context), and demonstrating that both produce systematic distortions in evaluation output.[28]

The instability of LLM-as-judge evaluation is equally concerning. Presenting the same two responses in different order can change which is preferred. Running the same evaluation twice with identical inputs can produce different scores. These are not edge cases; they are structural properties of probabilistic language models used in evaluation roles. Any pipeline that relies on a single LLM judgement call without mitigation is producing scores that contain substantial evaluator noise.

3.2 Adjacent Approaches and Their Limitations

Several existing approaches address parts of the problem. None addresses the whole.

Natural language inference (NLI) frameworks compute entailment, contradiction, or neutrality between sentence pairs. Large benchmark datasets such as MultiNLI established this as a central task in sentence-level NLP.[29] However, NLI operates on isolated premise-hypothesis pairs, cannot assess the broader reasoning process across a full response, and was shown by McCoy et al. to rely frequently on shallow lexical heuristics rather than genuine semantic inference.[30] By 2026, with LLMs capable of handling substantially more complex evaluation tasks natively, NLI as a standalone approach to reasoning assessment is largely superseded.

Direct LLM-as-judge evaluation, sending a user's response and a scoring rubric to a language model and receiving a structured score is flexible, easy to deploy, and does not require labelled training data. Mirzakhmedova et al. showed that LLM judges can achieve high agreement with human annotators on structured evaluation tasks.[31] However, this approach inherits the bias and instability problems documented by Chen et al. and Koo et al., and it evaluates the surface text rather than the underlying reasoning structure. For short-form responses where the reasoning is relatively transparent, direct evaluation may be sufficient. For longer, more complex arguments where critical reasoning moves are implicit, it is not.

Implicit premise reconstruction, as formalised by Habernal et al., offers a route to evaluating the hidden structure of arguments by identifying unstated warrants and testing their robustness.[23] Stahl et al. further distinguished between identifying that a gap exists in an argument and correctly reconstructing what should fill it, showing these are separate problems with different failure modes.[32] This approach is analytically powerful but computationally expensive: a full reconstruction pipeline involves multiple sequential LLM calls, each dependent on the output of the previous step. At scale, this creates a tension between evaluation depth and feasibility that the pipeline architecture must explicitly manage.

3.3 What a Viable Pipeline Must Integrate

The literature review identifies five requirements that no single existing approach satisfies but that a viable reasoning evaluation pipeline must integrate:

First, the pipeline must handle the reconstruction problem without treating LLM-generated reasoning chains as ground truth. The user's text must remain the primary evidence; reconstruction is a query layer that illuminates structure but does not replace it.

Second, the pipeline must operate in at least two evaluation modes, one optimised for short-form, high-volume responses where direct evaluation is sufficient, and one capable of structural analysis for longer, more complex arguments where critical reasoning moves are implicit.

Third, the pipeline must mitigate evaluator bias through structural means, rubric anchoring, position blinding where possible, and regression testing on balanced datasets, rather than relying on model instruction alone.

Fourth, the pipeline must maintain a longitudinal reasoning profile that accumulates evidence over time, because two of the five target dimensions (calibration and updating behaviour) are only meaningfully assessable over a sequence of judgements.

Fifth, the pipeline must balance analytical depth against computational cost. A system that produces excellent evaluations at prohibitive expense is an interesting research artefact, not a deployable capability.

Section 4 presents an architecture designed to satisfy these requirements.

4. The Reasoning Evaluation Pipeline

This section presents the architecture of the proposed reasoning evaluation pipeline. We describe the design principles that govern the system (4.1), the three input modes across which it operates (4.2), the short-form evaluation pathway (4.3), the long-form evaluation pathway (4.4), the dimension scoring framework (4.5), and the longitudinal reasoning profile (4.6).

4.1 Design Principles

Five principles govern the pipeline's architecture, each derived from the philosophical foundations in Section 2 and the technical constraints identified in Section 3:

Environment over reconstruction. Where possible, the pipeline relies on evaluation contexts that force reasoning to the surface rather than on post-hoc reconstruction of hidden chains. Structured environments that require commitment before revelation, elicit confidence levels, and probe reasoning through follow-up interaction produce richer primary evidence than unconstrained text, reducing the pipeline's dependence on LLM inference about what the user was thinking.

User text as primary evidence. In all evaluation modes, the user's original text is the primary object of assessment. LLM reconstruction, where employed, serves as a query layer, a means of illuminating the structure of the text, not as ground truth about the user's reasoning process. Scores are attached to what the user demonstrably wrote or said, not to what the system inferred they might have been thinking.

Longitudinal profile over single-shot scoring. Two of the five target dimensions (calibration and updating behaviour) are not meaningfully assessable from a single response. The pipeline therefore maintains a lightweight, rolling reasoning profile for each user, weighted toward recent performance and the current evaluation context. This profile is updated deterministically after each evaluation event and does not require an LLM call.

Mode-specific weighting. The five dimensions are weighted differently depending on the input mode and the observability of each dimension in that context. A dimension that cannot be reliably assessed from the available input is weighted toward zero rather than scored on insufficient evidence.

Feasibility at scale. Every architectural decision is constrained by computational cost. The pipeline explicitly manages the trade-off between evaluation depth and cost, using the minimum number of LLM calls required for defensible scoring in each mode.

4.2 Three Input Modes

The pipeline receives input tagged with one of three mode identifiers, each corresponding to a distinct evaluation context with different observability characteristics:

Mode A: Structured interactive environment. The user produces responses within a designed game or challenge format that structures the information available, requires explicit commitment, may elicit confidence levels, and can deliver new information mid-session to observe updating behaviour. This mode offers the highest observability across all five dimensions. Evaluation operates in short-form mode (Section 4.3) for individual responses, with longitudinal dimensions computed at the session and profile level.

Mode B: Live speech and debate. The user produces reasoning in real time, under time pressure, without structured mechanics to force hidden reasoning to the surface. The input is a transcript. The pipeline does not control the information environment and cannot introduce new evidence to test updating behaviour. Observability is partial: accuracy, consistency, and evidence engagement can be assessed from the transcript; calibration and updating behaviour are only partially observable and are weighted lower. Evaluation operates in long-form mode (Section 4.4).

Mode C: Published written argument. The input is a written article, essay, or editorial produced by an author who had time to compose, revise, and structure their argument deliberately. There is no interaction. Calibration is not directly observable. Updating behaviour is not observable. Accuracy, consistency, and evidence engagement are the primary scorable dimensions. Evaluation operates in long-form mode (Section 4.4) with adjusted weighting.

4.3 Short-Form Evaluation Pathway

The short-form pathway handles responses produced in structured interactive environments (Mode A) where the input is relatively brief, the information context is known, and the pipeline can leverage game-specific metadata. The pathway consists of four steps:

Step 1: Input structuring. The user's response is packaged alongside its metadata into a standardised evaluation object: the prompt or question the user was responding to, the information they were provided, their response (text, selection, confidence level, prediction, or other input format), any verifiable ground truth, and the dimension weightings for this game type. This step is deterministic and requires no LLM call.

Step 2: Dimension evaluation. The evaluation object is sent to an LLM with a structured rubric describing each dimension and the weightings for this game type. The model returns a constrained JSON output: a numeric score per dimension within a defined range, and a brief textual justification for each score. The rubric is specific and concrete rather than abstract, anchored to observable features of the response, to minimise the evaluator bias documented by Chen et al.[27] This is a single LLM call.

Step 3: Output validation. The LLM's output is validated against a schema and subjected to basic sanity checks: do the justifications reference the actual response content, are dimensions weighted to zero scored zero, and do scores fall within the expected distribution. Malformed or anomalous outputs are rejected and the evaluation retried or flagged. This step is deterministic.

Step 4: Profile update. The validated scores are written to the user's longitudinal reasoning profile. Calibration is recalculated as a running comparison between stated confidence levels and actual accuracy across all responses with resolvable outcomes. Updating behaviour is recalculated where the session included delivery of new information and a subsequent response. The profile is weighted toward recency and the current evaluation context: recent performance on the current topic carries more weight than historical performance on unrelated topics. This step is entirely deterministic.

4.4 Long-Form Evaluation Pathway

The long-form pathway handles debate transcripts, speeches, and written articles (Modes B and C) where the input is complex, the reasoning structure is layered, and critical inferential moves are likely to be implicit. A single LLM-as-judge call against a rubric is insufficient for these inputs: the text is too long, the argument structure too layered, and the implicit assumptions too numerous for a single pass to produce defensible scores.

The full implicit premise reconstruction framework proposed by Habernal et al. involves eight sequential steps.[23] We compress this to four steps, preserving the analytical operations that contribute most to scoring accuracy while halving the number of LLM calls required. The compression is guided by the principle that each step must justify its computational cost against the scoring improvement it produces.

Step 1: Stance and structure extraction. A single LLM call takes the full input text and extracts: the position or positions being advanced, the core reasons offered in support of each, and a compressed gist of each reason. The output is structured JSON: a stance label (or multiple stance labels for texts advancing more than one position) plus an array of reason objects, each containing the original text span, the compressed gist, and the conclusion it most naturally supports. This step combines Habernal's stance identification, reason extraction, and reason gist summarisation into a single operation.

Step 2: Warrant reconstruction and stress test. For each extracted reason, a second LLM call identifies the hidden assumption, the implicit warrant, that connects the reason to the conclusion, and then tests whether the same reason could support the opposite conclusion under a different assumption. The output contains for each reason: the reconstructed warrant, an alternative warrant under which the reason supports a different conclusion, and a robustness assessment. This step combines Habernal's implicit warrant reconstruction, alternative interpretation generation, warrant validation, and comparative reasoning assessment. The compression works because the stress test question, 'could this reason support the opposite conclusion?' captures the core analytical value of the four separate steps it replaces.

Step 3: Dimension evaluation against structure. The scoring LLM evaluates the input against the structured output from Steps 1 and 2: the extracted stance, the reasons, the hidden assumptions, the stress test results, and the robustness assessments. This is a substantially richer input than the raw text and means the scoring is based on the reasoning structure rather than the surface language. The model returns constrained JSON scores per dimension with justifications, with dimension weightings adjusted for the input mode. This is the third LLM call.

Step 4: Validation and profile update. Identical to the short-form pathway: schema validation, sanity checks, and deterministic profile update. No additional LLM call.

The long-form pathway thus requires three LLM calls per evaluation: extraction, reconstruction and stress test, and scoring. This is approximately half the cost of the full eight-step Habernal framework. Empirical testing (Section 5) will determine whether this compression degrades scoring quality below acceptable thresholds.

4.5 Dimension Scoring Across Input Modes

Each of the five dimensions has different observability characteristics across the three input modes:

Accuracy is scorable in all three modes. In Mode A, the pipeline can compare the user's claims against the information set they were provided and any verifiable ground truth. In Modes B and C, the pipeline assesses whether factual claims within the argument are accurate and whether evidence is represented faithfully. This dimension is assessed per-response. The primary risk is that factual verification requires external knowledge; where ground truth is unavailable, the pipeline scores the internal coherence of evidential claims rather than their absolute accuracy.

Calibration is only meaningfully scorable in Mode A, where confidence levels can be elicited and outcomes subsequently resolved. In Modes B and C, linguistic markers of confidence provide weak signals that are weighted accordingly. Calibration is assessed at the longitudinal profile level as a running comparison between stated confidence and actual accuracy over many responses.

Consistency is assessable in all three modes at two levels. Within a single response, the pipeline detects internal contradictions. Across responses (requiring Mode A's repeated structured interactions), the pipeline detects whether the user reaches different conclusions on structurally identical problems presented under different frames. The within-response assessment is per-response; the across-response assessment is longitudinal.

Evidence engagement is assessable in all three modes. In Mode A, the pipeline evaluates whether the user's response demonstrates engagement with the information provided rather than reliance on prior conviction. In Modes B and C, the pipeline assesses whether the argument actively incorporates and addresses available evidence, counterarguments, and relevant information. This dimension is assessed per-response.

Updating behaviour is directly scorable only in Mode A, where the evaluation environment can deliver new information mid-session and observe whether the user's subsequent responses reflect proportional revision. In Mode B (live debate), updating may be partially observable if a

debate partner introduces new evidence and the speaker responds to it. In Mode C (written articles), updating behaviour is not observable and is weighted toward zero.

4.6 The Longitudinal Reasoning Profile

The longitudinal reasoning profile is the data structure that makes calibration and updating behaviour assessable beyond single responses. It is not a permanent, ever-growing record. It is a lightweight, rolling summary of recent reasoning behaviour, weighted toward the current evaluation context.

After each evaluation event, the profile is updated deterministically. No LLM call is required. The update computes: a recalculated calibration score based on the running comparison between confidence and accuracy across all responses with resolvable outcomes, a recalculated updating behaviour score where the session included new-information delivery, and dimension-level trend data indicating whether the user is improving, declining, or stable on each dimension over recent sessions.

The profile is weighted by recency and topical relevance. A user's performance on a geopolitics story last week carries more weight for the current geopolitics story than their performance on an economics story last month. This reflects the empirical reality that reasoning quality is not a fixed trait: it varies by domain, by the user's familiarity with the subject matter, and over time. The profile captures this variation rather than smoothing it away.

The computational cost of the profile update is negligible it is arithmetic over a small structured object, not a language model operation. This is architecturally important: it means the longitudinal intelligence of the pipeline does not scale with LLM cost. Only the per-response evaluations (one LLM call for short-form, three for long-form) incur model inference costs. The profile layer that makes the system genuinely longitudinal is effectively free.

5. Validation Framework

A pipeline that produces scores must demonstrate that those scores are meaningful. This section describes the empirical programme required to validate the reasoning evaluation pipeline,

organised into three components: human benchmark construction (5.1), bias testing (5.2), and longitudinal validation (5.3). We present this as a concrete methodology for planned empirical work rather than as completed results.

5.1 Human Benchmark Construction

The standard validation approach for automated evaluation systems is to compare their outputs against expert human judgement. For the reasoning pipeline, this requires constructing a benchmark dataset of natural language responses scored by multiple human annotators on each of the five dimensions using the same rubric the pipeline uses.

The benchmark dataset must include: responses across all three input modes (short-form game responses, debate transcripts, and written articles), responses spanning a range of reasoning quality from poor to excellent, and responses on politically and ideologically diverse topics to test for content bias. Each response is scored independently by a minimum of three trained annotators. Inter-annotator agreement is measured using Krippendorff's alpha, which handles ordinal scales and multiple annotators.

The pipeline's scores are then compared against the human consensus scores. The target is for the pipeline to achieve agreement with human annotators at a level comparable to the agreement between human annotators themselves. If the pipeline substantially underperforms human inter-annotator agreement on any dimension, that dimension's scoring mechanism requires revision.

5.2 Bias Testing

The pipeline scores reasoning quality on contested political and social topics. Actual bias is tested using a balanced benchmark dataset: high-quality and low-quality reasoning on both sides of contested political topics, constructed so that reasoning quality and political position are orthogonal. The pipeline is run on this dataset and scores are analysed for systematic skew. This test is run on every prompt version and every model change.

Three structural mitigations are built into the pipeline. First, rubric anchoring: the scoring rubric uses specific, observable criteria rather than abstract quality descriptors, reducing the

evaluative surface area available for bias. Second, position blinding where feasible: for dimensions that should not be affected by which side the user argues (consistency, calibration), the pipeline can neutralise specific political content before scoring. Third, regression testing: the balanced benchmark is maintained as a standing test suite and the pipeline's scores are monitored for drift over time, model updates, and prompt changes.

5.3 Longitudinal Validation

Validating the longitudinal dimensions requires a different methodology from single-response validation because these dimensions are, by design, properties of a sequence of judgements rather than of individual responses.

For calibration, the validation is inherently longitudinal: a user's calibration score should predict their future accuracy at stated confidence levels. A well-calibrated user who says they are 80% confident should be right approximately 80% of the time. This is testable by holding out recent responses and comparing predicted accuracy against actual accuracy.

For updating behaviour, the validation requires evaluation events that include delivery of new information. The pipeline's updating score should predict the degree to which a user revises subsequent judgements in response to evidence. Users scored as strong updaters should show larger, more proportional revisions than users scored as weak updaters.

6. Limitations and Open Problems

The pipeline architecture proposed in this paper is a theoretical contribution accompanied by a concrete implementation plan, not a report of empirical results. Several significant limitations and open problems must be acknowledged.

The reconstruction problem is irreducible. The pipeline mitigates but does not solve the fundamental gap between text and thought. Multiple reasoning chains can produce identical text outputs, and no evaluation system can determine with certainty which chain a person actually relied on. For long-form inputs where reconstruction is employed, the pipeline's scores are

assessments of the text's reasoning structure, not of the author's internal reasoning process. The distinction matters and should not be obscured.

The cold-start problem. A new user with no evaluation history presents the pipeline with minimal information. Calibration and updating behaviour are effectively unscoreable. The pipeline can assess accuracy, consistency, and evidence engagement from the first interaction, but the longitudinal dimensions that distinguish this system from a generic text evaluator require data that does not yet exist.

Cost-fidelity trade-off in long-form evaluation. The compressed four-step long-form pathway halves the computational cost of the full Habernal framework but may sacrifice analytical nuance. Whether the compression degrades scoring quality below acceptable thresholds is an empirical question that the validation programme must answer.

Evaluator model dependence. The pipeline's scoring quality is bounded by the capabilities of the LLM used for evaluation. Model updates, provider changes, and behavioural drift can all affect scoring consistency. Prompt versioning and regression testing mitigate this but do not eliminate it.

Cultural and linguistic scope. The philosophical grounding in Section 2 draws on cross-civilisational traditions, but the pipeline's rubrics, prompts, and evaluation criteria are designed for English-language inputs. Reasoning conventions, rhetorical norms, and the relationship between text and thought vary across languages and cultures. The New Confucian philosophers Mou Zongsan and Tang Junyi argued that the Western scientific tradition developed its capacity for detached, propositional reasoning partly by bracketing out moral and relational knowing, and that the Chinese tradition developed sophisticated moral epistemology at the cost of the detached scientific mode.[33,34] If reasoning quality is evaluated purely in terms of accuracy and calibrated probabilistic updating, the system may inadvertently privilege one culturally specific epistemic mode over others. Extending the pipeline to non-English inputs requires not merely translation of prompts but adaptation of the evaluation framework itself. This represents a substantial research agenda in its own right.

Empirical validation is outstanding. The validation framework described in Section 5 is a methodology, not a result. The benchmark dataset has not yet been constructed, the human

annotation study has not been conducted, and the bias testing has not been run. The claims made in this paper about the pipeline's capabilities are architectural and theoretical. They become empirical claims only when the validation programme is complete.

7. Conclusion

This paper has presented a reasoning evaluation pipeline, an architecture for quantifying human reasoning quality from natural language, grounded in a philosophically defensible definition and designed to operate at scale across multiple input contexts.

The foundational contribution is the definition itself. By synthesising traditions from Aristotelian formal logic through classical Indian epistemology, early Chinese philosophy of language, Enlightenment theories of bias and uncertainty, Peircean abduction, bounded rationality, and Bayesian belief management, we arrived at a formulation, reasoning as the process of forming and revising beliefs in proportion to evidence, consistently across contexts, under uncertainty, in the service of understanding the world as it actually is. That produces five measurable dimensions, each philosophically grounded and each operationalisable through the pipeline architecture.

The technical contribution is the hybrid pipeline itself. By operating in two evaluation modes, a lean, single-call pathway for short-form responses and a compressed three-call pathway for long-form argument that extracts reasoning structure before scoring - the pipeline manages the trade-off between analytical depth and computational feasibility. By maintaining a lightweight longitudinal reasoning profile weighted toward recency and topical relevance, the pipeline makes calibration and updating behaviour assessable without inflating per-evaluation costs. By tagging inputs with mode identifiers and adjusting dimension weightings accordingly, the pipeline handles the reality that reasoning manifests differently across structured environments, live speech, and published writing.

The honest limitation is that these are architectural claims, not empirical ones. The pipeline has been designed; it has not yet been validated against human judgement at the scale required to confirm that its scores are meaningful. The next phase of this work is to execute the validation plan described in Section 5: construct the benchmark dataset, conduct the human annotation

study, run the bias tests, and report whether the pipeline's scores achieve the agreement with expert human judgement that would justify deployment.

If they do, the result is a novel capability: the first system designed to evaluate not what a person believes, not how persuasively they argue, and not whether a machine can replicate a chain of logic - but how well a human being reasons. That capability has never existed at scale. Whether it can be built is an empirical question. This paper is the case that it is worth attempting, and the blueprint for how to try.

Notes

[1] Habernal, I. et al., "The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants," Proceedings of NAACL-HLT (2018), pp. 1930–1940.

[2] Suzgun, M. et al., "Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them," Findings of ACL (2023).

[3] Wachsmuth, H. et al., "Computational Argumentation Quality Assessment in Natural Language," Proceedings of EACL (2017), pp. 176–187.

[4] Aristotle. Prior Analytics, Book I. Trans. Robin Smith (1989). Indianapolis: Hackett. See also Striker, G., Aristotle's Prior Analytics Book I (Oxford: Oxford University Press, 2009).

[5] Mates, B. (1953). Stoic Logic. Berkeley: University of California Press, pp. 26–42. See also Bobzien, S., "Ancient Logic," in Stanford Encyclopedia of Philosophy.

[6] Nyāya-sūtra 1.1.1. See Vātsyāyana's Nyāyabhāṣya; discussed in Ganeri, J. (2001). Philosophy in Classical India: The Proper Work of Reason. London: Routledge, pp. 77–99.

[7] Matilal, B.K., Perception: An Essay on Classical Indian Theories of Knowledge (Oxford: Clarendon Press, 1986), pp. 1–40. See also Matilal, B.K., Epistemology, Logic, and Grammar in Indian Philosophical Analysis (1971; repr. Oxford: Oxford University Press, 2005).

[8] Ganeri, J. (ed.) (2017). The Oxford Handbook of Indian Philosophy. Oxford: Oxford University Press, Introduction.

[9] Fraser, C. (2011). "Knowledge and Error in Early Chinese Thought." Dao, 10(2), 127–148. See also Fraser, C., "Mohist Canons," in Stanford Encyclopedia of Philosophy.

[10] Bacon, F. (1620). Novum Organum, Book I, Aphorisms 38–68. Modern edition: The New Organon, ed. L. Jardine and M. Silverthorne (Cambridge: Cambridge University Press, 2000).

[11] Hume, D. (1739). A Treatise of Human Nature, Book I, Part III, Section VI. Standard edition: ed. T.L. Beauchamp (Oxford: Oxford University Press, 2000).

[12] Kant, I. (1781). Critique of Pure Reason, Second Analogy of Experience, A189/B232. Standard translation: ed. and trans. P. Guyer and A.W. Wood (Cambridge: Cambridge University Press, 1998).

[13] Boole, G. (1854). An Investigation of the Laws of Thought. London: Walton & Maberly.

[14] Frege, G. (1879). Begriffsschrift. Halle: Louis Nebert.

[15] Peirce, C.S. (1903). "Harvard Lectures on Pragmatism." In Collected Papers of Charles Sanders Peirce, Vol. 5, eds. Hartshorne, C. & Weiss, P. Cambridge: Harvard University Press, §171. See also "Deduction, Induction, and Hypothesis" (1878), vol. 2, §§619–644.

[16] Simon, H.A. (1955). "A Behavioral Model of Rational Choice." Quarterly Journal of Economics, 69(1), 99–118.

[17] Tversky, A. and Kahneman, D. (1974). "Judgment under Uncertainty: Heuristics and Biases." Science, 185(4157), 1124–1131.

[18] Kahneman, D. (2011). Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, pp. 19–30.

[19] Tetlock, P.E. (2005). Expert Political Judgment: How Good Is It? How Can We Know? Princeton: Princeton University Press. See also Tetlock, P.E. and Gardner, D. (2015). Superforecasting: The Art and Science of Prediction. New York: Crown, pp. 63–91.

[20] Jaynes, E.T. (2003). Probability Theory: The Logic of Science. Cambridge: Cambridge University Press, ch. 1. See also Howson, C. and Urbach, P. (2006). Scientific Reasoning: The Bayesian Approach, 3rd ed. Chicago: Open Court.

[21] Tversky, A. and Kahneman, D. (1974). Ibid., p. 1128.

[22] Mercier, H. and Sperber, D. (2011). "Why Do Humans Reason? Arguments for an Argumentative Theory." Behavioral and Brain Sciences, 34(2), 57–74.

[23] Habernal, I. et al. (2018). Op. cit.

[24] Turpin, M. et al. (2023). "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." arXiv preprint arXiv:2305.04388.

[25] Lanham, T. et al. (2023). "Measuring Faithfulness in Chain-of-Thought Reasoning." arXiv preprint arXiv:2307.13702.

[26] Gneiting, T. and Raftery, A.E. (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation." Journal of the Royal Statistical Society: Series B, 69(2), 243–268.

[27] Chen, G. et al. (2024). "Humans or LLMs as the Judge? A Study on Judgement Biases." Proceedings of EMNLP, pp. 474–489.

[28] Koo, R. et al. (2024). "Benchmarking Cognitive Biases in Large Language Models as Evaluators." Findings of ACL, pp. 29–45.

[29] Williams, A. et al. (2018). "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference." Proceedings of NAACL-HLT, pp. 1112–1122.

[30] McCoy, R.T. et al. (2019). "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference." Proceedings of ACL, pp. 3428–3448.

[31] Mirzakhmedova, N. et al. (2024). "How Well Can LLMs Negotiate? A Comprehensive Assessment of LLMs as Evaluators." arXiv preprint arXiv:2404.09696.

[32] Stahl, M. et al. (2023). "Detecting and Filling Missing Inferential Steps in Arguments." Findings of EMNLP, pp. 312–327.

[33] Rošker, J.S. (2021). "Modern Confucian Epistemology: From Reason to Intuition — And Back." In Springer Handbook of Chinese Philosophy, pp. 421–435.

[34] "Epistemology in Chinese Philosophy." Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/chinese-epistemology/