Contents

Chapter 6

Conclusions and Critical Reflections

We have now arrived at the closing part, where we will summarize this thesis and discuss its limitations, significance, and potential future directions. This chapter will present the main conclusions of this thesis along with critical and systematic reflections on the limitations of the thesis.

6.1 Conclusions and Contributions

While intelligence has long been a quest for human beings, we now stand at a critical point in time. We are experiencing an intelligence revolution, where in the envisioned future, intelligence can be packed into units that can be disseminated easily across time and space, akin to how Industrial Revolution packs our physical capabilities into units. In this revolution, knowledge plays a crucial role as it serves as the interface between our cognition and the reality. An accurate knowledge interface allows intelligent agents to conceptualize and model the reality effectively (even though it remains uncertain if humans experience the reality directly). The knowledge of actions and their consequences further allows the intelligent agents to intervene and transform their environments. Hence, building knowledge engines are essential to both natural and artificial intelligence.

There are two conventional paradigms to constructing knowledge engines: the structured and unstructured paradigm, exemplified respectively by knowledge graphs and large language models. Which one is better? This thesis aims to discuss this age-old debate in the context of recent findings in knowledge graphs and language models. We argue that the presence of structures is inevitable regardless of whether data is explicitly structured (as in knowledge graphs) or implicitly structured (as in large language models). Furthermore, we assert that not all structures are of positive roles. If focusing narrowly on structure formation, we can make models that are overly rigid. Thus, we motivate the necessity of destructuring, which improves models' plasticity so that they can learn rapidly with few examples in new environments.

In summary, this thesis presents a scientific journey to discover the commonalities between the two mainstream paradigms for building knowledge engines. Although these paradigms initially appear distinct, often perceived as separate approaches, this thesis demonstrates that deeper connections can be established through a functional examination of model training dynamics and analytical reformulation of model computations. The contributions of this thesis are fourfold as detailed below:

First, the thesis identifies new connections between the two seemingly disjoint paradigms as summarized in Table 6.1:

- The language modelling objective induces latent structures within model computations, supporting tasks such as knowledge base completion and the interpretation of large language models.
- Active forgetting enables inductive reasoning in *both* paradigms, facilitating efficient generalization to unseen entities in knowledge graphs and new languages in pretrained language models.

Second, we provide new insights into the role of structures in building general knowledge engines:

- Structures are indispensable for knowledge engines, though they can manifest in various forms – explicit in data or implicit within models. The structured paradigm explicitly specifies the structures in the data. For the unstructured paradigms, latent structures about the relationships among tokens can be directly extracted from model computations post-training (Chapter 3).
- However, overly encoding structures within models can hinder their ability to generalize to unseen scenarios, highlighting the importance of balancing structures and flexibility (Chapter 4).

Table 6.1: Comparison of the structured and unstructured paradigm through the dual forces of structure and destructure.

Force	**Structured Paradigm**	**Unstructured Paradigm**
Structure	Language modelling objectives induce structure into factorization models *(Chapter 2)*	Language modelling objectives induce structure into Transformers *(Chapter 3)*
Destructure	Active forgetting enhances generalization to unseen graphs *(Chapter 4)*	Active forgetting enhances generalization to unseen languages *(Chapter 5)*

Third, we obtain new understandings about the embedding layer, the often-overlooked components in *both* paradigms:

- The concept of the *embedding sandwich* emerges as a suitable architectural abstraction for models in both the structured and unstructured paradigm, e.g. transformers (Chapter 3 and Chapter 5).
- Embeddings should not be examined in isolation but rather in conjunction with their optimization dynamics. They serve as dynamic repositories where gradient descent accumulates, propagates, and stores symbolic interactions. (Chapter 4).

Finally, our findings advocate for a shift in focus from the surface-level distinctions of structured versus unstructured data to the underlying dynamics of structure formation and destructuring as indicated by Table 6.2:

- Rather than focusing solely on whether data is structured or unstructured, especially given modern datasets often contain data exhibiting varying degrees of structures, we highlight the need to study both the forces driving structure formation (Part I) and their opposing force, destructuring (Part II).
- Structuring promotes structure encoding in the models. Destructuring mitigates the rigidity of excessive structuring, enabling AI systems to adapt and reason effectively in dynamic, unseen environments—an ability we term *model plasticity*.

In conclusion, this thesis highlights that structure formation and its dual force, destructuring, are both essential components for building general knowledge engines.

Table 6.2: The unified paradigm seen through the mechanistic forces of structure and destructure.

Force	**The (Un)Structured Paradigm**
Structure	Language modelling induces structure in model computation *(Part I)*
Destructure	Active forgetting helps address unseen symbols and adapt to new environments *(Part II)*

6.2 Limitations and Flaws

While this thesis provides new insights into the bridging of structured and unstructured learning, the thesis contains several limitations and flaws in its current form. While individual chapters already provide discussion on their own limitation, this section acknowledges global limitations related to the topic of structured and unstructured learning so that the readers can have a rigorous assessment of the thesis.

6.2.1 Theoretical Scope

We discussed several key constructs and concepts used in our thesis, where broader notions of them are combed through.

Structure

The core concept in this thesis is *structure*. In the traditional discussion of structured and unstructured learning, the concept of structure mainly centres around the structures in the data. This thesis takes a step further to discuss the relationship between the structures in the data and the structures in the computational model: language modelling objective can induce the former into the latter; the latter can in turn be recovered to the former by rearranging model computation. In this sense, the thesis considers primarily structures in the context of relational learning and language modelling.

However, structures have other broader notions which the thesis could have engaged with. We enumerate a number of them to better contextualize our notion of structures.

First, structures are the obsession object for philosophers, psychologist, educators who strive to understand human minds. In this context, structures typically refer to mental or cognitive structures, with knowledge being perhaps one of the most important of these structures. We review several famous notions of structures under this category. As early as 1781, Kant discussed how the mind structures experience and how such abstractions form the basis for humans to understand the world [Kant, 1781]. Clinical methods were used to study how these cognitive structures form in children by Piaget in 1920s [Piaget, 1929]. Using controlled observations, Vygotsky further highlighted the mental structures are highly impacted by external factors such as language and culture [Vygotsky, 1934]. Bruner examined the role of mental structures in the learning process and showed how abstract thinking is necessary to organize new experiences and knowledge [Bruner, 1960]. More recently, Deleuze further argued that mental structures are not static but dynamic and emergent in his masterwork, *Difference and Repetition* [Deleuze and Patton, 1994]. This thesis can be seen as an effort towards implementing such a dynamic notion of mental structures in AI systems (in fact, to a certain extent, one can see our destructure process as Deleuze's difference process and our structure formation as Deleuze's repetition process), while more understandings into the difference and repetition processes are required to fully realize the flexible structures described by Deleuze.

Secondly, in programming, structures often mean data structures, the abstract models for organizing and storing data [Knuth, 1997]. In this case, structures refer to an abstraction where the physical implementation is often hidden, and developers interact with abstract representations of the data. In our thesis, the "structure" in the structured and unstructured paradigm refer to the structures in the training data. Specifically, in the case of structured paradigm, the structures are relational structures formatted in subject-relation-object triples [Ji et al., 2020]; in the unstructured paradigm, the texts are without such formatting, e.g. the first few tokens are not necessarily the subject rather they can play various grammar roles depending on the contexts.

Thridly, structures in mathematics, such as sets, groups, and graphs, are abstractions that represent relationships [Dummit et al., 2004, Hausdorff, 2021, Deisenroth et al., 2020]. These mathematical constructs are essential in modelling relationships and complex systems. In the structured paradigm, knowledge bases can be represented using graphs, known as knowledge graphs. The graph representation of knowledge bases enable easy visualization, comprehension, and efficient querying, reasoning [Noy et al., 2019, Ji et al., 2020]

Finally, structures have ample notions in machine learning. Most of these notions explcitly incorporate structures into learning and learnt structures are used to help reasoning. We discuss a couple of works with such explicit structure integration. In Bayesian learning, structures manifest as Bayesian networks, which are probabilistic graphical models consisting of variables and their conditional dependencies expressed by a directed acyclic graph [Neal, 2012]. Compared to the usual neural networks, Bayesian networks can be used for prediction with uncertainty. Similarly, in causal inference, structures often refer to causal graphs or structural causal models (SCMs) [Pearl, 1998]. These causal models describe causal relationships between variables, distinguishing causation from correlation [Pearl and Shafer, 1995, Pearl and Mackenzie, 2018]. Another branch of work, neuro-symbolic AI [Besold et al., 2021], integrates symbolic structures with neural networks. Since neural networks excel at pattern recognition and symbolic reasoning excels at abstract concepts and logic rules, neuro-symbolic AI aims to combine their strengths [Garcez et al., 2019]. In this domain, structures often refer to knowledge representations such as knowledge graphs and logic rules [Hamilton et al., 2024, Colelough and Regli, 2024]. Thus, the knowledge graphs based learning methods in this thesis can apply to some neuro-symbolic AI systems while more research are needed to extend the methods to complicated structures like logic rules. For example, it would be interesting to explore how active forgetting (Part II) could help models adapt logic rule templates by flexibly substituting different entities (i.e., performing variable instantiations) depending on the task. This could make reasoning systems more adaptable and task-specific.

All the above notions of structures are also meaningful structural objects to extend our methods with. A more comprehensive treatment of the broader notions of structures would require integrating toolkits from causal machine learning, general Bayesian learning, and neuro-symbolic AI.

Destructure

The concept of destructure introduced in this thesis, with active forgetting as one potential implementation, is not without limitations. While active forgetting specifically targets the embeddings, actively removing the structures captured in them, other components of the model could also be considered for forgetting. However, we have chosen embeddings as the primary target for this process, leaving other model components unexplored. One key limitation is the lack of automatic selection for which component should undergo forgetting during the pretraining process. Although it would be more convenient for users if such a mechanism were automated, this would add significant overhead to an already computationally expensive pretraining phase. Additionally, the idea of automating the schedule of forgetting frequency rather than treating it as a hyperparameter introduces further complexity that may increase the computational burden.

It is also worthwhile to compare with techniques such as dropout [Baldi and Sadowski, 2013, Srivastava et al., 2014] and iterative pruning [Frankle and Carbin, 2019]. These methods periodically erase weights, providing regularization and helping to prevent overfitting. However, they are not designed to specifically address generalization to unseen environments, which is a key goal of our proposed destructuring method. Theoretically, active forgetting and similar techniques could be linked to frameworks like Invariant Risk Minimization (IRM) [Arjovsky et al., 2019], which aims to reduce risks across different environments and improve generalization to unseen data points. However, further investigation is required to fully establish this connection.

Another limitation arises from the lack of study on one-time destructuring methods, which focus on removing unwanted structures in the models, directly patching problematic model behaviors. These methods, such as DPR [Karpukhin et al., 2020], RAG [Lewis et al., 2020b], model editing [Meng et al., 2022], and model unlearning [Liu et al., 2025], address specific issues like hallucinations or toxicity in LLM generations. However, they lack the ability to systematically and globally address these issues across the model in an integrated way. Instead, they rely on external interventions, which may not lead to the same depth of control over the model's learning and forgetting processes as the proposed approaches in this thesis. That said, these external methods are reactive and easy to deploy on-the-fly.

6.2.2 Methodological Constraints

While this thesis attempts to bridge structured and unstructured learning paradigms, it centres on the embedding and its role in caching symbolic relationships. The unification is limited to reframing neural embedding optimization as structural operations (message-passing over graphs). This unification is implicit rather than explicit. The alternative direction can be to estabish a mathematically rigorous framework or an empirical system that directly mixes both structured and unstructured inputs synthesizing neural networks and symbolic reasoning [Colelough and Regli, 2024]. For example, directly injecting structured data into the unstructured paradigm, making language models structure-aware [Li et al., 2023, Wu et al., 2024].

6.2.3 Evaluation, Scalability, and Computation Efficiency

In the destructuring experiments presented in Part II, we focused on evaluating the model's performance on unseen entities in knowledge graphs and unseen languages for pretrained language models. However, in real-world applications, there are many other potentially unseen scenarios that need to be considered. For pretrained language models, there exists a wide spectrum of linguistic shifts to which the model must adapt. These shifts include domain shifts [Gururangan et al., 2020], temporal evolution [Liska et al., 2022], task/-tool changes [Lu et al., 2024], and personalization for different users [Kirk et al., 2024]. All of these factors are important to test when applying active forgetting techniques to ensure that they are effective across a variety of scenarios.

Although latent structures in large language models were analysed in Chapter 3, extracting transparent higher-order n-grams remains a partially unsolved challenge. This issue requires further scaling up of our methods to handle more complex structures. Similarly, the analysis of computational paths can be extended to explore the cascading effects across multiple self-attention modules, a task that demands increased computational resources. We focused on extracting n-grams for a selected set of large language models. To gain a deeper understanding of more models and their internal knowledge, we must systematically examine a broader range of models. This would include potentially verifying the extracted structures against data distributions or real-world knowledge graphs to validate the models' generalization capabilities and alignment with external knowledge.

6.3 Significance and Implications

This thesis carries several important implications that extend beyond its theoretical contributions, shedding light on practical applications for both artificial and human cognitive systems.

6.3.1 Applications for Machine Minds

By bridging the structured and unstructured paradigms, this work enables cross-paradigm learning, allowing us to borrow strengths from both approaches.

One key application lies in improving the *controllability* of large language models (LLMs). For instance, by training models on relevant n-gram paths identified through our method in Chapter 3, we might be able to enhance their functional flexibility. This can be particularly useful in managing *tooling plasticity* [Lu et al., 2024], where tools can be interpreted as "neural circuits" that activate only under specific conditions. By identifying model paths related to particular tool usages, we could apply active forgetting from Chapter 5 to adjust or reset those paths as needed. Additionally, this work offers pathways to improve the *interpretability* and transparency of LLMs. As shown in Chapter 3, our n-gram interpretability requires only CPU-based post-training processing and no curated external datasets, making it more computationally efficient than other approaches. This method allows institutional actors to systematically audit LLMs, enhancing transparency and user trust in generative AI applications.

6.3.2 Applications for Human Minds

This thesis also implies a broader scientific inquiry into how humans build knowledge engines in their minds.

First, the research on embeddings and their function as "memory banks" during training (Chapter 4) provides insights into how human memory might work and be regulated, potentially linking to *engram cells* in neuroscience [Tonegawa et al., 2015, 2018, Ryan and Frankland, 2022, Guskjolen and Cembrowski, 2023]. Second, *active forgetting* techniques that help pretrained language models generalize to new languages with less data (Chapter 5) could inspire new research in language acquisition. This is especially relevant to studying phenomena such as the *critical period* for language learning [Constantinescu et al., 2024]. Third, this thesis highlights potential applications for *digital intervention*. Many recommender systems and social media platforms maintain persistent embeddings for individual users. While these technologies have become integral to daily life, they often accumulate implicit user preferences in embeddings, leading to issues like "brain rot" addiction and echo chamber. If platforms periodically reset user embeddings, we might be able to mitigate digital addiction, and promote healthier online interactions.

Bibliography

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O Alabi, Shamsuddeen H Muhammad, Peter Nabende, et al. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. In *2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4488–4508. Association for Computational Linguistics (ACL), 2022.

Khushbu Agarwal, Tome Eftimov, Raghavendra Addanki, Sutanay Choudhury, Suzanne R. Tamang, and Robert Rallo. Snomed2vec: Random walk and poincaré embeddings of a clinical knowledge base for healthcare analytics. *ArXiv*, abs/1907.08650, 2019. URL `https://api.semanticscholar.org/CorpusID:198147334`.

Divyanshu Aggarwal, Ashutosh Sathe, and Sunayana Sitaram. Exploring pretraining via active forgetting for improving cross lingual transfer for decoder language models. *arXiv preprint arXiv:2410.16168*, 2024.

Jethro Akroyd, Sebastian Mosbach, Amit Bhave, and Markus Kraft. Universal digital twin - a dynamic knowledge graph. *Data-Centric Engineering*, 2:e14, 2021. doi: 10.1017/dce.2021.10.

Ibrahim Alabdulmohsin, Hartmut Maennel, and Daniel Keysers. The impact of reinitialization on generalization in convolutional neural networks. *arXiv preprint arXiv:2109.00267*, 2021.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL `https://aclanthology.org/2022.coling-1.382`.

Prithviraj Ammanabrolu and Mark Riedl. Learning knowledge graph-based world models of textual environments. *Advances in Neural Information Processing Systems*, 34: 3720–3731, 2021.

Michael C. Anderson and Justin C. Hulbert. Active forgetting: Adaptation of memory by prefrontal control. *Annual Review of Psychology*, 72(1):1–36, 2021. doi: 10.1146/annurev-psych-072720-094140. URL `https://doi.org/10.1146/annurev-psych-072720-094140`. PMID: 32928060.

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. *Advances in neural information processing systems*, 29, 2016.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, 2022.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, 2020.

Various authors. Wikipedia, the free encyclopedia, 2024. URL https://www.wikipedia.org. A collaboratively edited, free online encyclopedia.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Ivana Balazevic, Carl Allen, and Timothy M. Hospedales. Tucker: Tensor factorization for knowledge graph completion. In *EMNLP/IJCNLP*, 2019.

Pierre Baldi and Peter J Sadowski. Understanding dropout. *Advances in neural information processing systems*, 26, 2013.

Jeffrey Barrett and Kevin JS Zollman. The role of forgetting in the evolution and learning of language. *Journal of Experimental & Theoretical Artificial Intelligence*, 21(4):293–309, 2009.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

Shawn Beaulieu, Lapo Frati, Thomas Miconi, Joel Lehman, Kenneth O Stanley, Jeff Clune, and Nick Cheney. Learning to continually learn. *arXiv preprint arXiv:2002.09571*, 2020.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Edward L Bennett, Marian C Diamond, David Krech, and Mark R Rosenzweig. Chemical and anatomical plasticity of brain: Changes in brain through experience, demanded by learning theories, are found in experiments with rats. *Science*, 146(3644):610–619, 1964.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Generative or discriminative? getting the best of both worlds. *Bayesian statistics*, 8(3):3–24, 2007.

Jacob A Berry, Dana C Guhle, and Ronald L Davis. Active forgetting and neuropsychiatric diseases. *Molecular Psychiatry*, pages 1–11, 2024.

Tarek R Besold, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, et al. Neural-symbolic learning and reasoning: A survey and interpretation 1. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, pages 1–51. IOS press, 2021.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for bert. *arXiv preprint arXiv:2104.07143*, 2021.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, J. Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, 2013.

Léon Bottou. *Stochastic Gradient Descent Tricks*, pages 421–436. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8_25. URL `https://doi.org/10.1007/978-3-642-35289-8_25`.

Thorsten Brants, Ashok Popat, Peng Xu, Franz Josef Och, and Jeffrey Dean. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference* *on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, 2007.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Richard E Brown. Hebb and cattell: The genesis of the theory of fluid and crystallized intelligence. *Frontiers in human neuroscience*, 10:606, 2016.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Jerome Bruner. *The Process of Education*. Harvard University Press, 1960.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. Imbalanced learning: A comprehensive evaluation of resampling methods for class imbalance. *arXiv preprint arXiv:1710.05381*, 2018. URL `https://arxiv.org/abs/1710.05381`.

Raymond B Cattell. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, 54(1):1, 1963.

Yihong Chen, Bei Chen, Xiangnan He, Chen Gao, Yong Li, Jian-Guang Lou, and Yue Wang. λopt: Learn to regularize recommender models in finer levels. In *KDD 2019 (Oral), Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 978–986, 2019.

Yihong Chen, Pasquale Minervini, Sebastian Riedel, and Pontus Stenetorp. Relation prediction as an auxiliary training objective for improving multi-relational graph representations. In *AKBC 2021*, 2021.

Yihong Chen, Pushkar Mishra, Luca Franceschi, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Refactor gnns: Revisiting factorisation-based models from a message-passing perspective. In *Advances in Neural Information Processing Systems*, 2022.

Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Ifeoluwa Adelani, Pontus Stenetorp, Sebastian Riedel, and Mikel Artetxe. Improving language plasticity via pretraining with active forgetting. In *NeurIPS 2023*, 2023.

Yihong Chen, Xiangxiang Xu, Yao Lu, Pontus Stenetorp, and Luca Franceschi. Jet expansions of residual computation, 2024. URL `https://arxiv.org/abs/2410.06024`.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Brandon C Colelough and William Regli. Neuro-symbolic ai in 2024: A systematic review. 2024.

Together Computer. Redpajama dataset. `https://www.together.xyz/blog/redpajama`, 2023. Accessed: 2023-12-12.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc.,

2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Paper-Conference.pdf.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. Investigating critical period effects in language acquisition through neural language models. *arXiv preprint arXiv:2407.19325*, 2024.

OpenWebText Contributors. The openwebtext dataset. https://github.com/jcpeterson/openwebtext, 2019. Accessed: 2023-12-12.

Moheb Costandi. *Neuroplasticity*. MIT Press, 2016.

Common Crawl. Common crawl corpus. https://commoncrawl.org, 2023. Accessed: 2023-12-12.

Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong. *Mathematics for machine learning*. Cambridge University Press, 2020.

Gilles Deleuze and Paul Patton. *Difference and Repetition*. Athlone, London, 1994.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Kevin P. Donnelly. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279–90, 2006. URL `https://api.semanticscholar.org/CorpusID:22491040`.

Pierluca D'Oro, Max Schwarzer, Evgenii Nikishin, Pierre-Luc Bacon, Marc G Bellemare, and Aaron Courville. Sample-efficient reinforcement learning by breaking the replay ratio barrier. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=OpC-9aBBVJe`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61): 2121–2159, 2011. URL `http://jmlr.org/papers/v12/duchi11a.html`.

David Steven Dummit, Richard M Foote, et al. *Abstract algebra*, volume 3. Wiley Hoboken, 2004.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Meza-Ruiz, et al. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, 2022.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Javier Ferrando and Elena Voita. Information flow routes: Automatically interpreting language models at scale. *arXiv preprint arXiv:2403.00824*, 2024.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-jussà. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*, 2024.

Matthias Fey, Jan Eric Lenssen, Frank Weichert, and Jure Leskovec. Gnnautoscale: Scalable and expressive graph neural networks via historical embeddings. In *ICML*, 2021.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJl-b3RcF7`.

Eberhard Fuchs and Gabriele Flügge. Adult neuroplasticity: more than 40 years of research. *Neural plasticity*, 2014(1):541870, 2014.

A Garcez, M Gori, LC Lamb, L Serafini, M Spranger, and SN Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Journal of Applied Logics*, 6(4):611–632, 2019.

Jean-Baptiste Gaya, Thang Doan, Lucas Caccia, Laure Soulier, Ludovic Denoyer, and Roberta Raileanu. Building a subspace of policies for scalable continual learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UKr0MwZM6fL.

Floris Geerts and Juan L Reutter. Expressiveness and approximation properties of graph neural networks. In *International Conference on Learning Representations*, 2021.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. URL https://arxiv.org/abs/2004.07780.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, 2022.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Siavash Golkar, Micheal Kagan, and Kyunghyun Cho. Continual learning via neural pruning. In *Real Neurons {\&} Hidden Units: Future directions at the intersection of neuroscience and artificial intelligence@ NeurIPS 2019*.

Joshua T Goodman. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434, 2001.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, 2:729–734 vol. 2, 2005.

C Shawn Green and Daphne Bavelier. Exercising your brain: a review of human brain plasticity and training-induced learning. *Psychology and aging*, 23(4):692, 2008.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, 2020.

Axel Guskjolen and Mark S Cembrowski. Engram neurons: Encoding, consolidation, retrieval, and forgetting of memory. *Molecular psychiatry*, 28(8):3207–3219, 2023.

Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. *Semantic Web*, 15(4):1265–1306, 2024.

William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159.

William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035, 2017.

Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

Oliver Hardt, Einar Örn Einarsson, and Karim Nader. A bridge over troubled water: Reconsolidation as a link between cognitive and neuroscientific memory research traditions. *Annual review of psychology*, 61(1):141–167, 2010.

Oliver Hardt, Karim Nader, and Lynn Nadel. Decay happens: the role of active forgetting in memory. *Trends in cognitive sciences*, 17(3):111–120, 2013.

Michael Hart and Project Gutenberg Volunteers. Project gutenberg online library, 1971–2024. URL `https://www.gutenberg.org`. Free eBooks from the public domain.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.234. URL `https://aclanthology.org/2022.acl-long.234`.

Felix Hausdorff. *Set theory*, volume 119. American Mathematical Soc., 2021.

Frederick Hayes-Roth, Donald A Waterman, and Douglas B Lenat. *Building expert systems*. Addison-Wesley Longman Publishing Co., Inc., 1983.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021a.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021b. URL `https://openreview.net/forum?id=XPZIaotutsD`.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. Measuring and manipulating knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.

F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *J. Math. Phys*, 6(1):164–189, 1927.

S Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation MIT-Press*, 1997.

Sara Hooker. The hardware lottery. *Communications of the ACM*, 64(12):58–65, 2021.

Roy Horan. The neuropsychological connection between creativity and meditation. *Creativity research journal*, 21(2-3):199–222, 2009.

John L Horn and Raymond B Cattell. Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of educational psychology*, 57(5):253, 1966.

Ian Horrocks. Owl: A description logic based ontology language. In *International conference on principles and practice of constraint programming*, pages 5–8. Springer, 2005.

Ian Horrocks, Peter F Patel-Schneider, and Frank van Harmelen. From shiq and rdf to owl: The making of a web ontology language. *Journal of Web Semantics*, 1(1):7–26, 2003. URL `https://doi.org/10.1016/j.websem.2003.07.001`.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, and Yizhou Sun. Gpt-gnn: Generative pre-training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2020.

David Hume. An enquiry concerning human understanding. 1748. *Classics of Western Philosophy*, pages 763–828, 1999.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, 2021.

Alex Iacob, Lorenzo Sani, Meghdad Kurmanji, William F Shen, Xinchi Qiu, Dongqi Cai, Yan Gao, and Nicholas D Lane. Dept: Decoupled embeddings for pre-training language models. *arXiv preprint arXiv:2410.05021*, 2024.

Maximilian Igl, Gregory Farquhar, Jelena Luketina, Wendelin Boehmer, and Shimon Whiteson. Transient non-stationarity and generalisation in deep reinforcement learning. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=Qun8fv4qSby`.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=6t0Kwf8-jrj`.

Prachi Jain, Sushant Rathi, Mausam, and Soumen Chakrabarti. Knowledge base completion: Baseline strikes back (again). *ArXiv*, abs/2005.00804, 2020a.

Prachi Jain, Sushant Rathi, Mausam, and Soumen Chakrabarti. Knowledge base completion: Baseline strikes back (again). *CoRR*, abs/2005.00804, 2020b. URL `https://arxiv.org/abs/2005.00804`.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. A survey on knowledge graphs: Representation, acquisition and applications. *arXiv preprint arXiv:2002.00388*, 2020.

Rudolf Kadlec, Ondrej Bajgar, and Jan Kleindienst. Knowledge base completion: Baselines strike back. In Phil Blunsom, Antoine Bordes, Kyunghyun Cho, Shay B. Cohen, Chris Dyer, Edward Grefenstette, Karl Moritz Hermann, Laura Rimell, Jason Weston, and Scott Yih, editors, *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*, pages 69–74. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-2609. URL `https://doi.org/10.18653/v1/w17-2609`.

Immanuel Kant. *Critique of Pure Reason (1st edition)*. Macmillan Company, Mineola, New York, 1781.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

Jill L Kays, Robin A Hurley, and Katherine H Taber. The dynamic brain: neuroplasticity and mental health. *The Journal of neuropsychiatry and clinical neurosciences*, 24(2): 118–124, 2012.

Seyed Mehran Kazemi and David Poole. Simple embedding for link prediction in knowledge graphs. *Advances in neural information processing systems*, 31, 2018.

Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, pages 381–388. AAAI Press, 2006.

Phillip Kent. Fluid intelligence: A brief history. *Applied Neuropsychology: Child*, 6(3): 193–203, 2017.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.

Byung-Hak Kim, Arvind Yedla, and Henry D Pfister. Imp: A message-passing algorithm for matrix completion. In *2010 6th International Symposium on Turbo Codes & Iterative Information Processing*, pages 462–466. IEEE, 2010.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Jeffrey A Kleim and Theresa A Jones. Principles of experience-dependent neural plasticity: implications for rehabilitation after brain damage. 2008.

Donald Ervin Knuth. *The art of computer programming*, volume 3. Pearson Education, 1997.

Stanley Kok and Pedro M. Domingos. Statistical predicate invention. In *ICML*, volume 227 of *ACM International Conference Proceeding Series*, pages 433–440. ACM, 2007.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018.

Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. Canonical tensor decomposition for knowledge base completion. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2869–2878. PMLR, 2018.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022.

Seungpil Lee, Woochang Sim, Donghyeon Shin, Wongyu Seo, Jiwon Park, Seokki Lee, Sanha Hwang, Sejin Kim, and Sundong Kim. Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. *ACM Trans. Intell. Syst. Technol.*, January 2025. ISSN 2157-6904. doi: 10.1145/3712701. URL `https://doi.org/10.1145/3712701`. Just Accepted.

Su Young Lee, Choi Sungik, and Sae-Young Chung. Sample-efficient deep reinforcement learning via episodic backward update. *Advances in neural information processing systems*, 32, 2019.

Benedetta Leuner and Elizabeth Gould. Structural plasticity and hippocampal function. *Annual review of psychology*, 61(1):111–140, 2010.

Benjamin J Levy, Nathan D McVeigh, Alejandra Marful, and Michael C Anderson. Inhibiting your native language: The role of retrieval-induced forgetting during second-language acquisition. *Psychological Science*, 18(1):29–34, 2007.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, 2020a.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020b.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. Implicit representations of meaning in neural language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.143. URL `https://aclanthology.org/2021.acl-long.143`.

Chenchen Li, Aiping Li, Ye Wang, Hongkui Tu, and Yichen Song. A survey on approaches and applications of knowledge representation learning. In *2020 IEEE Fifth*

International Conference on Data Science in Cyberspace (DSC), pages 312–319. IEEE, 2020.

Ren Li, Yanan Cao, Qiannan Zhu, Guanqun Bi, Fang Fang, Yi Liu, and Qian Li. How does knowledge graph embedding extrapolate to unseen data: a semantic evidence view. *CoRR*, abs/2109.11800, 2021b.

Xinze Li, Zhenghao Liu, Chenyan Xiong, Shi Yu, Yu Gu, Zhiyuan Liu, and Ge Yu. Structure-aware language model pretraining improves dense retrieval on structured data. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. Gated graph sequence neural networks. In *ICLR (Poster)*, 2016.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, pages 13604–13622. PMLR, 2022.

Chen Liu, Jonas Pfeiffer, Anna Korhonen, Ivan Vulić, and Iryna Gurevych. Delving deeper into cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2408–2423, 2023a.

Chunan Liu, Lilian Denzler, Yihong Chen, Andrew Martin, and Brooks Paige. Asep: Benchmarking deep learning methods for antibody-specific epitope prediction. In *NeurIPS 2024, Proceedings of the Thirty-eighth Conference on Neural Information Processing Systems, Datasets and Benchmarks*, 2024a.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023b.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*, 2024b.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14, 2025.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019b. URL `http://arxiv.org/abs/1907.11692`.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Jiarui Lu, Thomas Holleis, Yizhe Zhang, Bernhard Aumayer, Feng Nan, Felix Bai, Shuang Ma, Shen Ma, Mengyu Li, Guoli Yin, Zirui Wang, and Ruoming Pang. Toolsandbox: A stateful, conversational, interactive evaluation benchmark for llm tool use capabilities, 2024. URL `https://arxiv.org/abs/2408.04682`.

Scott Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Qingsong Lv, Ming Ding, Qiang Liu, Yuxiang Chen, Wenzheng Feng, Siming He, Chang Zhou, Jianguo Jiang, Yuxiao Dong, and Jie Tang. Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '21, page 1150–1160, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467350. URL `https://doi.org/10.1145/3447548.3467350`.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 23190–23211. PMLR, 7 2023. URL `https://proceedings.mlr.press/v202/lyle23b.html`.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Brian MacWhinney. A unified model of language acquisition. In Judith F. Kroll and Annette M.B. De Groot, editors, *Handbook of Bilingualism: Psycholinguistic Approaches*, pages 49–67. Oxford University Press, 2005.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.

Kelly Marchisio, Patrick Lewis, Yihong Chen, and Mikel Artetxe. Mini-model adaptation: Efficiently extending pretrained models to new languages via aligned shallow training. In *ACL 2023, Findings of the Association for Computational Linguistics*, 2023.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL `https://www.sciencedirect.com/science/article/pii/S0079742108605368`.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35: 17359–17372, 2022.

B. D. Mishra, Niket Tandon, and P. Clark. Domain-targeted, high precision knowledge extraction. *Transactions of the Association for Computational Linguistics*, 5:233–246, 2017.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2021.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR, 2022.

Sameh K Mohamed, Vít Nováček, Pierre-Yves Vandenbussche, and Emir Muñoz. Loss functions in knowledge graph embedding models. In *Proceedings of DL4KG2019-Workshop on Deep Learning for Knowledge Graphs*, page 1, 2019.

Aaron Mueller. Missed causes and ambiguous effects: Counterfactuals pose challenges for interpreting neural networks. *arXiv preprint arXiv:2407.04690*, 2024.

Deepak Nathani, Jatin Chauhan, Charu Sharma, and Manohar Kaul. Learning attention-based embeddings for relation prediction in knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4710–4723, 2019.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Q. Phung. A novel embedding model for knowledge base completion based on convolutional neural network. In *NAACL-HLT (2)*, pages 327–333. Association for Computational Linguistics, 2018.

Timothy Nguyen. Understanding transformers via n-gram statistics. *arXiv preprint arXiv:2407.12034*, 2024.

M. Nickel, Volker Tresp, and H. Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, 2011a.

M. Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104:11–33, 2016a.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, pages 809–816. Omnipress, 2011b.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proc. IEEE*, 104(1):11–33, 2016b.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. Holographic embeddings of knowledge graphs. In *AAAI*, pages 1955–1961. AAAI Press, 2016c.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *International Conference on Machine Learning*, pages 16828–16847. PMLR, 2022.

Evgenii Nikishin, Junhyuk Oh, Georg Ostrovski, Clare Lyle, Razvan Pascanu, Will Dabney, and Andre Barreto. Deep reinforcement learning with plasticity injection. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023. URL https://openreview.net/forum?id=O9cJADBZT1.

nostalgebraist. interpreting gpt: the logit lens, 2021. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens#HEf5abD7hqqAY2GSQ.

Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale knowledge graphs: Lessons and challenges: Five diverse technology companies show how it's done. *Queue*, 17(2):48–75, 2019.

Simon Nørby. Why forget? on the adaptive value of memory loss. *Perspectives on Psychological Science*, 10(5):551–578, 2015. doi: 10.1177/1745691615596787. URL https://doi.org/10.1177/1745691615596787. PMID: 26385996.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113: 54–71, 2019.

Denise C Park and Chih-Mao Huang. Culture wires the brain: A cognitive neuroscience perspective. *Perspectives on Psychological Science*, 5(4):391–400, 2010.

Bernhard Pastötter, Karl-Heinz Bäuml, and Simon Hanslmayr. Oscillatory brain activity before and after an internal context change—evidence for a reset of encoding processes. *NeuroImage*, 43(1):173–181, 2008.

Judea Pearl. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284, 1998.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Judea Pearl and Glenn Shafer. Probabilistic reasoning in intelligent systems: Networks of plausible inference. *Synthese-Dordrecht*, 104(1):161, 1995.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 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)*, pages 2463–2473, 2019.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL `https://aclanthology.org/2020.emnlp-main.617`.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Unks everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, 2021.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, 2022.

Jean Piaget. *The Child's Conception of the World.* Harcourt, Brace & World, 1929.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022.

BigCode Project. Starcoder dataset. `https://huggingface.co/bigcode`, 2023. Accessed: 2023-12-12.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

Vijaya Raghavan T Ramkumar, Elahe Arani, and Bahram Zonooz. Learn, unlearn and relearn: An online learning paradigm for deep neural networks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=WN1O2MJDST`.

Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.

Siamak Ravanbakhsh, Barnabás Póczos, and Russell Greiner. Boolean matrix factorization and noisy completion via message passing. In *International Conference on Machine Learning*, pages 945–954. PMLR, 2016.

Michael Reed and Barry Simon. *Methods of modern mathematical physics: Functional analysis*, volume 1. Gulf Professional Publishing, 1980.

Benjamin Reichman and Larry Heck. Dense passage retrieval: Is it retrieving?, 2024. URL https://arxiv.org/abs/2402.11035.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 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)*, pages 5418–5426, 2020.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Mark R Rosenzweig. Aspects of the search for neural mechanisms of memory. *Annual review of psychology*, 47(1):1–32, 1996.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024.

Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. You can teach an old dog new tricks! on training knowledge graph embeddings. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BkxSmlBFvr.

Tomás J Ryan and Paul W Frankland. Forgetting as a form of adaptive engram cell plasticity. *Nature Reviews Neuroscience*, 23(3):173–186, 2022.

Tara Safavi and Danai Koutra. Codex: A comprehensive knowledge graph completion benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8328–8350, 2020.

Rolf Sandell. Structural change and its assessment. *International Journal of Psychology and Psychoanalysis*, 5:042, 2019. doi: 10.23937/2572-4037.1510042.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1): 61–80, 2009.

Tom Schaul and Jürgen Schmidhuber. Metalearning. *Scholarpedia*, 5(6):4650, 2010.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer, 2018.

Hans-Jörg Schmid. *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge*. American Psychological Association, 2017.

Jürgen Schmidhuber. Powerplay: Training an increasingly general problem solver by continually searching for the simplest still unsolvable problem. *Frontiers in psychology*, 4:313, 2013.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. URL `https://arxiv.org/abs/2102.11107`.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International conference on machine learning*, pages 4528–4537. PMLR, 2018.

Harshay Shah, Andrew Ilyas, and Aleksander Madry. Decomposing and editing predictions by modeling model computation. *arXiv preprint arXiv:2404.11534*, 2024.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Yifei Shen, Yongji Wu, Yao Zhang, Caihua Shan, Jun Zhang, Khaled B Letaief, and Dongsheng Li. How powerful is graph convolution for recommendation? *arXiv preprint arXiv:2108.07567*, 2021.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Herbert A Simon et al. Invariants of human behavior. *Annual review of psychology*, 41 (1):1–20, 1990.

Burrhus Frederic Skinner. *Science and human behavior*. Number 92904. Simon and Schuster, 1965.

Luca Soldaini and Kyle Lo. peS2o (Pretraining Efficiently on S2ORC) Dataset. Technical report, Allen Institute for AI, 2023. ODC-By, `https://github.com/allenai/pes2o`.

Balasubramaniam Srinivasan and Bruno Ribeiro. On the equivalence between positional node embeddings and structural graph representations. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SJxzFySKwH`.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha Talukdar, and Yiming Yang. A re-evaluation of knowledge graph completion methods. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5516–5522, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.489. URL https://aclanthology.org/2020.acl-main.489.

Zhiqing Sun, Shikhar Vashishth, Soumya Sanyal, Partha P. Talukdar, and Yiming Yang. A re-evaluation of knowledge graph completion methods. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5516–5522. Association for Computational Linguistics, 2020b. URL https://www.aclweb.org/anthology/2020.acl-main.489/.

Anej Svete and Ryan Cotterell. Transformers can represent n-gram language models. *arXiv preprint arXiv:2404.14994*, 2024.

Ahmed Taha, Abhinav Shrivastava, and Larry S Davis. Knowledge evolution in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12843–12852, 2021.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401, 2020. URL https://api.semanticscholar.org/CorpusID:220936592.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Komal K. Teru, Etienne G. Denis, and William L. Hamilton. Inductive relation prediction by subgraph reasoning. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 9448–9457. PMLR, 2020.

Jonathan Thomm, Giacomo Camposampiero, Aleksandar Terzic, Michael Hersche, Bernhard Schölkopf, and Abbas Rahimi. Limits of transformer language models on learning to compose algorithms. In *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024. URL `https://arxiv.org/abs/2402.05785`.

Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

Susumu Tonegawa, Xu Liu, Steve Ramirez, and Roger Redondo. Memory engram cells have come of age. *Neuron*, 87(5):918–931, 2015.

Susumu Tonegawa, Mark D Morrissey, and Takashi Kitamura. The role of engram cells in the systems consolidation of memory. *Nature Reviews Neuroscience*, 19(8):485–498, 2018.

Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, pages 57–66, 2015.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1174. URL `https://www.aclweb.org/anthology/D15-1174`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Eric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2071–2080, New York, New York, USA, 6 2016. PMLR. URL https://proceedings.mlr.press/v48/trouillon16.html.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha Talukdar. Composition-based multi-relational graph convolutional networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BylA_C4tPr.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems*, 29, 2016.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Petar Veličković. Message passing all the way up, 2022. URL https://arxiv.org/abs/2202.11097.

Tom Veniat, Ludovic Denoyer, and Marc'Aurelio Ranzato. Efficient continual learning with modular networks and task-driven priors. *arXiv preprint arXiv:2012.12631*, 2020.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in large language models: Dead, n-gram, positional. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 1288–1301, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.75.

Lev Vygotsky. *Thought and Language*. MIT Press, 1934.

Jiongxiao Wang, Junlin Wu, Muhao Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. On the exploitability of reinforcement learning with human feedback for large language models. *arXiv preprint arXiv:2311.09641*, 2023.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D. Goodman. Hypothesis search: Inductive reasoning with language models. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL `https://arxiv.org/abs/2309.05660`.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359, 2021. URL `https://arxiv.org/abs/2112.04359`.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, 2022.

Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, 2018.

World Wide Web Consortium (W3C). RDF 1.2 Primer, 2024. URL `https://w3c.github.io/rdf-primer/spec/`. Accessed: 2024.

Shirley Wu, Shiyu Zhao, Michihiro Yasunaga, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis Ioannidis, Karthik Subbian, James Y Zou, and Jure Leskovec. Stark: Benchmarking llm retrieval on textual and relational knowledge bases. *Advances in Neural Information Processing Systems*, 37:127129–127153, 2024.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*. OpenReview.net, 2019.

Xiaoran Xu, Wei Feng, Yunsheng Jiang, Xiaohui Xie, Zhiqing Sun, and Zhi-Hong Deng. Dynamically pruned message passing networks for large-scale knowledge graph reasoning. In *ICLR*. OpenReview.net, 2020a.

Xiaoran Xu, Wei Feng, Yunsheng Jiang, Xiaohui Xie, Zhiqing Sun, and Zhi-Hong Deng. Dynamically pruned message passing networks for large-scale knowledge graph reasoning. In *International Conference on Learning Representations*, 2020b. URL `https://openreview.net/forum?id=rkeuAhVKvB`.

Jiaqi Xue, Mengxin Zheng, Yebowen Hu, Fei Liu, Xun Chen, and Qian Lou. Badrag: Identifying vulnerabilities in retrieval augmented generation of large language models. *arXiv preprint arXiv:2406.00083*, 2024.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR (Poster)*, 2015a.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR (Poster)*, 2015b.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.550. URL `https://aclanthology.org/2024.acl-long.550`.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 40–48. JMLR.org, 2016.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.

Jingwei Yi, Rui Ye, Qisi Chen, Bin Zhu, Siheng Chen, Defu Lian, Guangzhong Sun, Xing Xie, and Fangzhao Wu. On the vulnerability of safety alignment in open-access llms. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9236–9260, 2024.

Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. L2-gcn: Layer-wise and learned efficient training of graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2127–2135, 2020.

Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. GraphSAINT: Graph sampling based inductive learning method. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=BJe8pkHFwS`.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhao Zhang, Fuzhen Zhuang, Hengshu Zhu, Zhi-Ping Shi, Hui Xiong, and Qing He. Relational graph neural network with hierarchical attention for knowledge graph completion. In *AAAI*, pages 9612–9619. AAAI Press, 2020.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2): 1–38, 2024a.

Wanru Zhao, Yihong Chen, Royson Lee, Xinchi Qiu, Yan Gao, Hongxiang Fan, and Nicholas Donald Lane. Breaking physical and linguistic borders: Multilingual federated prompt tuning for low-resource languages. In *The Twelfth International Conference on Learning Representations*, 2024b.

Hattie Zhou, Ankit Vani, Hugo Larochelle, and Aaron Courville. Fortuitous forgetting in connectionist networks. In *International Conference on Learning Representations*, 2022.

Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal A. C. Xhonneux, and Jian Tang. Neural bellman-ford networks: A general graph neural network framework for link prediction. *CoRR*, abs/2106.06935, 2021. URL `https://arxiv.org/abs/2106.06935`.

George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books, 2016.

Difan Zou, Ziniu Hu, Yewen Wang, Song Jiang, Yizhou Sun, and Quanquan Gu. Few-shot representation learning for out-of-vocabulary words. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2019.