CONKE: Conceptualization-Augmented Knowledge Editing in Large Language Models for Commonsense Reasoning

Anonymous ACL submission

Abstract

Knowledge Editing (KE) aims to adjust a Large Language Model's (LLM) internal representations and parameters to correct inaccuracies and improve output consistency without incurring the computational expense of re-training the entire model. However, editing commonsense knowledge still faces difficulties, including limited knowledge coverage in existing resources, the infeasibility of annotating labels for an overabundance of commonsense knowledge, and the strict knowledge formats of current editing methods. In this paper, we address these challenges by presenting CONKE, a framework that integrates conceptualization and instantiation into the KE pipeline for LLMs to enhance their commonsense reasoning capabilities. CONKE dynamically diagnoses implausible commonsense knowledge within an LLM using another verifier LLM and augments the source knowledge to be edited with conceptualization for stronger generalizability. Experimental results demonstrate that LLMs enhanced with CONKE successfully generate commonsense knowledge with improved plausibility compared to other baselines and achieve stronger performance across multiple question answering benchmarks.

1 Introduction

Recent advancements in Large Language Models (LLMs;OpenAI, 2024b,a; Dubey et al., 2024; Chan et al., 2024) have led to Knowledge Editing (KE;Zhang et al., 2024; Wang et al., 2025), a computationally efficient strategy to correct inaccurate responses and update LLMs by modifying their internal weights or representations, without retraining the entire model. Such methods have been applied to various domains, including factual reasoning (Ju et al., 2024; Wang et al., 2024a), medical knowledge (Xu et al., 2024b), and commonsense reasoning (Huang et al., 2024), and have proven effective in enhancing domain-specific expertise.

Despite their success, current KE methods face several challenges, including limited knowledge coverage (Davis and Marcus, 2015) in existing commonsense knowledge bases (West et al., 2023; Fang et al., 2021b; Yang et al., 2023; Fang et al., 2021a, 2023; Ding et al., 2024; Xu et al., 2024a) which offer limited coverage and focus on isolated facts, rather than forming hierarchical structures that enable generalization through editing (Ma et al., 2021b; Wang et al., 2024e). Furthermore, the unstructured nature of commonsense knowledge complicates scaling, while the flexible representation of commonsense knowledge means that a single fact may manifest in multiple formats. This necessitates editing at the `(relation, tail)` pair level rather than at individual tokens.

To address these issues, we present CONKE, a novel knowledge editing framework tailored for editing commonsense knowledge within LLMs. We use VERA (Liu et al., 2023), an automated commonsense plausibility verifier, to assess the plausibility of commonsense knowledge in LLMs. For knowledge deemed erroneous and requiring edits, we integrate conceptualization and instantiation (Wang et al., 2023b,a) to enrich semantic coverage and support more generalizable editing, covering not only the targeted knowledge but also other potentially relevant yet implausible information within the LLM. This pipeline

To ensure flexibility, CONKE adopts an open-ended format for editing, enabling the handling of arbitrary knowledge structures rather than focusing solely on traditional `(h,r,t)` triplets. We go beyond traditional Knowledge Editing techniques by combining automated knowledge detection, conceptualization, and instantiation, enhancing the model's ability to generalize and adapt to diverse contexts. Experimental results on AbstractATOMIC (He et al., 2024) demonstrate that LLMs enhanced by CONKE generate commonsense knowledge with improved plausibility. Fur-

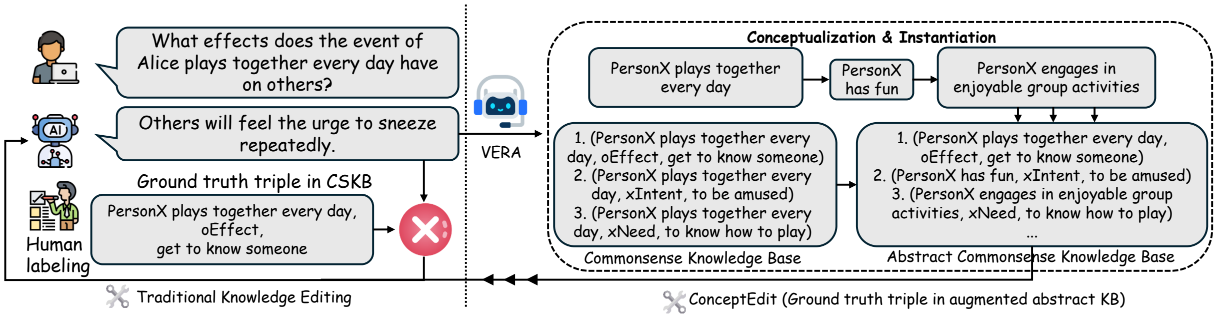

Figure 1: An overview of CONKE, which pipelines conceptualization and instantiation, knowledge editing, and LLM verification together for automated and scalable knowledge editing over commonsense knowledge.

ther evaluations across five commonsense question-answering benchmarks also show performance improvements. These experiments demonstrate the robustness and generalizability of our approach in enhancing commonsense reasoning across diverse architectures and tasks.

2 Related Works

2.1 Knowledge Editing

Knowledge editing (Cao et al., 2021) aims to update an LLM's internal knowledge without full retraining or relying solely on prompt engineering, is becoming increasingly crucial. Meng et al. (2022) propose ROME, which identifies and updates factual associations within specific MLP layers, achieving precise single-fact edits guided by causal mediation analysis. MEMIT (Meng et al., 2023) extends ROME's principles to handle large-scale edits simultaneously. By distributing updates across multiple layers and parameters, MEMIT efficiently integrates thousands of facts while maintaining specificity and fluency. GRACE (Hartvigsen et al., 2023), on the other hand, avoids internal parameter changes by integrating external dictionaries and adapters as a modular memory source. This approach allows flexible, inference-time access to new knowledge, though it may sacrifice some internal coherence and interpretability. In our work, we build upon these methods to enhance editing commonsense knowledge in LLMs.

2.2 Conceptualization in Commonsense

Conceptualization abstracts entities or events into general concepts, forming abstract commonsense knowledge (Murphy, 2004), while instantiation grounds these concepts into new instances, introducing additional commonsense knowledge. Previous work largely focused on entity-level conceptualization (Durme et al., 2009; Song et al., 2011, 2015; Liu et al., 2022; Peng et al., 2022), with He et al. (2024); Wang et al. (2023b,a) pioneering event-level conceptualization from WordNet (Miller, 1995) and Probase (Wu et al., 2012). For instantiation, Allaway et al. (2023) introduced a controllable generative framework that automatically identifies valid instances. In this work, we leverage the conceptualization distillation framework proposed by Wang et al. (2024d) to augment the knowledge being edited, ensuring broader semantic coverage and thereby improving the generalizability of edited knowledge.

3 The CONKE Framework

An overview of CONKE is presented in Figure 1. Our framework consists of three main components: (1) automated knowledge verification with VERA (Liu et al., 2023), (2) abstract knowledge acquisition via conceptualization and instantiation, and (3) LLM knowledge editing. We use the AbstractATOMIC (He et al., 2024) and CANDLE (Wang et al., 2024d) datasets for training and evaluation as two rich sources of abstract knowledge with conceptualization and instantiation. The training set of both datasets are used for editing and the testing sets are used for evaluation.

3.1 Automated Knowledge Verification

Since commonsense knowledge is vast, traditional human-in-the-loop methods for detecting and correcting erroneous outputs in LLMs are neither easily scalable nor adaptable. Inspired by recent advances in using LLMs as automated judges (Raina et al., 2024; Wang et al., 2024c), we propose a fully automated verification strategy to assess an LLM's internal commonsense knowledge. Our verification process involves VERA (Liu et al., 2023), a discriminative model trained to score the plausibility of arbitrary commonsense statements, as our evaluation tool. For each triple in the AbstractATOMIC (He et al., 2024) training set, we prompt the LLM with the head event and request it to generate the corresponding relation and tail. VERA then evaluates the plausibility of the generated knowledge by producing a score in the range $[0, 1]$, where values above 0.5 are considered plausible, and those below 0.5 are deemed implausible. By iterating over all triples, this process provides both the LLM's generated responses and VERA's discrimination results, pinpointing which portions of the generated knowledge are incorrect. Consequently, we can identify the exact "areas" within the LLM's internal knowledge that require editing. This automated pipeline eliminates the dependence on costly human annotations for error detection, enabling scalable and efficient improvements of the LLM's commonsense understanding.

3.2 Conceptualization and Instantiation

While existing approaches primarily integrate decontextualized commonsense knowledge into LLMs through KE techniques, we hypothesize that capturing the diverse patterns that the same piece of knowledge can exhibit under different contexts is equally important. However, repeated editing may result in knowledge drift, where successive modifications will lead to substle conflicts, causing the model's internal representation to become unstable. To this end, we augment the knowledge to be edited by implementing both conceptualization and instantiation, following Wang et al. (2024d). For each triple targeted for editing, we first abstract its instances into more general concepts by prompting GPT-4o, producing abstract knowledge triples (Figure 1). We then instantiate these abstract concepts into novel, context-specific instances, again using GPT-4o, thereby forming a rich knowledge base. This process yields approximately 160,000 commonsense knowledge triples, substantially improving the semantic coverage and contextual adaptability of the edited knowledge. Additionally, we are mindful of cascading effects that may arise when modifying a piece of commonsense knowledge. As noted in (Wang et al., 2024b), knowledge is highly interconnected, and modifying one fact can trigger unintended changes in related facts, leading to inconsistencies. To mitigate these cascading effects, we use conceptualization and instantiation to ensure that modifications to abstract concepts are consistently applied to their related instances, hence maintaining coherence and reducing the risk of introducing inconsistencies.

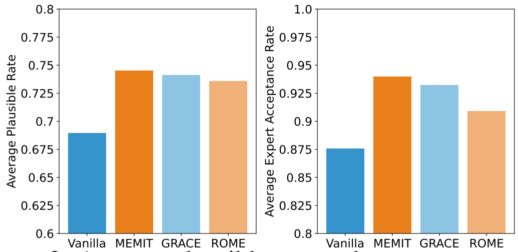

Figure 2: Average plausible rate and expert acceptance rate of LLMs' generation after CONKE.

3.3 LLM Knowledge Editing

Finally, we apply knowledge editing to the LLM using the enriched knowledge base generated through our conceptualization and instantiation processes, correcting errors identified by VERA. To accomplish this, we experiment with three established knowledge editing methods: MEMIT (Meng et al., 2023), ROME (Meng et al., 2022), and GRACE (Hartvigsen et al., 2023). For GRACE, which relies on adapters to determine whether and how to use an external dictionary, we adopt the original deferral mechanism implementation. We evaluate our framework with these editing methods on four representative LLM backbones: `Mistral-7B-Instruct-v0.2`(Jiang et al., 2023), `Meta-Llama-3-8B-Instruct`(Dubey et al., 2024), `Chatglm2-6b`(Zeng et al., 2024), and `GPT-J-6B`(Wang and Komatsuzaki, 2021).

4 Experiments and Analyses

In this section, we evaluate LLMs after applying CONKE through expert and automated assessments, illustrating improved performance on downstream tasks and present several ablation studies.

4.1 LLMs-After-Editing Evaluation

We first evaluate LLMs after editing via two measures. First, we prompt these LLMs with head events in the testing set of AbstractATOMIC and ask it to complete the commonsense knowledge. With the generations on the testing set, we ask VERA to score them again and we calculate the plausible ratio whose scores are above 0.5. Then, we sample a subset of 200 generations and recruit two expert annotators to conduct a manual analyses on the acceptance ratio of the plausible assertions that passed VERA's filtering. We compare models after being edited with MEMIT, GRACE, and ROME, and set another vanilla group as baseline comparison. As shown in Figure 2, both VERA and human evaluations exhibit consistent trends,

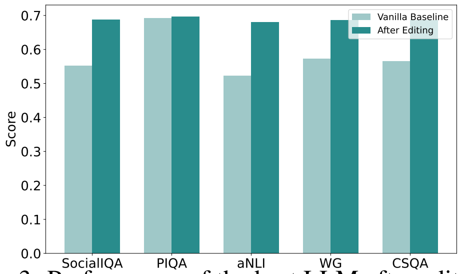

Figure 3: Performance of the best LLM after editing on five downstream tasks compared to the vanilla baseline.

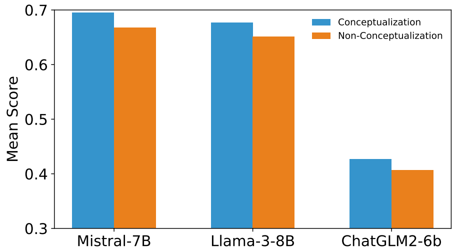

Figure 4: VERA evaluation scores of edited LLMs with and without integrating conceptualization.

with human raters tend to assign higher scores but identifying similar improvements. When applying MEMIT-based editing, both VERA and human evaluations show notable enhancements over the Vanilla baseline. Similarly, GRACE and ROME edits enhance plausibility scores, with MEMIT and GRACE achieving the highest overall performance. The strong results from expert annotations further validate the reliability of VERA's judgments, supporting the use of VERA in our framework as an effective commonsense evaluator to identify implausible knowledge requiring further editing. This approach reduces reliance on manual annotations while preserving robust assessment capabilities.

4.2 Downstream Improvements

To assess whether enhanced internal commonsense reasoning improves downstream task performance, we evaluate the edited models on multiple commonsense reasoning benchmarks. Following Ma et al. (2021a), we test our framework on the validation splits of five widely-used commonsense QA benchmarks: Abductive NLI (aNLI; Bhagavatula et al., 2020), CommonsenseQA (CSQA; Talmor et al., 2019), PhysicalIQA (PIQA; Bisk et al., 2020), SocialIQA (SocialIQA; Sap et al., 2019), and WinoGrande (WG; Sakaguchi et al., 2021). These benchmarks are designed to evaluate a range of knowledge types crucial for robust commonsense reasoning (Shi et al., 2023; Wang and Song, 2024). We compare the performance of the best LLM edited with CONKE against its corresponding vanilla baseline across all benchmarks, with the results visualized in Figure 3. The results show that models edited with CONKE achieve significant performance improvements across all benchmarks, with particularly notable gains in aNLI and SocialIQA. These findings demonstrate the effectiveness of CONKE in enhancing commonsense reasoning capabilities and suggest its potential for broader applications in improving LLM performance on real-world reasoning tasks.

4.3 Ablation Study

Finally, to validate the effect of conceptualization, we conducted an ablation study on MEMIT by removing the conceptualization step and comparing performance. In this setup, we edit LLMs both with and without the integration of conceptualization and instantiation, and evaluate their performance by examining the average VERA scores of the generated outputs on the testing set. The conceptualized variant leveraged enriched commonsense triples generated via abstraction and instantiation prior to the editing process, while the non-conceptualized variant directly applied MEMIT without these preprocessing steps.

Figure 4 demonstrates that the conceptualized variants consistently outperform their non-conceptualized counterparts, achieving higher plausibility and improved downstream task accuracy. These results suggest that the enriched conceptual patterns introduced before editing not only enhance plausibility but also enable the model to generalize commonsense knowledge to more complex reasoning tasks, ultimately boosting overall performance.

5 Conclusions

In this paper, we introduce CONKE, a novel knowledge editing framework designed to enhance commonsense reasoning in LLMs by addressing challenges of limited knowledge coverage and scalability, and by integrating automated verification through VERA and semantic enrichment via conceptualization and instantiation for more effective and generalizable editing. Experimental results demonstrate significant improvements in both knowledge plausibility and downstream task performance, validating the effectiveness of our approach. We envision that CONKE will inspire future research on scalable and context-aware knowledge editing, paving the way for LLMs to better handle the complexity and diversity of commonsense reasoning.

Limitations

Our approach, CONKE, advances LLM commonsense reasoning through conceptualization and iterative knowledge editing, yet several challenges persist. First, editing one piece of knowledge can cascade through related concepts, creating non-linear interactions that are difficult to detect and manage, especially as the knowledge base scales up. Second, iterative updates risk knowledge drift, where successive edits subtly conflict with or overwrite prior facts, emphasizing the need for robust frameworks to maintain consistency. Finally, the lack of stable ground truth for commonsense, which is often context-sensitive and culturally variable, complicates standardization. Addressing these challenges will require globally coordinated editing mechanisms, improved theoretical frameworks, and systematic human-in-the-loop validation to ensure edits align with broader consensus and expert judgment.

Ethics Statement

In this paper, all datasets and models used are free and accessible for research purposes, aligning with their intended usage. The expert annotators are graduate students with extensive experience in NLP and commonsense reasoning research, and they voluntarily agreed to participate without compensation.

However, we recognize commonsense knowledge is inherently culturally and contextually variable, and there are ethical considerations related to the knowledge edited and propagated through the models. We must ensure that the knowledge inserted into the model doesn't favor certain views over others, especially in sensitive cases such as healthcare or law applications. To mitigate this, we implement a robust process of cross-validation with human experts to monitor for biases. Moreover, we propose regular audits of the system's performance, to ensure that its performance remains fair.

References

Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen R. McKeown, Doug Downey, and Yejin Choi. 2023. Penguins don't fly: Reasoning about generics through instantiations and exceptions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2610–2627. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. 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*, pages 6491–6506. Association for Computational Linguistics.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of chatgpt on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 684–721. Association for Computational Linguistics.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Wenxuan Ding, Weiqi Wang, Sze Heng Douglas Kwok, Minghao Liu, Tianqing Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, Chen Luo, Qingyu Yin, Bing Yin, Junxian He, and Yangqiu Song. 2024. Intentionqa: A benchmark for evaluating purchase intention comprehension abilities of language models in e-commerce. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 2247–2266. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab

AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Benjamin Van Durme, Phillip Michalak, and Lenhart K. Schubert. 2009. Deriving generalized knowledge from corpora using wordnet abstraction. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, pages 808–816. The Association for Computer Linguistics.

Tianqing Fang, Quyet V. Do, Sehyun Choi, Weiqi Wang, and Yangqiu Song. 2023. CKBP v2: An expert-annotated evaluation set for commonsense knowledge base population. *CoRR*, abs/2304.10392.

Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. Benchmarking commonsense knowledge base population with an effective evaluation dataset. 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*, pages 8949–8964. Association for Computational Linguistics.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. DISCOS: bridging the gap between discourse knowledge and commonsense knowledge. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.

Tom Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. Aging with GRACE: lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2024. Acquiring and modeling abstract commonsense knowledge via conceptualization. *Artif. Intell.*, 333:104149.

Xiusheng Huang, Yequan Wang, Jun Zhao, and Kang Liu. 2024. Commonsense knowledge editing based on free-text in llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 14870–14880. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Tianjie Ju, Yijin Chen, Xinwei Yuan, Zhuosheng Zhang, Wei Du, Yubin Zheng, and Gongshen Liu. 2024. Investigating multi-hop factual shortcuts in knowledge editing of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8987–9001. Association for Computational Linguistics.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1264–1287. Association for Computational Linguistics.

Jingping Liu, Tao Chen, Chao Wang, Jiaqing Liang, Lihan Chen, Yanghua Xiao, Yunwen Chen, and Ke Jin. 2022. Vocsk: Verb-oriented commonsense knowledge mining with taxonomy-guided induction. *Artif. Intell.*, 310:103744.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021a. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13507–13515. AAAI Press.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Satoru Ozaki, Eric Nyberg, and Alessandro Oltramari. 2021b. Exploring strategies for generalizable commonsense reasoning with pre-trained models. 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*, pages 5474–5483. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Gregory Murphy. 2004. *The big book of concepts*. MIT press.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence. *OpenAI*.

OpenAI. 2024b. Hello gpt-4o. *OpenAI*.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. COPEN: probing conceptual knowledge in pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5015–5035. Association for Computational Linguistics.

Vyas Raina, Adian Liusie, and Mark J. F. Gales. 2024. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot LLM assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7499–7517. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. 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*, pages 4462–4472. Association for Computational Linguistics.

Haochen Shi, Weiqi Wang, Tianqing Fang, Baixuan Xu, Wenxuan Ding, Xin Liu, and Yangqiu Song. 2023. QA-DYNAMICS: training dynamics-driven synthetic QA diagnostic for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15329–15341. Association for Computational Linguistics.

Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. 2011. Short text conceptualization using a probabilistic knowledgebase. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 2330–2336. IJCAI/AAAI.

Yangqiu Song, Shusen Wang, and Haixun Wang. 2015. Open domain short text conceptualization: A generative + descriptive modeling approach. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 3820–3826. AAAI Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. 2024a. Cross-lingual knowledge editing in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11676–11686. Association for Computational Linguistics.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024b. Knowledge editing for large language models: A survey. *Preprint*, arXiv:2310.16218.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2025. Knowledge editing for large language models: A survey. *ACM Comput. Surv.*, 57(3):59:1–59:37.

Weiqi Wang, Limeng Cui, Xin Liu, Sreyashi Nag, Wenju Xu, Sheikh Sarwar, Chen Luo, Yang Laurence Li, Hansu Gu, Hui Liu, Changlong Yu, Jiaxin Bai, Yifan Gao, Haiyang Zhang, Qi He, Shuiwang Ji, and Yangqiu Song. 2024c. EcomScriptBench: A multi-task benchmark for e-commerce script planning via step-wise intention-driven product association. *CoRR*.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: Conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13520–13545, Singapore. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Jiayang Cheng, Chunkit Chan, and Yangqiu Song. 2024d. CANDLE: iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*

2024, Bangkok, Thailand, August 11-16, 2024. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Haochen Shi, Baixuan Xu, Wenxuan Ding, Liyu Zhang, Wei Fan, Jiaxin Bai, Haoran Li, Xin Liu, and Yangqiu Song. 2024e. On the role of entity and event level conceptualization in generalizable reasoning: A survey of tasks, methods, applications, and future directions. *CoRR*, abs/2406.10885.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13111–13140. Association for Computational Linguistics.

Weiqi Wang and Yangqiu Song. 2024. MARS: benchmarking the metaphysical reasoning abilities of language models with a multi-task evaluation dataset. *CoRR*, abs/2406.02106.

Peter West, Ronan Le Bras, Taylor Sorensen, Bill Yuchen Lin, Liwei Jiang, Ximing Lu, Khyathi Chandu, Jack Hessel, Ashutosh Baheti, Chandra Bhagavatula, and Yejin Choi. 2023. Novacomet: Open commonsense foundation models with symbolic knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1127–1149. Association for Computational Linguistics.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 481–492. ACM.

Baixuan Xu, Weiqi Wang, Haochen Shi, Wenxuan Ding, Huihao Jing, Tianqing Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, Chen Luo, Qingyu Yin, Bing Yin, Long Chen, and Yangqiu Song. 2024a. MIND: multimodal shopping intention distillation from large vision-language models for e-commerce purchase understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7800–7815. Association for Computational Linguistics.

Derong Xu, Ziheng Zhang, Zhihong Zhu, Zhenxi Lin, Qidong Liu, Xian Wu, Tong Xu, Wanyu Wang, Yuyang Ye, Xiangyu Zhao, Enhong Chen, and Yefeng Zheng. 2024b. Editing factual knowledge and explanatory ability of medical large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 2660–2670. ACM.

Zonglin Yang, Xinya Du, Erik Cambria, and Claire Cardie. 2023. End-to-end case-based reasoning for commonsense knowledge base completion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 3491–3504. Association for Computational Linguistics.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from GLM-130B to GLM-4 all tools. *CoRR*, abs/2406.12793.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. 2024. A comprehensive study of knowledge editing for large language models. *CoRR*, abs/2401.01286.