



UNIVERSITEIT  
VAN AMSTERDAM

INDE lab

# Conceptual Engineering Using Large Language Models

Bradley P. Allen

PhAI 2023, Erlangen

2023-12-16

# How are philosophy and AI alike?

For the past 400 years, the domain of philosophy has been shrinking. ... Physics, geology, chemistry, economics, biology, anthropology, sociology, meteorology, psychology, linguistics, computer science, cognitive science. Each of those subject matters was a part of philosophy a mere 400 years ago.

Kevin Scharp. Philosophy as the study of defective concepts. In Conceptual engineering and conceptual ethics, pages 396–416. Oxford University Press, 2020.

... the promoters of AI may not have delivered exactly what they announced ... but what they delivered is astounding. Many breakthroughs in computer science ... can directly be traced to work in AI. Part of the problem is a phenomenon that I heard John McCarthy himself describe: *“As soon as it works, no one calls it AI anymore.”*

Bertrand Meyer. "John McCarthy". <https://cacm.acm.org/blogs/blog-cacm/138907-john-mccarthy/fulltext>, 2011. Accessed: 2023-03-03.

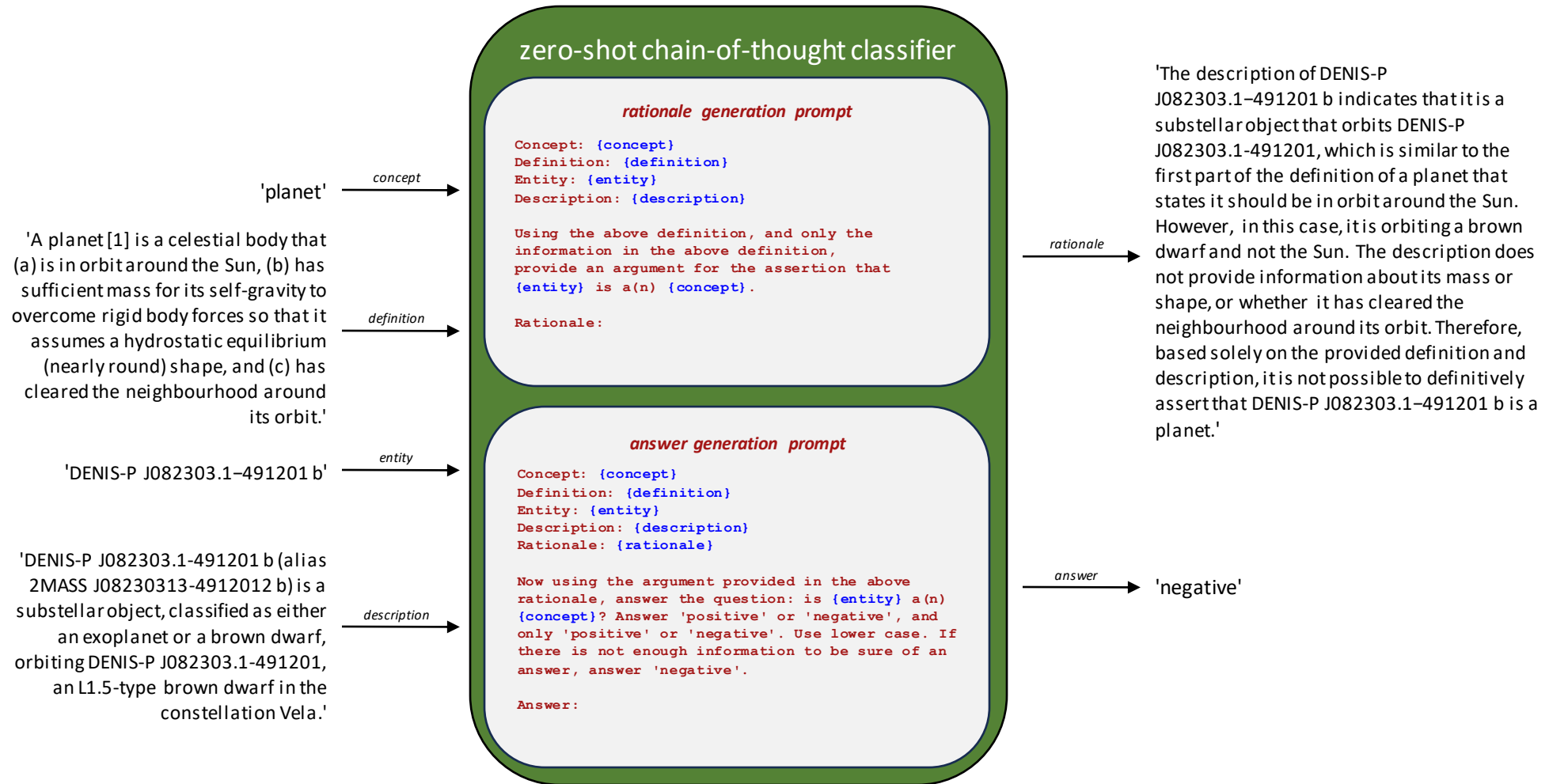
# How are conceptual engineering and knowledge engineering alike?

- Philosophers have different ideas about the best way to do conceptual engineering (CE), but a common element is the definition and analysis of concepts using natural language
- Knowledge engineering (KE) focuses the transformation of knowledge (including concept definitions) expressed in natural language into a structured formal language suitable for automated reasoning
- CE and KE share the same task: *(re)defining concepts to achieve a normative goal*
- We assert that large language models (LLMs) raise the possibility that CE has relevance for the practice of KE, and vice versa
- This work is an initial exploration of one side of that potential relationship

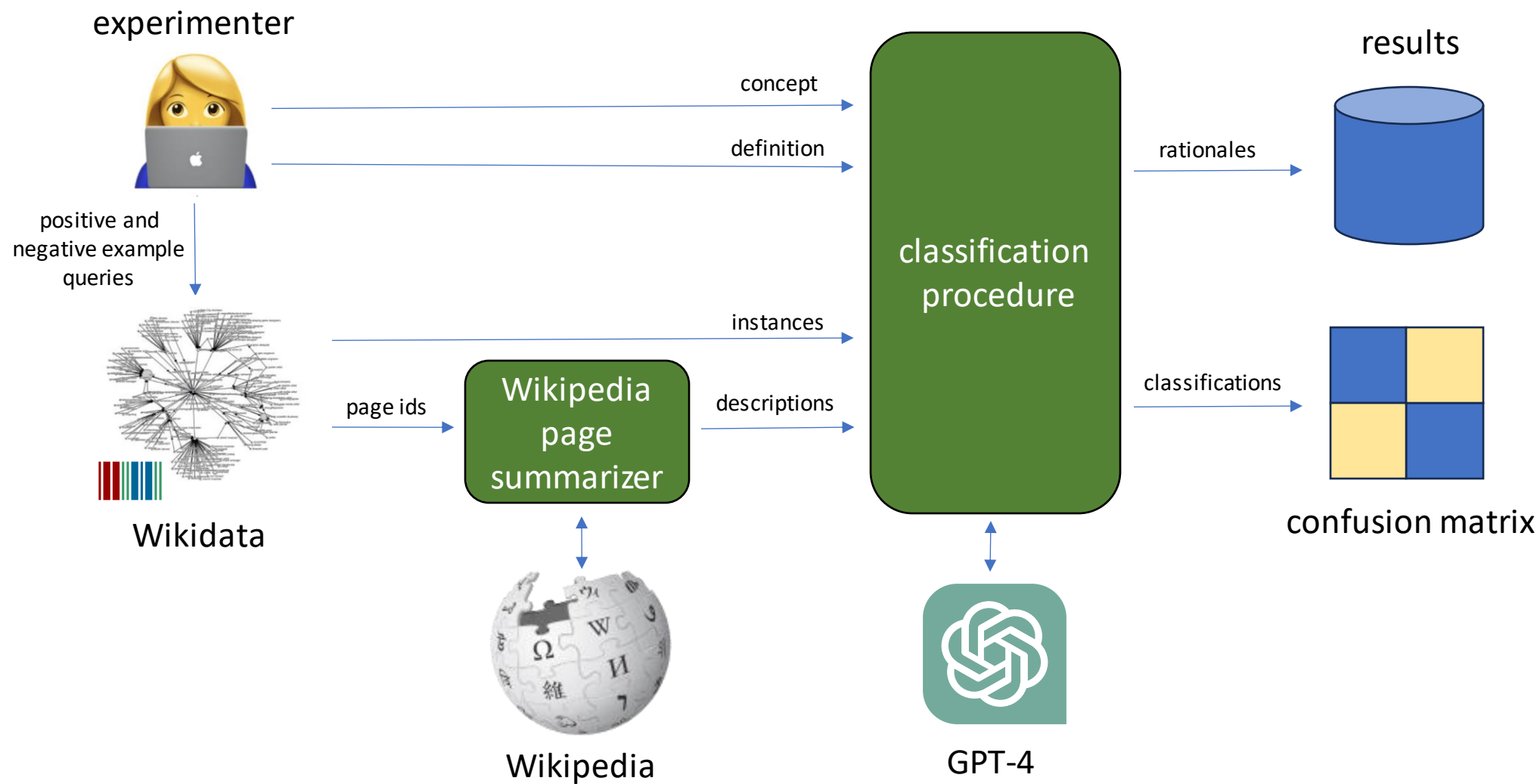
# Approach

- An important question for a theory of CE is the nature of its targets, i.e., "what conceptual engineers are (or should be) trying to engineer"
- (Nado 2023) proposes as targets *classification procedures* (CPs), defined as abstract 'recipes' which sort entities "into an 'in'-group and an 'out'-group"
- We build on this idea by defining a method that uses prompt engineering of LLMs to implement CPs using natural language intensional definitions of concepts
- We then evaluate CPs built using this method by leveraging a *knowledge graph* (KG) as a source of positive and negative examples of elements in the extension of a concept

# Example



# Experimental framework



# Results

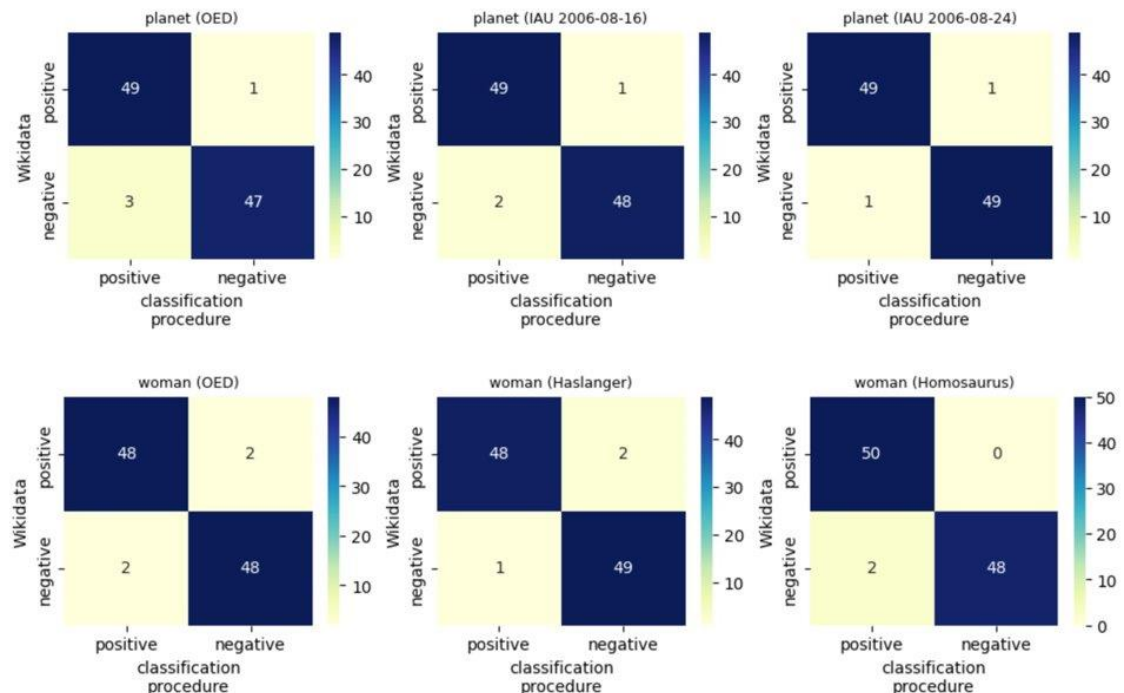
- PLANET

- Definitions: OED, Haslanger, Homosaurus
- Positive examples: 50 instances with sex or gender (P21) either female (Q6581072) or trans woman (Q1052281)
- Negative examples: 50 instances with sex or gender either male (Q6581097), non-binary (Q48270), or trans man (Q2449503)

- WOMAN

- Definitions: OED, IAU 2006-08-16, IAU 2006-08-24
- Positive examples: 50 instances (P31) of planet (Q634)
- Negative examples: 50 instances of substellar object (Q3132741) that are not instances of planet

concept	definition	Cohen's kappa	F1 macro	FN	FP
PLANET	IAU 2006-08-24	<b>0.96</b>	<b>0.98</b>	1	1
	IAU 2006-08-16	0.94	0.97	1	2
	OED	0.92	0.96	1	3
WOMAN	Homosaurus	<b>0.96</b>	<b>0.98</b>	0	2
	Haslanger	0.94	0.97	2	1
	OED	0.92	0.96	2	2

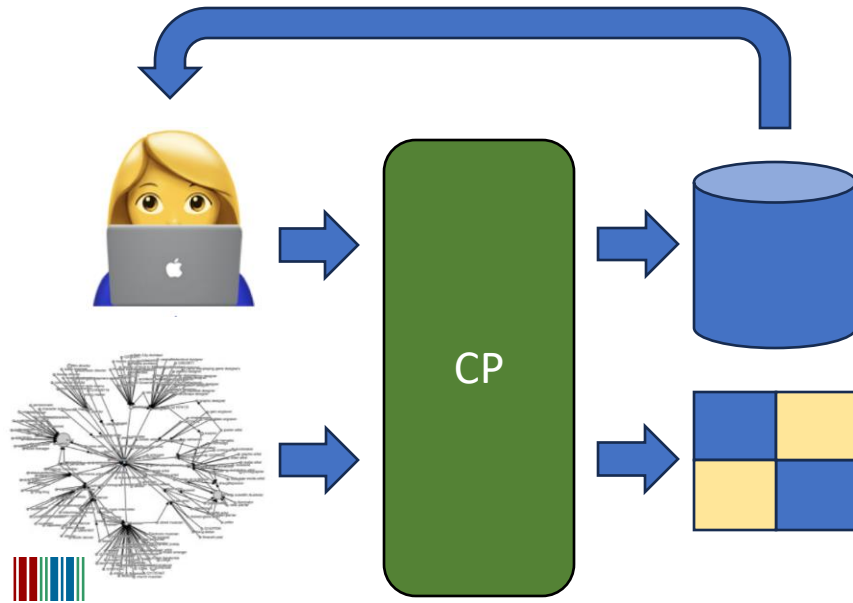


# Findings

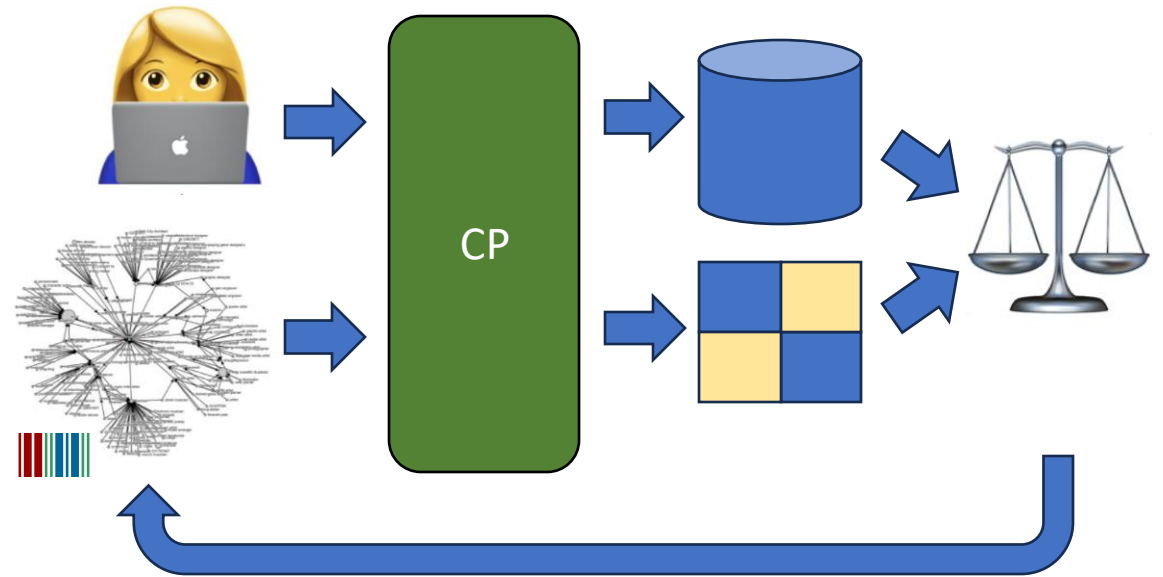
- Rationales generated by the classification procedures were sound
- Answers were faithful to their rationales
- Rationales frequently contained statements about issues with definitions or descriptions
- For PLANET, most errors were false positives relating to trans-Neptunian objects
- For WOMAN:
  - Homosaurus performed best, possibly because it is the most inclusive definition
  - Haslanger had two false negatives, the rationales of which ascribed to entity descriptions lacking evidence of systematic subordination



# Two potential uses



Generative AI assistance  
for conceptual engineers



Ameliorative analysis of  
knowledge graphs

# Discussion

- We've shown how a CE project can incorporate an empirical, data-driven activity using LLMs and KGs
- However, the use of LLMs as they exist today raises transparency, reproducibility and safety concerns
- Further work is needed to evaluate our method with respect to these issues, with a specific focus on evaluating explanation faithfulness



UNIVERSITEIT  
VAN AMSTERDAM

INDE lab

**Thank you!**

GitHub repository: <https://github.com/bradleypallen/zero-shot-classifiers-for-conceptual-engineering>