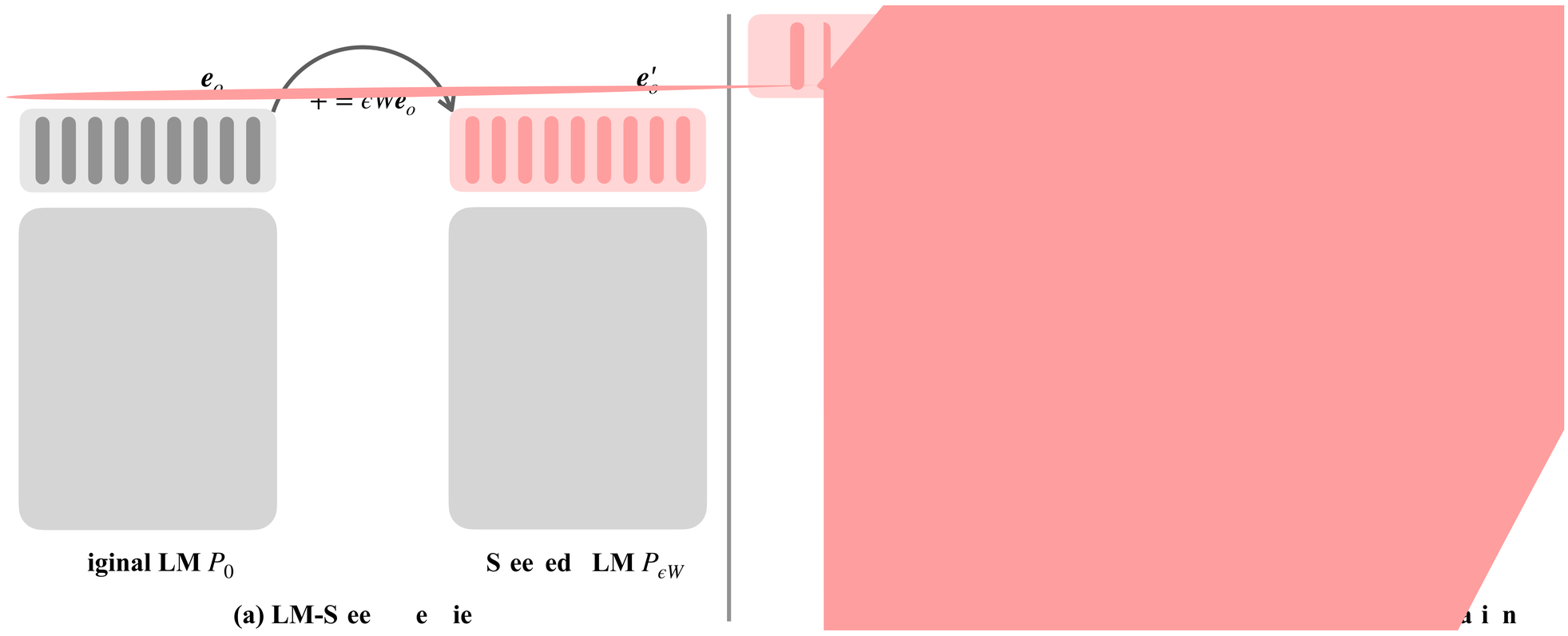
LM-Steer: Word Embeddings Are Steers for Language Models

Chi Han, Jialiang Xu, Manling Li,
Yi Fung, Chenkai Sun, Nan Jiang,
Tarek Abdelzaher, Heng Ji

paper codes author



(a) LM-S ee e ie