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

Marburg University

TEXAS The University of Texas at Austin

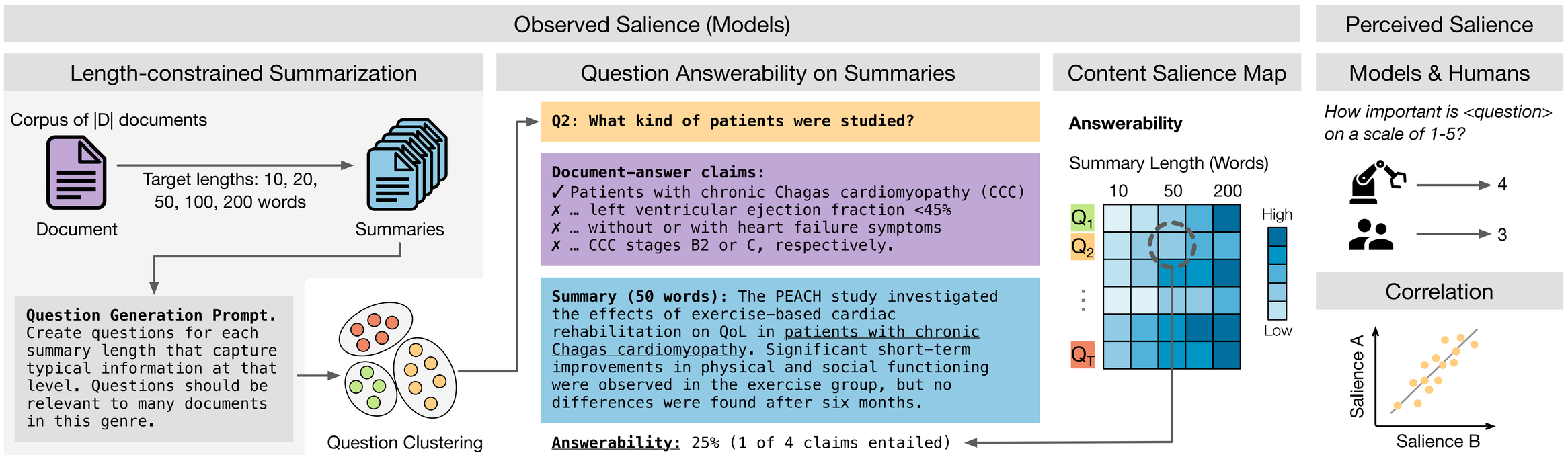
Paper Link

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.

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$^{3.1}_{70b}$	GPT-4o
Consistency of Salience Estimates					
LLM-perceived	–0.05	0.20*	0.54**	0.71**	**0.76****
LLM-observed	0.92**	**0.99****	**0.99****	**0.99****	0.98**
Correlation of Salience Estimates					
LLM-perceived vs. Observed	0.03	0.12	0.37*	0.47*	**0.50***
Correlation of Model and Human Salience					
LLM-perceived vs. Human	0.07	0.16	0.41*	0.46**	**0.53****
LLM-observed vs. Human	0.20	0.25	0.33*	**0.36***	0.25