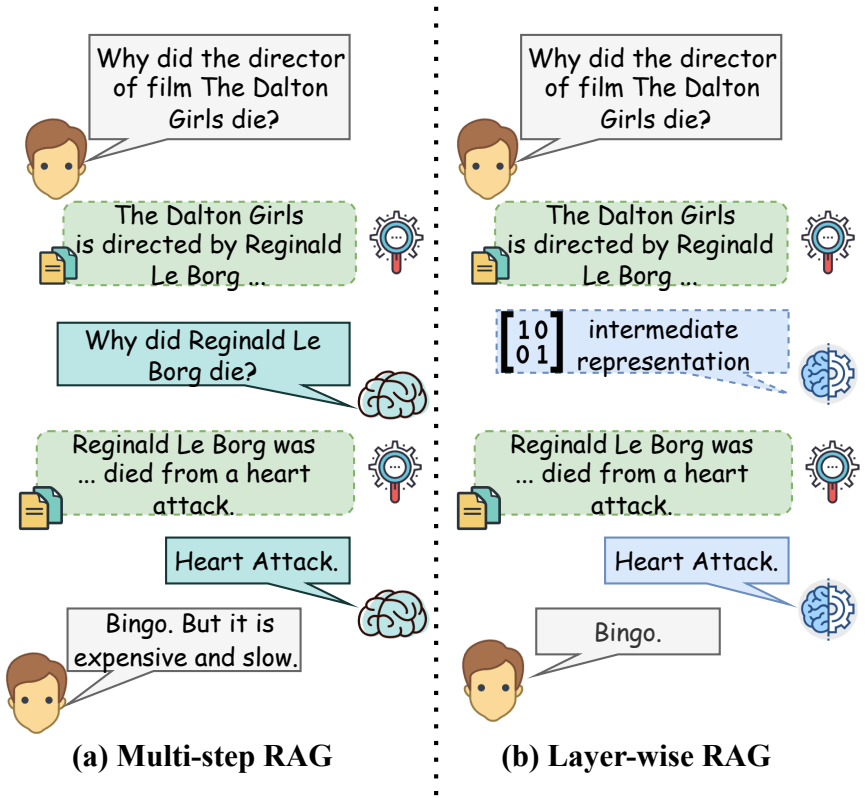
Optimizing Multi-Hop Document Retrieval Through Intermediate Representations



Jiaen Lin, Jingyu Liu

Tsinghua University

Motivation



Q1: How can we build a efficient and effective retriever?

A1: Use the information contained in intermediate representation.

Q2: How can we determine that the intermediate representation has more information?

Figure 1: Motivation

Model Analysis



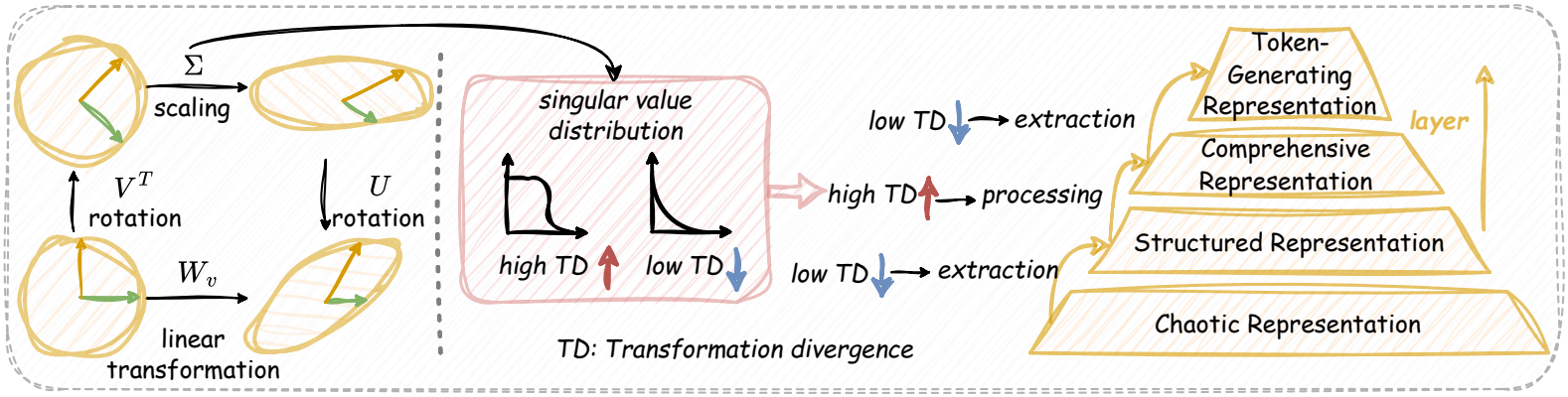


Figure 2: Weight matrix transformation divergence (TD) and information processing in LLM reasoning.

$$
\begin{aligned}
\boldsymbol{W}_v &= \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T \\
&= \boldsymbol{u}_1\sigma_1\boldsymbol{v}_1^T + \cdots + \boldsymbol{u}_r\sigma_r\boldsymbol{v}_r^T \\
&= \sigma_1\boldsymbol{u}_1\boldsymbol{v}_1^T + \cdots + \sigma_r\boldsymbol{u}_r\boldsymbol{v}_r^T,
\end{aligned} \tag{1}
$$

Definition (Transformation Direction)

$$\text{Direction}(\boldsymbol{W}_v) = [\boldsymbol{u}_1\boldsymbol{v}_1^T, \ldots, \boldsymbol{u}_r\boldsymbol{v}_r^T]$$

Definition (Transformation Divergence)

$$p_i = \frac{\sigma_i}{\sum_{j=1}^{r} \sigma_j} \tag{2}$$

$$\text{TD}(\boldsymbol{W}_v) = -\sum_{i=1}^{r} p_i \log p_i, \tag{3}$$

where r denotes the rank of $\boldsymbol{W}_v$.

Transformation Divergence Test

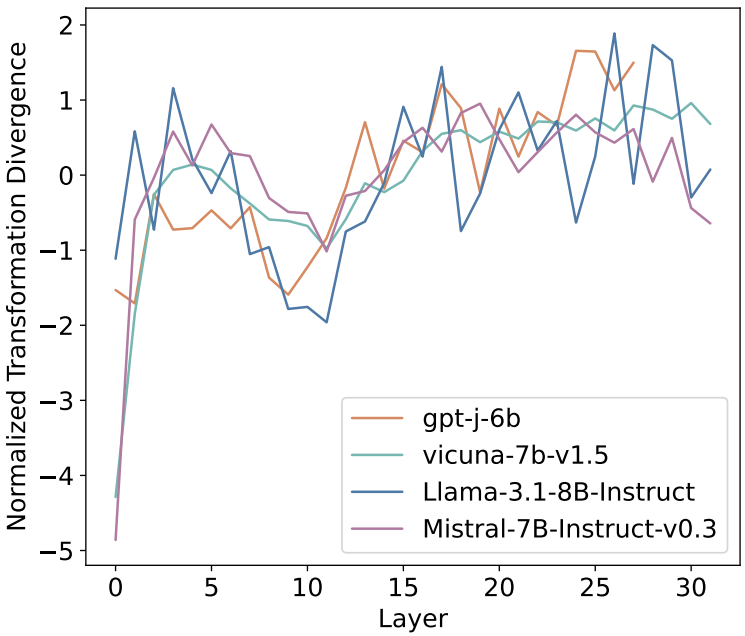


Figure 3: Transformation Divergence TD($\boldsymbol{W}_v$) of the weight matrix across various models.

- **initial low** TD($\boldsymbol{W}_v$) **phase**: LLMs extract information from disorganized data
- **subsequent high** TD($\boldsymbol{W}_v$) **phase**: processing these information
- **final low** TD($\boldsymbol{W}_v$) **phase**: aligns the representation with the token prediction requirements

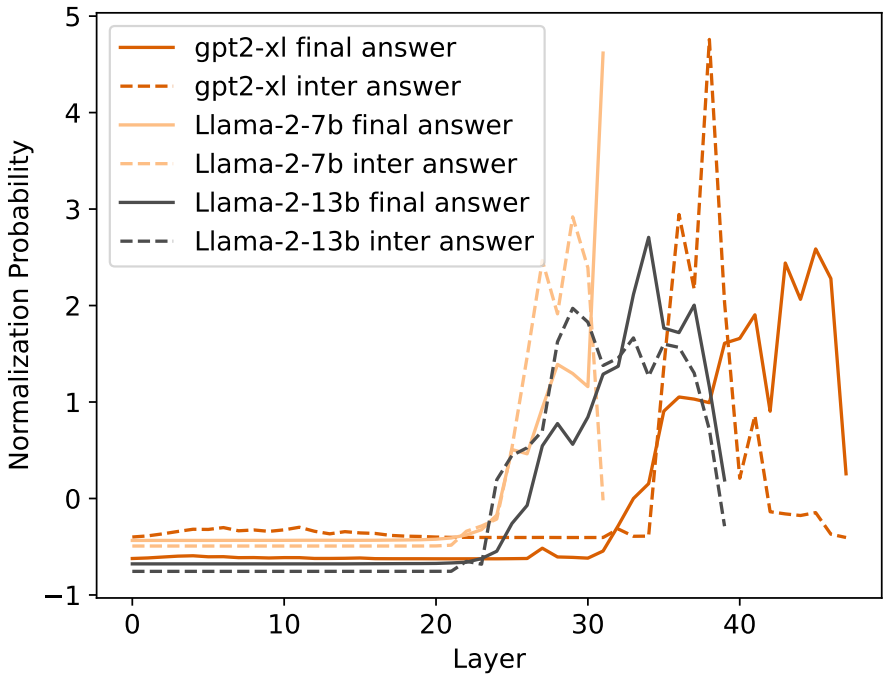


Figure 4: Logit Lens analysis conducted at multiple layers within different LLMs. Internal answer always occurs earlier than final answer.

Method



Layer-wise RAG(L-RAG)

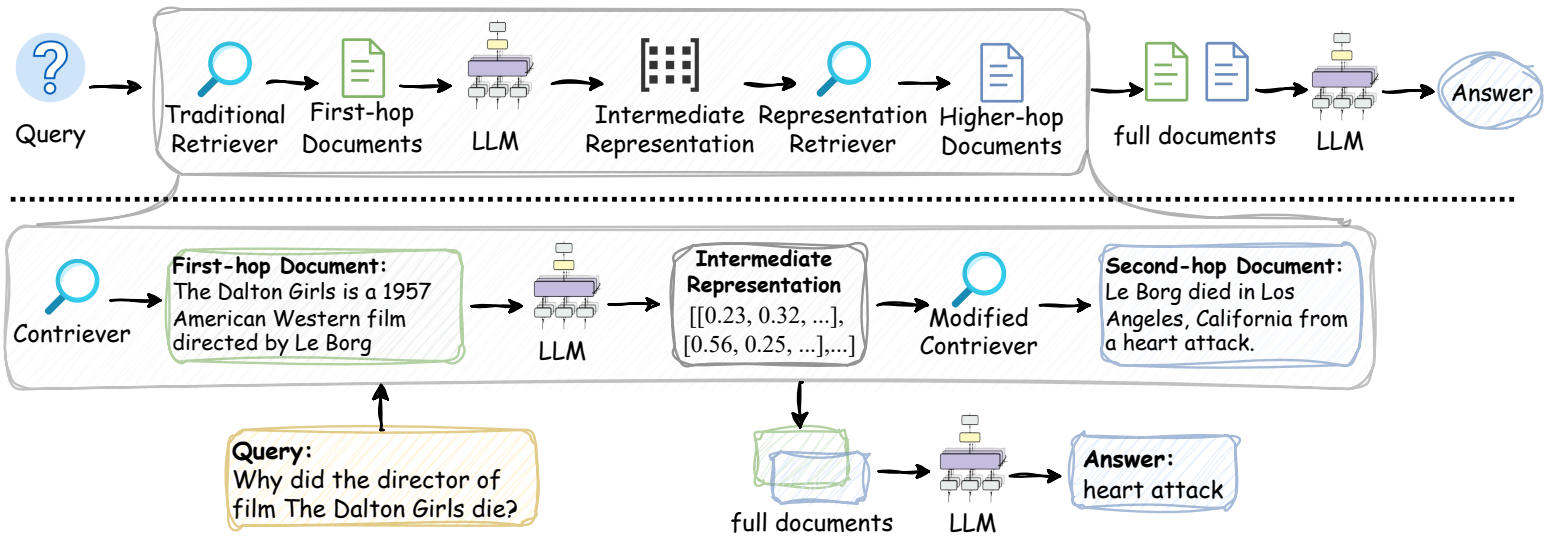


Figure 5: An overview and example of the L-RAG framework, which consists of three components: a traditional retriever, an LLM for generating intermediate representations, and the representation retriever.

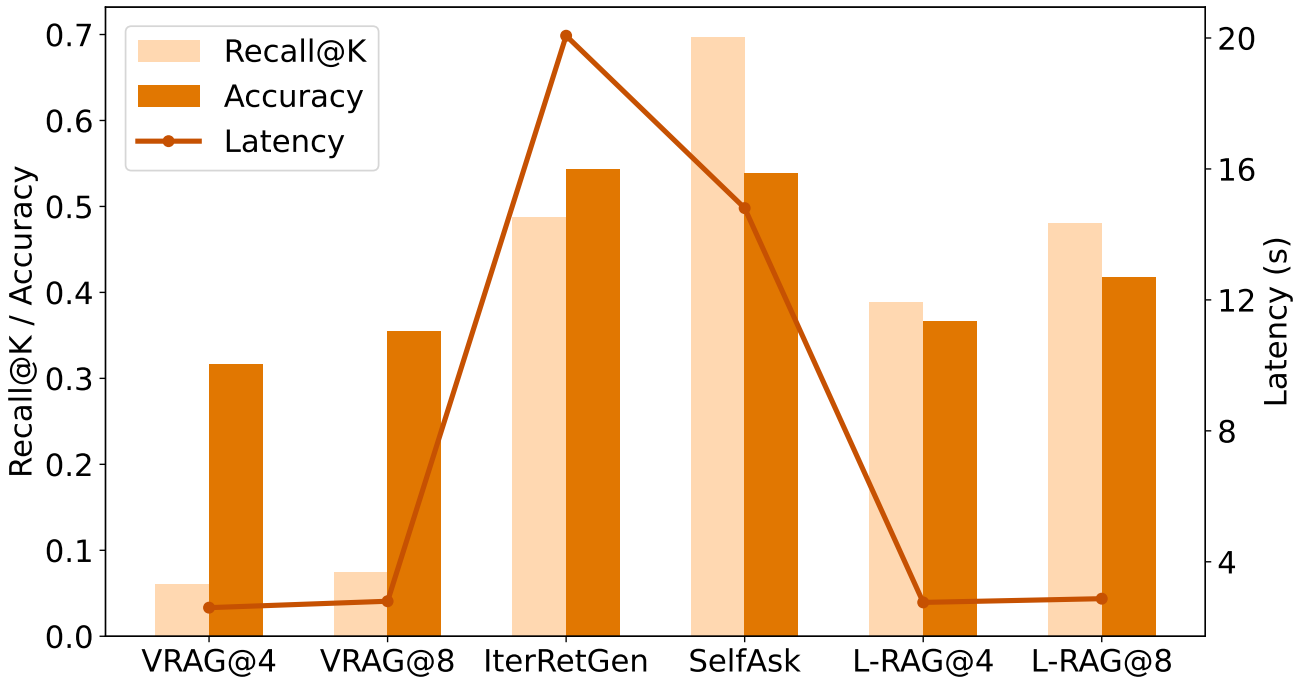


Figure 6: The recall rate and accuracy of different methods in the 2WikiMQA dataset. VRAG stands for Vanilla RAG.