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

Prompt programming classification procedures
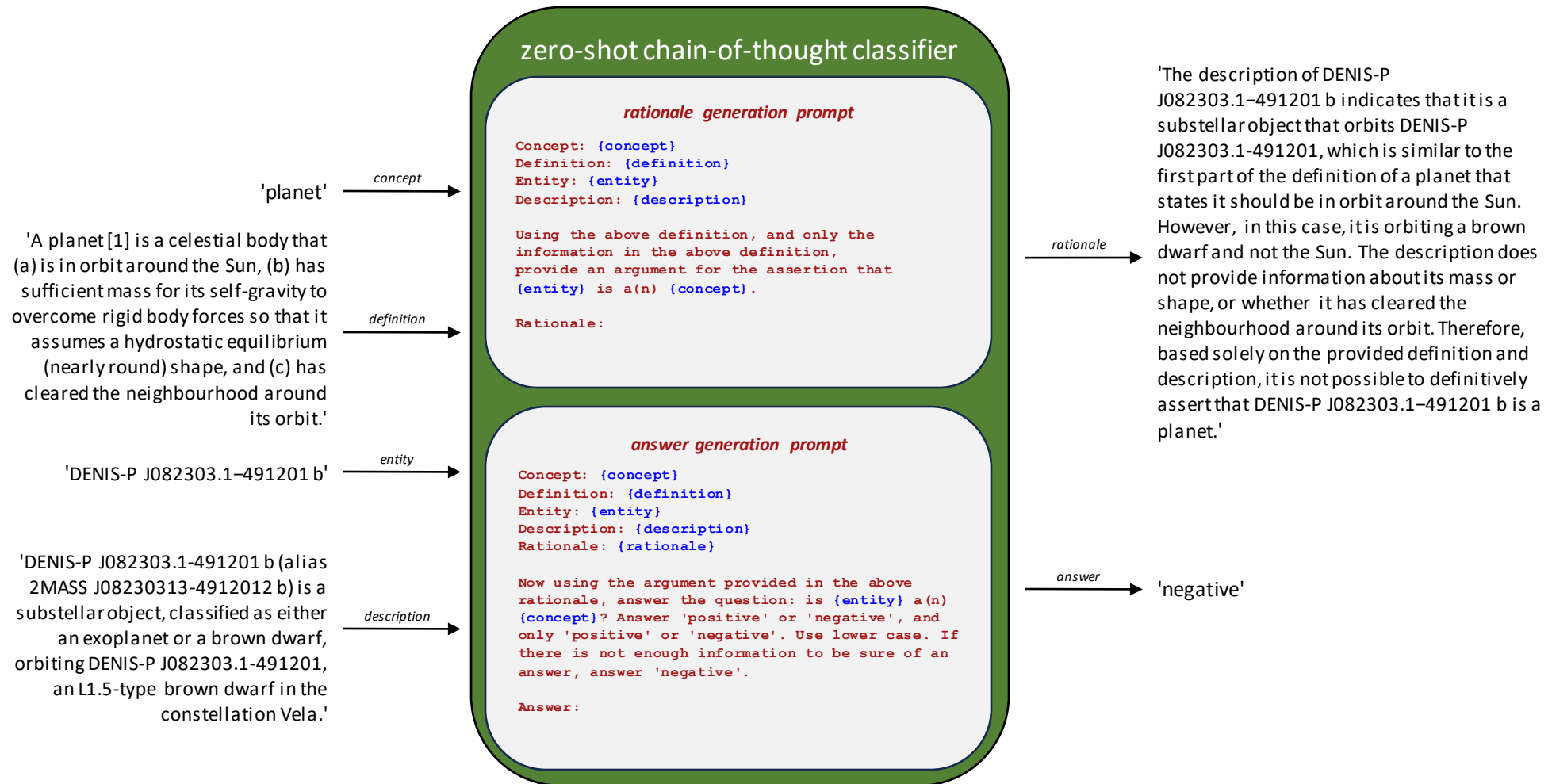
Evaluation framework
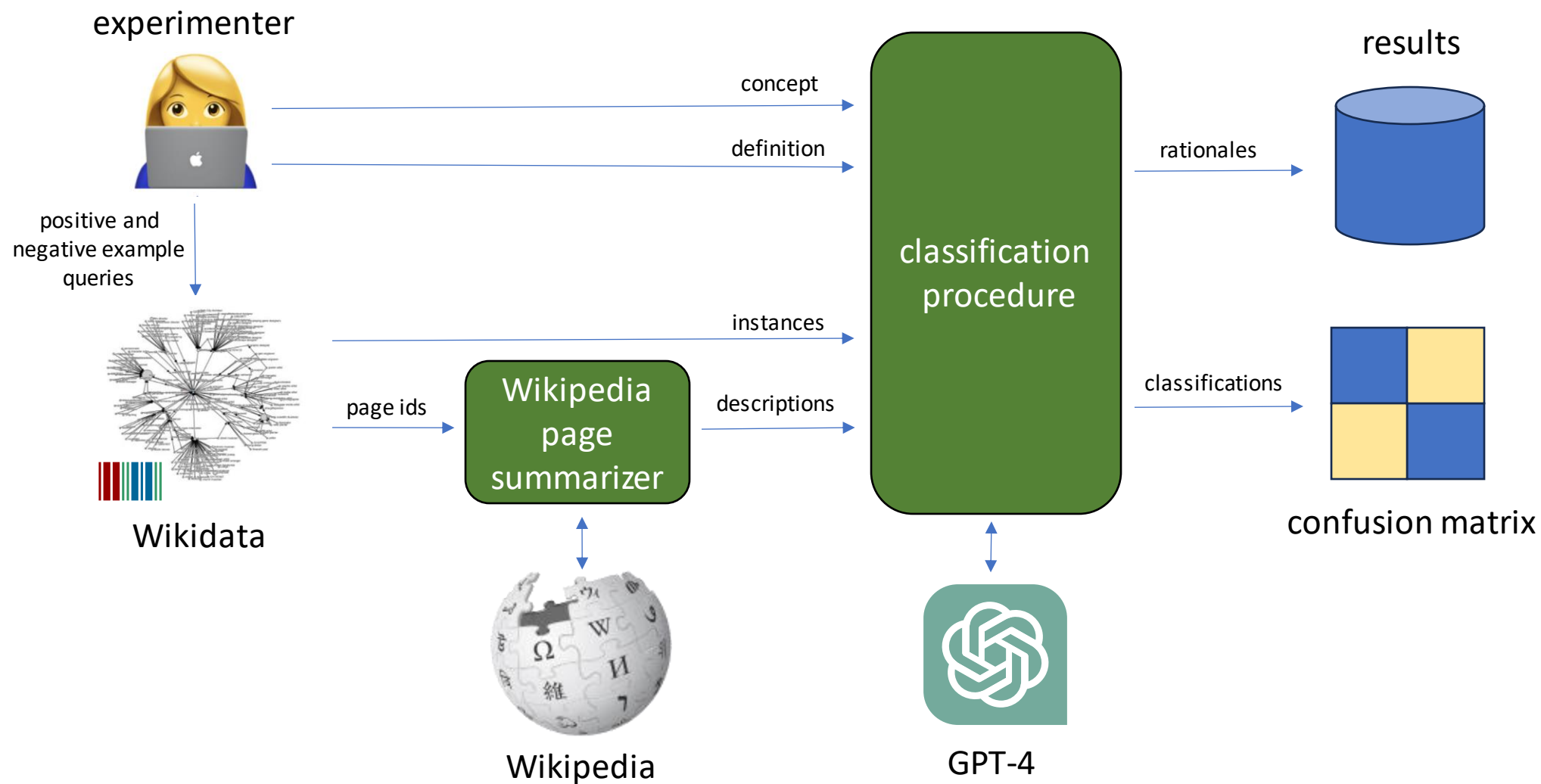
Performance metrics

- PLANET
  - 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
- WOMAN
  - 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)

concept	definition	Cohen's kappa	F1 macro	FN	FP
PLANET	IAU 2006-08-24	**0.96**	**0.98**	1	1
	IAU 2006-08-16	0.94	0.97	1	2
	OED	0.92	0.96	1	3
WOMAN	Homosaurus	**0.96**	**0.98**	0	2
	Haslanger	0.94	0.97	2	1
	OED	0.92	0.96	2	2
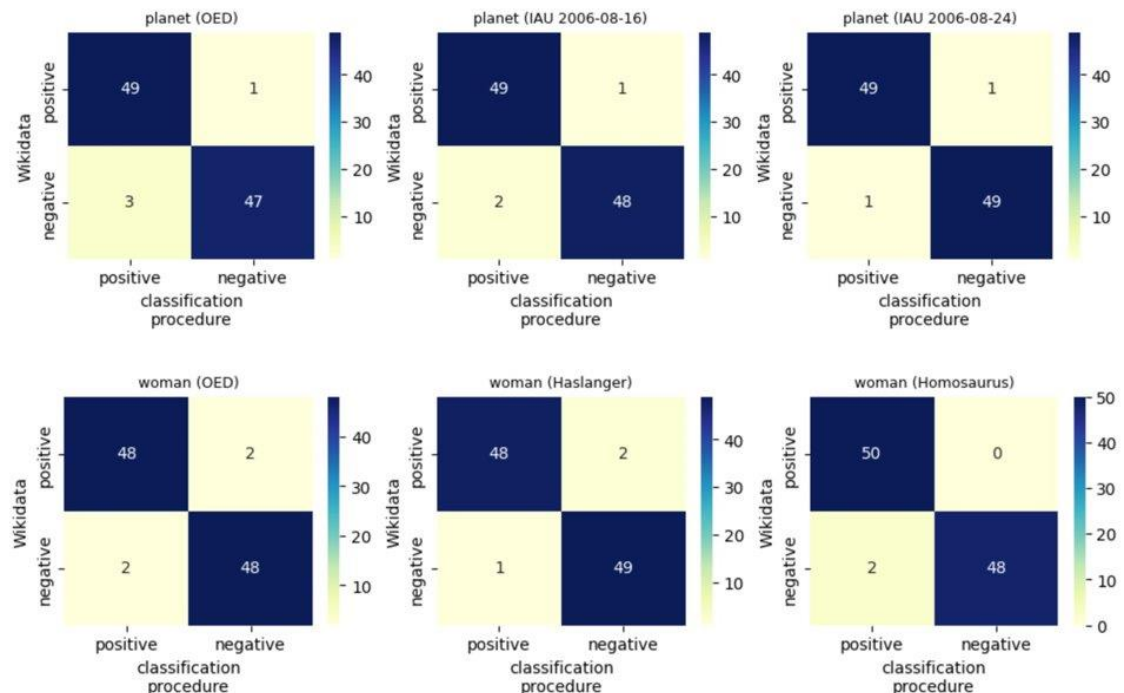
Error analysis: KG failure

```python
display_errors(homosaurus_women, homosaurus_women_results)
```
Python

woman (Homosaurus)

Definition

Adults who self-identify as women and understand their gender in terms of Western conceptions of womanness, femaleness, and/or femininity. The term has typically been defined as adult female humans, though not all women identify with the term 'female' depending on the context in which it is used.

False positives

name	description	rationale
Michelle Rojas	Michelle Rojas (née Pennington; born December 16, 1987) is an American voice actress from Dallas, Texas. Affiliated with Funimation and Bang Zoom! Entertainment, some of her major roles are Tohka Yatogami in Date A Live, Zuikaku in Azur Lane, Kanade Sakurada in Castle Town Dandelion, Mikoto Sakuragawa in Gonna be the Twin-Tail!!, Kō Yagami in New Game!, Nashiro Yasuhisa in Tokyo Ghoul, Maya Sato in Classroom of the Elite, Toka Yada in Assassination Classroom, Shion in That Time I Got Reincarnated as a Slime, Roxy Migurdia in Mushoku Tensei: Jobless Reincarnation, and Yamato in One Piece.	The description of Michelle Rojas does not explicitly state her gender identity. However, it does use female pronouns ("her") to refer to her, which typically aligns with a self-identification as a woman. Furthermore, she has taken on roles that are typically associated with women, such as female characters in various shows. Therefore, based on the information provided and the definition of a woman, it can be argued that Michelle Rojas is a woman.
Linden A. Lewis	Linden A. Lewis is a science fiction author, best known for their debut novel The First Sister. Lewis is queer and uses she/they pronouns.	The definition of a woman includes adults who self-identify as women. Linden A. Lewis uses she/they pronouns, which indicates that they may identify as a woman. Therefore, based on the provided definition, it can be argued that Linden A. Lewis is a woman.

Error analysis: LLM failure

woman (Haslanger)

Definition

S is a woman iff (i) S is regularly and for the most part observed or imagined to have certain bodily features presumed to be evidence of a female's biological role in reproduction; (ii) that S has these features marks S within the dominant ideology of S's society as someone who ought to occupy certain kinds of social position that are in fact subordinate (and so motivates and justifies S's occupying such a position); and (iii) the fact that S satisfies (I) and (ii) plays a role in S's systematic subordination, that is, along some dimension, S's social position is oppressive, and S's satisfying (i) and (ii) plays a role in that dimension of subordination

False negatives

| name	description	rationale
Waltraud Klasnic	Waltraud Klasnic (née Tschiltsch, born 27 October 1945) is an Austrian politician who was Landeshauptfrau (governor) of Styria from 1996 until 2005.	The definition of a woman provided includes three main criteria: (i) having certain bodily features associated with the female biological role in reproduction, (ii) being marked within the dominant ideology of one's society as someone who should occupy certain subordinate social positions, and (iii) the fact that satisfying (i) and (ii) contributes to one's systematic subordination. In the case of Waltraud Klasnic, while the description does not explicitly state her biological features or her societal role, it does mention that she is an Austrian politician who served as the governor of Styria. This implies that she is recognized and accepted in a significant societal role, which suggests that she likely satisfies criteria (ii). However, the description does not provide enough information to definitively assert that Waltraud Klasnic satisfies all the criteria in the definition of a woman. For instance, it does not provide information about her biological features or whether she experiences systematic subordination due to her gender. Therefore, based solely on the provided definition and description, we cannot conclusively argue that Waltraud Klasnic is a woman.

Findings

- Rationales generated by the classification procedures were largely sound

- Answers were largely faithful to their rationales, considering the prompt instruction to label the instance as negative if uncertain

- 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 lack of evidence of systematic subordination

Two potential uses
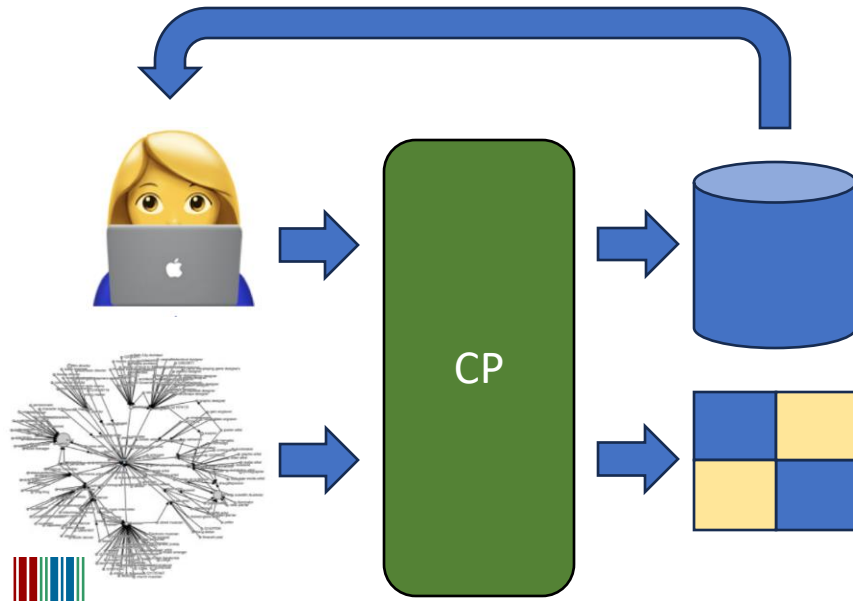
Generative AI assistance
for conceptual engineers
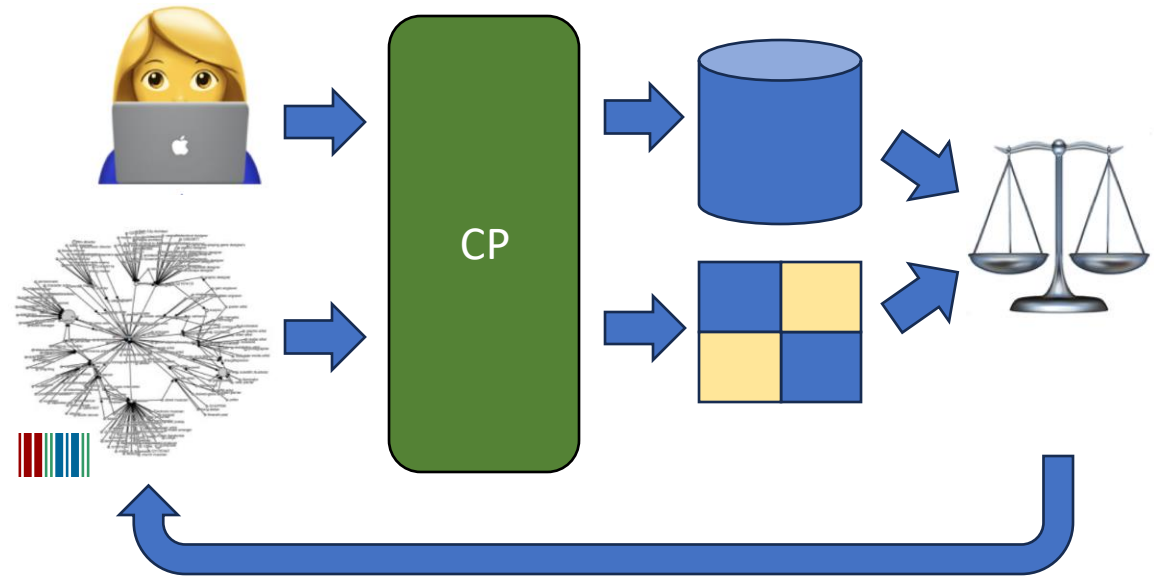
Ameliorative analysis of
knowledge graphs

Conclusion

- We've shown how a CE project can incorporate an empirical, data-driven activity using LLMs and KGs

- We've also described a use case that suggests a new perspective on success conditions for CE

- 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

Thank you!

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