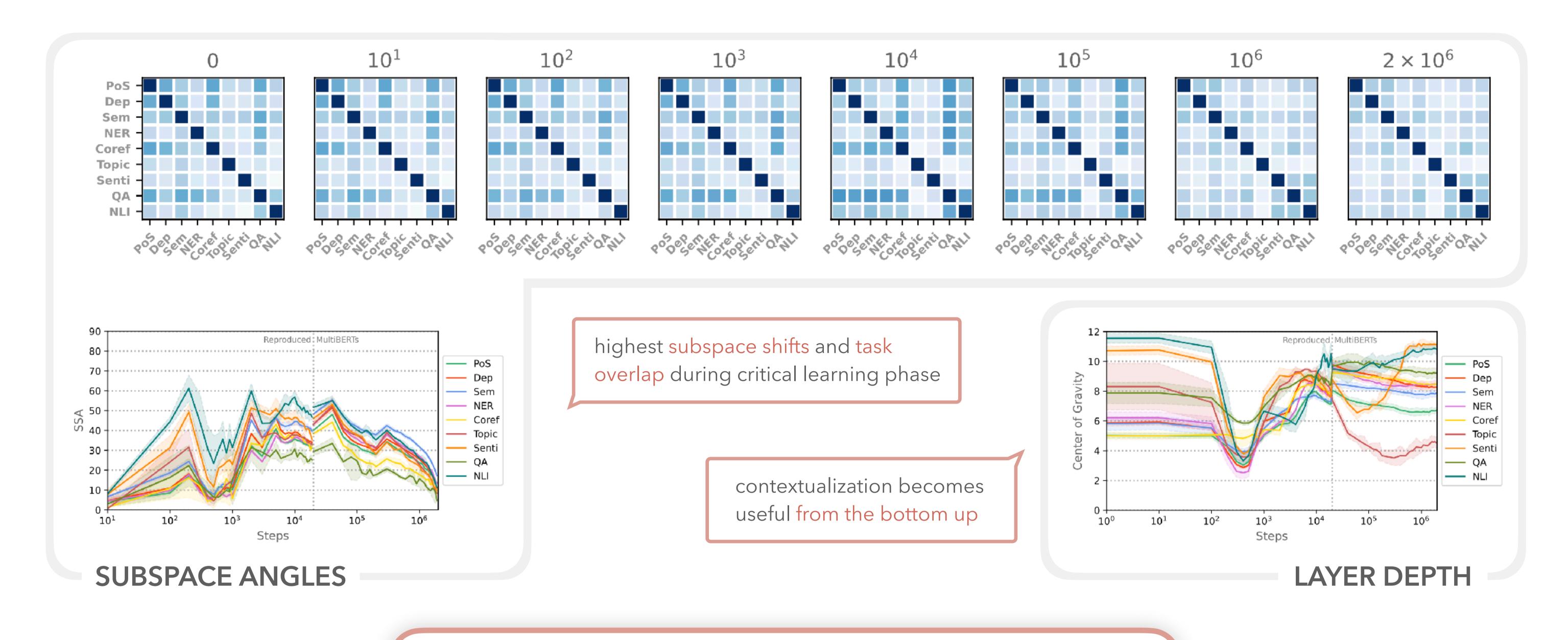


How Linguistic Information Emerges, Shifts and Interacts during Language Model Training



Task_A Treating the probe θ as $L = -\mathbb{E}_{\theta \sim \beta} \left[\sum_{x,y \in \mathcal{D}} \log_2 p_{\theta}(y|x) \right]$ a subspace within which Task_B **F1**₂ task-relevant information $+ \text{KL}(\beta||\gamma)$ Task_C is particularly salient allows for comparisons Information-Theoretic Probing across tasks and time. (Voita and Titov, 2020) LM_0 LM_1 LM_2 • • • **Pre-training Corpus**

