BERTology in the Modern World

Michael Li[♠] **Nishant Subramani**[♠]
[♠]Carnegie Mellon University - Language Technologies Institute
{ml6, nishant2}@cs.cmu.edu

Abstract

Large transformer-based language models dominate modern NLP, yet our understanding of how they encode linguistic information is rooted in studies of early models like BERT and GPT-2. To better understand today's language models, we investigate how both classical architectures (BERT, DeBERTa, GPT-2) and contemporary large language models (Pythia, OLMo-2, Gemma-2, Qwen2.5, Llama-3.1) represent lexical identity and inflectional morphology. We train linear and nonlinear classifiers on layer-wise activations to predict word lemmas and inflectional features. We discover that models concentrate lexical information linearly in early layers and increasingly nonlinearly in later layers, while keeping inflectional information uniformly accessible and linearly separable throughout the layers. Further analysis reveals that these models encode inflectional morphology through generalizable abstractions, but rely predominantly on memorization to encode lexical identity. Remarkably, these patterns emerge across all 16 models we test, despite differences in architecture, size, and training regime (including pretrained and instruction-tuned variants). This consistency suggests that, despite substantial advances in LLM technologies, transformer models organize linguistic information in similar ways, indicating that these properties could be fundamental for next token prediction and are learned early during pretraining. Our code is available at https://github.com/ml5885/model_internal_sleuthing