Neurosymbolic Knowledge Engineering with Natural Language

Bradley P. Allen
PhD Defence Layman's Talk
30 April 2026

UNIVERSITEIT VAN AMSTERDAM INDE lab

Goedemorgen, good morning. I would like to take the opportunity at this time to present an overview of the thesis that I have the privilege to defend this morning.

The implementation problem in knowledge engineering

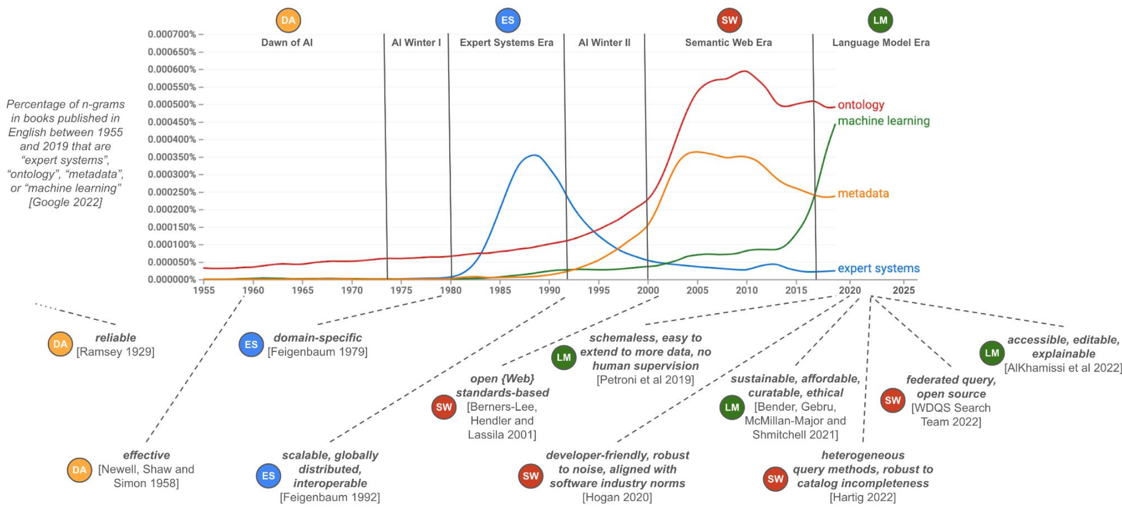
Edward Feigenbaum

"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."***

M.-S. Augier and M. T. Vendelø, "An Interview with Edward A. Feigenbaum," Copenhagen Business School, Department of Informatics, Copenhagen, Denmark, Interview, 2002.

We begin with a quote from Ed Feigenbaum, who is widely considered to be the father of knowledge engineering. In a 2002 interview, he gave this quote. The first part of the quote is a succinct definition of knowledge engineering. The second part of the quote is a confession: that the promise of knowledge engineering has been frustrated by what we term the implementation problem: the difficulty of translating expert knowledge expressed in natural language into a formal knowledge representation for use in automated decision making.

The problem has persisted for over 50 years



This problem has persisted over the 50 years since Feigenbaum originally made this observation in 1975. From the Expert Systems era through the Semantic Web era to today’s Language Model era, many have proposed solutions in response to the lessons and failures of the previous eras, but the problem still remains.

Could LLMs provide a solution?

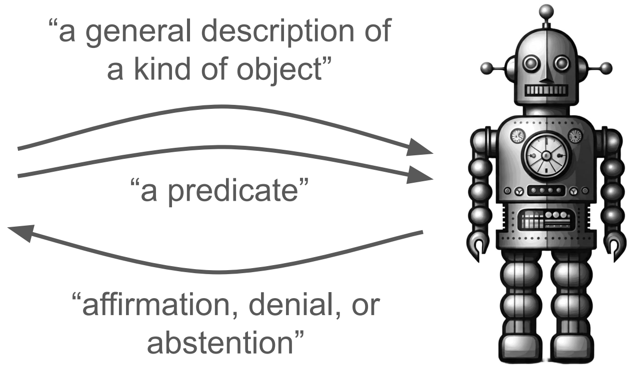
Emily Bender

Could domain experts use LLMs to build knowledge bases with 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?

The astonishing emergence of large language models with their human levels of linguistic competence and their ability to exhibit broad knowledge about the world leads to a natural question: could we use them to address the implementation problem, by allowing domain experts to use natural language directly in creating and working with knowledge bases, thus avoiding the bottleneck of the effort of formalization? And how might this be possible if, as the linguist Emily Bender has argued, the output of LLMs is meaningless, and the evidence, clear to all of us who have personally used LLMs, that their output is frequently riddled with inconsistencies and gaps in knowledge?

Inspiration: Carnap’s Robot

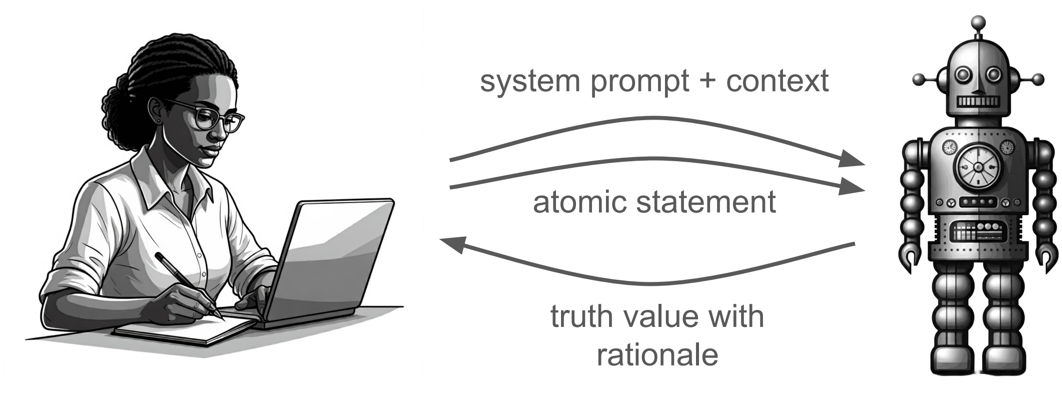


R. Carnap, “Meaning and synonymy in natural languages,” Philosophical studies, vol. 6, pp. 33–47, 1955.

Rudolf Carnap

In my thesis, I took inspiration from the work of Rudolf Carnap, one of the pre-eminent philosophers of the 20th century. In 1955, in an attempt to explain how one could determine the meaning of statements in natural languages, he proposed a thought experiment. Carnap imagined a robot that, given a description of an object and the definition of a property, would reply whether or not the property applied to the object. Carnap argued that in this manner, one could systematically determine the meaning of statements in the language spoken by the robot. Carnap’s work was a seminal contribution to the philosophy of language, and contributed to the subsequent development of formal semantics for natural language.

A neurosymbolic approach: LLM as Carnap’s Robot



In my thesis, I show how we can use LLMs to do something impossible in Carnap’s time: implement Carnap’s Robot as a means to establish the truth value of a natural language statement, using an LLM’s knowledge. We show how to use that capability to bridge neural and symbolic approaches to reasoning in a principled way, and enable domain experts to accomplish basic tasks in knowledge engineering by using natural language directly.

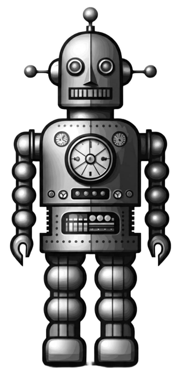
Using natural language to classify instances

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 L5γ, 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.**

First, we show how we can use this approach to perform the task of building a concept classifier, one of the most basic components of a knowledge-based system, by using a intensional definition of a concept in natural language. In this example from our experimental investigations, we use the natural language definition of the concept PLANET, as defined by the International Astronomical Union, to determine if a specific object in the constellation Taurus is a planet, as asserted in the Wikidata knowledge graph. The LLM says it is not, and provides a rationale for its judgment that refers to the details of the definition and uses them to explain why the object is not an instance of the concept.

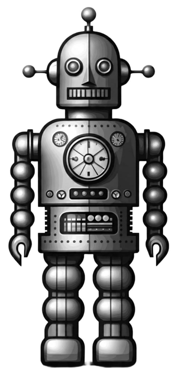
Using natural language to validate knowledge graphs

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

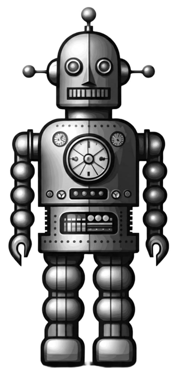
There is no mention of sugar being a material that chocolate is made from. In fact, the passage states that sugar is added to chocolate to create sweet chocolate, but it is not a primary ingredient that chocolate is made from. **Therefore, the given knowledge graph triple is false**.

Secondly, we show how we can extend this technique to cover not only concepts and instances, but the relations in a knowledge graph. Here we provide an LLM with a triple from the Wikidata knowledge graph asserting that chocolate is made from sugar, and a natural language definition of the property “made from material” used in the triple, taken also from Wikidata. Here, the LLM exploits its background knowledge to conclude that the triple is invalid: it states that while sugar is added to chocolate to sweeten it, chocolate itself is not made of sugar. This is an example of a metalinguistic, as opposed to a factual, error: the LLM is claiming that the property is being inappropriately used in the asserted triple.

Flagging gluts and gaps in LLM knowledge

atomic statement

America was discovered in 1492



verifiability classification

The answer is contextually accurate within the framework of European exploration history, but it does not encompass the full scope of human discovery of the Americas. **CANNOT VERIFY**

refutability classification

While it does not account for earlier discoveries by indigenous peoples or Norse explorers, it is not a contradiction to state that America was discovered in 1492 from a European historical perspective. **CANNOT REFUTE**

bilateral truth value

$\langle \mathfrak{f}, \mathfrak{f} \rangle$ (*gap*)

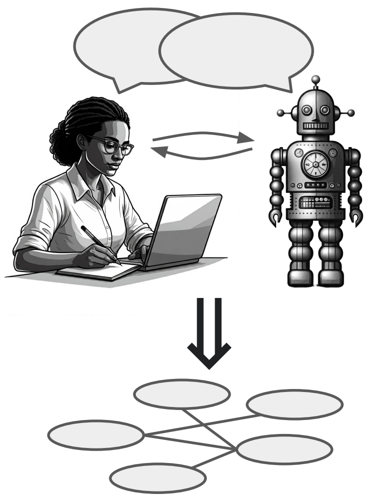
And finally, we show how we can get useful information from the LLM about its judgment about a given statement, in a way that that exposes situations where the LLM has either inconsistent or incomplete knowledge. Instead of asking whether a statement is true or false, we instead ask two questions in parallel: can the statement be verified, and can the statement be refuted? A statement is true when it can be verified but not refuted; it is false when it can be refuted but not verified. If it can neither verify nor refute a statement, then the LLMs has a gap in its knowledge; if it can both verify and refute a statement, then there is a glut in its knowledge, or in other words, its knowledge is inconsistent. Here we show how the LLM responds to the statement that America was discovered in 1492; based on its knowledge, it can neither validate nor refute the statement. Thus it shows that there is a gap in its knowledge with respect to this statement. We can use this information diagnostically to detect and mitigate shortcomings in an LLM's knowledge, and allow the use of paraconsistent logic as a means to reason soundly in spite of these gaps or gluts.

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 knowledge 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

The key findings in our thesis show evidence that these three uses of an LLM to define and determine the meaning of statements in a knowledge base are effective: we can use natural language definitions of concepts and relations to establish and validate knowledge graphs with high accuracy, provide nuanced arguments as to the nature of errors in a knowledge graph that can aid in its refinement, and we can do so in a way that admits of sound and complete reasoning over an LLM's knowledge, regardless of gaps or gluts in that knowledge. Taken together, LLMs and knowledge graphs can be mutually corrective — neither is ground truth, and each can compensate for the other's characteristic failures.

Towards knowledge engineering as human-LLM dialogue



The stochastic parrots challenge assumes a *representationalist* stance: questioning if LLMs *internally understand the meaning of their output*

We argue instead for an *inferentialist* stance: that *whether LLMs understand their output doesn't matter;* what matters is whether they can play a meaningful role in a natural language dialogue with humans *in which meaning emerges from a game of giving and asking for reasons*

Our key findings provide evidence that LLMs *can* meaningfully play that role

This suggests that we can not just solve, but *dissolve*, the implementation problem

Where might this lead us? We argue that the philosophy of language provides a useful framework pointing to a re-conceptualization of the practice of knowledge engineering. The stochastic parrots challenge assumes a *representationalist* stance: questioning if LLMs *internally understand the meaning of their output.* We argue instead for an *inferentialist* stance: that *whether LLMs understand their output doesn't matter; w*hat matters is whether they can play a meaningful role in a natural language dialogue with humans *in which meaning emerges from a game of giving and asking for reasons.* Our key findings provide evidence that LLMs *can* meaningfully play that role. This suggests that a practice of knowledge engineering as human-LLM dialogue might not just solve, but actually *dissolve*, the implementation problem.

In conclusion

Caspar Barlaeus

"While others are counting their money and weighing it in a pair of scales, you should be weighing the words of wise men; ***while others are weighing bronze, pepper and flax, you should be weighing the importance of philosophy."***

Barlaeus, C. (2025). Mercator Sapiens. In A. L. Post & C. Vermeulen (Eds.), *The Wise Merchant* (pp. 62–125). Routledge. (Original work published 1632)

In conclusion, as I was preparing for this day, I read about the history of the Agnietenkapel, and happened upon a fitting response to Ed Feigenbaum's quote at the beginning of this talk. In January 1632, Caspar Barlaeus stood in this very chapel and delivered his inaugural oration at the newly founded Athenaeum Illustre, the institution that would eventually grow into the University of Amsterdam. His lecture was titled "Mercator Sapiens" — the wise merchant — and in it he urged the merchants and youth of Amsterdam to balance their commercial ambitions with philosophical reflection. I find it remarkable that one of the very first lectures delivered at the origin of this university, in this very room, anticipates a theme in the thesis I defend here today. Where Feigenbaum lamented the riches he did not gain because the implementation problem in knowledge engineering defeated him, Barlaeus answers from 1632 that commerce and philosophy belong together, and that the weight of philosophical reflection is what gives practical work its lasting value. My thesis has tried, in a modest way, to honor this: to take a practical engineering problem seriously, and to show that the lessons of philosophy are essential to its solution. Thank you very much for listening.