Behavioral Analysis of Information Salience in Large Language Models

Jan Trienes[1] **Jörg Schlötterer**[1,2] **Junyi Jessy Li**[3] **Christin Seifert**[1]
[1]Marburg University [2]University of Mannheim
[3]The University of Texas at Austin
{jan.trienes,joerg.schloetterer,christin.seifert}@uni-marburg.de
jessy@utexas.edu

Abstract

Large Language Models (LLMs) excel at text summarization, a task that requires models to select content based on its importance. However, the exact notion of salience that LLMs have internalized remains unclear. To bridge this gap, we introduce an explainable framework to systematically derive and investigate information salience in LLMs through their summarization behavior. Using length-controlled summarization as a behavioral probe into the content selection process, and tracing the answerability of Questions Under Discussion throughout, we derive a proxy for how models prioritize information. Our experiments on 13 models across four datasets reveal that LLMs have a nuanced, hierarchical notion of salience, generally consistent across model families and sizes. While models show highly consistent behavior and hence salience patterns, this notion of salience cannot be accessed through introspection, and only weakly correlates with human perceptions of information salience.[1]

1 Introduction

Large Language Models (LLMs) significantly advanced text synthesis tasks, including text summarization, which they perform well even under zero-shot conditions (Goyal et al., 2023; Zhang et al., 2024). The nature of the summarization task requires models to do content selection: picking the most salient pieces of information for inclusion in a summary (Mani and Maybury, 1999). However, it remains unclear what underlying notion of salience the models have internalized.

Prior work investigated information salience from several angles. Theories of discourse structure have been used to induce content salience (Marcu, 1999; Louis et al., 2010), and a large body of summarization research uses word distribution or centrality as the main signal for content selection (Nenkova and McKeown, 2012; Nazari and Mahdavi, 2019). Peyrard (2019) laid out a theoretical perspective for content salience in summarization, though the exact notion remains largely latent and aloof; rather, pre-LLM summarization work uses human summaries as supervision signals to learn what to include (Gehrmann et al., 2018; Chen and Bansal, 2018; Liu and Lapata, 2019, *inter alia*). Yet, none of these accounts explains why LLM zero-shot summarization works so well on the one hand, while missing key elements on the other (Kim et al., 2024; Trienes et al., 2024; Huang et al., 2024).

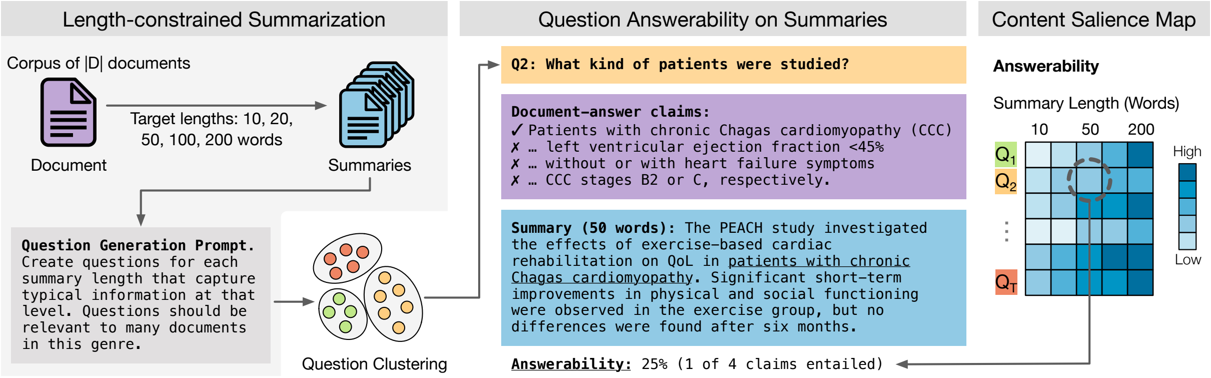
To begin to make sense of this behavior, we need to understand how models internalize salience: whether it is a *consistent* notion within and across models, *how* they prioritize information, and whether LLMs' notion of salience *aligns* with prior theories or human intuitions.

In this paper, we present a novel explainable framework to systematically derive and investigate LLMs' grasp of information salience through their summarization behavior. Our method combines **two key ideas**. First, we can use length-constrained summarization (Fan et al., 2018; He et al., 2022) as a behavioral probe into the content selection process of LLMs. Intuitively, when there is a limited length budget for a summary, we posit that the least important information is dropped first.

Second, we can describe what is salient as the answerability of domain-relevant Questions Under Discussion (QUDs; Van Kuppevelt, 1995; Benz and Jasinskaja, 2017; Wu et al., 2023a). QUDs can be thought of as representations of a coherent unit of information in the form of information-seeking questions, e.g., *Who are the participants of this study?* We use such questions — and hence their answers extracted from documents, according to alternative semantics (Hamblin, 1973; Karttunen, 1977; Groenendijk and Stokhof, 1984) — as the

[1] We release code, model outputs and human annotations at https://github.com/jantrienes/llm-salience.

Findings of the Association for Computational Linguistics: ACL 2025, pages 23428–23454
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics



Question Generation Prompt. Create questions for each summary length that capture typical information at that level. Questions should be relevant to many documents in this genre.

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)

Figure 1: Framework overview, conceptualizing content salience as question answerability. **Left:** Given a corpus, we derive questions that are typically answered in summaries. Length-controlled summarization acts as a probe into the content-selection process of LLMs. Question paraphrases are clustered by semantic intent. **Middle:** Answerability is calculated as the fraction of document-answer claims entailed by the summary. **Right:** The content salience map tracks answerability at each summary length. More salient questions remain answerable even in shorter summaries.

primary unit of analysis, making our framework interpretable and customizable.

Taken together, by gradually decreasing the length budget available for a summary and by systematically tracing question answerability throughout, we can derive a *proxy* for how models prioritize information. See Figure 1 for an overview.

Using this framework (§ 2), we empirically study LLMs' content selection *behavior*, and its alignment with *perceived* notions of salience. Through experiments on 13 models and four datasets (§ 3), we aim to answer the following research questions:

RQ1 What notion of salience have LLMs learned in different domains?

RQ2 Do LLMs of different families/sizes have a similar notion of salience?

RQ3 When models introspect, does their perceived notion of salience align with their summarization behavior?

RQ4 To what extent does model salience align with human perceived salience?

We find that LLMs have a nuanced notion of salience prioritizing information hierarchically across summary lengths (§ 4.1). Also the notion of salience is generally compatible between models even of different families and sizes, though more recent/bigger LLMs correlate more strongly with GPT-4o (§ 4.2). Furthermore, models show highly consistent behavior and hence notions of salience, but it cannot be elicited through introspection (i.e., directly prompting for salience of topics; § 5.2). Lastly, we find that model behavior only weakly aligns with human perceptions of salience (§ 5.2).

2 Method: Analyzing Content Salience

To analyze content salience we need a way to both *observe* what content models consider important (§ 2.1), and to *describe* it in an interpretable manner (§ 2.2). Figure 1 illustrates the framework.

2.1 Length-constrained Summarization as a Content Salience Probe

To elicit content-selection decisions from models, we use length-constrained summarization as a probe. Our key intuition is that, under a limited length budget, well-behaving models will drop the least important information first, while preserving the most salient content.

Summary Generation. Given a corpus D, and a set of target lengths L specified in words, we generate summaries $S = \{s_{d,l} \mid d \in D, l \in L\}$ for all documents and length targets. We consider $L = \{10, 20, 50, 100, 200\}$ to capture a range of typical summary lengths.

Tracing Content-selection Decisions. To understand how summary content changes with varying length budgets, we introduce *Content Salience Maps (CSMs)* as a structured representation to systematically track the inclusion and exclusion of topics. Formally, let T be a set of topics of interest, and let $f : T \times S \to [0, 1]$ be a function that measures to what extent topic t is present in summary s. For a document d, CSM(d) is a $|T| \times |L|$ matrix, where each entry is defined as:

$$\text{CSM}(d)_{t,l} = f(t, s_{d,l}). \tag{1}$$

We define the corpus-level $\text{CSM}(D)$ as the average of document-level measurements:

$$\text{CSM}(D)_{t,l} = \frac{1}{|D^t|} \sum_{d \in D^t} f(t, s_{d,l}), \qquad (2)$$

where D^t is the set of documents that contain topic t. We also define *topic prevalence* as $|D^t|/|D|$, representing the fraction of documents in the corpus that contain topic t.

Below, we describe a concrete instantiation of this framework, where the set of topics T is represented as QUDs, and the inclusion measure f is defined as question answerability. However, we note that the framework is highly customizable in terms of the definitions of T and f.

2.2 Question-based Content Analysis

We represent the topics $t \in T$ as Questions Under Discussion (QUD), a linguistic representation for topics in discourse (Van Kuppevelt, 1995). In our setup, each QUD represents a possible *answer space* across different documents in the same genre. This aligns with alternative semantics, where questions are viewed as the set of possible answers (Hamblin, 1973; Karttunen, 1977; Groenendijk and Stokhof, 1984). In addition to the interpretability provided by natural language questions, we can also quantify content salience through *question answerability*: questions which remain answerable even with shorter summaries are more salient than questions which can only be answered with longer summaries. Below (also Algorithm 1), we describe a four-step pipeline to implement this approach.

Step 1: Question Generation. We design a question-generation prompt inspired by (Laban et al., 2022). Given summaries of varying lengths from a random sample of documents, we prompt an LLM to generate n questions which each summary answers in a unique way. The prompt specifies two requirements: (1) the questions should be answerable by most documents in the given genre, and (2) they should highlight meaningful differences between summaries of different lengths (full prompt in Appendix G). For example, in movie reviews, most summaries will answer questions such as *"What is the main plot of the movie?"*, but naturally the answers will be different for each review. We repeat this process for all documents and associated summaries in the corpus.[2]

Algorithm 1 Content Salience Map (CSM) Derivation

```
Input:  Corpus: $D = \{d_1, d_2, ..., d_{|D|}\}$
        Lengths: $L = \{10, 20, 50, 100, 200\}$ (words)
        Models: $M_{\text{Sum}}, M_{\text{QG}}, M_{\text{Emb}}, M_{\text{QA}}, M_{\text{ClaimSplit}}, M_{\text{NLI}}$
Output: Corpus-level $\text{CSM}_D$
for $(d, l) \in D \times L$ do                         ▷ Step 0: Summarization
    $S[d, l] \leftarrow M_{\text{Sum}}(d, l)$
for $d \in D$ do                                       ▷ Step 1: Question Generation
    $Q \leftarrow Q \cup M_{\text{QG}}(d, S[d, :])$
$T \leftarrow$ Cluster($M_{\text{Emb}}(Q)$)            ▷ Step 2: Question Clustering
$T \leftarrow$ ManualReview($T$)
$T \leftarrow$ SelectClusterRepresentatives($T$)
for $(d, t) \in D \times T$ do                         ▷ Step 3: QA and Claim Split
    $\text{ans}_{\text{ref}} \leftarrow M_{\text{QA}}(d, t)$
    if $\text{ans}_{\text{ref}} \neq \varnothing$ then $A[d, t] \leftarrow M_{\text{ClaimSplit}}(\text{ans}_{\text{ref}})$
    else $A[d, t] \leftarrow \varnothing$
for $(t, l) \in T \times L$ do                         ▷ Step 4: Answerability
    for $d \in D$ do
        $s, A_t \leftarrow S[d, l], A[d, t]$           ▷ (summary, claims)
        $\text{CSM}_d[t, l] \leftarrow \text{avg}([M_{\text{NLI}}(a, s) \mid a \in A_t])$ ▷ Eq. 3
    $D_t \leftarrow \{d \in D \wedge A[d, t] \neq \varnothing\}$
    $\text{CSM}_D[t, l] \leftarrow \text{avg}([\text{CSM}_d[t, l] \mid d \in D_t])$   ▷ Eq. 2
return $\text{CSM}_D$
```

Step 2: Clustering. We then cluster questions with the same semantic intent. For instance, *"Is the soundtrack effective?"* and *"How does the music contribute to the film's atmosphere?"* are considered equivalent, as they ask for the same information. We select the question closest to the mean embedding of each cluster as its representative. These questions form the topics T.

Step 3: Question-Answering and Claim Decomposition. For each `(original document, question)` pair, we first obtain a *reference answer* using a QA-model. We then decompose each answer into a set of atomic claims A_t (see Figure 1 for an example). These claims support the answerability calculation (described next), and a fine-grained analysis of summary similarity and consistency through claim entailment patterns (§ 4.2).

Step 4: Answerability Estimation. We measure how well a summary answers a question by the fraction of reference answer claims it entails. This naturally accounts for questions that are only partly answerable with a given summary. Formally, let A_t be the set of answer claims for a given question. The answerability score is then calculated as:

$$f(t, s) = \frac{1}{|A_t|} \sum_{a \in A_t} e(a, s), \qquad (3)$$

where $e : A \times S \rightarrow \{0, 1\}$ is a natural language inference (NLI) model that determines if claim a is entailed (1) or not entailed (0) by summary s. This

[2] We ran question generation over GPT-4o, Llama 3.1 (8B), and Mistral summaries. As the resulting questions were highly similar, we did not include additional models.

Statistic	RCT	CL	Astro	QMSum
Documents	200	185	106	90
Words/doc	290	459	703	10,837
Questions	21	14	13	10
Answered/doc	84.1%	86.2%	96.5%	91.9%
Words/answer	30.9	53.0	70.3	161.5
Claims/answer	6.5	11.4	12.4	29.6
Claims (total)	23,124	25,353	16,430	24,459

Table 1: Dataset overview. Number of words is calculated as whitespace-separated tokens.

practice of claim-entailment is commonly used in similar settings such as fact checking (Kamoi et al., 2023; Min et al., 2023; Stacey et al., 2024).[3]

Implementation. For question generation, we found it necessary to use a strong model (i.e., GPT-4o). For clustering, following Lam et al. (2024), we represent questions using sentence embeddings (Reimers and Gurevych, 2019), followed by a dimensionality reduction and density-based clustering with HDBSCAN (McInnes et al., 2017) which requires minimal parameter tuning and does not presuppose a fixed number of clusters.[4] After an initial round of clustering, we found several overlapping clusters which were merged manually. For question-answering and answer-claim splitting, we use Llama 3.1 8B (see Appendix G for the prompts). For claim entailment, we use the efficient `MiniCheck` (Tang et al., 2024).

3 Experimental Settings

Datasets. We analyze LLM salience across several technical and scientific domains using four datasets (Table 1). We designed slightly unconventional summarization tasks because of their limited "oracle" summaries in common LLM training data. This allows us to analyze how LLMs handle texts without strong priors, and how salience judgments vary across genres and discourse types (technical writing, academic discourse, and dialogue).

(1) Randomized Controlled Trials (RCT). We draw a random sample of 200 abstracts of RCTs published Jan–Apr 2024 from PubMed. These documents follow established conventions to describe the conduct and outcomes of clinical studies. The task is to further summarize the abstracts.

(2) Computation and Language (CL). The second task is to summarize the *related work* sections of NLP/CL papers published on arXiv. Although CL paper summarization is common, summarizing the related work section itself is not. We convert raw LaTeX sources to Markdown and only consider documents up to 2,000 tokens to fit the context window of smaller models. A random sample of 185 documents published in October 2024 is drawn.

(3) Astrophysics (Astro). The third dataset contains *discussion* sections of astrophysics papers published on arXiv. These documents interpret key results of theoretical and empirical astrophysics research. Similar to the CL portion, summarizing only the discussion sections is uncommon. A random sample of 106 documents is drawn, with pre-processing analogous to CL.

(4) Meetings (QMSum). Lastly, we consider meeting transcript summarization. We randomly sample 90 documents balanced across three domains from QMSum (Zhong et al., 2021): product design, research and political discussions. We format transcripts as `[Speaker]: [Utterance]` turns, separated by newlines. We only experiment with long-context models ($\geq$ 32k tokens) on this dataset.

Summarization Models. We experiment with 13 LLMs of different scales: **OLMo** (7B; 02/24, 07/24; Groeneveld et al., 2024), **Mistral** (7B; v0.3; Jiang et al., 2023), **Mixtral** (8x7B; v0.1, Jiang et al., 2024), **Llama 2** (7B, 13B, 70B; Touvron et al., 2023), **Llama 3** (8B, 70B), and **Llama 3.1** (8B, 70B; Grattafiori et al., 2024). For API-based models, we use **GPT-4o-mini** (07/24) and **GPT-4o** (08/24; OpenAI et al., 2024). We also include 3 baselines to contextualize results: **Lead-N**, **Random** and **TextRank** (Mihalcea and Tarau, 2004), all adjusted to meet summary length budgets. To assess consistency across multiple rounds of decoding, we generate 5 summaries per document and target length with temperature $\tau = 0.3$. We use a zero-shot summarization prompt (Appendix G).

Before analyzing salience in these models, we validate two key assumptions: *(i)* generated summaries should approximately meet the target length, and *(ii)* longer summaries should expand on shorter ones ("incremental consistency"). Additionally, we analyze how greater τ affect those criteria. Our analysis confirms that models largely meet above criteria, with newer and bigger models showing better length control. Higher τ results in stable

[3] Since longer answers tend to include more claims, answer length may affect salience scores. In practice, we observe a weak negative relationship (see discussion in Appendix D).

[4] We use `all-mpnet-base-v2` for sentence embeddings, UMAP for dimensionality reduction, and cluster with HDBSCAN (leaf clustering, min size = 15, defaults: $\epsilon = 0$, $\alpha = 1$).

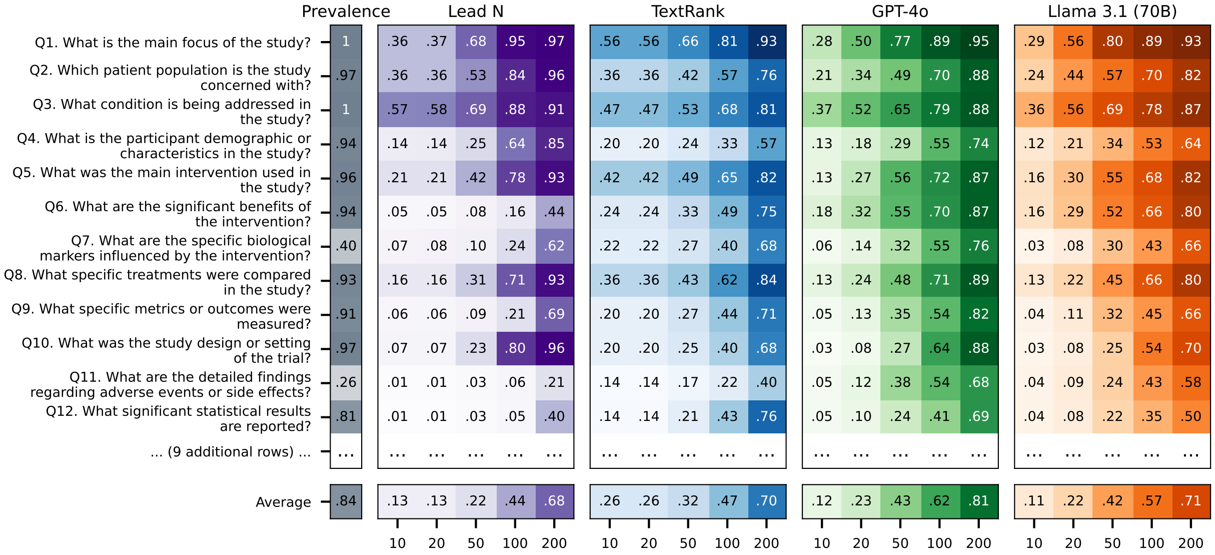


Figure 2: Corpus-level content salience map for *RCT* summaries by four methods (continued in Figure 11).

average summary length at the corpus level, but greater length variance at the document level (up to 10% difference), along with a slight decline in incremental consistency (details in Appendix A).

4 Observed Salience

4.1 RQ1: What notion of salience have LLMs learned in different domains?

To understand how LLMs prioritize different information, we consider average question answerability as a proxy for salience. We show the results for the *RCT* dataset as a representative example in Figure 2, and include other datasets in Appendix B.

Models prioritize information hierarchically. We observe a clear hierarchy in how information is prioritized across summary lengths. For example, fundamental aspects such as the focus of a study (*Q1*), and the condition being treated (*Q3*) consistently achieve higher scores, even at 10-word summaries. In contrast, more specific and technical information like the study design (*Q10*) and the statistical significance of results (*Q12*) are primarily discussed in longer summaries ($\geq$ 100 words).

Information frequency is not in itself predictive of salience. When we consider how frequently a question is answered by documents in the corpus (leftmost column of Figure 2), we find that even relatively rare questions such as biological markers and adverse effects (*Q7/11*, prevalence 40%/26%) maintain a consistent representation in summaries. This suggests that LLMs do not simply prioritize information based on its frequency in a genre.

Summaries progressively get more detailed, and information density differs across models. As expected, longer summaries consistently include more information as shown by the higher average answerability (bottom row in Figure 2). However, the absolute scores differ across models. GPT-4o has a notably higher answerability score than Llama 3.1, particularly at longer summaries (0.81 vs. 0.71 at the 200-word length). Given that both models generate summaries of similar lengths (cf. Figure 5), this suggests that GPT-4o conveys information more efficiently.

4.2 RQ2: Do LLMs of different families and sizes have a similar notion of salience?

We want to understand to what extent different models (e.g., families, scales) have a shared notion of information salience in a given domain. We define a fine-grained similarity metric that compares models' content-selection decisions.

Intuitively, two models are more similar if their summaries include the same answer claims. More formally, for each summary length l, we compile all atomic claims derived from question-answers along with their entailment labels (cf. § 2.2). These form a binary vector $\mathbf{v}_{M,l}$ indicating which claims model M includes in its summaries. We then measure agreement between two models using Krippendorff's alpha: $\alpha(\mathbf{v}_{M_1,l}, \mathbf{v}_{M_2,l})$. This claim-level agreement metric is stricter than comparing aggregate answerability scores, as it requires models to consistently include or exclude the same claims at

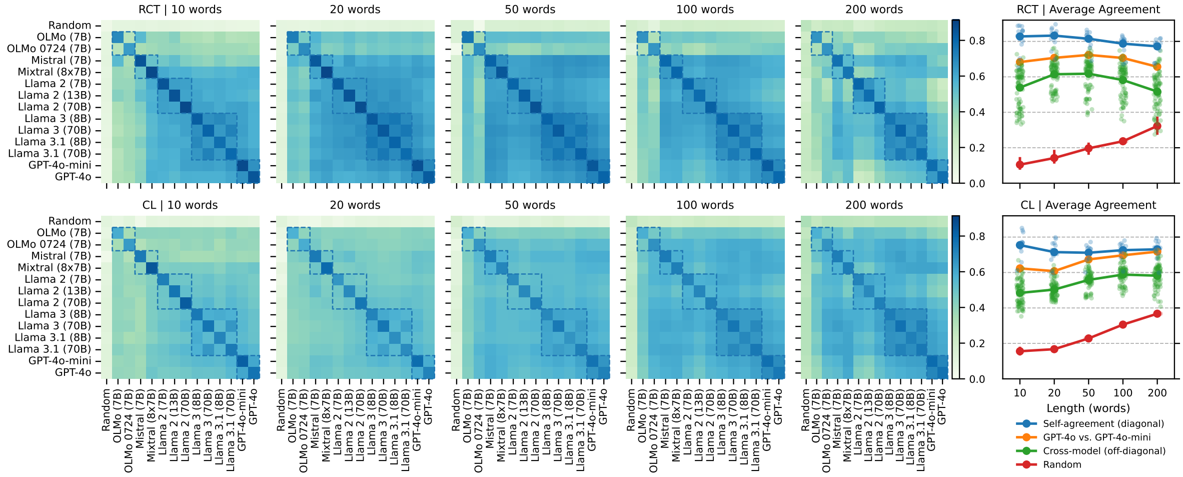


Figure 3: Do LLMs share a similar notion of salience? Heatmaps show agreement of content-selection at the atomic-claim level (Krippendorff's α). Dashed bounding boxes indicate models of the same family. The diagonal shows self-agreement over multiple generations. Top row: *RCT*, Bottom row: *CL*.

each summary length.[5] Figure 3 shows the model-model agreement for the *RCT* and *CL* datasets.

High agreement across multiple runs suggests models apply salience notion consistently. The diagonal in Figure 3 shows the average pairwise agreement across 5 model runs. Overall, self-agreement is the highest for *RCT* ($\approx$ .80), while it is slightly lower for *CL*, *Astro* and *QMSum* ($\approx$ .75). We observe a slight decline in self-agreement as the summary length increases. We hypothesize that each document has a tail of medium- to low-salient topics which may or may not be included as the length budget gives more "freedom" to the models.

In sum, high self-agreement suggests that models apply salience consistently, which is beneficial for downstream users who depend on predictable summarization behavior. Additionally, this result serves as a validation of our method and enables the following cross-model analyses which would be meaningless without high self-agreement.

Models of the same family or size do *not consistently* have a higher agreement than any other model. We next inspect the off-diagonal agreements, comparing one model family with another model family. Overall, we find that within-family agreement is not consistently higher than cross-family agreement. While there are isolated cases of a higher within-family agreement (e.g., Llama 3.1 and GPT-4o on *RCT*), this trend cannot be confirmed for all families and datasets.

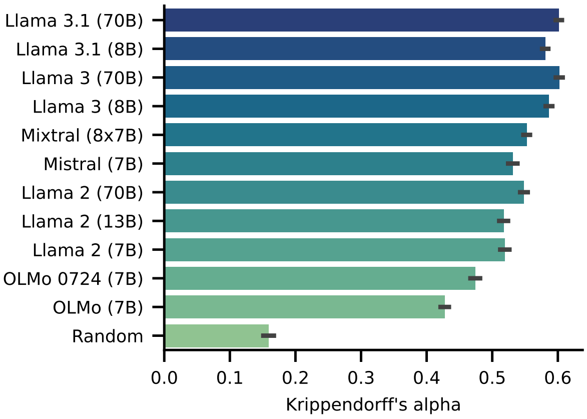


Figure 4: Agreement with GPT-4o-mini, averaged over all datasets and summary lengths.

Agreement by summary length and with GPT-4o-mini. We observe that certain summary-lengths have higher agreement than others, though the peak is different for each dataset (e.g., agreement on *RCT* is highest for 50 word summaries, whereas on *CL* it peaks at 100 words). There could be a "natural" summary length for each dataset where model more easily agree. Lastly, we find that more recent and bigger models agree better with GPT-4o-mini which suggests a clear scaling effect and that open-weights models are getting closer in capabilities to large proprietary models (Figure 4).

5 Perceived Salience and Alignment

In addition to the *observational salience* analysis, we elicit *perceived salience* by having humans and models directly rate the salience of each question. This study has two purposes: (1) to understand whether model behavior aligns with human expectations, and (2) to see if the summarization behavior of LLMs can be approximated by direct prompting.

[5] In contrast, similar answerability scores can result from selecting a similar *number* of claims.

5.1 Setup

Human salience annotation. We recruited 18 experts across the four domains through our network (3 for *RCT*, and 5 each for *Astro, CL, QMSum*).[6] Experts rated the relative salience of each question on a 5-point Likert scale (ranging from 1: least important, to 5: most important). Annotators were asked to motivate their rating through a brief rationale to encourage thoughtful judgments and to allow post-hoc analysis of their decision-making process. To establish a shared understanding between annotators of what content a question may elicit, each question is accompanied by an example answer from a randomly drawn document in the domain. To ensure high annotation quality, we conducted two pilot rounds with four annotators to refine our annotation guidelines (see Appendix H).

Importantly, the human annotations cannot be regarded as a gold standard for salience. The ratings represent how humans *perceive* question salience, which may not be reflective of how humans actually write summaries. As an initial step toward analyzing human salience through summarization behavior, we explore the application of our framework to human-written summaries in Appendix E.

Model-based salience ratings (LLM-perceived). We prompt LLMs to directly rate question salience. The prompt includes the question list for a given domain and instructions that closely mirror the human annotation guidelines to allow for direct comparison (i.e., 5-point Likert scale and rationales). Each model is prompted 5 times with a shuffled question list to mitigate position bias and to quantify consistency. See Appendix G for the full prompt.

Analysis method. We use Spearman's rank correlation coefficient (ρ) to quantify alignment between three measures: human-perceived salience, LLM-perceived salience (both 5-point Likert), and LLM-observed salience (continuous $[0, 1]$).[7] For groups with multiple ratings, we report averaged pairwise correlation and test for statistical significance with the harmonic-mean p-value (Wilson, 2019).

Human correlation. Inter-human correlation varies by domain, with meeting summarization (QMSum, $\rho = 0.60$) and RCT abstracts ($\rho = 0.46$) showing a moderate to strong correlation (Table 2). These domains presumably have established conventions about summary content. In contrast, correlation is weak for summarization of related work (CL, $\rho = 0.26$) and discussion sections (Astro, $\rho = 0.16$). Documents in these domains may vary significantly in the type of content they present (i.e., certain questions may be more relevant to theoretical vs. empirical papers). While our annotation protocol aims to control for this aspect through the example answers by question, there remains annotator subjectivity related to their personal interests.

Dataset	Questions	Raters	Spearman	Std.
QMsum	10	5	0.60^{*}	0.18
RCT	21	3	0.46^{*}	0.06
CL	14	5	0.26^{**}	0.29
Astro	13	5	0.16	0.44

Table 2: Inter-annotator correlation for question salience rating. Significance: * ($p < 0.05$) and ** ($p < 0.01$).

5.2 Results

To understand if LLMs can reliably rate question salience, we study three conditions. First, as a reference point, we measure consistency of the observational and perceived salience measures estimated over 5 model runs (LLM-observed, LLM-perceived). Second, we study the correlation of LLM-perceived and LLM-observed to measure if models' explicit ratings align with their summarization behavior (RQ3). Third we correlate LLM-derived salience in human perceived salience (RQ4). We report results for the three conditions in Table 3 and provide qualitative examples in Table 4.

RQ3: When models introspect, does their perceived notion of salience align with their summarization behavior? LLMs have strong and consistent *implicit* notions of salience, but they are unreliable when explicating these preferences in rating tasks. We detail these observations below.

Observational salience is highly stable. We find that observational question salience leads to highly stable scores for all models ($\rho \geq 0.98$). This suggests that LLMs' underlying summarization process is highly deterministic despite the stochastic nature of language models. Also, it suggests that our proposed approach is a reliable tool for analyzing model behavior.

Models fail to have consistent perceived salience. We find that the consistency of direct

[6] Trained physicians (*RCT*), graduate students/faculty (*Astro*), and graduate students (*CL, QMSum*) based in US/Europe.

[7] We take observed salience scores at the 200-words summary length which correlated on average most strongly with human salience. Other scores are explored in Appendix C.

Measure	**Random**	**OLMo**	**Mixtral**	**Llama$^{3.1}_{8b}$**	**Llama$^{3.1}_{70b}$**	**4o-mini**	**4o**	**Average**
				Consistency of Salience Estimates				
LLM-perceived	–0.05	0.20*	0.54**	0.37*	0.71**	0.73**	**0.76****	0.47**
LLM-observed	0.92**	**0.99****	**0.99****	0.98**	**0.99****	0.98**	0.98**	0.97**
				Correlation of Salience Estimates				
LLM-perceived vs. -observed	0.03	0.12	0.37*	0.36**	0.47*	**0.56****	0.50*	0.35*
				Correlation of Model and Human Salience				
LLM-perceived vs. Human	0.07	0.16	0.41*	0.31*	0.46**	0.51**	**0.53****	0.35**
LLM-observed vs. Human	0.20	0.25	0.33*	0.35*	**0.36***	0.34*	0.25	0.30*

Table 3: Spearman rank correlations between salience estimates, averaged across datasets. Per-dataset values in Table 6. Significance: * ($p < 0.05$) and ** ($p < 0.01$); row-wise maxima are bolded.

Question	$L_o/L_p/H$	🤖 **Rationale LLM-perceived**	👥 **Rationales Human**
(#1, RCT 🧑‍⚕️) What significant statistical results are reported?	2.5/4.0/5.0	Statistical results are essential for assessing the validity and reliability of the findings.	(H$_{\{1,2,3\}}$ = 5) Main outcome that physicians are looking for; very important to be included as measure of success; without it reporting is lacking
(#2, Astro ★) What broader context or field does the study contribute to or address?	4.6/3.4/2.4	Understanding the broader context helps situate the study within the larger scientific discourse.	(H$_{\{2,5\}}$ = 1) not relevant, experts would fill context; (H$_1$ = 3) inferable from another question; (H$_3$ = 4) context is important although not always addressed; (H$_4$ = 3) Situating findings within literature, existing models or theories is important
(#3, RCT 🧑‍⚕️) What was the main intervention used in the study?	4.2/4.8/5.0	The main intervention is central to understanding the study's focus and findings.	(H$_{\{1,2,3\}}$ = 5) determines whether study influences decision making; part of PICO therefore fundamental; essential for communicating purpose/focus of the trial

Table 4: Example questions, salience scores by LLM-observed (L_o, rescaled to 1-5), LLM-perceived (L_p), humans (H) and summarized rationales. Additional examples in Table 7.

salience ratings varies significantly for all models and datasets. Generally, strong instruction-following models have more consistent perceived salience than weaker models (avg. ρ ranges from 0.20 for OLMo to 0.76 for GPT-4o). This finding mirrors recent results in the LLM-as-a-judge literature which demonstrated instability in ratings due to various factors including position bias (Wang et al., 2024; Stureborg et al., 2024).

Perceived $\neq$ observed salience. Lastly, we find only a weak to moderate correlation between perceived and observed salience (highest: avg. $\rho = 0.56$ for GPT-4o-mini, lowest: $\rho = 0.12$ for OLMo). Again, stronger instruction-following models show higher correlations, indicating a clear scaling effect. This gap echoes broader findings where generative abilities may not reflect an underlying understanding in models (West et al., 2024).

RQ4: To what extent does model salience align with human perceived salience? We find that both LLM-salience estimates only show a weak to moderate correlation with human salience perception. Direct rating for question salience correlates more than observed salience (highest LLM-perceived: avg. $\rho = 0.53$ for GPT-4o, highest LLM-observed: avg. $\rho = 0.36$ for Llama 3.1 70B). Weak correlation between models and humans holds for all dataset, also those where humans agree more strongly among themselves (Table 6).

In sum, LLM users should carefully consider if a model is appropriate for their summarization task, or provide explicit signals about content priority through prompts or during model training.

6 Related Work

Evaluating and Interpreting Summarization. Recent work suggests that LLMs match or surpass human performance in news summarization (Zhang et al., 2024). However, traditional evaluation protocols remain unreliable especially for LLM-generated summaries (Fabbri et al., 2021; Goyal et al., 2023). This spurred interest in analyzing summarization model behavior. Studies found biases towards content near the beginning/end of documents (Ravaut et al., 2024; Laban et al., 2024). Others analyze training dynamics of summarization models to identify when skills like content selection are learned (Goyal et al., 2022). Extract-then-abstract pipelines (Gehrmann et al., 2018; Li et al., 2021) aim for interpretable text summarization but this interpretability is limited to the document-level (Dhaini et al., 2024). Our research complements prior work by providing a *global in-*

terpretation of what topics LLMs consider important through the lens of text summarization.

Explainable Topic Modeling. Our analysis method draws inspiration from the interpretable topic modeling literature. While classical topic models such as LDA (Blei et al., 2003) have long been used to explain latent themes in text corpora, they are often difficult to interpret (Chang et al., 2009). Recent work showed that LLMs can effectively be used to generate natural language descriptions of latent themes in text mining, clustering and concept induction workflows (Pham et al., 2024; Zhong et al., 2024; Wang et al., 2023; Lam et al., 2024). Our framework uses LLMs to describe salient summary content in form of information-seeking QUDs. The use of QUDs as a representation of information units was shown successful in a wide range of tasks (Newman et al., 2023; Laban et al., 2022; Trienes et al., 2024; Wu et al., 2023b). Finally, in the context of summarization, our work shares theoretical foundations with Wu et al. (2024), who explore human curiosity through inquisitive QUDs. They observe that answering salient questions is a quality indicator for news summaries.

7 Conclusion

We propose an interpretable framework to systematically derive and analyze LLMs' notion of information salience, a previously latent concept that is nonetheless crucial for text synthesis applications. Our work builds on two key ideas: using length-controlled summarization as a behavioral probe for content selection, and describing what is salient as the answerability of questions. We found that LLMs have a highly consistent notion of salience which is largely compatible across models. We further found that LLMs cannot directly rate the salience of questions, and that model salience weakly aligns with human perceptions. Our work opens new directions to study how LLM salience emerges during training, and for diagnosing content selection challenges in text synthesis tasks.

Limitations

We consider zero-shot prompting with temperature-based decoding to generate summaries. While these settings are common defaults for LLM users, it is conceivable that different prompting styles (e.g., chain-of-density) or decoding methods influence salience patterns. Future work should explore how these techniques affect salience, particularly in adjacent information-seeking tasks such as query-based summarization.

While our experiments cover diverse disciplines (medicine, astrophysics, computational linguistics, and meetings) and discourse types (structured writing, academic discourse, and dialogue), the texts are primarily technical. Since our framework is designed to be domain-agnostic, we believe it is an exciting direction for future work to explore less technical genres such as fiction (Kim et al., 2024).

Our user study assumed a uniform background and interests among participants, which is a simplification of practical applications. Additionally, the specialized nature of two tasks (i.e., summarizing related work and discussion sections) may have contributed to variability in responses, as even domain experts may not have strong priors on how these texts should be summarized. Future work could explore how differences in expertise and prior knowledge shape perceptions of salience.

Acknowledgments

We thank Hsin-Pei Chen, Khawla Elhadri, Arya Farahi, Juan P. Farias, Cheng Han Hsieh, Sebastian Joseph, Ramez Kouzy, Michael Muzinich, Van Bach Nguyen, Juan Diego Rodriguez, Paul Torrey, Manya Wadhwa, Barry Wei, and Paul Youssef for their participation in the salience annotation study. We also thank Dennis Aumiller and Philippe Laban for early feedback on this research. This work was partially supported by the US National Institutes of Health (NIH) grant 1R01LM014600-01, and the US National Science Foundation grants IIS-2107524, IIS-2145479, and Cooperative Agreement 2421782 and Simons Foundation MPS-AI-00010515 (NSF-Simons AI Institute for Cosmic Origins[8]).

References

Anton Benz and Katja Jasinskaja. 2017. Questions under discussion: From sentence to discourse.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22.

[8] CosmicAI, https://www.cosmicai.org/

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686.

Mahdi Dhaini, Ege Erdogan, Smarth Bakshi, and Gjergji Kasneci. 2024. Explainability meets text summarization: A survey. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 631–645.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of GPT-3. *Preprint*, arXiv:2209.12356.

Tanya Goyal, Jiacheng Xu, Junyi Jessy Li, and Greg Durrett. 2022. Training dynamics for text summarization models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2061–2073.

Aaron Grattafiori et al. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jeroen Antonius Gerardus Groenendijk and Martin Johan Bastiaan Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, Univ. Amsterdam.

Dirk Groeneveld et al. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809.

Charles L Hamblin. 1973. Questions in montague english. *Foundations of language*, 10(1):41–53.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*.

Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 570–593.

Albert Q. Jiang et al. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang et al. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583.

Lauri Karttunen. 1977. Syntax and semantics of questions. *Linguistics and philosophy*, 1:3–44.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. FABLES: Evaluating faithfulness and content selection in book-length summarization. In *First Conference on Language Modeling*.

Philippe Laban, Alexander Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context LLMs and RAG systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9885–9903.

Philippe Laban, Chien-Sheng Wu, Lidiya Murakhovs'ka, Xiang Chen, and Caiming Xiong. 2022. Discord questions: A computational approach to diversity analysis in news coverage. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5180–5194.

Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using LLooM. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24.

Haoran Li, Arash Einolghozati, Srinivasan Iyer, Bhargavi Paranjape, Yashar Mehdad, Sonal Gupta, and Marjan Ghazvininejad. 2021. EASE: Extractive-abstractive summarization end-to-end using the information bottleneck principle. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 85–95.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156.

Inderjeet Mani and Mark T Maybury. 1999. *Advances in automatic text summarization*. MIT press, Cambridge, MA.

Daniel Marcu. 1999. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, pages 123–136.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11).

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Narges Nazari and MA Mahdavi. 2019. A survey on automatic text summarization. *Journal of AI and Data Mining*, 7(1):121–135.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. *Mining text data*, pages 43–76.

Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A question answering framework for decontextualizing user-facing snippets from scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3194–3212.

OpenAI et al. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073.

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Oana-Maria Camburu, and Marek Rei. 2024. Atomic inference for NLI with generated facts as atoms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10188–10204.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *Preprint*, arXiv:2405.01724.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jan Trienes, Sebastian Joseph, Jörg Schlötterer, Christin Seifert, Kyle Lo, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2024. InfoLossQA: Characterizing and recovering information loss in text simplification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4263–4294.

Jan Van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of Linguistics*, 31(1):109–147.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.

Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023. Goal-driven explainable clustering via language descriptions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10626–10649.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin

Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2024. The generative AI paradox: "What it can create, it may not understand". In *The Twelfth International Conference on Learning Representations*.

Daniel J. Wilson. 2019. The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200.

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023a. QUDeval: The evaluation of questions under discussion discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363.

Yating Wu, Ritika Rajesh Mangla, Alex Dimakis, Greg Durrett, and Junyi Jessy Li. 2024. Which questions should I answer? Salience prediction of inquisitive questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19969–19987.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023b. Elaborative simplification as implicit questions under discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

Ruiqi Zhong, Heng Wang, Dan Klein, and Jacob Steinhardt. 2024. Explaining datasets in words: Statistical models with natural language parameters. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

A Length-instruction Following

We analyze to what extent length-controlled summarization is a consistent probe for content selection. Ideally, we expect the following behavior of summarization models: (1) the generated summary length matches approximately the target length, and (2) as we increase the length budget, summaries should provide all content of the shorter version in addition to expanding on it. We define two measures for these desiderata.

Target length ratio (TLR). We quantify the length deviation of a generated summary (s_l) from the target word count (l) as follows:

$$\textbf{TLR}(s_l) = \frac{|s_{d,l}|}{l}. \tag{4}$$

Where $|\cdot|$ is the summary length (whitespace separated tokens). A value of 1 indicates perfect length match, while values greater or smaller than 1 indicate over- or under-generation, respectively.

Incremental consistency (IC). Longer summaries should contain a proper superset of claims found in the adjacent shorter version. Formally, for each document d and topic t recall that we have a set of atomic claims A_t (§ 2.2). We first identify the set of claims that are entailed at least once across any summary length:

$$A_{\text{entailed}}(d,t) = \{a \in A_t \mid \exists l \in L, e(a, s_{d,l}) = 1\},$$

where e is an NLI model indicating whether claim a is entailed by summary $s_{d,l}$ of length l. Next, we determine if a claim is included consistently across increasing summary lengths (monotonicity condition).

$$e(a, s_{d,l_1}) \leq e(a, s_{d,l_2}) \; \forall l_1 < l_2$$

We then define the set of consistent claims where this condition holds:

$$\begin{aligned} A_{\text{consistent}}(d,t) = \{&a \in A_{\text{entailed}}(d,t) \\ &\mid \text{monotonicity holds } \forall l \in L\}. \end{aligned}$$

Finally, the overall incremental consistency for summaries of a corpus D is given as the fraction of consistent claims:

$$\textbf{IC}(D) = \frac{\sum_{d \in D} \sum_{t \in T} |A_{\text{consistent}}(d,t)|}{\sum_{d \in D} \sum_{t \in T} |A_{\text{entailed}}(d,t)|}. \tag{5}$$

This metric ranges from 0 to 1, where 1 indicates perfect monotonicity (longer summaries always include all information found in shorter ones).

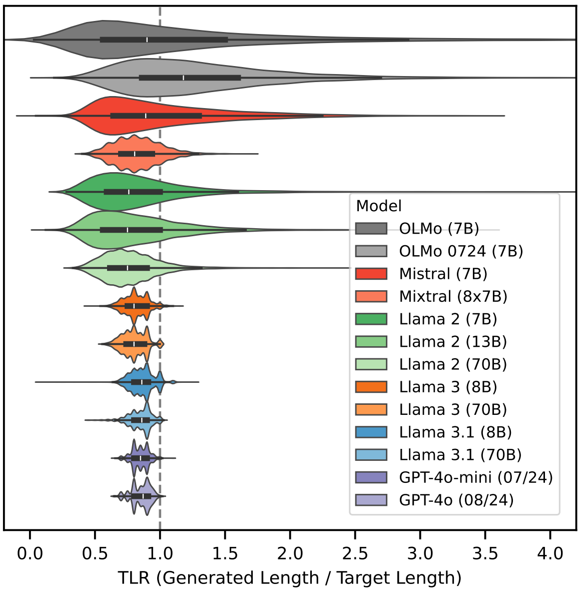


Figure 5: Distribution of target length ratios over all generated summaries (aggregating lengths and datasets).

Do models meet the target length? We find that all models generally undershoot the length target (Figure 5). However, more recent models match the target length more closely and consistently, showing a clear scaling effect. The best performing models are Llama 3.1 and GPT-4o, while OLMo is unable to follow length-instructions, presumably because this was not part of the instruction tuning data. Surprisingly, we do not find substantial differences across datasets. This suggests that the ability of models to follow length-instructions is mostly invariant to the input document length, even if they are considerably long (e.g., meeting transcripts). See Figure 10 for an analysis of summary length stratified by dataset and target length.

How incrementally consistent are summaries? We report the average incremental consistency by dataset and model in Figure 6. We observe that all models are substantially more consistent than the random summarization baseline. Furthermore, incremental consistency decreases with more difficult datasets, likely because there is more freedom on what content to include in a summary. Similar to the ability of following length instructions, we observe a scaling effect where stronger models have a higher incremental consistency.

Influence of temperature sampling. The main results in this paper are obtained with a temperature of $\tau = 0.3$. To assess how temperature affects summary length and incremental consistency, we perform a temperature sweep on the RCT dataset

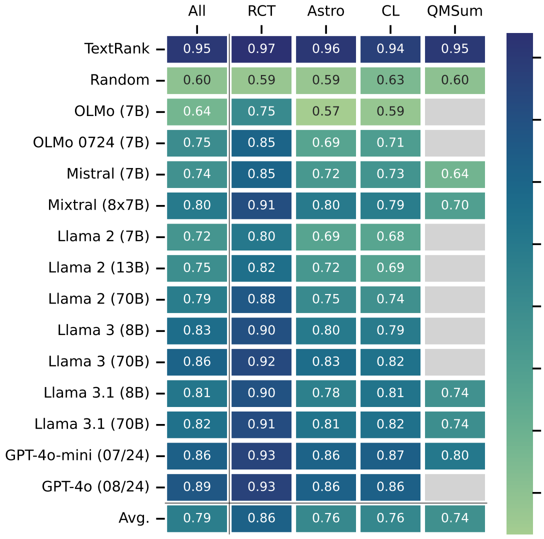


Figure 6: Incremental consistency by model and dataset.

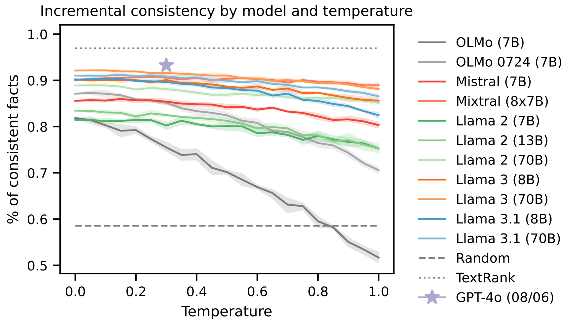


Figure 7: Incremental consistency by temperature.

for all open-weights models (20 settings in $[0, 1]$). Surprisingly, higher temperatures do not affect the *average* summary length on a dataset-level, but lead to greater variance at the document level (up to 10% length difference between generations, Figure 9). Furthermore, higher temperatures lead to a slight decline in incremental consistency for all models that adequately follow length instructions (a drop of 1% to 9%, Figure 7).

Summary. Overall, we find that strong models are able to follow length-instructions and that they consistently expand the summary content with increasing length budgets. As our salience analysis assumes this behavior of models, it may be less reliable for weaker models (OLMo, Mistral, Llama 2).

B Salience Analysis

The corpus-level salience analysis for PubMed, Astro, CL, and QMSum is given in Figure 11, Figure 12, Figure 13, and Figure 14, respectively. We also provide a fully-worked example of the content salience analysis in Figure 18.

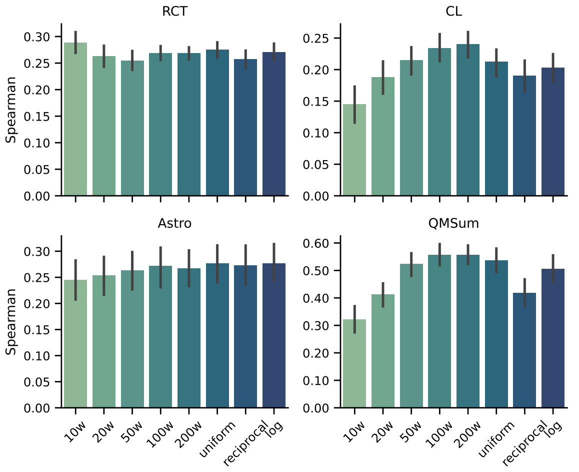


Figure 8: Correlation of different salience scores with human salience. Here we aggregate over all LLMs which showed similar trends.

C Ablation: Salience Score

We analyze how different salience scores derived from the CSM correlate with human salience. Recall that the $\text{CSM}(D)_{t,l}$ tracks the average answerability of question $t \in T$ at summary length $l \in L = \{10, 20, 50, 100, 200\}$. We take raw salience scores at each summary length. Additionally, we calculate several question-wise aggregate scores. Intuitively, questions which are more answerable at shorter summaries score higher under the aggregated scheme. Formally, we aggregate scores as follows:

$$\text{CSM}_{\text{agg}}(D)_t = \frac{\sum_{l \in L} w_l \cdot \text{CSM}(D)_{t,l}}{\sum_{l \in L} w_l},$$

where w_l is a weighting term. We experiment with three weighting functions: uniform ($w_l = 1$), reciprocal length ($w_l = 1/l$), and logarithmic decay ($w_l = 1/\log(1 + l)$).

Figure 8 shows the Spearman rank correlation coefficient (ρ) with human salience for each salience score. Overall, we find that all salience scores correlate similarly with human salience ratings on *RCT* and *Astro*, while the 200 words salience score correlates most strongly on *CL* and *QMSum*.

D Ablation: Effect of Answer Length on Question Salience

We calculate question salience as the fraction of answer claims entailed by the summary (Equation 3). Naturally, some questions can be answered succinctly (e.g., *"What is the goal of the study?"*) while others require more elaboration (e.g., *"What*

Summary	RCT	Astro	CL	QMsum
10 words	–0.03	–0.11**	–0.07**	–0.22**
20 words	–0.10**	–0.17**	–0.10**	–0.31**
50 words	–0.18**	–0.21**	–0.19**	–0.36**
100 words	–0.26**	–0.28**	–0.22**	–0.41**
200 words	–0.31**	–0.31**	–0.26**	–0.45**

Table 5: Spearman rank correlation between *answer length* and *question salience* for summaries generated by Llama 3.1 (70B). Significance: ** ($p < 0.01$).

were the detailed findings?"). This raises the question how answer length influences salience. To better understand this relationship, we compute the Spearman rank correlation between answer length (measured in whitespace-separated tokens) and question salience. Table 5 reports results for summaries generated by Llama 3.1 (70B), with similar trends for other models.

We observe a weak negative correlation between answer length and question salience for shorter summaries (10–20 words), and a weak-to-moderate negative correlation for longer ones ($\geq$ 100 words). This suggests that answer length explains some variance in question salience, but cannot fully account for it. A hypothesis is that model-generated summaries are abstractive, possibly conveying information more densely than the reference answers.

E Pilot Study: Human Salience in News Summaries

To test the generality of our framework, we use it as a tool to analyze the salience notions encoded in human-written summaries from a standard summarization benchmark. We focus on news articles from the CNN/DM dataset (Hermann et al., 2015).

Method. First, we run question generation over a sample of 200 random documents to identify QUDs for this domain (see Steps 1 and 2 in § 2.2). Next, we use the resulting questions to analyze human summaries (see Steps 3 and 4 in § 2.2). Since each article in CNN/DM has only one reference summary, we select `(document, summary)` pairs with summaries approximating our target lengths of $L = \{10, 20, 50, 100, 200\}$ words, allowing for a delta of $\pm 10\%$.[9] Finally, from each length bucket, we draw a random sample of 200 `(document, summary)` pairs for analysis.

Results. We present the content salience map for both human and model summaries in Figure 15. We observe consistent trends in question salience. For example, questions about the main event (*Q1*) or its magnitude (*Q13*) and consequences (*Q6*) consistently achieve higher scores than more detailed questions about reactions (*Q7*), expert opinions (*Q10*) or additional stakeholders (*Q3*). While the salience scores of human and model summaries are not directly comparable due to differing document samples, they exhibit similar trends.

In sum, this pilot study demonstrates the versatility of the framework, and suggests that it could be used in future work to understand human notions of salience on a larger scale.

F Responsible NLP Considerations

Compute Requirements. Experiments were conducted on NVIDIA A100 80GB GPUs, requiring approximately 20 GPU hours per dataset, and an additional 360 GPU hours for the temperature sweep on the RCT dataset, totaling 440 GPU hours. We ran inference using VLLM (`docs.vllm.ai`). GPT-4o models were accessed through the OpenAI API with inference costs $\leq$ 100$.

Salience Annotation Study. Participants joined on a volunteer basis, gave informed consent and agreed that their annotations will be shared in anonymized form in the paper repository. According to our institutional policies, this study did not require institutional review board (IRB) approval.

Data Licensing. We obtain RCT abstracts in accordance with fair use principles through the PubMed Entrez API.[10] Related work sections of CL and Astro papers were collected via the arXiv API.[11] While the majority of papers on arXiv is published under the arXiv license[12] retaining *copyright* with the original author(s), the *use* of paper contents for research is explicitly granted and encouraged in the arXiv API terms & conditions.[13] We reused meeting transcripts from QMSum (Zhong et al., 2021).[14] All meeting transcripts are under an open use license, such as CC BY 4.0 (academic meetings and product meetings) or Open Government License Version 3 (parliament commitee meetings).[15,16,17,18]

[9] Of the 287,113 documents in the CNN/DM training set, 0.28/1.60/26.84/3.50/0.04% fall into the respective buckets.

All URLs accessed 2025-05-15.

[10] `www.ncbi.nlm.nih.gov/home/develop/api/`

[11] `info.arxiv.org/help/api/index.html`

[12] `arxiv.org/licenses/nonexclusive-distrib/1.0/license.html`

[13] `info.arxiv.org/help/api/tou.html`

[14] `github.com/Yale-LILY/QMSum`

[15] `creativecommons.org/licenses/by/4.0/legalcode`

[16] `groups.inf.ed.ac.uk/ami/icsi/license.shtml`

[17] `groups.inf.ed.ac.uk/ami/corpus/license.shtml`

[18] `www.nationalarchives.gov.uk/doc/open-government-licence/`

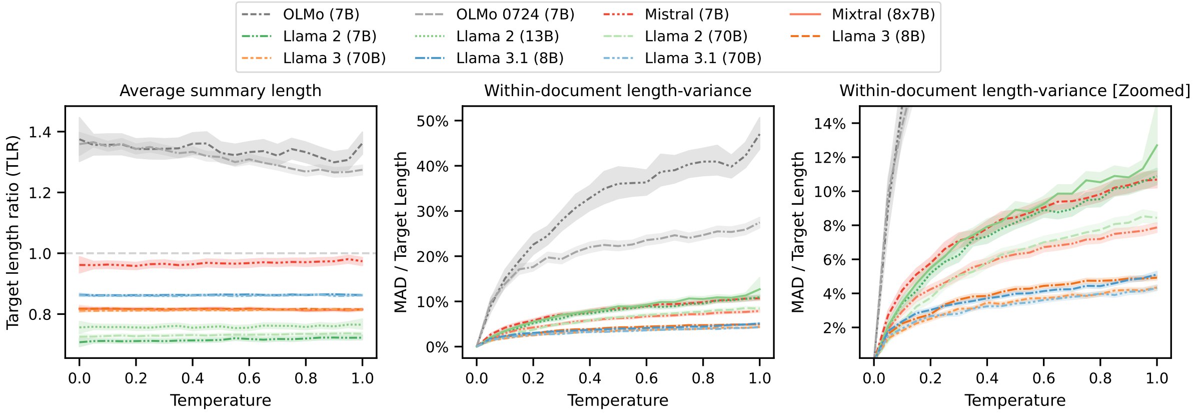
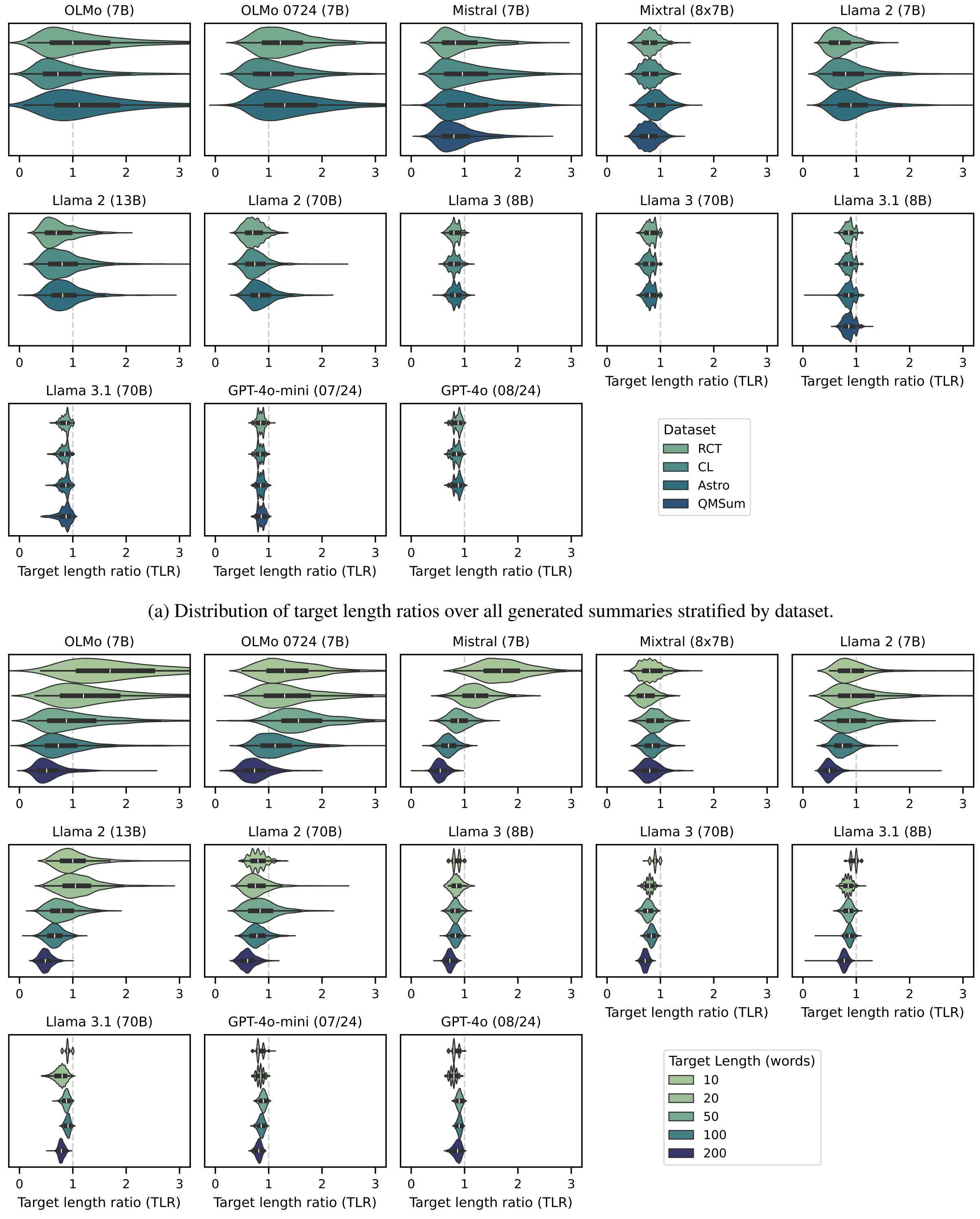


Figure 9: Influence of temperature on generated summary length. **Left:** target length-ratio. **Center:** "within-document length variance" calculated as the mean deviation from the average summary length of 5 summaries for the same document (MAD). MAD is normalized to be comparable across length targets. **Right:** zoomed version.

Measure	Dataset	Random	OLMo	Mixtral	$\text{Llama}^{3.1}_{8b}$	$\text{Llama}^{3.1}_{70b}$	4o-mini	4o	Average
				Consistency of Salience Estimates					
LLM-perceived	RCT	–0.06	0.34*	0.48**	0.29*	0.61**	0.75**	**0.80****	0.46**
	Astro	0.02	0.07	0.41*	0.56*	0.64**	0.76**	**0.86****	0.47**
	CL	–0.10	–0.05	0.65**	0.29	**0.73****	0.70**	0.57**	0.40**
	QMsum	–0.07	0.42**	0.61*	0.36	**0.87****	0.72**	0.80**	0.53**
	Average	–0.05	0.20*	0.54**	0.37*	0.71**	0.73**	**0.76****	0.47**
LLM-observed	RCT	0.94**	0.98**	**0.99****	**0.99****	**0.99****	**0.99****	**0.99****	**0.99****
	Astro	0.91**	0.99**	**1.00****	0.97**	0.99**	0.97**	0.97**	0.97**
	CL	0.96**	**0.99****	**0.99****	0.96**	**0.99****	**0.99****	0.97**	0.98**
	QMsum†	0.87**	—	0.99**	0.99**	**1.00****	0.99**	—	0.97**
	Average	0.92**	**0.99****	**0.99****	0.98**	**0.99****	0.98**	0.98**	0.97**
				Correlation of Salience Estimates					
LLM-perceived vs. LLM-observed	RCT	–0.06	0.10	0.25	0.25	0.37**	0.41**	**0.51****	0.28*
	Astro	0.11	0.09	0.31	0.56**	0.50*	**0.65****	0.58*	0.40**
	CL	–0.08	0.16	0.44	0.47*	0.38	**0.58***	0.41	0.34
	QMsum†	0.11	—	0.46*	0.16	**0.63***	0.60*	—	0.39*
	Average	0.03	0.12	0.37*	0.36**	0.47*	**0.56****	0.50*	0.35*
				Correlation of Model and Human Salience					
LLM-perceived vs. Human	RCT	–0.03	0.22	0.38**	0.34*	0.49**	0.48*	**0.56****	0.35**
	Astro	0.07	0.12	0.30**	0.31*	0.27	**0.45****	0.44*	0.28**
	CL	0.06	–0.03	0.41*	0.22	**0.48***	0.44*	0.46**	0.29**
	QMsum	0.14	0.34	0.54*	0.36*	0.62*	**0.67****	**0.67****	0.48*
	Average	0.07	0.16	0.41*	0.31*	0.46**	0.51**	**0.53****	0.35**
LLM-observed vs. Human	RCT	0.31	0.28	0.27	0.25	0.25	**0.34**	0.24	0.27
	Astro	0.11	0.25*	0.27*	0.29*	**0.31**	0.26	0.25*	0.25*
	CL	**0.30**	0.23	0.23	0.24	0.26	0.25	0.24	0.25
	QMsum†	0.16	—	0.53*	0.58**	**0.59****	0.51**	—	0.48*
	Average	0.20	0.25	0.33*	0.35*	**0.36***	0.34*	0.25	0.30*

Table 6: Spearman rank correlations between salience estimates, split by dataset. Significance: * ($p < 0.05$) and ** ($p < 0.01$); row-wise maxima are bolded. †Results for QMSum not available due to limited context window (OLMo) and budget constraints (GPT-4o).



(a) Distribution of target length ratios over all generated summaries stratified by dataset.

(b) Distribution of target length ratios over all generated summaries stratified by target summary length.

Figure 10: Analysis of length-instruction following. The target length ration (TLR) indicates to what extent models match the provided length. A value of 1 indicates perfect length match, while values greater or smaller than 1 indicate over- or under-generation, respectively.

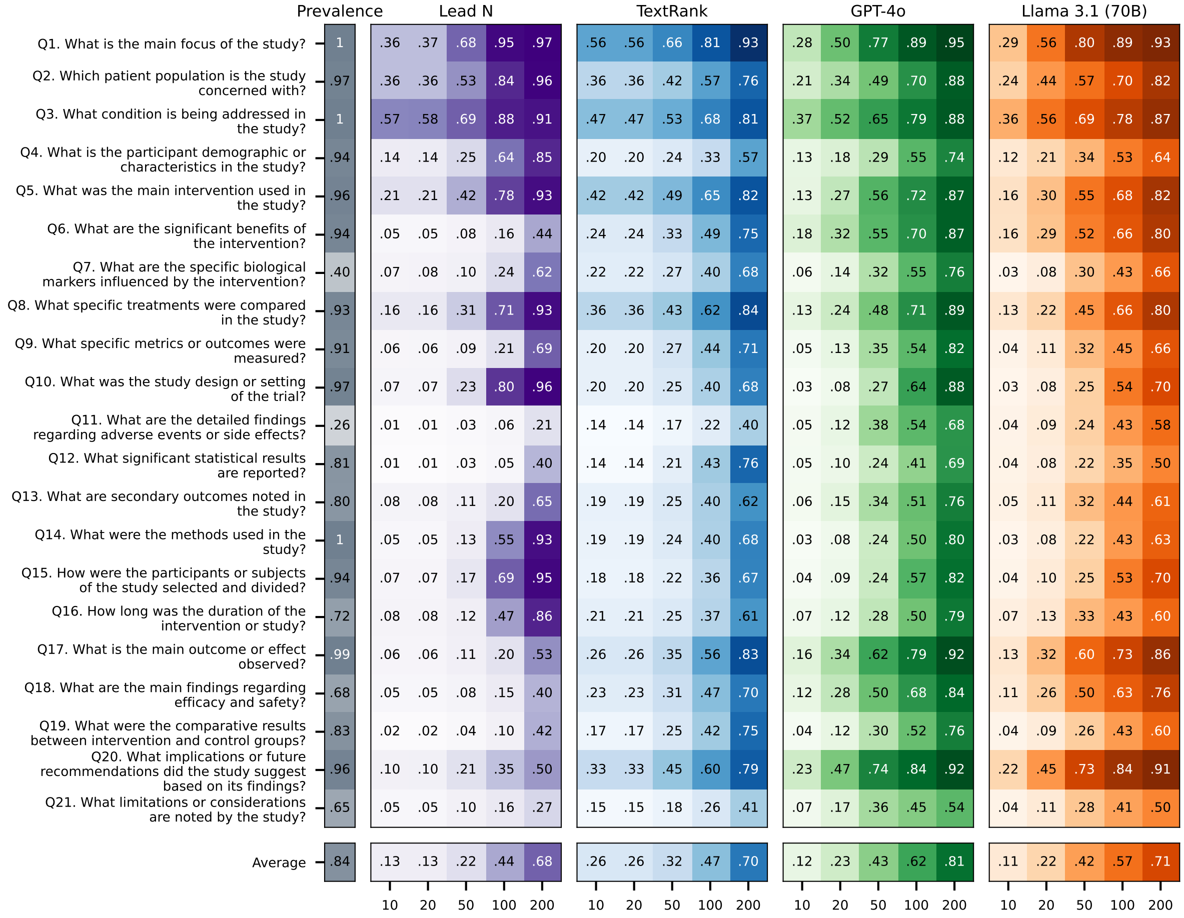


Figure 11: Corpus-level content salience map for *RCT* summaries by four methods.

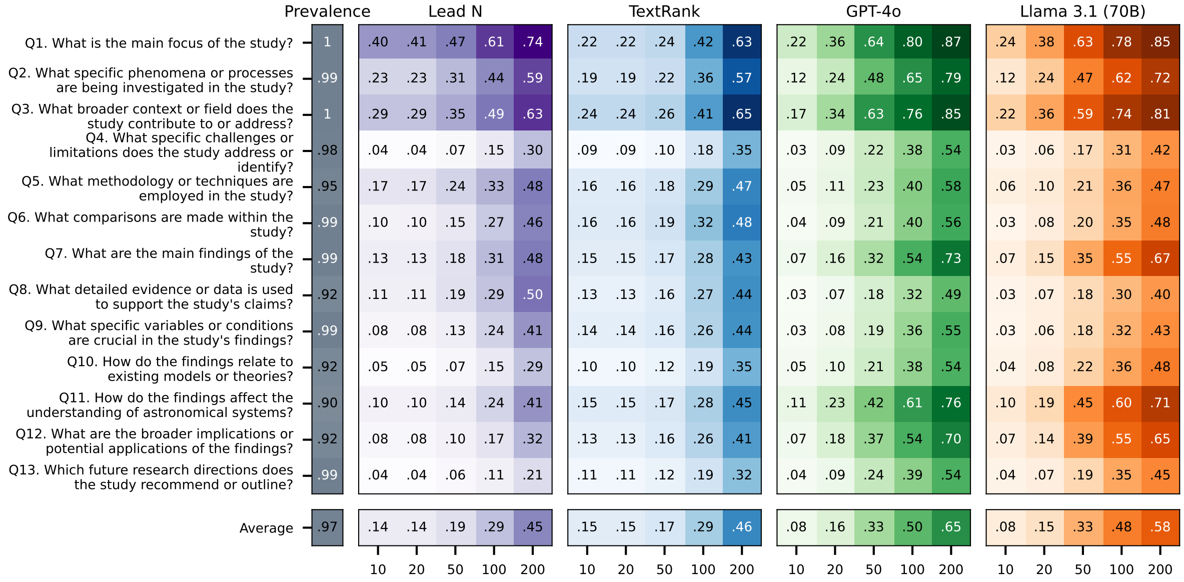


Figure 12: Corpus-level content salience map for *Astro* summaries by four methods.

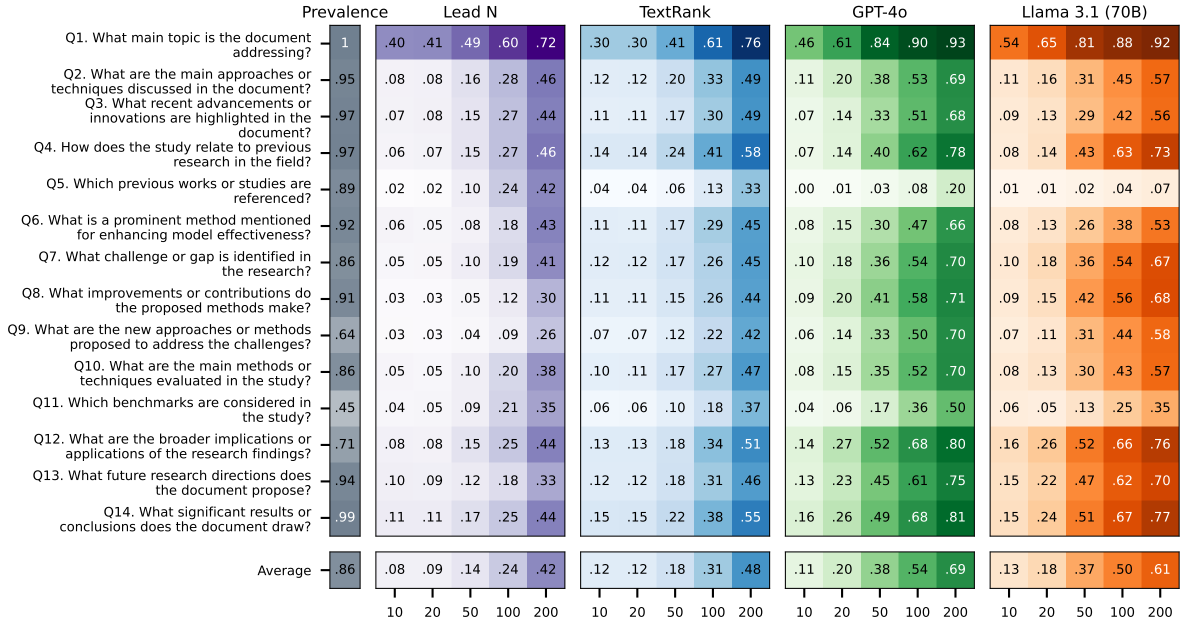


Figure 13: Corpus-level content salience map for *CL* summaries by four methods.

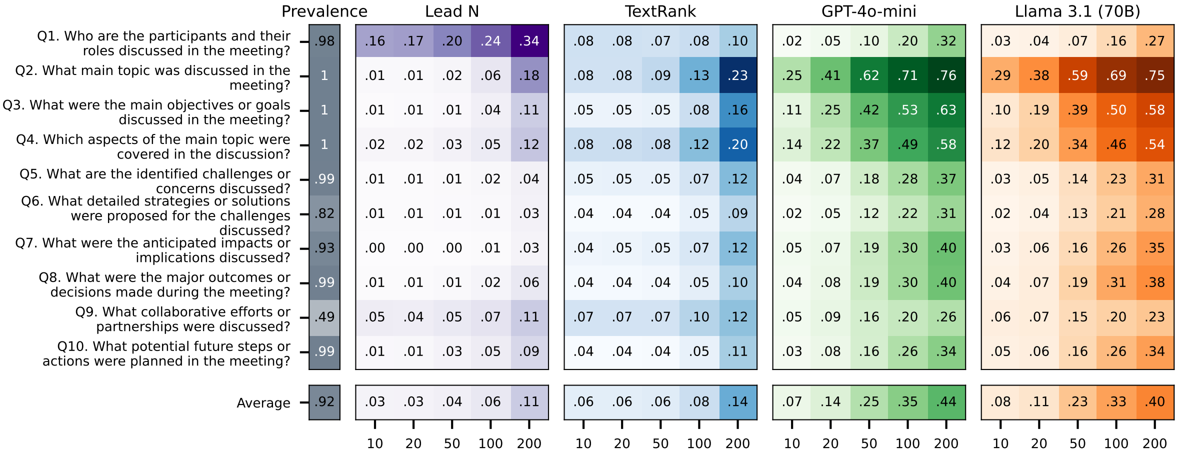


Figure 14: Corpus-level content salience map for *QMSum* summaries by four methods.

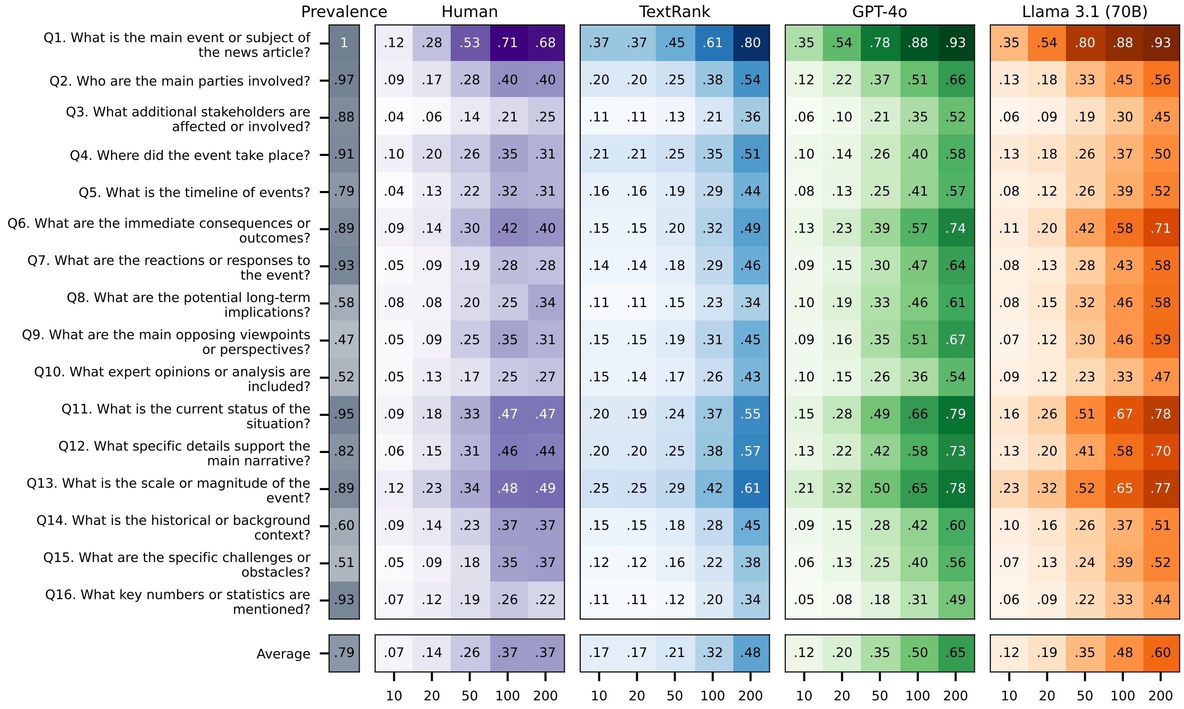
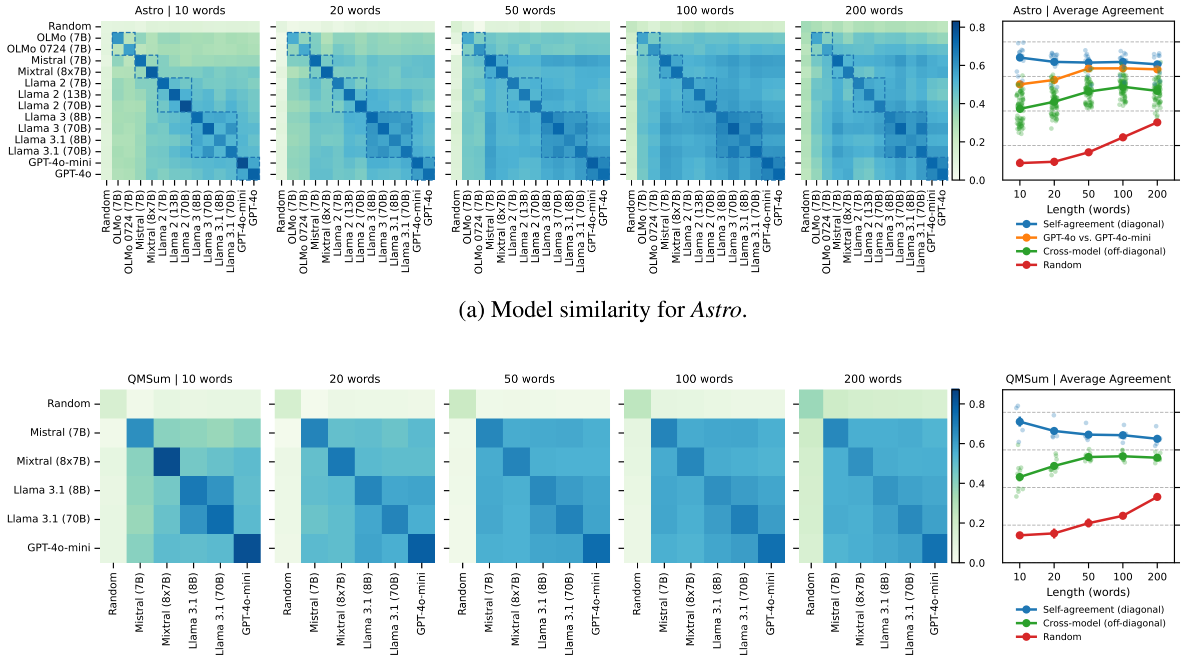


Figure 15: Corpus-level content salience map for *CNN/DM*. Salience scores for human summaries and model summaries are derived form different document samples, so cannot be directly compared.



(a) Model similarity for *Astro*.

(b) Model similarity for *QMSum*.

Figure 16: Do models share a similar notion of salience? Heatmaps show agreement on topic inclusion by summary length (Krippendorff's alpha calculated column-wise for the CSM pairs).

Question	$L_o/L_p/H$	🤖 **Rationale LLM-perceived**	👥 **Rationales Human**
		continued from Table 4	
(#4, RCT 👤) What were the comparative results between intervention and control groups?	3.2/4.6/5.0	Comparative results are key to understanding the effectiveness of the intervention.	($H_{\{1,2,3\}}$ = 5) Brief comparison between groups allows for readers to quickly assess statistical significance; Comparative results are one of the primary outputs reflecting effectiveness; essential for understanding the findings.
(#5, RCT 👤) What implications or future recommendations did the study suggest based on its findings?	4.6/3.6/2.0	Implications and recommendations are important for understanding the study's impact and future research directions.	(H_1 = 1) This is a job for the reader to decide for himself or herself; (H_2 = 3) helps the reader understand the paper. Useful for at the end of a summary. (H_3 = 2) While helpful, they are not necessarily always based on the opinion of the authors; would have the reader get to that conclusion themselves.
(#6, RCT 👤) What is the main outcome or effect observed?	4.6/5.0/5.0	The main outcome is critical to understanding the study's primary conclusion.	($H_{\{1,2,3\}}$ = 5) determines whether a paper can be used to influence decision making; primary information when reading summary
(#7, RCT 👤) How were the participants or subjects of the study selected and divided?	3.7/3.2/2.3	Participant selection and division are important for assessing the study's validity.	(H_1 = 2) This helps a reader determine external validity of a study; (H_2 = 1) Not necessary as fairly standardized; (H_3 = 4) Crucial to evaluate the study's validity and relevance
(#8, Astro ★) What limitations or considerations are noted by the study?	1.1/3.6/2.3	Limitations are crucial for interpreting the study's reliability and potential biases.	(H_1 = 3) important but best reserved for readers who choose to critically examine the study in its totality; (H_2 = 1) Typically addressed in the discussion section of a paper, but is not necessary for a summary; (H_3 = 3) would note if I had the space, however, I would also implore the readers to think about limitations themselves.
(#9, Astro ★) How do the findings relate to existing models or theories?	1.6/4.0/3.8	This is important for understanding the study's impact on current scientific understanding.	(H_1 = 4) I think this is important to describe the finding in context of the research question, initial hypothesis, or a theory. (H_2 = 3) This is important for the broader context of the study, however it is more important to present the study results in a clear way. Only after that is clear I'd compare to other studies.(H_3 = 5) This forms the main body of a discussion section. (H_4 = 3) Situating the findings within the literature by relating to existing models or theories is important.(H_3 = 4) Comparing with existing literature is very important for the discussion section.
(#10, Astro ★) What are the main findings of the study?	3.5/5.0/4.8	The main findings are the core of the study and must be included in any summary.	(H_1 = 5) key takeaway from the paper and should be included regardless of what the paper is about (H_2 = 5) most important information of the summary (H_3 = 4) The main findings should be briefly addressed in the summary of a discussion for the reader's quick follow-up (H_4 = 5) The main findings of the study, along with the main focus, form the two most important elements of an article summary. (H_5 = 5) Important to state the main findings and then discuss them in details.
(#11, Astro ★) What specific challenges or limitations does the study address or identify?	1.6/3.2/2.6	Understanding the challenges or limitations provides context for the study's reliability and areas for improvement.	(H_1 = 1) I most likely do not include challenges and limitations. These examples focused on the future needs not an existing open question. The focus will be on the findings in the context of a hypothesis, conjecture, or a theory. (H_2 = 1) Level of detail that a reader would need only if interested in full paper. Some challenges can be identified if the methods and scope of the paper are summarized clearly. (H_3 = 5) This forms the main body of a discussion section. (H_4 = 2) depends upon the significance of those challenges or limitations (H_5 = 4) Identify the limitations and challenges of the study is very important

Table 7: Example questions, salience scores by LLM-observed (L_o, rescaled to 1-5), LLM-perceived (L_p), humans (H) and summarized rationales.

G LLM Prompts

This section provides all prompts used throughout the experiments. Summarization (Listings 1 and 2), question generation (Listing 3), question answering (Listing 4), answer claim splitting (Listing 5), and introspection (Listing 6).

Listing 1: Summarization prompt

```
## Document
{{ text }}

## Instruction
Please summarize the above document. Use up to {{ length_target }} words. Respond exactly in following
JSON format:

{
    "summary": "(the {{ length_target }} words summary)"
}
```

Listing 2: Summarization prompt for meeting transcripts

```
## Meeting Transcript
{{ text }}

## Instruction
Please summarize the above meeting transcript. Use up to {{ length_target }} words. Respond exactly in
following JSON format:

{
    "summary": "(the {{ length_target }} words summary)"
}
```

Listing 3: Question generation prompt

```
Your task is to analyze summaries of different lengths within a given genre. Your goal is to create
question-answer pairs that capture the essence of information typically included in various summary
lengths. Below is the dataset where each document was summarized in 5 different lengths.

# Dataset

{% for document in documents %}
## Document {{ loop.index }}

{% for length, text in document.items() %}
### Summary {{ length }}
{{ text }}

{% endfor %}
{% endfor %}

# Genre
The genre of the documents:
{% if dataset == "rct" %}
    Randomized controlled trials (RCT) in the clinical domain.
{% elif dataset == "astro-ph" %}
    Discussion section in astrophysics papers.
{% elif dataset == "cs-cl" %}
    Related work section in NLP papers.
{% elif dataset == "qmsum" %}
    Meeting transcripts.
{% endif %}

# Task
Each text in the dataset has been summarized in 5 different lengths (in 10, 20, 50, 100, and 200 words).
Your task is to analyze the summaries and identify the types of information typically included at each
summary length. To do this, please proceed as follows:

1. Carefully read the summaries, paying attention to what information is included or omitted.
2. For each summary length, create a set of question-answer pairs that represent typical information
included at this length. The questions should be general enough to apply to many documents in this genre,
while the answers will naturally be different across documents.
```

```
Important guidelines:
- Ensure that your questions are relevant to the genre and capture information that would be commonly
found in texts of this type.
- It is really important that the questions are answerable with most documents in this genre, not just
with the ones presented here! To this end, state a prototypical answer to each question.
- The questions should be unique to each length. That means, do not repeat a question if it is already
sufficiently covered at the shorter length.
- Start with question words (What, How, Why, Which) rather than 'Can you'
- Make the topic the grammatical subject of a question.
- Keep the questions concise and focused.
- Create at least 3-5 questions for each summary length.

Structure your response as a valid json object with the following format:

{
    ”questions_10_words”: [
        {
            ”question”: ””,
            ”example_answer”: ””,
        }
    ],

    [... truncated for brevity ...]
}
```

Listing 4: Question answering prompt

```
Answer the following question given the text. If the question cannot be answered with the text, reply ”no
answer”.

## Text
{{ text }}

## Question
{{ question }}

First, carefully read and analyze both the text and the question. Then provide the answer. Please follow
these guidelines:
- If the question cannot be answered, reply with ”no answer”
- Use only information explicitly stated in or directly implied by the text
- Do not include any external knowledge or personal opinions
- Aim for concise answers that include all important points relevant to the question

Please use this format for your response:
Question: [restate the question exactly]
Answer: [the answer based on the text or ”no answer”]
```

Listing 5: Claim splitting prompt

```
You split sentences into a list of facts that we explicitly know from the sentence. Make each fact as
atomic as possible.

Sentence: Protein-rich nutrition is necessary for wound healing after surgery.
Output:
[
    ”Protein-rich nutrition is necessary for wound healing.”,
    ”Wound healing occurs after surgery.”
]

Sentence: In this study, the benefit of preoperative nutritional support was investigated for non-small
cell lung cancer patients who underwent anatomic resection.
Output:
[
    ”The study investigated the benefit of preoperative nutritional support.”,
    ”The study considers patients with non-small cell lung cancer.”,
    ”The study considers patients who underwent anatomic resection.”
]

[... 12 few-shot examples truncated for brevity ... ]

Here is a new sentence. Please split it into a list of facts that we explicitly know from the sentence.
Make each fact as atomic as possible. Output the facts as Python list. Only output the list, nothing more.
```

```
Sentence: {sent}
Output:
```

Listing 6: Introspection prompt

```
## Task
{% if dataset == ”rct” %}
    You are a research expert in randomized controlled trials (RCTs). Imagine you are asked to summarize a
paper describing the results of an RCT 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.
{% elif dataset == ”astro-ph” %}
    You are a research expert in astrophysics. Imagine you are asked to summarize the discussion section
of an astrophysics 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.
{% elif dataset == ”cs-cl” %}
    You are a research expert in natural language processing (NLP). Imagine you are asked to summarize the
related work section of an NLP 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.
{% elif dataset == ”qmsum” %}
    You are an expert in communications and meetings. Imagine you are asked to summarize a meeting
transcript (e.g., research group meetings) for a typical reader of these texts. The summary should provide
enough context to stand alone, since the reader will only see your summary and not the full meeting
transcript.
{% endif %}

The summary length is constrained, requiring you to think about what content to prioritize. Ask yourself:
what are some key questions you want the summary to answer? Your task is to rate the relative importance
of a list of questions that the summary could answer.

## Questions
Here is the list of questions you should evaluate.

{% for question in questions %}
{{ loop.index }}. {{ question }}
{% endfor %}

## Rating
Please use the following scale, going from least important to most important.

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.

For each rating, please provide a brief (1-sentence) rationale explaining your decision or highlighting
any considerations or uncertainties you had.

Important considerations:
- Use the full scale (1-5) to express relative importance.
- Remember that space in the summary is limited, so not everything can be included, and you CANNOT rate
all questions as 5.
- Make sure to rate all given questions.

Please respond as a valid JSON list with following format:

[
    {
        ”id”: ”[the question number]”,
        ”question”: ”[repeat the exact question]”,
        ”rationale”: ”[your one-sentence rationale for the rating]”,
        ”rating”: ”[your numeric rating, 1-5]”
    }
]
```

H Question Salience Annotation Guidelines

Motivation. When summarizing long texts, we must consciously decide what information to include or exclude from a summary. These decisions are grounded in a notion of information salience, or how important we consider the information for our intended audience. We study this phenomenon in the context of automatic text summarization systems. Specifically, we aim to understand how well these systems replicate the judgments of domain experts regarding what information is most relevant.

Task. Imagine you are asked to **summarize a paper describing the results of a randomized controlled trial (RCT)** 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. Furthermore, the summary length is constrained, requiring you to think about what content to prioritize. In this study, we frame content as questions that a summary could answer.

Ask yourself: **What are some key questions you want the summary to answer?** Your task is to rate the relative importance of a list of questions on the following 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.

Rationale. For each rating, please provide a brief (1-sentence) rationale explaining your decision or highlighting any considerations or uncertainties.

Example answers. To give you a feeling for the kind of content a question might elicit, all questions have an illustrative answer sourced from a randomly chosen document (= RCT paper). Please keep the following in mind:

- *Answer length* does not determine the question's importance.
- *Phrasing and selection.* The precise answer phrasing can be different in the summary, and not all answer content must appear in the summary.
- *Overlap.* Some questions may elicit overlapping answers. Therefore, focus on the essence of each question. Remember that in an actual summary, overlapping answer information would only be stated once, so don't worry about it (see below).
- *Relevance.* The questions are answerable with most documents in this genre. Do your rating on the assumption that the document talks about this information.

Suggested process.

1. Read all questions first.
2. Identify questions that seem most/least important, and rate these as "anchor points."
3. Then, rate the remaining questions.

Finally, there are no right or wrong ratings. Use your best judgment and intuition. Thank you for participating!

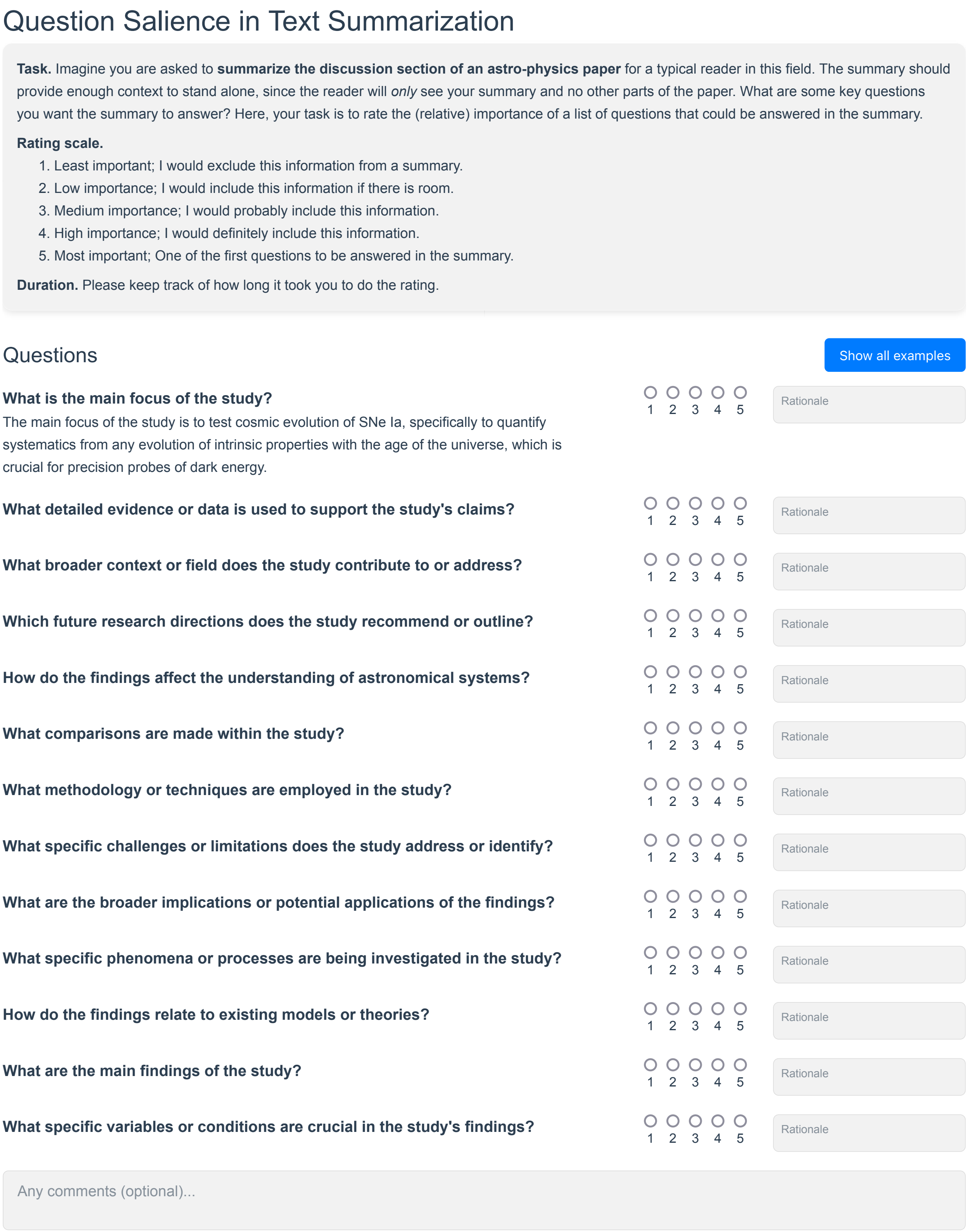
Appendix: Example of overlapping answers. Consider questions Q1-Q3 below. Each question asks for a distinct unit of information, but the answer of Q3 overlaps with the answer of Q1 and Q2. The overlapping information is highlighted in orange while the *essence of the question* is highlighted in green. Base your rating on the essence of the question.

Example of overlapping answers

Q1. What was the study design or setting of the trial? This trial is a multicentre, randomized, double-blind, phase 3 study.

Q2. What specific treatments were compared in the study? DBPR108 100 mg, sitagliptin 100 mg, and placebo.

Q3. How were the participants or subjects of the study selected and divided? In this multicentre, randomized, double-blind, phase 3 study, adult patients with type 2 diabetes were randomly assigned to receive either DBPR108 100mg, sitagliptin 100mg, or placebo once daily. A total of 766 patients were enrolled and divided into three groups: DBPR108 100mg (n=462), sitagliptin 100mg (n=152), or placebo (n=152).



Question Salience in Text Summarization

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? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

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.

What detailed evidence or data is used to support the study's claims? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

What broader context or field does the study contribute to or address? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

Which future research directions does the study recommend or outline? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

How do the findings affect the understanding of astronomical systems? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

What comparisons are made within the study? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

What methodology or techniques are employed in the study? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

What specific challenges or limitations does the study address or identify? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

What are the broader implications or potential applications of the findings? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

What specific phenomena or processes are being investigated in the study? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

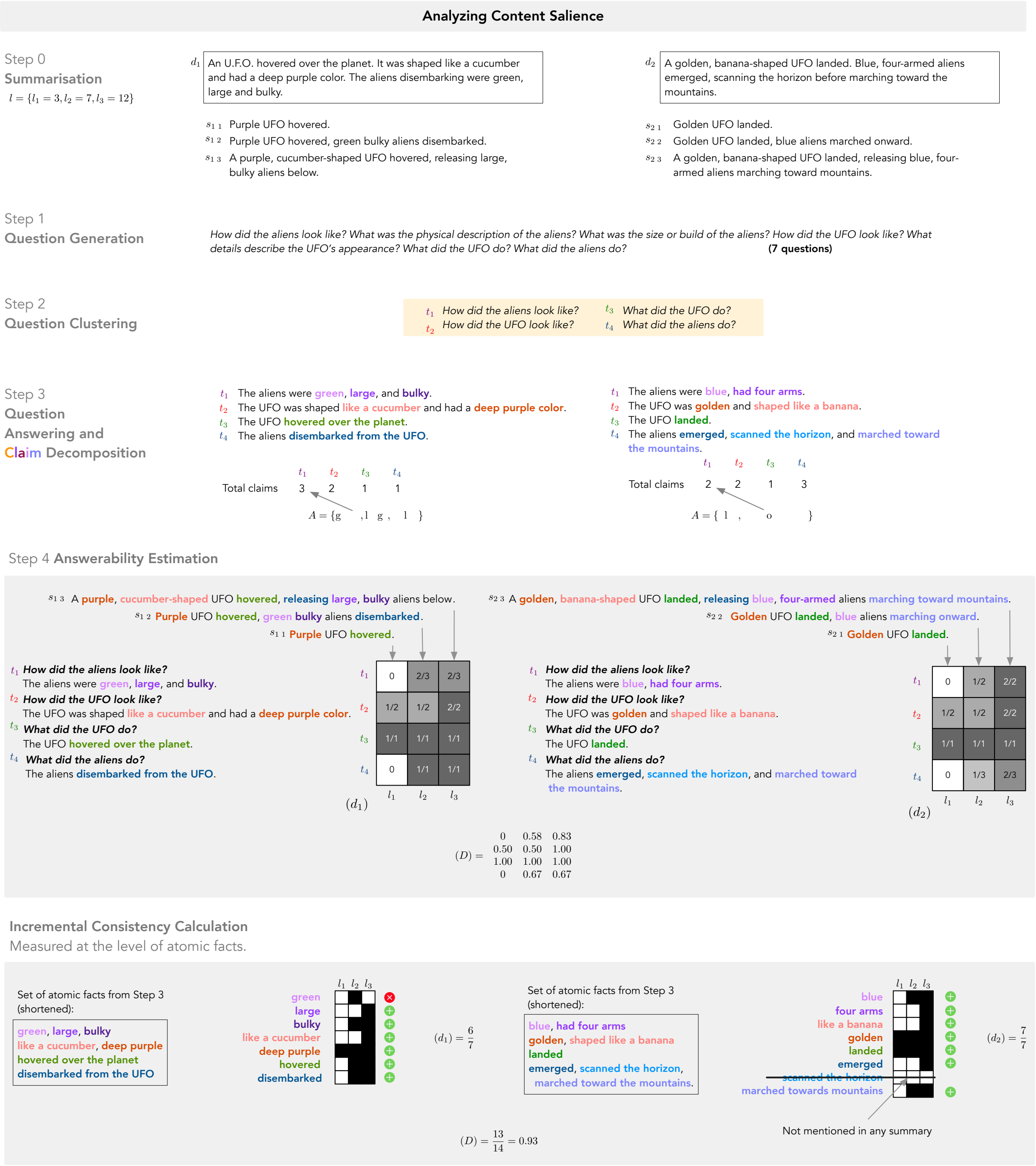
How do the findings relate to existing models or theories? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

What are the main findings of the study? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

What specific variables or conditions are crucial in the study's findings? ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Rationale

Any comments (optional)...

Figure 17: Interface for question salience annotation. Each question can be expanded to show an illustrative answer sourced from a randomly chosen document. The questions shown here are for the *Astro* dataset.



Analyzing Content Salience

Step 0 **Summarisation**

$l = \{l_1 = 3, l_2 = 7, l_3 = 12\}$

d_1 An U.F.O. hovered over the planet. It was shaped like a cucumber and had a deep purple color. The aliens disembarking were green, large and bulky.

$s_{1\,1}$ Purple UFO hovered.
$s_{1\,2}$ Purple UFO hovered, green bulky aliens disembarked.
$s_{1\,3}$ A purple, cucumber-shaped UFO hovered, releasing large, bulky aliens below.

d_2 A golden, banana-shaped UFO landed. Blue, four-armed aliens emerged, scanning the horizon before marching toward the mountains.

$s_{2\,1}$ Golden UFO landed.
$s_{2\,2}$ Golden UFO landed, blue aliens marched onward.
$s_{2\,3}$ A golden, banana-shaped UFO landed, releasing blue, four-armed aliens marching toward mountains.

Step 1 **Question Generation**

How did the aliens look like? What was the physical description of the aliens? What was the size or build of the aliens? How did the UFO look like? What details describe the UFO's appearance? What did the UFO do? What did the aliens do? **(7 questions)**

Step 2 **Question Clustering**

t_1 *How did the aliens look like?*
t_2 *How did the UFO look like?*
t_3 *What did the UFO do?*
t_4 *What did the aliens do?*

Step 3 **Question Answering and Claim Decomposition**

t_1 The aliens were green, large, and bulky.
t_2 The UFO was shaped like a cucumber and had a deep purple color.
t_3 The UFO hovered over the planet.
t_4 The aliens disembarked from the UFO.

	t_1	t_2	t_3	t_4
Total claims	3	2	1	1

$A = \{g \;\; ,l \;\; g \;,\;\; l \;\; \}$

t_1 The aliens were blue, had four arms.
t_2 The UFO was golden and shaped like a banana.
t_3 The UFO landed.
t_4 The aliens emerged, scanned the horizon, and marched toward the mountains.

	t_1	t_2	t_3	t_4
Total claims	2	2	1	3

$A = \{ \; l \;\; , \quad o \qquad \}$

Step 4 **Answerability Estimation**

$s_{1\,3}$ A purple, cucumber-shaped UFO hovered, releasing large, bulky aliens below.
$s_{1\,2}$ Purple UFO hovered, green bulky aliens disembarked.
$s_{1\,1}$ Purple UFO hovered.

t_1 ***How did the aliens look like?***
The aliens were green, large, and bulky.
t_2 ***How did the UFO look like?***
The UFO was shaped like a cucumber and had a deep purple color.
t_3 ***What did the UFO do?***
The UFO hovered over the planet.
t_4 ***What did the aliens do?***
The aliens disembarked from the UFO.

(d_1)	l_1	l_2	l_3
t_1	0	2/3	2/3
t_2	1/2	1/2	2/2
t_3	1/1	1/1	1/1
t_4	0	1/1	1/1

$s_{2\,3}$ A golden, banana-shaped UFO landed, releasing blue, four-armed aliens marching toward mountains.
$s_{2\,2}$ Golden UFO landed, blue aliens marching onward.
$s_{2\,1}$ Golden UFO landed.

t_1 ***How did the aliens look like?***
The aliens were blue, had four arms.
t_2 ***How did the UFO look like?***
The UFO was golden and shaped like a banana.
t_3 ***What did the UFO do?***
The UFO landed.
t_4 ***What did the aliens do?***
The aliens emerged, scanned the horizon, and marched toward the mountains.

(d_2)	l_1	l_2	l_3
t_1	0	1/2	2/2
t_2	1/2	1/2	2/2
t_3	1/1	1/1	1/1
t_4	0	1/3	2/3

$$(D) = \begin{matrix} 0 & 0.58 & 0.83 \\ 0.50 & 0.50 & 1.00 \\ 1.00 & 1.00 & 1.00 \\ 0 & 0.67 & 0.67 \end{matrix}$$

Incremental Consistency Calculation
Measured at the level of atomic facts.

Set of atomic facts from Step 3 (shortened):

green, large, bulky
like a cucumber, deep purple
hovered over the planet
disembarked from the UFO

$(d_1) = \frac{6}{7}$

Set of atomic facts from Step 3 (shortened):

blue, had four arms
golden, shaped like a banana
landed
emerged, scanned the horizon,
marched toward the mountains.

$(d_2) = \frac{7}{7}$

$(D) = \frac{13}{14} = 0.93$

Figure 18: Fully worked example of the question-based content analysis. Two documents in a fictional domain are each summarized at three lengths. Afterwards Steps 1 – 4 are analogous to § 2.2. Summary claims are color-coded.