EMNLP2023 Best Paper "Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning"

我觉得这个文章好的一点是他完整呈现了一个研究问题的思路，从提出假设到验证假设，再根据假设设计新的方法和错误分析。文章其实很好理解，比较容易读懂，大概就是下面的结构。

1. Hypothesis

Information Flow with Labels as Anchors

H1: In **shallow layers**, label words **gather the information of demonstrations** to form semantic representations for deeper layers. 浅层将上下文信息聚合到label words representation

H2: In **deep layers**, the model **extracts the information from label words** to form the final prediction. 深层利用聚合后的representation进行最终的预测

2. Hypothesis verification

Methods1: saliency technique (Simonyan et al., 2013) Attention可视化

Methods2: isolate the label words in different layers and observe the performance

Results analysis: shallow layers and deep layers

3. Proposed new methods

Anchor Reweighting: Add a trainable reweighting vector and train the parameters on an auxiliary training set.

Anchor-Only Context Compression: By concatenating at the front in each layer during inference, instead of using the full demonstration, we can speed up inference.

Anchor Distances for Error Diagnosis: extract the components of the key vectors along the directions with significant variations in qq,