# 参考文献

1. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
2. Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.
3. Bai J, Bai S, Chu Y, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.
4. Zou A, Wang Z, Carlini N, et al. Universal and transferable adversarial attacks on aligned language models[J]. arXiv preprint arXiv:2307.15043, 2023.
5. Xie Y, Yi J, Shao J, et al. Defending ChatGPT against jailbreak attack via self-reminders[J]. Nature Machine Intelligence, 2023, 5: 1486–1496.
6. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[J]. Advances in neural information processing systems, 2022, 35: 27730-27744.
7. Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback[J]. arXiv preprint arXiv:2204.05862, 2022.
8. Liu Y, Deng G, Xu Z, et al. Jailbreaking chatgpt via prompt engineering: An empirical study[J]. arXiv preprint arXiv:2305.13860, 2023.
9. Liu Y, Jia Y, Geng R, et al. Formalizing and benchmarking prompt injection attacks and defenses[C]//33rd USENIX Security Symposium (USENIX Security 24). 2024: 1831-1847.
10. Qi X, Zeng Y, Xie T, et al. Fine-tuning aligned language models compromises safety, even when users do not intend to![J]. arXiv preprint arXiv:2310.03693, 2023.
11. Jain N, Schwarzschild A, Wen Y, et al. Baseline defenses for adversarial attacks against aligned language models[J]. arXiv preprint arXiv:2309.00614, 2023.
12. Alon G, Kamfonas M. Detecting language model attacks with perplexity[J]. arXiv preprint arXiv:2308.14132, 2023.
13. Liu S, Cui S, Bu H, et al. JailBench: A Comprehensive Chinese Security Assessment Benchmark for Large Language Models[J]. arXiv preprint arXiv:2502.18935, 2025.
14. Liu X, Xu N, Chen M, et al. Autodan: Generating stealthy jailbreak prompts on aligned large language models[J]. arXiv preprint arXiv:2310.04451, 2023.
15. Wang H, Li H, Huang M, et al. Asetf: A novel method for jailbreak attack on llms through translate suffix embeddings[J]. arXiv preprint arXiv:2402.16006, 2024.
16. Kiela D, Firooz H, Mohan A, et al. The hateful memes challenge: Detecting hate speech in multimodal memes[J]. Advances in neural information processing systems, 2020, 33: 2611-2624.
17. Hada R, Sudhir S, Mishra P, et al. Ruddit: Norms of offensiveness for English Reddit comments[J]. arXiv preprint arXiv:2106.05664, 2021.
18. Zampieri M, Malmasi S, Nakov P, et al. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)[J]. arXiv preprint arXiv:1903.08983, 2019.
19. Valerio Basile, Cristina Bosco, Elisabetta Fersini, et al. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter[C]// Proceedings of the 13th International Workshop on Semantic Evaluation. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019: 54–63.
20. Markov T, Zhang C, Agarwal S, et al. A holistic approach to undesired content detection in the real world[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(12): 15009-15018.
21. Xie Y, Fang M, Pi R, et al. GradSafe: Detecting Jailbreak Prompts for LLMs via Safety-Critical Gradient Analysis[J]. arXiv preprint arXiv:2402.13494, 2024.
22. Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP[C]//International conference on machine learning. PMLR, 2019: 2790-2799.
23. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
24. Wang A, Singh A, Michael J, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding[J]. arXiv preprint arXiv:1804.07461, 2018.
25. Pfeiffer J, Kamath A, Rücklé A, et al. Adapterfusion: Non-destructive task composition for transfer learning[J]. arXiv preprint arXiv:2005.00247, 2020.
26. Karimi Mahabadi R, Henderson J, Ruder S. Compacter: Efficient low-rank hypercomplex adapter layers[J]. Advances in Neural Information Processing Systems, 2021, 34: 1022-1035.
27. Sarlin P E, DeTone D, Malisiewicz T, et al. Superglue: Learning feature matching with graph neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 4938-4947.
28. Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.
29. Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[J]. arXiv preprint arXiv:2104.08691, 2021.
30. Liu X, Ji K, Fu Y, et al. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks[J]. arXiv preprint arXiv:2110.07602, 2021.
31. Zaken E B, Ravfogel S, Goldberg Y. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models[J]. arXiv preprint arXiv:2106.10199, 2021.
32. Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. ICLR, 2022, 1(2): 3.
33. Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
34. Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
35. Sanh V, Webson A, Raffel C, et al. Multitask prompted training enables zero-shot task generalization[J]. arXiv preprint arXiv:2110.08207, 2021.
36. Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners[J]. arXiv preprint arXiv:2109.01652, 2021.
37. Wang Y, Mishra S, Alipoormolabashi P, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks[J]. arXiv preprint arXiv:2204.07705, 2022.
38. Yang A, Yang B, Hui B, et al. Qwen2 technical report[J]. arXiv preprint arXiv:2407.10671, 2024.
39. Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.
40. Su J, Ahmed M, Lu Y, et al. Roformer: Enhanced transformer with rotary position embedding[J]. Neurocomputing, 2024, 568: 127063.
41. Hendrycks D, Gimpel K. Gaussian error linear units (gelus)[J]. arXiv preprint arXiv:1606.08415, 2016.
42. Shazeer N, Mirhoseini A, Maziarz K, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer[J]. arXiv preprint arXiv:1701.06538, 2017.
43. Fedus W, Zoph B, Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity[J]. Journal of Machine Learning Research, 2022, 23(120): 1-39.
44. Lepikhin D, Lee H J, Xu Y, et al. Gshard: Scaling giant models with conditional computation and automatic sharding[J]. arXiv preprint arXiv:2006.16668, 2020.
45. Du N, Huang Y, Dai A M, et al. Glam: Efficient scaling of language models with mixture-of-experts[C]//International conference on machine learning. PMLR, 2022: 5547-5569.
46. Wang W, Bi B, Yan M, et al. Structbert: Incorporating language structures into pre-training for deep language understanding[J]. arXiv preprint arXiv:1908.04577, 2019.
47. Hu S, Tu Y, Han X, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies[J]. arXiv preprint arXiv:2404.06395, 2024.
48. Huang Y, Bai Y, Zhu Z, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models[J]. Advances in Neural Information Processing Systems, 2023, 36: 62991-63010.
49. Li H, Zhang Y, Koto F, et al. Cmmlu: Measuring massive multitask language understanding in chinese[J]. arXiv preprint arXiv:2306.09212, 2023.
50. Wu X, Yao W, Chen J, et al. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning[J]. arXiv preprint arXiv:2310.00492, 2023.
51. Zhang S, Dong L, Li X, et al. Instruction tuning for large language models: A survey[J]. arXiv preprint arXiv:2308.10792, 2023.
52. Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners[J]. Advances in neural information processing systems, 2022, 35: 22199-22213.
53. Zheng Y, Zhang R, Zhang J, et al. Llamafactory: Unified efficient fine-tuning of 100+ language models[J]. arXiv preprint arXiv:2403.13372, 2024.
54. Kwon W, Li Z, Zhuang S, et al. Efficient memory management for large language model serving with pagedattention[C]//Proceedings of the 29th Symposium on Operating Systems Principles. 2023: 611-626.
55. Ran D, Liu J, Gong Y, et al. Jailbreakeval: An integrated toolkit for evaluating jailbreak attempts against large language models[J]. arXiv preprint arXiv:2406.09321, 2024.
56. Guo D, Yang D, Zhang H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning[J]. arXiv preprint arXiv:2501.12948, 2025.
57. Grattafiori A, Dubey A, Jauhri A, et al. The llama 3 herd of models[J]. arXiv preprint arXiv:2407.21783, 2024.
58. Yang A, Xiao B, Wang B, et al. Baichuan 2: Open large-scale language models[J]. arXiv preprint arXiv:2309.10305, 2023.
59. Yang A, Yang B, Zhang B, et al. Qwen2. 5 technical report[J]. arXiv preprint arXiv:2412.15115, 2024.
60. Bhatt M, Chennabasappa S, Nikolaidis C, et al. Purple llama cyberseceval: A secure coding benchmark for language models[J]. arXiv preprint arXiv:2312.04724, 2023.
61. Perez E, Huang S, Song F, et al. Red teaming language models with language models[J]. arXiv preprint arXiv:2202.03286, 2022.
62. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.
63. 李南,丁益东,江浩宇,等.面向大语言模型的越狱攻击综述[J].计算机研究与发展,2024,61(05):1156-1181.
64. 台建玮,杨双宁,王佳佳,等.大语言模型对抗性攻击与防御综述[J].计算机研究与发展,2025,62(03):563-588.