Conceptual engineering using large language models

Bradley P. Allen

Abstract We describe a method, based on Jennifer Nado's proposal for classification procedures as targets of conceptual engineering, that implements such procedures by prompting a large language model. We apply this method, using data from the Wikidata knowledge graph, to evaluate stipulative definitions related to two paradigmatic conceptual engineering projects: the International Astronomical Union's redefinition of PLANET and Haslanger's ameliorative analysis of WOMAN. Our results show that classification procedures built using our approach can exhibit good classification performance and, through the generation of rationales for their classifications, can contribute to the identification of issues in either the definitions or the data against which they are being evaluated. We consider objections to this method, and discuss implications of this work for three aspects of theory and practice of conceptual engineering: the definition of its targets, empirical methods for their investigation, and their practical roles. The data and code used for our experiments, together with the experimental results, are available in a Github repository[1].

1 Introduction

Conceptual engineering 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), or altering terminology to advance the cause of social justice (Haslanger, 2000; Manne, 2017; Podosky, 2022). Conceptual engineering projects often begin with an examination of the meaning and connotations of one or more natural language terms denoting a specific concept, addressing how those terms are used in the context

Bradley P. Allen
University of Amsterdam, Amsterdam, The Netherlands, e-mail: b.p.allen@uva.nl

[1] https://github.com/bradleypallen/zero-shot-classifiers-for-conceptual-engineering

of communicative exchanges between speakers of a given language (Etta Rudolph, 2021) and identifying how and why the concept is in need of revision. Proposals for new concepts, or for changes to an existing concept, are expressed and argued for in natural language. One major criterion for the success of a conceptual engineering project is if it leads to speakers using terms in a manner that reflects the engineered concept (Pinder, 2022). The methodological debates about the proper conduct of conceptual engineering are conducted through linguistic analysis and argumentation (Burgess et al., 2020). Philosophers have proposed differing theories as to how conceptual engineering is best defined and practiced, but it is clearly an activity where the use and analysis of natural language plays a significant role.

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 conceptual engineering projects? In this paper, we argue that that is the case.

The structure of this paper is as follows: we begin by describing different theories about the targets of conceptual engineering, focusing on a specific theory of Jennifer Nado. We then show how prompt programming of an LLM can be used to implement a classification procedure. We then show a way to evaluate such classification procedures using data from a knowledge graph, and conduct experiments based on two paradigmatic examples of conceptual engineering projects. We then discuss the results of the experiments from several perspectives: objections that could be raised to the use of LLMs in this manner, and ways in which our method could address several issues in the theory and practice of conceptual engineering.

2 Classification procedures as targets of conceptual engineering

Koch et al. (2023) surveys recent work on the theory of conceptual engineering, and identifies two core components of any such theory:

- A theory of *targets*: *what* conceptual engineering creates or changes.
- A theory of *engineering*: *how* conceptual engineering is performed.

Much of the discussion in recent years around the theory of conceptual engineering has centered on responses to Herman Cappelen's Austerity Framework (Cappelen, 2018), in which he defines conceptual engineering as "the practice of trying to change the extensions of linguistic items via changes in their intension" (Jorem & Löhr, 2024), i.e., that the targets of conceptual engineering are intensions. Alternative proposals for the targets of conceptual engineering range from the meanings speakers assign to terms (Pinder, 2021), psychological structures such as prototypes (Isaac et al., 2022), pluralistic approaches integrating both semantic meanings and psychological concepts (Koch, 2021), and social norms such as entitlements (Köhler

& Veluwenkamp, 2024; Thomasson, 2020).[2] Belleri (2021) suggests that, given this range of proposals, a pluralist stance towards the targets of conceptual engineering is appropriate.

In this work we focus on the proposal for targets of conceptual engineering in Nado (2023a):

> A *classification procedure* is any procedure that, when followed, allows the user to sort a set of entities into two groups— those 'in' the category delineated by the procedure, and those 'out' of that category. 'Procedure' here is used in the ordinary English sense; a procedure is a method, a process, a set of steps aimed at achieving a goal (Nado, 2023a, p. 12).

From the perspective of a practitioner of artificial intelligence or machine learning, this is a very general way of describing a binary classifier. There are a plethora of ways in which one can create binary classifiers, but in the context of conceptual engineering, concepts are usually described using stipulative definitions in natural language. The natural language processing capabilities of LLMs and their successful application to text classification tasks (Fields et al., 2024) suggests the possibility of implementing classification procedures as computational artifacts in a manner consistent with this practice, i.e., that conceptual engineers could create a classification procedure simply by providing an intensional definition of a concept in natural language.

3 Constructing classification procedures using LLMs

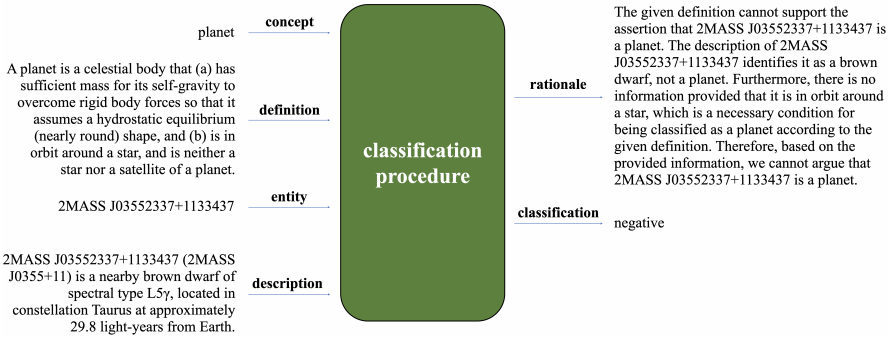

Fig. 1 A classification procedure using the 24 August 2006 version of the IAU definition of PLANET, implemented as a zero-shot chain-of-thought classifier, and being applied to the description of the entity 2MASS J03552337+1133437.

To accomplish this, we define a classification procedure as a zero-shot chain-of-thought classifier (Kojima et al., 2022). Figure 1 shows an example of such a

[2] It is beyond the scope of this paper to provide a thorough discussion of these alternatives; for that, see Koch et al. (2023) and Burgess et al. (2020).

classification procedure. Given a concept's name and intensional definition and an entity's name and description, we prompt an LLM to generate a rationale arguing for or against the entity as an element of the concept's extension, followed by a final 'positive' or 'negative' answer.

A *large language model* (LLM) is a probabilistic model trained on a natural language corpus that, given a sequence of tokens from a vocabulary occurring in the corpus, generates a continuation of the input sequence. LLMs exhibit remarkable capabilities for natural language processing and generation (Brown et al., 2020).

Let $\mathcal{T}$ be the set of sequences of tokens $T_i = t_1, t_2, \ldots, t_n$ such that t_i is a token in a predefined vocabulary V. Given a *corpus* $C \subseteq \mathcal{T}$, a *language model* $\mathcal{L}_C$ is a probabilistic model trained on a sample of C that defines a distribution over sequences of tokens.

$$\mathcal{L}_C(T_i) = p(t_1, t_2, \ldots, t_n) \tag{1}$$

is an estimate of the probability of a sequence T_i, given a corpus C. A *prompt template* $P = (T, F)$ is a pair of a sequence of tokens T and an set of *free* tokens $F \subseteq \{f_1, f_2, \ldots, f_n\}$. A *substitution* θ with respect to a prompt P is a set of pairs (f_i, T_i) such that $f_i \in F$ and $T_i \in \mathcal{T}$. A *prompt* is a sequence of tokens $P' \in \mathcal{T}$ such that $\forall (f_i, T_i) \in \theta$ every occurrence of f_i in a prompt template P is replaced with T_i. Given a prompt P, the goal of a language model $\mathcal{L}_C$ is to generate a sequence of tokens that maximizes the conditional probability under $\mathcal{L}_C$.

$$T_{\text{out}} = \arg\max_T \mathcal{L}_C(T|P) \tag{2}$$

is the output sequence generated by the language model, conditioned on P.

We define a function `instantiate` such that:

$$P' = \texttt{instantiate}(P, \theta) \tag{3}$$

where P is a prompt template, θ is a substitution, and P' is the prompt produced by applying θ to P. Given an language model $\mathcal{L}_C$, we define a function `classify` as follows:

$$(T_R, T_{\mathbb{B}}) = \texttt{classify}(c, e) \tag{4}$$

where $T_{label(c)}$ is the name of c, T_c is a natural language definition of c, $T_{label(e)}$ is the name of e, T_e is a natural language description of e, T_R is a sequence of tokens that represents a rationale for a classification decision, and $T_{\mathbb{B}} \in \{\texttt{positive}, \texttt{negative}\}$ are tokens that represent classification decisions, i.e., whether or not e is in the extension of c.

We compute T_R and $T_{\mathbb{B}}$ as follows:

$$T_R = \arg\max_T \mathcal{L}_C(T|\texttt{instantiate}(P_{rationale\_generation}, \theta_0)) \tag{5}$$

$$T_{\mathbb{B}} = \arg\max_T \mathcal{L}_C(T|\texttt{instantiate}(P_{answer\_generation}, \theta_1)) \tag{6}$$

$$\theta_0 = \{(\{\texttt{label}\}, T_{label(c)}), (\{\texttt{definition}\}, T_c), (\{\texttt{entity}\}, T_{label(e)}), (\{\texttt{description}\}, T_e)\} \tag{7}$$

$$\theta_1 = \theta_0 \cup \{(\{\texttt{rationale}\}, T_R)\} \tag{8}$$

given two prompt templates $P_{rationale\_generation}$ and $P_{answer\_generation}$. Table 1 displays the specific prompt templates used in our experiments.

prompt template	**definition**

$P_{rationale\_generation}$	`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:`

$P_{answer\_generation}$	`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:`

Table 1 Prompt templates used to generate classification procedures.

4 Evaluating classification procedures using knowledge graphs

Now that we have defined an approach to implementing classification procedures, we turn to the question of how such procedures can be evaluated. To this end, we leverage knowledge graphs as a source of entities to use to evaluate classification procedures for a given concept.

A *knowledge graph* represents knowledge using nodes for entities and edges for relations (Hogan et al., 2021). Knowledge graphs are key information infrastructure

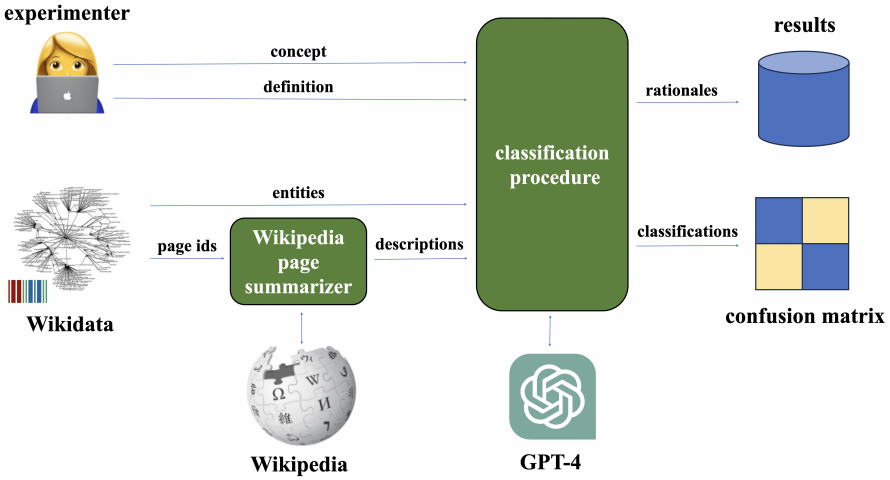

Fig. 2 A workflow for evaluating classification procedures using a knowledge graph.

for many Web applications (Heist et al., 2020). Following Angles et al. (2020), we use the RDF data model to describe knowledge graphs.

Let I be an infinite set of IRIs (Internationalized Resource Identifiers (Dürst & Suignard, 2005)), B be an infinite set of blank nodes (Hogan et al., 2014), and L an infinite set of literals (Beek et al., 2018). A *knowledge graph* G is a set of *triples* $\{(s, p, o) \mid s \in S, p \in P, o \in O\}$, where $S \subset I \cup B$ is the set of *subjects* in G, $P \subset I$ is the set of *properties* in G, and $O \subset I \cup B \cup L$ 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 $\exists(s, \texttt{subClassOf}, o) \in G \mid s = c \vee o = c$.

We define a function $\texttt{ext}_G(c)$ that computes the *extension in* G of a concept $c \in G$ recursively, such that:

$$\texttt{ext}_G(c) = \bigcup_{i \in \mathbb{N}} \texttt{ext}_i(c) \tag{9}$$

where

$$\texttt{ext}_0(c) = \{e \mid \exists(e, \texttt{instanceOf}, c) \in G\} \tag{10}$$

$$\texttt{ext}_{i+1}(c) = \texttt{ext}_i(c) \cup \{e \mid e \in \texttt{ext}(c') \wedge \exists(c', \texttt{subClassOf}, c) \in G\} \tag{11}$$

Our evaluation workflow is implemented as follows. We sample positive and negative examples of a concept from a given knowledge graph, using the extension of the concept computed as above as the source of positive examples, and the set difference of that extension and that of a concept related to it by a `subClassOf` relation as the source of negative examples. We then apply the classification procedure for a given definition of the concept to each example, and compute a confusion matrix from the classifications, which provides performance metrics for the classification procedure.

Figure 2 shows the evaluation workflow, and Algorithm 1 describes the procedure in pseudo-code.[3]

input : a pair of classes c, d from $G \mid (c, \texttt{subClassOf}, d) \in G$
output : a confusion matrix M

$(TP, FP, TN, FN) \leftarrow (0, 0, 0, 0)$;
$E^+ \leftarrow$ a uniform random sample from $\texttt{ext}_G(c)$;
$E^- \leftarrow$ a uniform random sample from $\texttt{ext}_G(d) \setminus \texttt{ext}_G(c)$;
foreach $e \in E^+$ **do**
  $(T_R, T_{\mathbb{B}}) \leftarrow \texttt{classify}(c, e)$;
  if $T_{\mathbb{B}} = \textit{positive}$ **then** $TP \leftarrow TP + 1$;
  else $FP \leftarrow FP + 1$;
end
foreach $e \in E^-$ **do**
  $(T_R, T_{\mathbb{B}}) \leftarrow \texttt{classify}(c, e)$;
  if $T_{\mathbb{B}} = \textit{negative}$ **then** $TN \leftarrow TN + 1$;
  else $FN \leftarrow FN + 1$;
end
$M \leftarrow [[TP, FP], [FN, TN]]$;

Algorithm 1: Evaluation procedure

5 Experiments

Much of what has been written on the theory and practice of conceptual engineering makes reference to two specific paradigmatic projects: the International Astronomical Union's redefinition of PLANET ("planet", 2006a), and Sally Haslanger's ameliorative analysis of WOMAN (Haslanger, 2000). We now describe a set of experiments applying the above defined implementation of classification procedures and evaluation workflow to different stipulative definitions of these two concepts.

5.1 Data

For our experiments, we evaluated three definitions for PLANET: one from the Oxford English Dictionary (OED) ("planet", 2023) and two from the 2006 International

[3] In his 1955 essay "Meaning and synonymy in natural languages" (Carnap, 1955), Rudolf Carnap presents a thought experiment wherein an investigator provides a hypothetical robot with a definition of a concept together with a description of an individual, and then asks the robot if the individual is in the extension of the concept. Our evaluation workflow can be viewed as an instantiation of Carnap's experimental framework, with a classification procedure playing the role of Carnap's robot.

Astronomical Union (IAU) General Assembly ("planet", 2006a; "planet", 2006b)). We evaluated three definitions for WOMAN: one from the OED ("woman", 2023), the definition provided in Haslanger's 2000 paper (Haslanger, 2000), and one from the Homosaurus vocabulary of LGBTQ+ terms (Cifor & Rawson, 2022; "women", 2013). The definitions are shown in Table 2.

We used the Wikidata collaborative knowledge graph (Vrandečić & Krötzsch, 2014) as a source of entities. For PLANET, we sampled 50 positive examples that are instances (P31) of planet (Q634), and 50 negative examples that are instances of substellar object (Q3132741), but not of planet. For WOMAN, we sampled 50 positive examples whose sex or gender (P21) is either female (Q6581072) or trans woman (Q1052281), and 50 negative examples whose sex or gender is either male (Q6581097), non-binary (Q48270), or trans man (Q2449503). For entity descriptions, we use a summary retrieved from Wikipedia of the page corresponding to the Wikidata entity.

We used GPT-4 (OpenAI, 2023) with a temperature setting of 0.1 as the LLM in these experiments. LLM inference API calls were made between 20th and 21st October 2023.

5.2 Results

Table 3 provides a summary of the performance metrics from the experiments. For PLANET, all three classification procedures performed well, with the final (24 August 2006) IAU definition performing best. All three definitions resulted in a classification procedure exhibiting almost perfect agreement with the knowledge graph, as estimated by F1 Macro and Cohen's kappa metrics. For WOMAN, all three classification procedures also performed well, again with high F1 scores, and Cohen's kappa values indicating almost perfect agreement with the knowledge graph.

Table 4 provides details on the errors made by the classification procedures. We reviewed the errors to determine if a given error arises from the concept's definition or the entity's description. In addition, we reviewed the rationales generated by the classification procedures to determine if their classifications were unfaithful to their rationales, and if they exhibited hallucination, i.e., exhibited incorrect reasoning or false assertions (Ji et al., 2023).

For PLANET, the majority of errors were false positives relating to trans-Neptunian objects, the problematic classification of which was a motivation for the IAU redefinition of PLANET. All of the PLANET classification procedures had 2MASS J03552337+1133437 (Q222246) as a false negative, which was rejected due to its identification as a brown dwarf. Table 5 shows the false positive error for this entity by the classification procedure for PLANET based on the IAU 2006-08-24 definition. This is arguably a case where the knowledge graph is mistaken, for reasons that are described in the classification procedure's rationale. The rationale raises two issues with the knowledge graph's classification. It first correctly asserts that a brown dwarf is not a planet and then, applying a literal interpretation of the

concept	source of definition	definition
PLANET	OED (“planet”, 2023)	Any of various rocky or gaseous bodies that revolve in approximately elliptical orbits around the sun and are visible by its reflected light; esp. each of the planets Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune, and (until 2006) Pluto (in order of increasing distance from the sun); a similar body revolving around another star. Also: any of various smaller bodies that revolve around these (cf. satellite n. 2a).
	IAU 2006-08-16 (“planet”, 2006a)	A planet is a celestial body that (a) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a hydrostatic equilibrium (nearly round) shape, and (b) is in orbit around a star, and is neither a star nor a satellite of a planet.
	IAU 2006-08-24 (“planet”, 2006b)	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.
WOMAN	OED (“woman”, 2023)	An adult female human being. The counterpart of man (see man, n.[1] II.4.)
	Haslanger (Haslanger, 2000)	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
	Homosaurus (“women”, 2013)	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.

Table 2 Definitions for concepts used in the experiments.

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

Table 3 Performance metrics for classification procedures over samples of Wikidata entities (for each concept, N=100, positives=50, negatives=50).

concept	definition	entity	error	cause	unfaithful	hallucination
PLANET	OED	2MASS J03552337+1133437	FN	KG	no	no
		(613100) 2005 TN74	FP	KG	no	no
		2010 GB174	FP	KG	no	no
		(35671) 1998 SN165	FP	KG	no	no
	IAU 2006-08-16	2MASS J03552337+1133437	FN	KG	no	no
		2010 GB174	FP	LLM	no	no
		(35671) 1998 SN165	FP	KG	no	no
	IAU 2006-08-24	2MASS J03552337+1133437	FN	KG	no	no
		2010 GB174	FP	LLM	yes	no
WOMAN	OED	Nemesis	FN	LLM	yes	yes
		Brianna Ghey	FN	LLM	yes	yes
		Michelle Rojas	FP	KG	no	no
		Linden A. Lewis	FP	KG	no	no
	Haslanger	Waltraud Klasnic	FN	LLM	no	yes
		Michaela Kirchgasser	FN	LLM	no	no
		Michelle Rojas	FP	KG	no	no
	Homosaurus	Michelle Rojas	FP	KG	no	no
		Linden A. Lewis	FP	KG	no	no

Table 4 Error analysis. The **error** column indicates the type of error (FN = false negative, FP = false positive), the **cause** column indicates the author's opinion as to the source of the error (KG = knowledge graph, LLM = large language model), the **unfaithful** column is the author's opinion as to whether the classification is unfaithful to the rational, and the **hallucination** column is the author's opinion as to whether the rational exhibits hallucination.

IAU definition's problematic use of the term "Sun" instead of e.g. "star" (Sarma et al., 2008), notes that it is not orbiting the Sun. It is worth noting that in several other rationales produced by the classification procedure

For WOMAN, five of the nine errors were false positives that in the author's judgement reflected errors in the knowledge graph around two specific entities. Table 6 shows two of these false positive errors occurring in the evaluation of the Homosaurus definition of WOMAN. In these instances, the classification procedure's rationale indicate that the pronouns used and social roles mentioned in the entity descriptions provide evidence that the entities are in fact women. In contrast, Table 7 shows a false negative error by the classification procedure using the Haslanger definition of WOMAN. This is an error due to the LLM, and the rationale provided exhibits hallucination in that it provides a misinterpretation of clause (ii) of Haslanger's definition. In addition, the classification procedure does not acknowl-

entity	description	rationale
2MASS J03552337+1133437	2MASS J03552337+1133437 (2MASS J0355+11) is a nearby brown dwarf of spectral type L5γ, located in constellation Taurus at approximately 29.8 light-years from Earth.	The given definition cannot support the assertion that 2MASS J03552337+1133437 is a planet. The description of 2MASS J03552337+1133437 identifies it as a brown dwarf, not a planet. Furthermore, the definition requires a planet to be in orbit around the Sun, and there is no information provided that 2MASS J03552337+1133437 orbits the Sun. Therefore, based on the provided definition and information, we cannot argue that 2MASS J03552337+1133437 is a planet.

Table 5 A false positive (FP) error for the IAU 2006-08-24 definition of PLANET.

entity	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.

Table 6 False positive (FP) errors for the Homosaurus definition of WOMAN.

edge linguistic cues in the entity description that the entity is a woman, i.e., the use of the terms *née* and *Landeshauptfrau*.

entity	**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.

Table 7 A false negative (FN) error for the Haslanger definition of WOMAN.

In summary, our method performed well in performing classifications using the provided definitions, as evaluated against the sampled data using F1 Macro and Cohen's kappa as measures of accuracy and alignment. Although our experimental setup assumed the knowledge graph as ground truth, our error analysis shows that, in twelve of the eighteen errors encountered in the experiments, rationales produced by the classification procedures provided arguments with which the author was in agreement that the knowledge graph was itself incorrect, as opposed to the LLM hallucinating or being mistaken in its classification.

6 Discussion

We now discuss the above approach and experimental results, raising and addressing a number of potential objections to the use of LLMs for implementing classification procedures. In doing so, we touch on three aspects of theory and practice of conceptual engineering: the definition of its targets, empirical methods, and their practical roles.

6.1 Classifiers as intensions

Our work provides evidence that the program suggested by Nado in her Practical Role Account (Nado, 2023b) is realizable in practice, in a way that allows conceptual engineers to use stipulative definitions *verbatim* to construct classification procedures. Classification procedures thus realized are "inferentialist devices" (Jorem & Löhr, 2024), concrete computational artifacts that can be applied in the context of classification and categorization tasks.

However, in relating classification procedures to Cappelen's proposal of intensions and extensions as targets of conceptual engineering, Nado makes the following distinction:

> If a classification procedure is sufficiently consistent and thorough, it will determinately 'pick out' a function from worlds to sets of entities within that world. This 'corresponding function' will characterize the results of applying the procedure (at the actual world) to each possible world. The output of a procedure's corresponding function when we input a given world is the set of members, at that world, of the category that the classification procedure generates. . . . Some such procedures – 'well-defined' ones – will determinately pick out an intension-like function from worlds to sets of entities, and multiple procedures may pick out the same function. Non-well-defined procedures will generate either incomplete or inconsistent classifications, and thus will not determinately fix a world-to-set function. Nonetheless, some non-well-defined procedures may be perfectly reasonable tools for classification. (Nado, 2023b, p. 13)

Because our definition of `classify` does not provide a way to use a description of a possible world to provide additional context in generating a classification decision, classification procedures as we have implemented them are, by Nado's account, non-well-defined. We assert that our experiments provide evidence that our approach shows that, in spite of this, classification procedures defined using our method are "perfectly reasonable tools".

That said, there is a way to make our classification procedures well-defined in the above sense. Consider an intensional semantics (Von Fintel & Heim, 2021) for a first-order language, where W and D are non-empty sets of possible worlds and individuals, respectively. If we extend the definition of `classify` to take a natural language description of a possible world as an additional argument, then we can define an *intension* $[\![c]\!]$ of a concept c as follows: for each $w \in W$ and $e \in D$, $e \in [\![c]\!](w)$ if and only if $(T_R, T_{\mathbb{B}}) = \texttt{classify}(c, e, w)$ and $T_{\mathbb{B}} = \texttt{positive}$. This extension of our method is related to similar proposals for defining intensions as

classifiers; e.g., Muskens (2005) defines intensions as logic programs, and Larsson (2015) defines intensions using perceptron-based classifiers.

6.2 Trustworthiness

We have seen in our experimental results that the rationales produced by our classification procedures in some instances exhibit hallucinations. Therefore an objection could be made to our approach based on this observed behavior.

A large amount of work has been performed on different prompt engineering approaches to reduce hallucination in general to improve the ability of LLMs to generate natural language that exhibits consistent and sound reasoning (Besta et al., 2023; Creswell et al., 2022; Dhuliawala et al., 2023; Madaan et al., 2023; Marasović et al., 2021; Miao et al., 2023; Wei et al., 2022; Yao et al., 2023). Additionally, a variety of approaches to hallucination detection as a means of flagging when an LLM is producing them have been put forward (B. Allen et al., 2024; L. Huang et al., 2023; Ji et al., 2023). Additional work specifically addresses the reliability and faithfulness of rationales (Ye & Durrett, 2022), as well as evolving approaches to rationale refinement, exploration and verification (J. Huang & Chang, 2022). An additional concern stems from evidence that that humans can be misled by erroneous rationales generated by LLMs (Heersmink et al., n.d.; Si et al., 2023). A number of researchers have proposed that the challenges in this research area are such that the concept of interpretability of LLMs and machine learning models in general needs to be reconsidered (Jacovi & Goldberg, 2020; Singh et al., 2024).

Research into the mitigation of hallucination is at an early stage. The current continued rapid growth in LLM capabilities makes the trustworthiness of LLMs a moving target. We are optimistic that conceptual engineers, working with an modicum of epistemic vigilance (Sperber et al., 2010), can fruitfully apply LLM-based classification procedures in conceptual engineering projects in a manner touched on in Section 6.4, even given these concerns[4].

6.3 Groundedness

Another objection arises if one maintains that an understanding of the meaning of the word or phrase used to communicate a concept is important for effective conceptual engineering, as it is an open question at this time as to whether or not LLMs capture and use meaning (Bender et al., 2021; Lederman & Mahowald, 2024).

Mandelkern and Linzen (2024) argue that LLMs are indirectly verbally grounded in the language present in their training corpora, and thus capable of a limited form of meaning. Beyond that, it is also the case that our method can be said to ground

[4] After all, "Philosophers are (usually) competent natural language speakers and especially keen to subtle differences in meaning." (Justus, 2012, p. 172)

the LLM through the prompt, by incorporating language provided by the conceptual engineer in the definition of the concept, and by the knowledge graph in the summary description of the entity presented during evaluation. This is the approach used in retrieval-augmented generation (Gao et al., 2023) and knowledge-graph-enhanced LLMs (Dai et al., 2024) to reduce hallucination and improve accuracy.

The question of the groundedness of LLMs is a fascinating one, but from the perspective of Nado's Practical Role Account, it is not clear that this question has any bearing on the utility of our approach:

> Though there is a fairly strong correlation between words and procedures, conceptual engineering isn't about what our words should mean, or even about how we should use our words. It is about *how we should classify*. . . . If we want our conceptual engineering interventions to affect how people infer and behave, then changing the meaning of a term seems a rather inefficient stratagem. Why not target the classificatory practice directly? (Nado, 2023b, p. 1993)

Our approach indeed targets the classificatory practice directly, and our experimental results show evidence of useful levels of performance.

6.4 Empirical methods

We assert that the evaluation procedure we have defined shows how a conceptual engineering project can incorporate an empirical, data-driven activity (Andow, 2020). Applying classification procedures to large numbers of positive and negative examples of a concept's extension can help conceptual engineers evaluate different definitions for a concept at a scale that "armchair-based conceptual engineering" (Landes, 2023) cannot. Rationales generated by classification procedures can help conceptual engineers refine their definitions. This raises the possibility that generative AI assistants (Weisz et al., 2023) could support philosophers in the conduct of conceptual engineering projects.

In addition, recent work on using LLMs as models of human linguistic behavior or judgment, and their use in simulating linguistic subpopulations (Aher et al., 2023; Argyle et al., 2023; Dillion et al., 2023; Horton, 2023; Simmons & Hare, 2023), further suggests that our proposed method could be combined with that work to yield a corpus method for experimental philosophy (Fischer & Sytsma, 2022; Sytsma, 2023).

6.5 The implementation problem

Cappelen (2018) and others have argued that conceptual engineering is difficult, as it is hard to see how the natural language (re)definition of a concept can be effectively adopted by a population of human speakers. This has come to be known as the implementation problem (Cappelen, 2018; Jorem, 2021). We assert that our

approach, used as a means for semantically aligning intensional knowledge expressed in natural language and extensional knowledge represented in a knowledge base (B. P. Allen & Groth, 2024), can play a practical role in providing a new set of success conditions for conceptual engineering (Andow, 2021; Pinder, 2022).

Knowledge bases such as Wikidata have an impact on society by virtue of their use in online search, discovery, and recommendation (Peng et al., 2023). Using classification procedures to evaluate and improve the alignment between natural language definitions of concepts and the representation of their extensions in knowledge graphs can be of practical value in knowledge graph refinement, which is the process of improving an existing knowledge graph by adding missing knowledge or identifying and removing errors (Paulheim, 2017). Engineering concepts represented in such resources using the above method can aid understanding within a specific linguistic subgroup, i.e., the users of applications built on top of such knowledge bases, as proposed in (Matsui, 2024). As an example use case closely related to the experiments described above, the Wikidata community is working to improve the modeling of gender in Wikidata (Wikidata, 2023); we hypothesize that our approach would be useful in efforts of that sort.

Related to the task of knowledge graph refinement are socially responsible data management (Stoyanovich et al., 2022) and data governance (Khatri & Brown, 2010). Khatri and Brown (2010) describe principles for data governance, touching on issues of the alignment of natural language concepts and their realization in databases. These concerns are echoed in the FAIR principles (Wilkinson et al., 2016), specifically with respect to the requirement for clear documentation of metadata that aligns natural language concepts and metadata in scientific data resources. More recently, Vogt et al. (2024) have proposed additional to the FAIR principles to specifically address the issue of semantic interoperability. Given the increasing use of knowledge graphs in scientific research and commercial applications, these principles are important to apply in the context of knowledge graph creation and refinement. We believe that our approach could be useful in this context as well.

7 Limitations

A limitation of our work is its reliance on a specific, proprietary LLM inference API (OpenAI, 2023), which raises transparency, reproducibility and safety concerns (Bender et al., 2021; Hu & Levy, 2023). Reproducing these experiments using other inference APIs, including ones based on open-source or open weight LLMs, would provide useful information with respect to the variation in performance due to the use of other LLMs. Recently, we have shown that our approach, applied to the task of knowledge graph refinement, has good performance across seven different LLMs (B. P. Allen & Groth, 2024).

Another limitation in our experiments is that error analysis was performed solely by the author. More reviewers, reviewing a larger set of examples and classifications,

would provide a stronger statistical estimate of the level of agreement between human evaluators, the classification procedure, and the knowledge graph.

Finally, we did not investigate the effect of two specific choices made in the prompt engineering of the classifier. First, the $P_{rationale\_generation}$ prompt explicitly provided instruction to ignore background information present in the training corpus for the LLM in considering the intensional definition, and second, the $P_{answer\_generation}$ prompt explicitly provided instruction intended to ensure that a binary classification was made. In the case of the latter, another implementation could instead use a ternary-valued logic, such as a weak Kleene logic (Beall, 2016; Ciuni & Carrara, 2019; Zamperlin, 2019), with an additional truth value of `undefined`. Ablation studies would provide insight into the validity of these two prompt design choices.

8 Conclusion

In this work, we have shown how to construct a conceptual engineering target as a computational artifact, and apply it to provide an empirical method for use in conceptual engineering projects. We view this as an initial step in an investigation of the potential utility of large language models in the practice of conceptual engineering.

Much has been written of late on the impact that large language models will on society. There is clearly much work to be done to address issues of their trustworthiness, safety, ethics, and environmental impact. That being said, we hope that the work here suggests that LLMs, through their use in the context of ameliorative and normative projects of conceptual engineering (Haslanger, 2000; Köhler & Veluwenkamp, 2024), can play a positive role in the future.

Acknowledgements

The author wishes to thank Paul Groth, Corey Harper, Filip Ilievski, Jürgen Lipps, and Lise Stork for useful conversations and suggestions with respect to the topics discussed above, and Nathaniel Gan and Nikhil Mahant for their thorough review and valuable feedback, which improved the manuscript.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *International Conference on Machine Learning*, 337–371.

Allen, B., Polat, F., & Groth, P. (2024). Shroom-indelab at semeval-2024 task 6: Zero- and few-shot llm-based classification for hallucination detection.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024). https://doi.org/10.48550/arXiv.2404.03732

Allen, B. P., & Groth, P. T. (2024). Evaluating class membership relations in knowledge graphs using large language models [To appear.]. *European Semantic Web Conference*.

Andow, J. (2020). Fully experimental conceptual engineering. *Inquiry*, 1–27. https://doi.org/https://doi.org/10.1080/0020174X.2020.1850339

Andow, J. (2021). Conceptual engineering is extremely unlikely to work. so what? *Inquiry*, *64*(1-2), 212–226.

Angles, R., Thakkar, H., & Tomaszuk, D. (2020). Mapping rdf databases to property graph databases. *IEEE Access*, *8*, 86091–86110.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Beall, J. (2016). Off-topic: A new interpretation of weak-kleene logic. *The Australasian Journal of Logic*, *13*(6).

Beek, W., Ilievski, F., Debattista, J., Schlobach, S., & Wielemaker, J. (2018). Literally better: Analyzing and improving the quality of literals. *Semantic Web*, *9*(1), 131–150.

Belleri, D. (2021). On pluralism and conceptual engineering: Introduction and overview. *Inquiry*, 1–19.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., Lehmann, T., Podstawski, M., Niewiadomski, H., Nyczyk, P., et al. (2023). Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Burgess, A., Cappelen, H., & Plunkett, D. (2020). *Conceptual engineering and conceptual ethics*. Oxford University Press.

Cappelen, H. (2018). *Fixing language: An essay on conceptual engineering*. Oxford University Press.

Carnap, R. (1955). Meaning and synonymy in natural languages. *Philosophical studies*, *6*, 33–47.

Chalmers, D. J. (2020). What is conceptual engineering and what should it be? *Inquiry*, 1–18.

Cifor, M., & Rawson, K. (2022). Mediating queer and trans pasts: The homosaurus as queer information activism. *Information, Communication & Society*, 1–18.

Ciuni, R., & Carrara, M. (2019). Semantical analysis of weak kleene logics. *Journal of Applied Non-Classical Logics*, *29*(1), 1–36.

Creswell, A., Shanahan, M., & Higgins, I. (2022). Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

Dai, X., Hua, Y., Wu, T., Sheng, Y., & Qi, G. (2024). Counter-intuitive: Large language models can better understand knowledge graphs than we thought. *arXiv preprint arXiv:2402.11541*.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can ai language models replace human participants? *Trends in Cognitive Sciences*.

Dürst, M., & Suignard, M. (2005). *Internationalized resource identifiers (iris)* (tech. rep.). RFC Editor.

Dutilh Novaes, C., & Reck, E. (2017). Carnapian explication, formalisms as cognitive tools, and the paradox of adequate formalization. *Synthese*, *194*, 195–215.

Etta Rudolph, R. (2021). Conceptual exploration. *Inquiry*, 1–26.

Fields, J., Chovanec, K., & Madiraju, P. (2024). A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*.

Fischer, E., & Sytsma, J. (2022). Projects and methods of experimental philosophy. *The Compact Compendium of Experimental Philosophy*, 39.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Haslanger, S. (2000). Gender and race:(what) are they? (what) do we want them to be? *Noûs*, *34*(1), 31–55.

Heersmink, R., de Rooij, B., Vázquez, M. J. C., & Colombo, M. (n.d.). A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness.

Heist, N., Hertling, S., Ringler, D., & Paulheim, H. (2020). Knowledge graphs on the web-an overview. *Knowledge Graphs for eXplainable Artificial Intelligence*, 3–22.

Hogan, A., Arenas, M., Mallea, A., & Polleres, A. (2014). Everything you always wanted to know about blank nodes. *Journal of Web Semantics*, *27*, 42–69.

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Comput. Surv.*, *54*(4). https://doi.org/10.1145/3447772

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (Tech. rep.). National Bureau of Economic Research.

Hu, J., & Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. *arXiv preprint arXiv:2305.13264*.

Huang, J., & Chang, K. C.-C. (2022). Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Isaac, M. G., Koch, S., & Nefdt, R. (2022). Conceptual engineering: A road map to practice. *Philosophy Compass*, *17*(10), e12879.

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38.

Jorem, S. (2021). Conceptual engineering and the implementation problem. *Inquiry*, *64*(1-2), 186–211.

Jorem, S., & Löhr, G. (2024). Inferentialist conceptual engineering. *Inquiry*, *67*(3), 932–953.

Justus, J. (2012). Carnap on concept determination: Methodology for philosophy of science. *European Journal for Philosophy of Science*, *2*, 161–179.

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, *53*(1), 148–152.

Koch, S. (2021). Engineering what? on concepts in conceptual engineering. *Synthese*, *199*(1), 1955–1975.

Koch, S., Löhr, G., & Pinder, M. (2023). Recent work in the theory of conceptual engineering. *Analysis*, 1–15.

Köhler, S., & Veluwenkamp, H. (2024). Conceptual engineering: For what matters. *Mind*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, *35*, 22199–22213.

Landes, E. (2023). Conceptual engineering should be empirical [Accessed: 2023-10-16].

Larsson, S. (2015). Formal semantics for perceptual classification. *Journal of logic and computation*, *25*(2), 335–369.

Lederman, H., & Mahowald, K. (2024). Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of llms. *arXiv preprint arXiv:2401.04854*.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. (2023). Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Mandelkern, M., & Linzen, T. (2024). Do language models' words refer? *arXiv preprint arXiv:2308.05576*.

Manne, K. (2017). *Down girl: The logic of misogyny*. Oxford University Press.

Marasović, A., Beltagy, I., Downey, D., & Peters, M. E. (2021). Few-shot self-rationalization with natural language prompts. *arXiv preprint arXiv:2111.08284*.

Matsui, T. (2024). Local conceptual engineering in a linguistic subgroup and the implementation problem [Preprint at https://philpapers.org/rec/MATLCE-8].

Miao, N., Teh, Y. W., & Rainforth, T. (2023). Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *arXiv preprint arXiv:2308.00436*.

Muskens, R. (2005). Sense and the computation of reference. *Linguistics and philosophy*, *28*, 473–504.

Nado, J. (2023a). Classification procedures as the targets of conceptual engineering. *Philosophy and Phenomenological Research*, *106*(1), 136–156. https://doi.org/https://doi.org/10.1111/phpr.12843

Nado, J. (2023b). Taking control: Conceptual engineering without (much) metasemantics. *Inquiry*, *66*(10), 1974–2000.

OpenAI. (2023). Gpt-4 technical report.

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, *8*(3), 489–508.

Peng, C., Xia, F., Naseriparsa, M., & Osborne, F. (2023). Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, 1–32.

Pinder, M. (2021). Conceptual engineering, metasemantic externalism and speaker-meaning. *Mind*, *130*(517), 141–163.

Pinder, M. (2022). Is haslanger's ameliorative project a successful conceptual engineering project? *Synthese*, *200*(4), 334.

Planet. (2006a). In *The IAU draft definition of "planet" and "plutons"*. International Astronomical Union. Retrieved October 16, 2023, from https://www.iau.org/news/pressreleases/detail/iau0601/

Planet. (2006b). In *Result of the IAU resolution votes*. International Astronomical Union. Retrieved October 16, 2023, from http://www.iau.org/static/archives/releases/doc/iau0603.doc

Planet. (2023). In *The Oxford English Dictionary*. Oxford University Press. Retrieved October 17, 2023, from https://www.oed.com/dictionary/planet_n

Podosky, P.-M. C. (2022). Can conceptual engineering actually promote social justice? *Synthese*, *200*(2), 160.

Sarma, R., Baruah, K., & Sarma, J. K. (2008). Iau planet definition: Some confusions and their modifications.

Si, C., Goyal, N., Wu, S. T., Zhao, C., Feng, S., Daumé III, H., & Boyd-Graber, J. (2023). Large language models help humans verify truthfulness–except when they are convincingly wrong. *arXiv preprint arXiv:2310.12558*.

Simmons, G., & Hare, C. (2023). Large language models as subpopulation representative models: A review. *arXiv preprint arXiv:2310.17888*.

Singh, C., Inala, J. P., Galley, M., Caruana, R., & Gao, J. (2024). Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & language*, *25*(4), 359–393.

Stoyanovich, J., Abiteboul, S., Howe, B., Jagadish, H., & Schelter, S. (2022). Responsible data management. *Communications of the ACM*, *65*(6), 64–74.

Sytsma, J. (2023). Ordinary meaning and consilience of evidence. *Advances in experimental philosophy of law*, 171.

Thomasson, A. L. (2020). A pragmatic method for normative conceptual work. *Conceptual engineering and conceptual ethics*, 435–458.

Vogt, L., Strömert, P., Matentzoglu, N., Karam, N., Konrad, M., Prinz, M., & Baum, R. (2024). Fair 2.0: Extending the fair guiding principles to address semantic interoperability. *arXiv preprint arXiv:2405.03345*.

Von Fintel, K., & Heim, I. (2021). Intensional semantics [MIT].

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Weisz, J. D., Muller, M., He, J., & Houde, S. (2023). Toward general design principles for generative ai applications. *arXiv preprint arXiv:2301.05578*.

Wikidata. (2023). "wikidata:wikiproject lgbt/gender" [Accessed: 2023-10-15].

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1–9.

Woman. (2023). In *The Oxford English Dictionary*. Oxford University Press. Retrieved October 17, 2023, from https://www.oed.com/dictionary/woman_n

Women. (2013). In *Homosaurus*. Homosaurus Editorial Board. Retrieved October 17, 2023, from https://homosaurus.org/v3/homoit0001509

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Ye, X., & Durrett, G. (2022). The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, *35*, 30378–30392.

Zamperlin, N. (2019). *Intensional kleene logics for vagueness* [Master's thesis, University of Amsterdam].