# Causal Interventions Reveal Shared Structure Across English Filler–Gap Constructions

**Sasha Boguraev**[1] **Christopher Potts**[2] **Kyle Mahowald**[1]
[1]The University of Texas at Austin [2]Stanford University
{sasha.boguraev,kyle}@utexas.edu cgpotts@stanford.edu

## Abstract

Large Language Models (LLMs) have emerged as powerful sources of evidence for linguists seeking to develop theories of syntax. In this paper, we argue that causal interpretability methods, applied to LLMs, can greatly enhance the value of such evidence by helping us characterize the abstract mechanisms that LLMs learn to use. Our empirical focus is a set of English filler–gap dependency constructions (e.g., questions, relative clauses). Linguistic theories largely agree that these constructions share many properties. Using experiments based in Distributed Interchange Interventions, we show that LLMs converge on similar abstract analyses of these constructions. These analyses also reveal previously overlooked factors – relating to frequency, filler type, and surrounding context – that could motivate changes to standard linguistic theory. Overall, these results suggest that mechanistic, internal analyses of LLMs can push linguistic theory forward.