Knowledge-Enhanced Language Models Are Not Bias-Proof: Situated Knowledge and Epistemic Injustice in AI

Angelie Kraft
angeline.kraft@uni-hamburg.de
Universität Hamburg, Department of Informatics, Semantic Systems, Germany
Hamburg, Germany
Leuphana Universität Lüneburg, Institute for Information Systems, Artificial Intelligence and Explainability
Lüneburg, Germany

Eloïse Soulier
eoloise.soulier@uni-hamburg.de
Universität Hamburg, Department of Informatics, Ethics in Information Technology
Hamburg, Germany

ABSTRACT
The factual inaccuracies (“hallucinations”) of large language models have recently inspired more research on knowledge-enhanced language modeling approaches. These are often assumed to enhance the overall trustworthiness and objectivity of language models. Meanwhile, the issue of bias is usually only mentioned as a limitation of statistical representations. This dissociation of knowledge-enhancement and bias is in line with previous research on AI engineers’ assumptions about knowledge, which indicate that knowledge is commonly understood as objective and value-neutral by this community. We argue that claims and practices by actors of the field still reflect this underlying conception of knowledge. We contrast this assumption with literature from social and, in particular, feminist epistemology, which argues that the idea of a universal disembodied knower is blind to the reality of knowledge practices and seriously challenges claims of “objective” or “neutral” knowledge.

Knowledge enhancement techniques commonly use Wikidata and Wikipedia as their sources for knowledge, due to their large scales, public accessibility, and assumed trustworthiness. In this work, they serve as a case study for the influence of the social setting and the identity of knowers on epistemic processes. Indeed, the communities behind Wikidata and Wikipedia are known to be male-dominated and many instances of hostile behavior have been reported in the past decade. In effect, the contents of these knowledge bases are highly biased. It is therefore doubtful that these knowledge bases would contribute to bias reduction. In fact, our empirical evaluations of RoBERTa, KEPLER, and CoLAKE demonstrate that knowledge enhancement may not live up to the hopes of increased objectivity. In our study, the average probability for stereotypical associations was preserved on two out of three metrics and performance-related gender gaps on knowledge-driven task were also preserved.

We build on these results and critical literature to argue that the label of “knowledge” and the commonly held beliefs about it can obscure the harm that is still done to marginalized groups. Knowledge enhancement is at risk of perpetuating epistemic injustice, and AI engineers’ understanding of knowledge as objective per se conceals this injustice. Finally, to get closer to trustworthy language models, we need to rethink knowledge in AI and aim for an agenda of diversification and scrutiny from outgroup members.

CCS CONCEPTS
• Computing methodologies → Natural language generation; Reasoning about belief and knowledge; Philosophical/theoretical foundations of artificial intelligence.

KEYWORDS
natural language processing, language models, knowledge graphs, knowledge enhancement, bias, fairness, representation, epistemology, feminism

ACM Reference Format:

1 INTRODUCTION
One of the currently most discussed limitations of large language models (LLMs) is their tendency to produce false statements [35]. While LLMs are capable of generating text with great fidelity to information from external sources, like knowledge graphs [3, 77] they frequently produce errors by associating events with the wrong dates or fabricating claims about real people, for instance. Such errors can yield negative impacts on society. It can affect the integrity of science and education [68] or influence the outcomes of democratic elections, by producing false claims about political candidates [87] and thus misleading voters.

This lack of factual accuracy is commonly attributed to the implicitness with which knowledge is stored in language models (LMs) and has sparked new interest in ways to enhance LMs with explicit information from external sources, like knowledge graphs [3, 77].

2Factual inaccuracies or false statements produced by language models are often referred to as ‘hallucinations’. We reject this term as it falsely implies a similarity of such models to the human mind.
We particularly focus on Wikimedia Foundation’s knowledge bases with traditional theories of knowledge [1]. Drawing from previous discussions, we believe that it does not significantly affect our argument. Although the distinction between propositional knowledge and knowledge-how and its consequences for knowledge databases are certainly relevant to this discussion, it feeds into knowledge-related injustices [1].

A contribution of this interdisciplinary work is to illustrate some of the related discourse within philosophy and, on this basis, question the prevalent assumptions about knowledge in the AI community. We discuss how dominant conceptions may disguise biases, and, as a consequence, perpetuate injustices. By that, we aim to motivate a rethinking of knowledge as situated and to emphasize the necessity for diversification.

In Section 2, we discuss the evolution of the approach to knowledge from traditional (Western) epistemology to social and feminist epistemology. The latter coined the concept of situated knowledge [30], which emphasizes the importance of social situatedness to practices of knowledge. We compare this philosophical discussion to AI engineers’ conceptions of knowledge and argue that the pervasive understanding of knowledge as objective and value-neutral may disguise the power dynamics that structure knowledge production [1]. Publications about knowledge-enhanced language modeling usually mention the risk of bias as a distinguishing property of statistical representations [2, 3, 106], implying that explicit knowledge is not susceptible to bias. This depiction can be misleading: In Section 3, we discuss empirical evidence for biases of popular knowledge resources and knowledge-enhanced language models. We particularly focus on Wikimedia Foundation’s knowledge bases Wikipedia and Wikidata [100], which play a major role in language model training and knowledge enhancement and were shown to exhibit coverage gaps and stereotypical biases along different social dimensions [13, 90, 95]. We found that knowledge-enhanced language modeling on the basis of Wikidata preserves the biases of the original language model. We maintain that knowledge sources and knowledge-enhanced language models should not per se be expected to be less biased than other datasets and AI models. In Section 4, we argue that trusting ‘knowledge data’ more than other types of data may wrongfully disguise these issues and contributes to perpetuate the specific kind of injustice that Miranda Fricker has dubbed epistemic injustice [22], that is, a kind of injustice that harms us specifically as knowers. Including more diverse voices is not only a way to tackle these injustices but also the only way we may strive towards objectivity [31, 52].

2 ASSUMPTIONS ABOUT KNOWLEDGE IN AI

In this paper, we argue that AI engineers commonly assume knowledge to be subject-independent, which corresponds to more traditional philosophical theories of knowledge. To this end, we start by briefly sketching the evolution from traditional Western epistemology and the figure of the universal knower, to recent approaches from social and feminist epistemology, which emphasize the central role of the social situation of the knower. Finally, we detail how these philosophical theories map to the conceptions of knowledge held by AI engineers and presumably influence modern-day research and practices related to knowledge in AI.

2.1 Philosophical Roots of the “View from Nowhere” and Critique

2.1.1 Traditional Western Philosophy. The idea that knowledge could depend on the identity and social situation of the knower has only relatively recently been theorized in Western philosophy. Traditionally, Western epistemology has seen knowledge as a relationship between an individual knower and an object of knowledge, and concentrated its efforts on characterizing this relationship of knowledge, theorizing what distinguishes knowledge from non-knowledge. This distinction often has to do with justification: A belief or perception only becomes knowledge with proper justification. In fact, in analytic philosophy, knowledge is often defined as ‘justified true belief’ [94] and the justification problem phrased as ‘S knows that p when [relevant justification]’, where S is a single undetermined knower [1]. What constitutes proper justification is part of the philosophical debate, but justification is often considered valid only if internal. For example, Descartes considers knowledge coming from others as unreliable [14]. This is in line with the general representation in Western philosophy, usually associated with figures of the Enlightenment such as Kant, that mature thinking and knowing is about autonomy [40]. In this perspective, knowledge is acquired independently and rationally, it is universal, independent from the knower’s embodied identity, social situation and interests. In Sandra Harding’s (critical) words: “In order to achieve the status of knowledge, beliefs are supposed to break free of – to transcend – their original ties to local, historical interests, values, and agendas” [31, p. 438].

2.1.2 Feminist and Social Epistemology. In the last decades, feminist and social epistemology have challenged this traditional approach to knowledge, arguing that knowers are always socially situated, and that this social situation mattered to the kind of knowledge they could produce. Social epistemologists have emphasized that the production of knowledge is an inescapably social activity [53]. In John Hardwig’s terms, we are epistemically dependent:

---

1Here objective is understood as subject-independent. The remaining of the paper elaborate on the necessity to challenge this understanding of objectivity.

2In using the term ‘knowledge’ throughout this article, we are aware of the abysmal amount of ink that has been spilled over this term, and of the differences that exist between disciplines and within epistemology as to what it encompasses. We understand knowledge here as content - not as a cognitive state - and as propositional. Although the distinction between propositional knowledge and knowledge-how and its consequences for knowledge databases are certainly relevant to this discussion, they are out of the scope of this article. We do not consider it crucial either in the context of this paper to draw a distinction between scientific and common knowledge, as we believe that it does not significantly affect our argument.

3Characterizing and summarizing “traditional Western philosophy” in one paragraph is a difficult task, considering that what is usually referred to as “Western thought” is itself a Western post hoc construction. What we mean here is a conceptual framework considered to have crystallized during the Enlightenment, which has been significantly challenged in the last half century by critical theories.
We detail in Section 4 the ways in which this can lead to injustices. The “view from nowhere” supposed to characterize objectivity, AI technology, we instead use the term “AI engineer” for the sake of consistency.

Feminist standpoint theorists have argued that we are limited in what we can know by our social situation, and “some social situations – critically unexamined dominant ones – are more limiting than others in this respect” [31, p. 443]. In other words, we are particularly constrained in what we are able to know when our social situation is dominant, and therefore seldom questioned. The “view from nowhere” [72] supposed to characterize objectivity, in Haraway’s words, actually “signifies the unmarked positions of Man and White” [30, p. 581]. Different feminist approaches disagree on the extent to which we are epistemically limited by our social situation, and the depth to which scientific frameworks should be questioned. We leave the detail of these discussions out of this short account, as we do not believe it is necessary to take sides in order to draw from these different theorists for the problem at hand. Note that related arguments have been made by decolonial epistemologists: These scholars have emphasized the geopolitical situation of knowledge under the persistent regime of coloniality [80], and the necessity for subjects of colonial oppression to think not only from their perspective, but outside of Western epistemic resources [28, 80]. We give this account of the evolution of the field of epistemology, as we consider it reasonable to assume that the influence of modern epistemology still has a bearing on contemporary conceptions of knowledge. In the following we focus on the group of AI engineers, as they are the relevant category to the object of this article, but we do not believe these representations to be limited to this group.

2.2 AI Engineers and the “View From Nowhere”
2.2.1 Forsythe’s Anthropological Study. Three decades ago, in 1993, Diana E. Forsythe published one of the first in-depth investigations of AI engineers’ conceptions of knowledge [21]. She had observed and interviewed a group of engineers whose task it was to elicit the knowledge of domain experts and translate it into a machine-readable representation for use in AI systems. Back then, it was already envisioned that AI would at some point “duplicate human expertise” [21, p. 1], i.e., that AI systems would gain the same capabilities that humans have. Without more critical scrutiny of what constitutes knowledge, the AI engineers in Forsythe’s study described it as universal, a constant that does not change with context, is purely cognitive and conscious in nature. Forsythe [21] also mentions the ways in which AI engineers’ assumptions differ from those held by social scientists. The latter believe knowledge to be a problematic subject of research that is highly dependent on social and otherwise contextual factors. They consider a lot of

capacity Descartes’s ideal of the independent knower, we cannot but rely on others’ testimony to know most of what we know, even in scientific contexts where the standards on what counts as knowledge are taken to be higher [32]. If knowledge necessarily involves relying on other’s testimony, then power dynamics within society are relevant to the production and dissemination of knowledge [22] and to the possibility to accept a claim as knowledge [88]. Indeed, these power dynamics determine whose knowledge will be heard. We detail in Section 4 the ways in which this can lead to injustices.

Feminist standpoint theorists have argued that we are limited in what we can know by our social situation, and “some social situations – critically unexamined dominant ones – are more limiting than others in this respect” [31, p. 443]. In other words, we are particularly constrained in what we are able to know when our social situation is dominant, and therefore seldom questioned. The “view from nowhere” [72] supposed to characterize objectivity, in Haraway’s words, actually “signifies the unmarked positions of Man and White” [30, p. 581]. Different feminist approaches disagree on the extent to which we are epistemically limited by our social situation, and the depth to which scientific frameworks should be questioned. We leave the detail of these discussions out of this short account, as we do not believe it is necessary to take sides in order to draw from these different theorists for the problem at hand. Note that related arguments have been made by decolonial epistemologists: These scholars have emphasized the geopolitical situation of knowledge under the persistent regime of coloniality [80], and the necessity for subjects of colonial oppression to think not only from their perspective, but outside of Western epistemic resources [28, 80]. We give this account of the evolution of the field of epistemology, as we consider it reasonable to assume that the influence of modern epistemology still has a bearing on contemporary conceptions of knowledge. In the following we focus on the group of AI engineers, as they are the relevant category to the object of this article, but we do not believe these representations to be limited to this group.

2.2.2 Adam’s Epistemological Analysis. In “Deleting the Subject: A Feminist Reading of Epistemology in Artificial Intelligence”, Alison Adam [1] compares AI engineers’ beliefs to the traditional Western take on knowledge (see Section 2.1.1). She points out that AI systems are built on the assumption of knowledge as a universal “view from nowhere” (as introduced by Nagel [72]) and thereby dismiss the importance of the identity of the knower. She argues that this effectively obscures an “implicit hierarchy of knowers”, i.e., the power dynamics which grant a specific demographic the privilege to represent its knowledge in AI systems and others not. Following an analysis of the Cyc commonsense knowledge base, Adam [1] formulates two main points of criticism: Firstly, the system did not allow to represent contradictory information and, thus, could only represent one world view at a time. She explains this with the presumably pervasive understanding of AI engineers “that there is an independent world that can be accessed through perception and also that everyone will agree on what the real world is like” [1, p. 241]. Again, this understanding disregards that individual knowers are limited in how they view the world (by their identity and situation), which means that different perceptions of the world co-exist. Her second point of criticism relates to the underlying hierarchy of knowers: Ultimately, whose knowledge would be considered the right one was determined only by the developers of Cyc, whose demographic was described as the “middle-class, Western, professional man” [1, p. 241]. Again, including their knowledge exclusively in a system like Cyc is to certify it as more legitimate than other knowledges.11

2.2.3 Understanding Modern Conceptions. The dominance of the “view from nowhere” and its harmful consequences are still frequently discussed in the context of modern Machine Learning and AI [26, 29, 42, 49]. The current discourse on the capabilities of AI indicate that engineers pre-dominantly focus on the technical challenges of knowledge extraction from data [57], benchmarking the knowledge of AI models [21] [39, 86, 107], and ways to embed more of it [77]. Yet, the provenance of this knowledge remains largely

For a detail of the different approaches in feminist philosophy of science, see e.g. [5] Forsythe [21] uses the term “knowledge engineer” to designate the participants’ profession. However, as they are described as researching and developing (symbolic) AI technology, we instead use the term “AI engineer” for the sake of consistency.

Knowledge regarding everyday situations and cause-effect relationships.
11https://cyc.com/
12“Artificial intelligence” has been, ever since the expression appeared in the 50s, associated with an anthropomorphic aim to replicate human capabilities. Even though the term is currently often associated with strictly technical definitions (for example, the definition that will most likely figure in the upcoming European AI Act: https://data.consilium.europa.eu/doc/document/ST-14954-2022-BIT/en/pdf), it remains a common way of understanding “artificial intelligence”. In the Google campaign, their Knowledge Graph was seen as a step towards “building the next generation of search, which […] understands the world a bit more like people do” (https://blog.google/products/search/introducing-knowledge-graph-things-not/). With the recent development of sophisticated AI systems, researchers in the philosophy of AI have been inquiring the ways in which concepts so far exclusively applied to humans and some other animals could be extended to AIs in a non-metaphorical sense. These reflections include whether an AI can “know” [11], or “believe” [85] but also “love” [75] or exert “agency” [19]. We are not concerned with these questions in this article. When we talk about what a LM knows, we mean metaphorically which – and importantly whose – knowledge it embeds, not in which sense it might be said to know something itself. This is not to say that this question is irrelevant to our main concern, as it seems possible that anthropomorphizing the AI itself might further contribute to the
unexamined. In a review on AI throughout history, Jiang et al. [36] claim that “knowledge describes regular patterns and abstract facts that human understands [sic]” [p. 9] and thereby attribute universality to knowledge. The authors continue by stating that, “therefore, it is usually semantic and embedded in books and research articles. To be interpretable and useful for machines, it needs to be modelled, transformed, and generated” [36, p. 9]. This quote refers to automated knowledge acquisition approaches, which are widely established. It points to an understanding of knowledge as subject-independent and is similar to the beliefs held by Forsythe’s participants, who had desired exactly this kind of automation to avoid having “to mine those jewels of knowledge out of their heads one by one” [21, p. 454]. In his vision paper, Marcus [56] argues that the next decade in AI should focus on “a hybrid, knowledge-driven, reasoning-based approach, centered around cognitive models, that could provide the substrate for a richer, more robust AI than is currently possible” [p. 1]. Without addressing the social conditions under which knowledge resources are created, he claims that having more of it embedded in AI models will make these models more robust. LeCun predicts that AI will become a “repository of all human knowledge”, claiming that such a repository would be the “ultimate solution to the problem of misinformation.” He, however, emphasizes that automation alone will not suffice and instead proposes Wikipedia-style crowd-sourcing, implying that the more people contribute, the closer we will get to a representation of the sum of all knowledge. As we will discuss in more detail in Section 3.3, Wikipedia, in fact, clearly exemplifies that crowd-sourcing processes are not immune to the influence of social power structures without appropriate countermeasures. While we agree on the importance of improving the factual accuracy of AI systems and on the value of crowd-sourcing as a basis for this, we believe that a more nuanced understanding of knowledge is needed to come closer to just and objective knowledge production in the long term.

3 CONNECTING THE DEBATES ON KNOWLEDGE ENHANCEMENT AND SOCIAL BIAS

In the following, we take a closer look at the bias issue in Wiki-media knowledge bases to exemplify the influence of the social setting on collective epistemic processes. To this end, we firstly explain the idea behind knowledge-enhanced language models. We then develop the connection between knowledge enhancement and social bias and later detail the representation issues in Wikimeda knowledge bases. Finally, we demonstrate how the biases of said knowledge bases can be adopted by technology. We do this at the example of language models enhanced with knowledge from Wikidata.

3.1 Knowledge Enhancement and the Dichotomy of Explicit and Implicit Knowledge in AI

Hybrid AI systems or knowledge-enhanced models are attempts to combine the strengths of statistical AI and explicit representations of knowledge. Statistical AI subsumes approaches that model patterns and rules implicitly from (large-scale) data sets, instead of following hard-coded rules. Such approaches allow to process enormous amounts of information with minimal human involvement (compared to mostly manually created symbolic systems) and are more generalizable to new areas and tasks [77]. Statistical AI is the currently dominating paradigm and AI-based language models are part of this category [38]. One limitation of these approaches is that the knowledge represented can no longer be accessed directly and can only be interpreted and quantified through dedicated decoding procedures [79, 107].

The effort to represent explicit knowledge content in machine- and human-readable form and perform inference based on hard-coded rules is commonly denoted symbolic AI, which was the most prominent AI paradigm for most of the second half of the 20th century. Knowledge graphs (KGs) are a type of symbolic representation that are still used to represent the semantic relationships between things in the world across various topical domains. A KG is a graph where each node represents a real-world entity, and edges represent the relationship between the entities. A KG-based system allows the efficient processing of knowledge in a machine-readable, and transparent.

Since statistical LMs always output the most likely next word, they may generate results that seem linguistically sound, even when the content is not accurate or appropriate [35]. This phenomenon is frequently observed, since the large-scale web-scraped datasets that LMs are trained on usually contain false information, inaccuracies, and gaps. In other cases, the perceived input may lack important contextual information for the model to produce contextually accurate results. To tackle this shortcoming, explicit, relevant, fine-grained knowledge can be incorporated [3]. A large variety of knowledge enhancement approaches exist to implement this idea. For example, the mention of an entity (a person, a place, an event, etc.) may be combined with additional background information during model training, so that an enriched representation of the entity is learned [97, 104]. Another common approach is to give the model access to an external knowledge base to retrieve relevant information from during runtime [46].

3.2 Why We Need to Talk About Knowledge Enhancement and Social Bias

Social bias is observed when language models “systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others” [23, p. 332]. It takes form in reproduced stereotypes [71], negative valuations of groups [91], or systematic performance differences based on sensitive attributes [15, 43]. Social bias is another widely discussed limitation of language models [8, 48, 96]. Both social bias and factual inaccuracies are considered obstacles to the trustworthiness of LMs [55, 101] but are usually investigated in isolation to each other. Factual inaccuracies
are countered by adding knowledge, i.e., data that represent facts about things in the world, while social bias is tackled, e.g., through data balancing, manipulation of the embedding space, or constraining the factual accuracy of LMs through knowledge enhancement could positively impact bias issues in the same instance, since knowledge is highly trusted and curated.16,17 This corresponds to our observation that, in the context of knowledge-enhanced language modeling, the issue of bias is usually only mentioned as a limitation of statistical AI and its unstructured training databases [2, 3, 106].18 The fact that highly curated and structured KGs, like Wikidata and DBpedia, reproduce the same societal biases mostly goes unmentioned [44]. This omission is unjustified and potentially harmful. That is, misconceiving of knowledge as objective and an antithesis to bias, value judgements, and uncertainty, grants anything under the label of knowledge potentially undeserved legitimacy. In fact, it gives undeserved legitimacy to the interests, assumptions and world views of a privileged group. In the case of both the work of Adam [1] and the KGs discussed here, this is predominantly the group of educated Western men [44].

In the next section, we summarize representation-related issues in Wikidata and Wikipedia, which are examples of crowd-sourced knowledge bases. As mentioned before, the creation or extension of knowledge graphs is also oftentimes based on or supported by automated processing [89], e.g., through automatic knowledge extraction [57] and knowledge integration [69]. Other works are even inspecting the possibility to extract knowledge directly from language models to utilize them as knowledge bases [79]. It is important to remember here that automatic approaches of course also mirror the values of their developers. Firstly, many of these mentioned natural language processing (NLP) approaches are affected by social biases [16, 25, 41, 61, 67]. Secondly, they are more frequently applied for the more represented languages. For instance, more bots are used to populate Germany-related content in Wikidata than Vietnam-related content [54], further amplifying existing coverage gaps. So, while the automatic creation and extension of knowledge bases may save a lot of time and effort (and avoid potential frustrations caused by social interactions [21]), they may amplify biases and further occlude the social conditions of knowledge production.

3.3 The Biases of Wikidata and its Hierarchy of Knowers

Most research articles that present new techniques for KG-based enhancement of language models utilize English Wikidata [e.g., 81, 97, 102, 103, 109], since it is the largest publicly accessible open-domain KG [104]. A wide range of non-KG approaches are developed on the basis of Wikipedia, e.g., many Retrieval-Augmented Generation (RAG) approaches [24, for an overview]. These knowledge bases19 are more curated and reviewed than most other data sources involved in the training of language models.20 That is, users populate the knowledge bases collaboratively, engage in discussions on the content, and constantly work on updates and refinement. Agarwal et al. [2] imply that KGs have less limited coverage of the world knowledge than text corpora. The authors used a dedicated data-to-text model to verbalize all triples in the English Wikidata KG and thereby created a synthetic natural-language corpus called the Kelm corpus (Corpus for Knowledge-Enhanced Language Model Pre-training) which is intended for integration with natural language training datasets to improve LM performance on knowledge-intensive tasks. In a blog post, the authors claim that “KGs are factual in nature because the information is usually extracted from more trusted sources, and post-processing filters and human editors ensure inappropriate and incorrect content are removed.”21 These claims strike us as particularly interesting in the face of prevalent issues with Wikimedia’s knowledge bases: Wikidata exhibits significant coverage gaps for different genders [13, 108], races, and citizenships [90]. We analyzed Wikidata and the Kelm corpus and found that women make up only approximately 20% and other genders make up less than 1% (see Table 3 in Appendix A). Representational biases are not only manifested in coverage gaps: Wikidata entries about German personalities are significantly more often edited than entries about Vietnamese personalities [54]. This indicates that the latter undergo less deliberation and may be less trustworthy [98, 22]. The narration style used to describe different demographics also differs in stereotypical ways. For example, on Wikipedia, women are more likely to be described with regards to personal life events (even within the “Career” section) than men [95]. Popular KGs like Wikidata use inappropriate and derogatory terms to indicate, e.g., ethnicity, sexual identity or orientation [74]. The cause of these representation issues can be found in the power hierarchies that characterize the community behind these efforts. Menking and Rosenberg [65] argue that there is a mismatch between the ideal scenario implied by the Five Pillars of Wikipedia, i.e., the guiding principles, and the reality of its epistemic community. “While anyone can edit Wikipedia, there are several barriers to becoming a Wikipedian. For example, newcomers must learn how to navigate any number of technical, organizational, and social hurdles they encounter when performing a substantial edit.” [65, p. 458]. Examples for said social hurdles are manifold: Members of marginalized communities face higher standards for notability, which is

16https://blog.research.google/2021/05/kelm-integrating-knowledge-graphs-with.html
17https://www.searchenginejournal.com/google-kelm-408151/
18We found one exception in Lewis et al. [46, p. 10], where it is stated that “Wikipedia, or any potential external knowledge source, will probably never be entirely factual and completely devoid of bias […]” and that “[i]n order to mitigate these risks, AI systems could be employed to fight against misleading content […]”. This suggestion fails to address the real-world source of the problem and instead points in the direction of techno-solutionism [70].
an eligibility requirement for coverage in Wikipedia and Wikidata [99]. Women editors’ articles are more likely to be reverted, especially in the early phases of their participation [45, 50]. Editors who identify as women and/or LGBTQIA+ are trolled, harassed, and receive death threats, and become victims of doxxing [63, 64]. Thus, it is not surprising that only 13% of all active Wikimedia editors are women and 4% gender-diverse, according to a 2023 report. The same report also showed that active editors are highly educated – 82% hold at least a post-secondary degree – and most US and UK editors are white (disproportionately more than in the general population). The geographic distribution of editors is skewed towards Western Europe, making up more than 50% (as of 2018).[26] These observations show how knowledge production is shaped by the situation of the knowers. Their identities and values influence the interactions leading to agreement (or disagreement) on what to consider knowledge. We focused on Wikipedia and Wikidata because they are prevalent resources in NLP research and a lot is known about the communities behind them. However, our criticism extends to other knowledge bases, like DBpedia and Freebase, which exhibit similar gaps.[44].

3.4 Knowledge Enhancement Does Not Solve the Bias Issue

Quantitatively, the effect of knowledge enhancement on bias was so far only shown for commonsense knowledge: Melotte et al. [62] fine-tuned different generative language models – GPT-2 [82], T5-base, and T5-small [83] – with commonsense KGs – Wikidata-CS [34] and ConceptNet [93] – to allow the models to predict an object from a given subject-predicate pair (e.g., “gentleman”, “is capable of”). The authors measured bias regarding origin, gender, religion, and profession via classifiers for sentiment and regard, which can identify whether or not an output sequence is a positive or negative portrayal. T5-small tuned on ConceptNet created more-than-average negative depictions of, e.g., “Cuban”, “Afghans”, and “Indians”. Occupations like “teacher”, “doctor”, and “professor”, were more likely depicted in positive ways, whereas “prosecutors” were more often depicted negatively. The results showed an increase of bias with the scale of the KG.

In the following, we present a preliminary analysis of social bias in language models enhanced with encyclopedic knowledge. We evaluated KEPLER (Knowledge Embedding and Pre-trained Language Representation) [104] and CoLAKE (Contextualized Language and Knowledge Embedding) [97] in comparison to RoBERTa (Robustly Optimized BERT Pretraining Approach) [51]. KEPLER and CoLAKE are both modified versions of the popular RoBERTa language model and incorporate Wikidata. More detailed explanations of these models are provided in Appendix B. To validate the knowledge enhancement effect, we compared the performance of the models on a suite of knowledge-intensive evaluation tasks, called the LAMA (LAnguage Model Analysis) probe [79], and present the results and more details on the probe in Appendix C. We investigated two kinds of bias: stereotypes, i.e., learned systematic associations between individuals/groups and classes of professions or other attributes, and secondly, performance differences on knowledge-related tasks that might arise from imbalanced representation of individuals or groups in the dataset.

3.4.1 Stereotypical Bias Analysis. We use three common stereotype measures to compare the biases across models: 1. SEAT (Sentence Embedding Association Test) [12, 59] measures the associations between certain demographics and certain attributes, which are often discussed in stereotypical portrayals of said demographics and their respective opposites. The significance of the association is determined via a permutation test and its effect size is interpreted as an indicator of the bias magnitude. Lower effect sizes indicate less bias. 2. CrowS-Pairs (Crowdsourced Stereotype Pairs) [73] is comprised of crowd-sourced stereotypical descriptions of historically disadvantaged groups in the United States. The test computes the percentage of instances where a stereotypical description is preferred over a less or non-stereotypical description by a given LM. For a random score of 50%, no systematic association is observed and the model is considered unbiased. 3. StereoSet follows a similar idea [71] and compares the likelihood of stereotypical, anti-stereotypical, and unrelated responses (example: “Girls tend to be more ___ than boys”; response options: “soft” (stereotypical), “determined” (anti-stereotypical), and “fish” (unrelated)). The idealized context association score (ICAT) is a stereotype metric based on the relative number of samples for which the stereotype is preferred over the anti-stereotypical option, scaled by the model’s language modeling capability (percentage of cases, where the model does not opt for the unrelated response).[30]

Table 1 shows the final bias metrics for all three models. On the SEAT metric, KEPLER and CoLAKE yield larger effect sizes than RoBERTa on two out of three bias dimensions, namely race and religion. On the gender bias dimension, CoLAKE outperforms RoBERTa by a large margin, causing CoLAKE to receive the best average score. For CrowS, the models again exhibit different strengths: While RoBERTa is least biased regarding race/color, nationality, age, and physical appearance, KEPLER and CoLAKE exhibit less stereotypical attributions in the case of other dimensions, like gender, religion, sexual orientation, and disability. On average, across all dimensions, all models prefer the stereotypical over the anti-stereotypical option in 58% of the cases. On StereoSet (ICAT), RoBERTa slightly outperforms the knowledge-enhanced models. In conclusion, these inconsistent results indicate that simply adding knowledge to language models does not solve the bias problem. Instead, two of the metrics used, CrowS-Pairs and StereoSet, indicate a preservation of the average probability for stereotypical associations.

3.4.2 Performance Bias Analysis. To investigate the models’ biases on a knowledge-intensive task, we performed a disaggregated
Table 1: Bias metrics for RoBERTa and its knowledge-enhanced variants KEPLER and CoLAKE. Bold scores indicate the most optimal model according to the respective metric. For SEAT, scores closer to 0 are less biased. For CrowS-Pairs, scores closer to 50 are more optimal and for StereoSet, ideal scores are ICAT=100.

<table>
<thead>
<tr>
<th></th>
<th>RoBERTa</th>
<th>KEPLER</th>
<th>CoLAKE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SEAT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td>.940</td>
<td>.789</td>
<td>.329</td>
</tr>
<tr>
<td>race</td>
<td>.307</td>
<td>.374</td>
<td>.340</td>
</tr>
<tr>
<td>religion</td>
<td>.127</td>
<td>.890</td>
<td>.332</td>
</tr>
<tr>
<td>average</td>
<td>.458</td>
<td>.684</td>
<td>.334</td>
</tr>
<tr>
<td><strong>CrowS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender/gender identity</td>
<td>60.15</td>
<td>59.39</td>
<td>54.41</td>
</tr>
<tr>
<td>race/color</td>
<td>63.57</td>
<td>64.92</td>
<td>63.53</td>
</tr>
<tr>
<td>religion</td>
<td>60.00</td>
<td>50.48</td>
<td>58.10</td>
</tr>
<tr>
<td>socioeconomic status/occupation</td>
<td>49.78</td>
<td>48.76</td>
<td>44.03</td>
</tr>
<tr>
<td>age</td>
<td>49.43</td>
<td>51.87</td>
<td>55.61</td>
</tr>
<tr>
<td>sexual orientation</td>
<td>61.10</td>
<td>59.52</td>
<td>61.90</td>
</tr>
<tr>
<td>physical appearance</td>
<td>60.83</td>
<td>57.59</td>
<td>55.56</td>
</tr>
<tr>
<td>disability</td>
<td>67.80</td>
<td>71.19</td>
<td>66.10</td>
</tr>
<tr>
<td>average</td>
<td>58.65</td>
<td>58.17</td>
<td>58.50</td>
</tr>
<tr>
<td><strong>StereoSet (ICAT)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td>60.48</td>
<td>68.63</td>
<td>70.43</td>
</tr>
<tr>
<td>race</td>
<td>68.93</td>
<td>63.96</td>
<td>65.09</td>
</tr>
<tr>
<td>religion</td>
<td>62.89</td>
<td>68.25</td>
<td>69.81</td>
</tr>
<tr>
<td>profession</td>
<td>67.42</td>
<td>66.06</td>
<td>66.49</td>
</tr>
<tr>
<td>overall</td>
<td>67.11</td>
<td>65.30</td>
<td>66.45</td>
</tr>
</tbody>
</table>

4 HOW CAN WE DO BETTER? DRAWING FROM PHILOSOPHICAL INSIGHTS

We used the example of Wikidata because it is a very popular database. Therefore, the biases described should be alarming in themselves. However, we do not expect these issues to be specific to Wikidata. As we have argued in Section 2, the conception of knowledge that seems to prevail in the AI community has been the object of philosophical reappraisal. Thus, we consider it fruitful to draw from feminist epistemology to better grasp the ways in which the social dimension of knowledge production in general can lead to injustices, but also how we can strive for better practices.

4.1 Including More Diverse Voices

The main insight we draw from feminist epistemology is that knowledge production is not immune to the power dynamics that structure society. This is what Miranda Fricker has famously theorized in her 2007 book “Epistemic Injustice, Power and the Ethics of Knowing” [22]. The fact that we are, as knowers, social beings that stand in power relations to each other, Fricker argues, makes knowledge practices the locus of a specific type of injustice: epistemic injustices. Fricker describes epistemic injustice as having two main types:
aspects: testimonial injustice and hermeneutical injustice. Testimonial injustice is a consequence of identity prejudice: We usually assign credibility automatically to speakers, and in this unreflective process, identity prejudice can unjustly lead us to grant less credibility to some speakers, typically from marginalized groups. Their contribution is dismissed, and they are harmed in their dignity and their capacity to participate in knowledge production and transmission. Hermeneutical injustice has to do with knowledge gaps: Because marginalized groups are less given the ability to participate in knowledge production, because their experiences are less the object of collective interest and study, their experiences and knowledge are not represented in our collective hermeneutical resources. Fricker gives the example of the concept of "sexual harassment", the absence of which long prevented some women from making sense of what they were experiencing. This understanding of hermeneutical injustice has however been nuanced among others by Rebecca Mason [58]. To Mason, hermeneutical injustice is not only a matter of marginalized groups not having the hermeneutical resources to articulate their experience, but also of dominant groups willfully, or at least blameworthy ignoring this experience. Dominant groups bear an important responsibility for these "blanks where there should be a name for an experience" [22, p. 160].

The mechanisms of exclusion from the Wikimedia community described in Section 3.3 are arguably examples of testimonial injustice contributing to hermeneutical injustice. Some contributors’ testimony is dismissed because of identity prejudice, and this results in gaps in the knowledge resource. As we have shown in Section 3.4, feeding such knowledge databases to LMs does not make them objective, but instead embeds these hermeneutical gaps in the technology. Epistemic injustices have to do with the possibility to participate in knowledge production and to be represented in collective resources. Working against these injustices is important to justice and non-discrimination, but it is also crucial for epistemic resources. Fricker gives the example of the concept of "sexual harassment", the absence of which long prevented some women from making sense of what they were experiencing. This understanding of hermeneutical injustice has however been nuanced among others by Rebecca Mason [58]. To Mason, hermeneutical injustice is not only a matter of marginalized groups not having the hermeneutical resources to articulate their experience, but also of dominant groups willfully, or at least blameworthy ignoring this experience. Dominant groups bear an important responsibility for these "blanks where there should be a name for an experience" [22, p. 160].

The mechanisms of exclusion from the Wikimedia community described in Section 3.3 are arguably examples of testimonial injustice contributing to hermeneutical injustice. Some contributors’ testimony is dismissed because of identity prejudice, and this results in gaps in the knowledge resource. As we have shown in Section 3.4, feeding such knowledge databases to LMs does not make them objective, but instead embeds these hermeneutical gaps in the technology. Epistemic injustices have to do with the possibility to participate in knowledge production and to be represented in collective resources. Working against these injustices is important to justice and non-discrimination, but it is also crucial for epistemic reasons. However strong a stance one takes on the way our situatedness epistemically limits us, it remains that our knowledge resources are enriched by including diverse contributions, particularly from marginalized groups. This is arguably not the case – yet – for Wikidata or Wikipedia.

Networks like Art+Feminism and FemNetz provide safe spaces for Wikipedia contributors with feminist visions. They organize regular events, e.g. edit-a-thons, to improve the platform’s coverage of knowledge relevant to all genders and increase the use of inclusive and anti-discriminatory language. These initiatives exemplify how epistemic injustice may be tackled bottom-up. However, against the backdrop of a community dominated by groups who resist the inclusion of certain experiences by violent means, participation can only be realized at high cost [63] or sometimes not at all. The founders of the German web encyclopedia Equalpedia initially raised public funds to build an editorial team that would contribute information about women and persons from the LGBTQIA+ community to Wikipedia. But, targeted by edit wars, they ultimately failed to prevail against the existing power structures and resorted to building their own platform instead. While institutions and individuals developing AI and respective data corpora should work towards solutions and pro-actively invite underrepresented views, the involvement of diverse voices should be approached in reciprocal and empowering ways [9]. The reality of modern AI is largely determined by powerful technology companies that gather information without consent to their own financial benefit. Especially historically exploited communities should (co-)determine how these resources are created, disseminated, and utilized [9]. Hence, refusal of participation in open access knowledge bases, like Wikipedia, is a legitimate alternative that should be supported, as well. Inclusion should always be approached with the perspective that hermeneutical injustice does not result from innocent knowledge gaps, but is motivated by group interest as an integral part of a pervasive system of social oppression [66]. Power dynamics shape discourses and practices of inclusion themselves [33], and we believe that inclusion should be approached critically, and not as the ultimate fix to structural injustice [10].

4.2 Reflexivity and Intersubjective Criticism: Objectivity Is Hard Work

Underlying this discussion is the question of whether there can be such a thing as knowledge that would be perspective-independent, and how we can strive for that or towards that goal. Viewpoints within feminist epistemology differ on this matter. However, we believe it is possible to draw some common lessons from them that are useful for AI engineers.

Feminist empiricists like Helen Longino or Elizabeth Anderson have argued that it is inevitable that moral and political values play a role in scientific inquiry [4, 52]. They play a role in determining what will be researched, but also with which methods. They influence according to which background theory facts will be interpreted and which facts will be considered significant. What still protects scientific knowledge from arbitrariness and preserves the possibility for objectivity – at least as a horizon – is, Longino says, the possibility for intersubjective criticism of commonly available phenomena and methodologies [52]. This supposes among others avenues for criticism, shared standards on the formulation of these criticisms, responsiveness to criticism, and equal intellectual authority among qualified practitioners. And the greater the number of points of view, the closer scientific practice gets to objectivity. In this sense, we consider interdisciplinary exchange and collaboration essential for critical decentering. In the case we have been discussing, engaging with different disciplines – for example during education [84] – and communities should contribute to fostering a more critical understanding of the concept of knowledge in the AI community.

Standpoint theorists share the conviction that beliefs and values are pervasive in every aspect of knowledge production. However, to them, there is no transcending our situatedness. Instead, it is precisely by theorizing this situatedness of subjects of knowledge and the values that underlie any knowledge-seeking endeavor that we can strive for what Harding calls "strong objectivity", a way

33https://artandfeminism.org/
35https://www.equalpedia.org/ueber-equalpedia/
to "maximize objectivity" through "strong reflexivity" [31, pp. 460-462]. This requires to think broader than the avenues for criticism organized by scientific communities (or any community that claims to create knowledge of some authority, for example a knowledge database). Indeed, the criteria that determine who is qualified to participate and according to which rules, should themselves be subjects of critical scrutiny. And those who are excluded from these groups are better situated to exercise this scrutiny. The consequence is that any claim to produce authoritative knowledge such as knowledge databases should not only imply organized practices of intersubjective criticism, but also actively seek the critical scrutiny of outgroup members.

In the absence of such strong standards, Harding calls objectivity a "mystifying notion", little more than an argument from authority that benefits dominant groups [31]. This article argues that in the same way, the term "knowledge" in the context of AI runs the risk of not being more than a mystification, if we do not strive for standards and practices that enable the resources in question to come closer to the ideal of objectivity associated with knowledge. Besides aforementioned efforts to facilitate more diverse contribution, we also need transparent documentation practices that allow scrutiny of knowledge bases and their original knowers [7, 27]. Institution-alizing (participatory) data collection through dedicated consortia to structure outreach to underrepresented groups as well as support and give visibility to their own initiatives are also important directions to consider [37].

5 CONCLUSION
Debates on the factual inaccuracy of language models and knowledge enhancement as a potential alleviation to it have given new relevance to the question, how engineers define knowledge and what attributes they associate with it. AI engineers seem to approach knowledge as a "view from nowhere", a conception prevalent in traditional Western epistemology. Based on this conception, knowledge enhancement strategies are advertised as inheriting increased trustworthiness from the objectivity and neutrality of their knowledge resources. We argue that this promotion of trust is unjustified and harmful. As feminist epistemologists have pointed out, dismissing the importance of the individual knowers behind this knowledge, their values and social settings, effectively conceals the power dynamics at play in knowledge production and dissemination, as well as resulting gaps and misrepresentations. Multiple reports and research studies have revealed such dynamics shaping the epistemic communities behind Wikipedia and Wikidata, knowledge bases which are essential to knowledge-enhanced language modeling. What is revealed is an underlying hierarchy of knowers, organized along dimensions of, e.g., gender, race, and geography. At Wikimedia, the testimony of women or persons from the LGBTQIA+ community is systematically disregarded on the basis of identity prejudice, yielding testimonial injustice. And, the consequence of this is hermeneutical injustice: The resulting knowledge bases primarily reflect the knowledge of and relevant to the dominant group.

Our first take-away is that a more nuanced understanding of knowledge is needed in the AI community. Researchers concerned with measures of knowledge in LMs and other AI systems should be aware of the social nature of knowledge and avoid assuming content labeled "knowledge" to be objective and neutral. Knowledge-enhanced language modeling serves as a case study for the relevance of the social situation to knowledge production. Commonly, comparisons between explicit knowledge resources and statistical AI models attribute bias-risks only to the latter and consider that adding explicit knowledge to statistical systems would make them more robust and less bias-prone. Our preliminary analyses provide evidence against this claim. We were able to show that knowledge enhancement on the basis of Wikidata does not remove biases on a stereotype and task performance level. This is in line with previous findings on biases in commonsense KG-enhanced language models [62], which is to our knowledge the only other work to analyze the relationship between bias and knowledge enhancement. Future work should follow-up with more detailed analyses, across different knowledge bases, LMs, and enhancement approaches. This also includes the currently popular RAG approaches. Understanding the issue at depth is vital as we strive for more trustworthy language models.

Our second take-away is that knowledge bases used in AI must include more diverse voices. More balanced contributions by members of marginalized or excluded groups must be fostered through dedicated structures [9, 37]. Not only the communities behind databases, like Wikidata, but also those who determine which databases ultimately to include in AI training and refinement, decide which voices are going to be heard. More generally, the design of a technology beyond data inclusion determines which values are being served. Hence, technical solutions that allow to encode more than one truth at a time are worth exploring [42]. AI engineers must recognize their own responsibility with regard to the ethical consequences of the technologies they develop [105]. They determine whose knowledge is legitimized, who is served hermeneutical resources, and whose perspectives are excluded, in turn. Diversity is also epistemically necessary to approach objectivity as a horizon. That is, only through intersubjective criticism and scrutiny of members from underrepresented groups can we hope to come closer to objective knowledge production.

Lastly, we would like to stress the importance of interdisciplinary work such as the one presented here and an overcoming of "disciplinary self-isolation" [84, p. 522]. Many ideas that are currently discussed in the AI field are by no means new to other disciplines, like philosophy, political science, or psychology, and in many instances even intentionally borrowed from them. We argue that a more comprehensive understanding of the original discourses provides important insights and, in certain cases, can avert harms.

6 LIMITATIONS
Even though statements and publications by important contemporary voices in the AI field indicate that the observations by Adam [1] and Forsythe [21] still apply (see Section 2.2.3), more up-to-date empirical research on the conceptions of knowledge held by different players in AI is needed and planned for future research. To debunk the prevalent association of knowledge to objectivity and
7 RESEARCHER POSITIONALITY

Both authors identify as Asian-White cis-gendered women, socialized and educated in Western Europe. Both share a background in Computer Science paired with Psychology or Philosophy. The first author considers herself to some extent part of the AI community, and the second author identifies as a researcher in the AI and Ethics of Technology communities. Both support and advocate for feminist and anti-racist values.

ACKNOWLEDGMENTS

We would like to thank Lieke Fröberg, Cedric Möller, Niclas Rautenberg, Jan Moritz Seliger, Ricardo Usbeck and Yan Xi for their very helpful feedback on earlier versions of this article. We would also like to thank our FAccT reviewers for their helpful suggestions. This work is supported by the German Research Foundation (DFG) project NFDI4DS under Grant No.: 460234259 and by the Volkswagen Stiftung (Az 9B331, 9B349). We utilized 2 x NVIDIA RTX A5000 24GB kindly provided by the NVIDIA Academic Hardware Grant Program.

REFERENCES


A DISTRIBUTION OF GENDERS IN WIKIDATA AND KELM

Table 3: Distribution of genders for all person entities in the English Wikidata and in the KELM corpus.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Wikidata</th>
<th>KELM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>Non-binary/...</td>
<td>1,017</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Trans female</td>
<td>1,387</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Trans male</td>
<td>310</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Female</td>
<td>1,988,388</td>
<td>19.47</td>
</tr>
<tr>
<td>Male</td>
<td>6,140,593</td>
<td>60.13</td>
</tr>
<tr>
<td>Unknown</td>
<td>2,080,256</td>
<td>20.37</td>
</tr>
<tr>
<td>Total</td>
<td>10,211,951</td>
<td>100.00</td>
</tr>
</tbody>
</table>

As described in Section 3.2, we investigated the distribution of genders across Wikidata (as of October 2022) and the KELM corpus. All human entities were filtered via relation instance_of and property Q5/human. For each of these, we retrieved property P21/gender or sex if existing. Where no gender was stored or the property value was "undisclosed", we counted the case as "Unknown". Table 3 shows that both datasets predominantly contain information about (cis-)male individuals.

B MODEL DETAILS

Knowledge-enhanced language models are language models with architectural, training, or inference-related adjustments made to increase the performance on knowledge-related tasks or reduce the likelihood of false fabrications during text generation [77]. KEPLER encodes KG entities and aligned text snippets in the same vector space and jointly optimizes for a knowledge embedding loss and a masked language modeling (MLM) loss [104]. This way, the model learns semantically richer representations for entities while preserving linguistic fluency. CoLAKE utilizes the same dataset and follows a similar idea: the input text is concatenated with subgraphs relating to the entities mentioned in the text [97]. Different type embeddings are assigned to the different occurring elements, i.e., words, entities, and relations. The training again follows the MLM objective. Both, KEPLER and CoLAKE are models that employ RoBERTa [51] as their backbone, which they outperform on knowledge-related tasks [97, 104].

We used the implementations and model weights provided through the GitHub repositories of KEPLER[39] and CoLAKE[40] and the HuggingFace implementation and weights of RoBERTa base[41]. We did not fine-tune or otherwise alter the models and ran inference with the original settings.

C VALIDATING ENHANCED PERFORMANCE ON LAMA

Table 4: LAMA evaluation results for different LMs (with and without knowledge enhancement). Numbers represent Mean P@1 scores (higher is better). Bold numbers indicate the best performing LM when comparing the original and their knowledge-enhanced variants.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Relation</th>
<th>RoBERTa</th>
<th>KEPLER</th>
<th>CoLAKE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google-RE</td>
<td>birth-place</td>
<td>11.56</td>
<td>11.90</td>
<td>10.32</td>
</tr>
<tr>
<td></td>
<td>birth-date</td>
<td>1.79</td>
<td>1.47</td>
<td>1.98</td>
</tr>
<tr>
<td></td>
<td>death-place</td>
<td>0.62</td>
<td>3.24</td>
<td>4.93</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>4.66</td>
<td>5.53</td>
<td>5.74</td>
</tr>
<tr>
<td>T-REx</td>
<td>1-1</td>
<td>57.99</td>
<td>57.32</td>
<td>58.68</td>
</tr>
<tr>
<td></td>
<td>N-1</td>
<td>20.32</td>
<td>22.55</td>
<td>23.29</td>
</tr>
<tr>
<td></td>
<td>N-M</td>
<td>19.96</td>
<td>21.43</td>
<td>21.13</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>22.02</td>
<td>23.81</td>
<td>24.17</td>
</tr>
<tr>
<td>SQuAD</td>
<td>Total</td>
<td>9.79</td>
<td>6.64</td>
<td>10.84</td>
</tr>
</tbody>
</table>

We used the LAMA probe [79] to check the effects of the knowledge enhancement on the task performance of the different models. The full probe comprises both encyclopedic and commonsense knowledge types. However, we leave out the commonsense evaluation since this is not the type of knowledge that is enhanced in the models evaluated here. We evaluate on the basis of facts from Wikipedia (Google-RE corpus), triples from Wikidata (T-REx), and question-answer sets derived from Wikipedia (SQuAD). Table 4 shows that KEPLER and CoLAKE slightly outperform their baseline on average for Google-RE and T-REx. For SQuAD, only CoLAKE surpasses RoBERTa. Again, the observed increases are rather small. As they serve only as additional evidence to the metrics reported in the original papers, we interpret these results as sufficient evidence for a successful knowledge enhancement and as providing a basis for further analyses.

Received 22 January 2024; revised 20 March 2024; accepted 30 March 2024

[39]https://github.com/THU-KEG/KEPLER
[40]https://github.com/ttxsun1997/CoLAKE
[41]https://huggingface.co/facebook/roberta-base