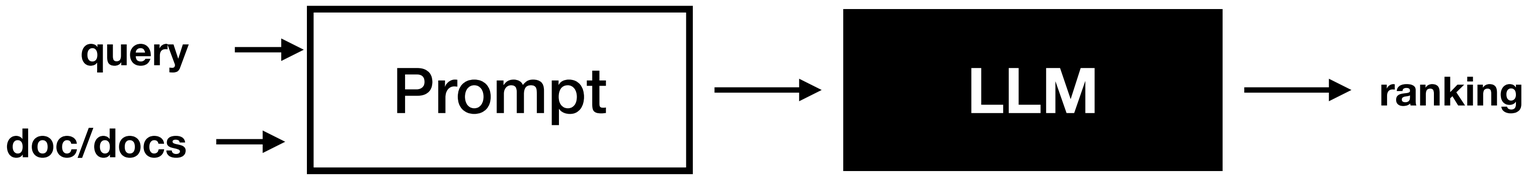
An Investigation of Prompt Variations for Zero-shot LLM-based Rankers

Shuoqi Sun[1*], Shengyao Zhuang[2], Shuai Wang[3], Guido Zuccon[3]

[1] RMIT University, Australia
[2] CSIRO, Australia
[3] The University of Queensland, Australia

[*] This work was conducted while Shuoqi Sun was a student at The University of Queensland.

LLM Rankers: Prompting LLMs to Rank Documents

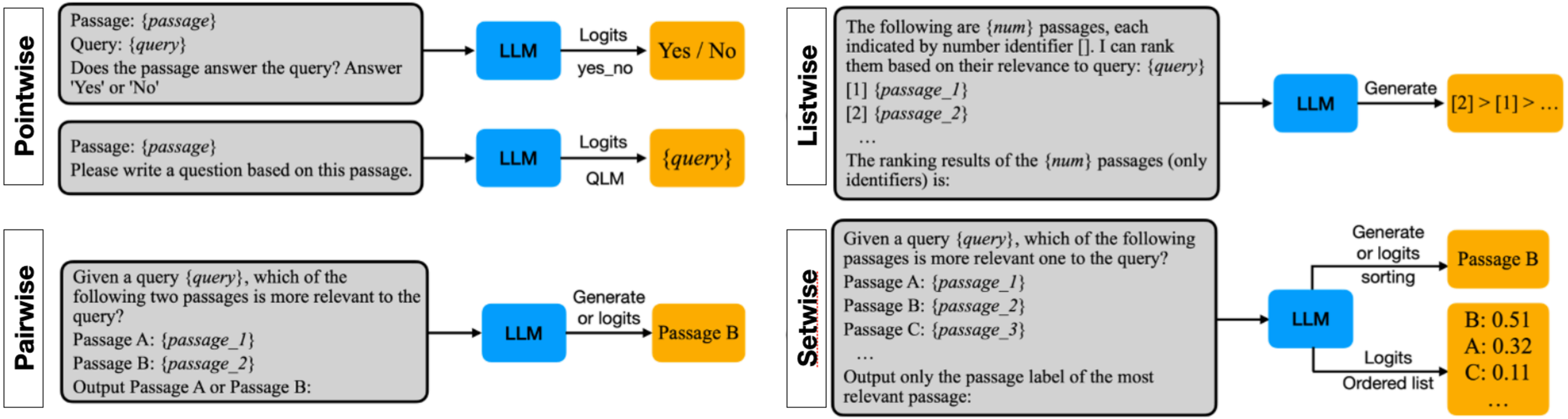


- All are “zero-shot”: i.e. once we obtained the pre-trained, instruction tuned LLM, no need to do SFT or RL for specific ranking task

  - There has no examples provided in the prompt

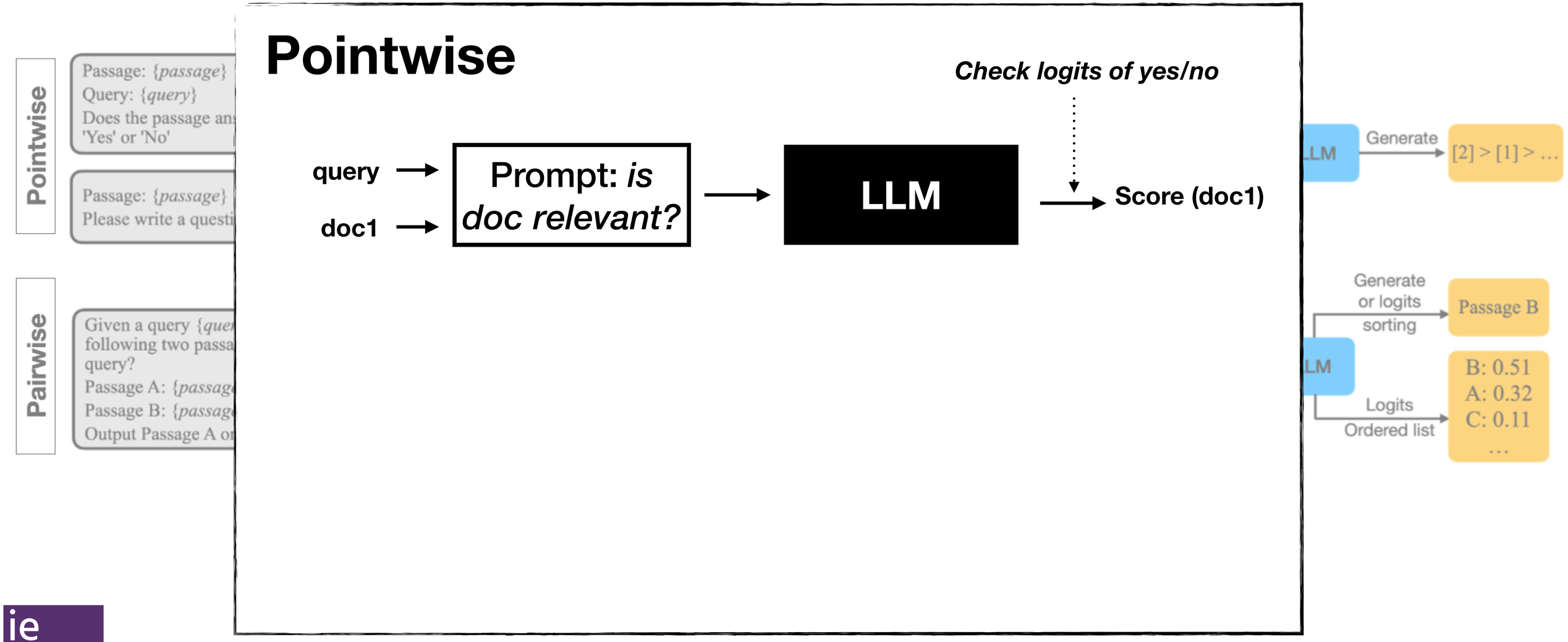
ie lab

Families LLM-based Rankers

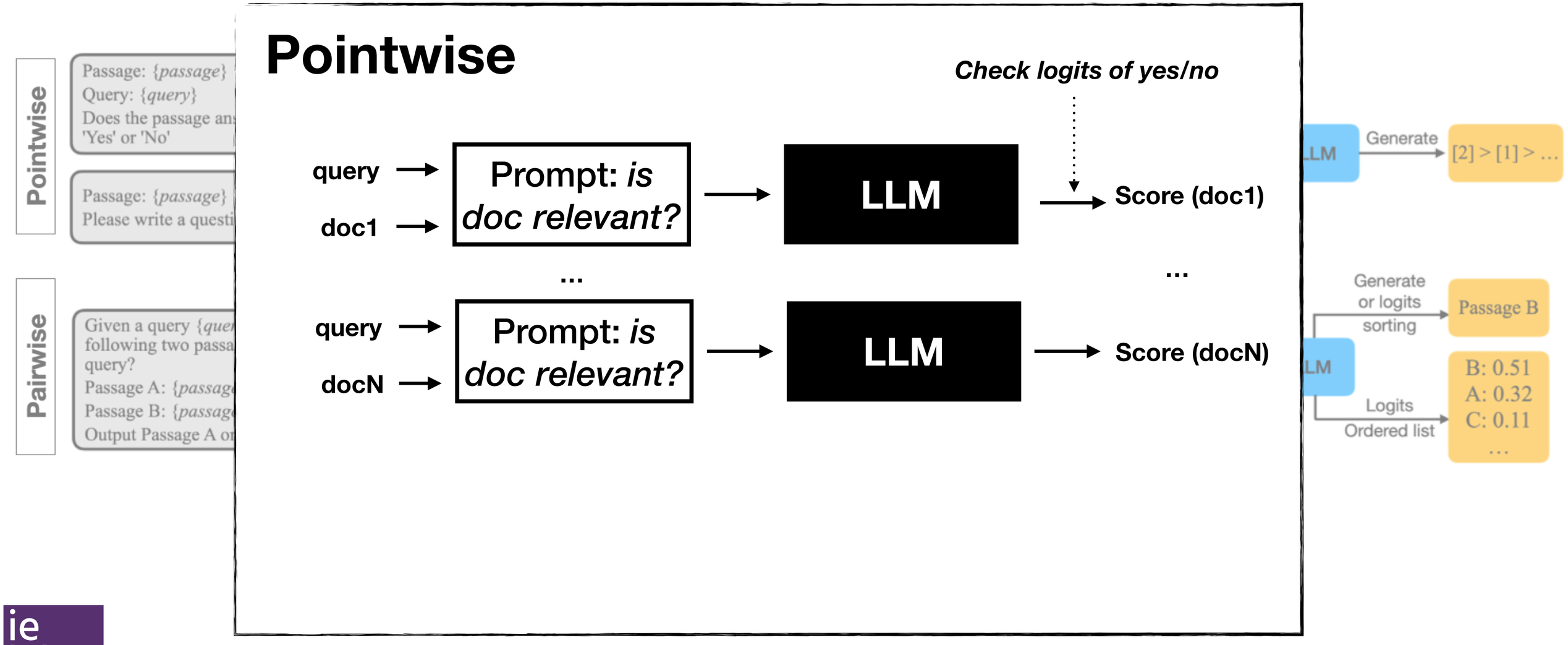


ie lab

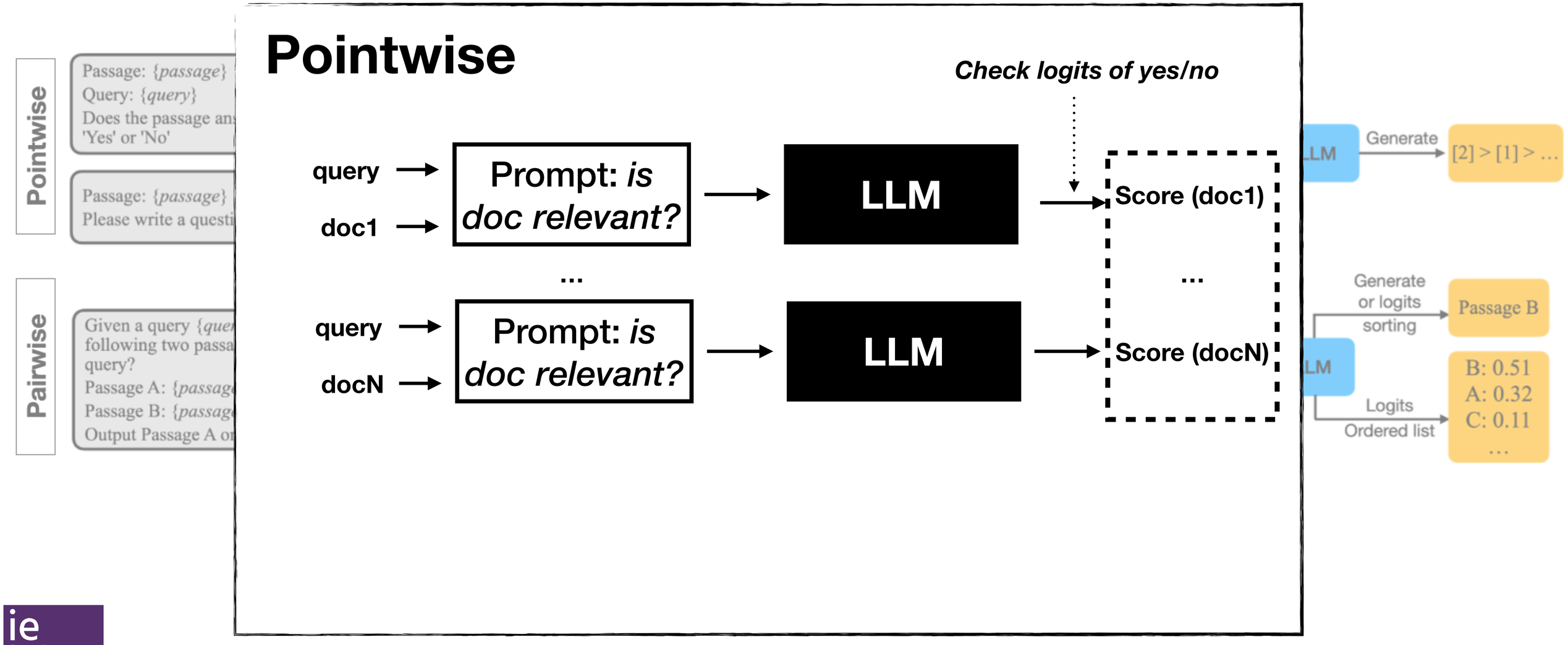
Families LLM-based Rankers



Families LLM-based Rankers

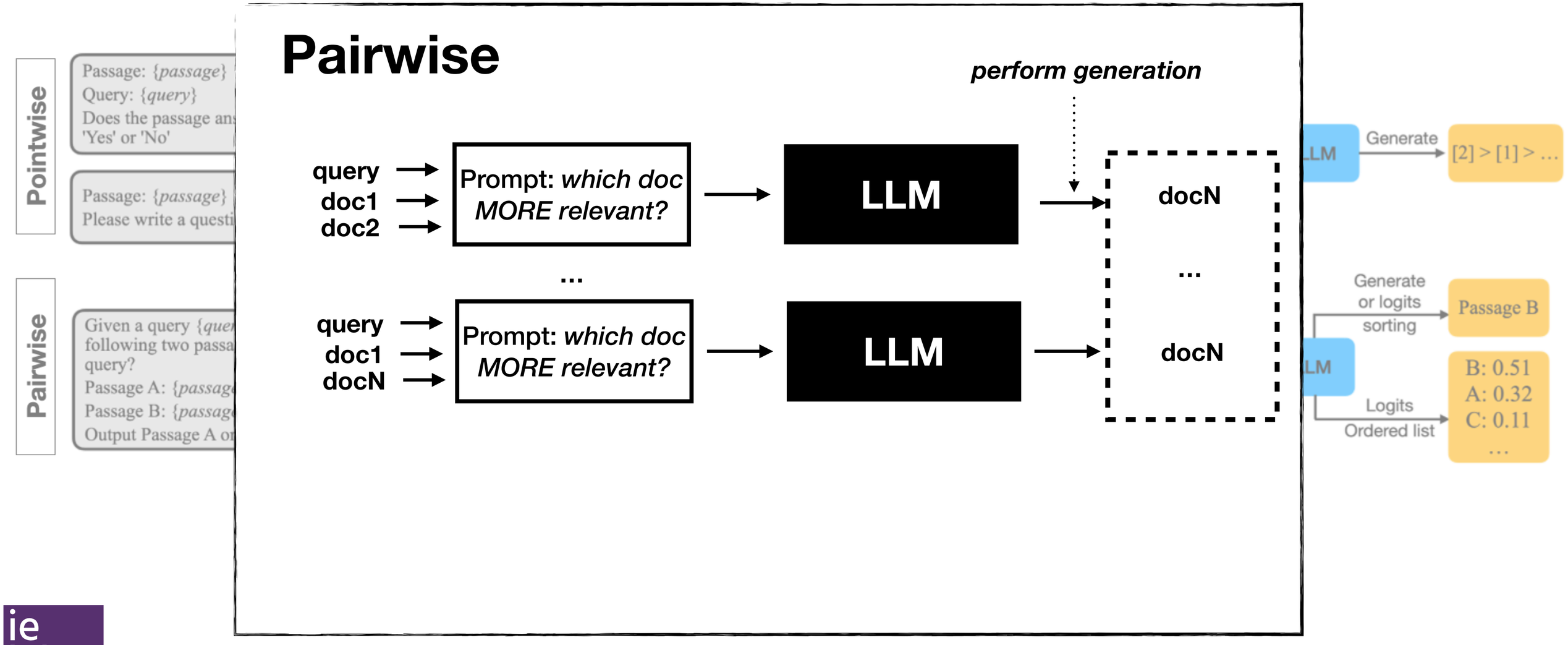


Families LLM-based Rankers

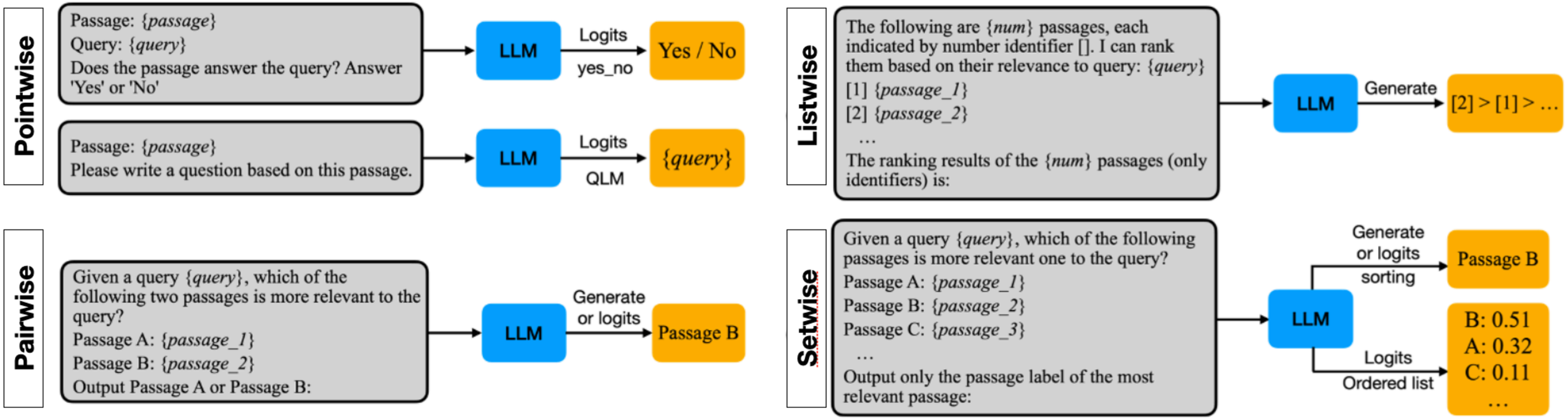


ie lab

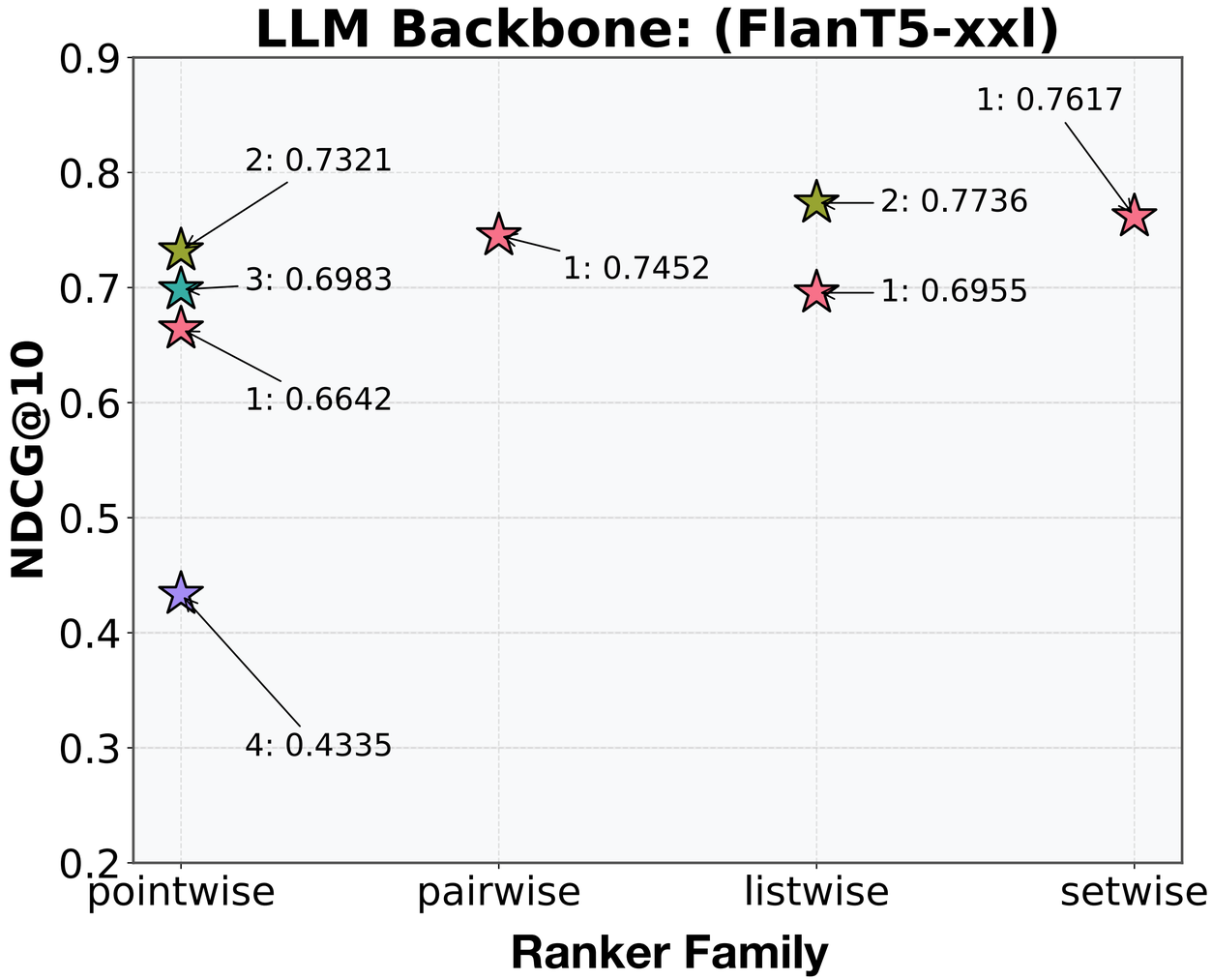
Families LLM-based Rankers



Families LLM-based Rankers

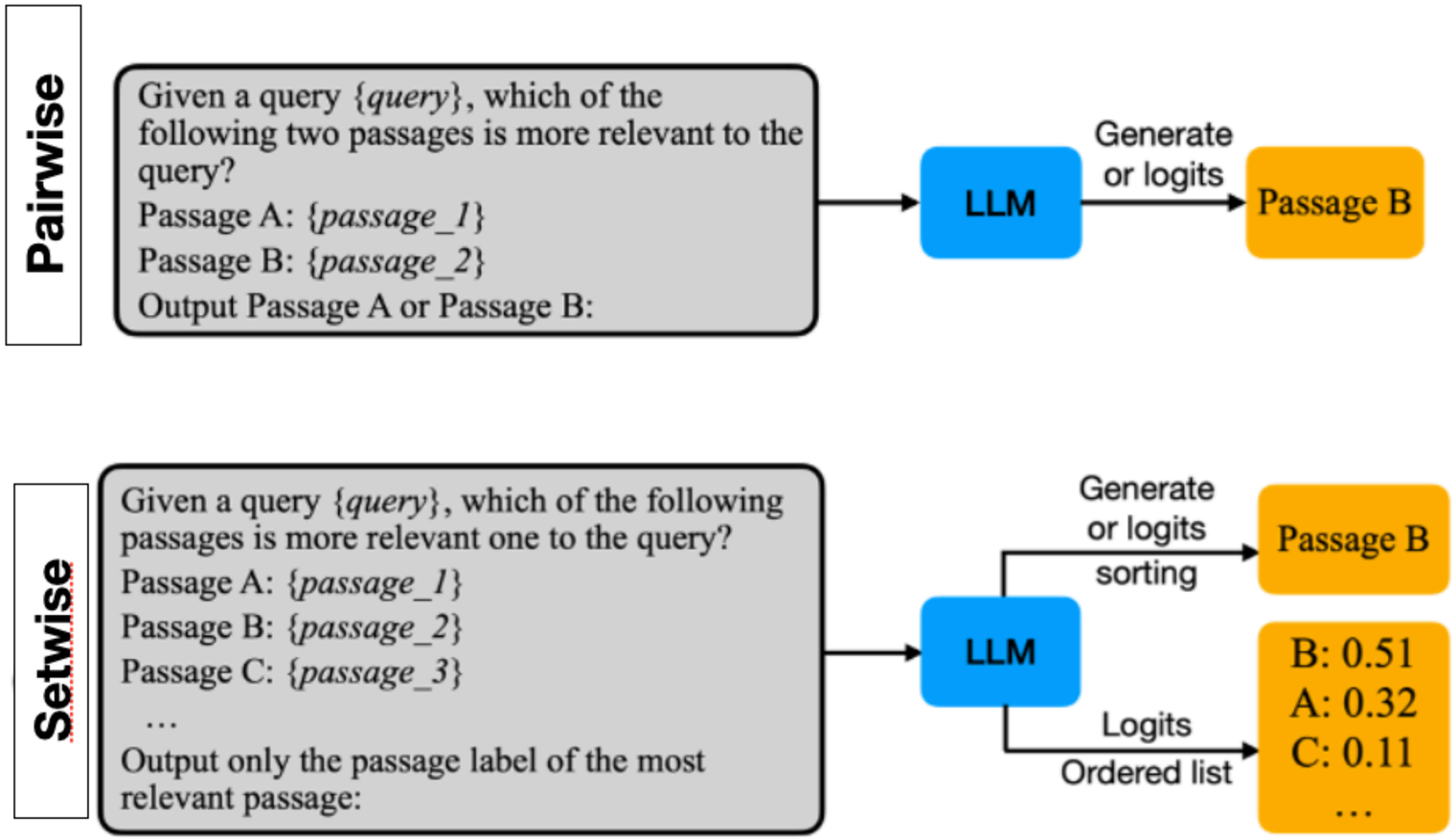


Different ranking mechanisms lead to different effectiveness



ie lab

Prompts don’t differ just because of ranking mechanism implemented



Prompts don’t differ just because of ranking mechanism implemented

The PRP Prompt

Passage: {text} Query: {query}
Does the passage answer the query?

The RankGPT Prompt

You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query. I will provide you with num passages, each indicated by number identifier []. Rank the passages based on their relevance to query: {query}. {PASSAGES} Search Query: {query}. Rank the num passages above based on their relevance to the search query. The passages should be listed in descending order using identifiers. The most relevant passages should be listed first. The output format should be [] > [], e.g., [1] > [2]. Only response the ranking results, do not say any word or explain.

ie lab

Prompts don’t differ just because of ranking mechanism implemented

Role Playing

Task Instructions

Evidence Ordering (wrt query) &
Position of Evidence (wrt instructions)

Output Type

Tone Words

The RankGPT Prompt

You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query. I will provide you with num passages, each indicated by number identifier []. Rank the passages based on their relevance to query: {query}. {PASSAGES} Search Query: {query}. Rank the num passages above based on their relevance to the search query. The passages should be listed in descending order using identifiers. The most relevant passages should be listed first. The output format should be [] > [], e.g., [1] > [2]. Only response the ranking results, do not say any word or explain.

ie lab

Prompts don’t differ just because of ranking mechanism implemented

What is the effect of differences in wording of these prompt components?

Role Playing

Task Instructions

Evidence Ordering (wrt query) &
Position of Evidence (wrt instructions)

Output Type

Tone Words

The RankGPT Prompt

You are RankGPT, an intelligent assistant

based on their relevance to query: {query}. {PASSAGES} Search Query: {query}. Rank the num passages above based on their relevance to the search query. The passages should be listed in descending order using identifiers. The most relevant passages should be listed first. The output format should be [] > [], e.g., [1] > [2]. Only response the ranking results, do not say any word or explain.

ie lab

Prompts don’t differ just because of ranking mechanism implemented

The RankGPT Prompt

You are RankGPT, an intelligent assistant

Tone Words

any word or explain.

What is the effect of differences in wording of these prompt components?

Is effectiveness differences b/w rankers due to:

- RQ1: the actual ranking mechanism, or the choice of words?
- RQ2: LLM characteristics such as backbone and size?

			Wording Alternatives				
Component	Ranker	None (0)	1	2	3	4	5
Task Instruction (TI)	pointwise	-	Does the passage answer the query?	Is this passage relevant to the query?	For the following query and document, judge whether they are relevant.	Judge the relevance between the query and the document.	-
	pairwise	-	Given a query, which of the following two passages is more relevant to the query?	-			
	listwise	-	Rank the {num} passages based on their relevance to the search query.	Sort the Passages by their relevance to the Query.	I will provide you with {num} passages, each indicated by number identifier []. Rank the passages based on their relevance to query.	-	
	setwise	-	Which one is the most relevant to the query.	-			
Output Type (OT)	pointwise		Judge whether they are "Highly Relevant", "Somewhat Relevant", or "Not Relevant".	From a scale of 0 to 4, judge the relevance.	Answer 'Yes' or 'No'.	Answer True/False.	-
	pairwise	-	Output Passage A or Passage B.	-			
	listwise	-	Sorted Passages = [	The passages should be listed in descending order using identifiers. The most relevant passages should be listed first. The output format should be [] >[], e.g., [1] >[2].	-		
	setwise	-	Output the passage label of the most relevant passage.	Generate the passage label.	Generate the passage label that is the most relevant to the query, then explain why you think this passage is the most relevant.	-	
Tone Words (TW)	All	✓	You better get this right or you will be punished.	Only output the ranking results, do not say any word or explanation.	Please	Only	Must
Role Playing (RP)	All	✓	You are RankGPT, an intelligent assistant that can rank passages based on their relevancy to the query.	-			

We control the prompt wording option, vary one option at a time

Table 3: Prompt templates that combine the five components with the four ordering options available. Q and P denote query text passage(s) text, respectively.

EO/PE	B	E
QF	RP+ TI (Q) + P + TW+ OT	RP+ TW+ OT + TI (Q) + P
PF	RP+ P + TI (Q) + TW+ OT	RP+ TW+ OT + P + TI (Q)

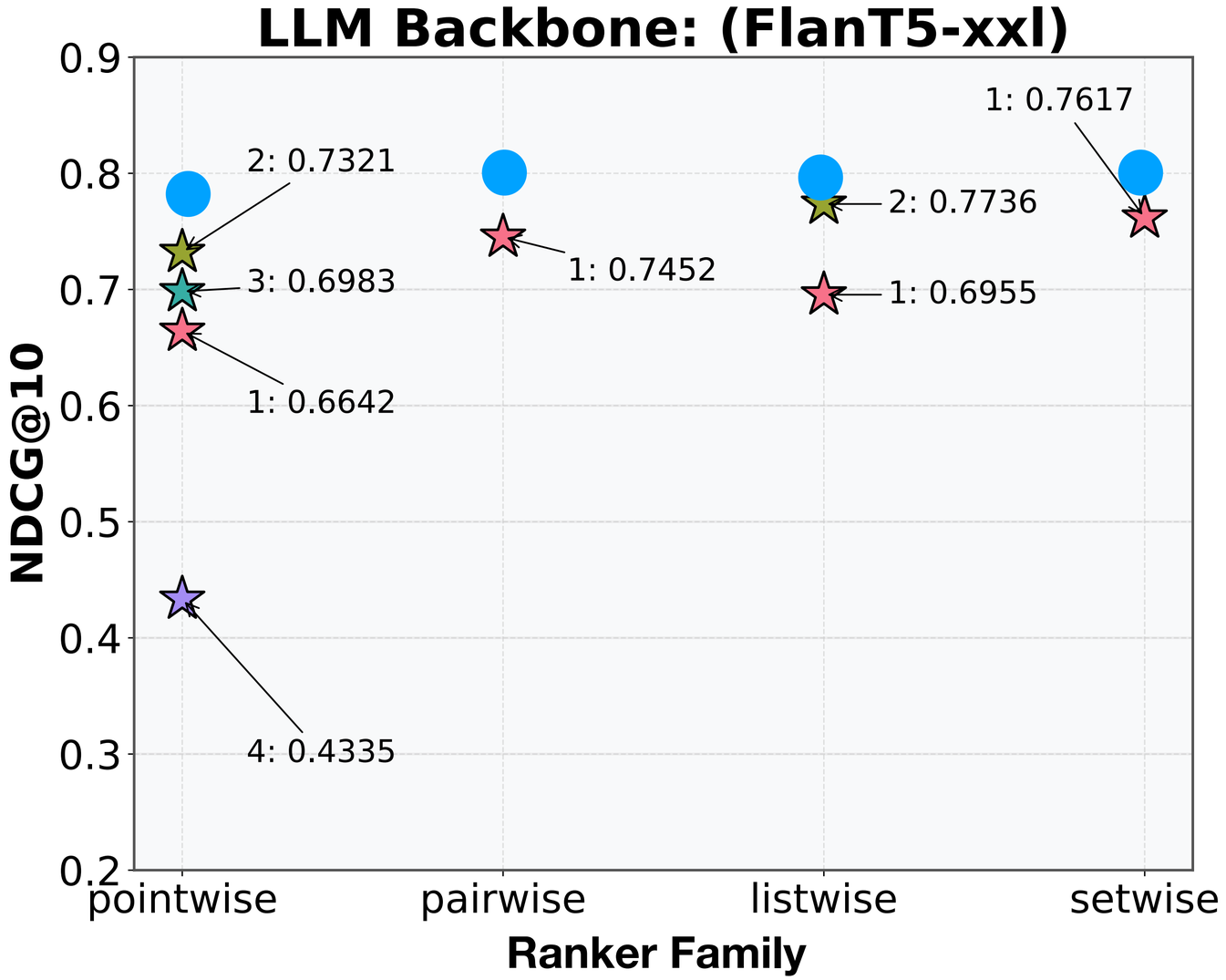
We explore variations in prompt components wordings and ordering

- 1,248 prompt variations
  - e.g. Tone Words: “Please”, “You better get this right or you will be punished”, ….
- 12,400+ GPU-hours, 12,000+ results analysed
- 3 LLM backbone families: FlanT5 (L, XL, XXL), Mistral-7B, Llama3-8B
- Experimented across DL 19, DL 20, COVID (BEIR)

ie lab

Prompts Better Than Original Ones Do Exist

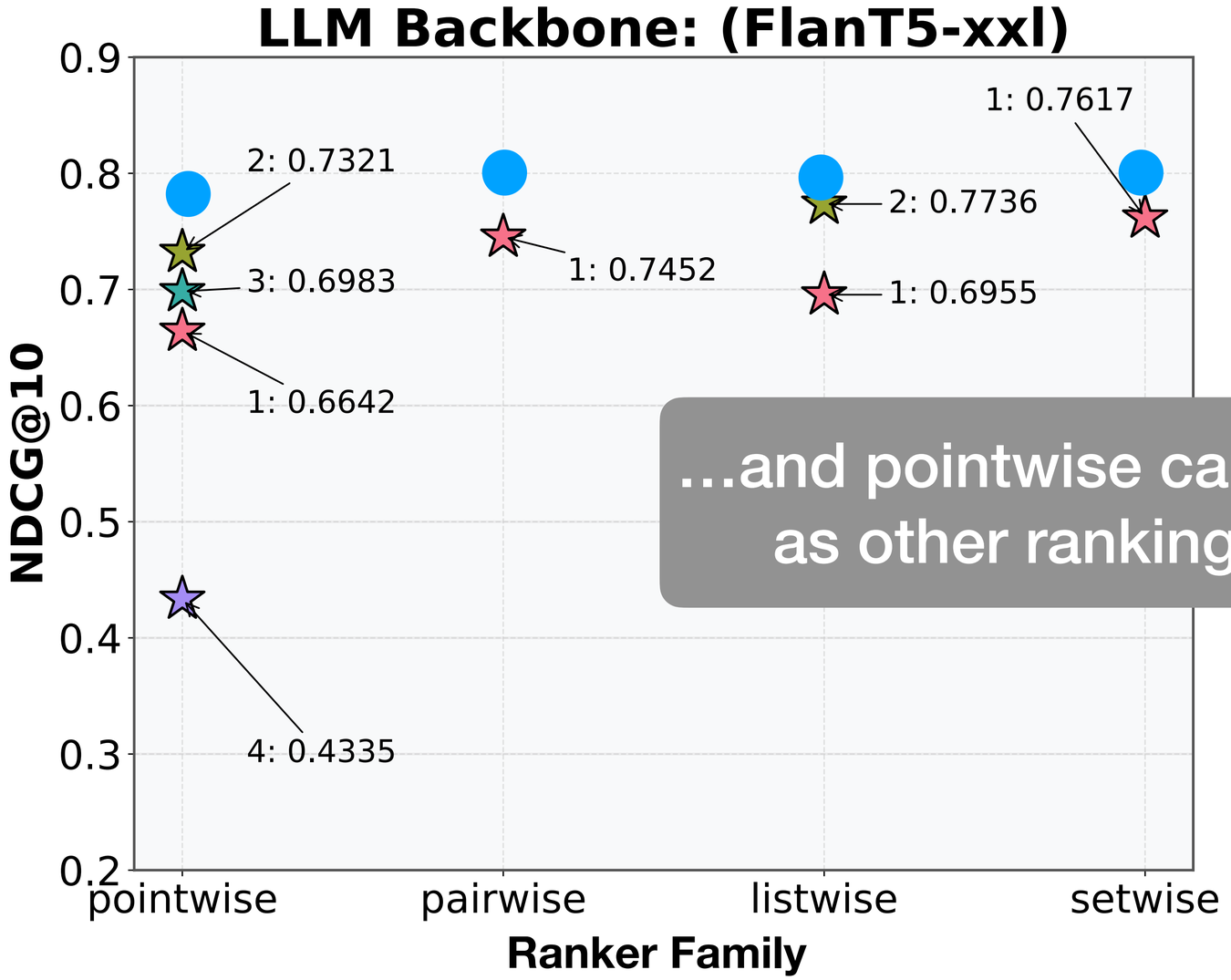
= best prompt variation per ranker family



ie lab

Prompts Better Than Original Ones Do Exist

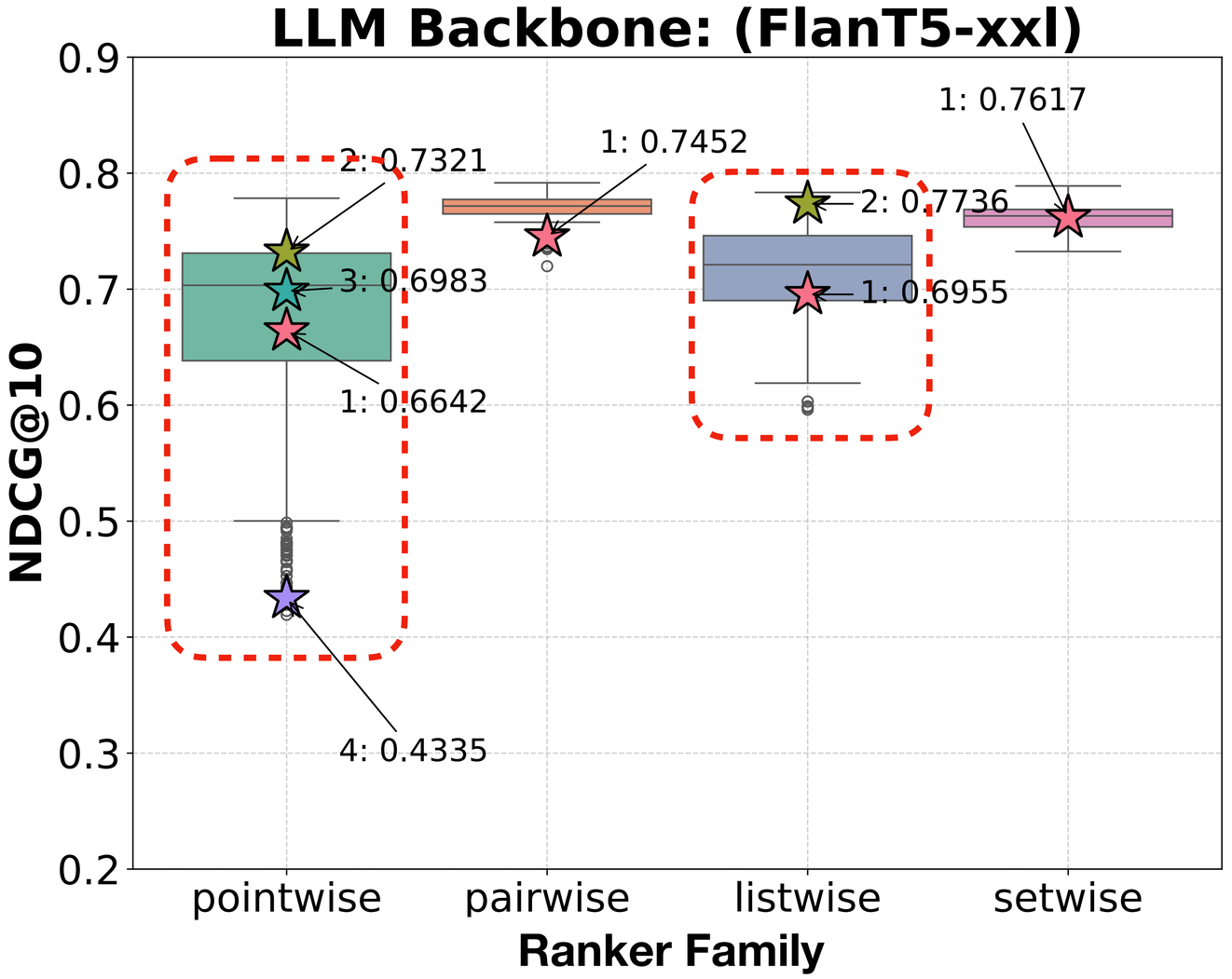
● = best prompt variation per ranker family



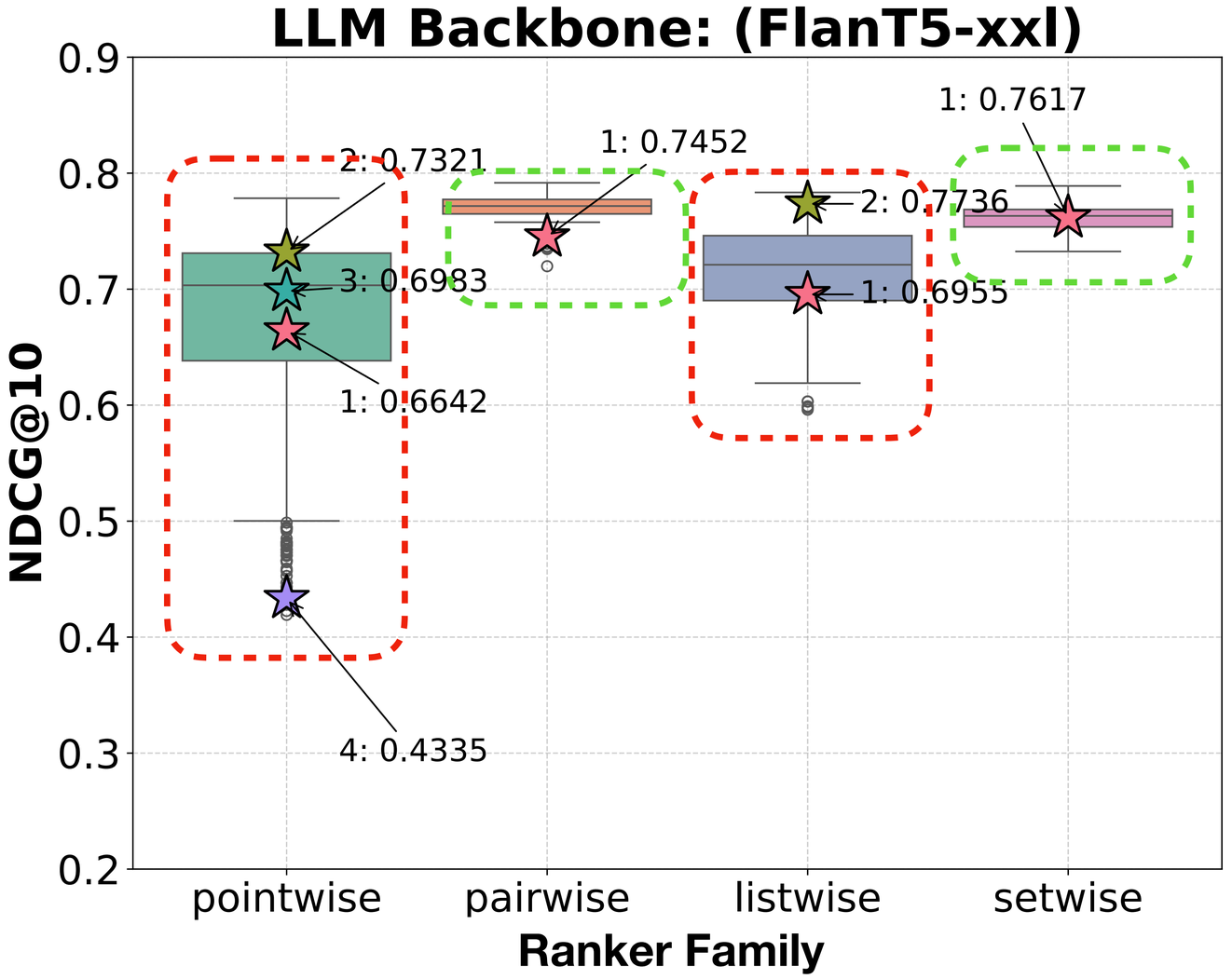
…and pointwise can be as good as other ranking methods

ie lab

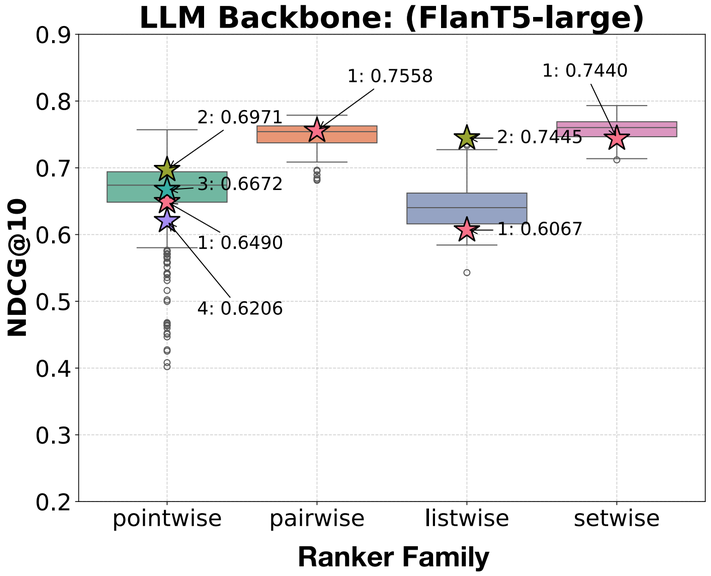
LLM Rankers Can Be (highly) Sensitive to Prompt Variations

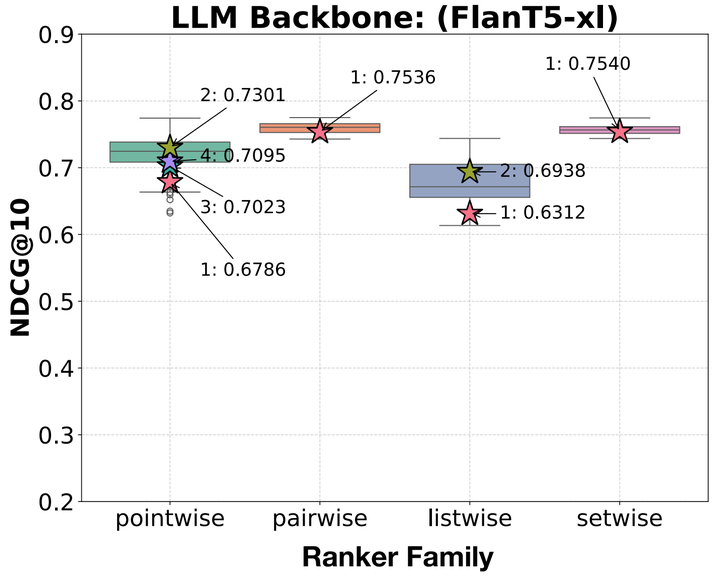


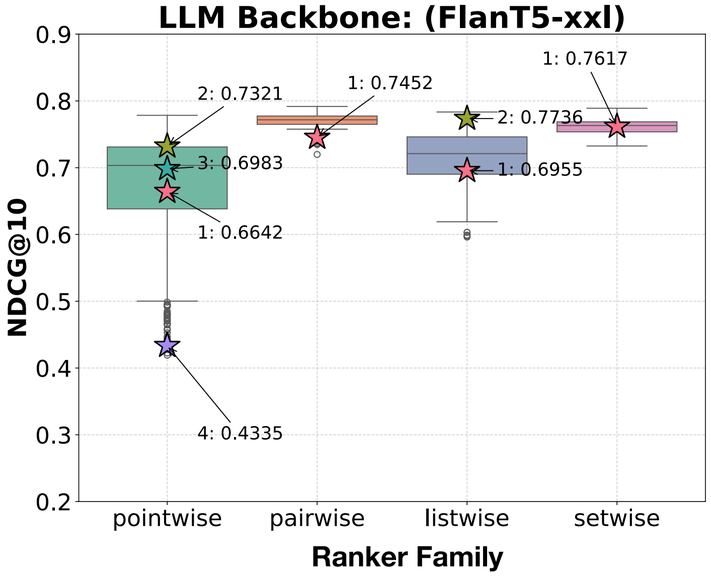
Different LLM Rankers Exhibit Different Variability

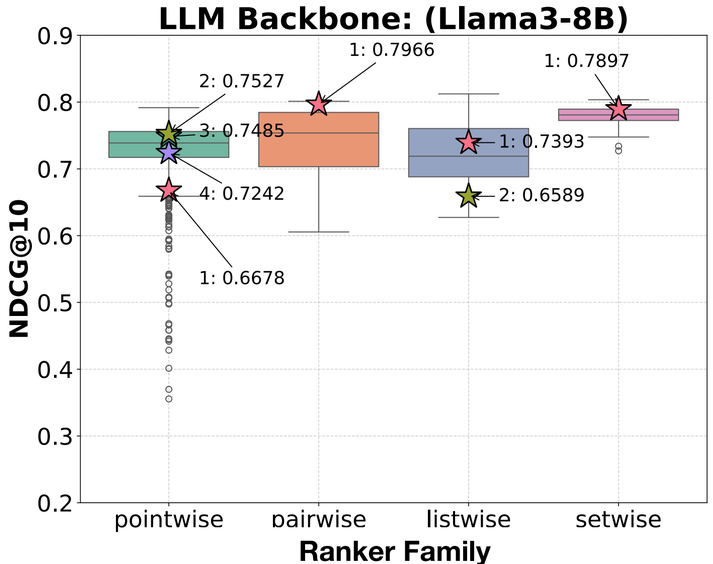


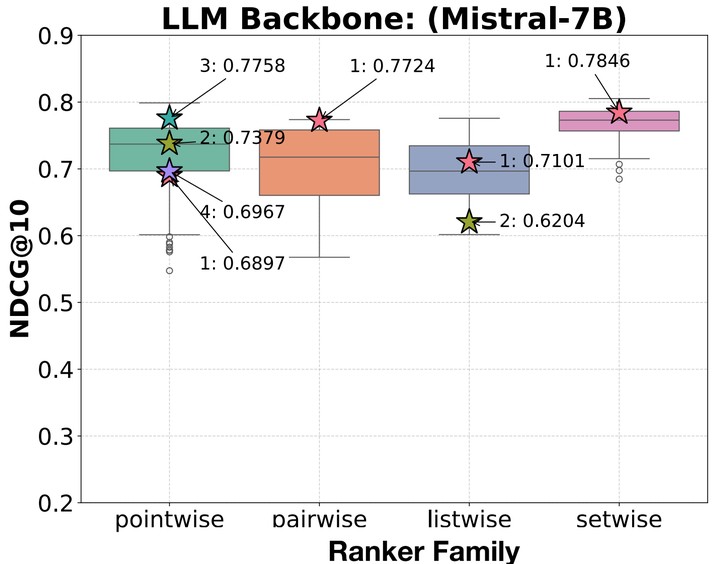
LLM Backbones Influence Effectiveness & Variations Differently Across LLM Rankers











ie lab

In the paper we also…

- show similar findings across datasets
- analyse role of each prompt component type, and instance within prompt component type

ie lab

Key Takeaways

Prompt components beyond ranking method significantly impact effectiveness

Each ranking method has distinct component preferences

No universal "best prompt" exists: depends on ranking method, dataset, and LLM

Key Takeaways

Prompt components beyond ranking method significantly impact effectiveness

Each ranking method has distinct component preferences

No universal "best prompt" exists: depends on ranking method, dataset, and LLM

Future work: *automatic prompt optimisation & prompt performance prediction*