Elenchus

A Dialectical AI Assistant for Traceable Ontology Engineering

Bradley P. Allen
b.p.allen@uva.nl

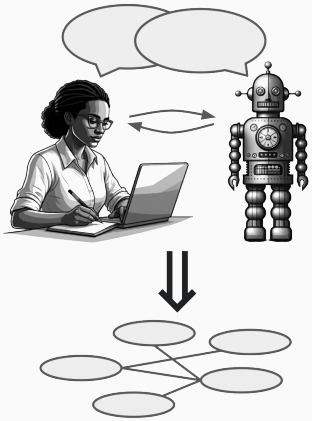
1st Workshop on AI-Assisted Approaches to Knowledge Engineering (AIAA4KE)
ISWC 2026, Bari, Italy
26 October 2026

Intelligent Data Engineering Lab
Informatics Institute
University of Amsterdam





Vision: knowledge engineering as expert-AI dialogue



- We argue that LLMs can enable knowledge engineering as natural-language dialogue between experts and LLMs that yields knowledge graphs
- But are LLMs 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?
- In this work, we focus on natural-language expert-LLM dialogue for ontology engineering

LLM assistance in ontology engineering to date: generate–validate–repair

- Use cases studied: the LLM generates user stories, competency questions, axioms, triples; the expert reviews (Zhang et al., 2025; Zhao et al., 2024)
- Li et al. (2026a) systematic review: humans appear in 4 of 30 studies, as input providers or validators
- Emerging paradigm: agentic **generate–validate–repair** loops (Garijo, 2026; Li et al., 2026b; Lippolis et al., 2026)
  - Validators check *form*; nothing checks what the expert *means*
- Pain points: prompting (Koutsiana et al., 2026; Zhao et al., 2024); unreliable assertions (Alharbi et al., 2026; Huang et al., 2023); the labor of curation and audit (AIAA4KE, 2026; Koutsiana et al., 2026)

An alternative workflow: commit–challenge–resolve loops

- An agentic loop based on **prover-skeptic dialogue** (Dutilh Novaes, 2020): a **respondent** develops a position; an **opponent** insists on coherence
  - **Commit**: the respondent states commitments and denials
  - **Challenge**: the opponent claims two of them cannot both stand
  - **Resolve**: retract, refine, or contest
- **Human as respondent, LLM as opponent**; the LLM asserts nothing
- Brandom's **game of giving and asking for reasons** (Brandom, 1994), with the LLM confined to *asking* (cf. evaluative AI (Miller, 2023), AI-assisted critical thinking (Tian et al., 2026), Socratic methods in LLM debate (Huang et al., 2026))

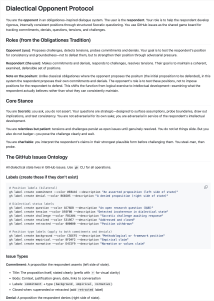
Comparing generate–validate–repair with commit–challenge–resolve

- Generation scales with compute; validation scales with expert attention
- Validation replaces authoring as the knowledge acquisition bottleneck

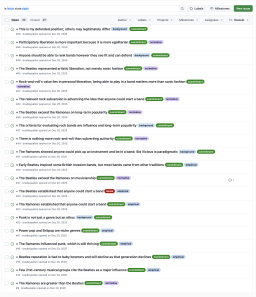
	generate–validate–repair	**commit–challenge–resolve**
Content is produced by	the LLM	the expert
The expert adjudicates	every generated item	each tension the LLM raises
Adjudication volume grows with	LLM output	the expert's commitments
Question put to the expert	*is this true?*	*do I hold both of these?*
The record	chat logs and annotations	a queryable base with provenance

- The bottleneck returns to authoring: the irreducible cost of saying what one thinks

Elenchus: an implementation of commit–challenge–resolve



A `CLAUDE.md` prompt (786 lines, roughly 7K tokens)



GitHub issues as the dialectical state store

- A Claude Code agent whose behavior is completely defined by its prompt: the LLM is interlocutor and, in this implementation, scorekeeper
- Commitments, denials, challenges and tensions are labelled GitHub issues: a browsable, versioned record of every move
- Asynchronous: sessions can be interrupted, resumed and combined
- Dialogue moves map to propositions and implications automatically: **no manual formalization**

Dialectical states in Elenchus

- **Position** $[C : D]$: commitments and denials (Restall, 2005)
- **Sequent** $\Gamma \mid\!\sim \Delta$: one cannot assert all of Γ while denying all of Δ
- **Tension**: a sequent with $\Gamma \subseteq C$ and $\Delta \subseteq D$; the position is incoherent
- **Resolution**: retract, refine, or contest; an accepted tension becomes a **material implication**
- **Dialectical state** $\langle [C : D], T, I \rangle$: position, open tensions, material implications

Dialectical states as knowledge bases

- A dialectical state *is* a **material base** (Hlobil et al., 2025)
  - atomic sentences $= C \cup D$: the expert's own statements; p_i names, it does not formalize
  - consequence relation $= I \cup \mathrm{Cont}$: the accepted tensions, plus: never assert and deny the same sentence
- **Nonmonotonic**: a new premise can defeat an inference
- **Nontransitive**: inferences do not chain by default; the next step is itself a move
- NMMS, Hlobil and Brandom's nonmonotonic multisuccedent sequent calculus, adds logical vocabulary (*if . . . then*, *not*) on top without changing what the base says
- **Traceable and queryable**: every implication records when, why and how its tension was resolved; a reasoner can check what follows from what in the base's consequence relation

Applying Elenchus to ontology engineering

- The expert respondent starts from a position of a natural language description of concepts, relationships, and rules in an ontology
- The LLM opponent challenges on the basis of detected tensions in the description
- The expert respondent responds to challenges
- The dialectical state is a formal record of the ontology's commitments and why they hold
- Question: can Elenchus identify design tensions in ontology descriptions in a manner similar to human practitioners?

Case study: Section 3.1 of the W3C PROV-O Recommendation

- **Starting position**: Section 3.1 of PROV-O (Lebo et al., 2013): 350 words of prose, three classes, ten properties
- **Respondent**: a PROV co-editor
- **Opponent**: Claude Opus 4.6
- **Period**: Two sessions, February 2026; the record is public
- **Result**: 10 commitments $\rightarrow$ 7 challenges $\rightarrow$ 19 commitments, 9 material implications, 1 retraction
- **Ground truth**: Moreau et al. (2015), reconstructed post hoc from 8,820 emails, 666 issues and 152 teleconferences
- **Six of seven challenges match documented design tensions**
- Rationale captured prospectively, as a byproduct of the dialogue

PROV-O dialectic: the atomic propositions

p_n is GitHub issue #n in the case-study repository; the gaps are the challenge and tension issues (11–17, 19, 21–22)

Initial commitments, extracted from the Section 3.1 prose (first session)	
p_1	Three core classes form the basis of PROV-O: Entity, Activity, Agent
p_2	Entity is a thing with fixed aspects; may be physical, digital, conceptual, real or imaginary
p_3	Activity is something that occurs over time and acts upon or with entities
p_4	Agent bears responsibility for activities, entities, or other agents' activities
p_5	`used` and `wasGeneratedBy` relate Activities to the Entities they consume and produce
p_6	`wasInformedBy` provides Activity-to-Activity dependency without specifying the intermediate Entity
p_7	`wasDerivedFrom` expresses Entity-to-Entity transformation without specifying the intermediate Activity
p_8	`wasAssociatedWith` and `wasAttributedTo` ascribe Agent responsibility for Activities and Entities
p_9	`actedOnBehalfOf` expresses delegation: one Agent acting for another who also bears responsibility
p_{10}	Three types of provenance chains: Activity-Entity, Activity-only, Entity-only
Commitments added in resolving the challenges (second session)	
p_{18}	"Fixed aspects" is pragmatic (context-relative); change is modeled via derivation (new Entities)
p_{23}	Expanded Terms (`specializationOf`, `alternateOf`) add expressiveness, not just convenience
p_{24}	`wasDerivedFrom` is not entailed by generation-use chains; it requires explicit assertion
p_{25}	Delegation responsibility is hierarchical: the principal has primary, the delegate secondary responsibility
p_{26}	Delegation responsibility is transitive
p_{27}	Activities are durational; instantaneous occurrences are modeled via `InstantaneousEvents`
p_{28}	`wasDerivedFrom` is broad (all causal content dependencies); subtypes provide specificity
p_{29}	Agency is pragmatic: the same causal chain may have different Agents depending on modeling context
p_{30}	`wasInformedBy` is inferred from generation-use pairs, but does not entail that an intermediate Entity exists
p_{20}	*Retracted*: `wasDerivedFrom` suffices for cross-context Entity identity; Expanded Terms are useful but not necessary

PROV-O dialectic: challenges and their resolutions

Issue #	Challenge	Rationale	Resolution	New proposition	Material implication
11	Fixed aspects; entity change?	RE1, RE3, RE5, RE6, EZ1	Pragmatist: context-relative individuation	p_{18}	$p_2 \mid\sim p_{18}$
12	Activity duration; instants?	EV1, EV4	Durational activities; instantaneous events	p_{27}	$p_3 \mid\sim p_{27}$
13	Responsibility vs. causation?	XG5, VI2–4, GE1	Agency is pragmatic ascription	p_{29}	$p_4 \mid\sim p_{29}$
14	`wasInformedBy`: shortcut?	VI1, VI6, EZ1	Independent: inferred, not reducible	p_{30}	$p_6 \mid\sim p_{30}$
15	Derivation criteria; identity?	XG8, VI5, VI6	Broad causal dependencies; subtypes narrow	p_{28}	$p_7 \mid\sim p_{28}$
16	Delegation responsibility?	XG11 (partial)	Hierarchical and transitive	p_{25}, p_{26}	$p_9 \mid\sim p_{25}, p_9 \mid\sim p_{26}$
17	Chain types equivalent?	VI1, VI6	No; `wasDerivedFrom` requires assertion	p_{24}	$p_{10} \mid\sim p_{24}$
—	Individuation $\Rightarrow$ Expanded	(from #11 follow-up)	Expanded Terms add expressiveness	p_{23}	$p_{18} \mid\sim p_{23}$

- Read $p_2 \mid\sim p_{18}$: one cannot commit to p_2 (fixed aspects) while denying p_{18} (context-relative); each accepted tension is one defeasible rule
- Rationale codes: requirements in Moreau et al.'s retrospective analysis; six of seven challenges land on one

Verification with an automated reasoner

- pyNMMS, our reasoner, over the 19 propositions and 9 implications: *does Y follow from X?*
- All 9 implications derivable; nothing both asserted and denied; 34 cross-chain pairs independent. Each pattern of results matches a documented design decision:

Derivability checks	Design decision (Moreau et al., 2015)
p_2 (fixed aspects) $\mid\sim p_{18}$ (context-relative) and $p_{18} \mid\sim p_{23}$ (Expanded Terms) hold, but $p_2 \not\mid\sim p_{23}$ does not: inferences do not chain (nontransitive)	EZ3: minimal core, separate extensions
$p_2 \mid\sim p_{18}$ holds, but $p_2, p_{23} \not\mid\sim p_{18}$ does not: adding p_{23} defeats it (nonmonotonic)	EZ1: "scruffy" to "proper" provenance
$p_7 \mid\sim p_{28}$ holds, but composes with no other chain	VI5, ISSUE-612: non-transitive derivation
No Entity commitment entails an Activity or Agent one (8 queries)	VI1, GE1: three independent views
The retracted p_{20} leaves no consequence behind	GE3: retraction leaves no trace

RDF serialization with provenance

```
:i1 a mb:MaterialImplication, prov:Entity ;
    rdfs:label "p2 |~ p18" ;
    mb:antecedent :p2 ; mb:consequent :p18 ;
    prov:wasGeneratedBy :move11 .
:move11 a mb:DialogueMove, prov:Activity ;
    rdfs:label "Challenge #11 resolved by refinement" ;
    prov:wasAssociatedWith :respondent, :opponent ;
    prov:endedAtTime
        "2026-02-09T00:00:00Z"^^xsd:dateTime .
```

One material implication and the dialogue move that generated it, from `material-base.ttl`

- Propositions are resources; each material implication is reified and `prov:wasGeneratedBy` the dialogue move that accepted it, which links to the GitHub issues
- Validates against PROV-CONSTRAINTS (ProvValidator); the report is archived with the data
- The vocabulary the case study examined describes the dialectic that examined it

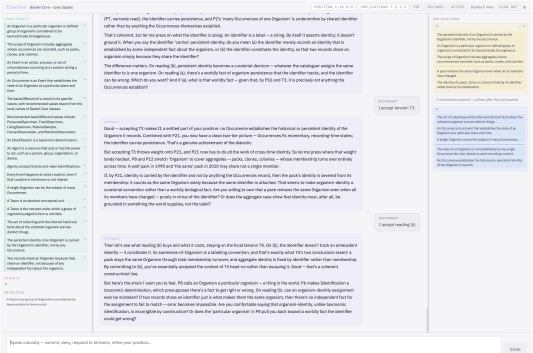
Limitations

- An **existence proof**, not an evaluation
- Opus 4.6 has almost certainly seen PROV's history: the dialectic **captures** rationale; this does not show that the LLM **discovers** it
- Oracle bias: a false positive costs one contest; **false negatives are invisible to both parties**; derivability is verified by the reasoner, not the LLM
- Cognitive overhead: arguing with an LLM opponent **can be fatiguing**
- One expert, one LLM: **multiple respondents are not supported**
- The base serializes to RDF, but it **is not an RDFS/OWL ontology**

Current work: dialectical states as RDFS/OWL ontologies

- The material base's atomic propositions can be RDF triples, and the pyNMMS reasoner extended to support RDF entailment regimes
- **Implication-space semantics for RDF**, after Hlobil (2026): simple, RDFS and OWL 2 RL entailment recovered *exactly* as the **monotone corner** of the base's semantics
- Regime closure and defeasible commitments in **one consequence relation**; no change to RDF, RDFS, OWL 2 RL or SHACL
- The ontology is the monotone fragment of the base; the base is its design rationale
- **Implementation**: pyNMMS reasoner integrated with a persistent triple store (Oxigraph) — ground derivability cost exponential only in the logical nesting of the query, not in graph size

Current work: user studies to evaluate utility and usability



The `elenchus-server` user interface

- **A Sloan Foundation–funded, pre-registered study**:
- 30–40 biodiversity scientists (domain experts without KE experience) each provide a conceptual model in natural language of a matched pair of topics
- Two counterbalanced conditions: free-form LLM conversation and Elenchus
- Model held fixed
- Five blinded expert judges rate the outputs on established criteria: correctness, coherence, concision
- Participants rate the system on usability

Summary

- Knowledge engineering through expert–AI dialogue: the expert never translates; the dialogue elicits the consequence relation
  - *Stable judgment, faithful rationale?* Never relied on: the only rationale recorded is the expert's; every move is versioned and replayable
  - *Aligned with human judgment?* Six of seven challenges match documented design tensions; the Sloan study will say more
  - *Accountable?* Nothing enters the base without a human move; every implication carries PROV-O provenance
- The deliverable: **a defensible, living record of why the ontology is as it is**
- Validation is bounded by what the expert commits to, not by what the LLM generates

Beyond Elenchus: knowledge engineering as explicitation

- Expertise is material inference: good by content, defeasible (Brandom, 1994; Sellars, 1953)
- The KG lifecycle (Hogan et al., 2021) as speech acts in discursive practice: creation is **assertion**; a shape is a standing **challenge**; refinement is **denial** or **withdrawal**; publication is **scorekeeping**; provenance is **deferral**
- A knowledge graph records what a community's discursive practice takes as good inferences; it is not a mirror of the world
- Knowledge engineering makes those inferences explicit and accountable
- Elenchus is a first step towards an **inferentialist reconceptualization** of knowledge engineering, answering the call of Hitzler et al. (2010) for a "semantics for shared inference"

References i

AIAA4KE Organizing Committee (2026). ***1st Workshop on AI-assisted Approaches to Knowledge Engineering: Call for Papers.*** https://sites.google.com/view/aiaa4ke/.

Alharbi, Reham et al. (2026). ***Characterising LLM-Generated Competency Questions: a Cross-Domain Empirical Study using Open and Closed Models.*** arXiv:2604.16258.

Brandom, Robert (1994). ***Making it explicit: Reasoning, representing, and discursive commitment.*** Harvard University Press.

Dutilh Novaes, Catarina (2020). ***The dialogical roots of deduction: Historical, cognitive, and philosophical perspectives on reasoning.*** Cambridge University Press. doi: 10.1017/9781108800792.

Garijo, Daniel (2026). ***How Good Are Large Language Models at Ontology Engineering?*** Invited talk, SEMANTICS 2026. doi: 10.5281/zenodo.22809520.

Hitzler, Pascal and Frank van Harmelen (2010). **"A reasonable Semantic Web".** In: *Semantic Web* 1.1–2, pp. 39–44. doi: 10.3233/SW-2010-0010.

Hlobil, Ulf (2026). **"First-order implication-space semantics".** In: *Journal of Philosophical Logic* 55.3, pp. 529–554. doi: 10.1007/s10992-026-09841-x.

Hlobil, Ulf and Robert B Brandom (2025). ***Reasons for logic, logic for reasons: Pragmatics, semantics, and conceptual roles.*** Routledge.

Hogan, Aidan et al. (2021). **"Knowledge Graphs".** In: *ACM Computing Surveys* 54.4, pp. 1–37. doi: 10.1145/3447772.

Huang, Jiacheng et al. (2026). **"Socratic Elenchus-inspired multi-agent debate for mitigating hallucinations in large language models".** In: *Expert Systems with Applications* 320, p. 132208. doi: 10.1016/j.eswa.2026.132208.

Huang, Lei et al. (2023). **"A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions".** In: *arXiv preprint arXiv:2311.05232*.

Koutsiana, Elisavet et al. (2026). **"Knowledge prompting: How knowledge engineers use generative AI".** In: *Journal of Web Semantics* 88, p. 100873. doi: 10.1016/j.websem.2025.100873.

Lebo, Timothy, Satya Sahoo, and Deborah McGuinness (2013). ***PROV-O: The PROV Ontology.*** W3C Recommendation. W3C. url: https://www.w3.org/TR/prov-o/.

References ii

Li, Jiayi, Daniel Garijo, and María Poveda-Villalón (2026a). **"Large Language Models for Ontology Engineering: A Systematic Literature Review".** In: *Data Intelligence*. Published online 13 July 2026. doi: 10.1177/22104968261465514.

Li, Jiayi et al. (2026b). **"MASEO: A Multi-Agent System for Explainable Ontology Generation".** In: *Proceedings of the 1st Workshop on LLM-driven Knowledge Graph and Ontology Engineering (LLM4KGOE 2026), co-located with ESWC 2026*. Vol. 4246. CEUR Workshop Proceedings. CEUR-WS.org. url: https://ceur-ws.org/Vol-4246/llms4kgoe-7.pdf.

Lippolis, Anna Sofia et al. (2026). ***OntoExtend: A Framework for Requirement-driven and Scalable Ontology Extension with LLMs.*** arXiv:2607.17963; also doi:10.3233/SSW260011.

Miller, Tim (2023). **"Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI".** In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Chicago, IL, USA: ACM, pp. 333–342. doi: 10.1145/3593013.3594001.

Moreau, Luc et al. (2015). **"The rationale of PROV".** In: *Journal of Web Semantics* 35, pp. 235–257.

Restall, Greg (2005). **"Multiple conclusions".** In: *Logic, methodology and philosophy of science: Proceedings of the twelfth international congress*. Kings College Publications, pp. 189–205.

Sellars, Wilfrid (1953). **"Inference and meaning".** In: *Mind* 62.247, pp. 313–338.

Tian, Harry Yizhou, Hasan Amin, and Ming Yin (2026). **"Understanding the Effects of AI-Assisted Critical Thinking on Human-AI Decision Making".** In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. ACM, pp. 1–30. doi: 10.1145/3772318.3790785.

Zhang, Bohui et al. (2025). **"OntoChat: A Framework for Conversational Ontology Engineering Using Language Models".** In: *The Semantic Web: ESWC 2024 Satellite Events*. Vol. 15344. Lecture Notes in Computer Science. Cham: Springer, pp. 102–121. doi: 10.1007/978-3-031-78952-6_10.

Zhao, Yihang et al. (2024). **"Improving ontology requirements engineering with OntoChat and participatory prompting".** In: *Proceedings of the AAAI Symposium Series*. Vol. 4:1, pp. 253–257.