|  |  |  |
| --- | --- | --- |
| # | Title | Topics |
| 1 | Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). **Chain-of-thought prompting elicits reasoning in large language models**. *Advances in Neural Information Processing Systems*, *35*, 24824-24837. | NLP, LLMs, prompting |
| 2 | Ojha, U., Li, Y., & Lee, Y. J. (2023). **Towards universal fake image detectors that generalize across generative models**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 24480-24489). | Cyber, Deepfake detection, Encoder-decoder |
| 3 | N. Vishwamitra, et al., "**Moderating New Waves of Online Hate with Chain-of-Thought Reasoning in Large Language Models**," in 2024 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2024 pp. 178-178. doi: 10.1109/SP54263.2024.00096 url: https://doi.ieeecomputersociety.org/10.1109/SP54263.2024.00096 | Cyber, Online abuse, online hate, LLMs |
| 4 | J. Pu, et al., "**Deepfake Text Detection: Limitations and Opportunities,**" in 2023 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 2023 pp. 1613-1630. doi: 10.1109/SP46215.2023.10179387 url: https://doi.ieeecomputersociety.org/10.1109/SP46215.2023.10179387 | Cyber, Machine-generated text detection, LLMs |
| 5 | Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., ... & Chi, E. (2023, April). **Least-to-Most Prompting Enables Complex Reasoning in Large Language Models**. In *ICLR 2023-11th International Conference on Learning Representations*. | NLP, LLMs, prompting |
| 6 | Sha, Z., Li, Z., Yu, N., & Zhang, Y. (2023, November). **De-fake: Detection and attribution of fake images generated by text-to-image generation models**. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (pp. 3418-3432). | Cyber, Text-to-Image models, Diffusion, Deepfake detection |
| 7 | Lin, H., Luo, Z., Gao, W., Ma, J., Wang, B., & Yang, R. (2024, April). **Towards Explainable Harmful Meme Detection through Multimodal Debate between Large Language Models**. In *Companion Proceedings of the ACM Web Conference 2024*. | Cyber, Explainable AI, hate memes, multimodal models |
| 8 | Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., ... & Dai, J. (2024). **Visionllm: Large language model is also an open-ended decoder for vision-centric tasks**. *Advances in Neural Information Processing Systems*, *36*. | Cyber, Computer Vision, Text-to-Image models, Diffusion, Deepfake detection |
| 9 | Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). **Retrieval-augmented generation for knowledge-intensive nlp tasks.** *Advances in Neural Information Processing Systems*, *33*, 9459-9474. | NLP, LLM, RAGs |
| 10 | Feng, S., & Chen, C. (2024, February). **Prompting Is All You Need: Automated Android Bug Replay with Large Language Models.** In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering* (pp. 1-13). | Cyber, NLP , bug replay, LLMs |
| 11 | Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). **Universal and transferable adversarial attacks on aligned language models**. *arXiv preprint arXiv:2307.15043*. | Cyber, NLP, Adversarial attacks, LLMs |
| 12 | Qu, Y., Shen, X., He, X., Backes, M., Zannettou, S., & Zhang, Y. (2023). **Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models**. *arXiv preprint arXiv:2305.13873*. | Cyber, Computer Vision, unsafe images, diffusion, text-to-image |