Open-World Biomedical Knowledge Probing and Verification

Xi Tian
Zhejiang University
Hangzhou, Zhejiang, China

Peng Wang
Zhejiang University
Hangzhou, Zhejiang, China

Shengyu Mao
Zhejiang University
Hangzhou, Zhejiang, China

ABSTRACT

Recently, there has been a surge of interest in the NLP community in using Pre-trained Language Models (PLMs) as general or domain-specific Knowledge Bases. However, previous evaluation settings are under the Closed-World Assumption. In this paper, we propose open-world knowledge probing and provide an empirical analysis of biomedical PLMs on three datasets. We find that the previous evaluation setting may underestimate the knowledge from the PLMs. We further integrate scientific knowledge into the prompt design and propose SciPrompt, leading to better performance for biomedical knowledge probing. We hope our work can better understand the knowledge learned from PLMs and inspire further research for scientific knowledge discovery.

CCS CONCEPTS

• **Computing methodologies** → *Natural language processing*.

KEYWORDS

Pre-trained Language Models, Probe, Verification, Open-world

ACM Reference Format:
Xi Tian, Peng Wang, and Shengyu Mao. 2023. Open-World Biomedical Knowledge Probing and Verification. In *Proceedings of The 12th International Joint Conference on Knowledge Graphs (IJCKG'23).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recent studies [1, 7, 14] have demonstrated that Pre-trained Language Models (PLMs) implicitly contain different kinds of knowledge [11] in their parameters without the need of human supervision. This is typically done by formulating knowledge triples as cloze-style queries with the objects being masked and using the PLM to fill the single or multiple [Mask] token(s) [3, 8, 20]. In the biomedical domain, it has been shown that specialized PLMs (e.g., BioBERT [9], Bio-LM [10]) potentially contain the implicit Knowledge Graphs (KGs) [2, 13, 18].

Existing research [16] on knowledge probing operates under the Closed-World Assumption (CWA) [15] in which all entities and relations already exist in the KG – they are only knowledge which has been discovered. However, the introduction of PLMs may bring in much-unseen knowledge, which is considered incorrect under CWA, wrongly lowering the knowledge probing performance. As shown in Figure 1, for a triple query *(Omicron, has_symptom, ?)*, the PLM model gives many correct tail entities, but only *Cough*, *Fatigue* and *Sore throat* are considered correct under CWA since it exists in KGs.

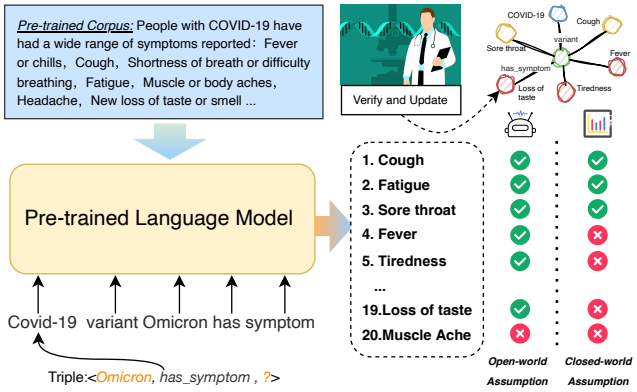

Figure 1: Illustration of open-world knowledge probing for biomedical PLMs.

In this work, we borrow the idea of open-world assumption [4, 17] to develop the open-world knowledge probing for PLMs [12]. With an empirical study by biomedical PLMs on three datasets, we notice that the previous evaluation setting may underestimate the performance of knowledge probing, and parts of the probed knowledge are correct via expert verification. These results indicate that previous knowledge probing methods miss many essential correct facts that may contribute to knowledge discovery from self-supervised pre-training. Moreover, we conduct a comprehensive analysis to find essential patterns of those "new" triples (not existing in KG under CWA). We observe that similar semantic facts account for lots of them, while some other facts are not related to existing KG, which is interesting for further investigation. We also propose a new simple probing method, SciPrompt, which utilizes scientific knowledge in prompt design and yields better performance.

The rest of the paper is organized as follows. Section 2 introduces our new probing method, SciPrompt, along with the knowledge verification procedure. Section 3 introduces our experimental setting and analyzes the intrinsic patterns of those "new" triples probed from PLMs under open-world assumption. Section 4 and Section 5 conclude the paper while discussing the existing challenges and future directions.

2 OPEN-WORLD KNOWLEDGE PROBING

In this section, we introduce the open and close-world assumptions and detail the probing method, SciPrompt, and knowledge verification procedure.

IJCKG'23, Dec 08–09, 2023, Toykyo, Japan

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

Table 1: Main results in CWA setting. We report Acc@1/Acc@5 of each model, including the macro average across three different knowledge sources. The highest scores are boldfaced.

Source	BioBERT			Bio-LM		
	Manual	Opti.	SciPrompt	Manual	Opti.	SciPrompt
Wikidata	3.67 / 11.2	3.21 / 10.75	3.50 / 11.17	**11.97** / 25.92	10.09 / 24.76	10.92 / **26.50**
UMLS	1.15 / 3.81	4.91 / 12.71	5.85 / 13.93	3.44 / 8.88	8.01 / 19.04	**9.37 / 20.90**
ProteinKG25	0.06 / 0.28	8.41 / 18.81	**8.43 / 24.60**	0.58 / 2.15	6.10 / 20.26	8.11 / 21.99

2.1 Open vs. Close-World Setting

Closed-world assumption (CWA) believes that the triples that do not appear in a given knowledge graph are wrong, which is essentially an approximation. Open-world assumption (OWA) assumes that the triples contained in the KG are not complete, which is more accurate and closer to the real scenario. The knowledge not in KGs is not false, but unknown. However, it requires additional human annotations to verify those triples carefully. In this paper, we introduce open-world knowledge probing for PLMs. We ask human experts to verify triples and tag them as **correct**, **incorrect**, and **unknown** (triples that can not be verified through resources on the Web). All annotated and verified data will be released for research purposes.

2.2 Knowledge Sources

Wikidata Wikidata[1] is a public Knowledge Base with many factual knowledge across various domains.
UMLS The UMLS Metathesaurus[2] is a biomedical knowledge graph that contains various vocabularies and concepts in the biomedical domain.
ProteinKG25 ProteinKG25[3] is a knowledge graph for protein science that contains descriptions and protein sequences (entity nodes). We use a sub-set of the ProteinKG25 with the relation *is a*, *is part of*, and *has part*.

2.3 Probing Methods

Vanilla Prompt. We use a fill-in-the-blank cloze statement for probing and adopt two baseline methods: manual prompts [14], and OptiPrompt [22]. For each relation, we follow Sung et al.[18] to create manual prompts with domain experts. In addition, OptiPrompt automatically obtains continuous embeddings, which are trained with disjoint instances.

SciPrompt. We additionally propose a new prompt-based method, SciPrompt, for probing. Specifically, we utilize scientific terms to construct discrete prompt tokens. We also add some continuous tokens, which make prompts using any vector in the embedding space. Formally, our prompt has the following form:

$$t_r = [\text{Term}]_1 \dots [\text{Term}]_m [\text{P}]_1 \dots [\text{P}]_n [\text{MASK}], \quad (1)$$

where each [Term] is a discrete token from scientific terms related to relation r, each $[\text{P}]_i \in \mathbb{R}^d$ is a dense vector with the same dimension as the LM's input embedding. We initialize the continuous embeddings with the embeddings of manual prompts, which worked consistently better than random initialization. With those scientific terms and learnable tokens, our SciPrompt can take advantage of domain knowledge as knowledge-informed prompts to elicit scientific factual knowledge. Following Sung et al.[18], we use gradient descent to minimize the negative log-likelihood for prompt optimization.

From Table 1, we observe that SciPrompt can obtain better performance, indicating the advantages of scientific knowledge guidance. Notably, manual prompt also perform well in Wikidata, possibly because Wikidata belongs to general domain. We leverage the best-performed model to generate candidate triples. For Top-1 triples that do not exist in the KG under OWA, we leverage knowledge verification addressed in the following section.

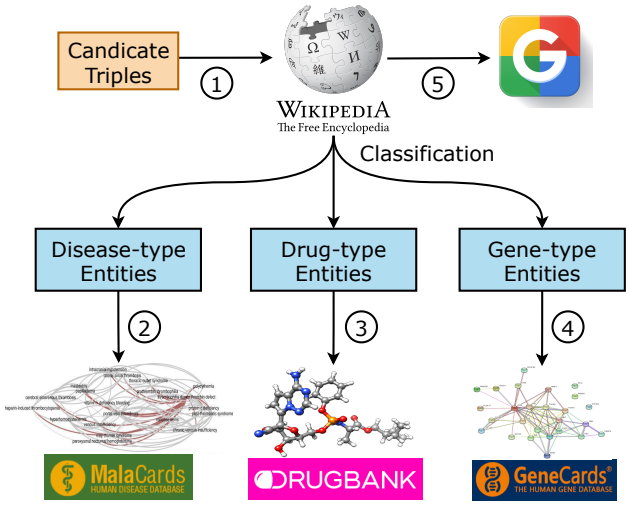

Figure 2: Procedure of human experts verification.

2.4 Knowledge Verification

During knowledge verification, we randomly sample 100 Top-1 triples for each relation that do not exist in the KG under CWA, resulting in a total number of 2.4K triples across three datasets. Each triple is annotated by 5 human experts with a biomedical background and proficiency in English as a second language. These human experts are trained for the well-designed knowledge verification procedure.

[1]https://wikidata.org/
[2]https://www.nlm.nih.gov/research/umls/
[3]https://zjunlp.github.io/project/ProteinKG25

Table 2: Main results in OWA setting. We report CR@1/PR@1 of each relation.

Relation ID	Relation Name	Subject	Object	#Triples	CR@1	PR@1
Wikidata						
P2175	medical condition treated	chemical	disease	704	20	21
P2176	drug used for treatment	disease	chemical	435	34	40
P2293	genetic association	gene	disease	830	11	13
P4044	therapeutic area	chemical	disease	344	14	14
P780	symptoms	disease	symptom	303	44	49
UMLS						
UR116	clinically associated with	disease	disease	773	14	24
UR124	may treat	disease	chemical	531	24	27
UR173	causative agent of	disease	vertebrate	560	63	66
UR180	is finding of disease	disease	body substance	419	12	20
UR211	biological process involves gene product	gene	function	736	27	56
UR214	cause of	disease	disease	548	21	26
UR221	gene mapped to disease	disease	gene	247	10	30
UR254	may be finding of disease	disease	symptom	614	33	46
UR256	may be molecular abnormality of disease	disease	genetic aberrant	262	7	58
UR44	may be prevented by	chemical	disease	507	23	24
UR45	may be treated by	chemical	disease	864	25	28
UR48	physiologic effect of	chemical	disease	834	1	37
UR49	mechanism of action of	chemical	function	755	5	38
UR50	therapeutic class of	chemical	type	738	13	91
UR588	process involves gene	gene	disease	750	16	77
UR625	disease has associated gene	gene	disease	465	4	7
ProteinKG25						
P0	is a	gene function	gene function	19891	7	96
P1	is part of	gene function	gene function	3374	12	94
P3	has part	gene function	gene function	301	31	99

As shown in Figure 2, we ask these human experts to verify those candidate triples under OWA. Human experts are asked to search the triple knowledge on Wikipedia to verify the correctness. For those candidate triples not existing in Wikipedia, human experts will search MalaCards[4](an integrated database of human maladies and their annotations) for disease-related entities, Drugbank[5](a comprehensive, freely accessible, online database containing information on drugs and drug targets) for drug-related entities, and GeneCards[6](a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes) for gene-related entities. Human experts will utilize Google to verify triples when the candidate triple does not exist in the database mentioned above. The triple is tagged with correct or wrong when there is a shred of solid evidence in the text from the existing authoritative biomedical corpus. The triple is tagged with unknown when there is no substantial evidence on the Web and these human experts are also unable to confirm.

[4]https://www.malacards.org/
[5]https://go.drugbank.com/
[6]https://www.genecards.org/

For each triple, we follow above annotation process and statistics votes of 5 human experts, ultimately concluding with the final result based on the majority consensus. The entire annotation procedure is an exhaustive endeavor, spanning approximately one month in total. To further assess the quality and agreement of annotations, we calculate the average inter-rater agreement between annotators using Fleiss' Kappa scores [5], finding that annotations show perfect agreement ($\kappa = 0.9$).

3 EXPERIMENTS

3.1 Settings

We follow Sung et al.[18] to reprocess all the datasets. For the open-world knowledge probing setting, we utilize **CR@1** and **PR@1** as evaluation metrics, which indicates **absolute correct** and **partial correct** (including absolute correct and those unknown triples), respectively. We argue that those unknown triples (triples that cannot be verified based on the present information on the Web) are valuable, and possibly part of them can be proved experimentally in the future.

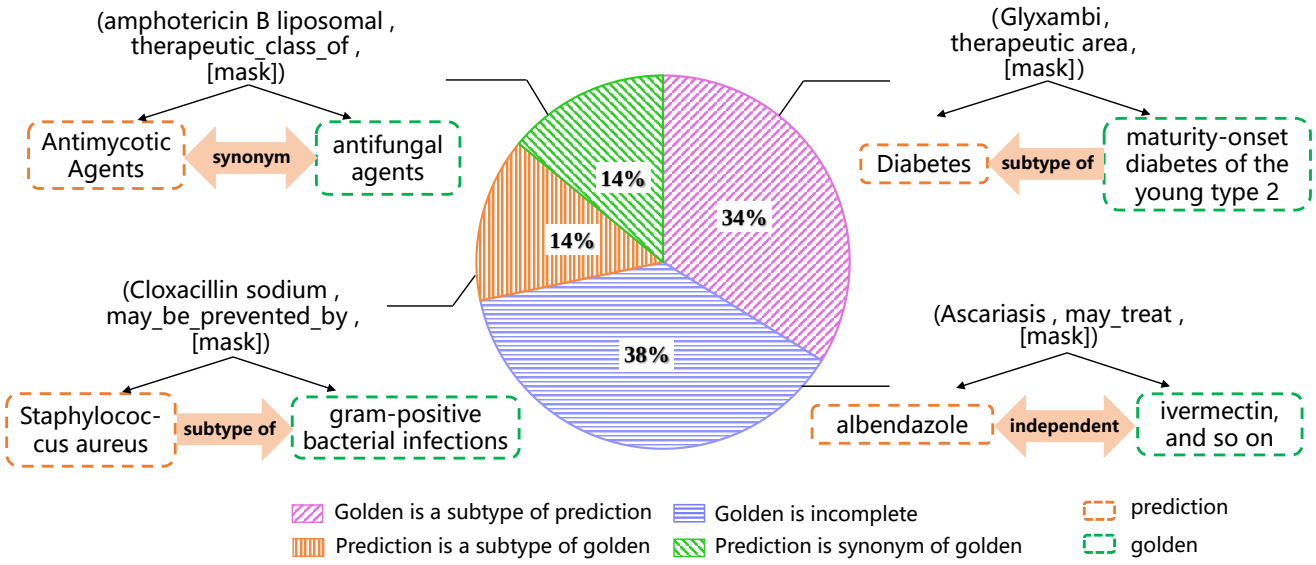

Figure 3: Case study for triples in OWA setting. We analyze those triples tagged correctly by human experts and divide them into four groups.

3.2 Results

Is this knowledge from PLMs correct? From Table 2, we notice that part of the knowledge from all datasets is actually correct when verified by human experts, which indicates that we may underestimate the rich knowledge from the PLMs. Note that the triples with *causative agent of* even obtain 63 CR@1 and 66 PR@1 scores, illustrating that **PLM has the potential ability to discover "new" scientific knowledge**.

What are the characteristics of this knowledge? To further investigate the intrinsic characteristic of this knowledge in the OWA setting, as shown in Figure 3, we conduct a manual analysis of these **annotated correct triples** and categorize them into four groups as follows :

Group 1: The predicted tail entity is independent of the golden entity. We observe that 38% of the tail entities in predicted triples have little relevance to the golden entities but are actually correct by human expert verification based on public information. We argue that this may be because, during pre-training, the PLMs have seen lots of patterns, leading to knowledge recalling.

Group 2: The predicted tail entity is a subtype of golden entities. We observe that 14% of the predicted entities are subtypes of golden entities, indicating that PLMs have the powerful ability to capture useful lexical-type knowledge, which is consistent with [6].

Group 3: The golden entity is a subtype of predicted tail entities. We observe that 34% of the golden entities are subtypes of predicted entities, demonstrating that PLMs can capture lots of hierarchy knowledge (e.g., a subtype of).

Group 4: The golden entity is a synonym of the golden entity. We observe that 14% of the predicted entities are synonyms of golden entity, which illustrates that PLMs have the powerful ability to memorize synonyms.

Overall, we notice that this knowledge with open-world knowledge probing from PLMs has part of intrinsic patterns but also contains **unknown emerging abilities of knowledge discovery** [21]. We think it is interesting to study the potential ability of large-scale PLMs via open-world knowledge probing, as it promises to uncover even more hidden knowledge.

4 DISCUSSION

Recently, PLMs have been suggested as a possible complement to KGs. However, previous studies focus on probing the existing knowledge (CWA Setting) rather than exploring the "new" (OWA Setting). We think this approach is complementary to the existing evaluation system (e.g., LAMA, BioLAMA) and can serve as a scaffold for investigating open knowledge in PLMs. We hope that by uncovering the potential ability of PLMs with open-world knowledge probing, we can continue to motivate exploring the positives of pre-training intrinsically and apply the technology for automatic science knowledge discovery.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose open-world knowledge probing for biomedical PLMs. We provide an empirical analysis of three datasets and find that the previous evaluation setting may underestimate the knowledge from the PLMs. We further propose SciPrompt, which integrates scientific knowledge into the prompt, obtaining better performance for biomedical knowledge probing. We hope that our work can contribute to a deeper understanding of the knowledge learned from PLMs and inspire further research for scientific knowledge discovery. In the future, we plan to study the evaluation strategies for open-world knowledge probing via fact verification [19] and systematically re-evaluate the existing PLMs to reveal their real performance.

REFERENCES

[1] Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A Review on Language Models as Knowledge Bases. *CoRR* abs/2204.06031 (2022). https://doi.org/10.48550/arXiv.2204.06031 arXiv:2204.06031

[2] Diego Calvanese, Davide Lanti, Tarcisio Mendes De Farias, Alessandro Mosca, and Guohui Xiao. 2021. Accessing scientific data through knowledge graphs with Ontop. *Patterns* 2, 10 (2021).

[3] Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 1860–1874. https://doi.org/10.18653/v1/2021.acl-long.146

[4] İsmail İlkan Ceylan, Adnan Darwiche, and Guy Van den Broeck. 2016. Open-World Probabilistic Databases. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016*, Chitta Baral, James P. Delgrande, and Frank Wolter (Eds.). AAAI Press, 339–348. http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12908

[5] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378. https://psycnet.apa.org/record/1972-05083-001

[6] Michael Hanna and David Marecek. 2021. Analyzing BERT's Knowledge of Hypernymy via Prompting. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2021, Punta Cana, Dominican Republic, November 11, 2021*, Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad (Eds.). Association for Computational Linguistics, 275–282. https://aclanthology.org/2021.blackboxnlp-1.20

[7] Benjamin Heinzerling and Kentaro Inui. 2020. Language Models as Knowledge Bases: On Entity Representations, Storage Capacity, and Paraphrased Queries. *CoRR* abs/2008.09036 (2020). arXiv:2008.09036 https://arxiv.org/abs/2008.09036

[8] Benjamin Heinzerling and Kentaro Inui. 2021. Language Models as Knowledge Bases: On Entity Representations, Storage Capacity, and Paraphrased Queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, 1772–1791. https://doi.org/10.18653/v1/2021.eacl-main.153

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.* 36, 4 (2020), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[10] Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Online, 146–157. https://doi.org/10.18653/v1/2020.clinicalnlp-1.17

[11] Peng Liu, Lemei Zhang, and Jon Atle Gulla. 2023. Pre-train, Prompt and Recommendation: A Comprehensive Survey of Language Modelling Paradigm Adaptations in Recommender Systems. *arXiv e-prints*, Article arXiv:2302.03735 (Feb. 2023), arXiv:2302.03735 pages. https://doi.org/10.48550/arXiv.2302.03735 arXiv:2302.03735 [cs.IR]

[12] Xin Lv, Yankai Lin, Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. Do Pre-trained Models Benefit Knowledge Graph Completion? A Reliable Evaluation and a Reasonable Approach. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 3570–3581. https://aclanthology.org/2022.findings-acl.282

[13] Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2022. Rewire-then-Probe: A Contrastive Recipe for Probing Biomedical Knowledge of Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 4798–4810. https://aclanthology.org/2022.acl-long.329

[14] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language Models as Knowledge Bases?. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 2463–2473. https://doi.org/10.18653/v1/D19-1250

[15] Raymond Reiter. 1977. On Closed World Data Bases. In *Logic and Data Bases, Symposium on Logic and Data Bases, Centre d'études et de recherches de Toulouse, France, 1977 (Advances in Data Base Theory)*, Hervé Gallaire and Jack Minker (Eds.). Plemum Press, New York, 55–76. https://doi.org/10.1007/978-1-4684-3384-5_3

[16] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 5418–5426. https://doi.org/10.18653/v1/2020.emnlp-main.437

[17] Baoxu Shi and Tim Weninger. 2018. Open-World Knowledge Graph Completion. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 1957–1964. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16055

[18] Mujeen Sung, Jinhyuk Lee, Sean S. Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can Language Models be Biomedical Knowledge Bases?. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 4723–4734. https://doi.org/10.18653/v1/2021.emnlp-main.388

[19] David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, United States, 61–76. https://doi.org/10.18653/v1/2022.findings-naacl.6

[20] Cunxiang Wang, Pai Liu, and Yue Zhang. 2021. Can Generative Pre-trained Language Models Serve As Knowledge Bases for Closed-book QA?. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 3241–3251. https://doi.org/10.18653/v1/2021.acl-long.251

[21] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *CoRR* abs/2206.07682 (2022). https://doi.org/10.48550/arXiv.2206.07682 arXiv:2206.07682

[22] Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual Probing Is [MASK]: Learning vs. Learning to Recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, 5017–5033. https://doi.org/10.18653/v1/2021.naacl-main.398