

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

Bradley P. Allen

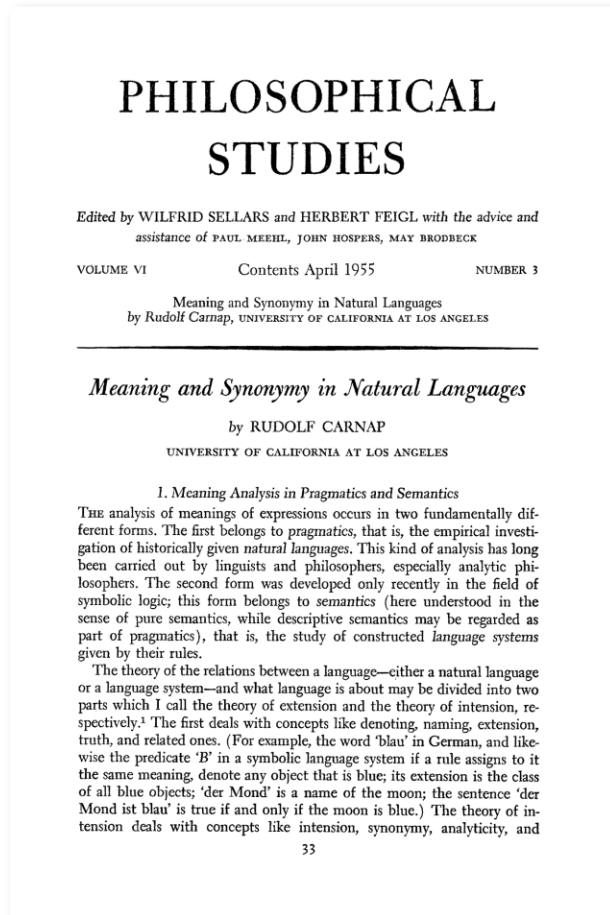
LLMs and Philosophy, Kanazawa

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
 - Intentions-as-classifiers as targets of conceptual engineering
 - Intentions-as-classifiers for the evaluation of knowledge graphs
- LLMs as cognitive tools for philosophy
 - Experimental conceptual engineering through simulation of linguistic groups
 - Ameliorative analysis of knowledge bases as a solution for the implementation problem in conceptual engineering
- Implications for work on LLMs and meaning
 - LLMs, metalinguistic disagreement, hyperintensionality, and two-component semantics

Carnap (1955)



- "I have tried to show in this paper that in a pragmatical investigation of a natural language there is not only, as generally agreed, an empirical method for ascertaining which objects are denoted by a given predicate and thus for determining the extension of the predicate, but also a method for testing a hypothesis concerning its intension (designative meaning)."
- "The intension of a predicate for a speaker X is, roughly speaking, the general condition which an object must fulfill for X to be willing to apply the predicate to it."
- "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's Robot



"the investigator of X's language L"

"a general description of a kind of object"

"a predicate"

"affirmation, denial, or abstention"



"robot X"

Carnap (1955) on mechanistic interpretability

- "There are in principle, although not always in practice, two ways for ascertaining whether a given thing or person X has the disposition D (at a given time t).
- "The first method may be called *behavioristic* (in a very wide sense); it consists in producing the condition C and then determining whether or not the response R occurs."
- "The second way may be called the method of *structure analysis*. It consists in investigating the state of X (at t) in sufficient detail such that it is possible to derive from the obtained description of the state with the help of relevant general laws (say of physics, physiology, etc.) the responses which X would make to any specified circumstances in the environment. Then it will be possible to predict, in particular, whether under the condition C X would make the response R or not; if so, X has the disposition D, otherwise not."

Intensions as classifiers

- 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



Formal semantics: token sequences, LLMs, and prompts

- Let Σ be a countable set of tokens and Σ^* be the set of finite sequences of tokens $\lceil t_0 \dots t_k \rceil$, 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 $P_{y_0, \dots, y_n}^{x_0, \dots, x_n}$ 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 $\Pi : L_C \times \Sigma^* \rightarrow \Sigma^*$ 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 .

Formal semantics: 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:

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

Formal semantics: 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 .

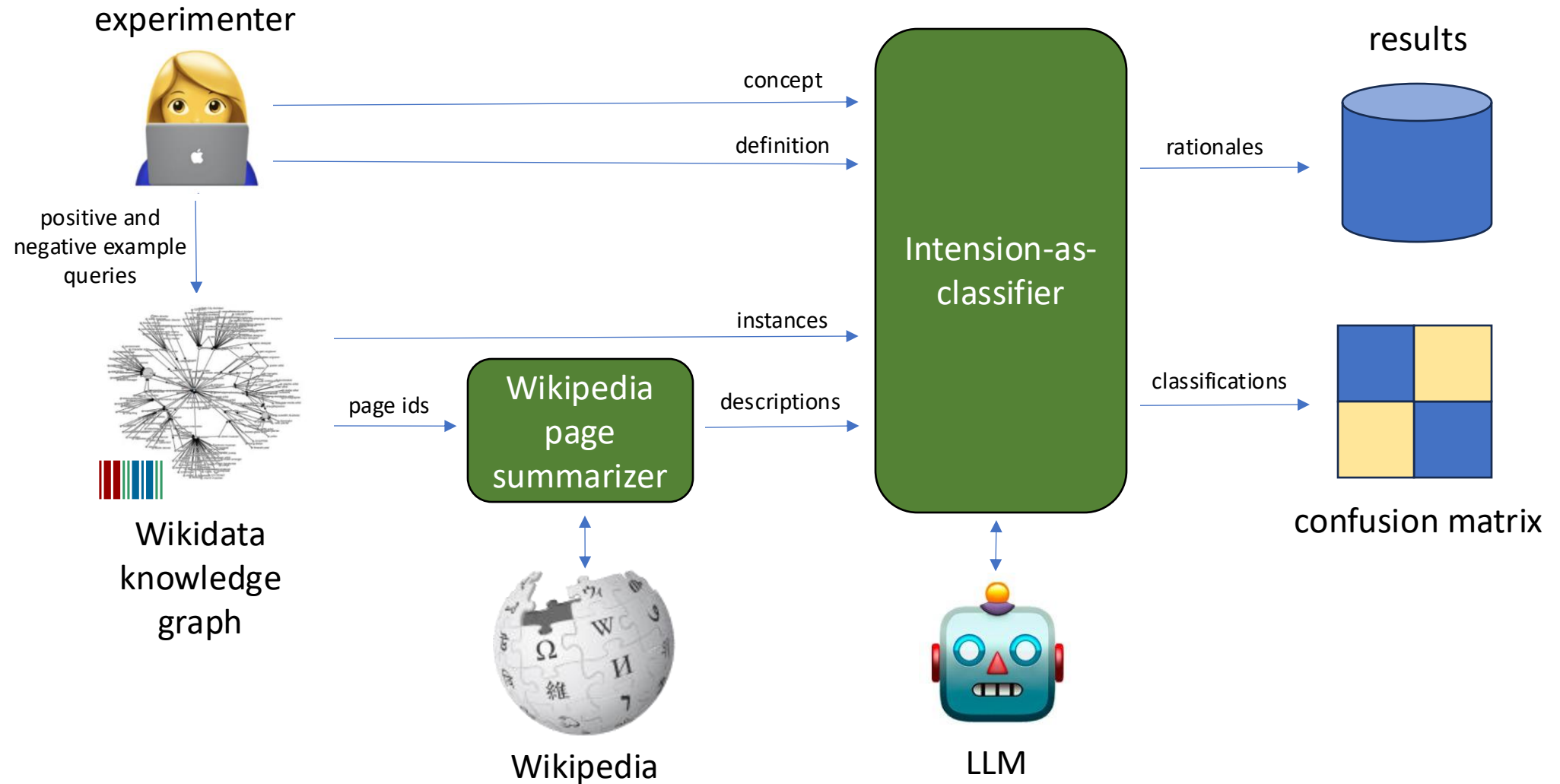
Formal semantics: intensions

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

Formal semantics: extensions

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

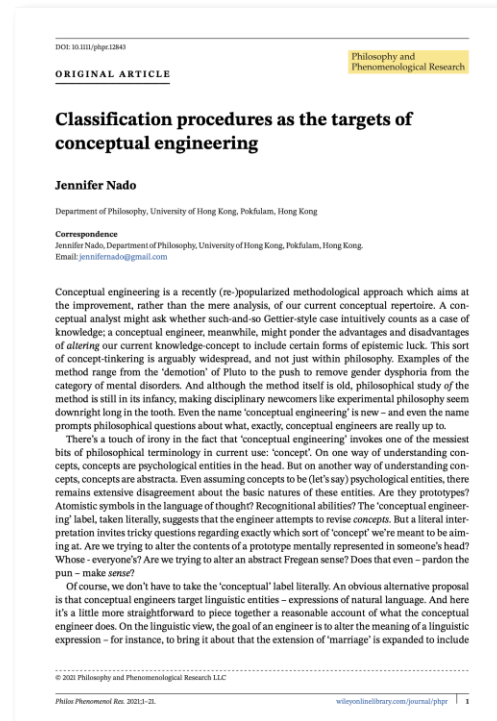
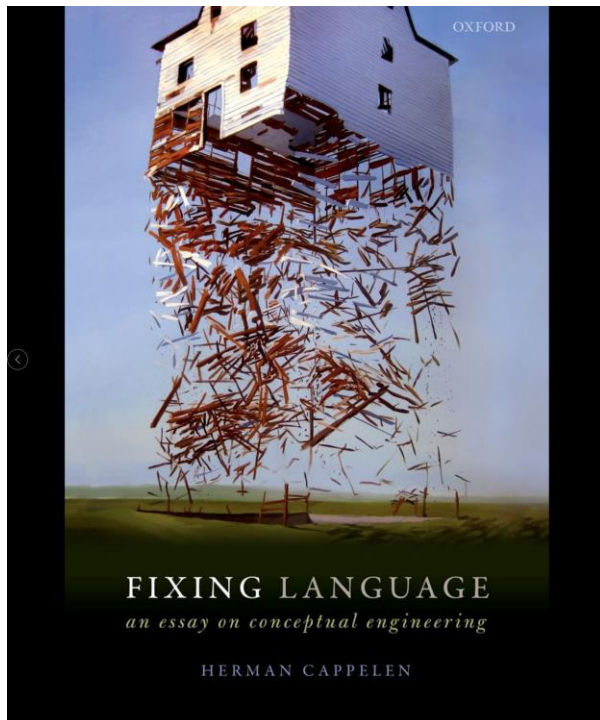
Carnap's Robot Redux



Investigations

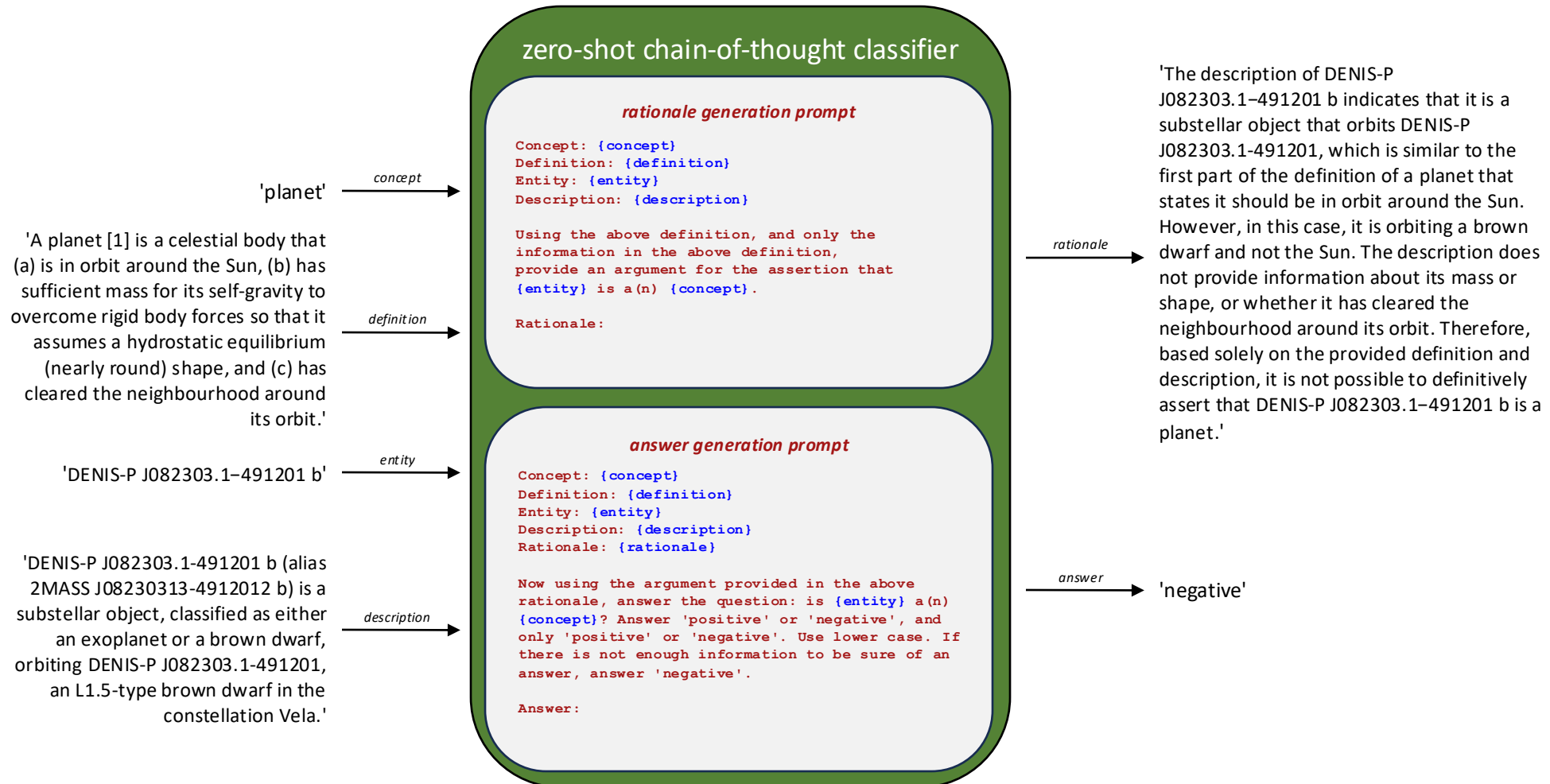
- Conceptual engineering
- Knowledge graph refinement

Classifiers as targets of conceptual engineering



- 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
- 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* (CPs), defined as abstract 'recipes' which sort entities "into an 'in'-group and an 'out'-group"

Implementing Nado's classification procedures



Experiments

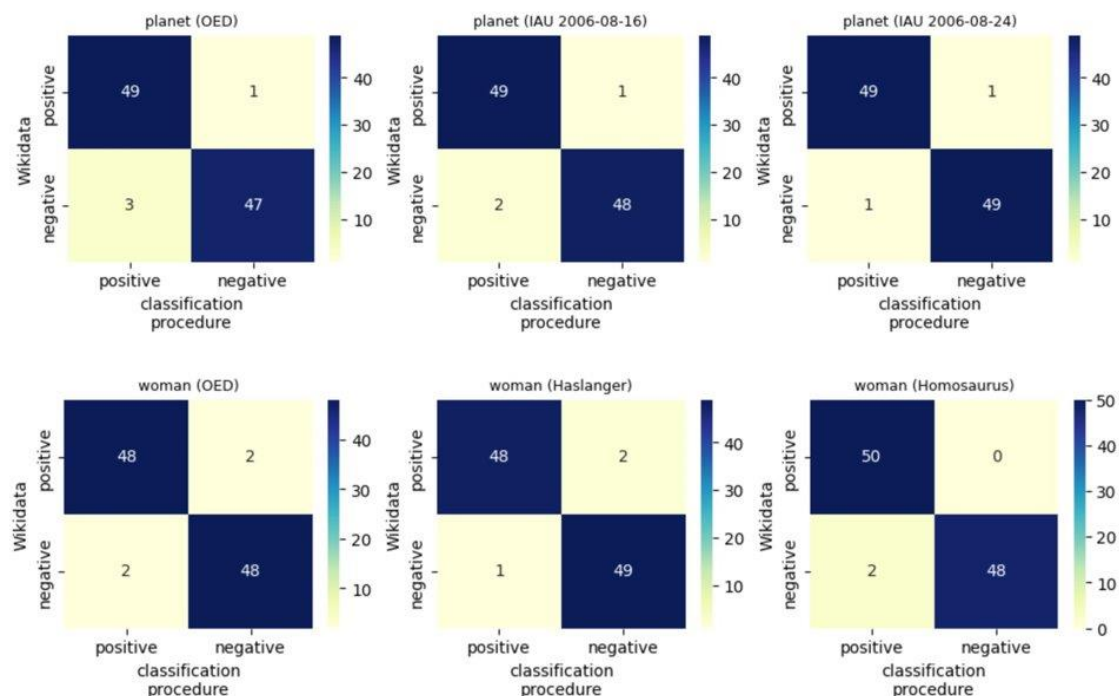
- PLANET

- 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

```
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.

Knowledge graph refinement

- Knowledge graph (KG) refinement is an increasingly important topic
- KG refinement is difficult
 - KGs can be messy (e.g., Wikidata quality issues)
 - Human evaluation is expensive
- Machine learning (ML) and natural language processing (NLP) may help by enabling automation of KG evaluation and refinement tasks
- Large language models (LLMs) are state-of-the-art ML/NLP technology with a broad range of application
- LLMs 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 and Brown]

Experiments

input : a pair of classes c, d from $G \mid (c, \text{subClassOf}, d) \in G$

output: a confusion matrix M

$(TP, FP, TN, FN) \leftarrow (0, 0, 0, 0);$

$E^+ \leftarrow$ a sample from $\text{ext}(c);$

$E^- \leftarrow$ a sample from $\text{ext}(d) \setminus \text{ext}(c);$

foreach $e \in E^+$ **do**

$(T_R, T_B) \leftarrow \text{classify}(c, e);$

if $T_B = \textit{positive}$ **then** $TP \leftarrow TP + 1;$

else $FP \leftarrow FP + 1;$

end

foreach $e \in E^-$ **do**

$(T_R, T_B) \leftarrow \text{classify}(c, e);$

if $T_B = \textit{negative}$ **then** $TN \leftarrow TN + 1;$

else $FN \leftarrow FN + 1;$

end

$M \leftarrow [[TP, FP], [FN, TN]];$

- LLMs:
 - gpt-4-0125-preview
 - gpt-3.5-turbo
 - gemma-7b-it
 - gemma-2b-it
 - Mixtral-8x7B-Instruct-v0.1
 - Mistral-7B-Instruct-v0.2
 - Llama-2-70b-chat-hf
- 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

Performance metrics

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

- 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

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%)

- Error analysis based on review by one of the authors
- 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

Incorrect relation (Wikidata)

T_c

Village (Japan): A village (村, mura) is a local administrative unit in Japan. It is a local public body along with prefecture (県, ken, or other equivalents), city (市, shi), and town (町, chō, sometimes machi). Geographically, a village's extent is contained within a prefecture. It is larger than an actual settlement, being in actuality a subdivision of a rural district (郡, gun), which are subdivided into towns and villages with no overlap and no uncovered area. As a result of mergers and elevation to higher statuses, the number of villages in Japan is decreasing. As of 2006, 13 prefectures no longer have any villages: Tochigi (since March 20, 2006), Fukui (since March 3, 2006), Ishikawa (since March 1, 2005), Shizuoka (since July 1, 2005), Hyōgo (since April 1, 1999), Mie (since November 1, 2005), Shiga (since January 1, 2005), Hiroshima (since November 5, 2004), Yamaguchi (since March 20, 2006), Ehime (since January 16, 2005), Kagawa (since April 1, 1999), Nagasaki (since October 1, 2005), and Saga (since March 20, 2006). Atarashiki-mura (which is an autonomous village community) is not included in the list below.

T_e

Tachikawa: Tachikawa (立川市, Tachikawa-shi) is a city located in the western portion of Tokyo Metropolis, Japan. As of 1 March 2021, the city had an estimated population of 184,383 in 93,428 households, and a population density of 7600 persons per km². The total area of the city is 24.36 square kilometres (9.41 sq mi).

$T_{\mathbb{B}}$

negative

T_R

Based on the information provided and the definition of a village in Japan, it is not possible to argue that Tachikawa is a village. Here's why, step by step:

- Definition of a Village in Japan:** A village in Japan, as defined, is a local administrative unit that is a subdivision of a rural district (郡, gun). It is one of the types of local public bodies, alongside prefecture, city, and town. Villages are characterized by being part of the rural district's administrative structure.
- Status of Tachikawa:** Tachikawa is explicitly described as a city (立川市, Tachikawa-shi). In Japan, the designation of "city" is distinct from that of a "village". Cities and villages are different types of administrative units with specific criteria for their classification, including population size, administrative structure, and economic functions.
- Population and Density:** The population and density of Tachikawa, with an estimated population of 184,383 and a population density of 7600 persons per km², far exceed what would typically be expected of a village. Villages, by their nature, tend to have smaller populations and lower population densities, reflecting their rural character.
- Geographical and Administrative Context:** Tachikawa's location within the Tokyo Metropolis and its designation as a city indicate that it is part of a highly urbanized and densely populated area. This context is inconsistent with the typical geographical and administrative setting of a village, which would be within a rural district (郡, gun).
- Lack of Overlap with Rural Districts:** The definition of a village mentions that it is a subdivision of a rural district with no overlap and no uncovered area. Tachikawa, being a city, does not fit this criterion as it is not a subdivision of a rural district but rather an independent administrative unit within the Tokyo Metropolis.

In conclusion, based on the provided definition of a village in Japan and the specific details about Tachikawa, it is clear that **Tachikawa cannot be classified as a village**. Its status as a city, along with its population size, density, and urban context, directly contradicts the criteria for being considered a village in Japan.

Missing relation (CaLiGraph)

T_c

The entity "**State leader in 2014**" is a class within the Caligraph knowledge graph's ontology. It is a specific subclass of the broader "State leader" class. The information for this entity is derived from the Wikipedia page titled "List of state leaders in 2014".

This class includes several individuals who were state leaders in 2014. These individuals are 'Akilisi Pōhiva, Armando Guebuza, Barry Haase, Hovik Abrahamyan, Michael Dunkley, Miloš Zeman, Nicos Anastasiades, Omar al-Hassi, Tsakhiagiin Elbegdorj, Adam Wood, Beji Caid Essebsi, Harold Martin, Jorge Carlos Fonseca, Joyce Banda, Mikhail Myasnikovich, and Hassan Rouhani. Each of these individuals is classified as an instance of the "State leader in 2014" class, indicating their role as state leaders during that year.

T_e

Zlatko Lagumdžija, also known as "Lagumdžija", is a Bosniak individual who was born in 1955 and is currently living. He hails from Sarajevo, Bosnia and Herzegovina. He has had a significant political career, serving as the Foreign Minister of Bosnia and Herzegovina in several years, including 2001, 2002, 2012, 2013, 2014, and 2015. In addition to his role as Foreign Minister, he also held a position as a state leader in 2001 and 2002. His contributions and roles have led to his recognition as a notable Bosniak of Bosnia and Herzegovina and a prominent politician from Sarajevo.

$T_{\mathbb{B}}$

positive

T_R

1. Definition of "State leader in 2014": The class "State leader in 2014" within the Caligraph knowledge graph's ontology is defined as a subclass of "State leader" that specifically includes individuals who were state leaders in the year 2014. This class is populated based on the roles these individuals held during that year, as derived from the Wikipedia page titled "List of state leaders in 2014".

2. Role of Zlatko Lagumdžija in 2014: According to the provided information, Zlatko Lagumdžija served as the Foreign Minister of Bosnia and Herzegovina in several years, including 2014. The role of a Foreign Minister is a significant position within the government, involving the oversight of international relations and diplomacy on behalf of the state. This position is typically considered a leadership role within the state's executive branch or government.

3. Criteria for being a "State leader": The role of Foreign Minister inherently involves leadership responsibilities at the state level, particularly in the realm of international affairs. Given that the class "State leader" would encompass individuals who held significant leadership positions within their respective states, serving as a Foreign Minister qualifies an individual as a state leader.

4. Conclusion: Based on the definition of "State leader in 2014" as individuals who were state leaders during that year, and given that Zlatko Lagumdžija served as the Foreign Minister of Bosnia and Herzegovina in 2014, it logically follows that Zlatko Lagumdžija meets the criteria to be classified as a "State leader in 2014". His role as Foreign Minister in 2014 directly aligns with the criteria for inclusion in this specific class within the Caligraph knowledge graph's ontology. **Therefore, Zlatko Lagumdžija is indeed a "State leader in 2014" according to the provided definition and information.**

Missing data (CaLiGraph)

concept

The entity "**FK Sarajevo player**" is a class within the ontology, specifically labeled as "FK Sarajevo player". It is a subclass of several other classes, including "RestrictionHasValue_team_FK Sarajevo", "Footballer in Bosnia and Herzegovina", and "Association football player in Bosnia and Herzegovina". This suggests that the class "FK Sarajevo player" is a specific type of football player who is associated with the team FK Sarajevo in Bosnia and Herzegovina. The information for this entity was derived from two Wikipedia pages: "Category:FK Sarajevo players" and "List of FK Sarajevo players". This suggests that the data for this class was gathered from these pages and includes all players listed there. Several individuals are classified under this entity, indicating that they are or were players for FK Sarajevo. These individuals include Brislav Arsenijević, Sead Jesenković, Darko Raca, Franjo Vuleta, Ibrahim Biogradlić, Vule Trivunović, Sretko Vuksanović, Vukašin Višnjevac, Marko Savić, Frank Liivak, Matej Kochan, Safet Nadarević, and Said Fazlagić. Each of these individuals is classified as an "FK Sarajevo player", reinforcing their association with this specific football team.

T_e

Darko Raca is a named individual, born in 1977, who is a professional football defender from Bosnia and Herzegovina. He has played for several football clubs including FK Sloboda Tuzla, FK ČSK Čelarevo, FK Kozara Gradiška, and FK Laktaši. His career has seen him play in various leagues such as the Serbian First League, Division 2 of Swedish football, and Ettan Fotboll. Darko Raca is not only recognized as an expatriate footballer who has played in Serbia, Serbia and Montenegro, and Sweden, but also as a Bosnia and Herzegovina expatriate sportsperson in these countries. He is further categorized as a foreign football player in Serbia. Darko Raca, a living person, hails from Sarajevo.

$T_{\mathbb{B}}$

negative

T_R

- 1. Identification of the Entity Class:** The entity class "FK Sarajevo player" is defined as a specific type of football player who is associated with the team FK Sarajevo in Bosnia and Herzegovina. This class is a subclass of broader categories such as "Footballer in Bosnia and Herzegovina" and "Association football player in Bosnia and Herzegovina", but it specifies the team association to FK Sarajevo.
- 2. Inclusion Criteria for the Entity Class:** To be classified under the "FK Sarajevo player" entity, an individual must have an association with the FK Sarajevo football team. This association is typically through playing for the team, as indicated by the sources of information (Wikipedia pages on FK Sarajevo players).
- 3. Analysis of Darko Raca's Profile:** Darko Raca is identified as a professional football defender from Bosnia and Herzegovina, with a career that includes playing for several football clubs. The clubs mentioned are FK Sloboda Tuzla, FK ČSK Čelarevo, FK Kozara Gradiška, and FK Laktaši. His career spans various leagues, including the Serbian First League and Division 2 of Swedish football.
- 4. Lack of Direct Mention of FK Sarajevo:** In the provided information about Darko Raca, there is no direct mention of him playing for FK Sarajevo. His association with other clubs is detailed, but FK Sarajevo is not listed among the clubs he has played for.
- 5. Conclusion Based on Provided Information:** Based solely on the information provided in the definition of "FK Sarajevo player" and the detailed profile of Darko Raca, there is no evidence to directly assert that Darko Raca is an FK Sarajevo player. The assertion that an individual is an FK Sarajevo player requires evidence of their association with FK Sarajevo, specifically playing for the team, which is not present in the information given about Darko Raca.
- 6. Final Argument:** Therefore, without additional information linking Darko Raca to FK Sarajevo directly, such as records of him playing for the team or being listed on the team's roster, **we cannot logically conclude that Darko Raca fits the classification of an "FK Sarajevo player"** based on the provided definitions and his profile information.

Incorrect reasoning (Wikidata)

 T_c

History of cities: Towns and cities have a long history, although opinions vary on which ancient settlements are truly cities. The benefits of dense settlement included reduced transport costs, exchange of ideas, sharing of natural resources, large local markets, and in some cases amenities such as running water and sewerage. Possible costs would include higher rate of crime, higher mortality rates, higher cost of living, worse pollution, traffic and high commuting times. Cities grow when the benefits of proximity between people and firms are higher than the cost.

 T_e

military history of Japan: The military history of Japan covers a vast time-period of over three millennia - from the Jōmon (c. 1000 BC) to the present day. After a long period of clan warfare until the 12th century, there followed feudal wars that culminated in military governments known as the Shogunate. Japanese history records that a military class and the Shōgun ruled Japan for 676 years - from 1192 until 1868. The Shōgun and the samurai warriors stood near the apex of the Japanese social structure - only the aristocratic nobility nominally outranked them. The sakoku policy effectively closed Japan from foreign influences for 212 years - from 1641 to 1853. Feudal militarism transitioned to imperialism in the 19th century after the arrival of Admiral Perry in 1853 and the elevation of Emperor Meiji in 1868. Western colonial powers and their imperialist policies impacted on Japan's outlook and led to Japanese colonialism and rampant imperialism (c. 1895 - 1945) until Japan's defeat in World War II. The 1947 Japanese Constitution prohibits Japan from offensively using war against other nations. This led to the establishment of the Japan Self-Defense Forces in 1954. The U.S.–Japan Alliance (1951 onwards) requires the United States of America to protect Japan and to conduct offensive duties. In 2015 the Constitution was reinterpreted to allow collective self-defense of Japan's allies. As of 1954 the Japan Self-Defense Forces (JSDF) consist of the Japan Ground Self-Defense Force (JGSDF), Japan Maritime Self-Defense Force (JMSDF) and Japan Air Self-Defense Force (JASDF). The Prime Minister is the commander-in-chief of the Japan Self-Defense Forces. Military authority runs from the Prime Minister to the cabinet-level Minister of Defense of the Japanese Ministry of Defense. The Prime Minister and Minister of Defense are advised by the Chief of Staff, Joint Staff, who heads the Joint Staff (統合幕僚監部, Tōgō Bakuryō Kanbu). The Chief of Staff, Joint Staff, is the highest-ranking military officer in the Japan Self-Defense Forces, and is the head of the Operational Authority over the JSDF, executing orders of the Minister of Defense with directions from the Prime Minister.

 $T_{\mathbb{B}}$

positive

To argue that the military history of Japan is a history of cities, we need to draw connections between the development, significance, and transformation of Japanese cities and their military history. The information provided offers several avenues to establish this relationship:

- 1. Feudal Wars and the Shogunate:** The period of clan warfare until the 12th century and the subsequent feudal wars that led to the establishment of military governments known as the Shogunate had profound impacts on the development and structure of Japanese cities. Cities during these times often centered around castles or fortifications, serving as administrative and military hubs. The Shogunate system, with its military class and samurai warriors, influenced the social and physical structure of cities, prioritizing defense and the needs of the ruling military elite.
- 2. Sakoku Policy:** The sakoku policy, which effectively closed Japan from foreign influences for 212 years, had significant implications for Japanese cities. This isolationist policy led to a period of internal peace and stability, allowing for urban development and growth without the external military threats that characterized earlier periods. Cities became centers of culture, commerce, and internal governance, reflecting the shift in focus from military defense to internal administration and development.
- 3. Transition to Imperialism:** The arrival of Admiral Perry and the elevation of Emperor Meiji marked the end of feudal militarism and the beginning of imperialism. This transition had a transformative effect on Japanese cities, as the country rapidly industrialized and modernized. Cities expanded and modernized, reflecting the new imperial ambitions and the influence of Western colonial powers. The growth of cities during this period was directly tied to Japan's military and imperial ambitions, as urban centers became crucial for industrial production, military recruitment, and administration of the growing empire.
- 4. Post-World War II Reconstruction and the Japan Self-Defense Forces:** Following Japan's defeat in World War II, the country underwent a period of reconstruction that again reshaped its cities. The establishment of the Japan Self-Defense Forces and the reinterpretation of the Constitution to allow collective self-defense led to the development of military bases and facilities within and around cities. The presence of the JSDF and the U.S.–Japan Alliance influenced the strategic importance of certain urban areas and their development post-war.
- 5. Modern Urban Development and Military Influence:** The structure and development of modern Japanese cities cannot be fully understood without considering their military history. From the feudal castles that formed the nuclei of historical cities to the modern bases of the JSDF that impact local economies and urban landscapes, the military history of Japan is deeply intertwined with the history of its cities. In conclusion, **the military history of Japan is indeed a history of cities**, as it has significantly influenced the development, structure, and transformation of Japanese urban centers throughout the centuries. From the feudal periods to modern times, military needs, policies, and transformations have played a central role in shaping Japanese cities, making the military history of Japan an integral part of the broader history of its urban development.

 T_R

Objections

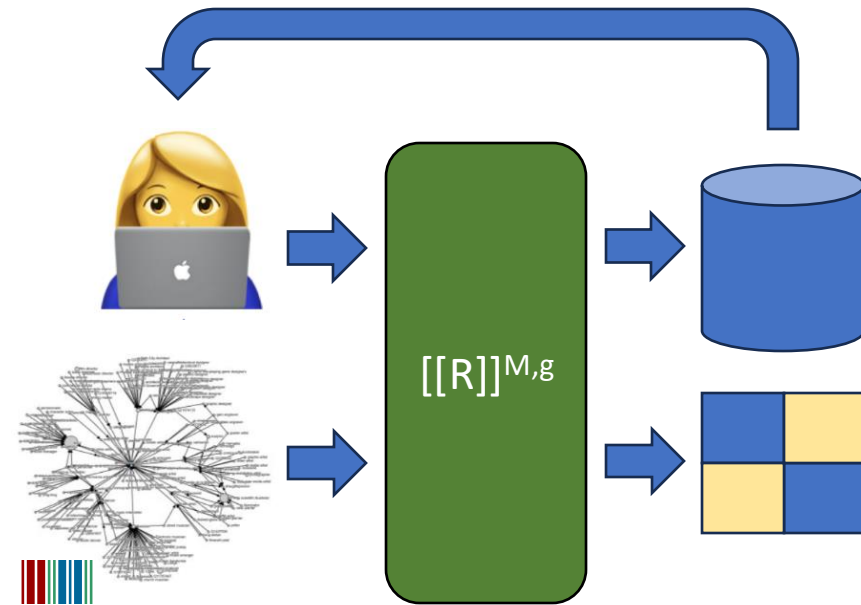
- Trustworthiness
 - Humans can be misled by rationales
- Groundedness
 - Can LLMs make meaningful statements?
 - Can LLMs *simulate* meaning in a useful way?
 - Does this matter for our purposes?

Applications

- LLMs as tools for experimental conceptual engineering
- Knowledge graph refinement with LLMs as a solution to the implementation problem

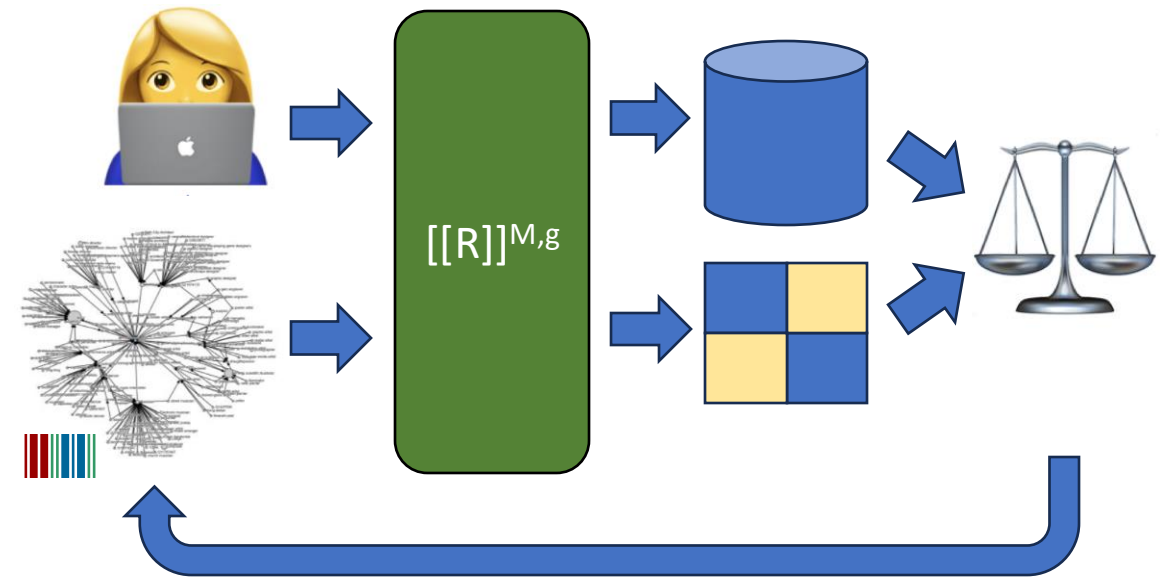
LLMs as tools for experimental conceptual engineering

- Assuming this, we argue that the above method can provide a useful cognitive tool (Menary & Gillett, 2022; Novaes, 2012) for conceptual engineers to compare the extension of a proposed concept definition to the extensional knowledge represented as facts in a given knowledge base (Allen & Groth, 2024).
- This provides an approach to calibrate trust in an LLM used in this manner, and can also be viewed as an instance of a corpus method for experimental philosophy (x-phi) (Fischer & Sytsma, 2022; Sytsma, 2023), with relevance to the relationship between x-phi, conceptual engineering, and Carnapian explication (Koch, 2019; Pinder, 2017; Shepherd & Justus, 2015).
- LLMs can be used to simulate human populations surveyed to express their judgment on concepts



The implementation problem in conceptual engineering

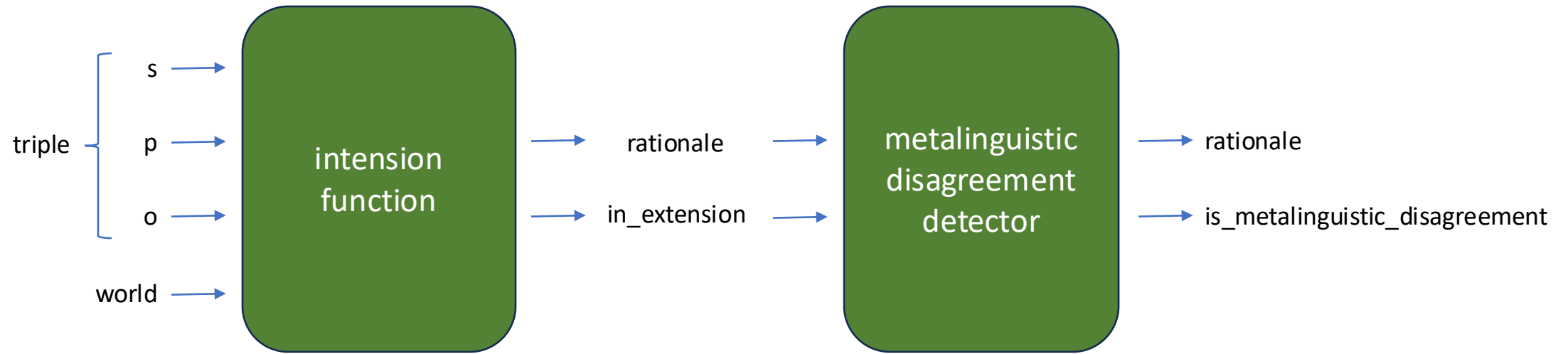
- We argue that the above method provides a possible solution to the implementation problem in conceptual engineering, which poses the question of whether (re)engineered concepts can be effectively adopted by a population of human speakers (Cappelen, 2018; Jorem, 2021).
- Online knowledge bases such as Wikidata (Vrandečić & Krötzsch, 2014) have a direct and material impact on society by virtue of their use in online search, discovery, and recommendation (Peng et al., 2023).
- Using the above method to guide changes to facts in a knowledge base to better align with the extension of a proposed definition provides an indirect method for shifting the semantic meaning of a concept for the specific linguistic subgroup (Matsui, 2024) constituted by users of such online knowledge bases.



Recent work: LLMs and metalinguistic disagreement

- *Metalinguistic disagreement* between an LLM and a KG presents a problem for the use of LLMs for text-to-fact extraction and fact-to-text verbalization
- LLMs for KE must address *metalinguistic negotiation* to resolve this challenge
- Our work shows how to
 - Identify concepts where metalinguistic disagreement exists
 - Quantify the degree of disagreement in terms of inter-annotator alignment

LLMs and metalinguistic disagreement



LLMs and 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

LLMs and 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.

Research questions

- 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?

Conclusions

- Carnap's Robot, a thought experiment from 70 years ago, turns out to be a useful tool for exploring the factuality and trustworthiness of LLMs
 - Carnap anticipated the central challenge associated with working with LLMs: can we use them as black boxes, or must we solve the problem of mechanistic interpretability?
- We would love to answer the questions around belief and meaning in LLMs, but we do not have to definitively answer those questions for them to be useful as cognitive tools for working with knowledge and meaning
 - LLMs open a path towards experimental conceptual engineering
 - LLMs used for knowledge base refinement can provide a way to address the implementation problem in conceptual engineering



UNIVERSITEIT
VAN AMSTERDAM

INDE lab

Thank you!

b.p.allen@uva.nl

<https://www.bradleypallen.org>