

- Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. 2019. [NumNet: Machine reading comprehension with numerical reasoning](#). *EMNLP*.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2021. [Measuring attribution in natural language generation models](#). *arXiv preprint arXiv:2112.12870*.
- Gabriel Recchia. 2021. [Teaching autoregressive language models complex tasks by demonstration](#). *arXiv preprint arXiv:2109.02102*.
- Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. [A recipe for arbitrary text style transfer with large language models](#). *ACL*.
- Laria Reynolds and Kyle McDonell. 2021. [Prompt programming for large language models: Beyond the few-shot paradigm](#). *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*.
- Subhro Roy and Dan Roth. 2015. [Solving general arithmetic word problems](#). *EMNLP*.
- Subhro Roy, Tim Vieira, and Dan Roth. 2015. [Reasoning about Quantities in Natural Language](#). *TACL*.
- Mohammed Saeed, Naser Ahmadi, Preslav Nakov, and Paolo Papotti. 2021. [RuleBERT: Teaching soft rules to pre-trained language models](#). *EMNLP*.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. [Multitask prompted training enables zero-shot task generalization](#). *ICLR*.
- Jianhao Shen, Yichun Yin, Lin Li, Lifeng Shang, Xin Jiang, Ming Zhang, and Qun Liu. 2021. [Generate & rank: A multi-task framework for math word problems](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. [CommonsenseQA: A question answering challenge targeting commonsense knowledge](#). *NAACL*.
- Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. [Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge](#). *NeurIPS*.
- Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. [CommonsenseQA 2.0: Exposing the limits of ai through gamification](#). *NeurIPS Track on Datasets and Benchmarks*.
- Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. [Unifying language learning paradigms](#). *arXiv preprint arXiv:2205.05131*.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. [LaMDA: Language models for dialog applications](#). *arXiv preprint arXiv:2201.08239*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022a. [Self-consistency improves chain of thought reasoning in language models](#). *arXiv preprint arXiv:2203.11171*.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. [Benchmarking generalization via in-context instructions on 1,600+ language tasks](#). *arXiv preprint arXiv:2204.07705*.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. [Finetuned language models are zero-shot learners](#). *ICLR*.

- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. [Emergent abilities of large language models](#). *Transactions on Machine Learning Research*.
- Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. *NAACL*.
- Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable NLP. *NeurIPS*.
- Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. [Measuring association between labels and free-text rationales](#). *EMNLP*.
- Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022a. [PromptChainer: Chaining large language model prompts through visual programming](#). *CHI Extended Abstracts*.
- Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022b. [AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts](#). *CHI*.
- Yujun Yan, Kevin Swersky, Danai Koutra, Parthasarathy Ranganathan, and Milad Hashemi. 2020. [Neural execution engines: Learning to execute subroutines](#). *NeurIPS*.
- Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. 2021. [Refining language models with compositional explanations](#). *NeurIPS*.
- Xi Ye and Greg Durrett. 2022. [The unreliability of explanations in few-shot in-context learning](#). *arXiv preprint arXiv:2205.03401*.
- Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2021. [Few-shot out-of-domain transfer learning of natural language explanations](#). *arXiv preprint arXiv:2112.06204*.
- Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using “annotator rationales” to improve machine learning for text categorization. *NAACL*.
- Wojciech Zaremba and Ilya Sutskever. 2014. [Learning to execute](#). *arXiv preprint arXiv:1410.4615*.
- Eric Zelikman, Yuhuai Wu, and Noah D. Goodman. 2022. [STaR: Bootstrapping reasoning with reasoning](#). *arXiv preprint arXiv:2203.14465*.
- Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). *ICML*.
- Wangchunshu Zhou, Jinyi Hu, Hanlin Zhang, Xiaodan Liang, Maosong Sun, Chenyan Xiong, and Jian Tang. 2020. [Towards interpretable natural language understanding with explanations as latent variables](#). *NeurIPS*.

Checklist

1. For all authors...
 - (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[Yes]**
 - (b) Did you describe the limitations of your work? **[Yes]** See Section 6 and Appendix A.2.
 - (c) Did you discuss any potential negative societal impacts of your work? **[Yes]** We don't expect negative societal impacts as a direct result of the contributions in our paper. One consideration, however, is that generated chain of thought is not always factual, which is noted as a limitation in Appendix D.1 (and note that we do not suggest using such chains of thought in a factual manner or in any real-world scenario).
 - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[Yes]**
2. If you are including theoretical results...
 - (a) Did you state the full set of assumptions of all theoretical results? **[N/A]**
 - (b) Did you include complete proofs of all theoretical results? **[N/A]**
3. If you ran experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[Yes]** We included inputs, outputs, and targets for LaMDA and GPT-3 in the supplementary material. Although we use proprietary models, we GPT-3 results are fully reproducible. Reproducibility is further discussed in Appendix E.1.
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[Yes]** Data splits were specified, N/A for hyperparams.
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[Yes]** Standard deviation for multiple seeds using LaMDA 137B, where each seed is a different random order of exemplars, is given in Table 6 and Table 7.
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[Yes]** Type of resources are described in Appendix E.2, though we did not estimate the total amount of compute.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - (a) If your work uses existing assets, did you cite the creators? **[Yes]** We used two models that we anonymized based on the recommendation of the NeurIPS chairs. These models will be cited in the camera-ready version of the paper.
 - (b) Did you mention the license of the assets? **[Yes]** See Appendix E.3.
 - (c) Did you include any new assets either in the supplemental material or as a URL? **[Yes]** The coinflip and last letter concatenation datasets are the only new assets, and they are given in the Supplementary Materials.
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[N/A]** No human data collected.
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[N/A]** No human data collected.
5. If you used crowdsourcing or conducted research with human subjects...
 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[N/A]**
 - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[N/A]**
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[N/A]**