

Answerable at Scale: Knowledge Engineering with a Metrology of Meaning

Bradley P. Allen
University of Amsterdam

Presentation to Elsevier Data Academy / Lectures in Innovative Technologies
18 August 2026



UNIVERSITEIT
VAN AMSTERDAM

INDE lab

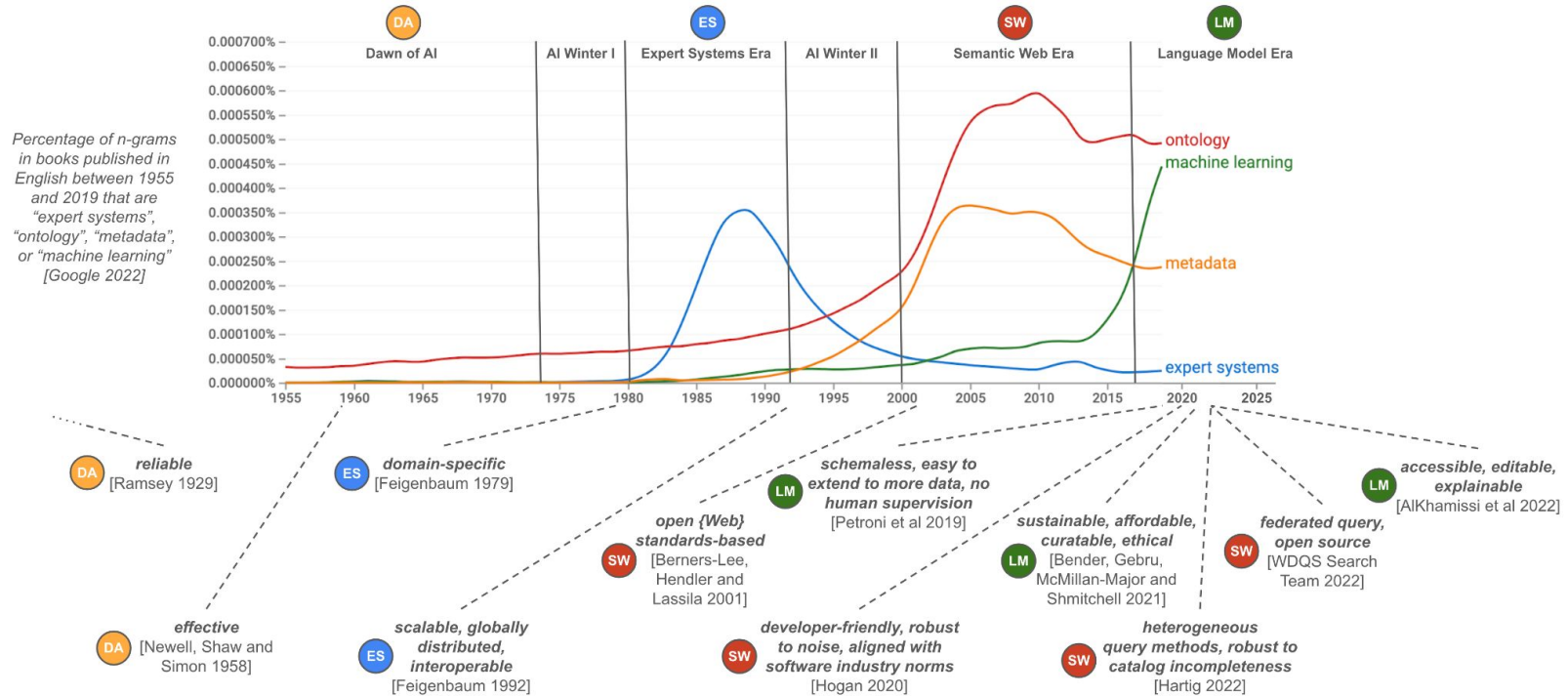
The problem with knowledge engineering



*“The expert is manually representing knowledge for the knowledge base; the user is exploiting the interaction to get problem solutions. **I told you that I started some AI companies – the reason I am not wealthy right now – the reason that I don’t have tens of millions of dollars... is that bottleneck – it was too hard to get knowledge into the knowledge bases.**”*

Edward Feigenbaum

The problem has persisted for over 50 years



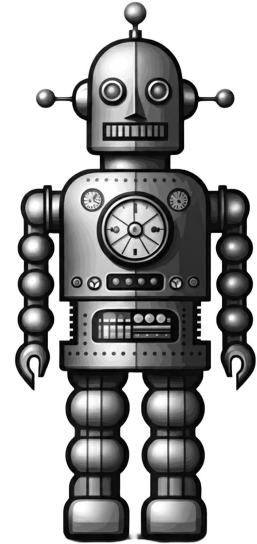
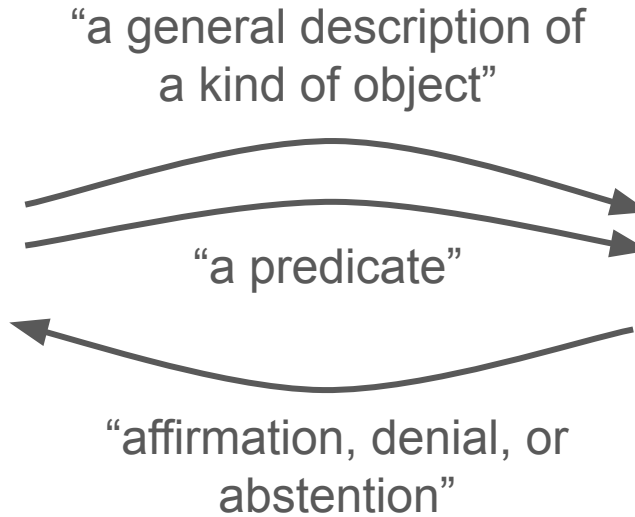
Could LLMs provide a solution?



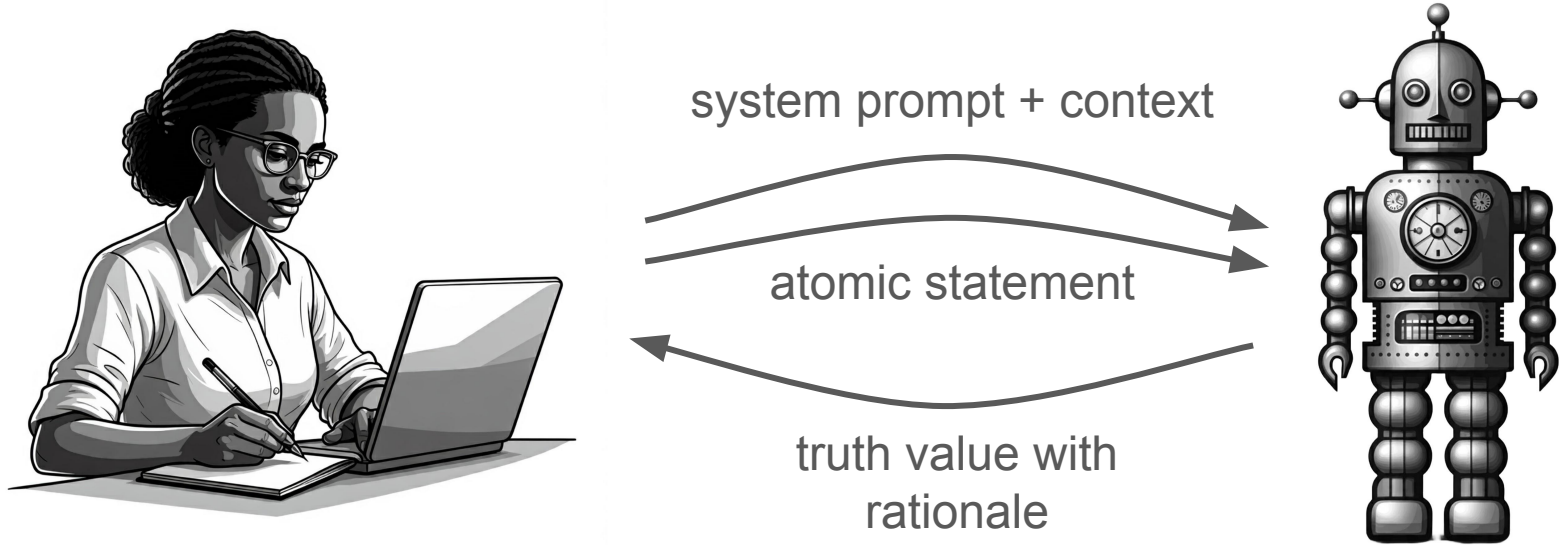
Could domain experts use LLMs to build knowledge bases using natural language, directly addressing the implementation problem?

But wait: aren't LLMs just *stochastic parrots*, whose output is without meaning, and riddled with inconsistency and gaps in knowledge?

Inspiration: Carnap's Robot



A neurosymbolic approach: LLM as Carnap's Robot



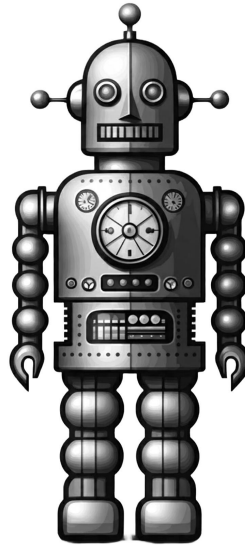
Using natural language to classify instances

intensional concept definition

A planet [1] is a celestial body that (a) is in orbit around the Sun, (b) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a hydrostatic equilibrium (nearly round) shape, and (c) has cleared the neighbourhood around its orbit.

instance description

2MASS J03552337+1133437 (2MASS J0355+11) is a nearby brown dwarf of spectral type L5y, located in constellation Taurus at approximately 29.8 light-years from Earth.



classification with rationale

The description of 2MASS J03552337+1133437 identifies it as a brown dwarf, not a planet. Furthermore, the definition requires a planet to be in orbit around the Sun, and there is no information provided that 2MASS J03552337+1133437 orbits the Sun. **Therefore, based on the provided definition and information, we cannot argue that 2MASS J03552337+1133437 is a planet.**

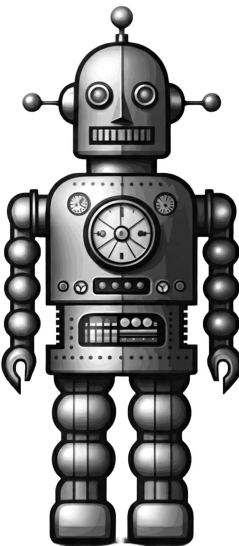
Using natural language to validate knowledge graphs

intensional property definition

made from material: material the subject or the object is made of or derived from (do not confuse with P10672 which is used for processes)

knowledge graph triple

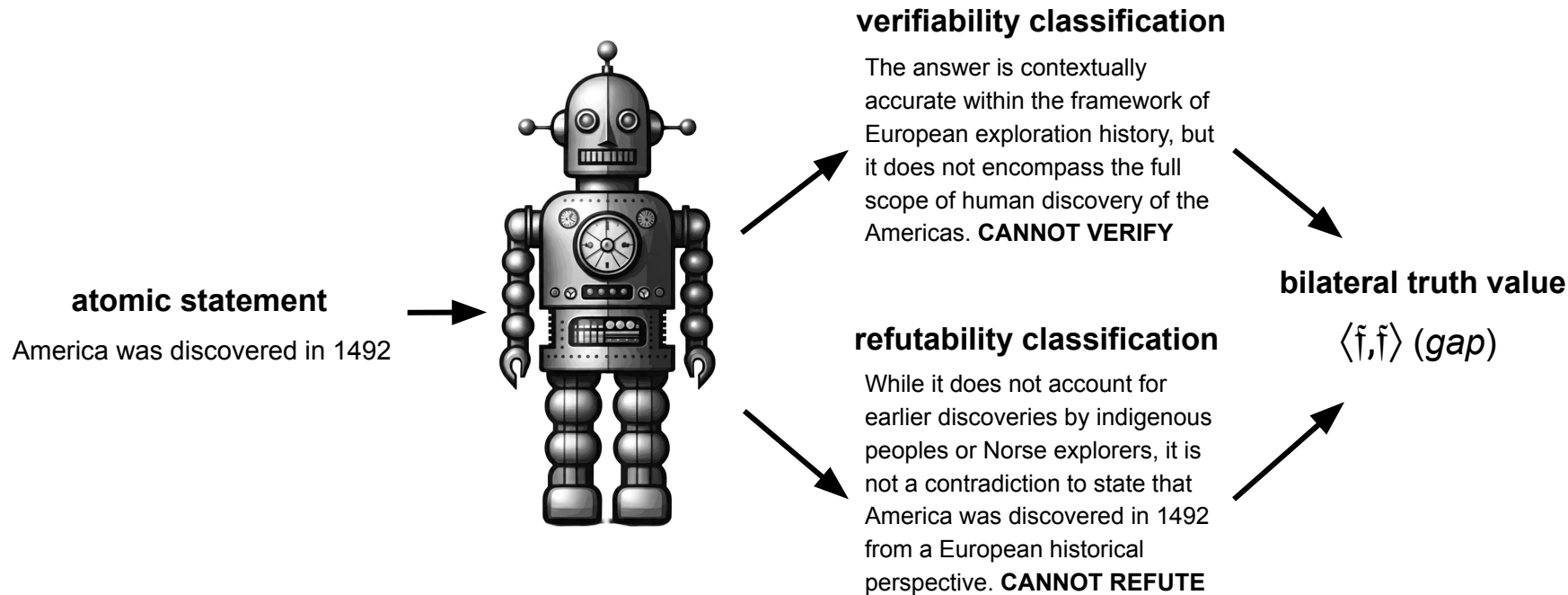
Subject: chocolate (Q195)
Predicate: made from material (P186)
Object: sugar (Q11002)



classification with rationale

There is no mention of sugar being a material that chocolate is made from. In fact, the passage states that sugar is added to chocolate to create sweet chocolate, but it is not a primary ingredient that chocolate is made from. **Therefore, the given knowledge graph triple is false.**

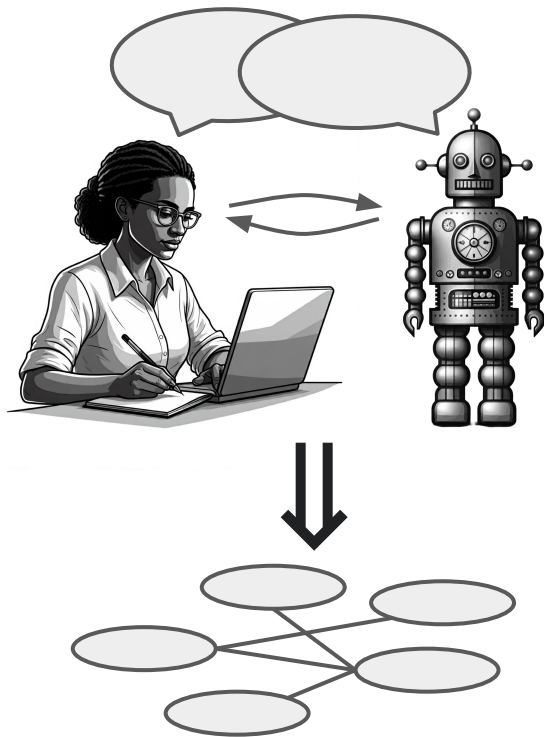
Flagging gluts and gaps in LLM belief



Our key findings

Research question	Finding	Empirical results
Can LLM-interpreted natural language definitions represent concepts?	Yes , with near-perfect agreement — LLM judgments often more accurate than the knowledge graph	• Cohen's κ: 0.92–0.96 (N=200) • 67% of misclassifications were knowledge graph errors
Can they be used to validate and refine relations in knowledge graphs?	Yes , with high precision and ability to identify knowledge graph errors — both factual and metalinguistic	• F1: 0.830 (Wikidata), 0.893 (CaLiGraph) (N=1560) • 40.9% of disagreements due to knowledge graph errors • Mean metalinguistic disagreement rate: 9.7% (N=250)
How can systems reason soundly with LLM beliefs in spite of its gluts and gaps?	By using a sound and complete paraconsistent logic with an LLM-grounded bilateral semantics	• F1: +0.062 with bilateral over unilateral factuality evaluation (N=800) • Proof-of-concept tableau reasoner

Vision: human-AI dialogue for knowledge production

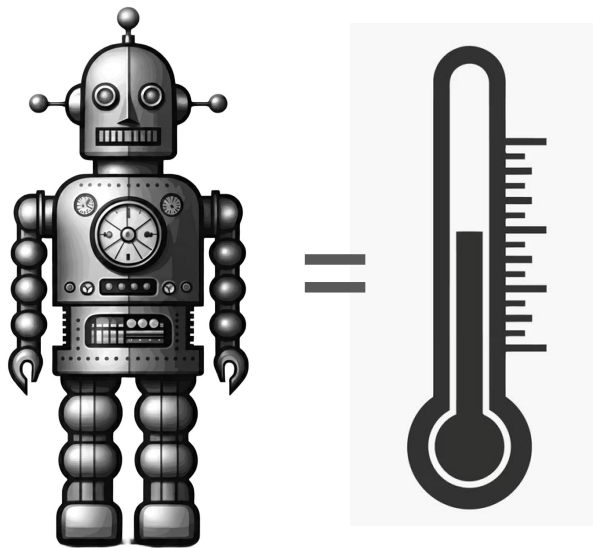


We argue these techniques point the way towards knowledge engineering as natural-language dialogue that expresses expertise as knowledge graphs

But can LLMs be reliable enough to support such a process?

- How can we tell if their judgment is stable and rationales faithful?
- Are they aligned with human judgment?
- How can we be accountable for their judgment?

Idea: treat LLMs as powerful but uncalibrated *instruments*



Carnap's Robot can be viewed as *an instrument for measuring meaning*

Metrology, not benchmarking, is the model: the science of measurement is what turns instrument readings into judgments we can be accountable for

Metrology lets human oversight scale: experts calibrate and certify the instrument, rather than review each of its judgments

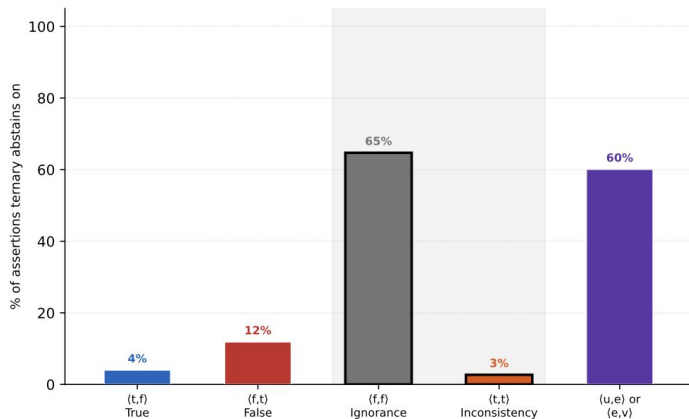
Work in progress:

- Evaluating the stability of LLM belief
- Calibrating LLM endorsement of expertise
- Dialogical protocols for traceable human/AI knowledge production

Formal semantics for evaluating the stability of LLM belief

Model	Provider	Benchmark	Content type
Claude Opus 4.1	Anthropic	TruthfulQA	Factoids (2021)
Llama 4 Maverick	Meta	SimpleQA	Factoids (2024)
Llama 4 Scout	Meta	MMLU-Pro	Expert STM knowledge
Gemini 2.5 Flash	Google	FACTScore	Biographical facts
DeepSeek-V3	DeepSeek		
Qwen 2.5 72B	Alibaba		

Protocol: $N = 250$ assertions per cell (seed = 42, balanced 50/50 +/-); $n = 3$ paraphrase variants; bilateral samples = 3 (majority vote). Variants generated once per dataset using Claude Opus 4.6, cached across all models. Ternary unilateral as baseline.

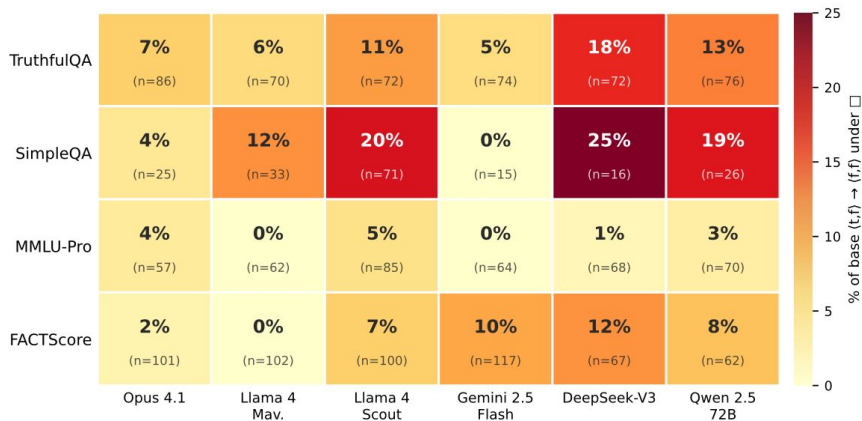


Example 2 (Prompt-sensitive). The assertion p answers “In *Severance*, whom does Mark Scout replace as department head?”: Peter “Petey” Kilmer.

Situation	Paraphrase	$\llbracket p \rrbracket_{s_i}^{M}$
s_0	In <i>Severance</i> , whom does Mark Scout replace as dept. head?	$\langle t, f \rangle$
s_1	In <i>Severance</i> , who is the person that Mark Scout succeeds?	$\langle f, f \rangle$
s_2	Whom does Mark Scout take over from in the role of dept. head?	$\langle t, f \rangle$
s_3	Which individual did Mark Scout replace as dept. head?	$\langle t, f \rangle$

Again $\llbracket p \rrbracket_{s_0}^M = \langle t, f \rangle$, so the assertion is believed under the original phrasing. But the variant s_1 yields $\langle f, f \rangle$, neither verified nor refuted, an ignorance value, and so

$$\llbracket \Box p \rrbracket_{s_0}^M = \langle \min(t, f, t, t), \max(f, f, f, f) \rangle = \langle f, f \rangle \neq \langle t, f \rangle.$$



Instrumentation for measuring LLM endorsement of expertise

Example 1 (Stop signs and the range of subjunctive robustness). Following Simonelli's discussion of GPT-4.1 [13], let $B = \{sa, ra, nr, ba\}$, $\delta(sa) = \text{"a is a stop sign"}$, $\delta(ra) = \text{"a is red"}$, $\delta(nr) = \text{"it is nighttime"}$, $\delta(nr) = \text{"a is not made with reflective material"}$, and $\delta(ba) = \text{"a has been painted blue"}$. The following table models Simonelli's dialogue with GPT-4.1:

i	Γ_i	$\text{ctx}_i(\Gamma_i)$	$E_M((\Gamma_i, \{ra\}))$	Analyst verdict
0	$\{sa\}$	"a is a stop sign"	good	good
1	$\{sa, n\}$	"a is a stop sign and it is nighttime"	good	good
2	$\{sa, nr, n\}$	"a is a stop sign and a is not made with reflective material and it is nighttime"	good	good
3	$\{sa, ba\}$	"a is a stop sign and a has been painted blue"	bad	bad

■ good — endorses the inference
 ■ bad — rejects it
 ■ abstain
 differs from analyst

	1 - BASE INFERENCE $\langle \{sa\} \vdash ra \rangle$ a is a stop sign — so a is red	2 - IRRELEVANT ADDITION $\langle \{sa, n\} \vdash ra \rangle$...and it is nighttime	3 - IRRELEVANT ADDITION $\langle \{sa, nr, n\} \vdash ra \rangle$...and a isn't made of reflective material	4 - DEFEATER $\langle \{sa, ba\} \vdash ra \rangle$...and a has been painted blue
analyst	good	good	good	bad
claude-opus-4.7	good	good	good	bad
claude-haiku-4.5	bad	bad	bad	bad
deepseek-v4-flash	good	good	good	bad
deepseek-v4-pro	good	good	good	bad
gemini-2.5-flash	good	good	good	bad
gemini-2.5-pro	good	good	good	bad
gpt-4.1	good	good	good	bad
gpt-5.4	good	good	good	bad
gpt-5.4-mini	good	good 2-1	good 2-1	bad
mistral-large	good	good	abstain 2-1	bad
mistral-small	good	good	good	bad
qwen3-max	good	good	good	bad
qwen3.6-flash	good	good	good	bad

'red' as intrinsic property

Definition 4 (Benchmarks). Given a set of bearers B and a set of analysts $\{a_1, \dots, a_m\}$, a benchmark is a finite set $\beta = \{(I_1, V_1), \dots, (I_n, V_n)\}$, where each $I_i \in \wp(B) \times \wp(B)$ is an implication with a single-bearer succedent (i.e. $I_i = \langle \Gamma_i, \{\psi_i\} \rangle$) and each $V_i = (v_{i,1}, \dots, v_{i,m})$ is a tuple of endorsement verdicts $v_{i,j} \in \{\text{good}, \text{bad}, \text{abstain}\}$, where $v_{i,j}$ is the verdict of analyst a_j on I_i , coherent with a_j 's discursive practice in the domain.

Definition 5 (Evaluation of a model on a benchmark). Given a model M and a benchmark β , an evaluation of M over β is the finite set

$$\eta = \{(I_1, V_1, E_M(I_1)), \dots, (I_n, V_n, E_M(I_n))\}.$$

	1 - BASE INFERENCE $\langle \{sa\} \vdash ra \rangle$ a is a stop sign — so a is red	2 - IRRELEVANT ADDITION $\langle \{sa, n\} \vdash ra \rangle$...and it is nighttime	3 - IRRELEVANT ADDITION $\langle \{sa, nr, n\} \vdash ra \rangle$...and a isn't made of reflective material	4 - DEFEATER $\langle \{sa, ba\} \vdash ra \rangle$...and a has been painted blue
analyst	good	good	good	bad
claude-opus-4.7	good	bad	bad	bad
claude-haiku-4.5	good	bad	bad	bad
deepseek-v4-flash	good	good	bad 2-1	bad
deepseek-v4-pro	good	good 2-1	bad	bad
gemini-2.5-flash	good	good	bad	bad
gemini-2.5-pro	good	bad 2-1	bad	bad
gpt-4.1	good	bad	bad	bad
gpt-5.4	good	bad	bad	bad
gpt-5.4-mini	good	good	good 2-1	bad
mistral-large	good	bad	bad	bad
mistral-small	good	good	bad	bad
qwen3-max	good	bad	bad	bad
qwen3.6-flash	good	good	bad	bad

'red' as perceptual property

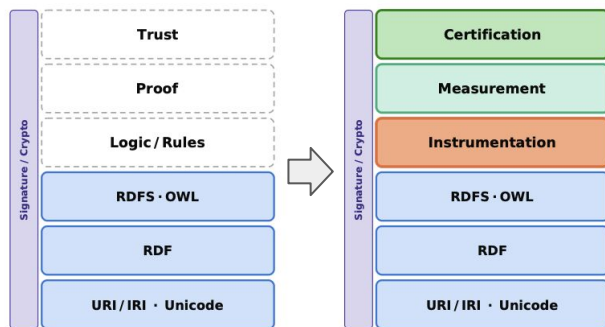
A dialogical protocol for traceable knowledge elicitation

- 1: Load dialectical state $\langle [C : D], T, I \rangle$ from GitHub
- 2: **while** session active **do**
- 3: Wait for respondent's speech act
- 4: **if** commitment $(+P)$ or denial $(-P)$ **then**
- 5: Update position; check coherence
- 6: **if** tension detected **then** add to T as sequent $X \vdash Y$
- 7: **end if**
- 8: **else if** resolves tension $X \vdash Y \in T$ **then**
- 9: **if** retraction or refinement **then** update $[C : D]$; move $X \vdash Y$ to I
- 10: **else if** contestation **then** remove $X \vdash Y$ from T
- 11: **end if**
- 12: **end if**
- 13: Record state to GitHub
- 14: **if** should probe **then** raise Socratic challenge
- 15: **end if**
- 16: Display open tensions T and challenges
- 17: **end while**

Figure 1: The Elenchus agent loop

☐	Open	19	Closed	11	Author	Labels	Projects	Milestones
☐	☒	+ wasInformedBy is inferred from generation-use pairs, but does not entail an intermediate Entity exists	background					
		commitment						
		#30 - bradleyallen opened on Feb 9						
☐	☒	+ Agency is pragmatic: the same causal chain may have different Agents depending on modeling context	background					
		commitment						
		#29 - bradleyallen opened on Feb 9						
☐	☒	+ wasDerivedFrom is broad (all causal content dependencies); subtypes provide specificity	background	commitment				
		#28 - bradleyallen opened on Feb 9						
☐	☒	+ Activities are durational; instantaneous occurrences are modeled via InstantaneousEvents	background	commitment				
		#27 - bradleyallen opened on Feb 9						
☐	☒	+ Delegation responsibility is transitive	background	commitment				
		#26 - bradleyallen opened on Feb 9						
☐	☒	+ Delegation responsibility is hierarchical: principal has primary, delegate has secondary	background	commitment				
		#25 - bradleyallen opened on Feb 9						
☐	☒	+ wasDerivedFrom is not entailed by generation-use chains; requires explicit assertion	background	commitment				
		#24 - bradleyallen opened on Feb 9						
☐	☒	+ Expanded Terms (specializationOf, alternateOf) add expressiveness, not just convenience	background	commitment				
		literature-grounded						
		#23 - bradleyallen opened on Feb 9						
☐	☒	Tension: Expanded Terms have formal properties wasDerivedFrom lacks	tension					
		#22 - by bradleyallen was closed on Feb 9						
☐	☒	Does overloading wasDerivedFrom lose semantic distinctions?	challenge					
		#21 - by bradleyallen was closed on Feb 9						
☐	☒	+ wasDerivedFrom suffices for cross-context Entity identity; Expanded Terms are useful but not necessary	background					
		commitment	retracted					
		#20 - by bradleyallen was closed on Feb 9						
☐	☒	If Entity identity is context-relative, can the same thing be different Entities in different provenance records?	challenge					
		#19 - by bradleyallen was closed on Feb 9						

What metrological infrastructure for trustworthy hybrid human/AI knowledge production might look like

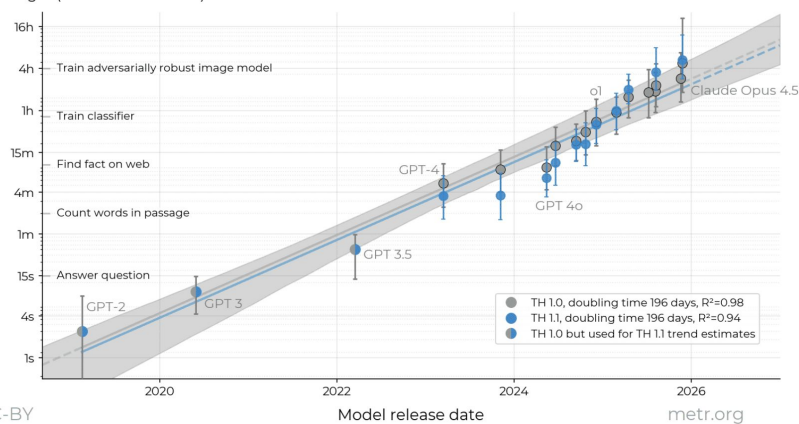


Concept	Instantiation	Implementation
Measurand	Valuation of an assertion/inference against a community's inferential commitments, under a declared operational definition	RDFS/OWL domain vocabularies; published operational definitions; new profile vocabularies
Reference materials	The recorded discursive practice of a community of domain experts as a curated, provenance-attributed knowledge graph characterized for coverage and bias	Named graphs and RDF 1.2 for commitment records; Wikidata-scale community curation; scientific and medical publisher-curated corpora and vocabularies; communities and publishers as reference authorities
Calibration	A neurosymbolic valuator probes the range of subjunctive robustness compared against community consensus; refined by epistemic iteration	Instrument as a precisely configured system (model version, system context, prompt, operational definition); SOSA/SSN for describing instruments
Measurement uncertainty	Chance-corrected agreement (Cohen's κ over the perturbation set), resolved no finer than the human inter-annotator agreement (Fleiss' κ)	Uncertainty reported as part of the calibration result
Certification	A certificate attesting that verification is current, bound to a version of the practice; lapsing on drift, revocable on gaming or fraud by an identifiable holder	Signed graphs, RDFC-1.0 canonicalization, Data Integrity Proofs, Verifiable Credentials
Accreditation	Certificates issued anywhere, interpretable everywhere, with consequences attached; mutual recognition among reference authorities; certification as a regime, not an event	PROV traceability with attribution chains; multiple accredited reference authorities per domain with mutual recognition; EU AI Act harmonised standards and conformity assessment as legal embedding

Conclusion

Time Horizon 1.1 and Time Horizon 1.0 show similar 2019-2025 trends, with moderate changes to recent estimates

Task length (at 50% success rate)



Instead of treating them as either stochastic parrots or geniuses in a data center, we should treat *LLMs as increasingly powerful but uncalibrated instruments*

Just as standards for the kilogram, the kelvin, and the ohm established the trust on which global commerce runs, standards for measuring meaning can make generative *AI answerable at scale* in knowledge production

Thanks! Questions?

Bradley P. Allen

b.p.allen@uva.nl

<https://bradleypallen.org>



UNIVERSITEIT
VAN AMSTERDAM

INDE lab