LLMs prioritize information consistently and hierarchically. **Behavioral probing** shows how.

Behavioral Analysis of Information Salience in Large Language Models

- Jan Trienes, Jörg Schlötterer, Junyi Jessy Li, Christin Seifert
- jan.trienes@uni-marburg.de







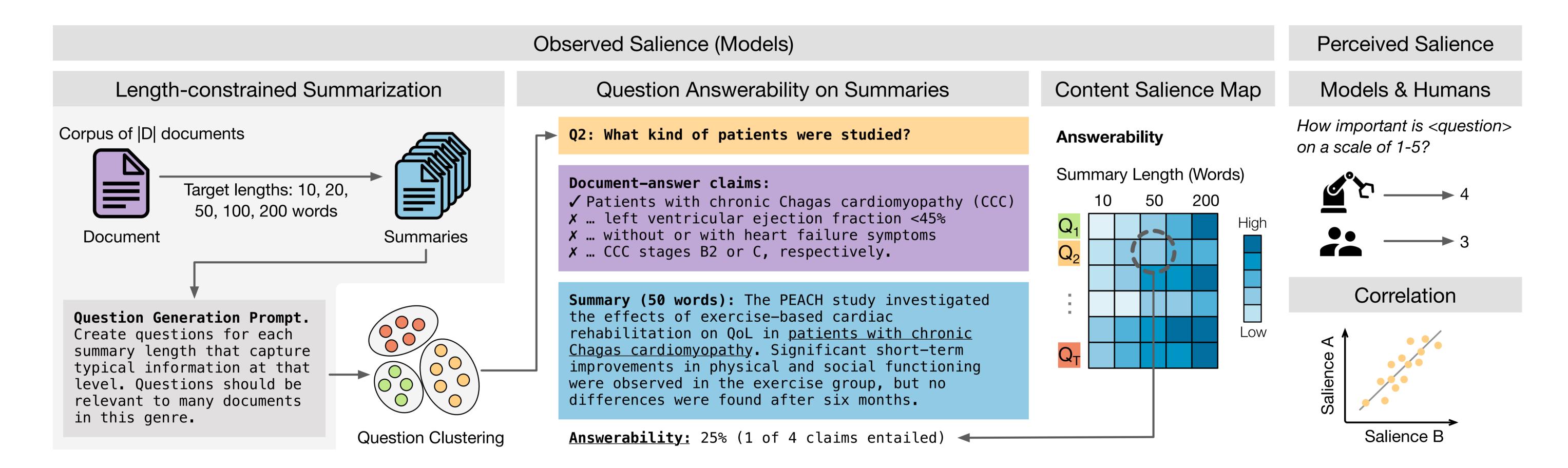
S Paper Link

Motivation

- LLMs excel at text summarization, but content selection issues persist.
- Prior work developed theoretical views of salience, but it remains a latent concept.
- RQ: What salience notion have LLMs learned?

Method

- Use length-controlled summarization as a probe to analyze content prioritization.
- Idea: shorter summaries = higher salience
- Systematically track answerability of questions at each length as *proxy* of content salience.



Results

- Content salience map gives an interpretable view on salience patterns.
- We find a hierarchical prioritization of questions: Some Qs are answered earlier/later
- Model behavior is highly **consistent** and correlates well with other models.
- However, model/human salience notions do align well with each other.

Prevalence GPT-40 Llama 3.1 (70B) .29 .56 .80 .89 .93 .28 .50 .77 .89 .95 Q1. What is the main focus of the study? Q2. Which patient population is the study _____.97 .21 .34 .49 .70 .88 .24 .44 .57 .70 .82 concerned with? Q3. What condition is being addressed in _ .37 .52 .65 .79 .88 .36 .56 .69 .78 .87 the study? .12 .21 .34 .53 .64 .13 .18 .29 .55 .74 characteristics in the study? Q5. What was the main intervention used in .13 .27 .56 .72 .87 .16 .30 .55 .68 .82 the study? Q6. What are the significant benefits of <mark>.18 .32 .55 .70 .87</mark> .16 .29 .52 .66 .80 the intervention? Q7. What are the specific biological -1.40.06 .14 .32 .55 .76 .03 .08 .30 .43 markers influenced by the intervention? .13 .24 .48 .71 .89 .13 .22 .45 .66 .80 in the study? Q9. What specific metrics or outcomes were ___.91 .05 .13 .35 .54 .82 .04 .11 .32 .45 measured? .27 .25 .54 .03 .08 .03 .08 of the trial? Q11. What are the detailed findings .05 .12 .38 .54 .24 .43 .58 .04 .09 regarding adverse events or side effects? Q12. What significant statistical results ______.81 .05 .10 .24 .41 .69 .04 .08 .22 .35 .50 are reported? ... (9 additional rows)11 .22 .42 .57 .71 .12 .23 .43 .62 .81 Average - .84 10 20 50 100 200

Conclusion

- We provide an interpretable framework for analyzing LLMs' notion of salience.
- Model behavior is highly consistent within and across families.
- However, we cannot directly elicit salience notions through introspection, and it only weakly aligns with human expectations.

Measure	Random	OLMo	Mixtral	$Llama_{70b}^{3.1}$	GPT-40
Consist	ency of Sa	alience E	Estimates	5	
LLM-perceived LLM-observed	-0.05 0.92**	0.20* 0.99 **	0.54** 0.99 **	0.71** 0.99 **	0.76 ** 0.98**
Correla	tion of Sa	alience E	Estimates	5	
LLM-perceived vs. Observed	0.03	0.12	0.37*	0.47*	0.50*
Correlatio	n of Mode	l and Hur	man Salie	ence	
LLM-perceived vs. Human LLM-observed vs. Human	0.07 0.20	0.16 0.25	0.41* 0.33*	0.46** 0.36 *	0.53 ** 0.25