中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

How do Large Language Models Understand Relevance? A Mechanistic Interpretability Perspective

Qi Liu
qiliu6777@gmail.com
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China

Jiaxin Mao*
maojiaxin@gmail.com
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China

Ji-Rong Wen
jrwen@ruc.edu.cn
Gaoling School of Artificial Intelligence
Renmin University of China
Beijing, China

arxiv: 25.04

于璐璐 25.06.10

QAs

1. 想探究模型评估相关性时不同模块的作用，是看对具体的**一个模块**做activation patching带来的影响。但是对该模块做activation patching后，该模块前向传播路径中的所有神经元的输出都会发生改变，这种分析方式是对这一个模块分析吗？

- 对一个模块做activation patching相当于模型编辑，编辑源头还是这个模块，所以是对这一个模块进行分析；

2. 如果同时对多个模块做activation patching，前向传播时彼此的干预会揉在一起，这种情况下起到什么作用？

3. 对比pointwise和pairwise两种prompt方式是否具有相同的模型可解释性机制这里，用序的一致性指标RBO可能不太合适，因为可能实际值上的差异很小；

4. evaluation on downstream tasks这部分加的document reranking任务，能反映出前边indirect effect得出的起重要作用的头，确实能在相关性评估上起到作用。

5. 训练数据里相关性方面的数据少，LLMs为什么具有相关性评估的能力？

- LLMs语言理解能力为相关性评估打底（相关性的含义 & 查询和文档的语义理解）

» Motivation

How do LLMs understand and operationalize the concept of relevance?

- Traditional relevance judgment:
  - BM25 -> term frequency & term weight
- LLMs can assess relevance.
  - The underlying mechanisms remain opaque.

=> **How different LLM modules (attention blocks & MLPs) contribute to relevance judgment?**

  - through mechanistic interpretability techniques: activation patching

Conclusion

1. LLMs may process and transmit information in a **progressive manner**.

2. The mechanism of **assessing relevance** within LLMs may serve as a **universal mechanism** and independent of specific prompts or tasks

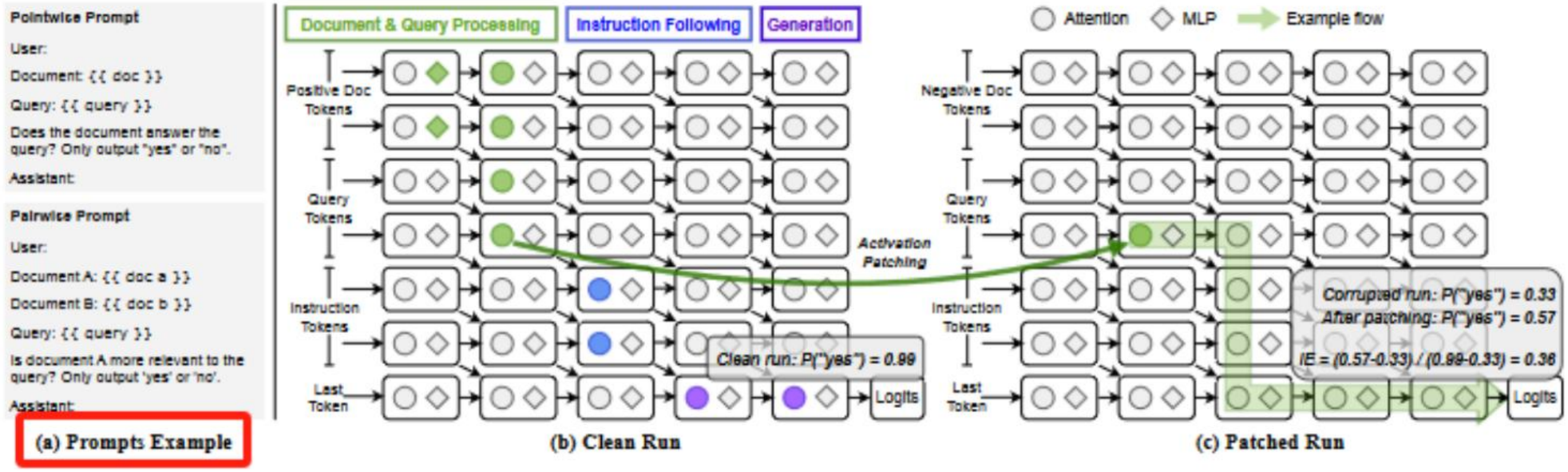
- 问题
  - 论文使用了activation patching的方式看评估相关性时LLMs本身哪些部分起到的作用，如果对LLMs进行微调，通过影响哪些部分影响了其评估相关性的能力
  - 很多实验结果，但是不太足以支撑结论，并且缺乏一些原因分析

Activation Patching

Visualization of an activation patching example

- Prompting LLMs to relevance judgment: pointwise, pairwise

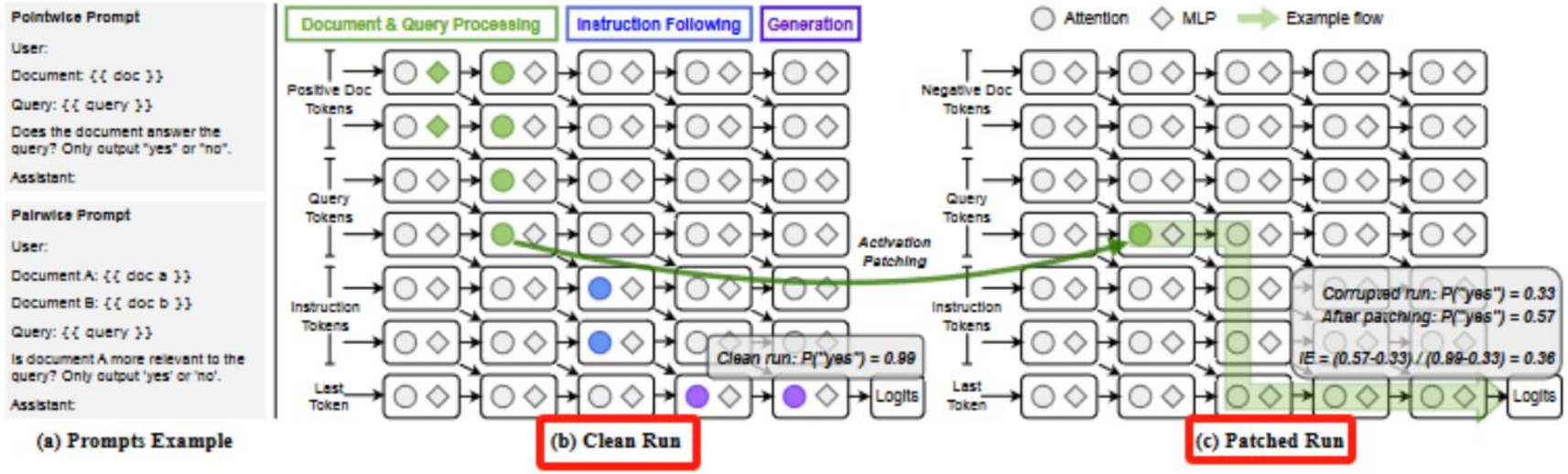
=> To analyze the differences in model behavior across various prompt formats.



Activation Patching

Visualization of an activation patching example

- **Activation patching procedure:**
  - Clean run: run the model on X_{clean} => doc / doc_a is **positive** [pointwise / pairwise]
  - Corrupted run: run the model on $X_{corrupted}$ => doc / doc_a is **negative** [pointwise / pairwise]
  - Patched run: run the model on $X_{corrupted}$ with a **specific activation** restored from the **clean run**

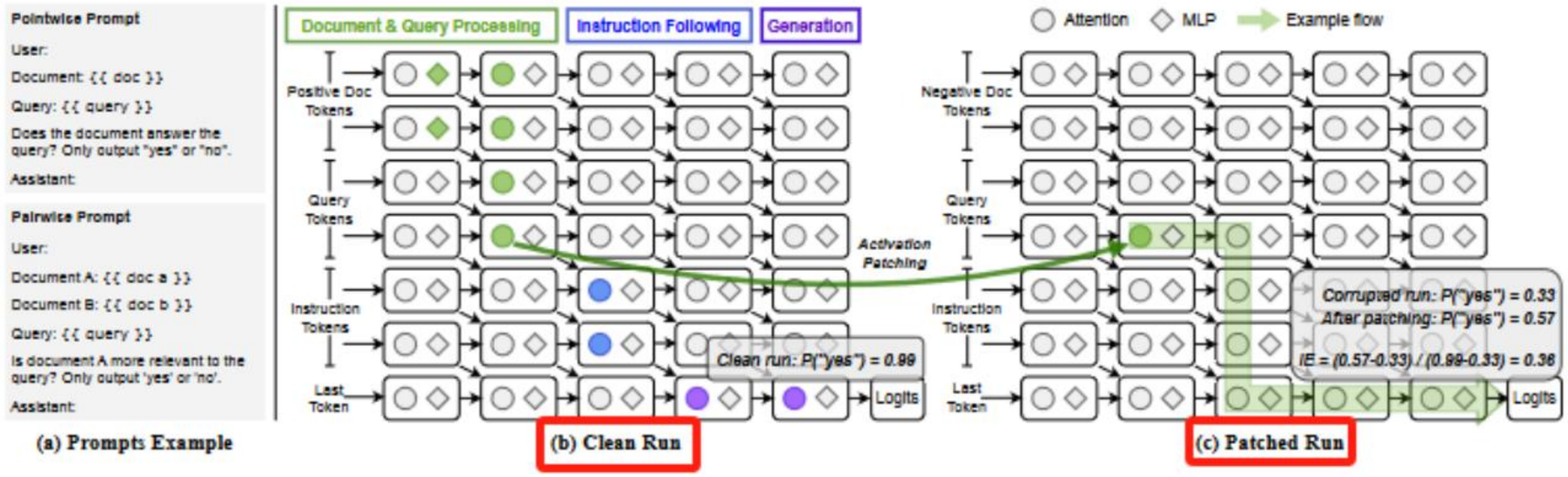


Activation Patching

Visualization of an activation patching example

- **Metrics**
  - log difference (LD) = Logits(“yes”) - Logits(“no”)
  - indirect effect (**IE**) = $\frac{LD_{patched} - LD_{corrupted}}{LD_{clean} - LD_{corrupted}}$

=> determine which component contributes more to the model behavior



Experimental Setup

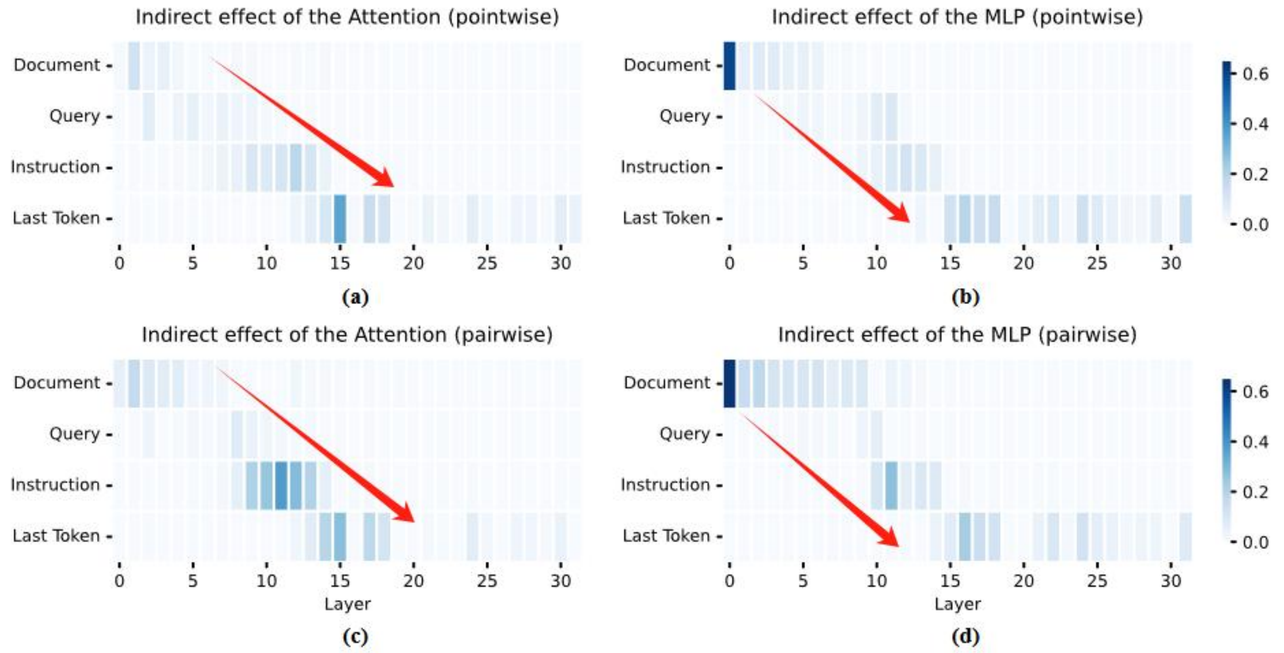
Models & Datasets

- **Models**
  - main: Llama3.1-8B-Instruct
  - supplementay: Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3
- **Datasets**
  - main: MS MARCO passage ranking dataset
  - supplementary: Natural Questions (NQ)
  - query num: 100 samples
  - positive & negative
    - **positive**: human-labeled; **negative**: sampled from top 100 documents retrieved by BM25

Experiments

Tracing information flow

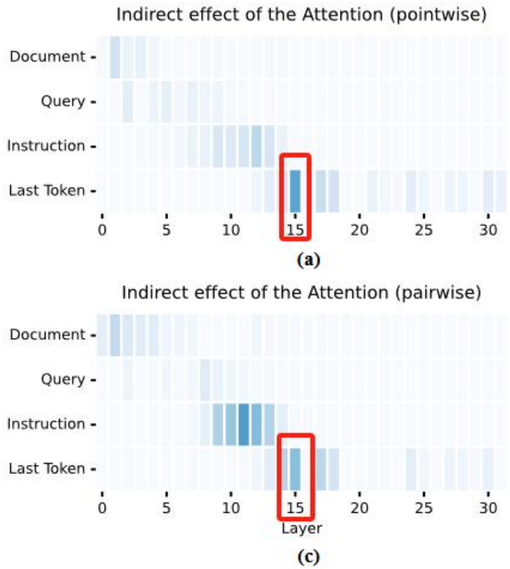
- The model processes and transmits information progressively.
  - **early layers**: extract query and document information
  - **middle layers**: process relevance information according to instructions
  - **later layers**: generate relevance judgments



Experiments

Tracing information flow

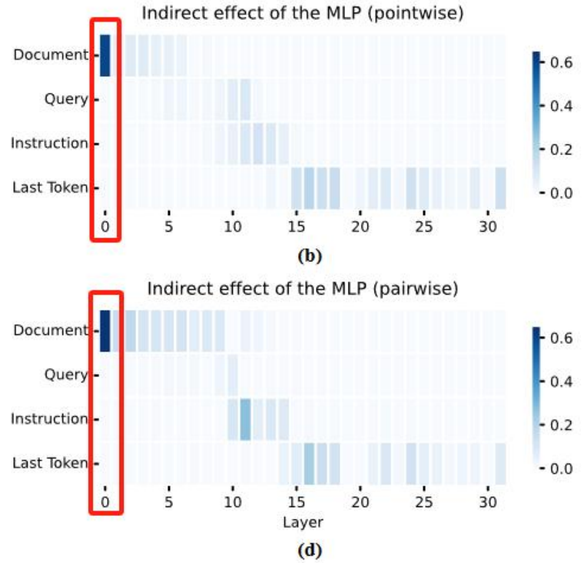
- Indirect effect of the **Attention**
  - **at layer 15**, transition towards the formulation of the final output



For MLPs, the influence of **query and instruction** tokens is lower than attention blocks.

Attention blocks seem pivotal for modeling complex interactions.

- Indirect effect of the **MLP**
  - **Early layers** store knowledge or semantic information about **documents**

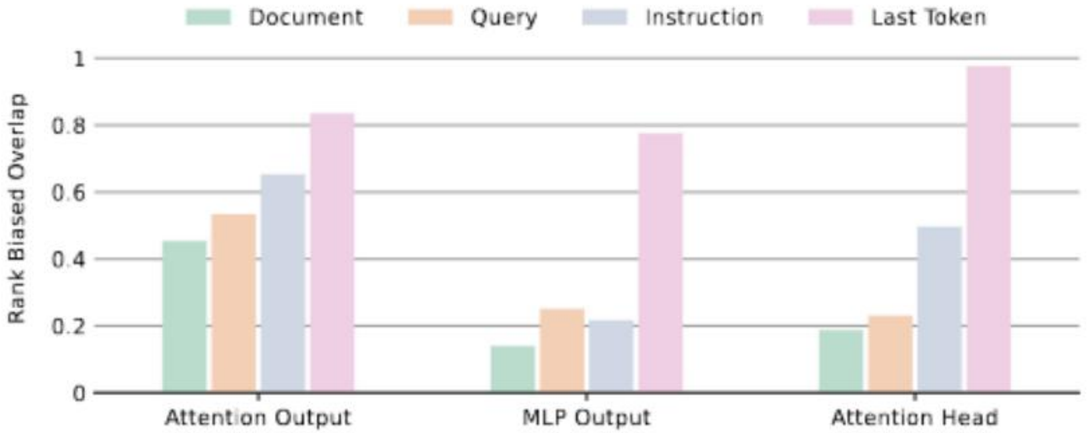


Experiments

Tracing information flow

- **pointwise v.s. pairwise**
  - To further quantify the **similarity**, compute the Rank Bias Overlap (RBO) [RBO=0 <=> disjoint]
  - For attention blocks: significant similarity
  - For MLPs: lower RBO may be attributed to **discrepancies** in rankings towards the **end**

=> Large language models possess a universal mechanism for assessing relevance internally.



Experiments

Analysis of individual attention heads

- conduct activation patching specifically on the outputs of **individual attention heads**

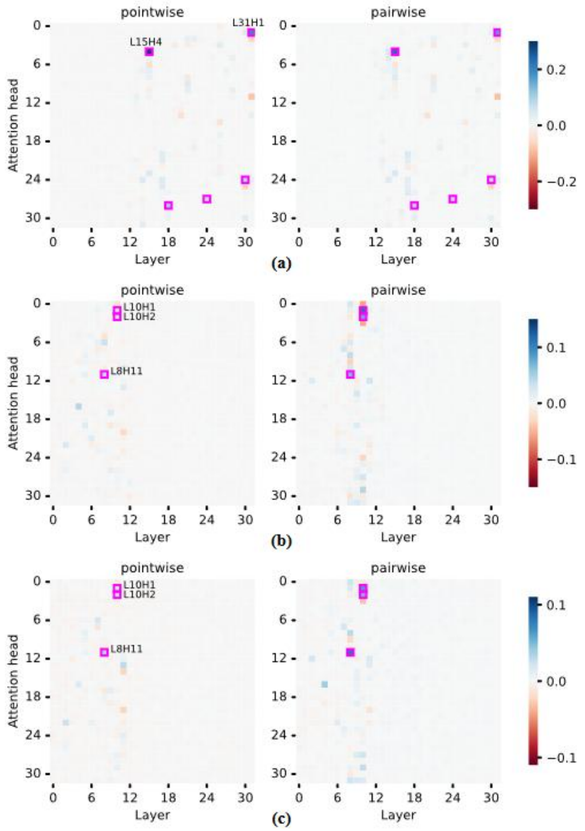


Figure 5: Indirect effect of individual head. (a) Head's output at last token. (b) Heads' output at the position of the query. (c) Heads' attention scores at the position of query-document. Several heads with the highest effects are highlighted in pink.

1. Investigate which attention heads significantly influence the **final output**.
   - highly sparse & particularly in the earlier layers, being nearly zero.
   - L15H4 & **L31H1**: critical components in controlling the output of "yes"?
     - compute $W_U a^{(h,l)} + b$ => vocab_dim => **check top token ("yes"?)**
       - **"yes" consistently exhibits the highest logit in L31H1.**

Experiments

Analysis of individual attention heads

- conduct activation patching specifically on the outputs of **individual attention heads**

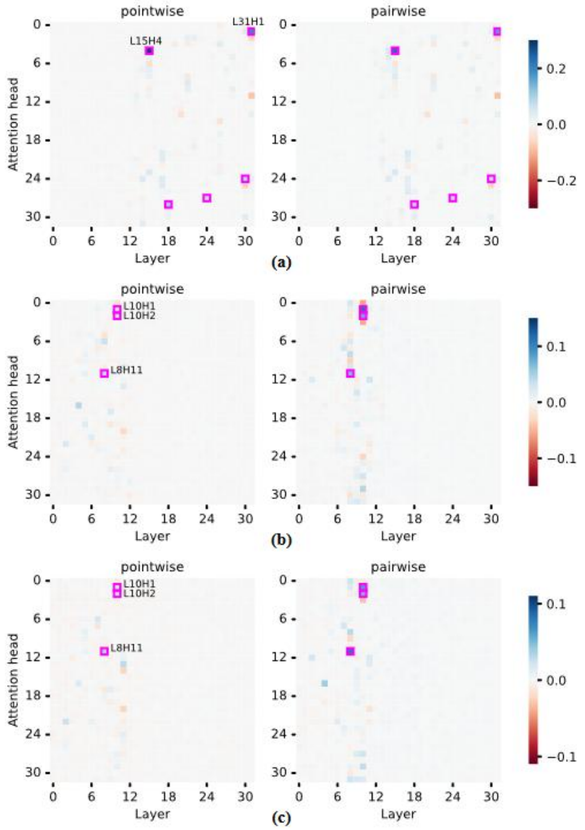


Figure 5: Indirect effect of individual head. (a) Head's output at last token. (b) Heads' output at the position of the query. (c) Heads' attention scores at the position of query-document. Several heads with the highest effects are highlighted in pink.

2. **Ascertain** whether there are any components employed to process **relevance signals.**

- conduct **two activation patching experiments**
  - patch the output of individual attention heads at the position of query
  - patch the attention scores where the query (i.e., target position) attends to the document (i.e., source position)

=> **similar distribution** of indirect effect

Experiments

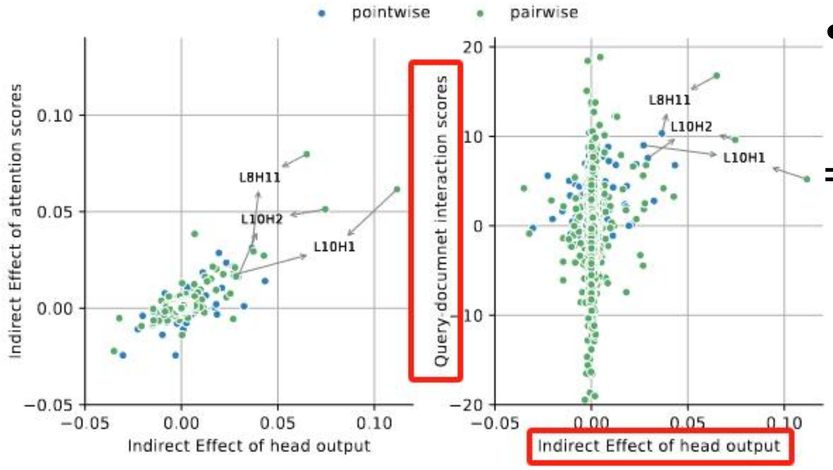
Analysis of individual attention heads

- conduct activation patching specifically on the outputs of **individual attention heads**

2. **Ascertain** whether there are any components employed to process **relevance signals.**

- use a **metric, attention interaction score,** to measure the interaction between query and document

$$s^{(l,j)}(q,d) = \sum_{i \in I_q} \max_{k \in I_d} \overbrace{A_{i,k}^{(l,j)}}^{\text{attention weight}} \quad \Rightarrow \quad S^{(l,h)} = \left(s^{(l,h)}(q, d_{pos}) - s^{(l,h)}(q, d_{neg}) \right)$$



- Those heads with the **highest effect** (L8H11 and L10H2) still **obtain high attention interaction scores.**

=> **several attention heads** engaged in processing relevance signals

Experiments

Generalize to different datasets and models

- **For different datasets**, from low to deep layers, the trend of effect changes is very similar for different components and positions.

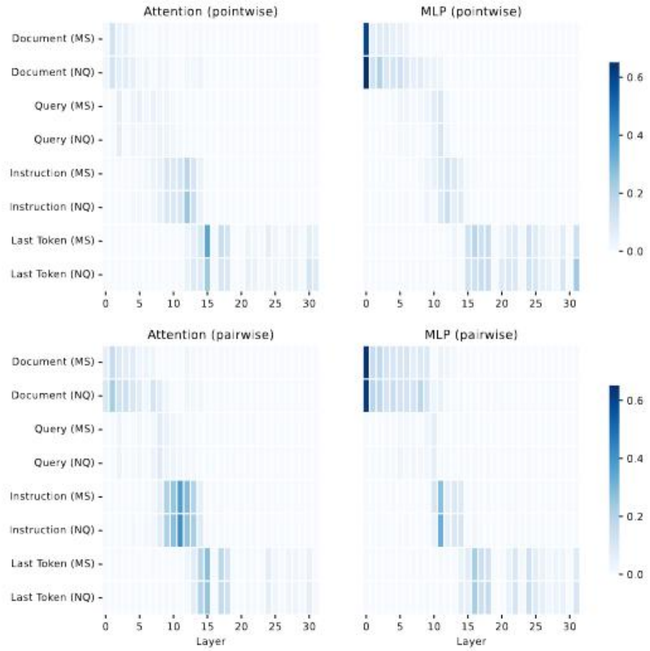
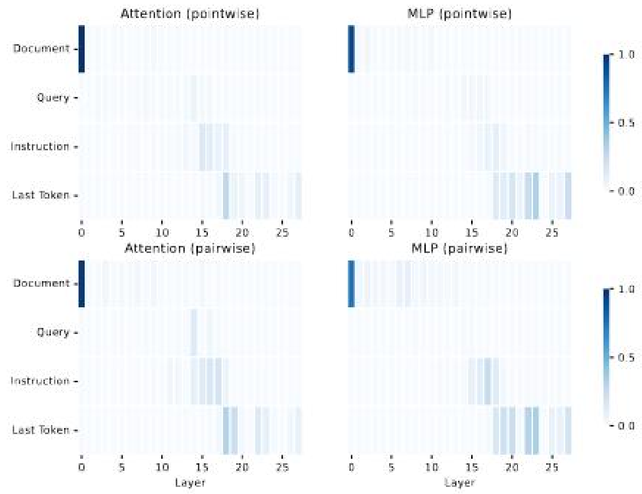
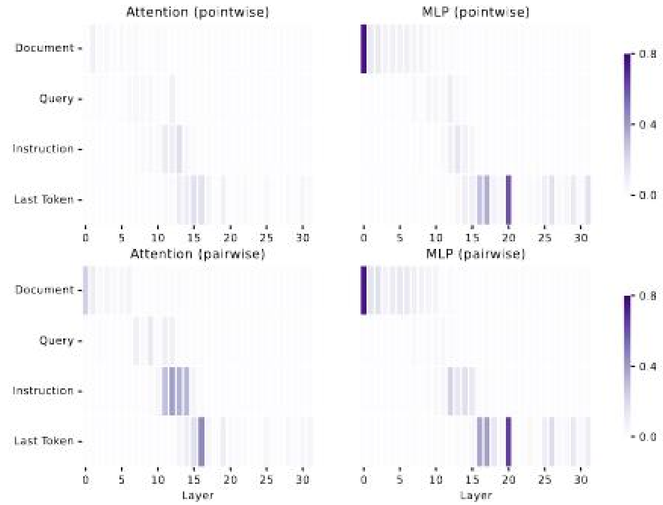


Figure 7: Indirect Effect of different components at different token positions within LLama-3.1-8B-Instruct on both MS MARCO (MS in short) and NQ. The results at the same token positions are placed on adjacent lines for easy comparison.

- **For different LLMs**, there are **similar mechanisms** within different large language models **for relevance judgment.**



Qwen2.5-7B-Instruct



Mistral-7B-Instruct-v0.3

Experiments

Evaluation on downstream tasks

- To **investigate** whether these observed components **are all important in different IR tasks** that is related to relevance assessment.
  - Evaluation experiments (at the level of attention head outputs) on downstream tasks:
    - relevance judgment & document reranking.
- Method: knockout technique => **mean ablation**
  - replace the activation with the mean activation value

Experiments

Evaluation on downstream tasks

Table 1: Mean ablation results of Llama-3.1-8B-Instruct on relevance judgment (F1-score as the metric) and document reranking (NDCG@10 as the metric, the NDCG@10 of the first-stage retrieval result is 0.51). The performance decrease ratio compared to the full model is indicated in parentheses.

	Relevance Judgment		Reranking	
	Pointwise	Pairwise	Pointwise	Pairwise
Full model	0.91	0.86	0.62	0.62
- Random-80	0.91 (-0.0%)	0.85 (-1.2%)	0.60 (-3.2%)	0.61 (-1.6%)
- Doc-20	0.90 (-1.1%)	0.85 (-1.2%)	0.60 (-3.2%)	0.60 (-3.2%)
- Query-20	0.81 (-11.0%)	0.81 (-4.7%)	0.58 (-6.5%)	0.58 (-6.5%)
- Inst-20	0.78 (-14.3%)	0.68 (-20.0%)	0.59 (-4.8%)	0.57 (-8.1%)
- Last-20	0.55 (-39.6%)	0.62 (-27.1%)	0.56 (-9.7%)	0.55 (-11.3%)
- Mixed-80	0.47 (-48.4%)	0.50 (-41.2%)	0.51 (-17.7%)	0.52 (-16.1%)

- Random-80: ablate 80 heads at all token positions that are **randomly sampled**
- Document-20: ablate 20 heads with the highest indirect effect at the **document positions**
- Mixed-80: use **all four types of top heads** listed above

- Knockout less than one-tenth of attention heads can completely render the model ineffective in relevance-related tasks.
- Ablation on the last token has the greatest impact on performance <= this position is directly related to the model output.

Conclusion

1. LLMs may process and transmit information in a **progressive manner**.

2. The mechanism of **assessing relevance** within LLMs may serve as a **universal mechanism** and independent of specific prompts or tasks

- 问题
  - 论文使用了activation patching的方式看评估相关性时LLMs本身哪些部分起到的作用，如果对LLMs进行微调，通过影响哪些部分影响了其评估相关性的能力
  - 很多实验结果，但是不太足以支撑结论，并且缺乏一些原因分析

感谢大家耐心倾听！