CS 839: Foundation Models
HW 3: Training Data Extraction and Multi-Modality

Nick Boddy

2024-11-18

All code and related files can be found on GitHub: `https://github.com/Nick-Boddy/CS839-HW3`

1 Training Data Extraction

1.1 Large Language Models' Training Data

What sorts of data are used to train LLMs like ChatGPT?

1.2 Manual Extraction

1.3 Automated Techniques

1.4 Implementation Details

2 Multimodal Model Limitations

3 Conclusion