

- Levy, O.; Remus, S.; Biemann, C.; and Dagan, I. 2015. Do Supervised Distributional Methods Really Learn Lexical Inference Relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*.
- Li, L.; Yin, Y.; Li, S.; Chen, L.; Wang, P.; Ren, S.; Li, M.; Yang, Y.; Xu, J.; Sun, X.; et al. 2023. M3IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning. *arXiv preprint arXiv:2306.04387*.
- Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.
- Liu, T.; Xin, Z.; Chang, B.; and Sui, Z. 2020a. HypoNLI: Exploring the Artificial Patterns of Hypothesis-only Bias in Natural Language Inference. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.
- Liu, T.; Xin, Z.; Ding, X.; Chang, B.; and Sui, Z. 2020b. An Empirical Study on Model-agnostic Debiasing Strategies for Robust Natural Language Inference. In *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics.
- Lu, Q.; Qiu, B.; Ding, L.; Xie, L.; and Tao, D. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint arXiv:2303.13809*.
- McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282.
- Min, S.; Wallace, E.; Singh, S.; Gardner, M.; Hajishirzi, H.; and Zettlemoyer, L. 2019. Compositional Questions Do Not Necessitate Multi-hop Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- OpenAI. 2022. Introducing ChatGPT.
- OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.
- Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction Tuning with GPT-4. *ArXiv*, abs/2304.03277.
- Schwartz, R.; Sap, M.; Konstas, I.; Zilles, L.; Choi, Y.; and Smith, N. A. 2017. The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL)*.
- Song, Y.; Wang, P.; Zhu, D.; Liu, T.; Sui, Z.; and Li, S. 2023a. RepCL: Exploring Effective Representation for Continual Text Classification. *arXiv preprint arXiv:2305.07289*.
- Song, Y.; Xiong, W.; Zhu, D.; Li, C.; Wang, K.; Tian, Y.; and Li, S. 2023b. RestGPT: Connecting Large Language Models with Real-World Applications via RESTful APIs. *arXiv preprint arXiv:2306.06624*.
- Sun, Z.; Shen, Y.; Zhou, Q.; Zhang, H.; Chen, Z.; Cox, D. D.; Yang, Y.; and Gan, C. 2023. Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision. *ArXiv*, abs/2305.03047.
- Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *CoRR*, abs/2305.04388.
- Wang, P.; Song, Y.; Liu, T.; Lin, B.; Cao, Y.; Li, S.; and Sui, Z. 2022. Learning Robust Representations for Continual Relation Extraction via Adversarial Class Augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6264–6278. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.
- Wang, P.; Xun, R.; Liu, T.; Dai, D.; Chang, B.; and Sui, Z. 2021. Behind the Scenes: An Exploration of Trigger Biases Problem in Few-Shot Event Classification. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1969–1978.
- Wang, Y.; Ivison, H.; Dasigi, P.; Hessel, J.; Khot, T.; Chandu, K. R.; Wadden, D.; MacMillan, K.; Smith, N. A.; Beltagy, I.; et al. 2023a. How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources. *arXiv preprint arXiv:2306.04751*.
- Wang, Y.; Yu, Z.; Zeng, Z.; Yang, L.; Wang, C.; Chen, H.; Jiang, C.; Xie, R.; Wang, J.; Xie, X.; et al. 2023b. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. *arXiv preprint arXiv:2306.05087*.
- Xia, H.; Wang, P.; Liu, T.; Lin, B.; Cao, Y.; and Sui, Z. 2023. Enhancing Continual Relation Extraction via Classifier Decomposition. In *Findings of the Association for Computational Linguistics: ACL 2023*, 10053–10062. Toronto, Canada: Association for Computational Linguistics.
- Xu, C.; Guo, D.; Duan, N.; and McAuley, J. 2023. Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data. *ArXiv*, abs/2304.01196.

Yuan, W.; Neubig, G.; and Liu, P. 2021. BARTScore: Evaluating Generated Text as Text Generation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 27263–27277.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; Zhang, S.; Ghosh, G.; Lewis, M.; Zettlemoyer, L.; and Levy, O. 2023. LIMA: Less Is More for Alignment.