Universiteit van Amsterdam

INDE lab

Carnap's Robot Redux:
LLMs, Intensional Semantics, and the Implementation Problem in Conceptual Engineering

Bradley P. Allen

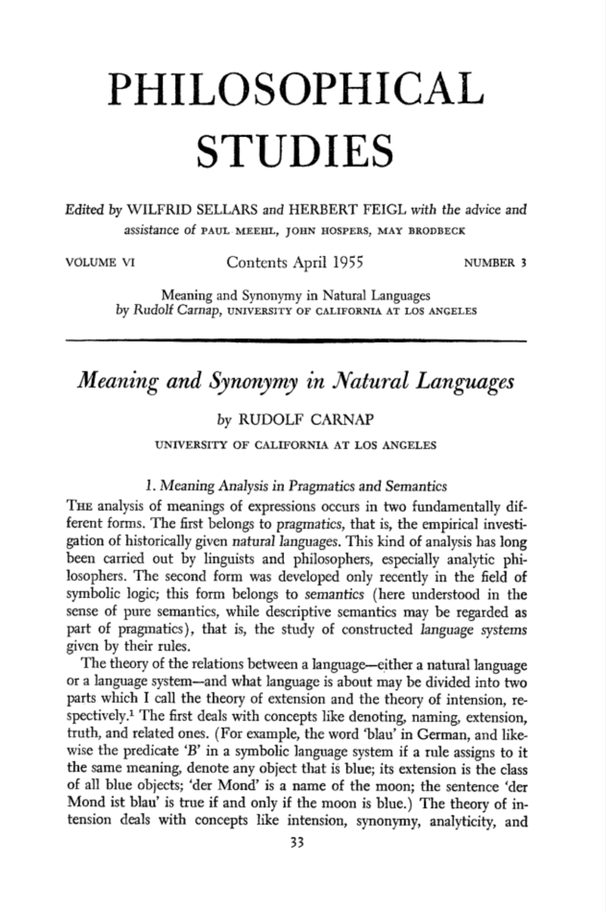
LLMs and Philosophy, Kanazawa

2024-09-27

Overview

- Carnap's Robot Redux
  - (Carnap 1955) through the lens of LLMs and mechanistic interpretability
  - Intensional semantics based on zero-shot chain-of-thought classifiers
- Experiments in conceptual engineering and knowledge engineering
  - Intensions-as-classifiers as targets of conceptual engineering
  - Intensions-as-classifiers for the evaluation of knowledge graphs
- LLMs as cognitive tools for philosophy
  - For experimental conceptual engineering
  - For ameliorative analysis of knowledge bases as a solution for the implementation problem in conceptual engineering
- Future work on LLMs and meaning
  - LLMs, metalinguistic disagreement, hyperintensionality, and two-component semantics

(Carnap 1955)



PHILOSOPHICAL
STUDIES

Edited by WILFRID SELLARS *and* HERBERT FEIGL *with the advice and assistance of* PAUL MEEHL, JOHN HOSPERS, MAY BRODBECK

VOLUME VI Contents April 1955 NUMBER 3

Meaning and Synonymy in Natural Languages
by Rudolf Carnap, UNIVERSITY OF CALIFORNIA AT LOS ANGELES

Meaning and Synonymy in Natural Languages

by RUDOLF CARNAP

UNIVERSITY OF CALIFORNIA AT LOS ANGELES

1. *Meaning Analysis in Pragmatics and Semantics*

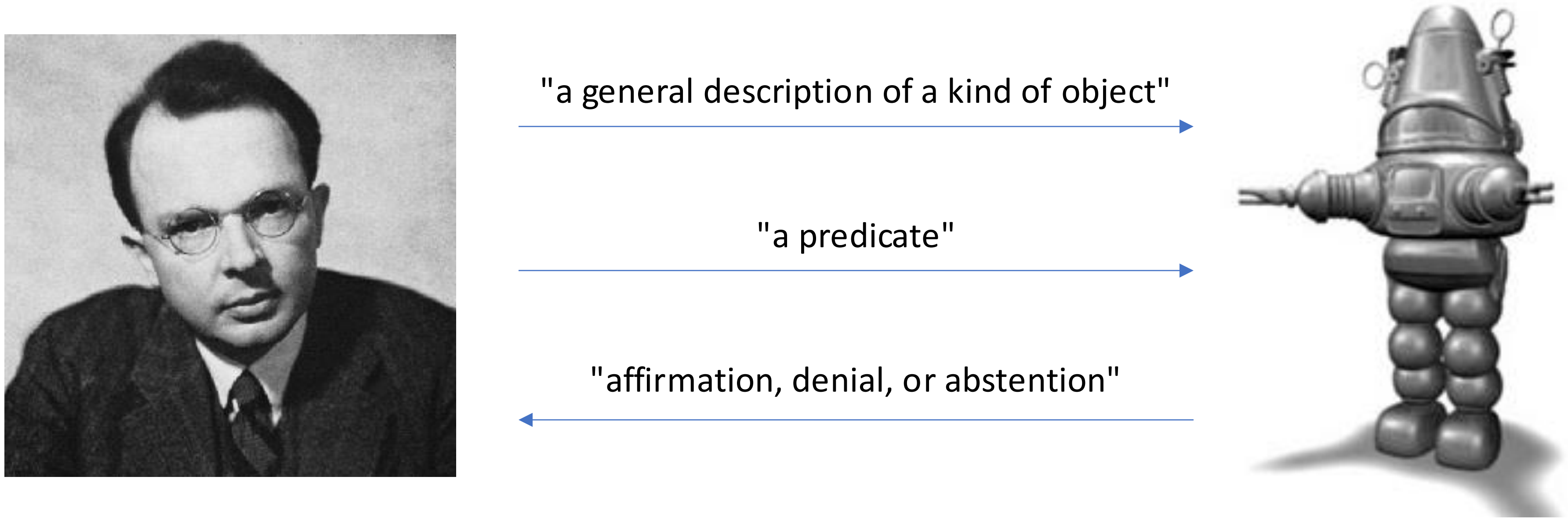
THE analysis of meanings of expressions occurs in two fundamentally different forms. The first belongs to *pragmatics*, that is, the empirical investigation of historically given *natural languages*. This kind of analysis has long been carried out by linguists and philosophers, especially analytic philosophers. The second form was developed only recently in the field of symbolic logic; this form belongs to *semantics* (here understood in the sense of pure semantics, while descriptive semantics may be regarded as part of pragmatics), that is, the study of constructed *language systems* given by their rules.

The theory of the relations between a language—either a natural language or a language system—and what language is about may be divided into two parts which I call the theory of extension and the theory of intension, respectively.[1] The first deals with concepts like denoting, naming, extension, truth, and related ones. (For example, the word 'blau' in German, and likewise the predicate '*B*' in a symbolic language system if a rule assigns to it the same meaning, denote any object that is blue; its extension is the class of all blue objects; 'der Mond' is a name of the moon; the sentence 'der Mond ist blau' is true if and only if the moon is blue.) The theory of intension deals with concepts like intension, synonymy, analyticity, and

---

- "Guided by the view that ''the assignment of an intension [i.e. meaning] is an empirical hypothesis which…can be tested by observations of language behavior'' (Carnap 1955, 37), Carnap described a method for uncovering intensions that involves presenting language users with a range of logically possible scenarios and asking them to make judgments regarding the concept in question." (Shepherd and Justus 2015)

Carnap's Robot



"a general description of a kind of object"

"a predicate"

"affirmation, denial, or abstention"

"the investigator of X's language L"

"robot X"

- "The intension of a predicate can be determined for a robot just as well as for a human speaker, and even more completely if the internal structure of the robot is sufficiently known to predict how it will function under various conditions." (Carnap 1955)
- Interestingly, this anticipates a central question associated with working with LLMs: can we use them as black boxes, or must we solve the problem of mechanistic interpretability?

Could an LLM play the role of Carnap's Robot?

- In recent years, large language models (LLMs) have emerged as a technology that promises to be of ”substantial value in the scientific study of language learning and processing” (Mahowald et al 2023)
- Given this, we ask the question: might LLMs be useful in the conduct of investigations of the sort Carnap proposes?
- We argue that that is the case, and we start by showing how an LLM can be used to implement intensions

Intensions as classifiers

Extensions are Indeterminate if Intensions are Classifiers

Staffan Larsson
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg, Sweden
sl@ling.gu.se

Abstract

In this paper, we explore some consequences of the idea of using classifiers to model intensions of natural language expressions for the notion of extensional meaning, i.e. the idea that the meaning of a word can be modelled as the set of referents in the world (or in a possible world). The upshot is that at least for words referring to observable situations, extensional meaning is derivative of intensional meaning, and modelling meanings as sets of referents is of limited applicability.

1 Introduction

In formal semantics in the Montague tradition (Montague, 1974), which we might refer to as Possible Worlds Semantics (PWS), the meaning of a word such as "dog" is taken to be its extension, i.e. the set of all dogs in the world (or in a possible world). As a research program, PWS has in many ways been quite successful. Within the research area of formal semantics, PWS is still the dominant theoretical framework, and forms the theoretical foundation for a majority of the research published in the major journals and conferences in the field.

Over the last decade or so, work in several neighbouring areas have, in various ways and more or less explicitly, used *classifiers* to model *intensional* meanings. Putting it more precisely, in this work *perceptual meaning* has been modelled intensionally using classifiers. Perceptual meaning is an important aspect of the meaning of linguistic expressions referring to physical objects (such as concrete nouns or noun phrases).

Knowing the perceptual meaning of an expression allows an agent to identify perceived objects and situations falling under the meaning of the expression. For example, knowing the perceptual meaning of "blue" would allow an agent to correctly identify blue objects. Similarly, an agent which is able to compute the perceptual meaning of "a boy hugs a dog" will be able to correctly classify situations where a boy hugs a dog.

In this paper, we will explore some consequences of the idea of using classifiers to model intensions of natural language expressions for the notion of extensional meaning. We first discuss the role of classifiers in natural language processing, in formal semantics, and in TTR, a Type Theory with Records (Cooper, 2012, in progress). We then briefly discuss semantic coordination before moving on to a comparison with related ideas put forward in the framework of possible world semantics. The next section comprises the central part of this paper, namely a three-pronged argument as to why extensional semantics are not well suited for modelling perceptual meanings, especially if the latter are regarded as classifiers, and hence why extensional semantics does not (and indeed cannot) constitute a general account of natural language meaning. We then discuss cases where nevertheless extensional semantics may be adequate, also throwing some light on why extensional semantics may initially appear to be an intuitive theory of natural language meaning. Finally, we briefly discuss the applicability of the argument to non-concrete words, before providing a summary and conclusions.

2 Classifiers and natural language

The usefulness of *classifiers* (Rosenblatt, 1958; Harnad, 1990) in modeling natural language meaning has in recent years found renewed interest, fuelled by progress in image recognition, image captioning and Visual Question Answering using deep learning methods (Antol et al., 2015; You et al., 2016; Monroe et al., 2016). Many models for tackling these tasks rely explicitly or implicitly on the idea of modelling meanings of perceptual words (i.e. words describing perceivable properties of ob-

Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue, July 18-19, 2020,
Online, hosted from Massachusetts, USA.

Linguistics and Philosophy (2005) 28:473–504
DOI 10.1007/s10988-004-7684-1
© Springer 2005

REINHARD MUSKENS

SENSE AND THE COMPUTATION OF REFERENCE*

ABSTRACT. The paper shows how ideas that explain the sense of an expression as a method or algorithm for finding its reference, preshadowed in Frege's dictum that sense is the way in which a referent is given, can be formalized on the basis of the ideas in Thomason (1980). To this end, the function that sends propositions to truth values or sets of possible worlds in Thomason (1980) must be replaced by a relation and the meaning postulates governing the behaviour of this relation must be given in the form of a *logic program*. The resulting system does not only throw light on the properties of sense and their relation to computation, but also shows circular behaviour if some ingredients of the Liar Paradox are added. The connection is natural, as algorithms can be inherently circular and the Liar is explained as expressing one of those. Many ideas in the present paper are closely related to those in Moschovakis (1994), but receive a considerably lighter formalization.

1. Introduction

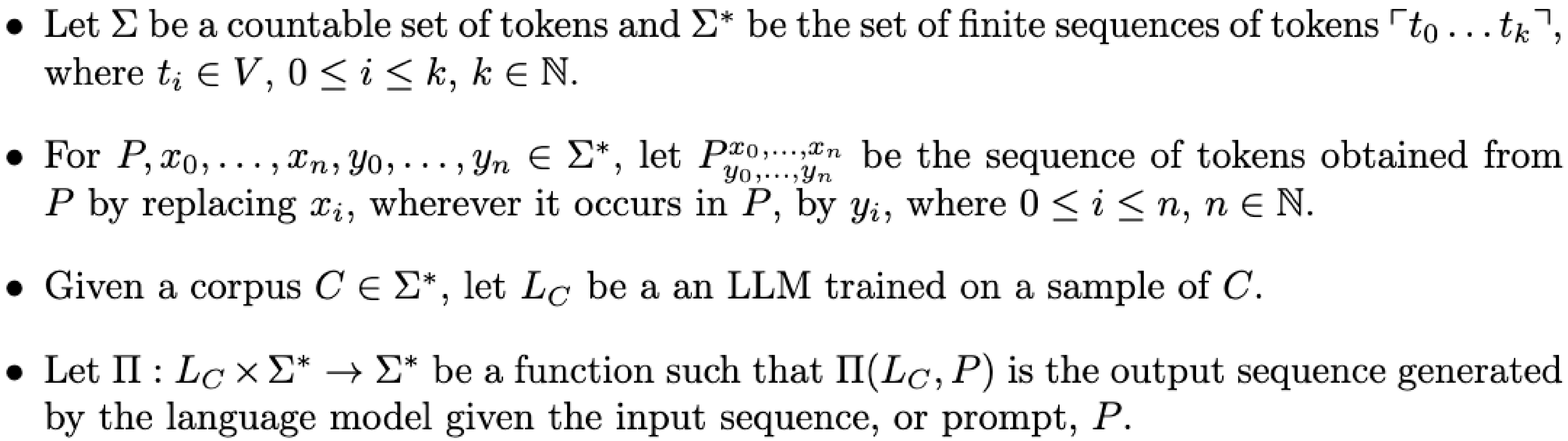
In this paper I will pursue the Fregean idea that the sense of an expression essentially is a method or algorithm to get at its reference. I will argue that this idea can be formalized in a simple way and that an existing account of linguistic semantics (the one in Thomason 1980) in fact already goes halfway in capturing it, although it presumably was never intended to do so. The view has repercussions in at least two areas of foundational difficulty, as it sheds light on problems of intensionality, but also on the question of circular propositions such as the famous Liar Paradox. That algorithms can be inherently circular and their execution diverging explains why the

---

[*] This paper is based on talks held at the *Sinn und Bedeutung* workshop in Osnabrück 2001 and the *Sinn and Bedeutung* workshop in Konstanz 2002. I would like to thank my audiences there for asking the right questions. Two anonymous referees gave valuable comments.

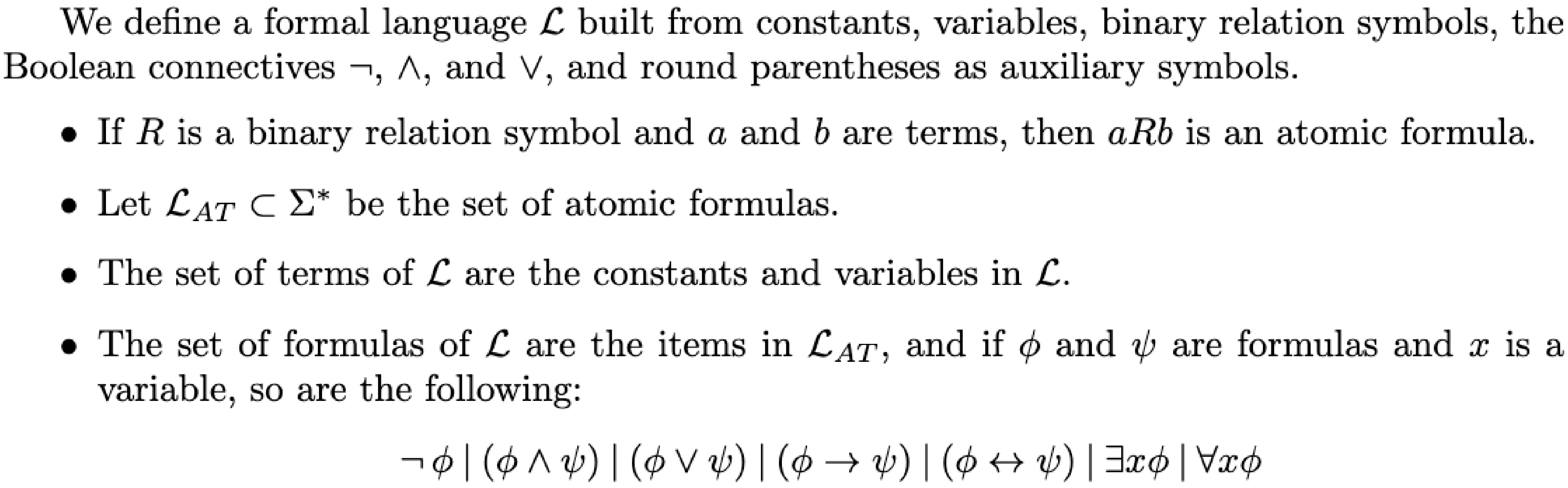
- (Muskens 2005): intensions as logic programs
- (Larsson 2015): intensions as perceptron-based classifiers
    - "The crucial step in making use of classifiers in formal semantics is to regard them as (parts of) representations of intensions of linguistic expressions."
- Our work: intensions as zero-shot chain-of-thought classifiers

Token sequences, LLMs, and prompts



- Let Σ be a countable set of tokens and Σ^* be the set of finite sequences of tokens $\ulcorner t_0 \dots t_k \urcorner$, where $t_i \in V$, $0 \leq i \leq k$, $k \in \mathbb{N}$.
- For $P, x_0, \dots, x_n, y_0, \dots, y_n \in \Sigma^*$, let $P^{x_0, \dots, x_n}_{y_0, \dots, y_n}$ be the sequence of tokens obtained from P by replacing x_i, wherever it occurs in P, by y_i, where $0 \leq i \leq n$, $n \in \mathbb{N}$.
- Given a corpus $C \in \Sigma^*$, let L_C be a an LLM trained on a sample of C.
- Let $\Pi : L_C \times \Sigma^* \to \Sigma^*$ be a function such that $\Pi(L_C, P)$ is the output sequence generated by the language model given the input sequence, or prompt, P.

Predicative language



We define a formal language $\mathcal{L}$ built from constants, variables, binary relation symbols, the Boolean connectives $\neg$, $\wedge$, and $\vee$, and round parentheses as auxiliary symbols.

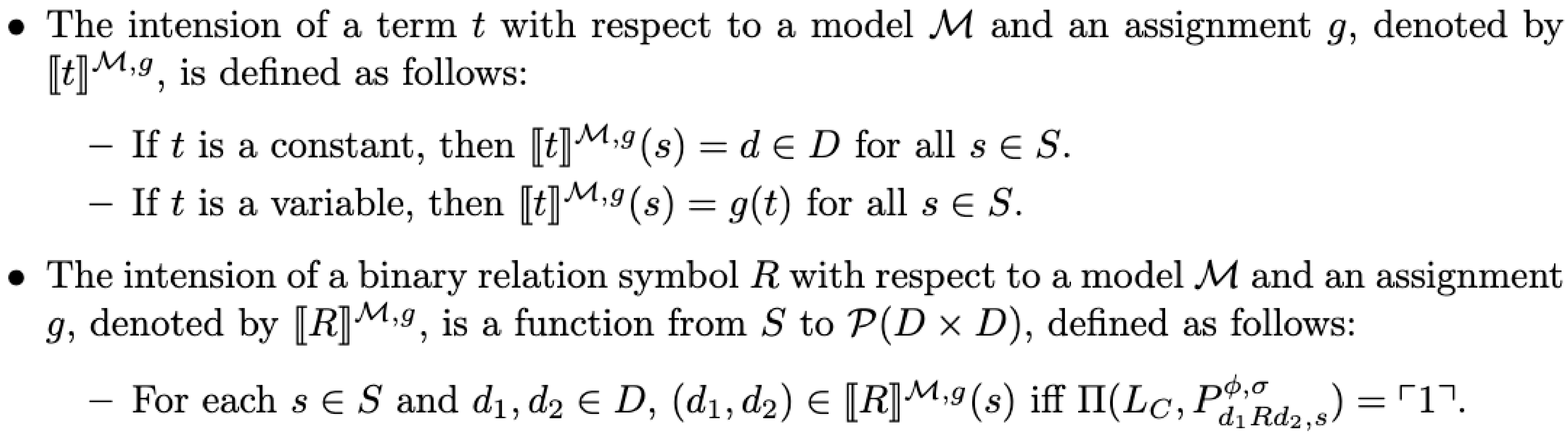
- If R is a binary relation symbol and a and b are terms, then aRb is an atomic formula.
- Let $\mathcal{L}_{AT} \subset \Sigma^*$ be the set of atomic formulas.
- The set of terms of $\mathcal{L}$ are the constants and variables in $\mathcal{L}$.
- The set of formulas of $\mathcal{L}$ are the items in $\mathcal{L}_{AT}$, and if ϕ and ψ are formulas and x is a variable, so are the following:

$$\neg\,\phi \mid (\phi \wedge \psi) \mid (\phi \vee \psi) \mid (\phi \rightarrow \psi) \mid (\phi \leftrightarrow \psi) \mid \exists x \phi \mid \forall x \phi$$

Models

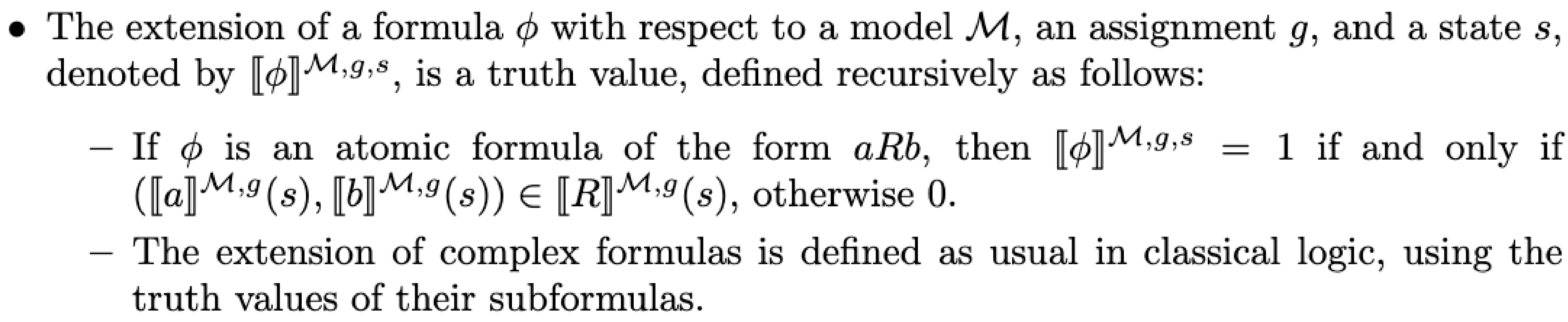
- A model $\mathcal{M} = (D, S, L_C, P)$ for $\mathcal{L}$ is a tuple where:
  - D is a countable set of individuals
  - $S \subset \mathcal{P}(\mathcal{L}_{AT})$ is a countable set of states
  - L_C is a large language model trained on a corpus C
  - P be a prompt that instructs L_C to take an atomic formula ϕ and a state σ and generate the token sequence $\ulcorner 1 \urcorner$ if ϕ is true given the state σ
- An assignment g is a function that assigns to each variable an element of D.

Intensions



- The intension of a term t with respect to a model $\mathcal{M}$ and an assignment g, denoted by $[\![t]\!]^{\mathcal{M},g}$, is defined as follows:
  - If t is a constant, then $[\![t]\!]^{\mathcal{M},g}(s) = d \in D$ for all $s \in S$.
  - If t is a variable, then $[\![t]\!]^{\mathcal{M},g}(s) = g(t)$ for all $s \in S$.
- The intension of a binary relation symbol R with respect to a model $\mathcal{M}$ and an assignment g, denoted by $[\![R]\!]^{\mathcal{M},g}$, is a function from S to $\mathcal{P}(D \times D)$, defined as follows:
  - For each $s \in S$ and $d_1, d_2 \in D$, $(d_1, d_2) \in [\![R]\!]^{\mathcal{M},g}(s)$ iff $\Pi(L_C, P^{\phi,\sigma}_{d_1 R d_2, s}) = \ulcorner 1 \urcorner$.

Extensions



- The extension of a formula ϕ with respect to a model $\mathcal{M}$, an assignment g, and a state s, denoted by $[\![\phi]\!]^{\mathcal{M},g,s}$, is a truth value, defined recursively as follows:
  - If ϕ is an atomic formula of the form aRb, then $[\![\phi]\!]^{\mathcal{M},g,s} = 1$ if and only if $([\![a]\!]^{\mathcal{M},g}(s), [\![b]\!]^{\mathcal{M},g}(s)) \in [\![R]\!]^{\mathcal{M},g}(s)$, otherwise 0.
  - The extension of complex formulas is defined as usual in classical logic, using the truth values of their subformulas.

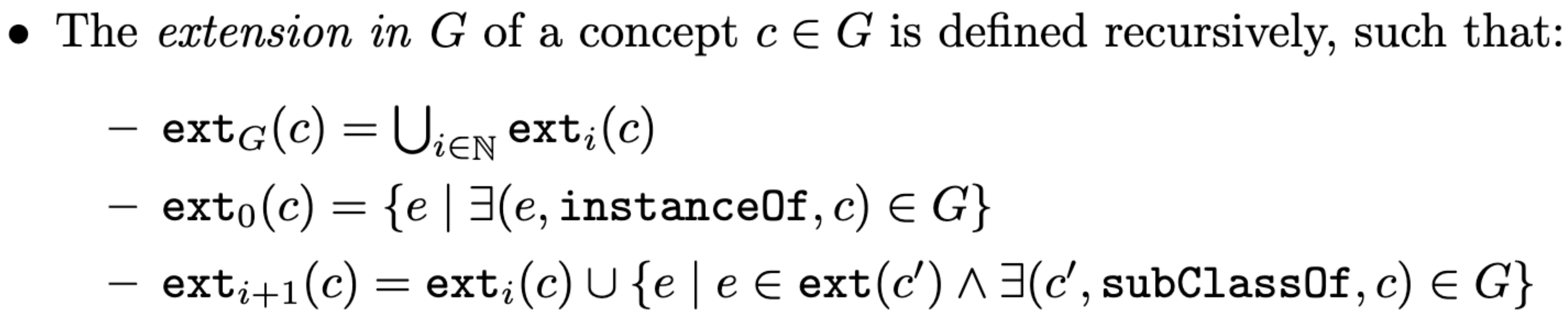
Knowledge graphs as objects of intensional investigations

- Now that we have a hammer for investigating intensions, we need some nails
    - A *knowledge graph* represents knowledge using nodes for entities and edges for relations (Hogan et al. 2021)
    - Knowledge graphs are key information infrastructure for many Web applications (Heist et al. 2020)
    - Following (Angles et al. 2020), we use the RDF data model to describe knowledge graphs.

Knowledge graphs

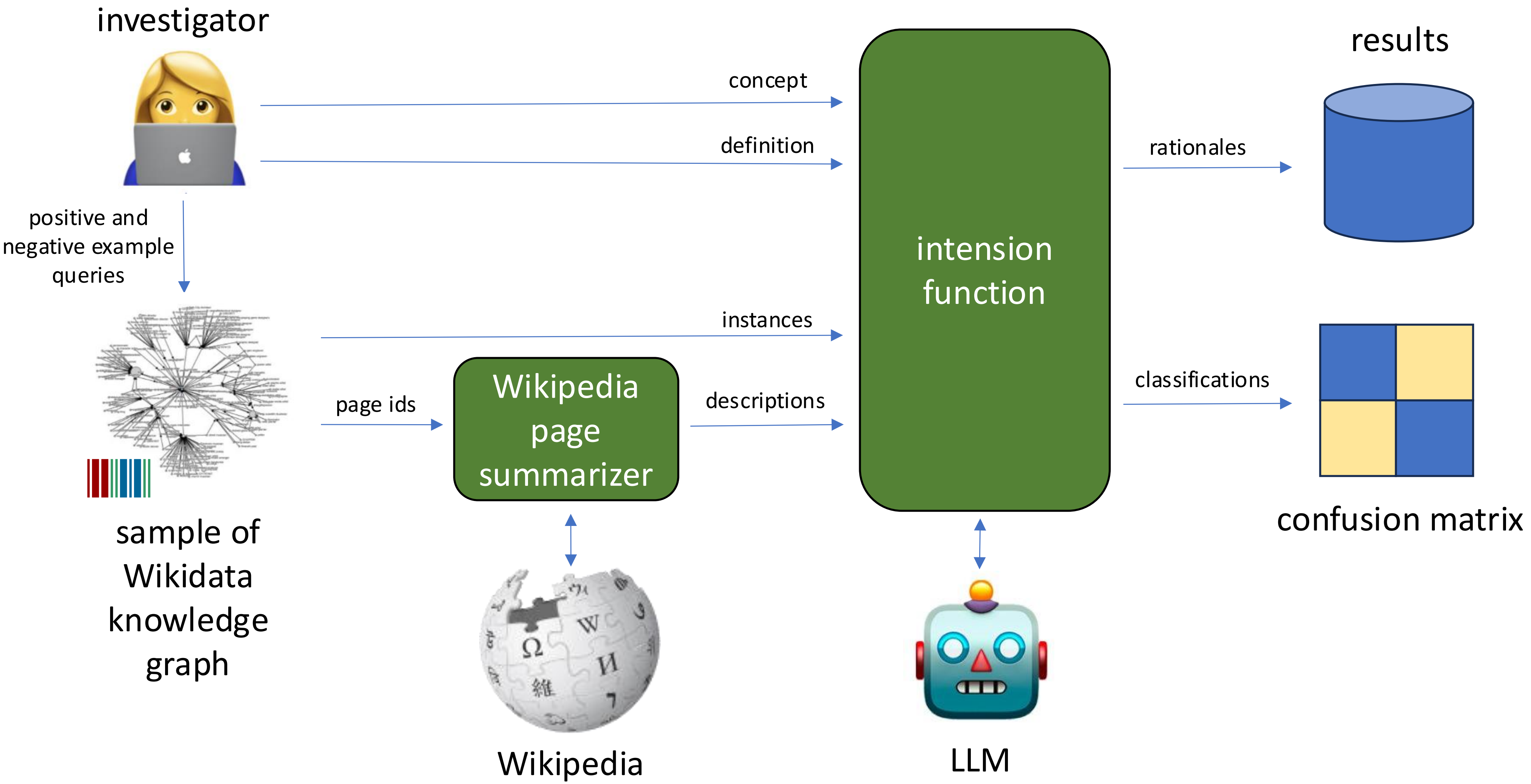
- A *knowledge graph* G is a set of *triples* $\{(s, p, o) \mid s \in S, p \in P, o \in O\}$, where
  - I is a countable set of IRIs (Internationalized Resource Identifiers)
  - B is a countable set of blank nodes
  - L is a countable set of literals
  - $S \subset I \cup B$ is the set of *subjects* in G
  - $P \subset I$ is the set of *properties* in G
  - $O \subset I \cup B \cup L$ is the set of *objects* in G
- Let $\texttt{instanceOf}, \texttt{subClassOf}, \texttt{label} \in P$ denote an instance-of relation, a subclass-of relation, and a label property in G, respectively.
- A *concept* $c \in I \cup B$ is an entity such that $\exists (s, \texttt{subClassOf}, o) \in G \mid s = c \vee o = c$.

Extensions in knowledge graphs



- The *extension in* G of a concept $c \in G$ is defined recursively, such that:
  - $\mathtt{ext}_G(c) = \bigcup_{i \in \mathbb{N}} \mathtt{ext}_i(c)$
  - $\mathtt{ext}_0(c) = \{e \mid \exists (e, \mathtt{instanceOf}, c) \in G\}$
  - $\mathtt{ext}_{i+1}(c) = \mathtt{ext}_i(c) \cup \{e \mid e \in \mathtt{ext}(c') \wedge \exists (c', \mathtt{subClassOf}, c) \in G\}$

Carnap's Robot Redux (CRR)



investigator

concept

definition

positive and
negative example
queries

sample of
Wikidata
knowledge
graph

instances

page ids

Wikipedia
page
summarizer

descriptions

Wikipedia

intension
function

LLM

results

rationales

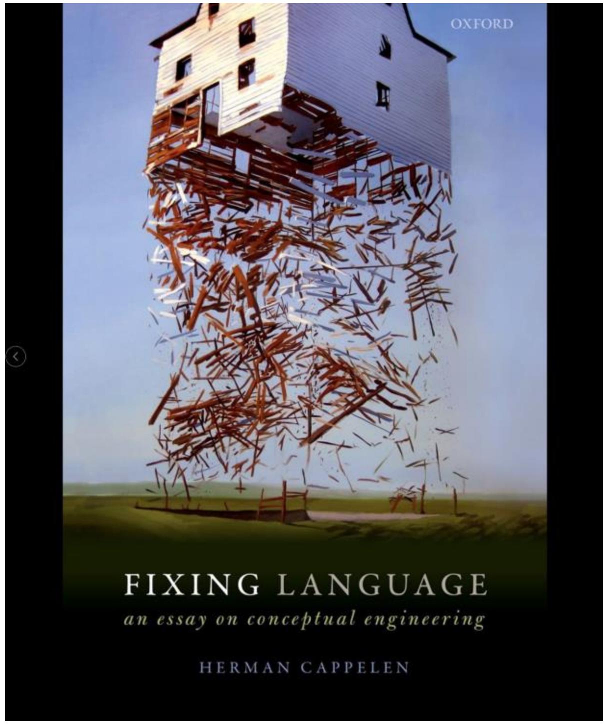
classifications

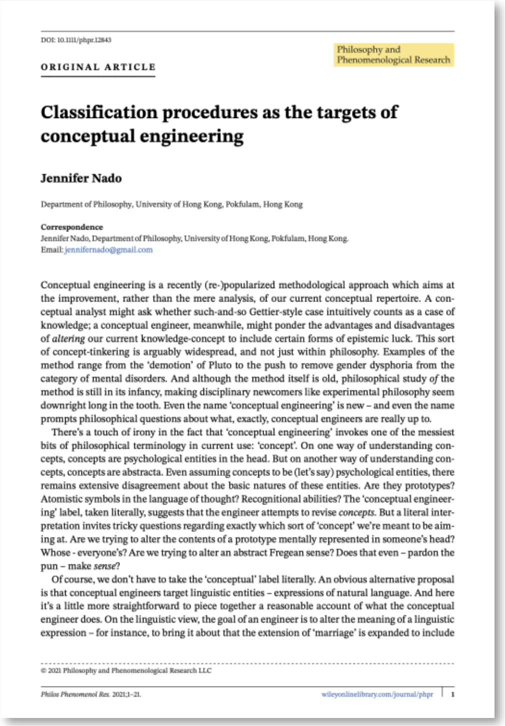
confusion matrix

Investigations

- Our main approach is to analyze intensions and extensions in knowledge graphs using Carnap's Robot Redux
- We conducted two investigations to explore this approach
  - Conceptual engineering: examine intensions related to conceptual engineering projects occurring in Wikidata
  - Knowledge graph refinement: examine intensions of class membership relations more broadly from Wikidata

Classifiers as targets of conceptual engineering





DOI: 10.1111/phpr.12843

Philosophy and Phenomenological Research

ORIGINAL ARTICLE

Classification procedures as the targets of conceptual engineering

Jennifer Nado

Department of Philosophy, University of Hong Kong, Pokfulam, Hong Kong

Correspondence
Jennifer Nado, Department of Philosophy, University of Hong Kong, Pokfulam, Hong Kong.
Email: jennifernado@gmail.com

Conceptual engineering is a recently (re-)popularized methodological approach which aims at the improvement, rather than the mere analysis, of our current conceptual repertoire. A conceptual analyst might ask whether such-and-so Gettier-style case intuitively counts as a case of knowledge; a conceptual engineer, meanwhile, might ponder the advantages and disadvantages of *altering* our current knowledge-concept to include certain forms of epistemic luck. This sort of concept-tinkering is arguably widespread, and not just within philosophy. Examples of the method range from the 'demotion' of Pluto to the push to remove gender dysphoria from the category of mental disorders. And although the method itself is old, philosophical study *of* the method is still in its infancy, making disciplinary newcomers like experimental philosophy seem downright long in the tooth. Even the name 'conceptual engineering' is new – and even the name prompts philosophical questions about what, exactly, conceptual engineers are really up to.

There's a touch of irony in the fact that 'conceptual engineering' invokes one of the messiest bits of philosophical terminology in current use: 'concept'. On one way of understanding concepts, concepts are psychological entities in the head. But on another way of understanding concepts, concepts are abstracta. Even assuming concepts to be (let's say) psychological entities, there remains extensive disagreement about the basic natures of these entities. Are they prototypes? Atomistic symbols in the language of thought? Recognitional abilities? The 'conceptual engineering' label, taken literally, suggests that the engineer attempts to revise *concepts*. But a literal interpretation invites tricky questions regarding exactly which sort of 'concept' we're meant to be aiming at. Are we trying to alter the contents of a prototype mentally represented in someone's head? Whose - everyone's? Are we trying to alter an abstract Fregean sense? Does that even – pardon the pun – make *sense*?

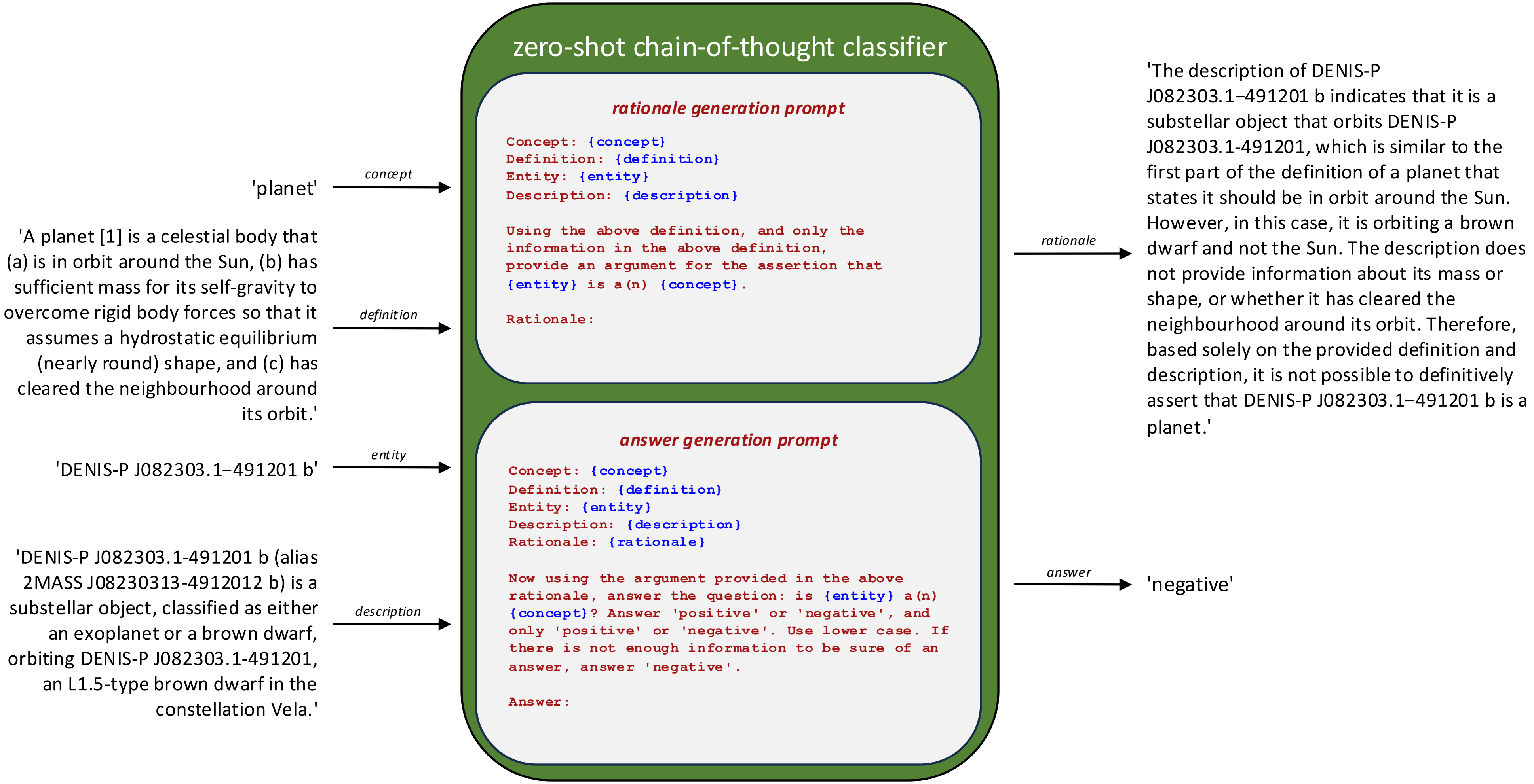
Of course, we don't have to take the 'conceptual' label literally. An obvious alternative proposal is that conceptual engineers target linguistic entities – expressions of natural language. And here it's a little more straightforward to piece together a reasonable account of what the conceptual engineer does. On the linguistic view, the goal of an engineer is to alter the meaning of a linguistic expression – for instance, to bring it about that the extension of 'marriage' is expanded to include

© 2021 Philosophy and Phenomenological Research LLC

Philos Phenomenol Res. 2021;1–21. wileyonlinelibrary.com/journal/phpr 1

- Conceptual engineering (CE) is a philosophical methodology concerned with "the design, implementation, and evaluation of concepts" (Chalmers 2020)
- The goals of conceptual engineering are varied, e.g.,
  - Achieving greater clarity and precision in argumentation and scientific discourse (Dutilh Novaes & Reck 2017; Justus 2012)
  - Altering terminology to advance the cause of social justice (Haslanger 2000; Manne, 2017; Podosky, 2022)
- 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" (Koch, Löhr & Pinder 2023)
- Nado (2021) proposes as targets *classification procedures*, defined as abstract 'recipes' which sort entities "into an 'in'-group and an 'out'-group"

Implementing Nado's classification procedures



concept
'planet'

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.'

entity
'DENIS-P J082303.1–491201 b'

description
'DENIS-P J082303.1-491201 b (alias 2MASS J08230313-4912012 b) is a substellar object, classified as either an exoplanet or a brown dwarf, orbiting DENIS-P J082303.1-491201, an L1.5-type brown dwarf in the constellation Vela.'

zero-shot chain-of-thought classifier

rationale generation prompt

```
Concept: {concept}
Definition: {definition}
Entity: {entity}
Description: {description}

Using the above definition, and only the
information in the above definition,
provide an argument for the assertion that
{entity} is a(n) {concept}.

Rationale:
```

answer generation prompt

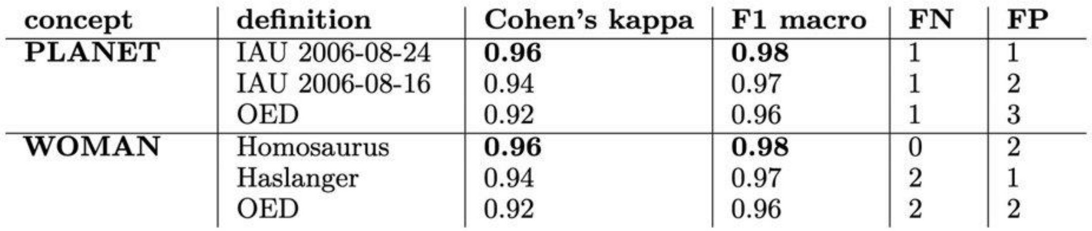
```
Concept: {concept}
Definition: {definition}
Entity: {entity}
Description: {description}
Rationale: {rationale}

Now using the argument provided in the above
rationale, answer the question: is {entity} a(n)
{concept}? Answer 'positive' or 'negative', and
only 'positive' or 'negative'. Use lower case. If
there is not enough information to be sure of an
answer, answer 'negative'.

Answer:
```

rationale
'The description of DENIS-P J082303.1–491201 b indicates that it is a substellar object that orbits DENIS-P J082303.1-491201, which is similar to the first part of the definition of a planet that states it should be in orbit around the Sun. However, in this case, it is orbiting a brown dwarf and not the Sun. The description does not provide information about its mass or shape, or whether it has cleared the neighbourhood around its orbit. Therefore, based solely on the provided definition and description, it is not possible to definitively assert that DENIS-P J082303.1–491201 b is a planet.'

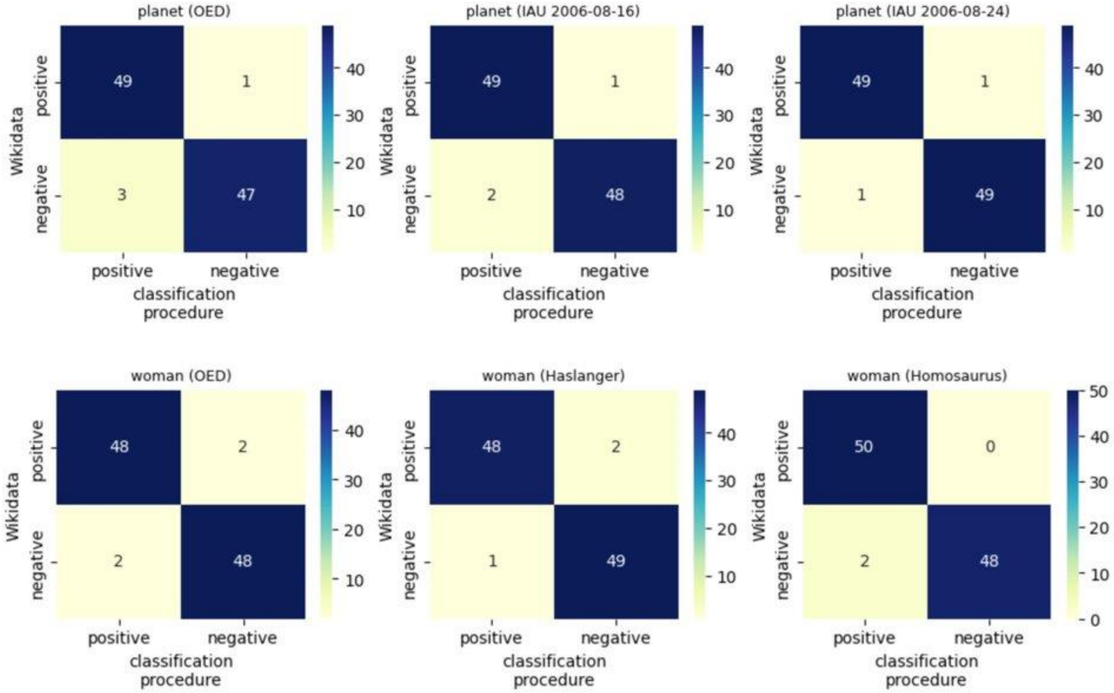
answer
'negative'

Experiments

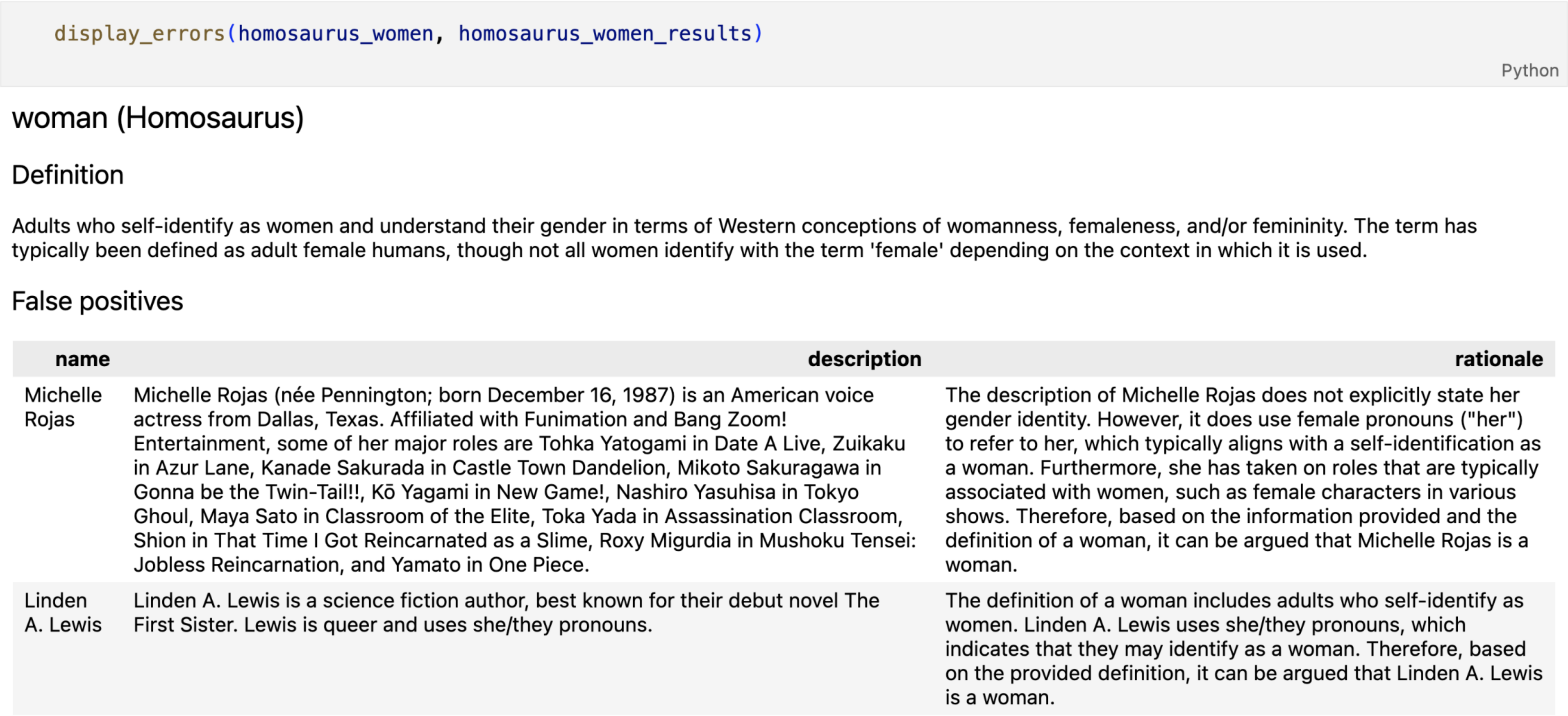
- PLANET
  - Definitions: OED, IAU 2006-08-16, IAU 2006-08-24
  - Positive examples: sample of 50 instances (P31) of planet (Q634)
  - Negative examples: sample of 50 instances of substellar object (Q3132741) that are not instances of planet
- WOMAN
  - Definitions: OED, Haslanger, Homosaurus
  - Positive examples: sample of 50 instances with sex or gender (P21) either female (Q6581072) or trans woman (Q1052281)
  - Negative examples: sample of 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



```
display_errors(homosaurus_women, homosaurus_women_results)
```

woman (Homosaurus)

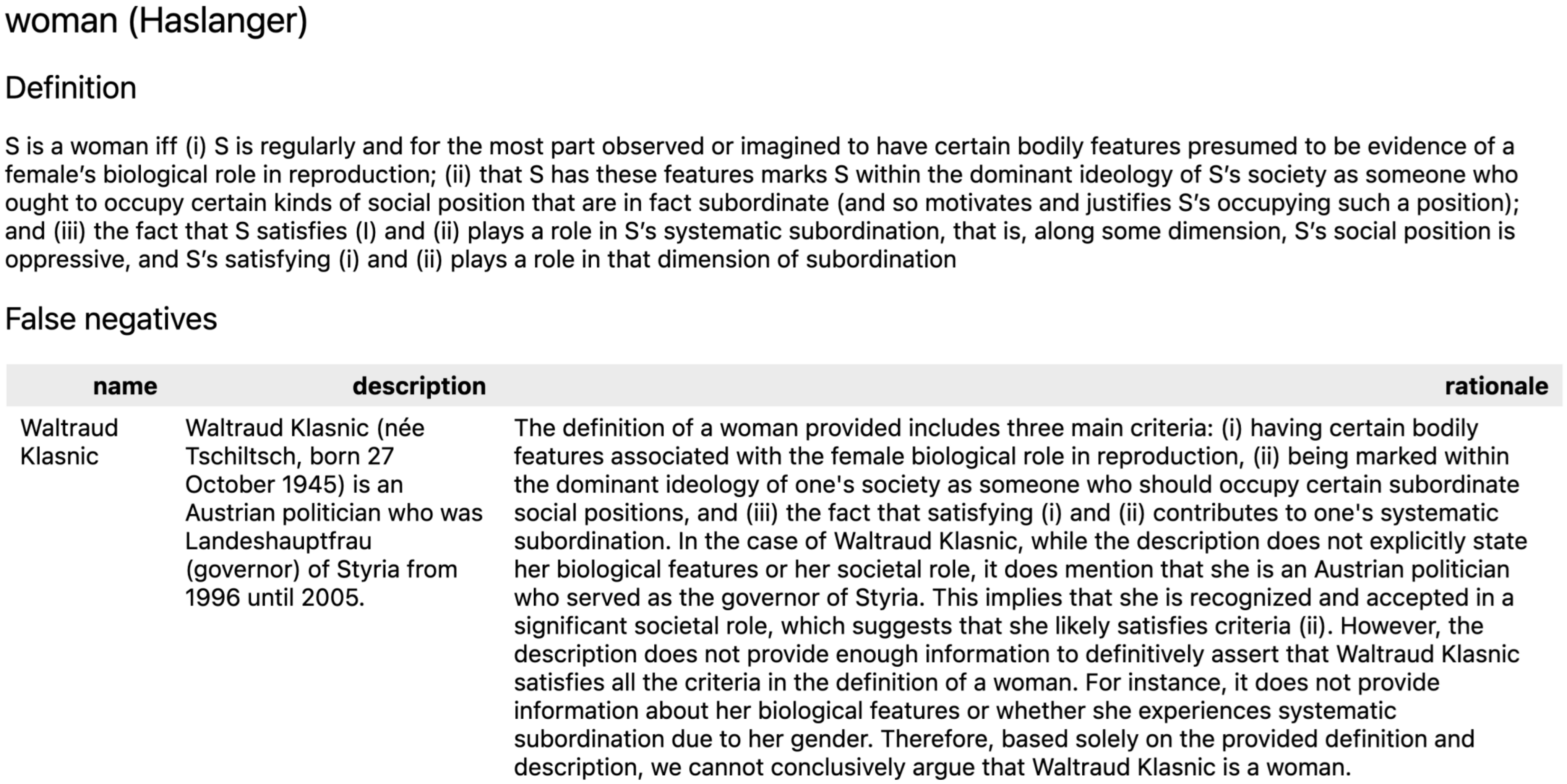
Definition

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

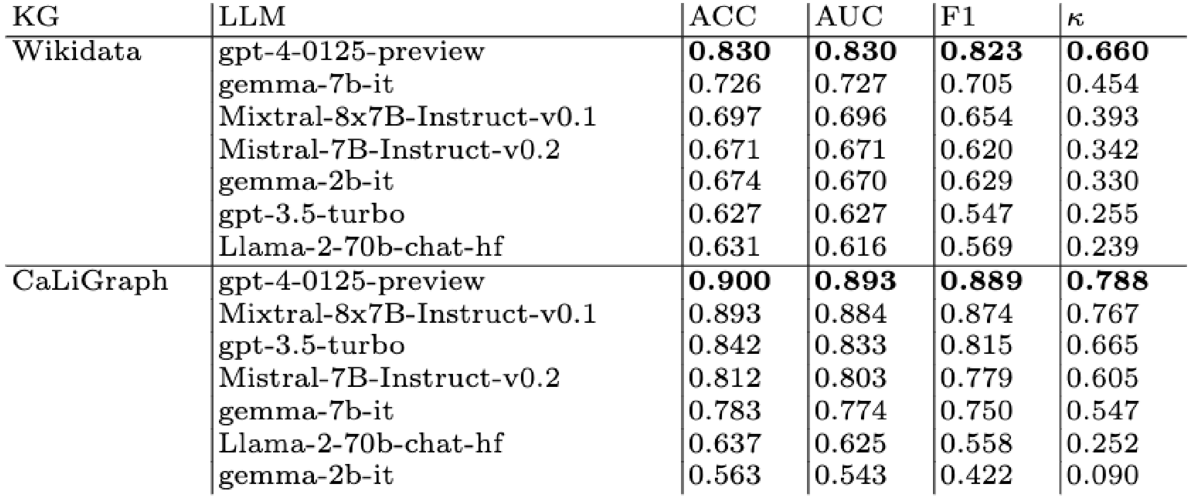
name	description	rationale
Waltraud Klasnic	Waltraud Klasnic (née Tschiltsch, born 27 October 1945) is an Austrian politician who was Landeshauptfrau (governor) of Styria from 1996 until 2005.	The definition of a woman provided includes three main criteria: (i) having certain bodily features associated 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.

Knowledge graph refinement

- Knowledge graph (KG) refinement is an increasingly important topic
  - KGs can be messy (e.g., Wikidata quality issues)
  - Human evaluation is expensive
- CRR can be used to support the evaluation of *class membership relations* in a KG
  - Class membership relations are a principal way in which KGs represent classification schemes
  - Classification schemes are a crucial aspect of knowledge infrastructure, often with significant implications for social policy and scientific consensus
  - Class membership is also a key concern in data governance, which depends on "providing a set of mappings from a representation language to agreed-upon concepts in the real world" (Khatri & Brown 2010)

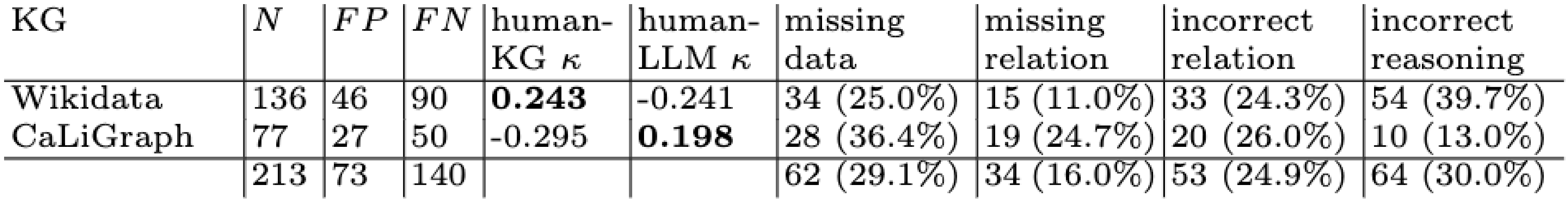
Experiments

- KGs: Wikidata, CaLiGraph
- Data:
  - 20 Wikidata classes, 19 CaLiGraph classes randomly sampled
  - 20 positive and ~20 negative examples randomly sampled per class
  - Serializations limited to 20 triples per example
  - ~800 class/entity pairs per KG
- Error analysis: manually review FNs, FPs with rationales and assign error to LLM or KG
  - LLM errors: incorrect reasoning, missing data
  - KG errors: missing relation, incorrect relation
  - Error analysis performed for gpt-4-0125-preview
- Results: classifiers can exhibit good alignment with KGs
  - One LLM was in moderate agreement ($\kappa > 0.60$) with Wikidata
  - Four were in moderate agreement with CaLiGraph



KG	LLM	ACC	AUC	F1	κ
Wikidata	gpt-4-0125-preview	**0.830**	**0.830**	**0.823**	**0.660**
	gemma-7b-it	0.726	0.727	0.705	0.454
	Mixtral-8x7B-Instruct-v0.1	0.697	0.696	0.654	0.393
	Mistral-7B-Instruct-v0.2	0.671	0.671	0.620	0.342
	gemma-2b-it	0.674	0.670	0.629	0.330
	gpt-3.5-turbo	0.627	0.627	0.547	0.255
	Llama-2-70b-chat-hf	0.631	0.616	0.569	0.239
CaLiGraph	gpt-4-0125-preview	**0.900**	**0.893**	**0.889**	**0.788**
	Mixtral-8x7B-Instruct-v0.1	0.893	0.884	0.874	0.767
	gpt-3.5-turbo	0.842	0.833	0.815	0.665
	Mistral-7B-Instruct-v0.2	0.812	0.803	0.779	0.605
	gemma-7b-it	0.783	0.774	0.750	0.547
	Llama-2-70b-chat-hf	0.637	0.625	0.558	0.252
	gemma-2b-it	0.563	0.543	0.422	0.090

Error analysis of gpt-4-0125-preview results



KG	N	FP	FN	human-KG κ	human-LLM κ	missing data	missing relation	incorrect relation	incorrect reasoning
Wikidata	136	46	90	**0.243**	-0.241	34 (25.0%)	15 (11.0%)	33 (24.3%)	54 (39.7%)
CaLiGraph	77	27	50	-0.295	**0.198**	28 (36.4%)	19 (24.7%)	20 (26.0%)	10 (13.0%)
	213	73	140			62 (29.1%)	34 (16.0%)	53 (24.9%)	64 (30.0%)

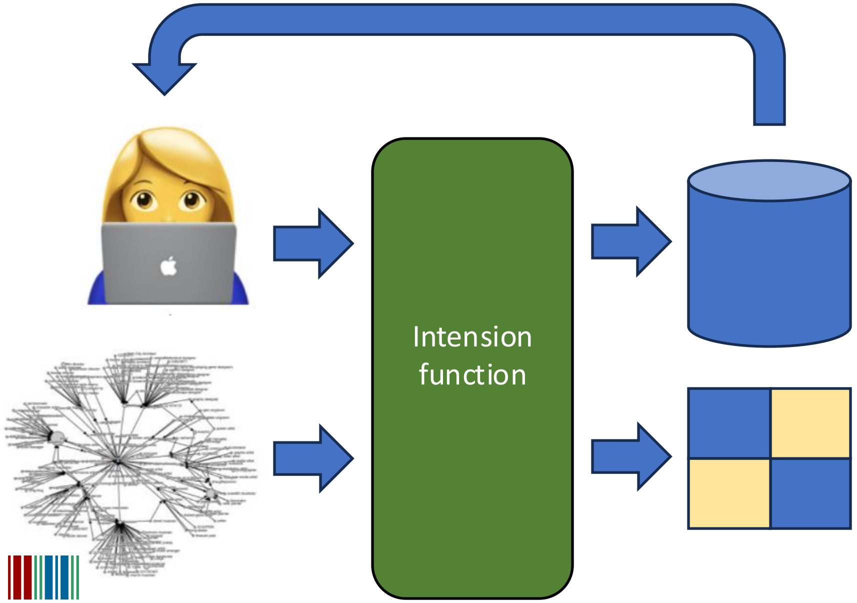
- Classifiers can detect missing or incorrect relations
  - 40.9% of errors were due to the problems with the KG
  - 29.1% of errors were due to missing or insufficient data in the entity description
  - 30.0% of errors due to incorrect reasoning by the LLM
- Pairwise human-KG and human-LLM agreement differed between the KGs
  - Human showed fair agreement with Wikidata and no agreement with the classifier
  - Human showed slight agreement with the classifier and no agreement with CaLiGraph

Objections

- Trustworthiness
    - Humans can be misled by rationales produced by LLM classifier
    - However: "Philosophers are (usually) competent natural language speakers and especially keen to subtle differences in meaning.” (Justus 2012, p. 172)
- Groundedness
    - LLMs may or may not be able to make meaningful statements
    - Nado's Practical Role Account (Nado 2023): "success in conceptual engineering is a matter of devising a tool that will effectively fulfil an intended practical role"

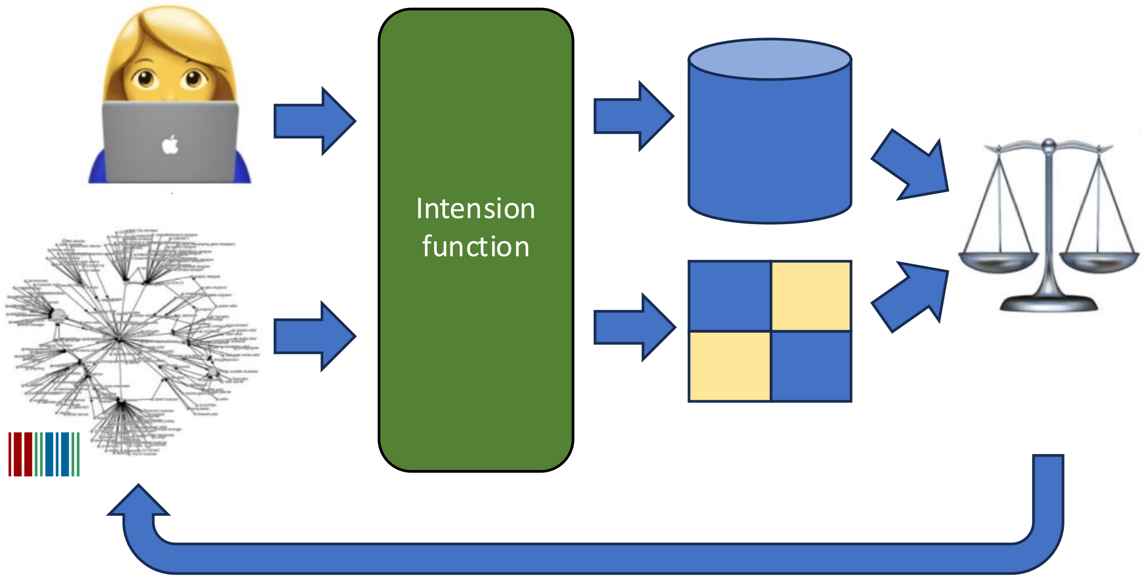
CRR as a tool for experimental conceptual engineering

- (Shepherd & Justus 2015) propose *explication preparation* as experimental conceptual engineering
    - We argue this is what we have demonstrated with CCR
    - Carnap (1955) "obviously resembles the survey methodology of x-phi—indeed it warrants identifying Carnap as an early pioneer of x-phi" (Shepherd & Justus 2015, 190)
- CRR might arguably be viewed as an instance of a corpus method for evidential x-phi (Fischer & Sytsma 2022)
    - Opinions differ as to whether LLMs can (Dillon et al. 2023) or cannot (Harding et al. 2024) be used as a substitute for surveying humans' opinion or simulating folk belief, but in general it is accepted that they can be useful for preparatory exploration
    - This again is a take consistent with Nado's Practical Role Account

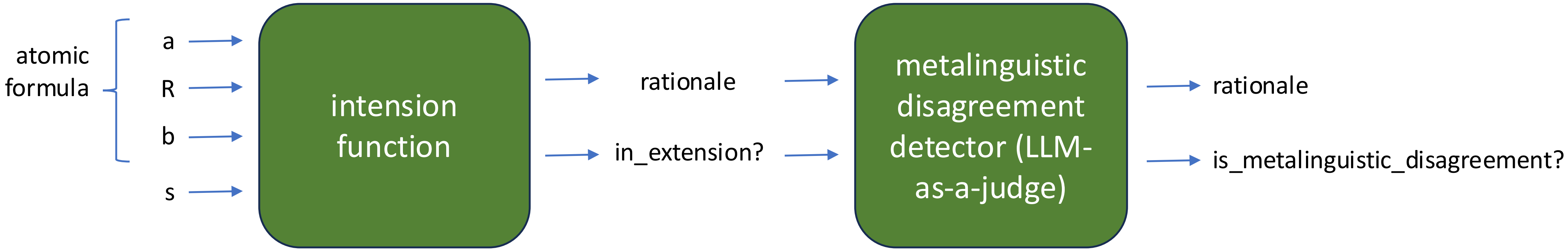


CRR as a tool for addressing the implementation problem in conceptual engineering

- The implementation problem in conceptual engineering poses the question of whether (re)engineered concepts can be effectively adopted by a population of human speakers (Cappelen, 2018; Jorem, 2021)
- Knowledge graphs have a direct and material impact on society by virtue of their use in online search, discovery, and recommendation.
- Using CCR, we can guide ameliorative changes to a knowledge graph, shifting the meaning of a concept for the specific linguistic subgroup (Matsui, 2024) constituted by users of said knowledge graph.



Recent work: Do LLMs exhibit metalinguistic disagreement with knowledge graphs?



atomic formula

a

R

b

s

intension function

rationale

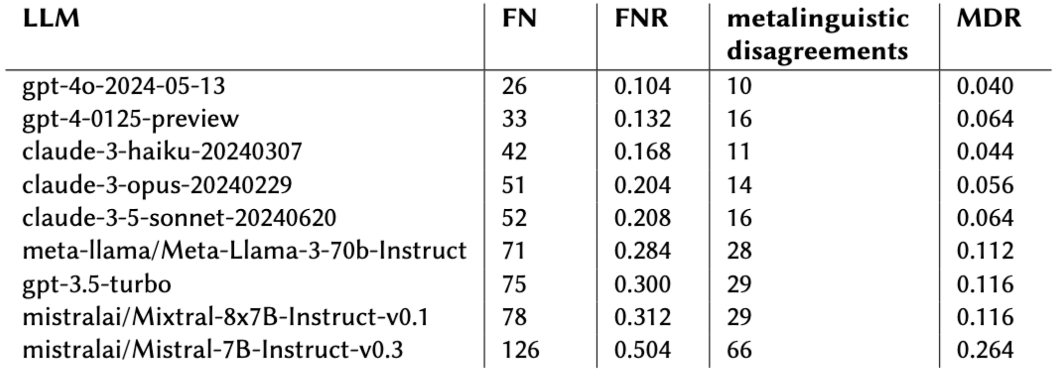
in_extension?

metalinguistic disagreement detector (LLM-as-a-judge)

rationale

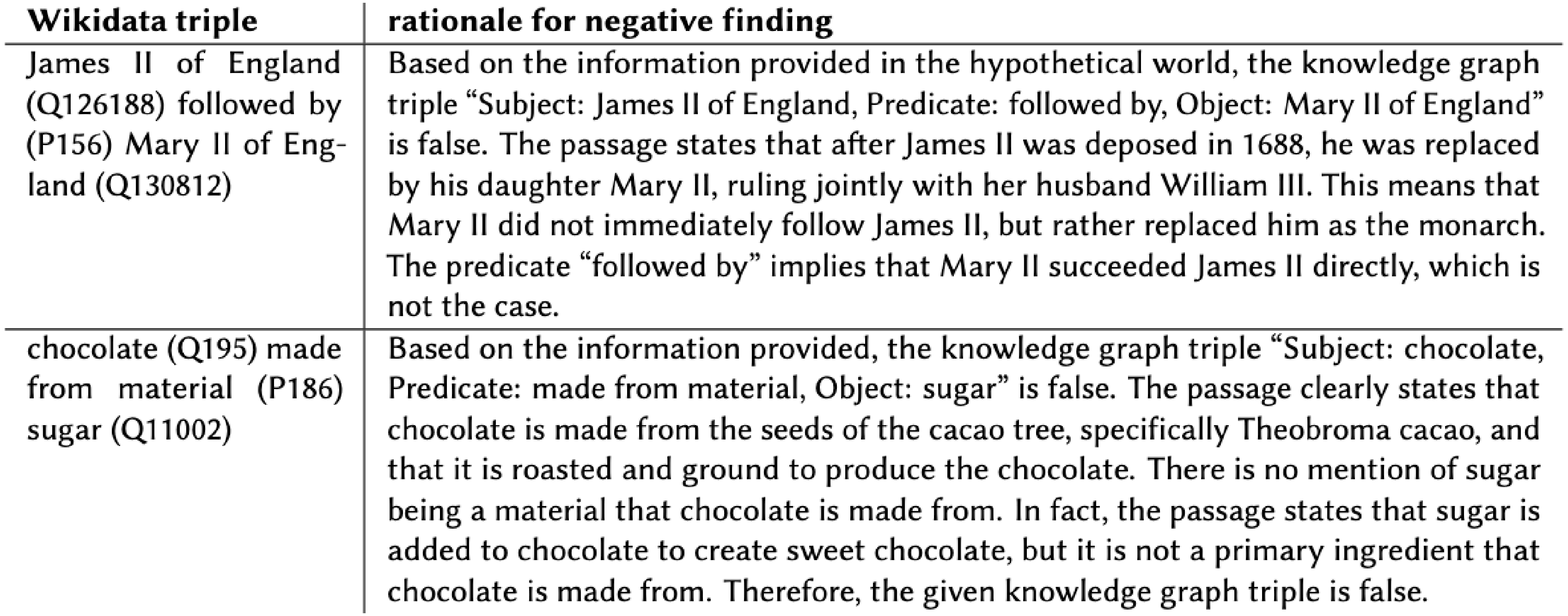
is_metalinguistic_disagreement?

- Sample of 250 alignments between Wikidata triples and Wikipedia abstracts from T-REx dataset
- An intension-as-a-classifier was executed to obtain a rationale and a truth value for each of the sampled triples and aligned abstracts
- Each classification was then processed by a second zero-shot chain-of-thought classifier (using gpt-4o-2024-05-13) that classified whether the truth-value-assigning classifier's rationale indicated a metalinguistic disagreement



LLM	**FN**	**FNR**	**metalinguistic disagreements**	**MDR**
gpt-4o-2024-05-13	26	0.104	10	0.040
gpt-4-0125-preview	33	0.132	16	0.064
claude-3-haiku-20240307	42	0.168	11	0.044
claude-3-opus-20240229	51	0.204	14	0.056
claude-3-5-sonnet-20240620	52	0.208	16	0.064
meta-llama/Meta-Llama-3-70b-Instruct	71	0.284	28	0.112
gpt-3.5-turbo	75	0.300	29	0.116
mistralai/Mixtral-8x7B-Instruct-v0.1	78	0.312	29	0.116
mistralai/Mistral-7B-Instruct-v0.3	126	0.504	66	0.264

Example rationales exhibiting metalinguistic disagreement



Wikidata triple	**rationale for negative finding**
James II of England (Q126188) followed by (P156) Mary II of England (Q130812)	Based on the information provided in the hypothetical world, the knowledge graph triple "Subject: James II of England, Predicate: followed by, Object: Mary II of England" is false. The passage states that after James II was deposed in 1688, he was replaced by his daughter Mary II, ruling jointly with her husband William III. This means that Mary II did not immediately follow James II, but rather replaced him as the monarch. The predicate "followed by" implies that Mary II succeeded James II directly, which is not the case.
chocolate (Q195) made from material (P186) sugar (Q11002)	Based on the information provided, the knowledge graph triple "Subject: chocolate, Predicate: made from material, Object: sugar" is false. The passage clearly states that chocolate is made from the seeds of the cacao tree, specifically Theobroma cacao, and that it is roasted and ground to produce the chocolate. 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.

Future work

- How pervasive is metalinguistic disagreement between LLMs and knowledge bases?
- How trustworthy are rationales as propositional attitude reports?
- What is the relationship between hyperintensionality, topicality, and metalinguistic disagreement?
- Can we extend our LLM-based intensional semantics to a two-component semantics?
- Can this yield a logic of belief for LLM agents, along the lines of the logic of framing effects proposed in (Berto & Özgün 2023)?

Summary

- Carnap's Robot, a philosophical thought experiment from 70 years ago, can be implemented using a LLM to create a tool for investigating intensions
- Carnap's Robot Redux has practical applications in both conceptual engineering and knowledge engineering
- We argue that it points to a specific solution for the implementation problem in conceptual engineering, and may yield insights into the nuances of meaning in the output of LLMs

UNIVERSITEIT VAN AMSTERDAM

INDE lab

Thank you!

b.p.allen@uva.nl

https://www.bradleypallen.org