Behavioral Analysis of Information Salience in Large Language Models

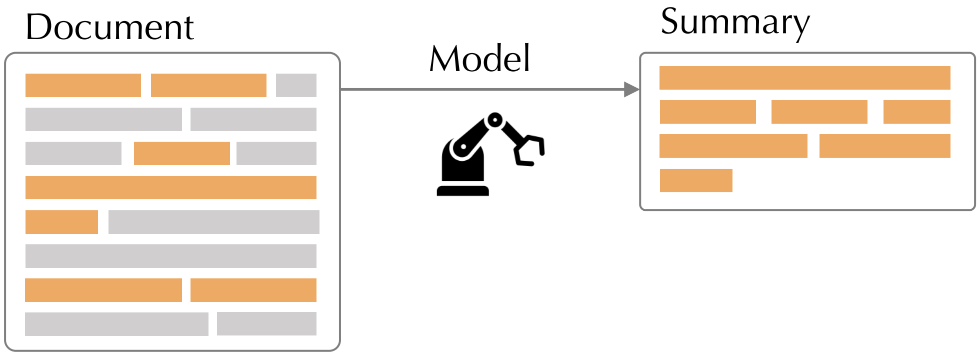
Jan Trienes, Jörg Schlötterer, Junyi Jessy Li, Christin Seifert

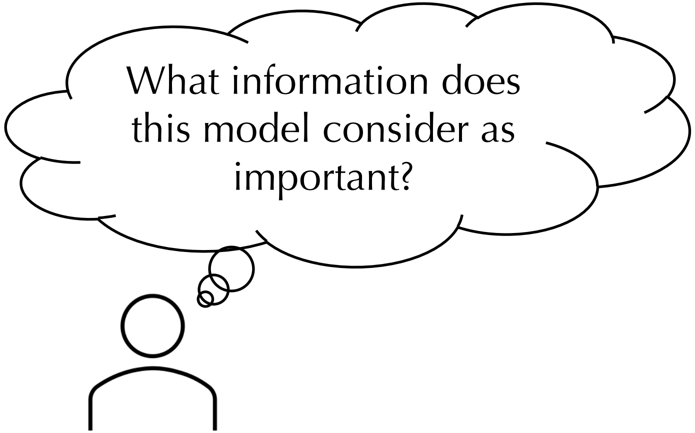
Marburg University

TEXAS
The University of Texas at Austin

ACL 2025 VIENNA

Summarization Needs a Model of Content Salience



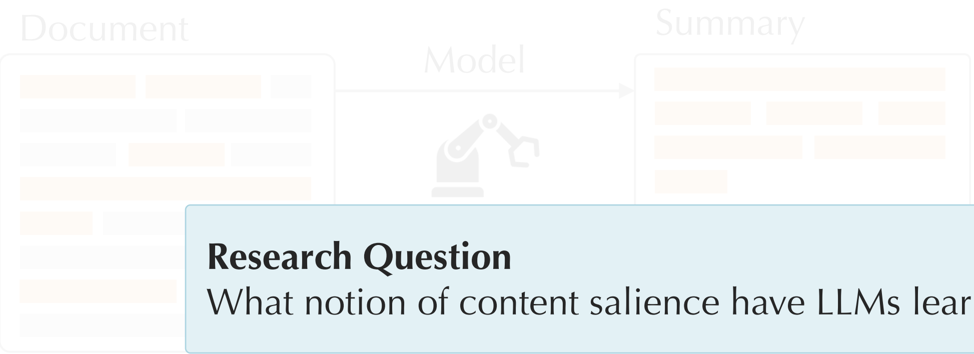


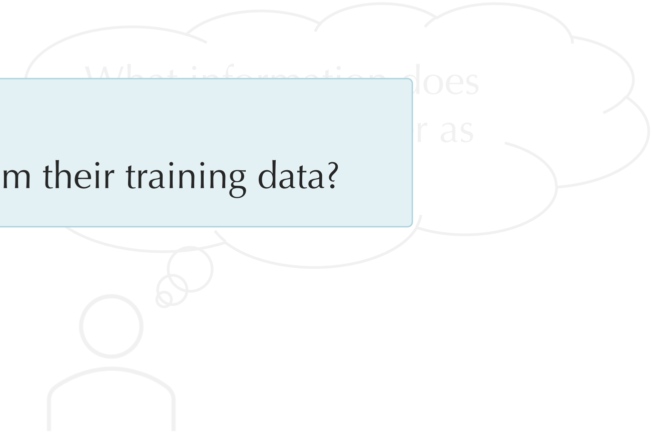
What information does this model consider as important?

Although LLMs are great at summarization, content selection issues remain:

- Book summarization [Kim ‘24, FABLES]
- Lay language [Trienes ‘24, InfoLossQA]
- Diverse opinions [Huang ‘24, DiverseSumm]

Summarization Needs a Model of Content Salience



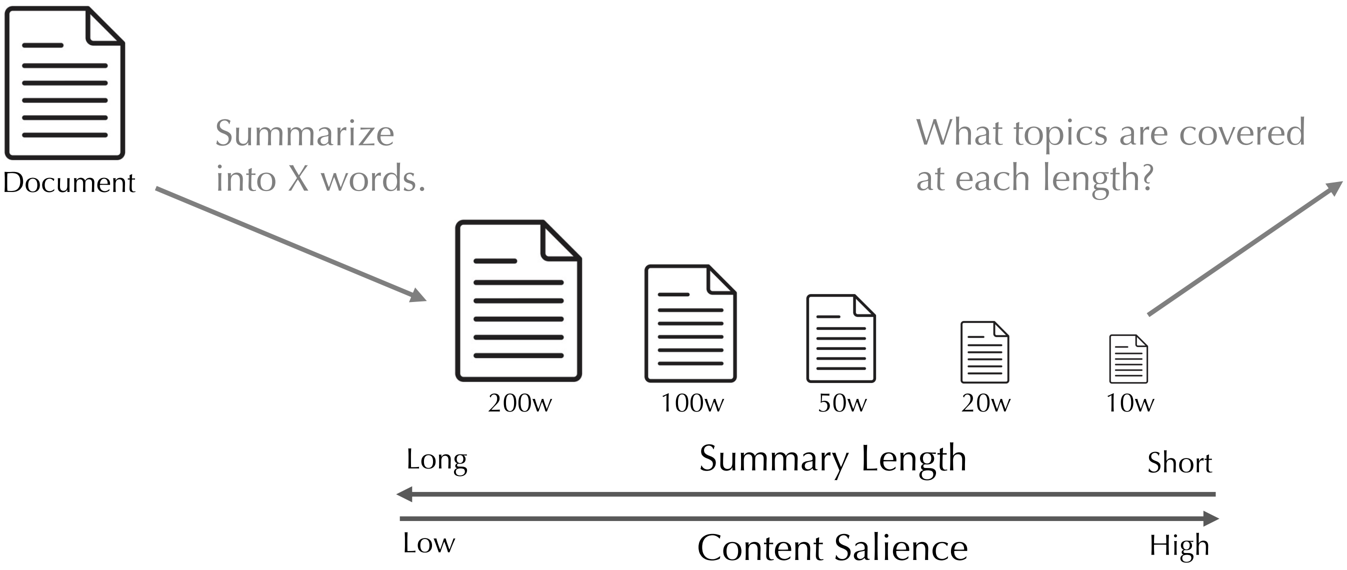


Research Question
What notion of content salience have LLMs learned from their training data?

Although LLMs are great at summarization, content selection issues remain:

- Book summarization [Kim '24, FABLES]
- Lay language [Trienes '24, InfoLossQA]
- Diverse opinions [Huang '24, DiverseSumm]

Using Length-controlled Summarization as a Probe

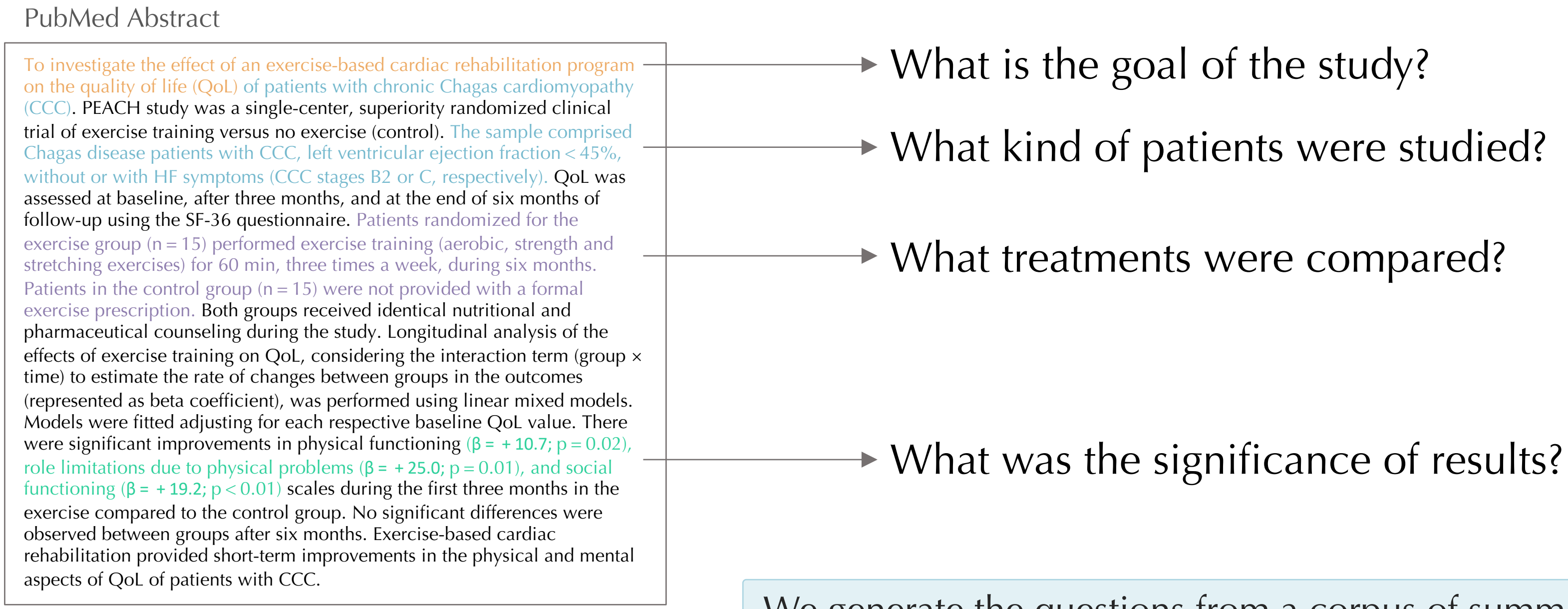


Topic	200	100	50	20	10
t_1	✓	✓	✓	✗	✗
t_2	✓	✓	✓	✓	✓
t_3	✗	✗	✗	✗	✗
t_4	✓	✗	✗	✗	✗
…	…	…	…	…	…

Questions

1. How can we make topics interpretable?
2. How to determine the presence of a topic?

Questions Under Discussion as Interpretable Topics



PubMed Abstract

To investigate the effect of an exercise-based cardiac rehabilitation program on the quality of life (QoL) of patients with chronic Chagas cardiomyopathy (CCC). PEACH study was a single-center, superiority randomized clinical trial of exercise training versus no exercise (control). The sample comprised Chagas disease patients with CCC, left ventricular ejection fraction < 45%, without or with HF symptoms (CCC stages B2 or C, respectively). QoL was assessed at baseline, after three months, and at the end of six months of follow-up using the SF-36 questionnaire. Patients randomized for the exercise group ($n = 15$) performed exercise training (aerobic, strength and stretching exercises) for 60 min, three times a week, during six months. Patients in the control group ($n = 15$) were not provided with a formal exercise prescription. Both groups received identical nutritional and pharmaceutical counseling during the study. Longitudinal analysis of the effects of exercise training on QoL, considering the interaction term (group × time) to estimate the rate of changes between groups in the outcomes (represented as beta coefficient), was performed using linear mixed models. Models were fitted adjusting for each respective baseline QoL value. There were significant improvements in physical functioning ($\beta = +10.7$; $p = 0.02$), role limitations due to physical problems ($\beta = +25.0$; $p = 0.01$), and social functioning ($\beta = +19.2$; $p < 0.01$) scales during the first three months in the exercise compared to the control group. No significant differences were observed between groups after six months. Exercise-based cardiac rehabilitation provided short-term improvements in the physical and mental aspects of QoL of patients with CCC.

What is the goal of the study?

What kind of patients were studied?

What treatments were compared?

What was the significance of results?

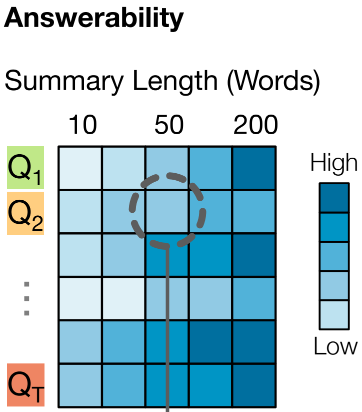
We generate the questions from a corpus of summaries. See paper for details.

Measuring Salience through Question Answerability

Q2: What kind of patients were studied?

Document-answer claims:
✓ Patients with chronic Chagas cardiomyopathy (CCC)
✗ … left ventricular ejection fraction <45%
✗ … without or with heart failure symptoms
✗ … CCC stages B2 or C, respectively.

Summary (50 words): The PEACH study investigated the effects of exercise-based cardiac rehabilitation on QoL in patients with chronic Chagas cardiomyopathy. Significant short-term improvements in physical and social functioning were observed in the exercise group, but no differences were found after six months.



Answerability: 25% (1 of 4 claims entailed)

Experiments

Summarization Tasks

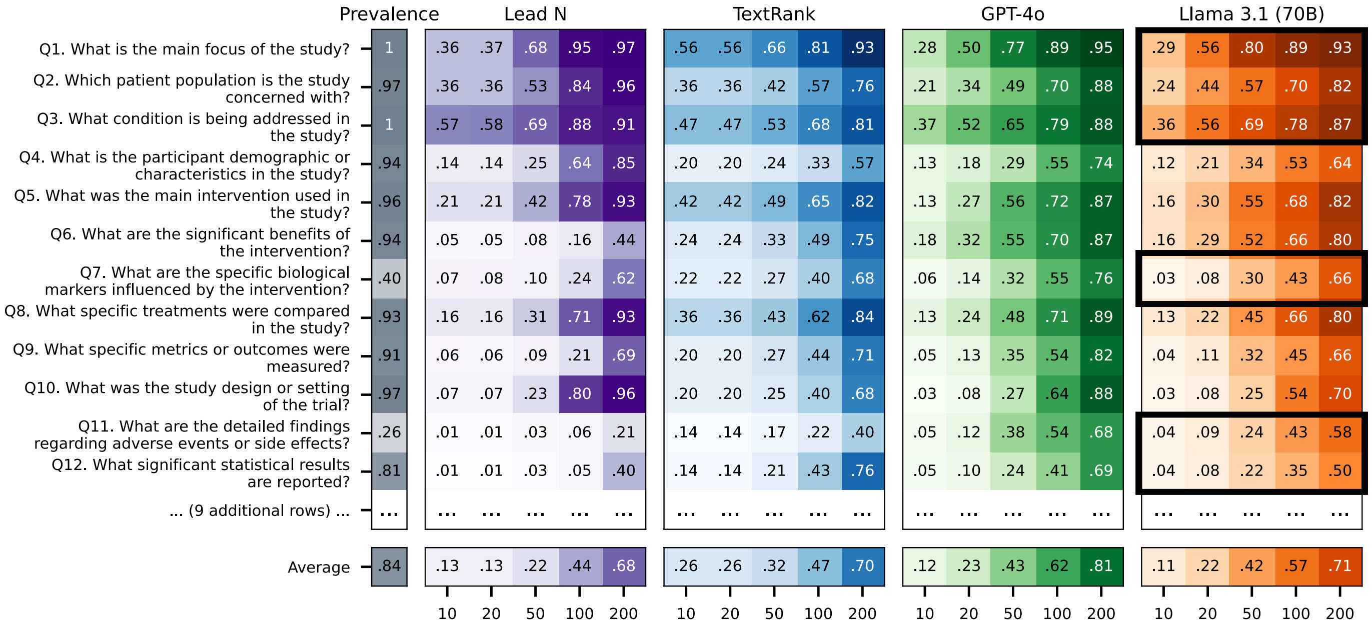
🏥 PubMed RCT abstracts
📝 Related work in NLP papers
🌌 Discussions in astrophysics papers
💬 Meeting transcripts (QMSum)

Summarization Models

OLMo (7B, v1)
Mistral (7B) and Mixtral (8x7B)
Llama 2 (7B, 13B, 70B)
Llama 3 (8B, 70B)
Llama 3.1 (8B, 70B)
GPT-4o and GPT-4o-mini

RQ1: What notion of salience have LLMs internalized?

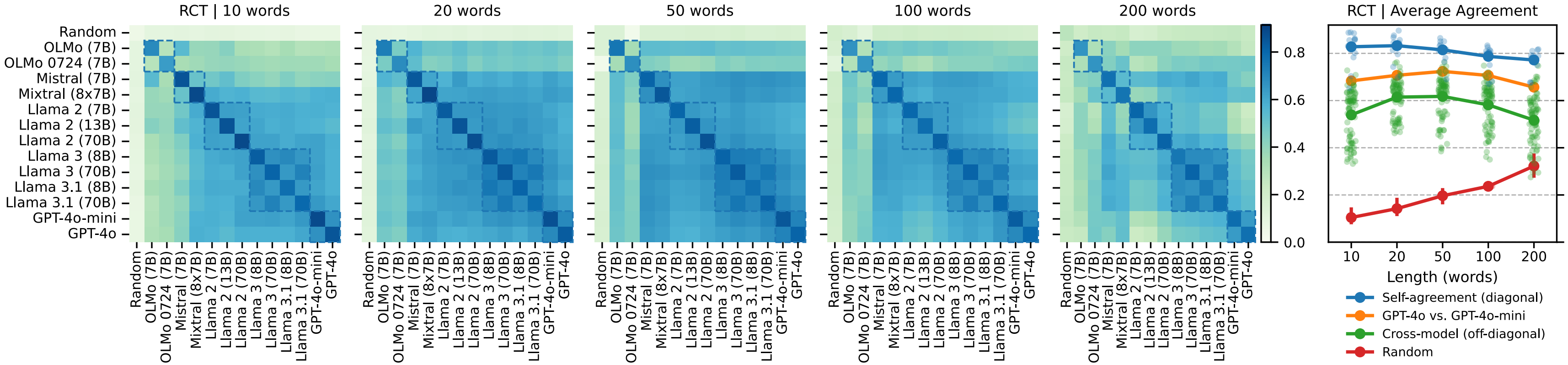
Question answerability by model and summary length.



	Prevalence	Lead N					TextRank					GPT-4o					Llama 3.1 (70B)				
		10	20	50	100	200	10	20	50	100	200	10	20	50	100	200	10	20	50	100	200
Q1. What is the main focus of the study?	1	.36	.37	.68	.95	.97	.56	.56	.66	.81	.93	.28	.50	.77	.89	.95	.29	.56	.80	.89	.93
Q2. Which patient population is the study concerned with?	.97	.36	.36	.53	.84	.96	.36	.36	.42	.57	.76	.21	.34	.49	.70	.88	.24	.44	.57	.70	.82
Q3. What condition is being addressed in the study?	1	.57	.58	.69	.88	.91	.47	.47	.53	.68	.81	.37	.52	.65	.79	.88	.36	.56	.69	.78	.87
Q4. What is the participant demographic or characteristics in the study?	.94	.14	.14	.25	.64	.85	.20	.20	.24	.33	.57	.13	.18	.29	.55	.74	.12	.21	.34	.53	.64
Q5. What was the main intervention used in the study?	.96	.21	.21	.42	.78	.93	.42	.42	.49	.65	.82	.13	.27	.56	.72	.87	.16	.30	.55	.68	.82
Q6. What are the significant benefits of the intervention?	.94	.05	.05	.08	.16	.44	.24	.24	.33	.49	.75	.18	.32	.55	.70	.87	.16	.29	.52	.66	.80
Q7. What are the specific biological markers influenced by the intervention?	.40	.07	.08	.10	.24	.62	.22	.22	.27	.40	.68	.06	.14	.32	.55	.76	.03	.08	.30	.43	.66
Q8. What specific treatments were compared in the study?	.93	.16	.16	.31	.71	.93	.36	.36	.43	.62	.84	.13	.24	.48	.71	.89	.13	.22	.45	.66	.80
Q9. What specific metrics or outcomes were measured?	.91	.06	.06	.09	.21	.69	.20	.20	.27	.44	.71	.05	.13	.35	.54	.82	.04	.11	.32	.45	.66
Q10. What was the study design or setting of the trial?	.97	.07	.07	.23	.80	.96	.20	.20	.25	.40	.68	.03	.08	.27	.64	.88	.03	.08	.25	.54	.70
Q11. What are the detailed findings regarding adverse events or side effects?	.26	.01	.01	.03	.06	.21	.14	.14	.17	.22	.40	.05	.12	.38	.54	.68	.04	.09	.24	.43	.58
Q12. What significant statistical results are reported?	.81	.01	.01	.03	.05	.40	.14	.14	.21	.43	.76	.05	.10	.24	.41	.69	.04	.08	.22	.35	.50
... (9 additional rows) ...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Average	.84	.13	.13	.22	.44	.68	.26	.26	.32	.47	.70	.12	.23	.43	.62	.81	.11	.22	.42	.57	.71

Finding: LLM's notion of salience is hierarchical. Some questions are answered earlier/later, and to different degrees.

RQ2: Is the salience notion consistent?



Finding 1: salience notion is highly consistent within the same model (diagonal).
Finding 2: high cross-family agreement suggests LLMs are converging (off-diagonal).

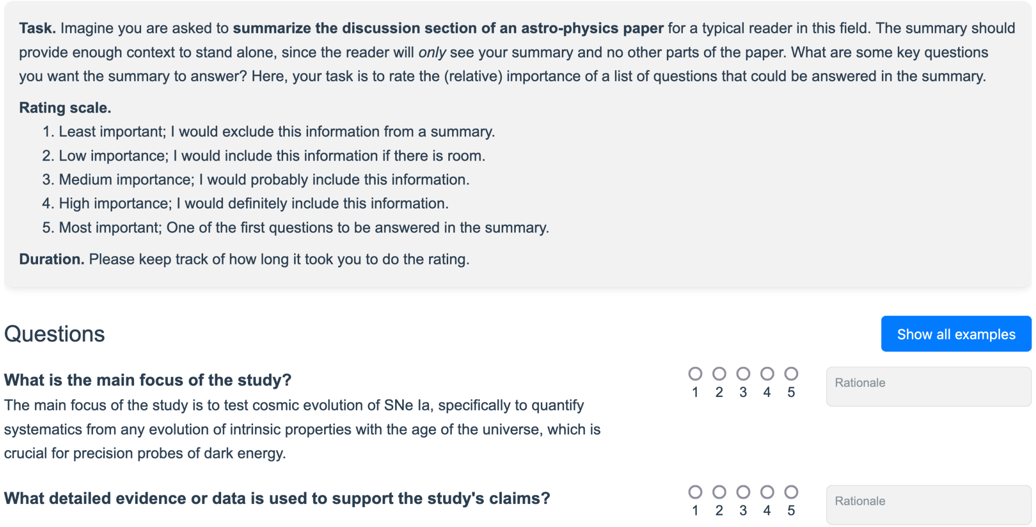
From Observed Salience to Perceived Salience

How does model salience relate to human expectations?

- Recruit 3-5 experts per task
- Rate salience of questions
- Correlate ratings

Additionally, prompt LLMs to rate questions.

- Does this approximate their behavior?
- Can they reason about salience?



Task. Imagine you are asked to **summarize the discussion section of an astro-physics paper** for a typical reader in this field. The summary should provide enough context to stand alone, since the reader will *only* see your summary and no other parts of the paper. What are some key questions you want the summary to answer? Here, your task is to rate the (relative) importance of a list of questions that could be answered in the summary.

Rating scale.

1. Least important; I would exclude this information from a summary.
2. Low importance; I would include this information if there is room.
3. Medium importance; I would probably include this information.
4. High importance; I would definitely include this information.
5. Most important; One of the first questions to be answered in the summary.

Duration. Please keep track of how long it took you to do the rating.

Questions

What is the main focus of the study?
The main focus of the study is to test cosmic evolution of SNe Ia, specifically to quantify systematics from any evolution of intrinsic properties with the age of the universe, which is crucial for precision probes of dark energy.

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

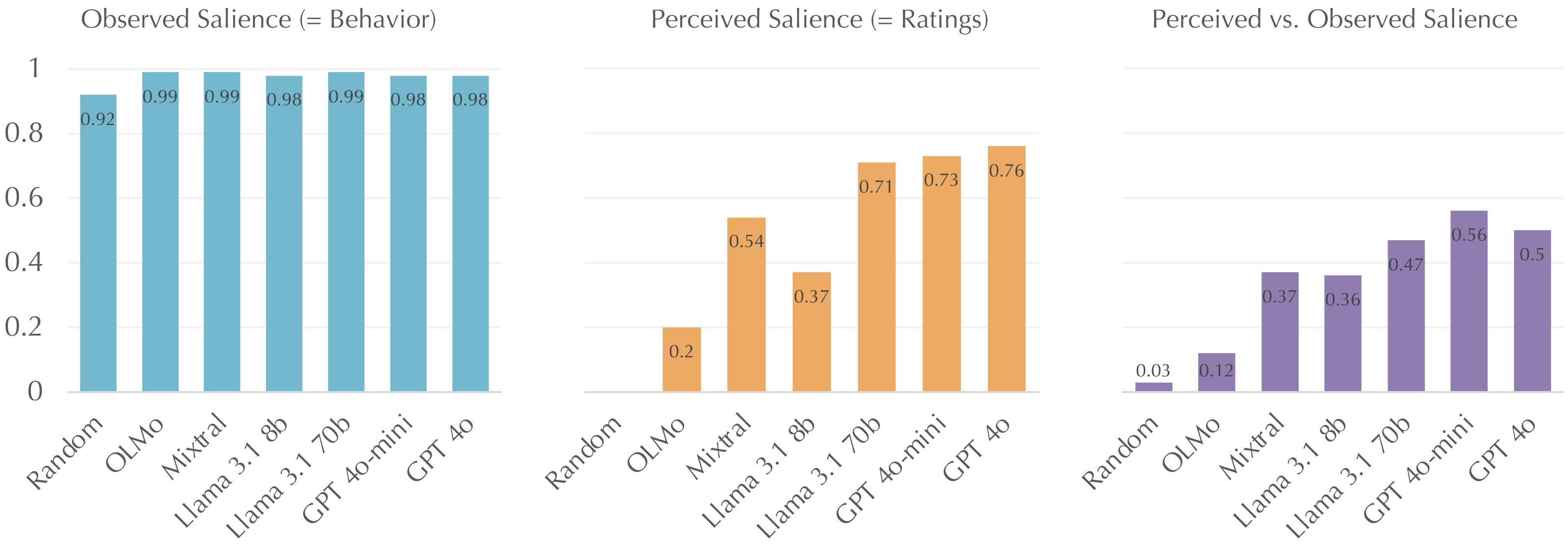
Rationale

What detailed evidence or data is used to support the study's claims?

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Rationale

RQ3: Can Models Reliably Rate Salience?



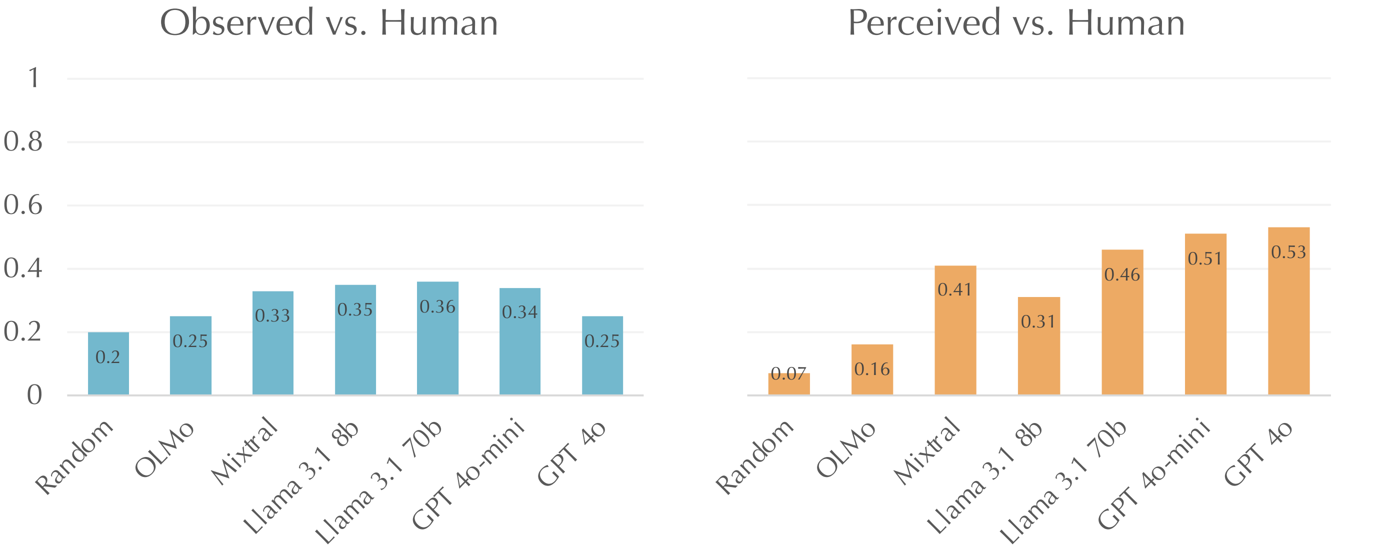
Observed Salience (= Behavior)

Perceived Salience (= Ratings)

Perceived vs. Observed Salience

Finding 1: models cannot consistently rate question salience
Finding 2: model behavior ≠ perceived notion of salience

RQ4: How does Model Salience Relate to Human Salience?



Finding: model salience appears misaligned from human expectations

Conclusion

We provide an **interpretable framework for analyzing** LLMs′ notion of content salience.

Model **behavior is highly consistent** within and across model families.

However, **we cannot directly elicit** internal salience notions, and it only **weakly aligns** with human expectations.

Thanks!

github.com/jantrienes/llm-salience

jan.trienes@uni-marburg.de